Proactive Quality Control based on Ensemble Forecast Sensitivity to Observations

Daisuke Hotta$^{1,2}$
Eugenia Kalnay$^1$, Yoichiro Ota$^2$

$^1$ University of Maryland
$^2$ Japan Meteorological Agency

July 16th, 2014 EMC Seminar
Outline

1. EFSO (Ensemble Forecast Sensitivity to Observations) and “Proactive QC”

2. EFSR (Ensemble Forecast Sensitivity to Observation Error Covariance matrix $R$) and tuning of $R$

3. Future Directions: Operational Applications

(Appendix: semi-implicit Lorenz $N$-cycle scheme)
Part 1:

**EFSO** (Ensemble Forecast Sensitivity to Observations)

and “Proactive QC”
Motivation:
The NCEP “forecast skill dropout” problem

- NCEP’s 5-day Forecast skill is generally very high (~ 0.9 level)
- However, it occasionally drops to a low level (= “dropout”)
- In some cases, all NWP centers suffer.
- But in some cases, NCEP does suffer while ECMWF does not.
Motivation: The NCEP “forecast skill dropout” problem

- “Culprit” is not the model but “bad observations” (or inability of DA system to properly assimilate them)

→ How can we detect those “flawed” observations?

From Kumar et al. (2009)
EFSO: Ensemble Forecast Sensitivity to Observations

- Quantifies **how much each observation improved/degraded the forecast**
- First invented for a variational DA-system using the *adjoint method* by Langland and Baker (2004)
- Liu and Kalnay (2008) adapted it to LETKF (*no adjoint*)
- Kalnay et. al (2012) gave an improved, simpler formulation
  - The new formulation is
    - more accurate
    - simpler and easier to implement
    - applicable to any formulation of EnKF
- Ota et al. (2013) successfully implemented the new EFSO into the NCEP’s operational GFS system

\[ \Delta e^2 = e^T_{t|0} C e_{t|0} - e^T_{t|-6} C e_{t|-6} \]

\[ \approx \frac{1}{K-1} \delta y_0^T R^{-1} Y_0^a X^T_{t|0} C (e_{t|0} + e_{t|-6}) \]

Reduction of forecast error by the assimilation of obs.

From Kalnay et al (2012)
Ota et al. (2013): Identification of “flawed” observations by 24-hour EFSO

Table 3. List of local 24-hour forecast failure cases (initial time from 00 UTC, 8 January 2012, to 18 UTC 7 February 2012)

<table>
<thead>
<tr>
<th>Initial</th>
<th>Area</th>
<th>Size</th>
<th>Rate</th>
<th>N</th>
<th>Denied observation (denied number/total number)</th>
<th>Change (estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>06 UTC JAN 12</td>
<td>50N–80N 145E–175W</td>
<td>1.99</td>
<td>1.36</td>
<td>5</td>
<td>AMSUA ch4, 5, 6 (2735/125063)</td>
<td>−8.7% (−19.5%)</td>
</tr>
<tr>
<td>00 UTC JAN 16</td>
<td>30N–60N 20W–0</td>
<td>2.71</td>
<td>1.35</td>
<td>6</td>
<td>GPSRO 600–950 hPa (50/4918)</td>
<td>−0.6% (−4.3%)</td>
</tr>
<tr>
<td>18 UTC JAN 27</td>
<td>30S–0 105E–120E</td>
<td>2.40</td>
<td>1.21</td>
<td>1</td>
<td>AIRS (19 908/670 041)</td>
<td>−0.2% (−6.0%)</td>
</tr>
<tr>
<td>00 UTC JAN 30</td>
<td>70S–40S 165E–165W</td>
<td>2.00</td>
<td>1.25</td>
<td>6</td>
<td>AMSUA ch1, 3, 4, 5, 15 (3822/164934)</td>
<td>−4.7% (−12.8%)</td>
</tr>
<tr>
<td>06 UTC FEB 2</td>
<td>50N–80N 150W–110W</td>
<td>3.01</td>
<td>1.22</td>
<td>5</td>
<td>GPSRO 250–400 hPa, 600–850 hPa (407/13092)</td>
<td>−11.7% (−8.7%)</td>
</tr>
<tr>
<td>06 UTC FEB 4</td>
<td>30N–60N 150W–130W</td>
<td>1.81</td>
<td>1.26</td>
<td>3</td>
<td>IASI (57 950/1 177 256), HIRS ch3, 4, 9, 11, 12, 14, 15 (785/73 419), Aircraft 950 hPa ~, 125–600 hPa (5794/100 896)</td>
<td>−25.5% (−81.6%)</td>
</tr>
<tr>
<td>18 UTC FEB 6</td>
<td>60N–90N 40E–100E</td>
<td>1.71</td>
<td>1.38</td>
<td>2</td>
<td>MODIS_Wind (10 970/43 452)</td>
<td>−28.4% (−47.7%)</td>
</tr>
</tbody>
</table>

- Identified 7 cases of potential “regional forecast skill dropouts”
- Rerun the analyses and forecasts without using “flawed” obs. identified by 24-hour EFSO
- The forecast errors were substantially reduced.
“Proactive QC”: Proposed Algorithm

Suppose we wish to identify and delete “flawed” obs. at 00h.

① Run regular DA cycle from -06h to 00h.
② Run regular DA cycle from 00h to 06h.
③ Detect “regional dropouts” using the information available from ① and ②.
④ Perform 6-hour EFSO to identify “flawed” obs. at 00h.
⑤ If “flawed” obs. are identified, repeat 00h analysis without using the detected “flawed” obs.
Key questions to be addressed in order for the Proactive QC to work

• Are 6 hours long enough for detecting “flawed observations”?
  – Forecast errors are computed as Forecast-minus-Analysis
  – When compared to the errors of very short-term forecast, analysis errors might not be small.
    → Estimation of forecast errors becomes more difficult.

• What is the best criterion for rejection of observations?
  Rejecting too many observations might lead to forecast degradation, but too few would make little difference.
    → How to strike the best balance?

• Does rejection of those observations really improve analysis and forecast?
Experiments with quasi-operational NCEP’s GFS/GSI system

Experimental Set-up

(Implemented on top of GFS/hybrid GSI ported to JCSDA’s S4 by Dr. Jim Jung)

- **Forecast Model**: NCEP’s GFS model
- **Resolution**: half of the operational:
  - T254L64 (deterministic), T126L64 (ensemble)
- **DA system**: hybrid GSI (as in the operational), but EnKF part replaced by LETKF
- **Observations**: same as the NCEP operational system
- **Period**: 34 days (Jan – Feb, 2012)
- **LETKF**: Covariance localization and inflation (same as the operational)
- **EFSO**:
  - **Localization**: same as LETKF + moving localization of Ota et al. (2013)
  - **Error norm**: Moist total energy norm
EFSO’s sensitivity to forecast lead time

(1) Time average

Average net observation impact for each observation type

- Lead-Time: 6 hrs.
- Lead-Time: 12 hrs.
- Lead-Time: 24 hrs.

- EFSO results are not very sensitive to the choice of evaluation lead time.
EFSO’s sensitivity to forecast lead time

(2) Individual Cases
Example: MODIS wind near the North Pole on Feb 06 18UTC, 2012

Geographical distribution of EFSO from each. obs.

Red: negative impact; Blue: positive impact
the size proportional to the magnitude

Lead-Time: 6 hrs.

Lead-Time: 24 hrs.

• Again, EFSO results are not very sensitive to the choice of evaluation lead time, even for individual cases.

• → 6 hours are long enough for detection of “flawed” observations.
Key questions to be answered

• Are 6 hours long enough for detecting “flawed observations”?
   → Yes. 6-hr EFSO is equally capable of detecting “flawed” obs. as 24-hr EFSO.

• What is the best criterion for rejection of observations?

• Does rejection of those observations really improve analysis and forecast?
Data Denial Experiments
Case study (1): MODIS case
2012-Feb-06-18Z, [60°N—90°N] x [40°E—100°E]

Net EFSO Impact by obs. types
measured with moist total energy norm

Units: J kg\(^{-1}\)

MODIS →

EFSO impact from each MODIS wind observation

- MODIS wind identified as “flawed” (i.e., with net negative impact).
- There are both helpful and harmful observations.
- How can we decide which / how many obs. should be denied?
Data Denial Experiments
Selection of the Obs. to be denied
→ Try four criteria, perform data denial for each

Distribution of EFSO values from each observation
Example: MODIS wind near the North Pole

- 0.01: reduces fcst. error → positive impact
- 0.001: increases fcst. error → negative impact

Units: J kg⁻¹

- allobs
- allneg
- one-sigma
- mean + σ
- netzero
How many obs. should we reject?

Case study (1): MODIS case

2012-Feb-06-18Z, [60°N—90°N] x [40°E—100°E]

Relative 24-hr fcst. improvement: $= \frac{e_{\text{beforeQC}}^f - e_{\text{afterQC}}^f}{e_{\text{beforeQC}}^f} \times 100 \%$

Data selection based on 6-hour EFSO

- **allobs**: overall improvement, but with several areas with degradation
- **allneg**: enhanced improvement, reduced degradation
- **one-sigma & netzero**: less improvement, but with further reduced degradation
### Data Denial Experiment

Summary of the relative 24-hr forecast improvement for 20 cases (20x4x2=160 experiments)

<table>
<thead>
<tr>
<th>Case #</th>
<th>max.imp.</th>
<th>max.deg.</th>
<th>avg.imp.</th>
<th>max.imp.</th>
<th>max.deg.</th>
<th>avg.imp.</th>
<th>max.imp.</th>
<th>max.deg.</th>
<th>avg.imp.</th>
<th>max.imp.</th>
<th>max.deg.</th>
<th>avg.imp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12%</td>
<td>-9%</td>
<td>0.0%</td>
<td>12%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>20%</td>
<td>-5%</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.1%</td>
</tr>
<tr>
<td>2</td>
<td>14%</td>
<td>-5%</td>
<td>-0.1%</td>
<td>14%</td>
<td>-5%</td>
<td>-0.1%</td>
<td>2%</td>
<td>0%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>3</td>
<td>13%</td>
<td>7%</td>
<td>0.0%</td>
<td>13%</td>
<td>7%</td>
<td>0.0%</td>
<td>2%</td>
<td>0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>4</td>
<td>25%</td>
<td>-5%</td>
<td>-0.6%</td>
<td>25%</td>
<td>-5%</td>
<td>-0.6%</td>
<td>0%</td>
<td>0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>5</td>
<td>15%</td>
<td>19%</td>
<td>-0.2%</td>
<td>15%</td>
<td>19%</td>
<td>-0.2%</td>
<td>1%</td>
<td>1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>6</td>
<td>9%</td>
<td>15%</td>
<td>0.0%</td>
<td>9%</td>
<td>15%</td>
<td>0.0%</td>
<td>2%</td>
<td>2%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>7</td>
<td>17%</td>
<td>13%</td>
<td>-0.0%</td>
<td>17%</td>
<td>13%</td>
<td>-0.0%</td>
<td>0%</td>
<td>0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>8</td>
<td>41%</td>
<td>41%</td>
<td>0.9%</td>
<td>41%</td>
<td>41%</td>
<td>0.9%</td>
<td>0%</td>
<td>0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>9</td>
<td>7%</td>
<td>8%</td>
<td>-0.6%</td>
<td>7%</td>
<td>8%</td>
<td>-0.6%</td>
<td>4%</td>
<td>4%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>10</td>
<td>25%</td>
<td>19%</td>
<td>1.1%</td>
<td>25%</td>
<td>19%</td>
<td>1.1%</td>
<td>2%</td>
<td>2%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

**Table 6.4: Relative improvement or degradation of 24-hour forecast by the denial of Data Denial Experiment**

**Case**

- **max.imp.** shows the maximum forecast improvement for each case.
- **max.deg.** shows the maximum degree of degradation for each case.
- **avg.imp.** shows the average forecast improvement for each case.

**Net Zero (netzero)**

- This column represents the net zero improvement or degradation compared to the baseline.

**Hemispheric-scale**

- Forecast error reduced in 18 out of 20 cases.
- Local improvement over 30% in 7 cases

- **With allneg:**
  - Hemispheric-scale forecast error reduced in 18 out of 20 cases.
  - Local improvement over 30% in 7 cases
Data Denial Experiment
Summary of the results for 20 cases

• Data selection based on 6-hour EFSO:
  – **allobs**: improvement mixed with degradation
  – **allneg**: enhanced improvement, reduced degradation
    Hemispheric-scale forecast error reduced in 18 out of 20 cases.
    Local improvement over 30% in 7 cases.
  – **one-sigma & netzero**: diminished improvement, but with further reduced degradation
  – For all of the 7 most successful cases, MODIS wind was identified as “flawed.”

• Data selection based on 24-hour EFSO: Similar to 6-hour EFSO, but with less improvements.
Key questions to be answered

• Are 6 hours long enough for detecting “flawed observations”?
  • Yes. 6-hr EFSO is equally capable of detecting “flawed” obs. as 24-hr EFSO is.

• What is the best criterion for rejection of observations?
  • A matter of trade-off: if some degradation is tolerable, “allneg” should be favorable; else “one-sigma” or “netzero” should be used.

• Does rejection of those observations really improve analysis and forecast?
  • Yes, with >30% local improvement in 7 out of 20 cases.
Summary for **Proactive QC**

- “Flawed” observations that potentially lead to forecast skill dropouts can be detected by EFSO diagnostics **after only 6 hours** from the analysis.
- **Proactive QC** does improve forecast and analysis.
- Proactive QC is **innovative**:
  - The first fully flow-dependent QC
  - based on whether observations actually improve/ degrade forecast
Part 2:  
**EFSR** (Ensemble Forecast Sensitivity to Observation Error Covariance matrix $R$) and **Tuning of $R$**
Motivation

• Data Assimilation combines information from background and observations with an “optimal weight.”

• The “optimal weight” is determined based on the background -and observation- error covariances $B$ and $R$.

• In EnKF, $B (=P^b)$ is dynamically estimated, but $R$ is still an external parameter.
  – Truth is unknown. $\rightarrow$ True $R$ is also unknown.
  – NWP centers specify it empirically and subjectively.

• $\rightarrow$ We need a systematic method for tuning $R$. 
EFSR Formulation

- We can formulate an *ensemble* version based on EFSO by Kalnay et al. (2012):

$$\left[ \frac{\partial e}{\partial R} \right]_{ij} \approx \frac{\partial e}{\partial y_i} z_j \approx -\frac{1}{K-1} \left[ R^{-1} Y_0^a X_T^f T C (e_{t|0} + e_{t|-6}) \right]_i \left[ R^{-1} \delta y^{o,a} \right]_j$$

- We know whether fcst. will be improved or degraded by the increase or decrease of $R$.
  - We can optimize $R$. 
EFSR: Experiments

• Perfect-model experiment with Lorenz ’96 system
  – Run two DA cycles, one with incorrect $R$, the other with correct $R$
  – Perform EFSR to the two experiments. Examine if EFSR can detect mis-specification of $R$.

• Real NWP system experiment & Tuning of $R$
  – Diagnose forecast $R$-sensitivity for each observation type by EFSR.
  – Tune $R$ based on EFSR and run the DA cycle again. Examine if the tuning improves the EFSO impacts of the tuned observation types.
Perfect-model Experiment: Experimental Setup

- **Model:** Lorenz ’96 model with $N=40$ and $F=8.0$

\[ \frac{dx_j}{dt} = x_j(x_{j+1} - x_{j-2}) - x_j + F. \]

- **DA method:** 40 member LETKF, no localization
- **EFSR:** no localization
- **Observations:** available at every grid point.
- **Specification of $R$:**

<table>
<thead>
<tr>
<th>Name</th>
<th>True obs error variance</th>
<th>Prescribed error variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPIKE</td>
<td>$\sigma_j^{o,true^2} = \begin{cases} 0.8^2 &amp; j = 11 \ 0.2^2 &amp; j \neq 11 \end{cases}$</td>
<td>$\sigma_j^{o^2} = 0.2^2$ everywhere</td>
</tr>
<tr>
<td>STAGGERED</td>
<td>$\sigma_j^{o,true^2} = \begin{cases} 0.1^2 &amp; j: \text{odd} \ 0.3^2 &amp; j: \text{even} \end{cases}$</td>
<td>$\sigma_j^{o^2} = 0.2^2$ everywhere</td>
</tr>
<tr>
<td>LAND-OCEAN</td>
<td>$\sigma_j^{o,true^2} = \begin{cases} 0.3^2 &amp; 1 \leq j \leq 20 \text{ (&quot;land&quot;)} \ 0.1^2 &amp; 21 \leq j \leq 40 \text{ (&quot;ocean&quot;)} \end{cases}$</td>
<td>$\sigma_j^{o^2} = 0.2^2$ everywhere</td>
</tr>
</tbody>
</table>

- Erroneous obs. variance only at the 11-th grid pt.
- DA system assumes constant $R$ for all grid pts.

Design is inspired by Liu and Kalnay (2008)
Perfect-model Experiment: Result (SPIKE experiment)

24-hr. EFSR sensitivity

Grid Number

0 5 10 15 20 25 30 35 40

\( \frac{de}{d\sigma^2} \)

1 0 -1 -2 -3 -4

incorrect-R correct-R

Negative sensitivity: forecast error can be reduced by increasing \( R \)  
\( \Rightarrow \) \( R \) is too small

- For “incorrect-\( R \),” EFSR detects the mis-specification of \( R \) at the 11\(^{th}\) grid point.  
  \( \Rightarrow \) We can detect mis-specified \( R \)
- For “correct-\( R \),” EFSR diagnoses almost-zero sensitivity.  
  \( \Rightarrow \) No “false alerts”
EFSR for GFS / GSI-LETKF hybrid

- Aircraft, Radiosonde and AMSU-A: large positive sensitivity
- MODIS wind: negative sensitivity
- **→ Tuning experiment:**
  - Aircraft, Radiosonde and AMSU-A: reduce R by 0.9
  - MODIS wind: increase R by 1.1

*R is too large (→ should reduce R)*
Tuning Experiment: Result

EFSO before/after tuning of R

• Aircraft, Radiosonde, AMSU-A:
  • significant improvement of EFSO-impact (as expected)

• MODIS wind:
  • No improvement in EFSO (contrary to expectation)

Why no improvement in MODIS?

• MODIS had “flawed” obs. along with “helpful” obs.
• The “flawed” obs. might have resulted in incorrect estimation of EFSR.
Excluding cases where MODIS wind had negative impact

- MODIS wind exhibited several negatively-impacting cases.
- Exclude negative cases
- \(\rightarrow\) EFSR for MODIS becomes neutral
- \(\rightarrow\) Consistent with the result of tuning experiment

Lesson:
- Before performing EFSR, we should remove “bad” obs.
Summary for EFSR

• EFSR gives information on whether we should increase/reduce prescribed observation error covariance $R$.

• Tuning of $R$ based on this diagnostics improves the EFSO.

• $\Rightarrow$ EFSR can be used to systematically optimize $R$. 
Future Directions:
Future plans

Immediate future:
• Implementation of Proactive QC into the real operational system
  – Can the operational system
    1. wait for 6 hours?
    2. afford to do analysis again?

Long-term Future Directions:
• Applications of EFSO and EFSR
  – Collaboration with instrument developers
  – Acceleration of development for assimilation of new observing systems
Implementation to the real operational system

(1) Can we wait for 6 hours?

Idea: Exploit the time lag between “early analysis” and “cycle (final) analysis”

(suggested by Dr. John Derber, 2013)

cycle (final) analysis: maintains analysis-forecast cycle
early analysis: provides initial condition for extended forecast
Implementation to the real operational system (1) we don’t need to wait 6 hours!

Idea: Exploit the time lag between “early analysis” and “cycle (final) analysis”

(suggested by Dr. John Derber, 2013)

**cycle (final) analysis**: maintains analysis-forecast cycle

**early analysis**: provides initial condition for extended forecast
Implementation to the real operational system (2) can we afford to do analysis twice?

Idea: Use approximated analysis rather than doing analysis again:

- Using the approximation to Kalman gain:

$$K \approx \frac{1}{K-1} X^a_0 X^a_0 H^T R^{-1} \approx \frac{1}{K-1} X^a_0 Y^a_0 H^T R^{-1}$$

the change in analysis by the denial of observations can be approximated by:

$$\vec{x}_{0,\text{deny}}^a - \vec{x}_0^a \approx -K \delta \vec{y}_{0,\text{deny}}^\text{ob} \approx -\frac{1}{K-1} X^a_0 Y^a_0 H^T R^{-1} \delta \vec{y}_{0,\text{deny}}^\text{ob}$$

- As inexpensive as EFSO.

→ No need to repeat analysis

→ Can minimize the time delay
Application of EFSO:
(1) Collaboration with instrument developers

• In our experiments, in several cases, MODIS wind showed large negative impacts that caused “regional dropouts.”

• EFSO can be used to build a database of “flawed” observations along with their relevant metadata.
  – Provide such database to instrument developers so that they can fix problems with their algorithms.
  – Collaborate with instrument developers to determine which metadata would be helpful to them.
Application of EFSO:
(2) Acceleration of development for assimilation of new observing systems

• Traditional approach: compare
  – Test: with a new observing system
  – Control: without a new observing system

• Difficulty with this approach:
  – Signals from a new observing system is obscured by the many observations that are already assimilated in Control.
    → Hard to obtain statistically significant results

→ **EFSO-based data selection will enable efficient determination of an optimal way to assimilate new observing systems.**

• An optimal specification of $R$ is also a difficult issue for assimilation of new observing systems.
  → **Our EFSR diagnostics should provide useful guidance.**
Conclusions

• 6-hour EFSO can successfully identify “flawed” observations.
• Rejecting them (Proactive QC) does improve the analysis and forecast.
• Database of “flawed” obs. can help instrument developers to improve their algorithms.

• **Proactive QC** can readily be implemented into the operational system.

• **EFSR** enables systematic tuning of R matrix.

• **EFSO and EFSR** together can accelerate development of the assimilation of new observing systems.
Acknowledgements

• My advisor Prof. Eugenia Kalnay for advice and encouragement
• Prof. Takemasa Miyoshi of RIKEN/AICS and University of Maryland for guidance and help
• JCSDA and its Director Dr. Sid Boukabara for access to the S4 supercomputer
• Dr. Jim Jung of NCEP for kind guidance on S4
• JPSS program for support
• Mr. Yoichiro Ota for EFSO and GFS-LETKF code
• Government of Japan, in particular Japan Meteorological Agency (JMA) for funding and support

• Dr. Jeff Whitaker, Dr. Sajal Kar and Dr. Jim Purser for discussion and advice (on the Appendix)

• ... and many others and all of you.

THANK YOU!