Data Assimilation

Data assimilation in different tongues

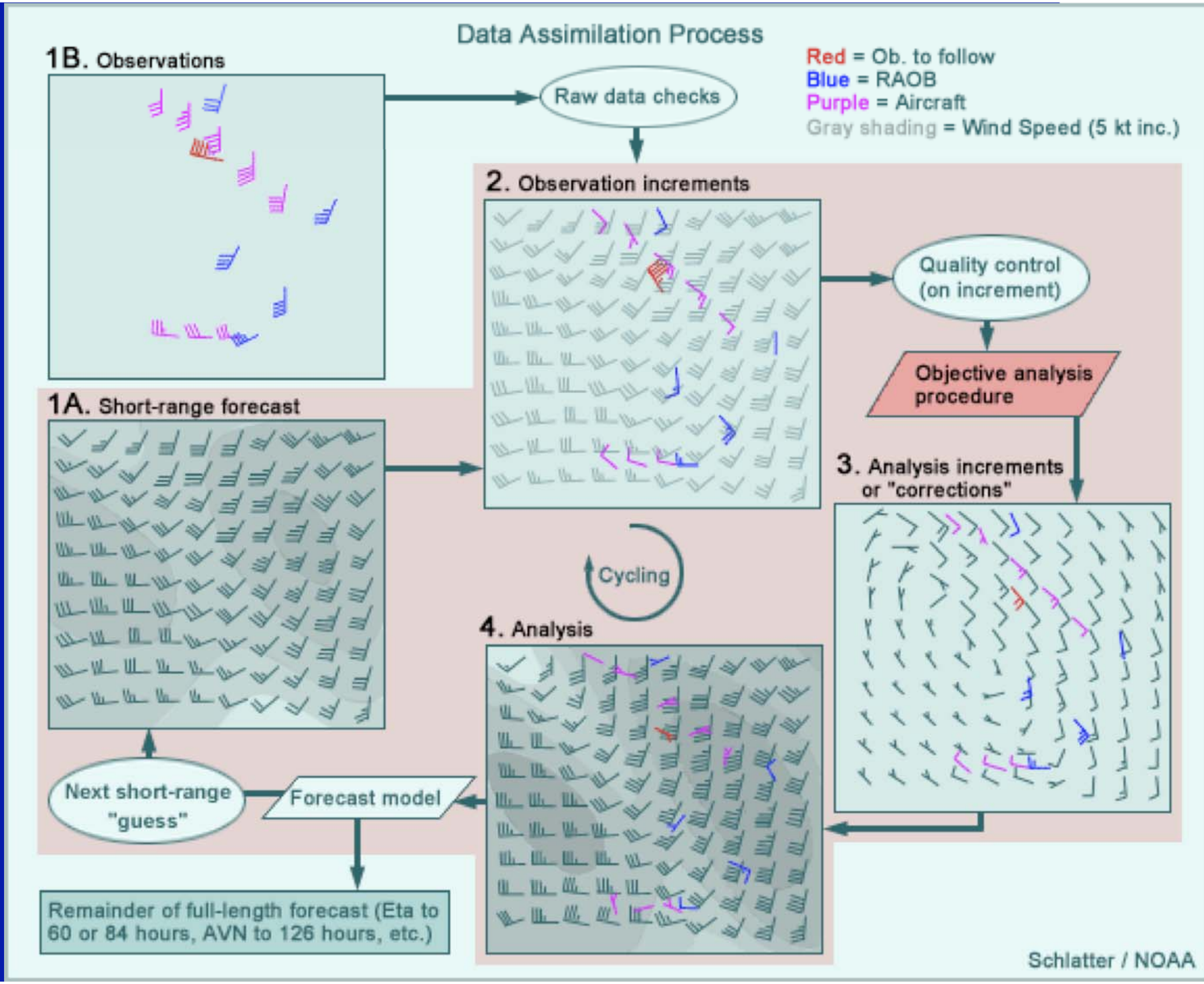
- "Data assimilation" - GFD
- "State estimation" - nonlinear dynamics
- "Inverse modeling" - geophysics et al.
- "Signal processing" – engineering
- "Chaos synchronization" – physics

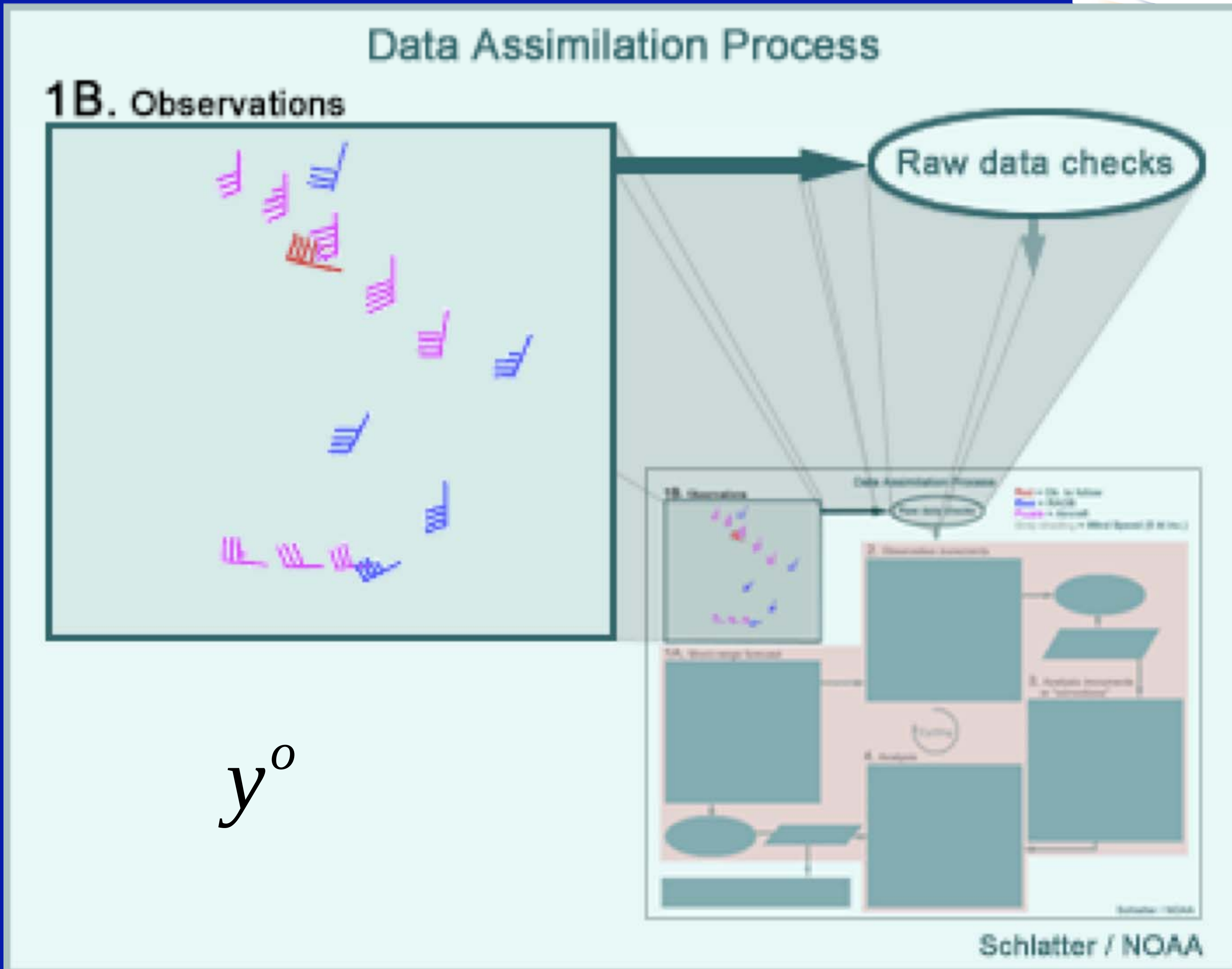
At root, it is blending/combining *multiple* sources of information to get a "best estimate."

Goals of data assimilation

- Keep a numerical model “close” to a set of observations over time
- Provide appropriate initial conditions for a forecast
- Provide an estimate of analysis errors
- Propagate information from observations to unobserved locations
- Tell us something about how the model behaves

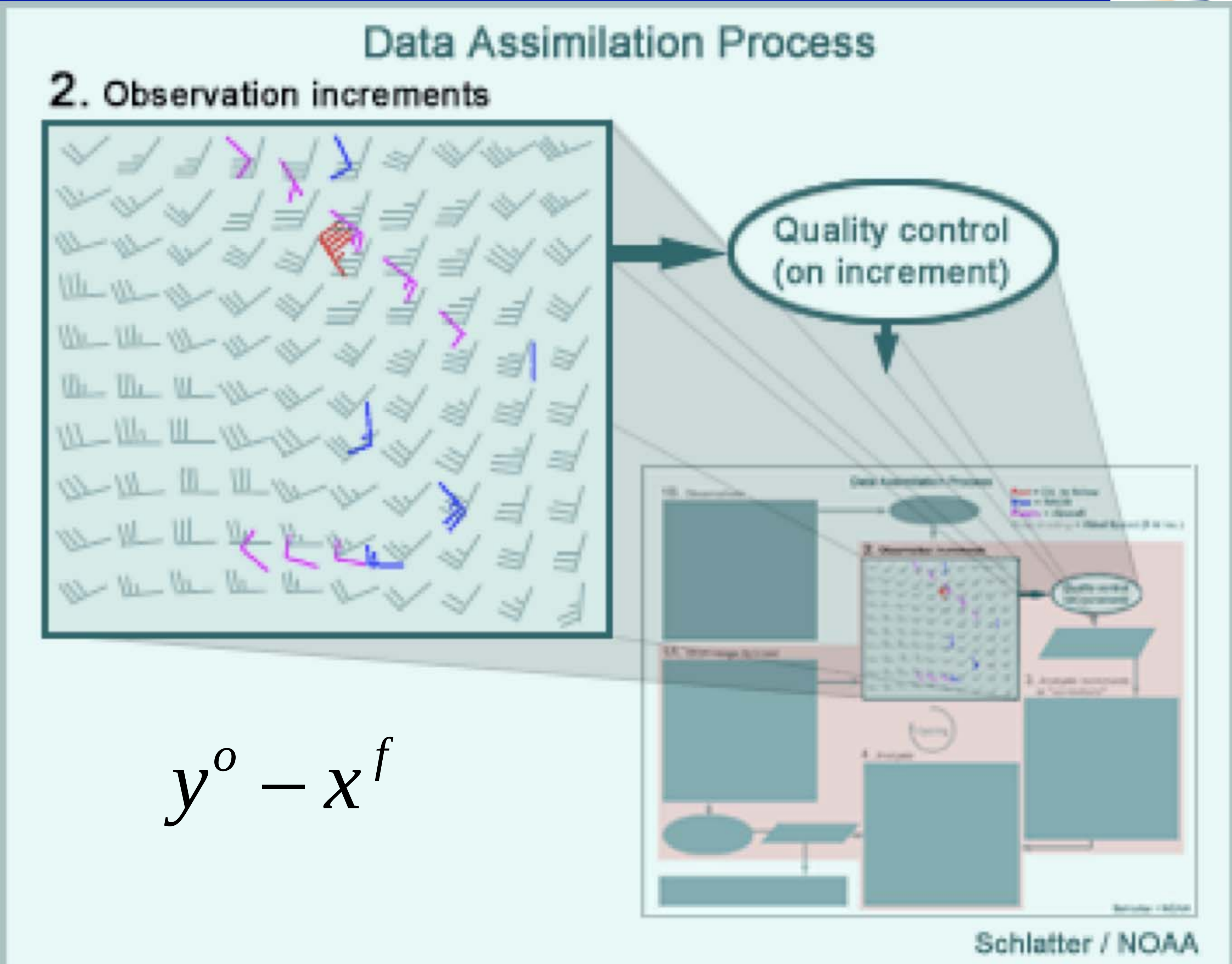
AR


Data Assimilation Process
Red = Ob. to follow
Blue = RAOB
Purple = Aircraft
Gray shading = Wind Speed (5 kt inc.)
1B. Observations
Raw data checks
2. Observation increments
Quality control (on increment)
Objective analysis procedure
3. Analysis increments or "corrections"
1A. Short-range forecast
Cycling
4. Analysis
Next short-range "guess"
Forecast model
Remainder of full-length forecast (Eta to 60 or 84 hours, AVN to 126 hours, etc.)
Schlatter / NOAA



Data Assimilation Process
1B. Observations
Raw data checks
Schlatter / NOAA

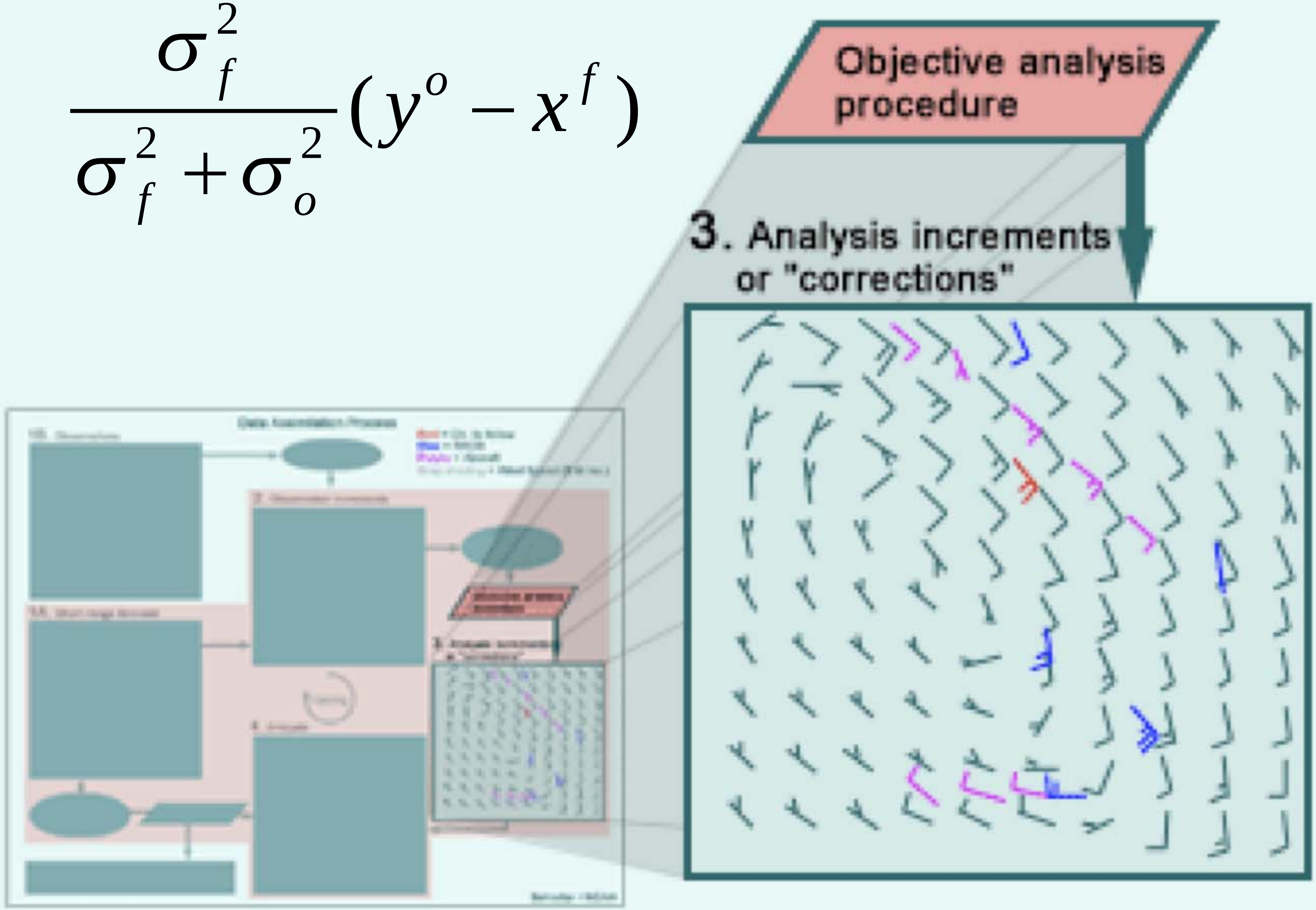
$$y^o$$


Data Assimilation Process
2. Observation increments
Quality control
(on increment)
Schlatter / NOAA

$$y^o - x^f$$

Data Assimilation Process

$$\frac{\sigma_f^2}{\sigma_f^2 + \sigma_o^2}(y^o - x^f)$$



Schlatter / NOAA

Review: main components

- Relating a gridded model state to observations
- Introducing them over some number of times (could be as few as one)
- (Initialization step)

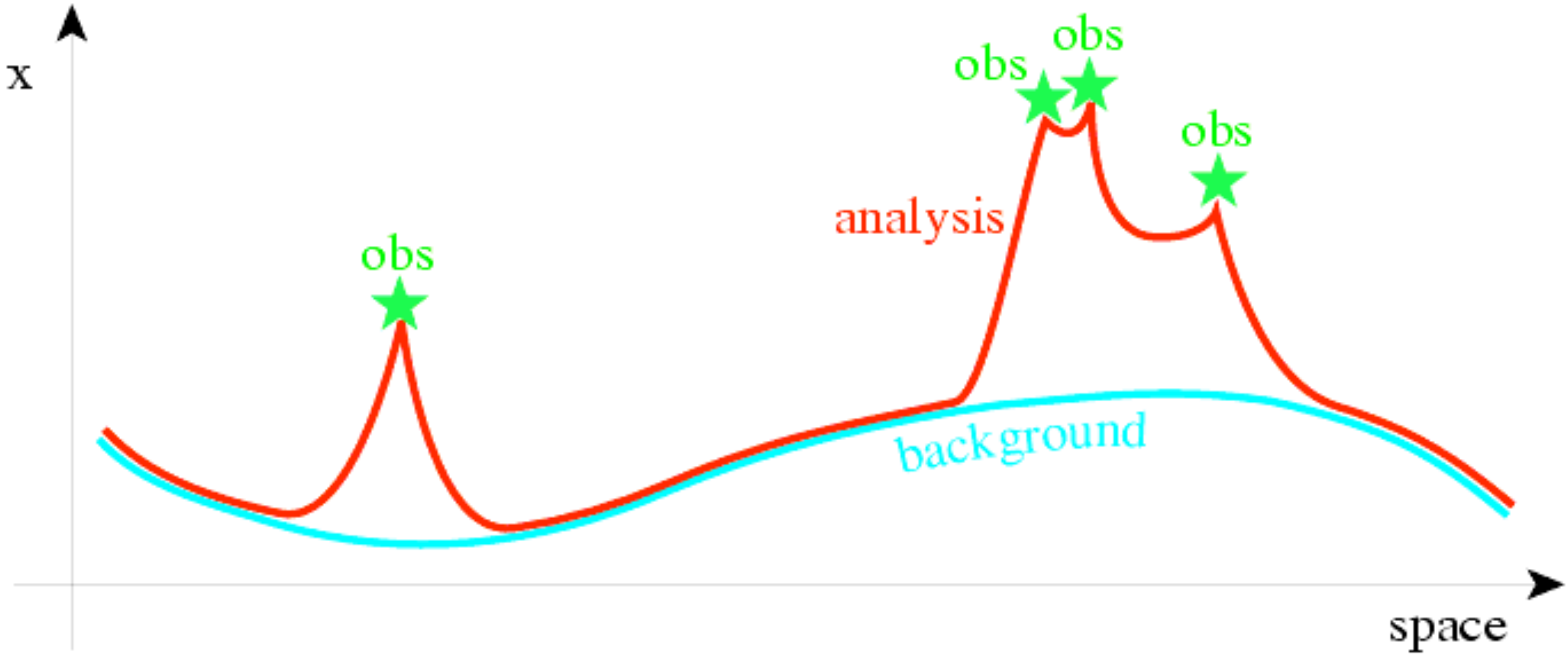
Information propagation into unobserved areas

- Want *good* information to propagate
- Depends on
  - Quality of the model
  - Quality of the observations
  - Observation frequency and location
  - Balance of initial conditions
- Any good DA system can achieve this "optimally" if properly set up

Single-observation examples

- Single-time updates:
  - Cressman
  - Statistical
- Weather is 4D
  - Sequential
  - Continuous

First steps: Cressman



Weighted-average of nearby observations based on the distance squared

Why not Cressman?

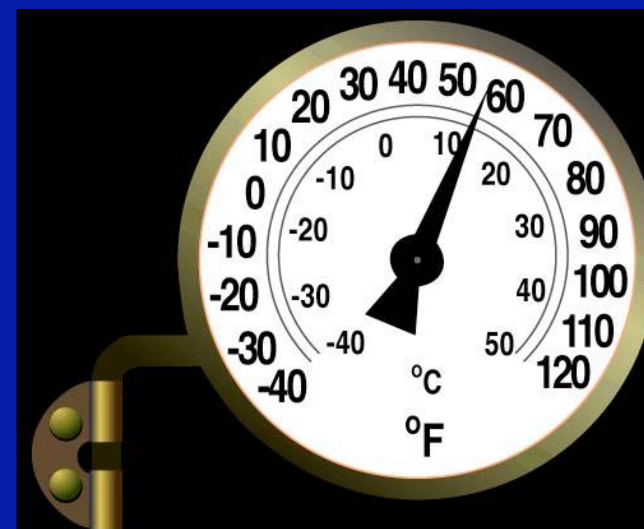
- If we have a preliminary estimate of the analysis with a good quality, we do not want to replace it by values provided from poor quality observations.
- When going away from an observation, it is not clear how to relax the analysis toward the arbitrary state, i.e. how to decide on the shape of the function.
- An analysis should respect some basic known properties of the true system, like smoothness of the fields, or relationship between the variables (e.g. hydrostatic balance). This is not guaranteed by the Cressman method: random observation errors could generate unphysical features in the analysis.

No background field!

A practical route to data assimilation…statistics

What temperature is it here?

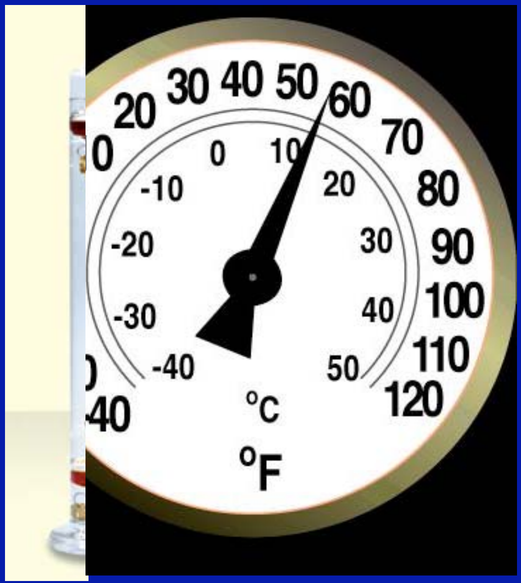
x_1 • x_2

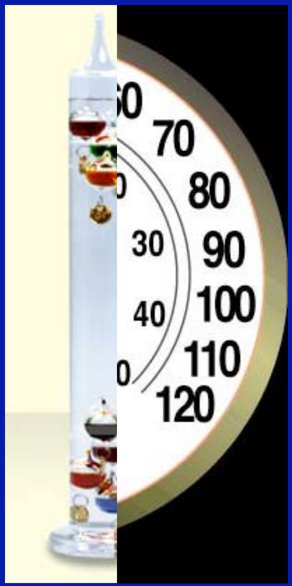


Form a linear combination of estimates

NCAR

$$x^a = \alpha\, x_1 + \left(1 - \alpha\right) x_2$$

$$x^a =$$ 

or 

?

$$x^a = x_2 + \alpha\left(x_1 - x_2\right)$$

Include the background as one "obs"

- Generalize to many observations, including the background (first guess)

$$x_1 \longrightarrow y^o$$

$$x_2 \longrightarrow x^f$$

Statistical assimilation

- Observations have errors
- The forecast has errors

$$y^o = x^t - \varepsilon^o$$

$$x^f = x^t + \varepsilon^f$$

The truth x^t is unknown

Statistical assimilation

- Best linear estimate: combination between the background x^f at observation locations and the observations y^o themselves
- Think in terms of *averages*
- We do not know the truth, so we look for the *maximum likelihood estimate*, or the *minimum-variance estimate*

Statistical assimilation

- We don't know truth, so we can't know the errors.
- We have at least a chance of estimating the *error variances*
- Making these estimates is the heart of statistical data assimilation

$$\sigma_o^2, \sigma_f^2$$

Statistical assimilation

- The goal can be restated: find the best α that minimizes the analysis error on average

$$\varepsilon^a = x^t - x^a$$

- The analysis is a combination between the observation and the forecast at the observation locations

$$x^a = \alpha x^f + (1-\alpha) y^o$$

Statistical assimilation

- It turns out that the best estimate is achieved when:

α

$$x^a = x^f + \frac{\sigma_f^2}{\sigma_f^2 + \sigma_o^2}(y^o - x^f)$$

Observation increment

Analysis increment

RTFDDA

NCAR

- How does it relate to statistical assimilation?

$$x^a = x^f + \frac{\sigma_f^2}{\sigma_f^2 + \sigma_o^2}(y^o - x^f)$$

$$\downarrow$$

$$x^a = x^f + G(y^o - x^f)$$

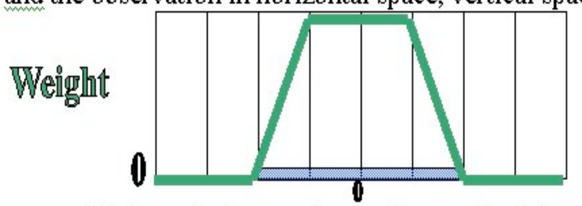
G includes it all

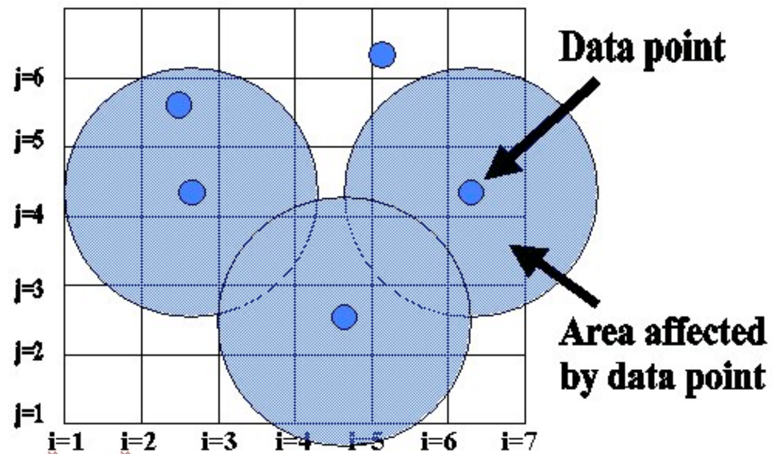
- Distance to the obs
- Time from the obs
- Expected obs error
- Quality control

Obs-nudging: Weighting Functions

$$G = W_{qf}\, W_{horizontal}\, W_{vertical}\, W_{time}$$

The weights can vary with the distance between the grid points and the observation in horizontal space, vertical space, and time.





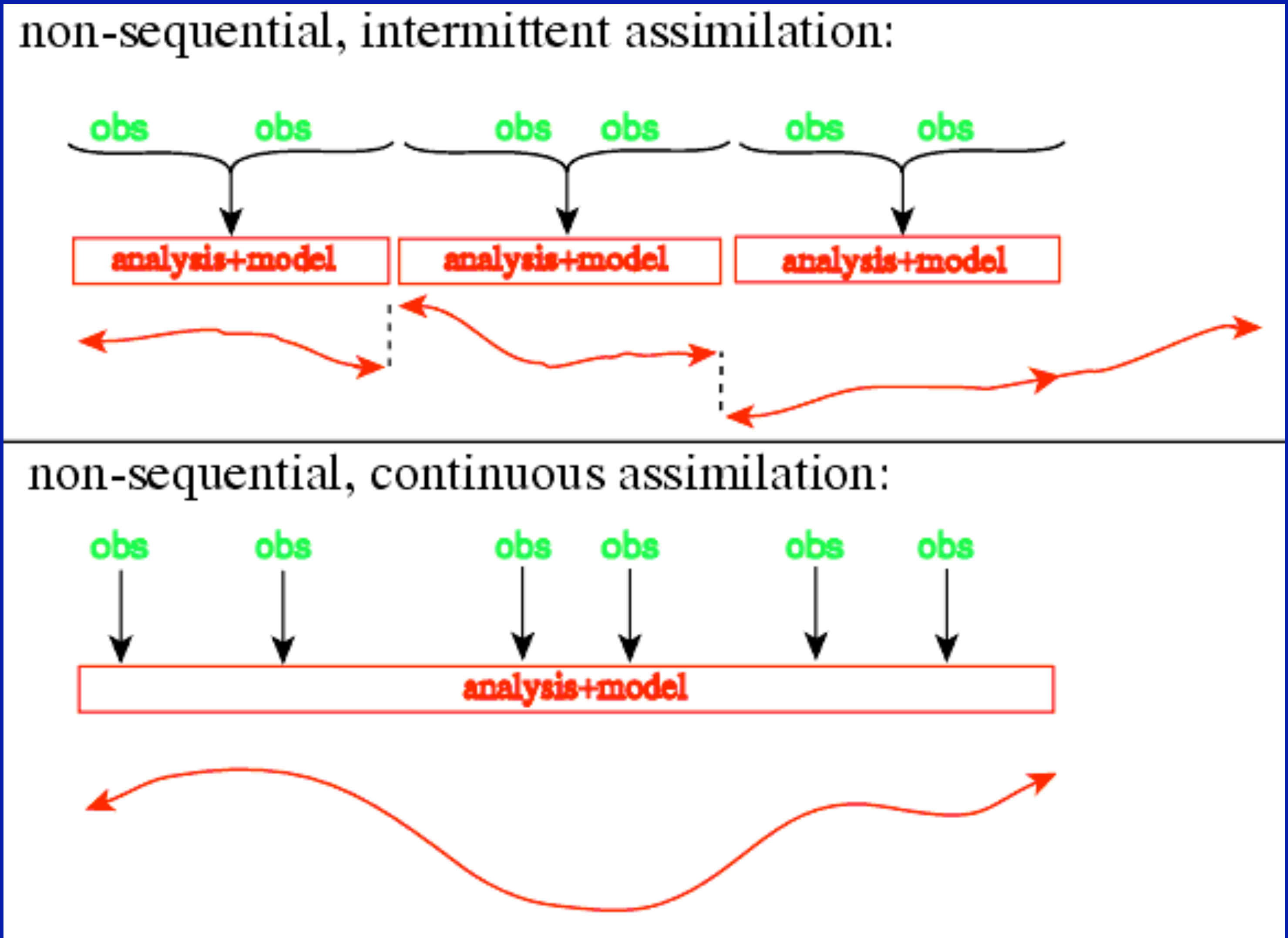
$$A_{ijk} = \frac{\sum_i W_i \cdot value_i}{\sum_i W_i}$$

At each grid point, the weighted effects from all the nearby observations are summed.

The fourth dimension: time

- Sequential assimilation
- Continuous assimilation
- Putting it all together

Variational methods

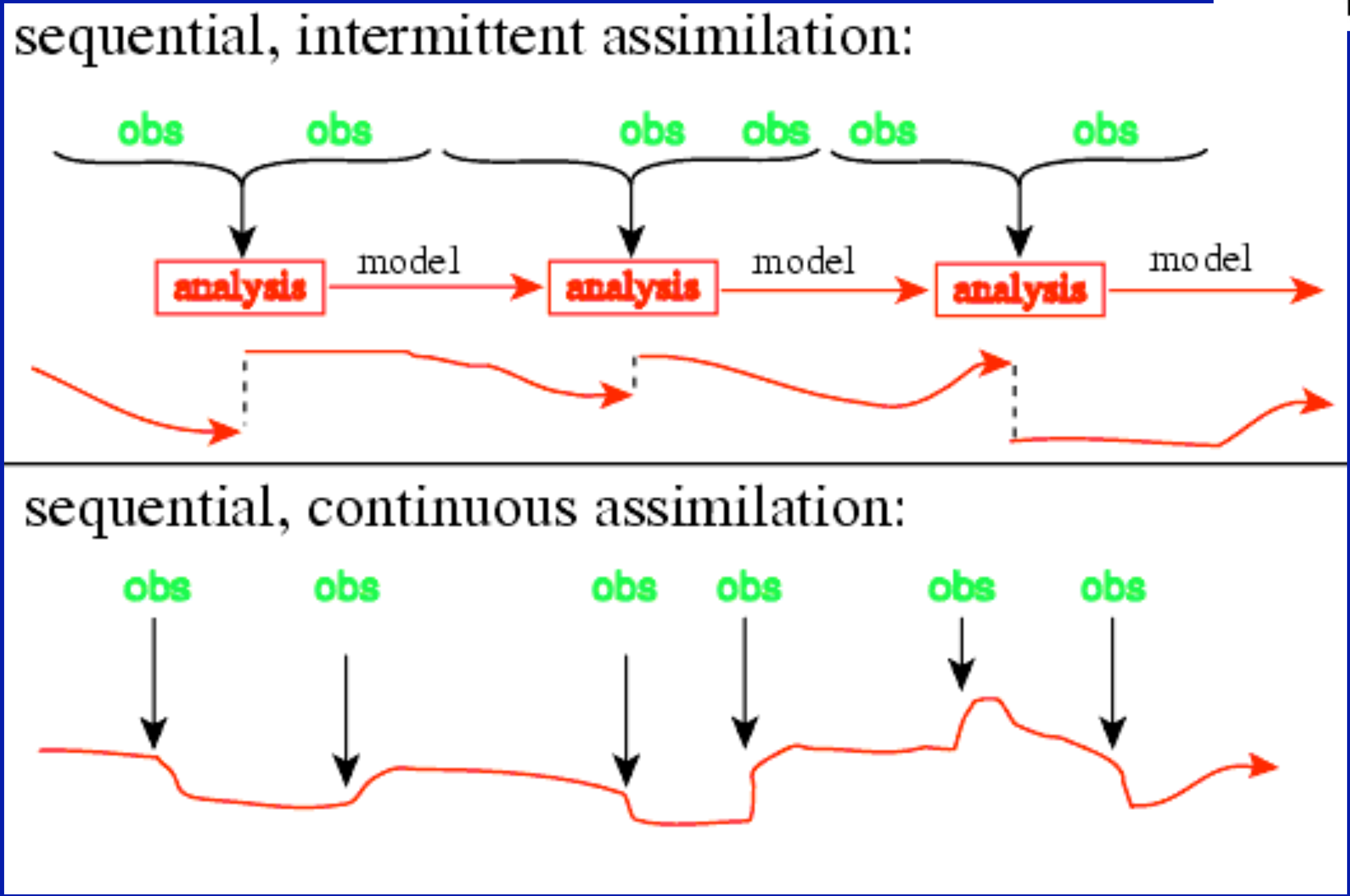


3DVAR

4DVAR

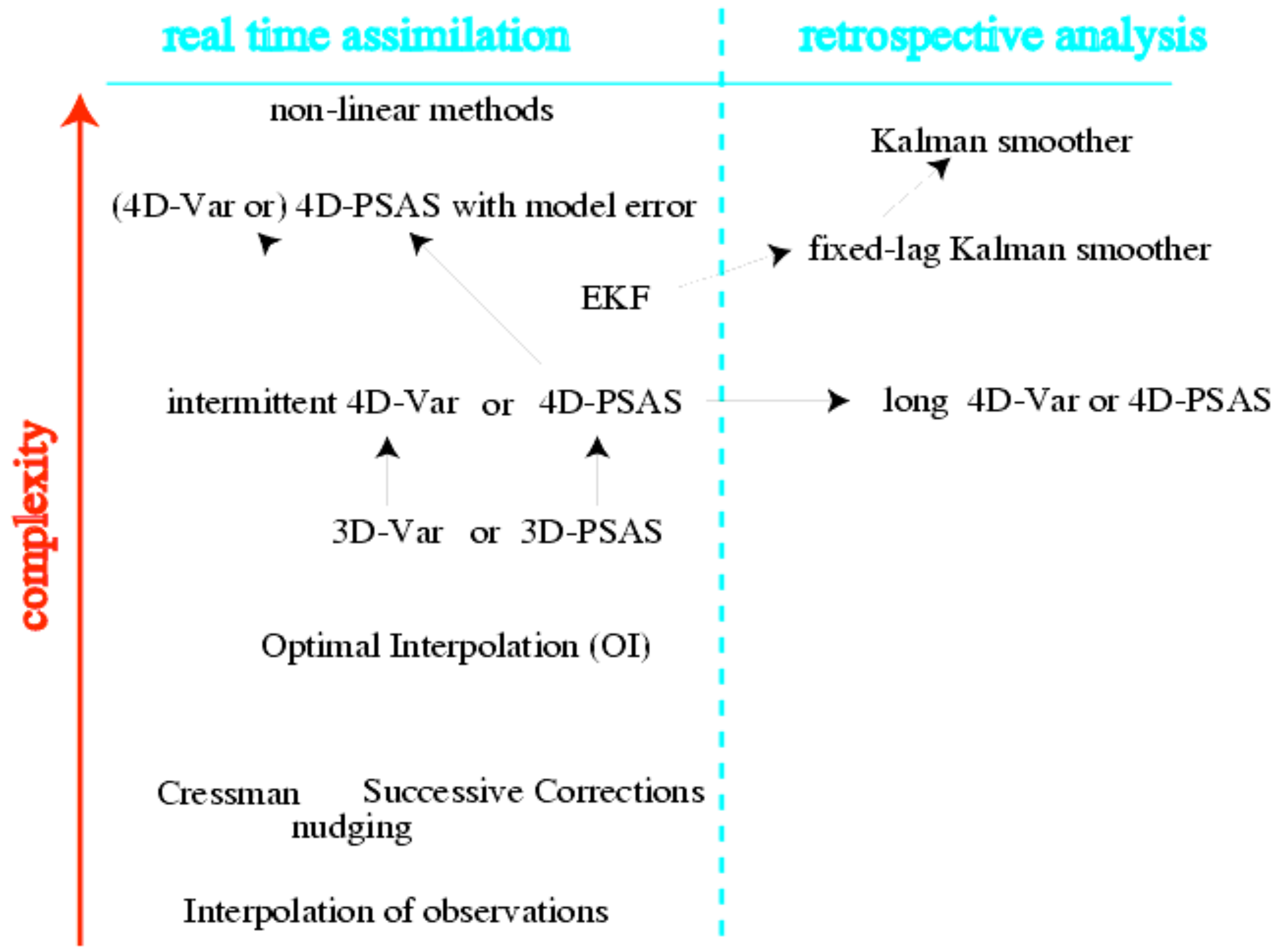
Very complex, maybe not better.

Sequential assimilation



This is RTFDDA!

Complexity


real time assimilation
retrospective analysis
complexity
non-linear methods
Kalman smoother
(4D-Var or) 4D-PSAS with model error
fixed-lag Kalman smoother
EKF
intermittent 4D-Var or 4D-PSAS
long 4D-Var or 4D-PSAS
3D-Var or 3D-PSAS
Optimal Interpolation (OI)
Cressman
Successive Corrections
nudging
Interpolation of observations