Validation of Complex Data Assimilation Methods

Hendrik Elber, Elmar Friese, Nadine Goris, Lars Nieradzik
and many others
Rhenish Institute for Environmental Research at the University of Cologne
and
Institute for Energy and Climate Research (Troposphere)
Forschungszentrum Jülich
Contents

1. Intro.: What are complex data assimilation methods?

2. Observability: Do observations sustain assimilation results?

3. Practical verification: Validation by forecast skills

4. A posteriori Validation: Is the analysis consistent?
What are complex data assimilation methods?
→ spatio-temporal techniques

2 types of assimilation algorithms: “smoother” and filter
The 4-dimensional variational technique: **Optimize over an assimilation window, then forecast**

**Emission Rate Optimization**

minimize cost function

\[
J(x(t_0), e) = \frac{1}{2} (x^b(t_0) - x(t_0))^T B_0^{-1} (x^b(t_0) - x(t_0)) + \\
\frac{1}{2} \int_{t_0}^{t_f} (e_b(t) - e(t))^T K^{-1} (e_b(t) - e(t)) dt + \\
\frac{1}{2} \int_{t_0}^{t_f} (y^0(t) - H[x(t)])^T R^{-1} (y^0(t) - H[x(t)]) dt
\]

- \( x^b(t_0) \): background state at \( t = 0 \)
- \( x(t) \): model state at time \( t \)
- \( e_b(t_0) \): background emission rate at \( t = 0 \)
- \( e(t) \): emission rate field at time \( t \)
- \( K \): emission rate error covariance matrix
- \( H[\ ] \): forward interpolator
- \( y^0(t) \): observation at time \( t \)
- \( B_0 \): background error covariance matrix

deviations from background initial state

deviations from a priori emission rates

model deviations from observations
Kalman filter: basic equations

Forecast steps:
a) the atmospheric state

\[ x^f(t_i) = M(t_i, t_{i-1})x^a(t_{i-1}) + \eta \]

b) the forecast error covariance matrix

\[ P^b_i = M(t_i, t_{i-1})P^a_{i-1}M^T(t_i, t_{i-1}) + Q \]

Analysis steps:
a) the atmospheric state

\[ x^a(t_i) = x^b(t_i) + K_id_i, \quad (1) \]

\[ K_i := P^b_iH_i^T(H_iP^b_iH_i^T + R_i)^{-1} \in \mathcal{R}^{n \times p_i} \quad (2) \]

and b) the analysis error covariance matrix

\[ P^a_i = (I - K_iH_i)P^b_i. \quad (3) \]
Computational challenge:
Background Error Covariance Matrix $P^b$

1. Ensemble approach: (e.g. Evensen, 1994)

$$B_{ij} = \frac{1}{K} \sum_{n=1}^{K} (x^n_i - \bar{x}_i)(x^n_j - \bar{x}_j)$$

Ensemble integration
$K = \#\text{ ensemble members}$
$i, j \text{ grid cells}$
2. Observability: Do observations sustain assimilation results?  
Observation network design

Is the forecasted system sensitive to available observations?

– Observation System Simulation Experiments (OSSEs)

– Targeted observations
Is NO\textsubscript{x} the key to ozone production?
And consequently, its observation\textsuperscript{the} key to better forecast?

Calculations
✓ within a fixed time span
✓ initial concentrations of NO / HCHO were varied
✓ change of final concentration is given by colour
✓ gradients (SVs) of maximyl ozone production given by arrows
# How can we optimize the observation configuration?

Given CTM (here RACM and EURAD-IM) acting as tan.-lin. model operator $\mathcal{L}$:

$$
\delta c(t_F) = \mathcal{L}_{t_I, t_F} \delta c(t_I), \quad \mathcal{L}_{t_I, t_F} = \frac{\partial M_{t_I, t_F}}{\partial c}
$$

| 1. Berliner et al., (1998) Statistical design: |
| “Minimize” the analysis error covariance matrix $\mathbf{A}$ (say, via trace): |

For this find maximal eigenvectors as observation operators $\mathbf{H}$, which configure observations.

| 2. Palmer (1995) Singular vector analysis: |
| Observe maximal SV configuration: |

$$
\min_{\mathbf{H}} \mathbf{A} = \mathbf{B} - \mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H}\mathbf{B}
$$

to be maximized by $\mathbf{H}$

$$
\mathcal{L}_{t_I, t_F} \mathbf{B} \mathcal{L}_{t_I, t_F}^T \mathbf{H}^T = \lambda \mathbf{H}^T
$$

$$
\max_{\delta c(t_I)} \frac{\|\delta c(t_I)\|^2_{\mathbf{B}}}{\|\delta c(t_I)\|^2_{\mathbf{B}}} = \max_{\delta c(t_I)} \frac{\delta c(t_I)^T \mathcal{L}_{t_I, t_F}^T \mathbf{B} \mathcal{L}_{t_I, t_F} \delta c(t_I)}{\delta c(t_I)^T \mathbf{B} \delta c(t_I)}
$$
Basic 0-D Regional Atm. Chemistry Mechanism („\(\mathcal{M}=\text{RACM}\)“)

- **Optimal perturbations (Singular Vectors) for** scenario MARINE
  - 1\textsuperscript{st} Grouped Singular Vectors (\(\delta\text{VOC}\))
  - 1\textsuperscript{st} Grouped Singular Vectors (\(\delta\text{NO}_x\))

- Forecasted time:
  - *sunrise*: 0.00E+00
  - *sunset*: 1.00E+00

- Initial time [h]:
  - 0.00E+00
  - 1.00E+00

- Initial time [h]:
  - 0.00E+00
  - 1.00E+00

- Final time [h]:
  - 0.00E+00
  - 1.00E+00

- Final time [h]:
  - 0.00E+00
  - 1.00E+00

- Not very important to observe
3. Practical verification: Validation by forecasts

Analysis of emissions by 4D-var (VERTIKO)

Observed and analysed ozone evolution at St. Poelten Vertical bars: ozone observations with error estimates.
- - - - - Control run without data assimilation.
- - - - - initial value optimisation.
- - - - - emission factor optimisation.
- - - - - joint initial value and emission factor optimisation (Strunk et al., 2011)
4. Focus: joint emission rate initial value optimisation

Semi-rural measurement site Eggegebirge


+ observations
no optimisation

initial value opt.

emis. rate opt.

joint emis + ini val opt.
How long does data assimilation have an impact?

Answer: Gas phase

12-24 hours, dependent on optimisation

**Bias**

<table>
<thead>
<tr>
<th>Observation Type</th>
<th>Method Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ observations</td>
<td>no optimisation</td>
<td>initial value opt.</td>
</tr>
<tr>
<td>emis. rate opt.</td>
<td></td>
<td>joint emis + ini val opt.</td>
</tr>
</tbody>
</table>

**Root Mean Square**

<table>
<thead>
<tr>
<th>Observation Type</th>
<th>Method Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ observations</td>
<td>no optimisation</td>
<td>initial value opt.</td>
</tr>
<tr>
<td>emis. rate opt.</td>
<td></td>
<td>joint emis + ini val opt.</td>
</tr>
</tbody>
</table>
Which is the requested resolution?

BERLIOZ grid designs and observational sites

(20.→21. 07.1998)

\[ \Delta x = 54 \text{ km} \]

\[ \Delta x = 18 \text{ km} \]

\[ \Delta x = 6 \text{ km} \]

\[ \Delta x = 2 \text{ km} \]
Some BERLIOZ examples of NOx assimilation (20.→21. 07.1998)

Time series for selected NOx stations on nest 2.
+ observations,
-- no assimilation,
- N1 assimilation (18 km),
- N2 assimilation (6 km),
grey shading: assimilated observations, others forecasted.
Validation by measurements withheld
(extract from MACC III EDA report draft)

**Forecast**

**O₃** Kosice

<table>
<thead>
<tr>
<th>ModelType</th>
<th>Meas</th>
<th>Modeled</th>
<th>Anal. 1yr</th>
<th>Anal. 2yr</th>
<th>Anal. 3yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>67.4</td>
<td>57.1</td>
<td>57.1</td>
<td>57.1</td>
<td>57.1</td>
</tr>
<tr>
<td>Eval</td>
<td>65.2</td>
<td>57.1</td>
<td>57.1</td>
<td>57.1</td>
<td>57.1</td>
</tr>
</tbody>
</table>

**Analyses**

**O₃** Kosice

<table>
<thead>
<tr>
<th>ModelType</th>
<th>Meas</th>
<th>Modeled</th>
<th>Anal. 1yr</th>
<th>Anal. 2yr</th>
<th>Anal. 3yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>67.4</td>
<td>57.1</td>
<td>57.1</td>
<td>57.1</td>
<td>57.1</td>
</tr>
<tr>
<td>Eval</td>
<td>65.2</td>
<td>57.1</td>
<td>57.1</td>
<td>57.1</td>
<td>57.1</td>
</tr>
</tbody>
</table>

**PM₁₀** Schmāuoecke

<table>
<thead>
<tr>
<th>ModelType</th>
<th>Meas</th>
<th>Modeled</th>
<th>Anal. 1yr</th>
<th>Anal. 2yr</th>
<th>Anal. 3yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>72.5</td>
<td>72.5</td>
<td>72.5</td>
<td>72.5</td>
<td>72.5</td>
</tr>
<tr>
<td>Eval</td>
<td>72.5</td>
<td>72.5</td>
<td>72.5</td>
<td>72.5</td>
<td>72.5</td>
</tr>
</tbody>
</table>

How long does data assimilation have an impact?

Answer: Aerosol phase aerosol data assimilation effects accumulate.

No previous assimilation

Only 1 day: 14. July 2003

Accumulation of retrieval information over 14 days
MOCAGE satellite data assimilation: IASI SOFRID O₃ re-analysis (CERFACS)

Validation of IASI analysis with ozonesonde data:
BIAS = model minus observations

- Bias reduced in the free troposphere
- Surface ozone impact is minor
- MOZAIC-IAGOS as additional validation? (only 2012 available)

O₃ profiles in July 2010:

O₃ profiles in Jan 2012:

Courtesy E. Emili, CERFACS
4. A posteriori Validation: Is the analysis consistent?

**a posteriori validation** of data assimilation results

Assumptions:

- Gaussian error distribution assumption sufficiently valid
- First guess not too far from “solution” (tangent-linear approximation must hold)
- A priori defined error covariances (background, observations)

**Necessary condition for a posteriori validation:**

Adjust $B$ and $R$ such that:

- Expectation
- Variance

\[
J_{\min} = \frac{1}{2} d^T (HBH^T + R)^{-1} d
\]

\[
d := y - Hx^a
\]

$p$ number of observations

\[
\mathbb{E}[J_{\min}] = \frac{p}{2}
\]

\[
\mathbb{V}[J_{\min}] = \frac{p}{2}
\]
Evaluating the Gaussian error distribution assumption

SACADA

O-F differences (left column) and

O-A differences (right column)

Dotted line represents a Gaussian with same variance as the data

HNO$_3$

HNO$_3$

ClONO$_2$

ClONO$_2$

O$_3$

O$_3$
**χ² validation MOCAGE**

**Comments:**
- **O₃**
  - the urban case is the only case with a distinct winter-summer behavior (higher χ² in winter)
  - presence of diurnal variability in all cases

- **NO₂**
  - large differences between rural/urban cases
  - strong variations in the rural case
  - presence of diurnal variability in all cases
  - no evidence of significant seasonality

**Graphs**
- Surface O₃ assimilated
- Surface NO₂ assimilated
- Winter period (1-2-2008, 6-8,2008)
- Summer period (1-8-2008, 6-8-2008)
- Only rural background sites assimilated
- Only urban background sites assimilated

*Courtesy E. Emili, CERFACS*
$\chi^2$ validation MOCAGE

What is the impact of a low $\chi^2$ in terms of validation with an independent dataset? Example: O$_3$ background urban sites assimilated in summer, validation against sites kept out from the assimilation, two choices of the background error variance $\sigma$

Comments:
Case 2 ($\sigma = 40\%$) has lower $\chi^2$ but better analysis scores. A better $\chi^2$ does not always imply a better analysis, because $\chi^2$ stats do not consider model biases.
Conclusions

• Atmospheric chemistry is a highly coupled nonlinear dynamic system, which is best addressed by spatio-temporal data assimilation.

• The system must be observed with respect to its sensitivity (NOx-VOX interaction).

• Forecasts must be shown to improve.

• The assimilation result must be consistent: proper balance between a priori and a posteriori knowledge ($\chi^2$-validation).
Additional illustrations
2. Focus: Can we identify flaws? 
A posteriori evaluation

1. $\chi^2$ – validation

2. a posteriori validation in observation space
Theoretical background on a posteriori evaluation

\[ J_{\text{min}} = \frac{1}{2} d^T E \hat{d} d^T d \]

\[ \mathbb{E}(J_{\text{min}}) = \frac{p}{2} \]
2. Focus: a posteriori validation

Aposteriori validation in observation space

Extended Kalman filter equations

Forecast step: \[ \mathbf{x}^b(t_i) = M_{i-1} \mathbf{x}^a(t_{i-1}) \]

\[ \mathbf{B}(t_i) = L_{i-1} A(t_{i-1}) L_{i-1}^T + Q(t_{i-1}) \]

Analysis step: \[ \mathbf{x}^a(t_i) = \mathbf{x}^b(t_i) + \mathbf{K}(t_i) (\mathbf{y} - H \mathbf{x}^b(t_i)) \]

\[ \mathbf{A}(t_i) = (\mathbf{I} - \mathbf{K}(t_i) H) \mathbf{B}(t_i) \]

where

- \( M_i := \) Model operator
- \( L_i := \) Tangent linear model operator
- \( \mathbf{K}(t_i) := \mathbf{B}(t_i) H^T \left[ R - H \mathbf{B}(t_i) H^T \right]^{-1} \)
- \( Q(t_i) := \) Model error covariance matrix
- \( \mathbf{B}(t_i) := \) Background error covariance matrix
- \( \mathbf{A}(t_i) := \) Analysis error covariance matrix
- \( \mathbf{R}(t_i) := \) Observation error covariance matrix

optimize R and B directly, and A indirectly
Diagnosis and Tuning of Error Covariances

(Desroziers et al. 2005)

\[
E \left\{ d_b^a d_b^{oT} \right\} = H \tilde{B} H^T
\]

\[
E \left\{ d_a^o d_b^{oT} \right\} = \tilde{R}
\]

\[
d_b^a := H(x^a) - H(x^b)
\]

\[
d_a^o := y - H(x^a)
\]

\[
d_b^o := y - H(x^b)
\]

If \( B \) and \( R \) are \textit{consistently} specified, then \( B = \tilde{B} \) and \( R = \tilde{R} \) and

\[
E \left\{ d_b^a d_a^{oT} \right\} = HAH^T
\]

Only a necessary, but not a sufficient condition is fulfilled: no unique solution
Tuning of Error Covariances in observation space
(Desroziers et al. 2005)

\[ E \left\{ d^a_b d^c_b T \right\} = HBH^T \]  \hspace{1cm} (1)

\[ E \left\{ d^c_a d^c_b T \right\} = R \]  \hspace{1cm} (2)

\[ E \left\{ d^c_b d^c_b T \right\} = HBH^T + R \]  \hspace{1cm} (3)

\[ E \left\{ d^a_b d^c_a T \right\} = HAH^T \]  \hspace{1cm} (4)

if $B$ and $R$ are correctly specified.

\[ d^a_b := H(x^a) - H(x^b) \]

\[ d^c_a := y - H(x^a) \]

\[ d^c_b := y - H(x^b) \]

in practice: Iterative approach
Practical estimate of diagonal elements of R and B

\[
(\tilde{\sigma}_i^b)^2 = (d_i^a)^T (d_i^o) = \sum_{j=1}^{p_i} (y_j^a - y_j^b)(y_j^o - y_j^b)/p_i
\]

\[
(\tilde{\sigma}_i^o)^2 = (d_i^o)^T (d_i^o) = \sum_{j=1}^{p_i} (y_j^o - y_j^a)(y_j^o - y_j^b)/p_i
\]

Estimate of off-diagonal elements of B

\[
(\tilde{\sigma}_{ij}^b)^2 = \sum_{i,j=1\atop i \neq j}^{p_{ij}} (y_i^a - y_i^b)(y_j^o - y_j^b)/p_{ij},
\]

Applied only along orbits in observation space

\[\Delta t < 10 \text{ min}\]
Geometrical representation of error components

H(x^t)  
\[ |\varepsilon^0| \]  
\[ |H(\varepsilon^b)| \]  
\[ |H(\varepsilon^a)| \]  
\[ |d_o^a| \]  
\[ |d_a^b| \]  
\[ |d_o^b| = |d_o^a| + |d_a^b| \]

Line of consistent definition of error covariance matrices

inconsistent formulation

amenable for a posteriori check

2. Focus: a posteriori validation