Pre- and Post-Processing in Data Assimilation

Florence Rabier
CNRM-GAME, Météo-France and CNRS
Introduction

• Data assimilation: art of combining model and observations

• It relies on a set of equations with a solid statistical basis

• Theoretical studies:
  • how to define optimally the various quantities
  • how to combine all the flow-dependent information

• In practice, a lot of attention has to be paid to details in
  • handling observations
  • possible filtering of the resulting analysis
Outline

Transforming the raw data
  Transforming into a different space
  Averaging the data
  Filtering the observations

Comparing model and observations
  Monitoring and choice of observations
  Bias correction
  Removing wrong data

Thinning the data
  Reducing data quantity and error correlation
  Choosing the most relevant local data
  Selective thinning depending on the flow

Filtering the analysis
  Initialisation methods
  Influence on the analysis
Outline

Transforming the raw data
  Transforming into a different space
  Averaging the data
  Filtering the observations

Comparing model and observations
  Monitoring and choice of observations
  Bias correction
  Removing wrong data

Thinning the data
  Reducing data quantity and error correlation
  Choosing the most relevant local data
  Selective thinning depending on the flow

Filtering the analysis
  Initialisation methods
  Influence on the analysis
Outline

Transforming the raw data
   Transforming into a different space
   Averaging the data
   Filtering the observations

Comparing model and observations
   Monitoring and choice of observations
   Bias correction
   Removing wrong data

**Thinning the data**
   Reducing data quantity and error correlation
   Choosing the most relevant local data
   Selective thinning depending on the flow

Filtering the analysis
   Initialisation methods
   Influence on the analysis
Outline

Transforming the raw data
  Transforming into a different space
  Averaging the data
  Filtering the observations

Comparing model and observations
  Monitoring and choice of observations
  Bias correction
  Removing wrong data

Thinning the data
  Reducing data quantity and error correlation
  Choosing the most relevant local data
  Selective thinning depending on the flow

Filtering the analysis
  Initialisation methods
  Influence on the analysis
From the raw data to observations

- Some observations are simple measurements: radiosondes

- Other observations are very indirect measurements:
  - Series of images from satellites to derive atmospheric motion vectors
Pre-processing the data to data easier to assimilate

- Radiances measure the electromagnetic spectrum
- Provide indirect information on temperature, humidity, surface, ozone…
- Can be used directly in data assimilation schemes (John Derber's talk)
- or via retrievals,
- or a mix, for some parameters and quality control
Averaging the data: spatially

Horizontal averaging performed by data producers

CSR = Clear-Sky radiances, averaged in boxes
typically 16*16 pixels for SEVIRI

Some averaging can be done at the user's level

• All-sky radiances at ECMWF: averaging observations
to create AMSR-E superobs (at 80km scale, Geer and Bauer, 2010)

• NRL produces superobs from satellite winds with a complex algorithm
  • Averaging in boxes (prisms of about 2° side)
  • Prism-quartering when high degree of variability
  • U and V obs have to agree within a certain range
  • … (Pauley, 2003)

and get more positive results than other centres
Averaging the data: spatially

Radar winds in the HIRLAM 3D-Var and at NCAR for WRF

( Lindskog et al, MWR, 2004) (Zhang et al, MWR, 2009)

Horizontal averaging to create super-obs from radial winds

Quality control steps

* At least 4 or 5 data in an bin
* Accepted only for low variance of the Vr values

Lindskog, Salonen, Järvinen, Michelson, MWR, 2004
Averaging the data: spatially

With a median concept
(Montmerle and Faccani, MWR, 2009)

* Median filter on boxes of 5*5 pixels
Replace value by the median
of neighbouring points

* and a « cleaner » filter, removing pixels
when large inconsistencies within boxes

---

Fig. 2. Example of radial velocities from the first elevation performed by the BOLL radar (Fig. 1): raw data (top) and after the application of median (middle) and cleaner filters (bottom). Positive velocities point towards the radar.
Averaging the data: temporally

Ground-based GPS
- Time-averaging of observations, 30 to 60 minutes.
  - Poli et al (JGR, 2007),
  - McPherson, Deblonde, Aparicio (MWR, 2008)

Radar data
- NCEP Stage IV radar and gauge precipitation data at ECMWF
  - Hourly data, but 6-hourly accumulations perform better
  - Correlation between departure computed in full trajectory (T799, full physics) and in first minimisation (T95, simplified physics): 0.2 to 0.7
  - 6-hour = compromise between linearity and observation usage over the 4D-Var 12-hour window (Lopez, MWR, 2011)
From the raw data to observations

Vertical choice of data for conventional observations: radiosondes

- Usually **selecting** all levels
- **Interpolating** between significant and mandatory levels (Benjamin et al, MWR, 2004)
- In the future, vertical **averaging** for radiosonde high-resolution profiles?
From the raw data to observations

- New challenges: transforming hyperspectral sounder data
- Channel selection for IASI (Collard, QJRMS, 2007)

Based on the information content
- Test which channel most improves DFS = \text{tr}(I-AB^{-1})
- Update the optimal A matrix
- Choose the next best channel
- ....
From the raw data to observations

Principal component compression (Collard et al, QJRMS, 2010)

PCs computed from a large set of spectra:
\[ C = \frac{1}{n} X X^T = L \Lambda L^T \]

Principal component amplitudes related to observed radiances:
\[ p = L^T y \]

Around 200 PCs required to represent signal, rest is noise.

Using leading PCs efficient for data transfer and noise filtering.

Figure 8. A comparison of the standard deviations of clear-sky departures from the same model background for real (dark/red) and reconstructed (light/green) radiances. Significant ‘denoising’ is seen in the 15 \( \mu \text{m} \) band where instrument noise is dominant over model error. One channel, at 8.07 \( \mu \text{m} \), has an apparent increase in departure standard deviation, but this is an artifact arising from the cloud detection scheme. This figure is available in colour online at wileyonlinelibrary.com/journal/qj
Monitoring and choice of relevant observations

Change in vertical resolution: better fit to high-peaking channels, AMSU-A ch 13

2006: Change from 41 to 46 levels: 5 more channels up to 0.05 hPa
2008: Change from 46 to 60 levels: more channels in stratosphere.
2010: Change from 60 to 70 levels: more channels in troposphere

Large reduction in STD when more levels in the stratosphere
In Honda and Yamada (SOLA, 2007):
Radar rain-gauge data assimilated in 4D-Var with simplified cloud microphysics

Exp A: no radar Rain-gauge data
Exp B: 1-hour rain > 0.5mm
Exp C: 1-hour rain > 0mm

Including more data can remove spurious precipitation
Monitoring and bias correction

Part of the biases seen in monitoring is attributed to observations

Bias correction scheme: from simple to elaborate

**GPS data:** Bias correction simply based on averaged deviation from model (Poli et al, JGR 2007), or on a 10-day running mean (McPherson et al, MWR, 2008)

**Radiosondes:** bias depends on a few factors

- Sonde type: Vaisala RS-80, RS-92, MODEM… sonde age
- Solar elevation: causes solar heating of the sensor
- Pressure level: the amount of solar radiation varies with pressure level
- Wetting of the radiosonde sensor in cloud can cause a wet bias at higher levels …

Monitoring and bias correction

Agusti-Panareda et al, QJRMS, 2009: bias-correction assuming the night-time RS-92 is bias-free, using the model as an intermediate.

Refined correction takes into account the dependence of the bias on the observed humidity.

CDF matching, then fitting four-sine wave components of a Fourier series.

The bias correction is computed using Equation (1) by subtracting the bias function of the reference sonde (sonde type BUFR code $s = 79$) from the bias function of the sonde to be corrected. The corrected RH ($R_{corr}$) for an observed RH value ($R_{obs}$) is given by:

$$R_{corr}(p, \theta, s) = R_{obs}(p, \theta, s) - [BIAS(R_{obs}, p, \theta, s) - BIAS(R_{obs}, p, \theta < 0, s = 79)] \quad (2)$$
Monitoring and bias correction

Radiances: a priori knowledge about the parameters affecting obs bias

Harris and Kelly (QJRMS, 2001) use scan-dependence and air-mass predictors
  Model thicknesses (1000-300hPa, 200-50hPa, …)
  Model surface temperature
  Model TCWV…

Regression coefficients are computed over a long time-series.

Can be adapted before each analysis off-line, or inside the assimilation (VarBC, see talk from John Derber)
Removing wrong data

Each observation is subject to a variety of errors

• biases from calibration…
• random errors
• representativeness errors
• gross errors: instrument malfunction, transmission error…

Data with gross errors are useless

Need for a quality control step
Removing wrong data

**Blacklist**

based on monthly monitoring generally,

can also be dynamically updated, based on gross-error statistics from the previous analyses (De Pondeca et al, WAF, 2011)

**Check for observation consistency**

« Buddy checks »

Check with observation consistent with neighbours (Benjamin et al, MWR, 2004)

Estimate of the innovation at the observation point from the innovations of a group of nearby observations

If the difference between the estimated and observed innovations exceeds a threshold, the observation is discarded
Removing wrong data

Check with model « First-guess check »

Gross check tests based on the comparison of departures with error estimates

\[(O-G)^2 < a (\sigma_o^2)\]  \hspace{1cm} \text{(De Pondeca et al, WAF, 2007)}

\[(O-G)^2 < a (\sigma_b^2)\]  \hspace{1cm} \text{(Benjamin et al, MWR, 2004)}

\[(O-G)^2 < a (\sigma_o^2 + \sigma_b^2)\]  \hspace{1cm} \text{(Lorenc and Hammon, QJRMS, 1988)}

\((\sigma_o^2 + \sigma_b^2)\) from accumulated statistics of departures (Cucurull et al, MWR, 2007),

or from the values used in the assimilation
Lorenc and Hammon. 1988

Figure 1. Probability density functions for background, observation, and Bayesian analysis, for four different observed values and a Gaussian observational error distribution.

Figure 2. As Fig. 1 for an observational error distribution equal to a Gaussian plus a small constant.
Removing wrong data: combination of tests

Different norms can be used (ex: Huber norm at ECMWF)

to represent departure statistics inside the assimilation

and adapt the prior FG-check

The pdf for the Huber norm is:

\[
p(y | x) = \begin{cases} 
\frac{1}{\sigma_o \sqrt{2\pi}} \exp \left( \frac{a^2}{2} - |a\delta| \right) & \text{if } a < \delta \\
\frac{1}{\sigma_o \sqrt{2\pi}} \exp \left( - \frac{1}{2} \delta^2 \right) & \text{if } a \leq \delta \leq b \\
\frac{1}{\sigma_o \sqrt{2\pi}} \exp \left( \frac{b^2}{2} - |b\delta| \right) & \text{if } \delta > b 
\end{cases}
\]

where \( \delta = \frac{y - H(x)}{\sigma_o} \)
Adaptive buddy check
flow-dependent tolerances for outlier observations
(Dee et al, QJRMS, 2001)

Dec 1999 storm

Removing wrong data: combination of checks

Non-adaptive

Adaptive

Adaptive buddy check

- rejected by buddy check
- passed buddy check
- passed FG check
Removing wrong data  Dependance on the errors of the day

Errors of the day provided by the Ensemble Data Assimilation.

New operational applications (in 2008 at Meteo-France for example)

Klaus: 24/01/2009 at 00h/03h

Errors for 3-hr fcst from the Ens Assim.

Berre and Desroziers, pers comm
Different analysis schemes use different temporal thinning of data

In 4D-Var, one groups observations in 30 or 60 minute time-slots and thin observations within each time-slot

In 3D-Var, select data closer to central analysis time (ex: +/- 1.5 hour for aircraft data)

In non-cycled schemes, choice of data really representative of analysis time
ex for the hourly Real-time Mesoscale Analysis (De Pondeca et al, WAF, 2011), time window of –12 to +12 minutes
For practical reasons, and avoiding obs error correlations not accounted for

Liu, 2002
\(\Delta x=100\text{km}, \sigma_b=\sigma_o=1\)
\(L_b=208\text{km}, L_o=100\text{km}\)

Optimal distance can be found

Evidence of error correlation exist in AMVs, radiances
(Bormann et al, QJRMS 2003; Bormann and Bauer, QJRMS, 2010)
Horizontal thinning

Generally, simple thinning by lat-lon boxes, with choice by quality criteria

(distance to guess, Quality Indicator, small value of radial wind variance in the superobs, max number of elevations which pass QC in radar profiles…).

Adaptive thinning: Ochotta et al, QJRMS, 2005

- Observations representative of clusters are inserted iteratively
- Or, removal of the observations from the full set, by removing redundant data

Figure 1. Concept of top-down clustering. (a) Observations are grouped to a cluster with a cluster centre (filled dot); (b) when the associated cluster error is too large, the cluster is split up by Principal Component Analysis, providing two new clusters; (c) this procedure is repeated until all cluster errors are below a given threshold, $t > 0$. The set of centroids is the reduced observation set.
Horizontal thinning

Optimal thinning distance investigated in the Met Office NWP system

(Dando, Thorpe and Eyre, QJRMS, 2007)

Control: thinning distance of 308km. Optimal distances found: 100-150km.

Detrimental to use thinning at 40-km distance, especially in Tropics (weak gradients in the fields)
Horizontal thinning

Optimal thinning using Singular vector information in Southern Hemisphere at ECMWF (Bauer et al, QJRMS, 2010). Different configurations, two seasons (JAS, DJF): 

EXP: global density of 1.25°

EXP-HI: Global High-density 0.625°

EXP-SV: High-density only in SV areas

EXP-CLI: High-density in SV-based climatological regions

EXP-RND: High-density in random areas
Horizontal thinning

Radius of Influence in EnKF, Zhang et al, MWR, 2009
Rad data assimilation, 3 domains D1 (40km) to D3 (4.5km)

**FIX1**: ROI = 1215km for D1, D2, D3  
**FIX2**: ROI = 405km for D1, D2, D3  
**FIX3**: ROI = 135km for D1, D2, D3

**CNTL**:  
ROI of 1215km for 10% of data in D1, D2, D3  
Then ROI of 405km for 20% of data in D2, D3  
Then ROI of 135km for 60% of data in D3

**DX30**:  
ROI of 1215km in D1  
Then ROI of 405km in D2  
Then ROI of 135km in D3

Better performance of **CNTL** and **DX30**

Hurricane Humberto,  
Forecast from 18UTC 12 Sep 2007
Post-processing: Filtering

The model can take time to adjust initial fields with respect to model equations. Dynamical adjustment by inertia-gravity waves, diabatic adjustment.

Ideally, balanced increments in the analysis (through B). There is also a possibility to include constraint terms inside the analysis (Gauthier and Thépaut, MWR, 2001).

Posterior filtering of the analysis is frequently performed.

Forces the initial state not to generate model tendencies that project onto high-frequencies model solutions

Different methods can be used:

**DFI**: Digital Filter Initialization (Lynch and Huang, MWR, 1990; Huang and Lynch, MWR, 1993)

**IAU**: Incremental Analysis Update (Lorenc et al, QJRMS, 1991; Bloom et al, MWR, 1996)
Initialization methods

- **DFI**:
  - Backward integration in time by Ndt, then forward integration by 2Ndt
  - Time series $X(n)$ is then filtered removing high frequencies
  - $X^* = \sum h(-n)X(n)$ where $h(n)$ are the filter coefficients
  - $h(n) = \left\{ \frac{\sin(n\pi/(N+1))}{n\pi/(N+1)} \right\} \ast \left\{ \frac{\sin(n\theta_c)}{n\pi} \right\}$
  - $\theta_c$ is the cutoff frequency

- **IAU**:
  - 3D-Var increment added gradually in the assimilation window

\[ \text{Transforming} \quad \text{Comparing} \quad \text{Thinning} \quad \text{Filtering} \]
Filtering

Imbalance depends on the quality of the analysis.

DFI applied to MM5 using either Cressman or 3D-Var analysis in Chen and Huang, MWR, 2006

DFI applied to both OI and 3D-Var versions of the RUC (Benjamin et al, MWR, 2004)

**Fig. 8.** The evolution of the mean absolute surface pressure tendency $N [\text{hPa (3 h)}^{-1}]$ in the first 12-h forecasts averaged from 14 cycles from 0000 UTC 21 Aug to 1200 UTC 27 Aug 2002.

**Fig. 2.** Noise parameter over a single time step (30 s) in the RUC model with 3DVAR or OI analysis, both with and without application of DFI. For the case with initial conditions at 1200 UTC 19 Nov 2002, data points taken every 30 min of integration.
Various flavours of DFI: diabatic versus adiabatic (Huang and Lynch, MWR, 1993), incremental versus non-incremental (Fischer and Auger, MWR, 2011)

**Standard DFI:** \( X_a^* = \text{DFI}(X_a) \)

Total increment for standard DFI is

\[
\text{DFI}(X_a) - X_b = \text{DFI}(X_a) - \text{DFI}(X_b) - (X_b - \text{DFI}(X_b))
\]

The total increment is the sum of

- a balanced increment
- a removal of the high frequencies in \( x_b \)

**Incremental DFI:**

\( X_a^* = X_b + \{ \text{DFI}(X_a) - \text{DFI}(X_b) \} \)

Total increment is \( \text{DFI}(X_a) - \text{DFI}(X_b) \)
Filtering not only for forecasts, but also for assimilation.

For rapid cycles, the assimilation cycle could be adversely affected by spurious waves

Brousseau, pers comm
Impact of initialization in AROME

Example of one precipitating event over SE France (Brousseau, pers comm)

- Date 15/06/2010:
  - 3-h and 1-h cycling perform similarly
  - IAU improves location
Conclusion

Le Bon Dieu est dans le détail, Gustave Flaubert, 1821-1880

Or

The devil is in the details