Lecture 1
Data Assimilation Basics

H. Elbern
Rhenish Institute for Environmental Research at the University of Cologne
and
Virt. Inst. for Inverse Modelling of Atmospheric Chemical Composition
Objective of this lecture

Understanding of

• the fundamental objectives of data assimilation and inverse modelling

• Theory of data assimilation
  – three basic sources of information
  – combination of information sources
  – the basic ideas of 4D-var and Kalman-filter
  – the limiting assumptions and a look forward
General textbook literature for data assimilation

  - Well established connection between statistics and practical data assimilation
  - Meanwhile behind cutting edge operational data assimilation implementations


  - Mathematically very sound, with focus on oceangraphy
  - Emphasis on science, rather than operational usage

  - Well presented pedagogical introduction to data assimilation theory

  - Mathematically rigorous formalism with Optimum Interpolation as central algorithm

Additionally, a collection of overview papers:
Underlying mathematics:


Terminology

Inverse Modelling

The inverse modelling problem consists of using the actual result of some measurements to infer the values of the parameters that characterize the system.

A. Tarantola (2005)
Objective of atmospheric data assimilation (1)

The ambitious and elusive goal of data assimilation is to provide a dynamically consistent motion picture of the atmosphere and oceans, in three space dimensions, with known error bars.

M. Ghil and P. Malanotte-Rizzoli (1991)
Objective of atmospheric data assimilation (2)

- "is to produce a regular, physically consistent four dimensional representation of the state of the atmosphere
- from a heterogeneous array of in situ and remote instruments
- which sample imperfectly and irregularly in space and time.

Data assimilation
- extracts the signal from noisy observations (filtering)
- interpolates in space and time (interpolation) and
- reconstructs state variables that are not sampled by the observation network (completion).“ (Daley, 1997)
Data assimilation has much of curve fitting!

Given

- the "model" $y = M(x; a, b) := ax + b$, and
- observations $(x_i, y_i), \quad i = 1, \ldots, M$,

provide a best estimate of model parameters $a, b$ in a sense that

$$\min_{a,b} \left[ \sum_i (y_i - (ax_i + b))^2 \bigg| \forall \text{ observations}(x_i, y_i) \right]$$

Writing the set of observations in column vector notation $\mathbf{y}, \mathbf{x}$, minimise

$$(\mathbf{y} - M(\mathbf{x}; a, b))^T (\mathbf{y} - M(\mathbf{x}; a, b))$$
Characteristics in data assimilation, in comparison to remote sensing retrievals

- high dimensional problem: $\dim(x) > 10^5$
- highly underdetermined system (few observations with respect to model freedom: $\dim(x) >> \dim(y)$)
- nonlinear dynamics
- constraints by physical laws are insufficient
Advanced data assimilation is an application of the principles of Data Analysis.

(As in satellite retrievals) we strive for estimating a latent, not apparent parameter set \( x \).

We dispose of
1. indirect information on the state and processes in terms of manifest, though insufficient or inappropriate data \( y \), and
2. a deterministic model \( M \) connecting \( x \) and \( y \), in some way.

Bayes’ rule gives access to the probability of state \( x \), given \( y \) and \( M \)

\[
\text{prob}(x|y,M) = \frac{\text{prob}(y|x,M)\text{prob}(x|M)}{\text{prob}(y|M)}
\]
DA in the satellite data application chain

- **Level 0**: Detector tensions
- **Level 1**: Spectra
- **Level 2**: Geolocated geophysical parameters
- **Level 3**: Geophysical parameters on regular grids
- **Level 4**: User applications

**Data Retrieval**
- Prior data
- Radiance
- Data assimilation
- Inversion
- Geophysical models

**Geophysical Parameters**
- Level 0: Detector tensions
- Level 1: Spectra
- Level 2: Geolocated geophysical parameters
- Level 3: Geophysical parameters on regular grids
- Level 4: User applications
DA and related algorithms: How can dynamic system understanding be assessed?

**Sensitivity**
- How probable/stable is the simulation/state?
- Compute leading singular vectors discontinuities

**Observability**
- By which observations “how unambiguous” can the state estimate be?
- Compute (some norm of) analysis error covariance matrix

**Predictability**
- Which credibility can the prediction claim?
- Compute probability density function (in some way by “optimal perturbations”)

**Data Assimilation**
- Inverse Modelling estimate most probable state parameter values

**Ill-posedness**
- optimal perturbations
Observation configuration optimisation
(targeted observations)

Given: \[ \delta x(t) = L(t_0, t) \delta x(t_0) \]
L(t₀,t) integration operator

Problem:
1. Find perturbation \( \delta x(t_0) \), which triggers the most uncertain system evolution \( \delta x(t) \)!

Stability measure:
\[
\frac{\| \delta x(t) \|^2}{\| \delta x(t_0) \|^2} = \frac{\delta x(t_0)^T L(t,t_0)^T L(t_0,t) \delta x(t_0)}{\delta x(t_0)^T \delta x(t_0)}
\]

Solution:
Calculate eigenvectors of \( L(t,t_0)^T L(t_0,t) \) with maximal eigenvectors

2. Infer the observation configuration, which minimizes the analysis error covariance matrix \( A \) (in some suitable sense): trace, determinant, max. EV.
Ad hoc (empirical) approaches (1)
Interpolation

Approximation of field $z$ at position $(x_1^{(i)}, x_2^{(i)})$ by a local algebraic polynomial

$$z(x_1, x_2) = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{20}x_1^2 + a_{11}x_1x_2 + a_{02}x_2^2$$

Find coefficients $a_{vw}, v, w \in \{0, 1, 2\}$
the optimisation problem is then:

$$\min_{a_{vw}} \sum_{k=1}^{K_i} w_k (y_k - z(x_1^{(i)}, x_2^{(i)}))^2$$

where $K_i$ is the number of adjacent observations,
$w_k$ the empirical weighting coefficient,
dependent on the distance between observation and model grid point, and the quality of the observation.
Ad hoc (empirical) approaches (2)

Nudging

Nudging is a data assimilation method affecting the model equations, by adding a "nudging term", which denotes a weighted observation-minus-model discrepancy (Newtonian cooling formulation) at the observation location $i$:

$$\frac{\partial x_i}{\partial t} = M(x) + c((y_o)_i - x_i).$$

For example, the zonal momentum equation with local nudging term reads

$$\frac{\partial u_i}{\partial t} = -v_i \cdot \nabla u_i + f v - \frac{\partial \phi}{\partial x} + \frac{((u_o)_i - u_i)}{\tau v}$$

with $\tau := 1/c$ the relaxation constant with unit 1/time.

Caution: prone to excitement of artificial modes
Optimality criteria: Which property can be attributed to our analysis result?
(Need for quantification)
• maximum likelihood:
  – maximum of probability density function
• minimal variance: $l_2$ norm
  – parameters optimal, for which analysis error spread is minimal (for Gaussian/normal and log-normal error distributions), Best Linear Unbiased Estimate (BLUE)
• minimax norm (discrete cases)
• maximum entropy
Data assimilation: Synergy of Information sources
example: 2 data sets with Gaussian distribution
(here: minimal variance = maximum likelihood)

Bayes’ rule:
\[ p(x|y_o) \propto p(y_o|x)p(x) \]

Analysis (=estimation) **BLUE**
Best Linear Unbiased Estimate

\[
\frac{x_a}{\sigma_a^2} = \frac{y_o}{\sigma_o^2} + \frac{x_b}{\sigma_b^2}
\]
\[
\frac{1}{\sigma_a^2} = \frac{1}{\sigma_o^2} + \frac{1}{\sigma_b^2}
\]

\[
p(x|y_o) =: N(x|x_a, \sigma_a^2)
\]
\[
= \frac{1}{\sqrt{2\pi}\sigma_a} \exp \left( -\frac{(x_a - x)^2}{2\sigma_a^2} \right)
\]

\[
p(y_o|x) =: N(y_o|x, \sigma_o^2)
\]
\[
= \frac{1}{\sqrt{2\pi}\sigma_o} \exp \left( -\frac{(y_o - x)^2}{2\sigma_o^2} \right)
\]

a priori (=prediction or climatology)  
observation

Day 5  Lecture 1  Data Assimilation  Hendrik Elber
The observation

\[ y_o = x_t + \epsilon_o, \quad (1) \]

with \( x \) true value, \( \epsilon_o \) observational error and additionally the error of representativity for an analysis grid resolution, is a realisation of the random variable \( X \).

Any other datum also be considered as a realization of a random experiment.

\[ x_b = x_t + \epsilon_b \quad (2) \]

Let us be given an observation \( y_o \) and a forecast \( x_b \) for the same location and time. We seek for a \( x_a \) based on \( y_o \) and \( x_b \) and their respective error variances \( \epsilon_o \) and \( \epsilon_b \). with the smallest analysis error.
A 2 information bits example (1): Assumptions

- the expectation (mean, average, first (statistical) moment)

\[ \mathcal{E}(\varepsilon_{o/b}) := \int_{\infty}^{\infty} \varepsilon_{o/b} P(\varepsilon_{o/b}) d\varepsilon_{o/b} = 0 \]  

which demands that there is no mean or systematic error (= unbiased)

- Variance (second central moment) of error as an inverse measure of accuracy

\[ \mathcal{E}(x_b^2) := \int_{\infty}^{\infty} (x_b - \mathcal{E}(x_b))^2 P(x_b) dx_b = \int_{\infty}^{\infty} \varepsilon_b^2 P(\varepsilon_b) d\varepsilon_b = \sigma_b^2 =: V_b \]  

and

\[ \mathcal{E}(y_o^2) := \int_{\infty}^{\infty} (y_o - \mathcal{E}(y_o))^2 P(y_o) dy_o = \int_{\infty}^{\infty} \varepsilon_o^2 P(\varepsilon_o) d\varepsilon_o = \sigma_o^2 =: V_o \]  

- the data \( y_o \) and \( x_b \) are independent, that is, there is no covariance between the observation and the a priori information:

\[ \mathcal{E}(\varepsilon_o \varepsilon_b) = 0 \]  

(4)
A 2 information bits example (2): estimation

We are seeking the estimator for $x_a$ as a weighted linear combination of two competing information elements $y_o$ and $x_b$, with as yet unknown weights $w_o$ and $w_b$ such that

$$x_a = w_o y_o + w_b x_b$$ \hfill (1)

minimal variance $V_a$.

With $x_a$ unbiased (as $y_o$ and $x_b$ are unbiased), we find

$$\mathcal{E}(x_a) = \mathcal{E}(x) = \mathcal{E}(w_o y_o + w_b x_b) = w_o x_t + w_b x_t.$$ \hfill (2)

It follows that

$$w_o + w_b = 1.$$ \hfill (3)

The optimality criterion selected is the minimal error variance of the estimator $X_a$. Our tunable parameters are the weights $w_o$ and $w_b$.

$$V_a = \sigma_a^2 = \mathcal{E}((x_a - \mathcal{E}(x_a))(x_a - \mathcal{E}(x_a)))$$
$$= \mathcal{E}((w_o y_o + w_b x_b - x)^2)$$
$$= \mathcal{E}((w_o (y_o - x_t) + w_b (x_b - x_t))^2)$$
$$= w_o^2 \mathcal{E}((y_o - x_t)^2) + w_b^2 \mathcal{E}((x_b - x_t)^2)$$
$$= w_o^2 \mathcal{E}(\epsilon_o^2) + w_b^2 \mathcal{E}(\epsilon_b^2)$$
$$= w_o^2 \sigma_o^2 + w_b^2 \sigma_b^2$$ \hfill (4)
A 2 information bits example (3): estimation

Define a cost function $J(w_o, w_b)$ to be minimised, subject to the constraint $(1 - w_o - w_b) = 0$ with the Lagrange multiplier $\lambda$

$$J(w_o, w_b) = \sigma_a^2 + \lambda(1 - w_o - w_b) = w_o^2 \sigma_o^2 + w_b^2 \sigma_b^2 + \lambda(1 - w_o - w_b), \quad (1)$$

given the necessary conditions for the observational (o) and background (b) case, independently

$$\frac{\partial J}{\partial w_o/b} = 2w_o/b \sigma_o^2/b - \lambda = 0, \quad (2)$$

from which follows

$$w_o/b = \frac{\lambda}{2 \sigma_o^2/b}. \quad (3)$$
A 2 information bits example (4): estimation

Then $\lambda$ can be expressed as

$$\lambda = \frac{2\sigma_o^2\sigma_b^2}{(\sigma_o^2 + \sigma_b^2)} = 2\phi(\sigma_o^{-2} + \sigma_b^{-2}). \quad (1)$$

We finally obtain the optimal weights.

$$w_o = \frac{\sigma_o^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}}, \quad w_b = \frac{\sigma_b^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}}. \quad (2)$$

Hence, the minimal variance estimator reads

$$x_a = \frac{\sigma_o^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}} y_o + \frac{\sigma_b^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}} x_b \quad (3)$$

and is called a **BLUE**, Best Linear Unbiased Estimator.
Single Observation, direct inversion

\[ w_o = \frac{\sigma_o^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}}, \quad w_b = \frac{\sigma_b^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}}. \] (1)

Hence, the minimal variance estimator reads

\[ x_a = \frac{\sigma_o^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}} y_o + \frac{\sigma_b^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}} x_b \] (2)

and is called a **BLUE**, Best Linear Unbiased Estimator. The minimal variance of the analysis itself directly follows

\[ \sigma_a^2 = \frac{1}{\sigma_o^{-2} + \sigma_b^{-2}} \] (3)

and hence

\[ \frac{1}{\sigma_a^2} = \frac{1}{\sigma_o^2} + \frac{1}{\sigma_b^2}, \quad \frac{x_a}{\sigma_a^2} = \frac{y_o}{\sigma_o^2} + \frac{x_b}{\sigma_b^2} \] (4)

it can be stated that

\[ V_a \leq \min(V_o, V_b) \] (5)
Single observation, Minimisation

Result obtained by variation of a cost function (distance function, objective function, test function) $J(x)$

$$J(x) = (y_o - x)^2 \sigma_o^{-2} + (x_b - x)^2 \sigma_b^{-2}$$  \hspace{1cm} (1)

$$0 = \frac{dJ(x)}{dx} \bigg|_{x=x_o} = 2(y_o - x)\sigma_o^{-2} + 2(x_b - x)\sigma_b^{-2}$$  \hspace{1cm} (2)

It follows

$$x_o = \frac{\sigma_o^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}} y_o + \frac{\sigma_b^{-2}}{\sigma_o^{-2} + \sigma_b^{-2}} x_b$$  \hspace{1cm} (3)
Optimal Interpolation

\[
J(x) = \frac{1}{2} [x^b - x]^T B_0^{-1} [x^b - x] + \frac{1}{2} \left\{ y^0 - H[x(t)] \right\}^T R^{-1} \left\{ y^0 - H[x] \right\}
\]

The gradient then reads

\[
\nabla J(x) = B_0^{-1} [x^b - x] + H^T R^{-1} \left\{ y^0 - H[x + (x_b - x_b)] \right\}
\]

where a trivial expansion is introduced for later manipulation.

At the minimum \( x =: x_a \)

\[
x_a - x_b = (B_0^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} \left\{ y^0 - H[x_b] \right\}
\]

\[
= B H^T (R + H^T B H)^{-1} \left\{ y^0 - H[x_b] \right\}
\]

with the latter result obtained after some manipulation.

Notation: Ide, K., P. Courtier, M. Ghil, and A. Lorenc,
Unified notation for data assimilation: operational sequential and variational,
Generalized Formulation (1)

\[ P(x|z) = \frac{P(z|x)P(x)}{\int P(z|x)P(x)dx} \]

with

\[ z := \Gamma x^t + \zeta \]

\[ P(x|z) \propto \exp\left[-\frac{1}{2} (\Gamma x - z)^T S^{-1} (\Gamma x - z) \right] \]

where \( S := \mathcal{E} (\zeta \zeta^T) \) is \( m \times m \) is the known “information” error covariance matrix
Generalized Formulation (2)

\[ \frac{1}{2} \frac{\partial}{\partial x} \langle z - \Gamma x, S^{-1}(z - \Gamma x) \rangle = -\Gamma^T S^{-1}(z - \Gamma x) = 0. \]

defining the analysis result \( x^a \)

\[ x^a = (\Gamma^T S^{-1} \Gamma)^{-1} \Gamma^T S^{-1} z, \]

The associated analysis error is

\[ x^a - x^t = (\Gamma^T S^{-1} \Gamma)^{-1} \Gamma^T S^{-1} (\zeta) \]

The associated analysis error covariance matrix is

\[ P^a := \mathcal{E} \left( (x^a - x)(x^a - x)^T \right) = (\Gamma^T S^{-1} \Gamma)^{-1} \]
Generalized Formulation (3)

\[ \frac{d\mathbf{x}}{dt} = \mathcal{M}(\mathbf{x}) + \eta, \quad (1) \]

with model error \( \eta \). In data assimilation parlance, model \( \mathcal{M} \) is often referred to as the ‘forward’ or ‘direct’ model. Upon differentiation with respect to \( \mathbf{x} \) we obtain

\[ \frac{d\delta\mathbf{x}}{dt} = \mathcal{M}'(\delta\mathbf{x}) = \mathbf{M}\delta\mathbf{x}, \quad (2) \]

where \( \mathcal{M}' = \mathbf{M} \) is the tangent–linear model to \( \mathcal{M} \). resolvent \( \mathbf{M}(t_j, t_i) \), which propagates a perturbation \( \delta\mathbf{x}(t) \) of the state variable \( \mathbf{x}(t) \) from time \( t_i \) to time \( t_j \), a stepwise tangent–linear model integration gives

\[ \delta\mathbf{x}(t_n) = \mathbf{M}(t_n, t_{n-1})\mathbf{M}(t_{n-1}, t_{n-2}) \ldots \mathbf{M}(t_1, t_0)\delta\mathbf{x}(t_0). \quad (3) \]

\[ \mathbf{y}_i^0 - \mathbf{H}_i\mathbf{x}(t_i) = \mathbf{H}_i\delta\mathbf{x}(t_i) = \mathbf{H}_i\mathbf{M}(t_i, t_0)\delta\mathbf{x}(t_0), \quad (4) \]

for each time step \([t_0, \ldots, t_N]\), \( \mathbf{G} := (\mathbf{H}_0, \mathbf{H}_1\mathbf{M}(t_1, t_0), \ldots, \mathbf{H}_N\mathbf{M}(t_N, t_0))^T \), with a concatenation of forward interpolator and tangent linear model resolvent

\[ \Gamma := \begin{pmatrix} \mathbf{I} \\ \mathbf{G} \end{pmatrix}, \quad \mathbf{z} := \begin{pmatrix} \mathbf{x}^b \\ \mathbf{y}_0^0 = \mathbf{G}\mathbf{x} + \zeta^b \end{pmatrix}. \quad (5) \]

\[ \mathbf{S} := \begin{pmatrix} \mathbf{P}^b & 0 \\ 0 & \mathbf{R} \end{pmatrix} \quad (6) \]
Generalized Formulation (4)

\[ P^a := (P^b^{-1} + G^T R^{-1} G)^{-1} = P^b - P^b G^T (G P^b G^T + R)^{-1} G P^b = (I - K G) P^b \] (1)

where

\[ K := P^b G^T (G P^b G^T + R)^{-1} \in n \times p \] (2)

is the Kalman gain matrix. With the Sherman–Morrison–Woodbury formula

\[
\begin{align*}
  x^a &= (P^b^{-1} + G^T R^{-1} G)^{-1} (I, G^T) \left( \begin{array}{cc} P^b & 0 \\ 0 & R \end{array} \right) \left( \begin{array}{c} x^b \\ y^0 \end{array} \right) \\
  &= (P^b^{-1} + G^T R^{-1} G)^{-1} P^b^{-1} x^b + (P^b^{-1} + G^T R^{-1} G)^{-1} G^T R^{-1} y^0 \\
  &= (I - K G) x^b + \\
  &\quad + (P^b^{-1} + G^T R^{-1} G)^{-1} (G^T R^{-1} G P^b G^T + G^T)(G P^b G^T + R)^{-1} y^0 \\
  &= (I - K G) x^b + K y^0.
\end{align*}
\]

Hence we find

\[ x^a = x^b + K d \] (3)

with \( d := y^0 - G x^b \) being the innovation vector.
Generalized Formulation (5)

\[ J(\xi) := \frac{1}{2}(\Gamma \xi - \mathbf{z})^T \mathbf{S}^{-1}(\Gamma \xi - \mathbf{z}). \]

After transfer into the form of two uncorrelated information sources "a priori" and "observations", the cost function then is

\[ J(\xi(t_0)) = \frac{1}{2}(\mathbf{x}^b(t_0) - \xi(t_0))^T \mathbf{P}^b^{-1}(\mathbf{x}^b(t_0) - \xi(t_0)) + \]

\[ \frac{1}{2} \sum_{i=0}^{N} (\mathbf{y}^0(t_i) - \mathbf{H}\xi(t_i))^T \mathbf{R}^{-1}(\mathbf{y}^0(t_i) - \mathbf{H}\xi(t_i)) \]  

(1)

\[ \nabla_{\xi(t_0)} J = -\mathbf{P}^b^{-1}(\mathbf{x}^b(t_0) - \xi(t_0)) - \sum_{m=0}^{N} \mathbf{M}^T(t_m, t_0)\mathbf{H}_m^T \mathbf{R}^{-1}(\mathbf{y}^0(t_m) - \mathbf{H}_m\xi(t_m)) \]  

(2)
Kalman filter: basic equations

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

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

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

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

and

\[ P^a_i = (I - K_i H_i)P^b_i. \quad (3) \]
Types of assimilation algorithms: “smoother” and filter

4D-var

Kalman Filter

Previous forecast

Corrected forecast

Analysis

Obs

X

X_a

X_b

3z  6z  9z  12z  15z

Time

Assimilation window
Why 4D-variational data assimilation?

• provides BLUE (Best Linear Unbiased Estimator).  
  *Remark: 3D-var → ingested in model does not!*

• Potential for “consistency” within the assimilation interval (O (1 day))
  
  *Remark: fast manifold perturbations (=“initialisation problem”) mitigated, but not removed. Inconsistencies at the end of the assimilation windows.*

• Allows for extensions to estimate analysis error covariance matrix and temporal correlations (red noise)
How does 4D-var work?
Derivation of 4D-var (1)

The distance function $\mathcal{J}$ may be defined as follows:

$$\mathcal{J}(\mathbf{x}(t)) = \frac{1}{2}(\mathbf{x}_b - \mathbf{x}(t_0))^T \mathbf{B}^{-1}(\mathbf{x}_b - \mathbf{x}(t_0)) + \frac{1}{2} \int_{t_0}^{t_N} (\hat{\mathbf{x}}(t) - \mathbf{x}(t))^T \mathbf{R}^{-1}(\hat{\mathbf{x}}(t) - \mathbf{x}(t))dt$$

(1)

where $\mathcal{J}$ is a scalar functional defined on the time interval $t_0 \leq t \leq t_N$ dependent on the vector valued state variable $\mathbf{x} \in \mathcal{H}$ with $\mathcal{H}$ denoting a Hilbert space. The first guess or background values $\mathbf{x}_b$ are defined at $t = t_0$, and $\mathbf{B}$ is the covariance matrix of the estimated background error. The observations are denoted $\hat{\mathbf{x}}$ and the observation and representativeness errors are included in the covariance matrix $\mathbf{R}$. 
Derivation of 4D-var (2)

Let the differential equation of the model $M$ be given by

$$\frac{dx}{dt} = M(x),$$

(1)

where $M$ acts as a generally nonlinear operator defining uniquely the state variable $x(t)$ at time $t$, after an initial state $x(t_0)$ is provided. The linear perturbation equation, giving the evolution of a small deviation $\delta x(t)$ from a model state $x(t)$ then reads

$$\frac{d\delta x}{dt} = M'\delta x,$$

(2)

where $M'$ is the tangent linear model of $M$. 
Derivation of 4D-var (3)

Denoting the inner product of $\mathcal{H}$ by bracketed parentheses $\langle \quad , \quad \rangle$, the operator $M'$, mapping from $\mathcal{H}$ into $\mathcal{H}$ itself, has the adjoint operator $M'^*$, which is defined by $\langle y, M'z \rangle = \langle M'^*y, z \rangle$ for all $y, z \in \mathcal{H}$. For the remainder we drop the background term, however introduce the model equation as a strong constraint with Lagrange multipliers $\lambda(t)$. Setting $1/2 (\hat{x}(t) - x(t))^T R^{-1} (\hat{x}(t) - x(t)) = \mathcal{O}(t)$ for notational convenience, we find

$$J(x(t)) = \int_{t_0}^{t_N} \mathcal{O}(t) + \langle \lambda, \frac{dx(t)}{dt} - Mx(t) \rangle dt$$  \hspace{1cm} (1)
Derivation of 4D-var (4)

\[
\delta J = \int_{t_0}^{t_N} \left( \langle \nabla_x O(t), \delta x(t) \rangle + \langle \delta \lambda, \frac{dx(t)}{dt} - Mx(t) \rangle + \langle \lambda, \frac{d\delta x(t)}{dt} - M' \delta x(t) \rangle \right) dt \\
= \int_{t_0}^{t_N} \left( \langle \nabla_x O(t) - \frac{d\lambda(t)}{dt} - M'^* \lambda(t), \delta x(t) \rangle + \langle \delta \lambda, \frac{dx(t)}{dt} - Mx(t) \rangle \right) dt + \lambda(t_N) \delta x(t_N) - \lambda(t_0) \delta x(t_0),
\]

where integration by parts was applied.
Derivation of 4D-var (5)

Introducing the extremal principle $\delta J = 0$, the integrand includes the inhomogeneous adjoint equation, which is forced by the observation term

$$-\frac{d\lambda(t)}{dt} - M'^* \lambda(t) = R^{-1}(\hat{x}(t) - x(t)).$$

(1)

The integration of the tangent linear equation gives the evolution of an initial perturbation $\delta x(t_0)$ at later times $t_n$ and can formally be expressed as

$$\delta x(t_n) = G(t_n, t_0)\delta x(t_0),$$

(2)

where the operator $G(t_n, t_0)$ denotes the resolvent of $M'$ for the time interval $[t_0, t_n]$ acting on the initial state $x(t_0)$. 
Derivation of 4D-var (6)

\[
\frac{d}{dt} \langle \lambda(t), \delta x(t) \rangle = \langle \lambda(t), M' \delta x(t) \rangle - \langle M'^* \lambda(t), \delta x(t) \rangle = 0. \tag{1}
\]

It is now possible to express \( \delta J \) as a function of \( x(t_0) \) alone

\[
\delta J(x(t_0)) = \int_{t_0}^{t_N} \langle \nabla_x \mathcal{O}(t), G(t, t_0) \delta x(t_0) \rangle dt = \left( \int_{t_0}^{t_N} S(t_0, t) \nabla_x \mathcal{O}(t) dt, \delta x(t_0) \right), \tag{2}
\]

which finally leads to the desired gradient of \( J \), given in terms of a discretized expression of the adjoint operator \( S(t_0, t) \)

\[
\nabla_{x(t_0)} J = \sum_{m=0}^{N} \tilde{S}(t_0, t_1) \tilde{S}(t_1, t_2) ... \tilde{S}(t_{m-1}, t_m) \nabla_x \mathcal{O}(t_m). \tag{3}
\]
Derivation of 4D-var (7)

It remains to be shown that $\nabla_{\mathbf{x}(t_0)} \mathcal{J} = \lambda(t_0)$. Defining the backward initial condition $\lambda(t_N) = 0$ and given a single instantaneous forcing $\hat{\lambda}(t') = -\nabla_{\mathbf{x}(t')} \mathcal{O}(t')$ at any time $t'$, $t \leq t' < t_N$, we can write, again in exact form for convenience, as

$$\frac{\partial}{\partial t} \mathbf{S}(t, t') \nabla_{\mathbf{x}(t')} \mathcal{O}(t') = -\frac{d\hat{\lambda}(t)}{dt} = -\mathbf{M}^{\prime \ast} \mathbf{S}(t, t') \nabla_{\mathbf{x}(t')} \mathcal{O}(t'), \quad (1)$$

to reveal that $\frac{\partial}{\partial t} \mathbf{S}(t, t') = -\mathbf{M}^{\prime \ast} \mathbf{S}(t, t')$.

Integration then gives

$$\hat{\lambda}(t) = -\mathbf{S}(t, t') \nabla_{\mathbf{x}(t')} \mathcal{O}(t'). \quad (2)$$

Since the resolvent $\mathbf{S}$ is a linear operator, all observations can be combined by adding up the corresponding equations, which demonstrates the preposition for $t = t_0$. 


Collection of assimilation formula

variational data assimilation (3D-var und 4D-var)

\[ J = (x - x_b)^T B^{-1} (x - x_b) + (y - Hx)^T R^{-1} (y - Hx) \]

\[
\begin{align*}
x_a &= (B^{-1} + H^T R^{-1} H)^{-1} (B^{-1} x_b + H^T R^{-1} y) \\
\varepsilon_a &= -x_* + (B^{-1} + H^T R^{-1} H)^{-1} (B^{-1} x_* + H^T R^{-1} y_*) \\
&
\quad + (B^{-1} + H^T R^{-1} H)^{-1} (B^{-1} \varepsilon_b + H^T R^{-1} \varepsilon_o) \\
\langle \varepsilon_a \rangle &= (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} (y_* - Hx_*) \\
x_a &= x_b + (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} (y - H(x_b))
\end{align*}
\]

Optimal Interpolation (OI)

\[ x^a(t_i) = x^b(t_i) + K_i(y^o - Hx), \]

\[ K_i := P_i^b H_i^T (H_i P_i^b H_i^T + R_i)^{-1} \in \mathbb{R}^{n \times p_i} \]
The “Initialisation” / Filtering Problem (1)

- Constraints by physical/chemical equations are insufficient, as they allow for physically/chemically admissible states, but partly very unlikely to occur:
  - high amplitude gravity waves (“fast manifolds”)
  - pronounced chemical imbalances

- Legacy procedure: introduce penalty term for fast changes (Normal Mode Initialisation, NMI)

\[ J_c(x) = \mu \left( \frac{dx}{dt} - \frac{dx_b}{dt} \right)^2 \]
The “Initialisation” / Filtering Problem (2)

- or filter function (Digital Filter Initialisation, DFI, here with Lanczos-filter), $x$ model state:

$$\bar{x}(0) = \sum_{n=-M}^{M} h_n x(t_n)$$

where

$$h_n := \frac{\sin(n\theta \Delta t) \sin(n\pi / M)}{(n\pi)^2 / M}$$
Outlook

• So far, only systems with Gaussian errors are treated.
• Sophisticated related algorithms (4D-var, KF) are far from being fully exploited.
• However, limits of Gaussian assumptions are obvious: assimilation for precipitation forecasts and aerosol process poor.
• Further, more general approaches necessary: MCMC, importance resampling, …