8B.4 **Assessing observation impact in ensemble data assimilation**

Hong Li*, Qin Wang

Shanghai Typhoon institute, Shanghai, China, 200030

1. INTRODUCTION

With the rapid development of exploring techniques, the number and variety of observations have increased significantly. Since the current operational forecast centers are lack of resource to assimilate all the available observations and in fact much of the data are redundant and even if some of them have negative impact on the forecast, the information of the impact that different observations have on the analyses and forecasts is crucial to better use the observations in a data assimilation system. Traditionally, a common method for assessing the observation impact is to carry out so-called Observation System Experiments (OSEs), in which a subset of observations used in the baseline experiment are removed from a data assimilation system. OSEs is a common way to estimate one or limited subsets of observations but prohibitive to investigate the impacts of large subsets of observations since each subset requires an independent experiment. Langland and Baker (2004) proposed an efficient way of estimating the impact of any subset of observation on short-range forecasts. However this procedure requires the adjoint of forecast model and adjoint of data assimilation, both of which are complicated to develop and not always available for a numerical weather prediction system. Inspired by this adjoint-based estimation, Liu and Kalnay (2008) (LK08) proposed an ensemble-based estimation method within the local ensemble transform Kalman filter by using nonlinear forecast model to evolve the forecast perturbations, avoiding the need of any adjoint system. The ensemble-based procedure is able to estimate impact of any set or subset of observations on measures of forecast by first calculating the impact for each individual observation and then aggregating the results in terms of instrument type, observed variable, channel, location or other category, all computed simultaneously based on a single assessment experiment, achieving the same goal of the adjoint-based method. Liu and Kalnay (2008) tested their method with the Lorenz 40-variable model. In this study, the LK08 method is applied to an AGCM global model to exam its performance with more realistic model and observations. The observation impact of simulated rawinsonde, cloud drift wind, and satellite retrieved temperature and humidity profiles are estimated.

* *Corresponding author:* Hong Li, 166 Puxi Road, Shanghai, China, 200030; e-mail: lih@mail.typhoon.gov.cn

2. LETKF DATA ASSIMILATION SCHEME

The Local Ensemble Transform Kalman Filter (LETKF) (Hunt et al. 2007) is an ensemble square-root filter in which the observations are assimilated to update only the ensemble mean (1), while the ensemble perturbations are updated by transforming the forecast perturbations through a transform matrix (2) introduced by Bishop et al (2001). The basic formulas used in the LETKF are:

$$\bar{x}^a = \bar{x}^b + X^b \tilde{P}^a (HX^b)^T R^{-1}[y^o - h(\bar{x}^b)] \quad (1)$$

$$X^a = X^b[(k-1)\tilde{P}^a]^{1/2} \quad (2)$$

Here X^a, X^b are the analysis and forecast ensemble perturbations, respectively (matrices whose columns are the difference between the ensemble members and the ensemble mean). The transform matrix $\tilde{P}^{a1/2}$ is the square-root of matrix $(K-1)\tilde{P}^a$ where $\tilde{P}^a$, the analysis error covariance in ensemble space, is given by

$$\tilde{P}^a = \left[(K-1)I + (HX^b)^T R^{-1}(HX^b)\right]^{-1} \quad (3)$$

It has dimension K by K where K is the ensemble size, which is generally much smaller than both the dimension of the model and the number of observations. Thus, the LETKF performs the analysis in the space spanned by the forecast ensemble members, which greatly reduces the computational cost. Furthermore, since the analysis is computed independently at each grid point, the LETKF computation can be performed in parallel.

3. ENSEMBLE-BASED METHOD (LK08) FOR ASSESSING THE IMPACT OF OBSERVATION

Like the adjoined-based method, the ensemble-based method assesses the observation impacts at the analysis time 00 on the forecast time t by comparing the forecast at time t starting from time 00 ($\bar{x}^f_{t|0}$ with forecast error $\varepsilon_{t|0}$) with the forecast stating from time -6 ($\bar{x}^f_{t|-6}$ with forecast error $\varepsilon_{t|-6}$) (Figure 1). Since the difference of these two forecast trajectories are from the different initial conditions due to updating the background $\bar{x}^f_{0|-6}$ to the analysis $\bar{x}^a_{00}$, the difference between the errors $(\varepsilon_{t|0} - \varepsilon_{t|-6})$ at time t is solely due to the assimilation of observations at time 00.

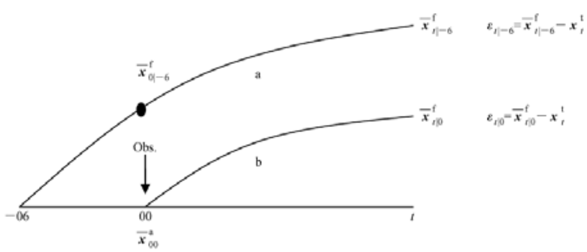

Fig. 1: A schematic of the impact of observations assimilated at time 00 on forecast at time t: observations assimilated at time 00 creating initial conditions for a new trajectory, which has forecast error $\varepsilon_{t|0}$ at time t and the old trajectory starting from time -06, which has forecast error $\varepsilon_{t|-6}$.

Define a cost function to measure the forecast error difference:

$$J = \frac{1}{2}(\varepsilon^{\mathrm{T}}_{t|0} \cdot C \cdot \varepsilon_{t|0} - \varepsilon^{\mathrm{T}}_{t|-6} \cdot C \cdot \varepsilon_{t|-6}) \qquad (4)$$

Here, a matrix C whose diagonal elements are the coefficients in the dry-air energy equation is introduced. Eq. (4) suggests that the negative value of J implies the overall positive impact of observation at time 00 on forecast at time t, and vice versa.

Liu and Kalnay (2008) re-wrote the expression of J by using the formulas of the LETKF to be:

$$J = \frac{1}{2}(2\varepsilon_{t|-6} + \mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}}v_0)^{\mathrm{T}} \cdot C \cdot \mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}}v_0 \qquad (5)$$

Note, here $\widetilde{K} = \widetilde{P}^a (HX^b)^T R^{-1}$ and $v_0 = y^o - h(\bar{x}^b)$ are the middle variables used in the LETKF data assimilation procedure (equation (1)).

Following eq. (5), the impact of each individual observation is calculated by:

$$J_i = v_{0(i)} \cdot [(\mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}})^{\mathrm{T}} \cdot C \cdot (\varepsilon_{t|-6} + 0.5\mathbf{X}^f_{t|-6}\widetilde{\mathbf{K}}v_0)]_{(i} \qquad (6)$$

Thus, the impact of a subset of observations can be obtained by adding each individual impact together within the selected subset.

4. EXPERIMENTAL SETUP

The SPEEDY model (Molteni 2003), an atmospheric general circulation model with simplified physical parameterization schemes is used in this study. It has a spectral primitive-equation dynamics and triangular truncation T30 at 7 sigma levels. Three types of observations including rawinsondes, cloud drift wind, and satellite retrieved temperature and humidity profiles are simulated by adding zero mean normally distributed noise to the SPEEDY model nature run from Jan 1, 1987 to Feb 15, 1987. The rawinsonde observations simulate the real sounding locations with errors of 1 m/s for u, v, 1 K for T, 10^{-4} kg/kg for q, and 1 hPa for Ps. The cloud drift wind data are available above 300hPa in the region of (50°S～50°N) with error of 3 m/s. The satellite retrieved temperature and humidity profiles are simulated on the model grid at every 2 grid points with errors 2 K for T, $2*10^{-4}$ kg/kg for q.

First, a cycled data assimilation experiment is performed in the period of Jan 1, 1987 to Feb 15, 1987 by assimilating the three types of data every 6-hour with the LETKF scheme. 30 ensemble members and a local patch of 4*4 grid box are used in the LETKF. Then we estimate the data impact on 6-hour forecast at each analysis step from Feb 1, 1987 to Feb 15, 1987 by using the assessment equation (6). To verify the results, the real error reduction is computed with eq. (4).

5. RESULTS

The map view of the global distribution of the estimated impact of each individual observation for different instrument types are given in figure 2. The overall impact is positive (more negative values) with more individual positive impact from rawinsonde and relative small but uniform impact from satellite retrievals. Cloud drift wind has largest impact in tropics.

We aggregate the results in terms of two Hemispheres (Fig. 3). Though the total impact in SH is much smaller than that in NH, the individual observation impact in the two hemispheres are similar.

We then group the results in terms of instrument type. Fig. 4 shows that the largest impact in the Northern Hemisphere is produced by rawinsondes, whereas it is produced by satellite retrieved profiles in the Southern Hemisphere.

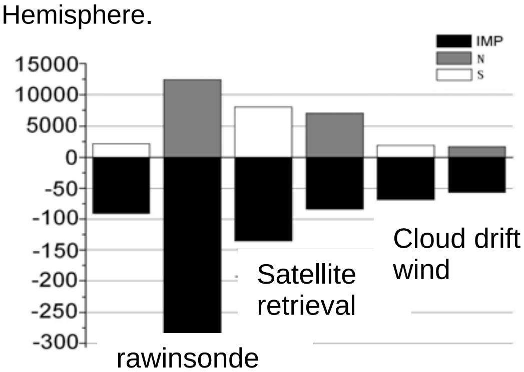

Fig. 4: Summed global observation impact for SH and NH with the y-axis below zero; and the total observation numbers for SH and NH with the y-axis above zero, partitioned by instrument type.

By comparing the impacts of two satellite derived data (Fig. 5), we see that the retrieved profiles are important in the poles while the cloud drift data is much valuable in tropics.

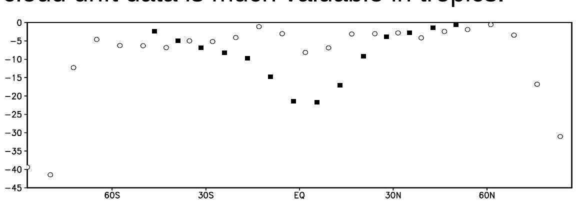

Fig 5: Longitude and vertical summed obs impact for satellite retrieval（○）and cloud-drift wind（■）（unit：J/kg）

To validate our assessment, we sum all the individual impact up and then compare this estimated total impact with the real forecast error reduction (i.e. the actual observation impact) calculated with eq. (4). The result (not shown) suggests that the total estimated global observation impact accounts for 70～80% of the value of actual impact and captures the variations of actual impact very well.

6. CONCLUTION

We have applied the LK08 method to the SPEEDY model to assess the observation impact of simulated rawinsondes, cloud drift wind, and satellite retrieved temperature and humidity profiles. Our results show that the LK08 procedure can successfully evaluate the impact of each observation on the forecast and the impact values can then be grouped and summed by various subsets of observations that may be of interest. The total observation impact in the Northern Hemisphere is bigger than that in the Southern Hemisphere. The largest impact in the Northern Hemisphere is produced by rawinsondes, whereas it is produced by satellite retrieved profiles in the Southern Hemisphere. Cloud drift wind has largest impact in tropics. The validation analysis shows that the estimation of global observation impact accounts for the majority of actual impact and captures the variations of actual impact very well.

REFERENCES

Hunt, B. R., E. Kostelich, and I. Szunyogh, 2007: Efficient Data Assimilation for Spatiotemporal Chaos: a Local Ensemble Transform Kalman Filter. Physica. D: Nonlinear phenomena, 230, 112-126.

Langland R H, N L. Baker, 2004. Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. Tellus, 56, 189-201.

Liu J., E. Kalnay 2008: Estimating observation impact without adjoint model in an ensemble Kalman filter. Quart. J. Roy. Meteor. Soc., 134, 1327-1335.

Molteni, F., 2003: Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. *Climate Dyn.*, **20**, 175-191.

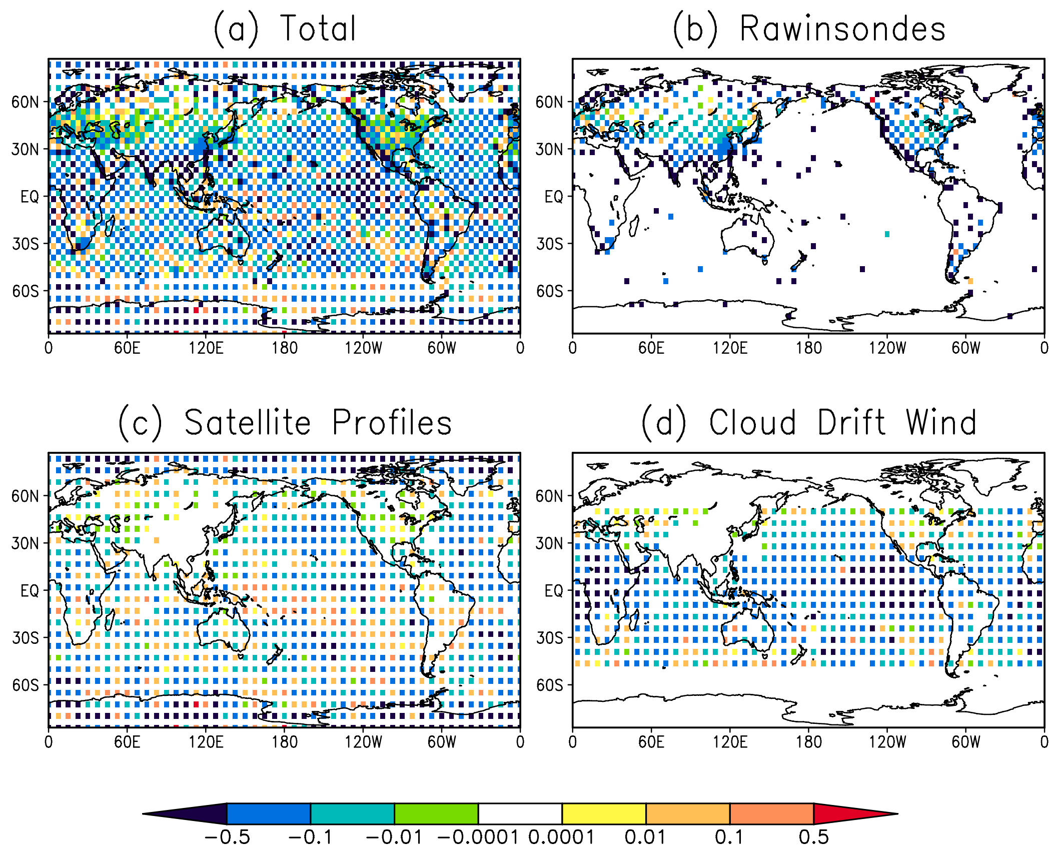

Fig. 2: Global distribution of the estimated impact of each individual observation for different instrument types, averaged for the period from 00Z, Feb01, 1987 to 18Z, Feb15, 1987. (Unit: J/kg)

Fig 3: Time series of summed observation impact for SH (solid line with marks) and NH (solid line) with the y-axis on the left; and time series of mean observation impact for SH (dash line with marks) and NH (dash line) with the y-axis on the right