Self-sensitivity calculation in an EnKF and its possible new applications

Junjie Liu[1], Eugenia Kalnay[2],
Takemasa Miyoshi[2], and Carla Cardinali[3]

[1]University of California at Berkeley

[2]University of Maryland

[3]ECMWF

Outline

- **Review of self-sensitivity and information content**
- Self-sensitivity calculation in:
  - 4D-Var
  - EnKF (new proposed method)
- Verification with cross validation
  - Lorenz-40 variable model
  - Local Ensemble Transform Kalman Filter (LETKF)
- Possible new applications of self-sensitivity:
  - Relationship between information content and data-denial exp.
  - Observation quality control
  - Calculation of the i^{th} forecast (not assimilating the i^{th} observation at the analysis time) based on self-sensitivity
  - Observation impact on the forecast accuracy.
- Conclusions and discussion

Review on influence matrix and self-sensitivity

- The analysis combines background and observations based on weighting matrix K:

$$\mathbf{x}^a = \mathbf{K}\mathbf{y}^o + (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{x}^b$$
$$\mathbf{y}^a = \mathbf{H}\mathbf{x}^a = \mathbf{H}\mathbf{K}\mathbf{y}^o + (\mathbf{I}_p - \mathbf{H}\mathbf{K})\mathbf{y}^b$$

- The analysis sensitivity with respect to the observations:

$$\mathbf{S}^o = \frac{\partial \mathbf{y}^a}{\partial \mathbf{y}^o} = \mathbf{K}^T\mathbf{H}^T = \mathbf{R}^{-1}\mathbf{H}\mathbf{P}^a\mathbf{H}^T$$

- The analysis sensitivity with respect to the background:

$$\mathbf{S}^b = \frac{\partial \mathbf{y}^a}{\partial \mathbf{y}^b} = \mathbf{I}_p - \mathbf{K}^T\mathbf{H}^T = \mathbf{I}_p - \mathbf{S}^o$$

* $\mathbf{S}^o$ is called the influence matrix, which reflects the sensitivity of the analysis to the observations. $\mathbf{S}^b$ is the sensitivity of the analysis to the background;

* Diagonal values of $\mathbf{S}_{ii}^o$ are self-sensitivity, indicating the sensitivity of $\mathbf{y}_i^a$ to $\mathbf{y}_i^o$

* Sensitivities to obs and to bkg are complementary $\mathbf{S}_{ii}^o + \mathbf{S}_{ii}^b = 1$

Cardinali et al. (2004) and Liu et al. (2009)

Self-sensitivity and the analysis value change

- The change in the i^{th} analysis value by leaving out the i^{th} observation is given by:

$$y_i^a - y_i^{a(-i)} = \frac{S_{ii}^o}{(1 - S_{ii}^o)}(y_i^o - y_i^a)$$

- The difference between the actual observation and the predicted observation based on “buddy” observations is given by:

$$(y_i^o - y_i^{a(-i)}) = \frac{(y_i^o - y_i^a)}{(1 - S_{ii}^o)}$$

Note: $y_i^{a(-i)}$ is the analysis value at the i^{th} point after leaving out the i^{th} observation during data assimilation. It is the best possible “buddy check”!

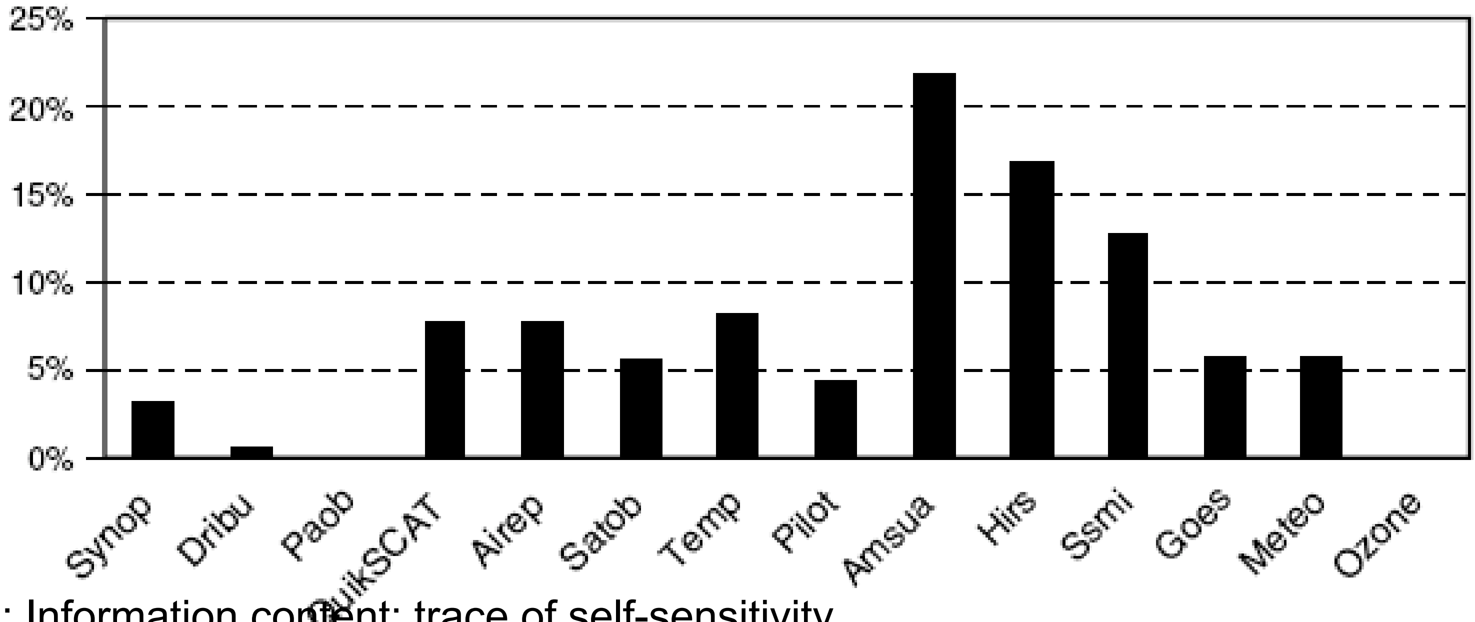
* Both quantities can be calculated from self-sensitivity without knowing $y_i^{a(-i)}$

* An abnormally large difference between the actual obs and the predicted obs may indicate problems with the quality of that observation.

Cardinali et al. (2004) and Liu et al. (2009)

Review of the applications of self-sensitivity

Information content of each major type observations over the total information content of all observations



Note: Information content: trace of self-sensitivity

* Self-sensitivity is a quantitative measure of the observation influence on analysis;

* The information content is qualitative consistent with the results from other studies.

* Information content is also used in channel selection in multi-thousand channel satellites (i.e., Rabier et al., 2002).

Cardinali et al. (2004)

Outline

- Review of self-sensitivity and information content;
- **Self-sensitivity calculation in:**
  - **4D-Var**
  - **EnKF (new proposed method)**
- Verification with cross validation
  - Lorenz-40 variable model
  - Local Ensemble Transform Kalman Filter (LETKF)
- Possible new applications of self-sensitivity:
  - Relationship between information content and data-denial exp.
  - Observation quality control
  - Calculation of the i^{th} forecast (not assimilating the i^{th} observation at the analysis time) based on self-sensitivity
  - Observation impact on the forecast accuracy.
- Conclusions and discussion

Self-sensitivity calculation in Variational approach

- Influence matrix S^o is a function of P^a and R.
- In Variational approach, P^a is not explicitly calculated.
- P^a is the inverse of the matrix of the second derivatives of the cost function J (Hessian) : $\mathbf{P}^a = (\mathbf{J}'')^{-1}$
- so $\mathbf{S}^o \simeq \mathbf{R}^{-1}\mathbf{H}(\mathbf{J}'')^{-1}\mathbf{H}^T$
- $(\mathbf{J}'')^{-1}$ is approximated with a truncated eigenvalue decomposition.

* The truncated eigenvalue decomposition makes S^o_{ii} larger than 1 in some cases, whereas it should be less than or equal than 1.

* The analysis value change by leaving out the i^{th} observation cannot be calculated from self-sensitivity because of this approximation.

Cardinali et al. (2004)

Calculation of self-sensitivity in EnKFs

Influence matrix valid in any data assimilation:

$$\mathbf{S}^o = \frac{\partial \hat{\mathbf{y}}^a}{\partial \mathbf{y}^o} = \mathbf{K}^T\mathbf{H}^T = \mathbf{R}^{-1}\mathbf{H}\mathbf{P}^a\mathbf{H}^T$$

In EnKFs, the calculation of influence matrix requires no approximation:

$$\mathbf{S}^o = \mathbf{R}^{-1}\mathbf{H}\mathbf{P}^a\mathbf{H}^T = \frac{1}{n-1}\mathbf{R}^{-1}(\mathbf{H}\mathbf{X}^a)(\mathbf{H}\mathbf{X}^a)^T$$

$$\mathbf{H}\mathbf{X}^{ai} \cong h(\mathbf{x}^{ai}) - \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}^{ai})$$

When the observation errors have no correlation, self-sensitivity S_{jj}^o and cross-sensitivity S_{jl}^o:

$$S_{jj}^o = \frac{\partial \hat{\mathbf{y}}_j^a}{\partial \mathbf{y}_j^o} = \left(\frac{1}{n-1}\right)\frac{1}{\sigma_j^2}\sum_{i=1}^{n}(\mathbf{H}\mathbf{X}^{ai})_j \times (\mathbf{H}\mathbf{X}^{ai})_j$$

$$S_{jl}^o = \frac{\partial \hat{\mathbf{y}}_j^a}{\partial \mathbf{y}_l^o} = \left(\frac{1}{n-1}\right)\frac{1}{\sigma_l^2}\sum_{i=1}^{n}[(\mathbf{H}\mathbf{X}^{ai})_j \times (\mathbf{H}\mathbf{X}^{ai})_l]$$

* In EnKFs, S_{jj}^o and S_{jl}^o require no approximation, and little computational time.
* S_{jj}^o is always within the theoretical range (0,1).

Liu et al. (Accepted by Q. J. R. Meteorol. Soc. 2009)

Outline

- Review of self-sensitivity and information content;
- Self-sensitivity calculation in:
  - 4D-Var
  - EnKF (new proposed method)
- **Verification with cross validation**
  - **Lorenz-40 variable model**
  - **Local Ensemble Transform Kalman Filter (LETKF)**
- Possible new applications of self-sensitivity:
  - Relationship between information content and data-denial exp.
  - Observation quality control
  - Calculation of the i^{th} forecast (not assimilating the i^{th} observation at the analysis time) based on self-sensitivity
  - Observation impact on the forecast accuracy.
- Conclusions and discussion

Validation of the self-sensitivity calculation in EnKFs

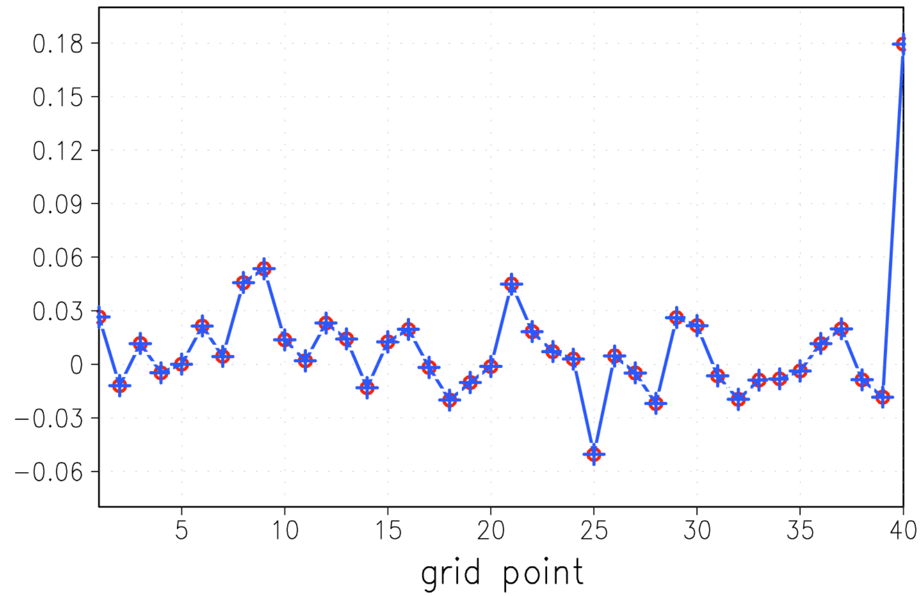
- Lorenz-40 variable model (Lorenz and Emanuel, 1996) with model error (F=8 for the nature run, and F=7.6 for the forecast);
- Local Ensemble Transform Kalman Filter (LETKF, Hunt et al., 2007);
- Observe every point;
- Observations are the nature run with random Gaussian error of 0.2.

Verification methods:

1. Compare $y_i^a - y_i^{a(-i)}$ (based on data-denial experiments) with $\frac{S_{ii}^o}{(1-S_{ii}^o)}(y_i^o - y_i^a)$
2. Compare $\sum_{i=1}^{m}(y_i^o - y_i^{a(-i)})^2$ (calculated by leaving out each obs in turn) with $\sum_{i=1}^{m}\frac{(y_i^o - y_i^a)^2}{(1-S_{ii}^o)^2}$

Analysis value change by leaving out one observation & that based on S_{ii}^o

$\circ: y_i^a - y_i^{a(-i)}$; $+: \frac{S_{ii}^o}{(1-S_{ii}^o)}(y_i^o - y_i^a)$



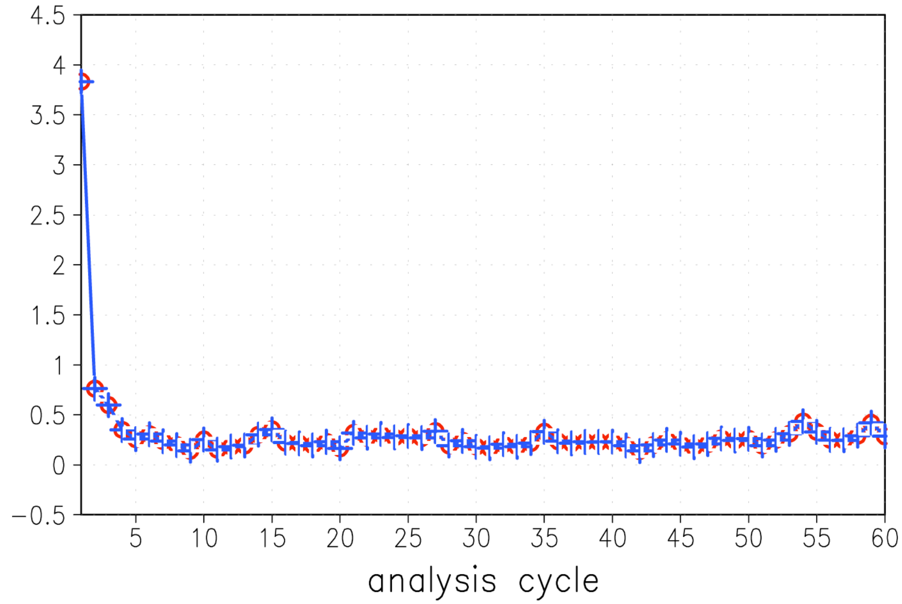
* Both quantities are instantaneous values: the two quantities are the same;

* The self-sensitivity calculation method we proposed is correct;

* The impact of the ith observation on the ith analysis value can be calculated without carrying out data-denial experiment (redoing the analysis w/o the ob).

Cross validation based on self-sensitivity

$$\circ : \sum_{i=1}^{m} (y_i^o - y_i^{a(-i)})^2; \quad + : \sum_{i=1}^{m} \frac{(y_i^o - y_i^a)^2}{(1 - S_{ii}^o)^2}$$



Note: m is the total number of observations.

* The two calculation methods give the same results;

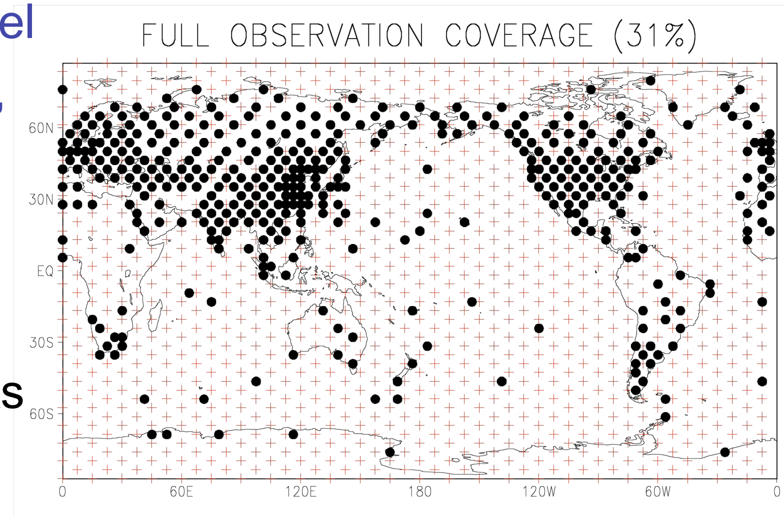
* Cross validation can be easily calculated from self-sensitivity.

Outline

- Review of self-sensitivity and information content;
- Self-sensitivity calculation in:
  - 4D-Var
  - EnKF (new proposed method)
- Verification with cross validation
  - Lorenz-40 variable model
  - Local Ensemble Transform Kalman Filter (LETKF)
- **Possible new applications of self-sensitivity**:
  - **Relationship between information content and data-denial exp.**
  - Observation quality control
  - Calculation of the i^{th} forecast (not assimilating the i^{th} observation at the analysis time) based on self-sensitivity
  - Observation impact on the forecast accuracy.
- Conclusions and discussion

The relationship between information content & the observation impact from data denial experiment

- Simplified PrimitivE Equation DYnamics model (SPEEDY) (Molteni, 2003, adapted by Miyoshi, 2005)
  - A global model with fast computation speed.
  - 96 grid points zonally, and 48 grid points meridionally, and 7 vertical level



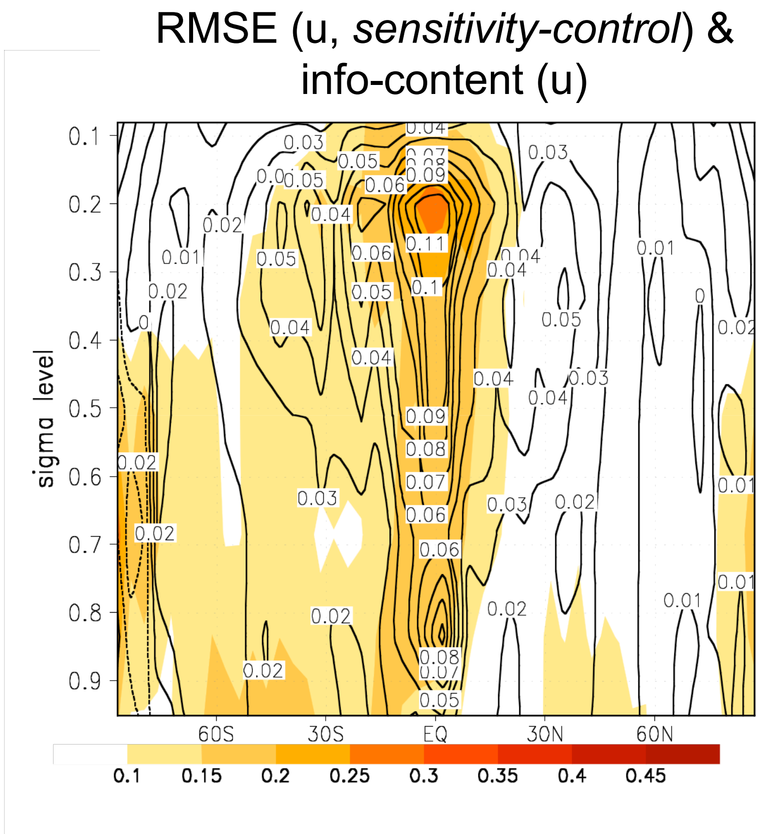
Data denial experiments:

Control run: all dynamical variables are observed in both red + and black dots.

Sensitivity experiment: winds not observed in locations with red +

• Compare information content (the trace of analysis sensitivity) of zonal wind at locations with red + from control run to the RMS error difference between sensitivity experiment and *control experiment.*

Information content (control, shaded) vs. RMSE difference (data-denial experiments, contour)



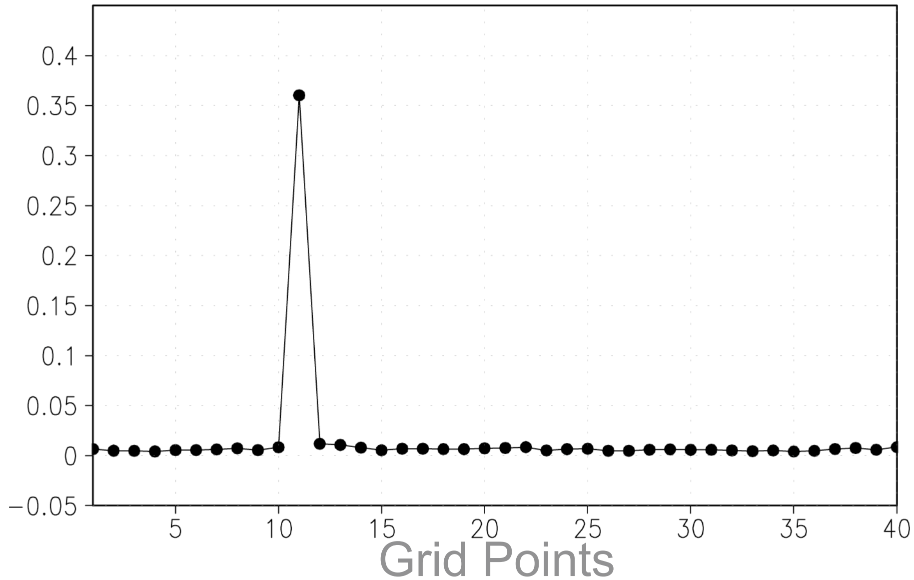
- Information content qualitatively reflects the actual observation impact from data-denial experiments.

Outline

- Review of self-sensitivity and information content;
- Self-sensitivity calculation in:
  - 4D-Var
  - EnKF (new proposed method)
- Verification with cross validation
  - Lorenz-40 variable model
  - Local Ensemble Transform Kalman Filter (LETKF)
- **Possible new applications of self-sensitivity**:
  - Relationship between information content and data-denial exp.
  - **Observation quality control**
  - Calculation of the i^{th} forecast (not assimilating the i^{th} observation at the analysis time) based on self-sensitivity
  - Observation impact on the forecast accuracy.
- Conclusions and discussion

Observation quality control

$$\frac{1}{T}\sum_{t=1}^{T}(y_i^o - y_i^{a(-i)})^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{(y_{i,t}^o - y_{i,t}^a)^2}{(1-S_{ii,t}^o)^2},\ i=1,\cdots,m$$

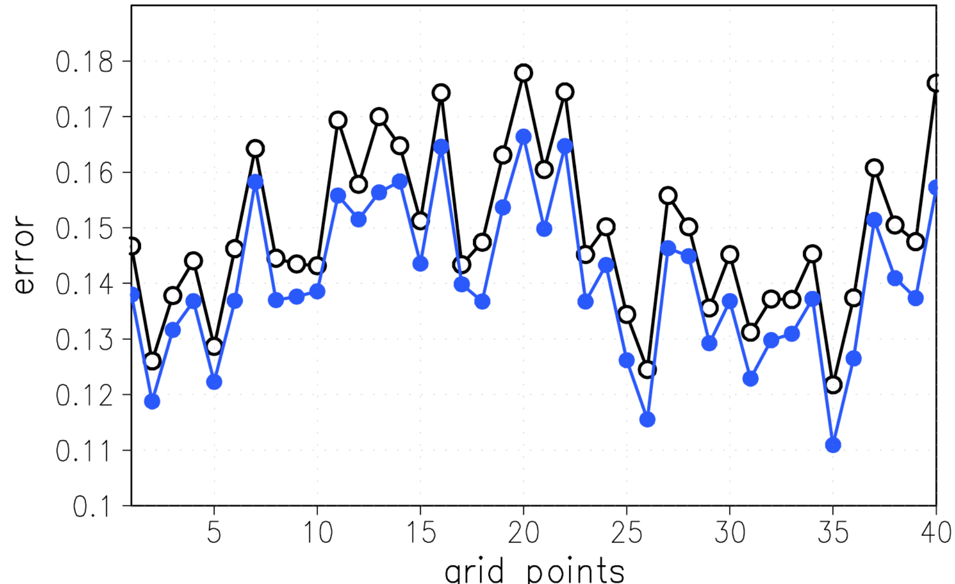


Experimental design: the observation error standard deviation at the 11th point is 4 times larger than the others.

- The difference between the predicted observation $y_i^{a(-i)}$ and the actual obs y_i^o is larger when the ith observation has larger error (11th point).
- Does not need much computational time.

Difference between y_i^b and $y_i^{a(-i)}$ and the implications for observation quality control

6-hour forecast error $[\frac{1}{T}\sum_{t=1}^{T}(y_{i,t}^b - y_{i,t}^{truth})^2]^{\frac{1}{2}}$ & error of predicted obs based on buddy analysis $[\frac{1}{T}\sum_{t=1}^{T}(y_{i,t}^{a(-i)} - y_{i,t}^{truth})^2]^{\frac{1}{2}}$



* 6-hour forecast error is similar to the error of the predicted obs based on the buddy obs, but the predicated obs is more accurate.

* Both the 6-hour forecast error and the error of the predicted obs have smaller error than the bad observation at the 11th point (stdv=0.80);

* The observation quality control based on 6-hour forecast and the predicted obs will give similar results.

Outline

- Review of self-sensitivity and information content;
- Self-sensitivity calculation in:
  - 4D-Var
  - EnKF (new proposed method)
- Verification with cross validation
  - Lorenz-40 variable model
  - Local Ensemble Transform Kalman Filter (LETKF)
- **Possible new applications of self-sensitivity**:
  - Relationship between information content and data-denial exp.
  - Observation quality control
  - **Calculation of the i^{th} forecast (not assimilating the i^{th} observation at the analysis time) based on self-sensitivity**
  - Observation impact on the forecast accuracy.
- Conclusions and discussion

The impact of the i^{th} observation on the forecast at the i^{th} observation point

The difference between the forecasts (at the i^{th} point) initiated from the analyses made with and without the i^{th} observation:

- $y_i^f - y_i^{f(-i)} = [M(\mathbf{y}^a) - M(\mathbf{y}^{a(-i)})]_i \cong [\mathbf{M}(\mathbf{y}^a - \mathbf{y}^{a(-i)})]_i \cong [\mathbf{M}(\mathbf{y}_i^a - \mathbf{y}_i^{a(-i)})]$

The approximation comes from two aspects:

1) Nonlinearity;

2) The impact of the change in the analysis of the points other than the i^{th} point (due to deletion of the i^{th} observation) on the i^{th} forecast.

$$y_i^f - y_i^{f(-i)} \simeq [\mathbf{M}S_{ii}^o(1 - S_{ii}^o)^{-1}\sigma_{ii}^{-2}(\mathbf{y}_i^o - \mathbf{y}_i^a)]_i$$

$$\simeq \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{HX}_i^{fj})(\mathbf{HX}_i^{aj})^T(1 - S_{ii}^o)^{-1}\sigma_{ii}^{-2}(\mathbf{y}_i^o - \mathbf{y}_i^a)$$

(n is # of ensemble members, σ_{ii}^{-2} is the inverse of observation error variance)

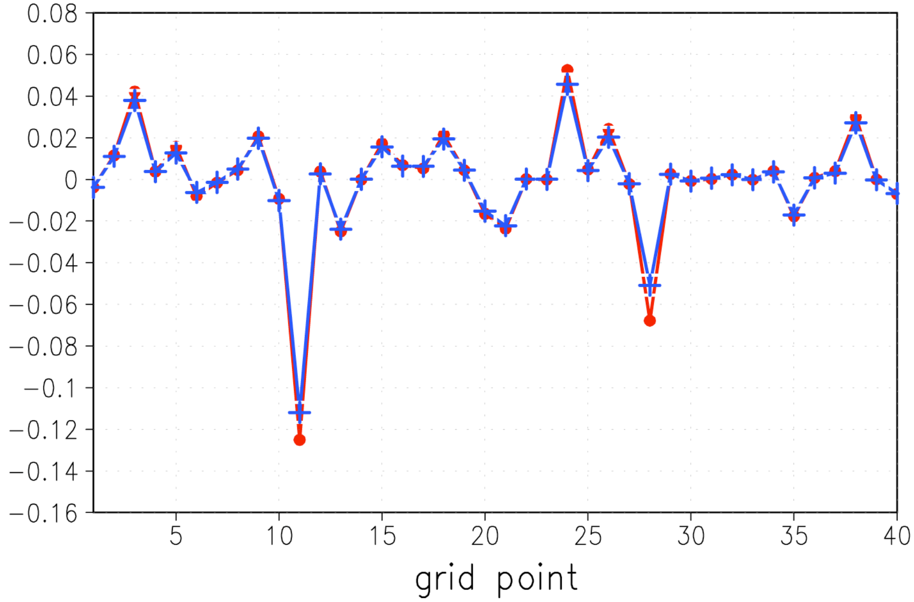
$$\mathbf{HX}^{ai} \cong h(\mathbf{x}^{ai}) - \frac{1}{n}\sum_{i=1}^{n}h(\mathbf{x}^{ai}) \qquad \mathbf{HX}^{fj} \cong h(\mathbf{x}^{fj}) - \frac{1}{n}\sum_{j=1}^{n}h(\mathbf{x}^{fj})$$

* $y_i^{f(-i)}$ can be approximately calculated without carrying out data denial experiment!

Forecast change by leaving out one observation & that based on S_{ii}^{o}

$\circ : y_i^f - y_i^{f(-i)}$ (data denial experiment);

$+$: Calculated from self-sensitivity & forecast perturbations



Note: the plot is instantaneous values at one analysis cycle; the forecast length is 24-hr.

* Forecast value change calculated from self-sensitivity and forecast perturbations is very close to the actual forecast value change from data denial exp.

Outline

- Review of self-sensitivity and information content;
- Self-sensitivity calculation in:
  - 4D-Var
  - EnKF (new proposed method)
- Verification with cross validation
  - Lorenz-40 variable model
  - Local Ensemble Transform Kalman Filter (LETKF)
- **Possible new applications of self-sensitivity**:
  - Relationship between information content and data-denial exp.
  - Observation quality control
  - Calculation of the i^{th} forecast (not assimilating the i^{th} observation at the analysis time) based on self-sensitivity
  - **Observation impact on the forecast accuracy.**
- Conclusions and discussion

Observation impact on the forecast accuracy

The forecast error changes by leaving out the i[th] observation:

$$J_i = [(y_i^f - y_i^a)^2 - (y_i^{f(-i)} - y_i^a)^2]$$

y_i^a **is the verification analysis.**

The statistics of J_i over a group of observations:

$$\sum_{i=1}^{N} J_i = \sum_{i=1}^{N} [(y_i^f - y_i^a)^2 - (y_i^{f(-i)} - y_i^a)^2]$$

When the observations improves the forecast accuracy,

the cost function is negative;

When the observations deteriorates the forecast,

the cost function is positive.

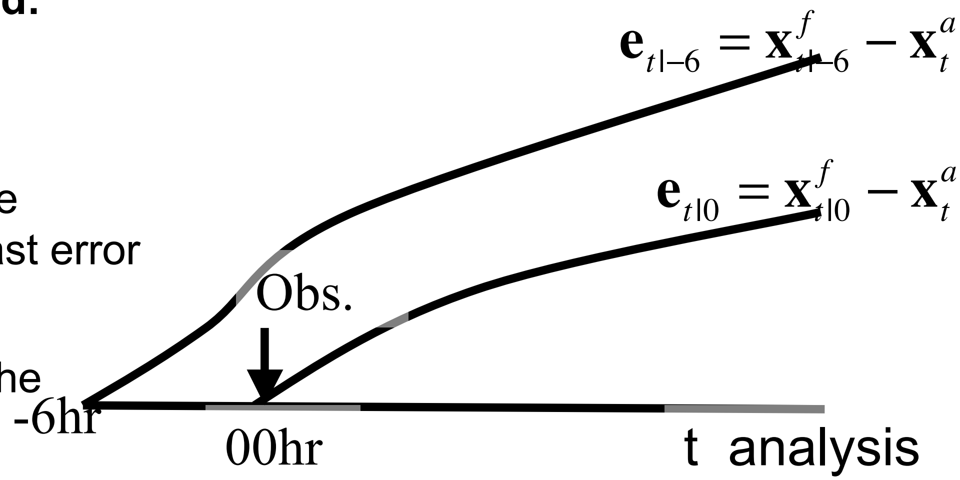
Observation impact based on self-sensitivity & the observation impact from adjoint and ensemble sensitivity method

The adjoint and ensemble sensitivity method:

$$J = [(\mathbf{x}^f_{t|0} - \mathbf{x}^a_t)^2 - (\mathbf{x}^f_{t|-6} - \mathbf{x}^a_t)^2]$$

* The cost function reflects the impact of all the observations assimilated at 00hr on the forecast error difference (model space) at time t.

* The cost function is rewritten as function of the observations assimilated at 00hr.



(Langland and Baker, 2004; Liu and Kalnay, 2008)

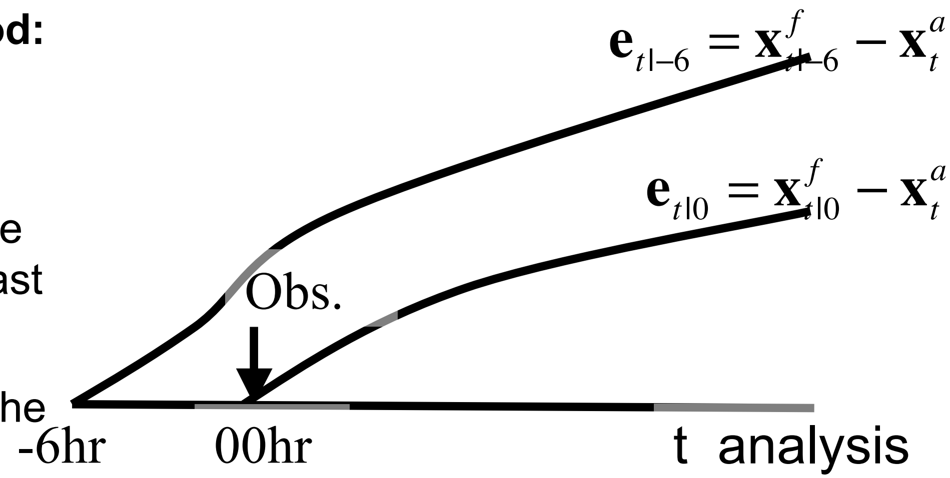
Observation impact based on self-sensitivity & the observation impact from adjoint and ensemble sensitivity method

The adjoint and ensemble sensitivity method:

$$J = [(\mathbf{x}^f_{t|0} - \mathbf{x}^a_t)^2 - (\mathbf{x}^f_{t|-6} - \mathbf{x}^a_t)^2]$$

* The cost function reflects the impact of all the observations assimilated at 00hr on the forecast error difference (model space) at time t.

* The cost function is rewritten as function of the observations assimilated at 00hr.

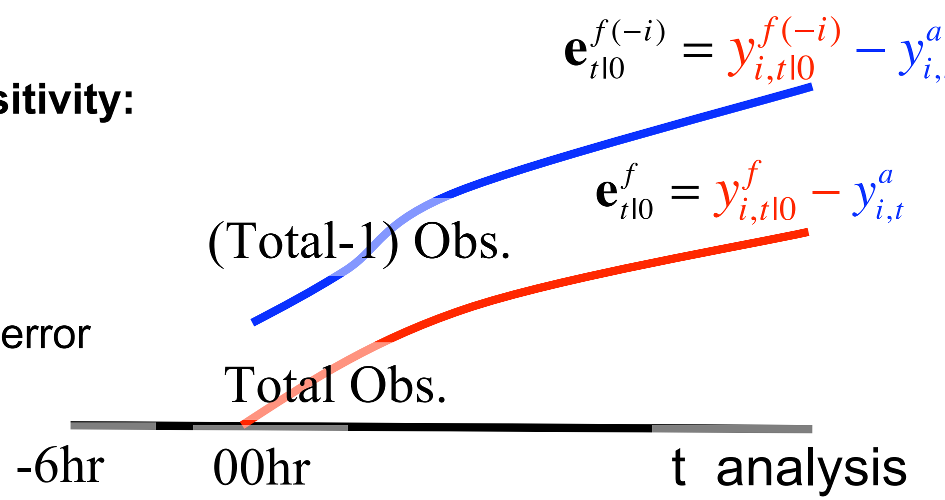


The observation impact based on self-sensitivity:

$$J_i = [(y^f_{i,t|0} - y^a_{i,t})^2 - (y^{f(-i)}_{i,t|0} - y^a_{i,t})^2]$$

* The cost function reflects the impact of the i^{th} observation assimilated at 00hr on the forecast error difference (observation space) at time t.

* There is no need to rewrite the cost function.

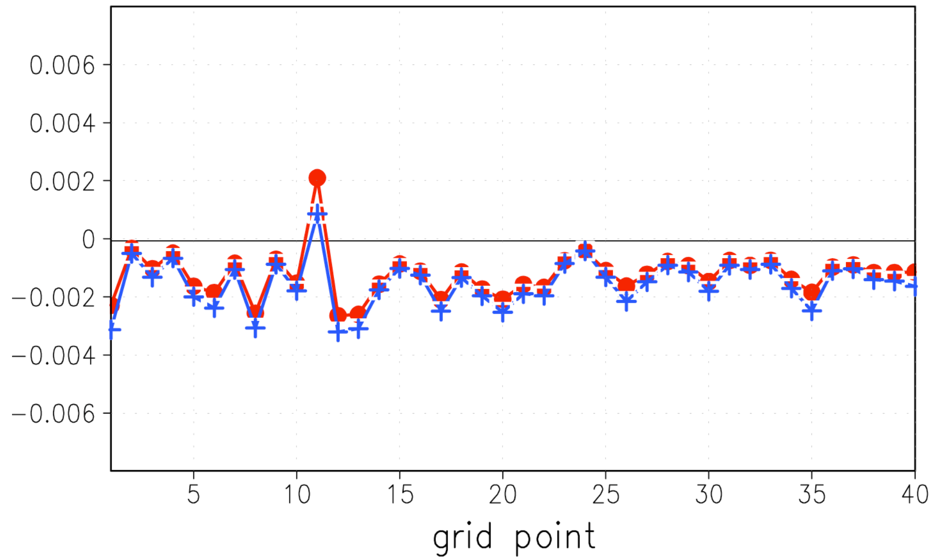


(Langland and Baker, 2004; Liu and Kalnay, 2008)

Detection of bad quality observation

Blue: time average of the cost function J_i with $y_i^{f(-i)}$ calculated from self-sensitivity, ensemble forecasts.

Red : time average of the cost function J_i with $y_i^{f(-i)}$ calculated by leaving out the ith observation during data assimilation.



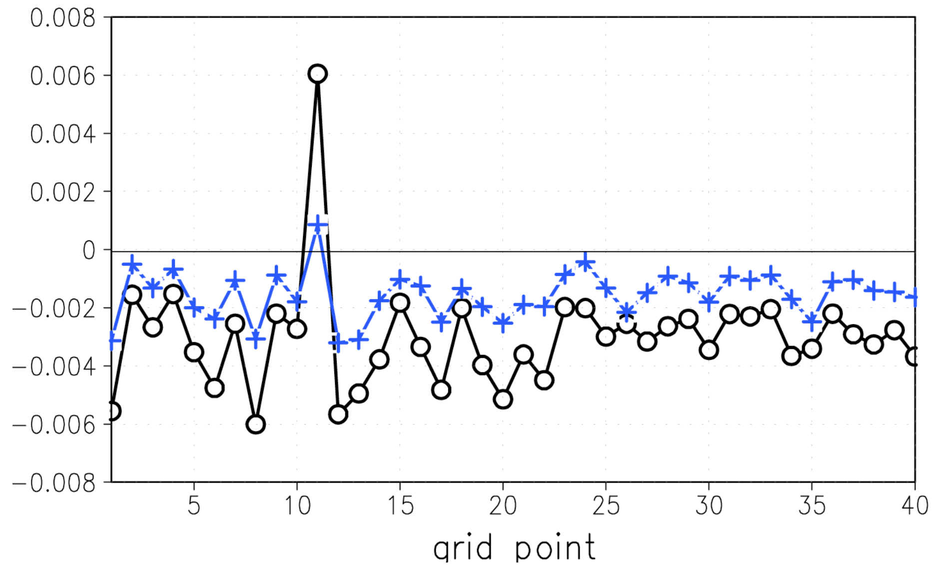
Note: the observation at the 11th point has 4 times larger random error than the others

* **Both cost function give similar results, and both detect the observation with bad quality.**

The impact of the accuracy of the verification state

$$J_i = (y_i^f - y_i^a)^2 - (y_i^{f(-i)} - y_i^a)^2 \qquad i = 1,\ldots,N \ \&$$

$$J_i = (y_i^f - y_i^t)^2 - (y_i^{f(-i)} - y_i^t)^2 \qquad i = 1,\ldots,N$$



The difference between black line and the blue line is the verification state. Both $y_i^{f(-i)}$ **are calculated from self-sensitivity and forecast ensemble forecasts.**

- **The cost function detects that the 11th observation makes the forecast worse.**
- **Different verification states make a big difference in the signal.**

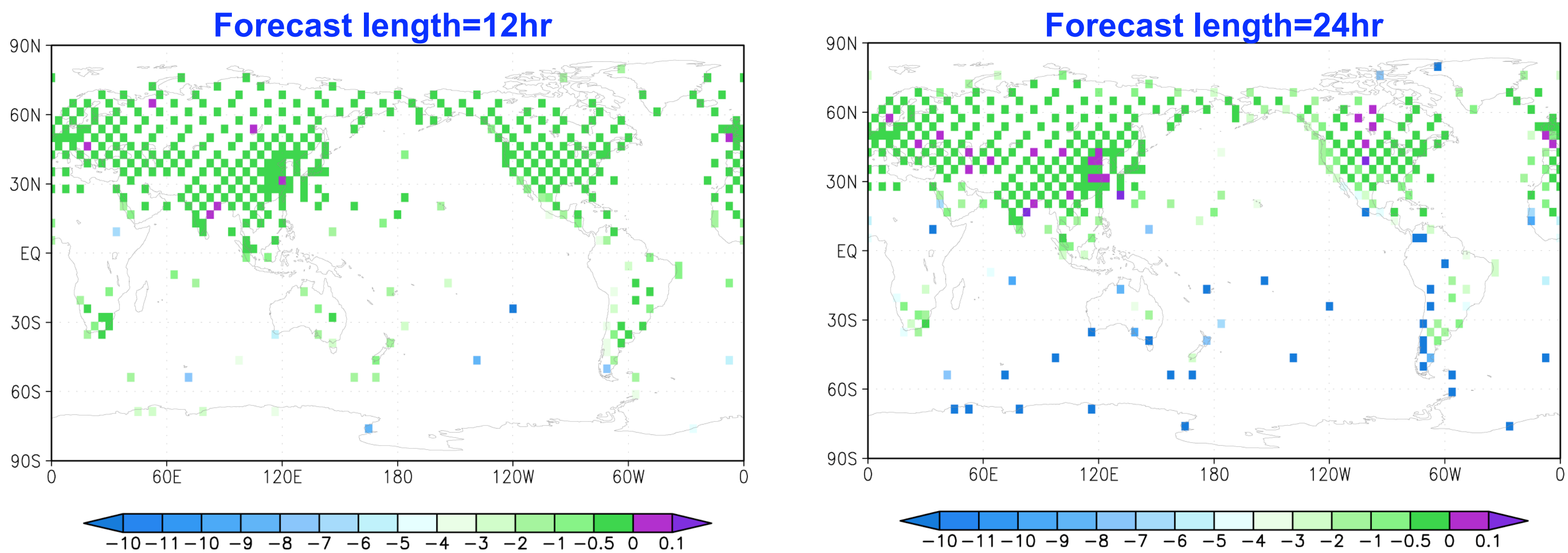
Observation impact on the forecast accuracy in a global model

- Numerical model: Simplified PrimitivE Equation DYnamics model (SPEEDY) (Molteni, 2003, adapted by Miyoshi, 2005)
- Local Ensemble Transform Kalman Filter (Hunt et al., 2007)
- OSSE experiments (perfect model);
- Observation error is about 30% of the natural variability of the model.
- Observed every vertical level in the rawinsonde observation location, except specific humidity (observed the lowest 5 vertical levels)

The zonal wind observation impact on the forecast accuracy

$$J_{i,j} = (y^{f}_{i,j} - y^{a}_{i,j})^2 - (y^{f(-i)}_{i,j} - y^{a}_{i,j})^2, i: longitude; j: latitude$$

Averaged over time, and all the vertical levels (unit: m^2/s^2)



Note: negative: the observation improves the forecast; positive: the observation makes the forecast worse.

*** A few points in the data dense area make the forecast worse just by chance.**

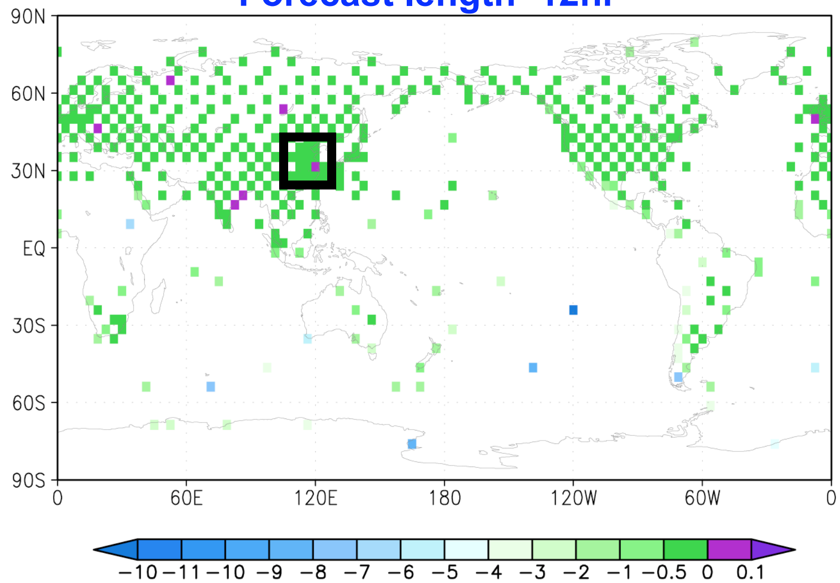
*** In the data sparse area, the obs impact on the 24hr forecast is larger than that on the 12hr forecast.**

Does the observation show different impact in data dense area?

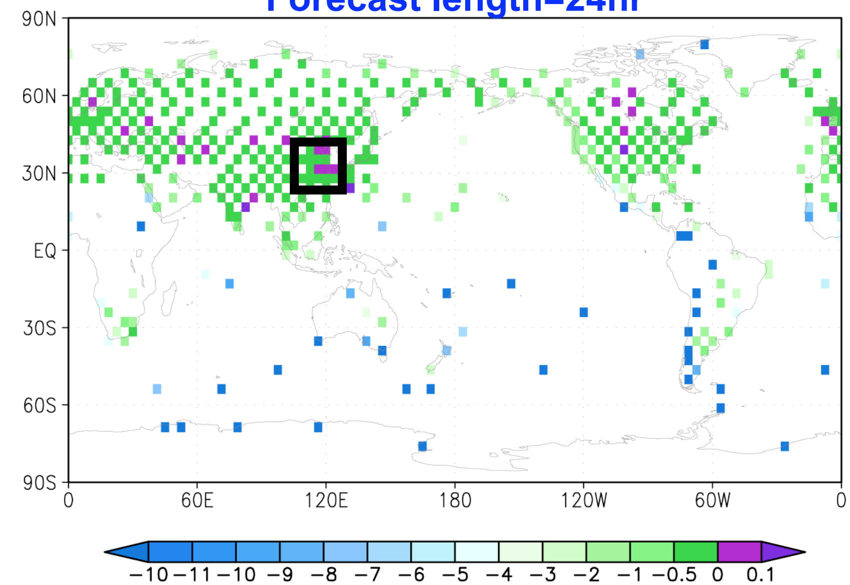
$$J_{i,j} = (y_{i,j}^{f} - y_{i,j}^{a})^2 - (y_{i,j}^{f(-i)} - y_{i,j}^{a})^2, i: longitude; j: latitude$$

Averaged over time, and all the vertical levels (unit: m^2/s^2)

Forecast length=12hr



Forecast length=24hr



Note: negative: the observation improves the forecast; positive: the observation makes the forecast worse.

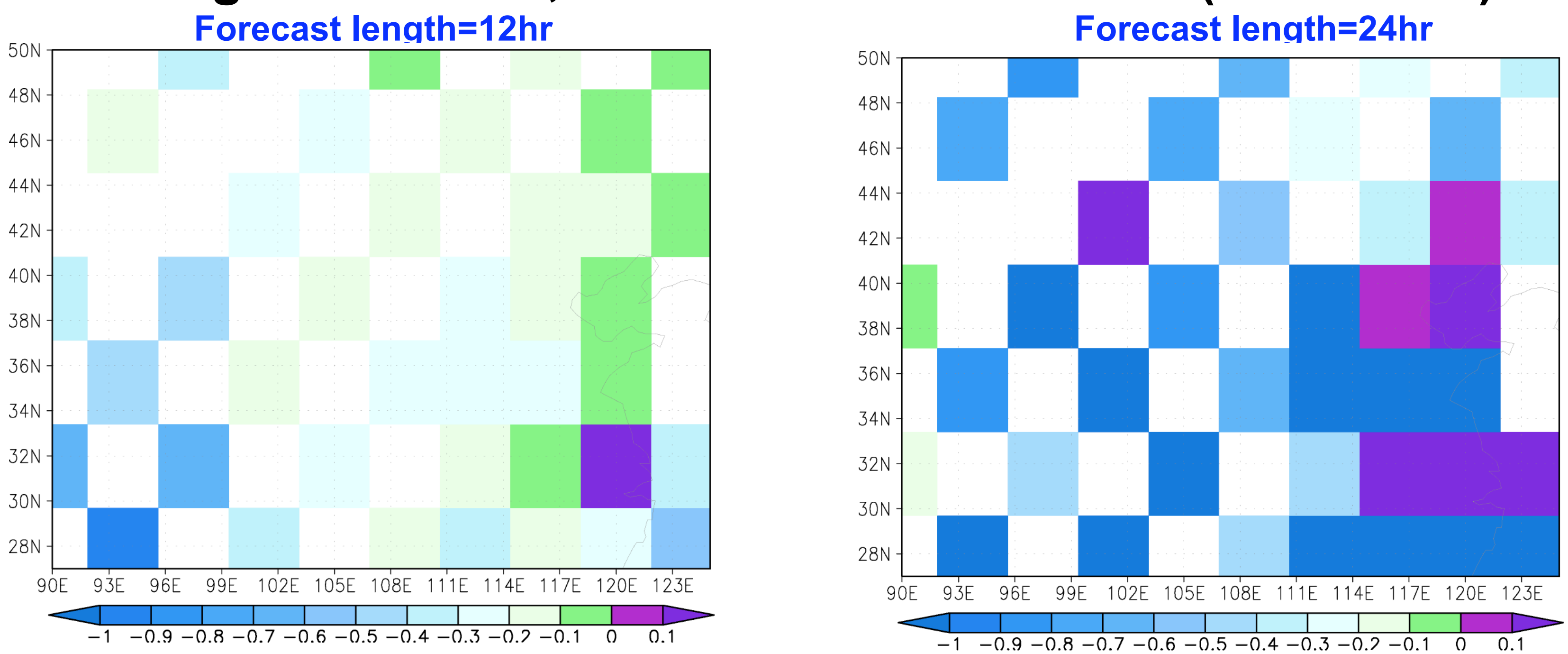
*** A few points in the data dense area make the forecast worse.**

*** In the data sparse area, the obs impact on the 24hr forecast is larger than that on the 12hr forecast.**

Zonal wind observation impact in the data dense area

$$J_{i,j} = 40 \times [(y_{i,j}^{f} - y_{i,j}^{a})^2 - (y_{i,j}^{f(-i)} - y_{i,j}^{a})^2], i: longitude; j: latitude$$

Averaged over time, and all the vertical levels (unit: m^2/s^2)



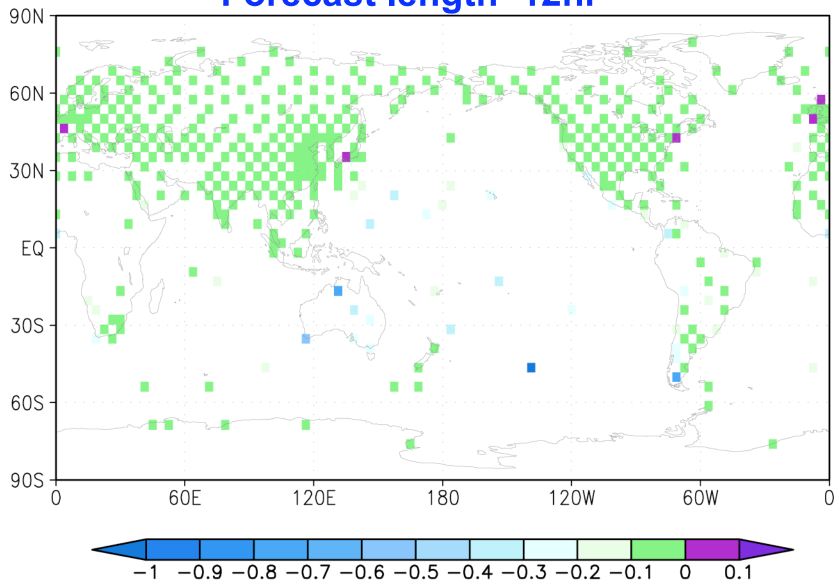
* **Observation impact shows difference in data dense area. The impact is much smaller than the observation in data sparse area.**
* **Observation impact on the 24hr forecast is larger than that on 12hr forecast.**

Specific humidity observation impact

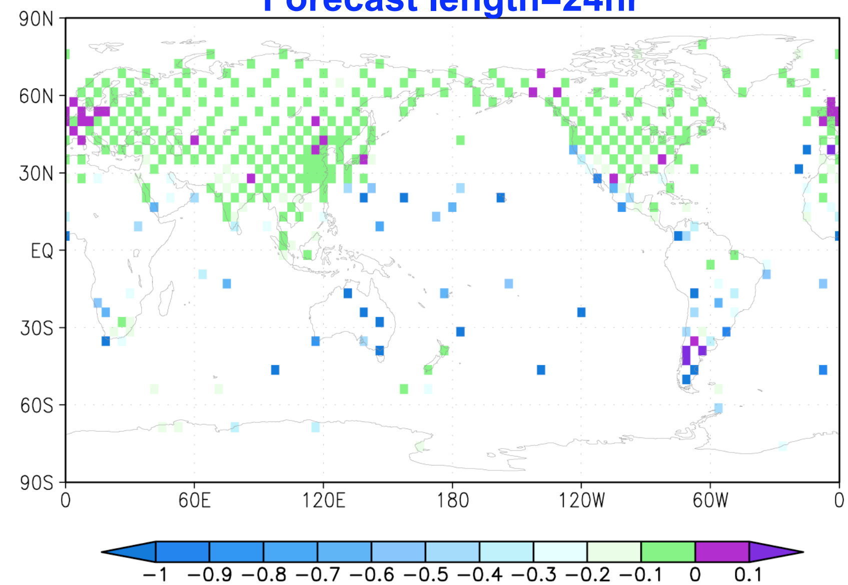
$$J_{i,j} = 1.0e7 \times (y_{i,j}^{f} - y_{i,j}^{a})^2 - (y_{i,j}^{f(-i)} - y_{i,j}^{a})^2, i: longitude; j: latitude$$

Averaged over time, and all the vertical levels (unit:$1e^7 kg^2/kg^2$)

Forecast length=12hr



Forecast length=24hr



* **Specific humidity involved in highly nonlinear process; specific humidity improves forecast in most places.**

* **Larger impact in data sparse areas.**

Outline

- Review of self-sensitivity and information content;
- Self-sensitivity calculation in:
  - 4D-Var
  - EnKF (new proposed method)
- Verification with cross validation
  - Lorenz-40 variable model
  - Local Ensemble Transform Kalman Filter (LETKF)
- Possible new applications of self-sensitivity:
  - Relationship between information content and data-denial exp.
  - Observation quality control
  - Calculation of the i^{th} forecast (not assimilating the i^{th} observation at the analysis time) based on self-sensitivity
  - Observation impact on the forecast accuracy.
- **Conclusions and discussion**

Conclusions and discussion

- A new method is proposed to calculate influence matrix and self-sensitivity in an EnKF;
- Influence matrix and self-sensitivity can be easily calculated in EnKFs: applying the observation operator on the ensemble analyses and carrying out scalar products;
- With no approximation needed in the calculation of self-sensitivity, the self-sensitivity remains within the theoretical range (0,1);
- The analysis value change by leaving out the i^{th} observation can be inexpensively calculated from self-sensitivity;
- Cross-validation can be easily calculated based on self-sensitivity without carrying out data-denial experiments.
- Information content qualitatively reflects the observation impact from data-denial experiments.
- Self-sensitivity could be used in observation quality control and the observation impact on the forecast accuracy.