Principles of data assimilation

From linear regression to 4DVAR

Examples of inverse modelling and state estimation
Data assimilation is the combination of measurements with any kind of model.

All we do is least squares fitting.
(Carl Wunsch)

All data assimilation methods are special cases of nudging.
(Andrew Bennett)
nudging

\[ \frac{\partial \psi}{\partial t} = \frac{\partial \psi^{\text{physics}}}{\partial t} - \gamma (\psi - d^{\text{measured}}) \]

\[ \gamma^{-1} \]

is a relaxation time
Inverse modelling

or:

given the answer, what was the question?
Inverse modelling

or:

given the answer, what was the question?

Where was the picture of the iceberg taken from?
Data assimilation: general formulation

\[ f(\psi | d) = \frac{f(d | \psi) f(\psi)}{f(d)} \]

NO INVERSION !!!
Wunsch’s statements:
Data must be valuable and contain signal not only noise.
Model must have some skill.
Model must perform differently than data significantly (or you are done).

Any method will help our understanding, they are optimal for your own purpose, i.e. doable in reasonable time.
Probability density functions (pdf = spread) should overlap, real pdf are always much broader than our simple estimates, we must be able to drive the model towards data (controllability) [observability comes later].

The assimilation problem is different from the forecast problem.
Use of data assimilation / inverse modelling

Perform sensitivity analysis, array design with adjoint system.

Analyse for systematic differences.

Analyse the trajectory and determine properties of the ocean/the system.

Use result to (iteratively) improve the model numerics/physics.
Use of data assimilation jargon

*inverse modelling* is stationary (e.g. 3DVAR),

*data assimilation* is time dependent,

*sequential* mean solve a sequence of subproblems in time,

*iterative*: let the computer do all the work at once,

*4DVAR* means use of adjoint model,

*state estimation*: determine all model variables and their temporal derivatives (usually over some period of time)
Use of data assimilation jargon

\[ j \] is

cost function, penalty function, merit function, objective function,
least squares sum, generalised distance,
negative exponent of pdf,
beauty principle,
which must be minimized
The best estimate is a weighted mean between model and data

\[ \psi^{opt} = a \psi^{model} + b \psi^{measured} \]

a and b can be operators
Estimation of a mean value

Probably the most simple estimation all of us have performed is the determination of the mean value $x$ of a set of observations $y_i$.

What we do is to solve for the value which is closest to all measurements $y_i$ in the least squares sense. In other words we determine the minimum of a cost function $j$ defined by

$$j = 0.5 \sum_{i=1,N} (x - y_i)^2$$
Estimation of a mean value

The minimum of $j$ can be found by setting the derivative of $j$ to zero: with the obvious solution

\[
\frac{dj}{dx} = \sum_{i=1,N} (x - y_i) = 0
\]

\[
Nx = \sum_{i=1,N} y_i \quad x = \frac{1}{N} \sum_{i=1,N} y_i
\]

\[
\frac{dx}{dy_i} = \frac{1}{N} \quad \text{var}(x) = \frac{1}{N} \text{var}(y)
\]
Update of a mean value

now suppose we have another estimate, $x_2$
this time averaged over $M$ values which we will assimilate

\[
x_2 = \frac{1}{M} \sum_{i=N+1}^{N+M} y_i
\]

the cost function $j$ for this problem is the sum

\[
j = 0.5 \sum_{i=1}^{N} (x - y_i)^2 + 0.5 \sum_{i=N+1}^{N+M} (x - y_i)^2
\]
update of a mean value

The minimum of $j$ is again found by setting the derivative of $j$ to zero:

$$\frac{dj}{dx} = \sum_{i=1}^{N} (x - y_i) + \sum_{i=N+1}^{N+M} (x - y_i) = 0$$

$$Nx + Mx = \sum_{i=1}^{N} y_i + \sum_{i=N+1}^{N+M} y_i$$

$$(N + M)x = Nx_1 + Mx_2$$

$$x = \frac{Nx_1 + Mx_2}{N + M}$$

$$= \frac{N}{N + M} x^{\text{model}} + \frac{M}{N + M} x^{\text{data}}$$
update of a mean value

what is the variance of the updated (posterior) mean?

\[
x = \frac{N}{N + M} \left(1 \sum_{i=1}^{N} y_i \right) + \frac{M}{N + M} \left(1 \sum_{i=N+1}^{N+M} y_i \right)
\]

\[
\frac{dx}{dy_i} = \frac{1}{N + M}
\]

\[
\text{var}(x) = \frac{1}{N + M} \text{var}(y) \leq \text{var}(\text{model})
\]

\[
\text{var}(x) \leq \text{var}(\text{data})
\]

Data assimiliation fundamentals 1
GODAE Summer School
fitting a straight line through data and determine $y = a + bx$ and residuals
fitting a straight line through data of the form \( y = a + bx \) with the cost function \( j \)

\[
j = 0.5 \sum_{i=1}^{N} \frac{(y_i - y)^2}{\sigma_i^2} = 0.5 \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_i^2}
\]

each data point \( y_i \) is weighted with the inverse of its estimated variance \( \sigma_i^2 \)
fitting a straight line through data
determination of a and b by minimizing j

\[
\frac{dj}{da} = \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2} = 0
\]

\[
\frac{dj}{db} = \sum_{i=1}^{N} \frac{x_i (y_i - a - bx_i)}{\sigma_i^2} = 0
\]
fitting a straight line through data

with the solution

\[ a = \frac{\sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i} \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i} - \left( \sum_{i=1}^{N} \frac{y_i}{\sigma_i} \right)^2} \]
fitting a straight line through data

with the solution

\[ b = \frac{\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i} \sum_{i=1}^{N} \frac{y_i}{\sigma_i} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - (\sum_{i=1}^{N} \frac{y_i}{\sigma_i})^2} \]
fitting a straight line through data

now we remember that the regression line is \( y - \bar{y} = b(x - \bar{x}) \)

\[
a = \bar{y} - b\bar{x}
\]

\[
b = \frac{\sum_{i=1}^{N} \left( x_i - \bar{x} \right) \left( y_i - \bar{y} \right)}{(\sum_{i=1}^{N} \frac{(x_i - \bar{x})}{\sigma_i})^2} = \frac{s_{xy}}{s_{x}^2}
\]
what if the data is not linear?

the residual tells
Fit to data

covariance function of residual
fitting a straight line through data

the variance of $a$ and $b$ can easily be calculated as before

\[
\frac{da}{dy_i} = \ldots.
\]

\[
\frac{db}{dy_i} = \ldots.
\]
fitting a straight line through data

a more elegant way of calculating the error covariance of a and b is via the Hessian matrix

cov\langle a, b \rangle = H^{-1}

\[
H = \begin{pmatrix}
\frac{\partial^2 j}{\partial a \partial a} & \frac{\partial^2 j}{\partial a \partial b} \\
\frac{\partial^2 j}{\partial b \partial a} & \frac{\partial^2 j}{\partial b \partial b}
\end{pmatrix}
= \begin{pmatrix}
\sum_{i=1}^{N} \frac{1}{\sigma_i^2} & \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \\
\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} & \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}
\end{pmatrix}
\]
fitting a straight line through data

student exercise (for volunteers)

1. generate data $y_i$ at fixed $x_i$,
2. use regression software to determine $a$ and $b$,
3. generate $k$ realizations for the data $y_i$, with a prescribed variance and determine $a$ and $b$ for each $k$,
4. calculate the error covariance $(a,b)$ from your results and compare with the inverse Hessian.
modelling of biology
The evolution of nutrients, plankton, detritus, zooplankton is described with a set of differential eq. They contain a number of “adjustable” parameters \( p \) (e.g. growth rate, efficiency of eating, sinking velocity for detritus…)
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Parameter</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_w$</td>
<td>attenuation coefficient of downwelling irradiance</td>
<td>0.04</td>
<td>m$^{-1}$</td>
</tr>
<tr>
<td>$k_c$</td>
<td>light attenuation due to phytoplankton</td>
<td>0.03</td>
<td>m$^2$ mmol$^{-1}$</td>
</tr>
<tr>
<td>$\beta_1, \beta_2, \beta_3$</td>
<td>zooplankton assimilation efficiency</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>$r_1$</td>
<td>zooplankton feeding preference</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>$r_2, r_7$</td>
<td>zooplankton feeding preference</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>$\delta$</td>
<td>fraction of zooplankton losses going to DON</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>fraction of zooplankton losses going to ammonium</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>$\eta$</td>
<td>ammonium:DON uptake ratio</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>$R_p$</td>
<td>carbon to nitrogen ratio for phytoplankton</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>$R_z$</td>
<td>carbon to nitrogen ratio for zooplankton</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>$R_B$</td>
<td>carbon to nitrogen ratio for bacteria</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Data assimilation fundamentals 1
GODAE Summer School
Table 1
Adjusted model parameters

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Parameter</th>
<th>First guess</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$</td>
<td>phytoplankton maximum specific mortality rate</td>
<td>0.05</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>$k_1, k_2$</td>
<td>half-saturation constants for nutrient and ammonium uptake</td>
<td>0.5</td>
<td>mmol N m$^{-3}$</td>
</tr>
<tr>
<td>$k_5$</td>
<td>phytoplankton mortality half-saturation constant</td>
<td>0.2</td>
<td>mmol N m$^{-3}$</td>
</tr>
<tr>
<td>$\psi$</td>
<td>nitrate uptake ammonium inhibition parameter</td>
<td>1.5</td>
<td>m$^3$ (mmol N)$^{-1}$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>initial slope of the $P-I$ curve</td>
<td>0.025</td>
<td>m$^2$ W$^{-1}$ day$^{-1}$</td>
</tr>
<tr>
<td>$g$</td>
<td>zooplankton maximum ingestion</td>
<td>1</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>zooplankton maximum loss rate</td>
<td>0.3</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>$k_3$</td>
<td>zooplankton ingestion half-saturation constant</td>
<td>1</td>
<td>mmol N m$^{-3}$</td>
</tr>
<tr>
<td>$k_6$</td>
<td>zooplankton loss rate half-saturation constant</td>
<td>0.2</td>
<td>mmol N m$^{-3}$</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>the bacterial excretion rate</td>
<td>0.05</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>$V_b$</td>
<td>bacterial maximum uptake rate</td>
<td>2</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>$k_4$</td>
<td>bacterial half-saturation constant for uptake</td>
<td>0.5</td>
<td>mmol N m$^{-3}$</td>
</tr>
<tr>
<td>$\mu_4$</td>
<td>detrital breakdown rate</td>
<td>0.05</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>$w_g$</td>
<td>detrital sinking rate</td>
<td>5.000</td>
<td>m day$^{-1}$</td>
</tr>
</tbody>
</table>
Analyse the possibility to determine $p$ via the Hessian matrix $H$

The posterior uncertainty of $p$ is described by (co-)variance

$$\text{cov}\langle p_l, p_m \rangle = H^{-1}$$

$$H_{l,m} = \frac{\partial^2 j}{\partial p_l \partial p_m}$$
Analyse the possibility to determine $p$ via the Hessian matrix $H$

A singular value analysis of $H$ reveals

$$H = USV^T$$

$$H^{-1} = US^{-1}U^T$$

where matrix $U$ contains the singular vectors and the diagonal matrix $S$ the singular values.
Singular vectors of $\mathbf{H}$

Fig. 3. Parameter resolution for experiment E1. Monthly measurements of nitrate and phytoplankton concentrations were employed.
Singular vectors of $\mathbf{H}$ (2)

Fig. 5. Parameter resolution for experiment E3. Monthly measurements of nitrate and phytoplankton concentrations were employed. Unit: $\mu g$.  

### Singular vectors of $\mathbf{H}$ (2)

- $\lambda_1 = 0.0007$
- $\lambda_2 = 0.017$
- $\lambda_3 = 0.067$
- $\lambda_4 = 0.12$
- $\lambda_5 = 0.34$
- $\lambda_6 = 0.98$
- $\lambda_7 = 14$
- $\lambda_8 = 46$
- $\lambda_9 = 1.1 \times 10^2$

---

39
FEMSECT:
inverse model to determine velocities and transports through a hydrographic section
equations of section inverse model

1. Geostrophy and hydrostatics:

\[ \mathbf{u} - \frac{g}{\rho_0 f} \int_{-H}^{z} (\mathbf{k} \times \nabla \rho) \, dz - \mathbf{u}_b = 0, \]  

(1)

where \( \mathbf{k} \) is the vertical unit vector, \( g \) is the gravitational acceleration, \( f \) is the Coriolis parameter and \( \mathbf{u}_b \) is the velocity at the bottom \( z = -H(x, y) \).

2. Equation of state

\[ \rho - \mathcal{R}(\rho, T, S) = 0, \]  

(2)

\[ \frac{\partial^2 \omega}{\partial z^2} - \frac{g}{\rho_0 f^2} (\nabla \rho \times \mathbf{k}) \cdot \nabla f = 0, \]  

(3)

with boundary conditions

\[ w(0) - \frac{1}{\rho} \text{curl} \tau + F_{\text{top}} = 0, \]

\[ w(-H) - [(\mathbf{u}_b \cdot \nabla H) + F_{\text{bot}}] = 0, \]
Tracer advective diffusive equations (N=8)

\[(u \cdot \nabla C_n) + w \frac{\partial C_n}{\partial x} - F_n = 0, \]

\[1 \leq n \leq N : \theta, S, C_3, ..., C_N,\]

unknowns in the inverse model are bottom reference velocities \(u_b\), tracers \(C_i\) on model grid and residuals \(F_i\)
The cost function $j$

$$j_0 = \frac{1}{2} \left[ \sum_{m,r} (\hat{C}_{m}(z) - \hat{C}_{m}^*(z'))^\dagger \sum_{z} \sum_{z'} W_{m,z}(z,z')(\hat{C}_{r}(z') - \hat{C}_{r}^*(z')) \right]$$

$$+ \sum^{*}(u - u^*)^\dagger W_u (u - u^*)$$

$$+ W_w \int_{x=0} (F_w)^2 dx + W_w \int_{x=-H} (F_w)^2 dx$$

$$+ \int_{\Omega} \sum_{n} F_n^\dagger W_F F_n d\Omega$$

$$+ \int_{\Omega} (\hat{S}_1 C_{r})^\dagger D_n (\hat{S}_1 C_{r}) d\Omega$$

$$+ \int_{\Omega} (\hat{S}_1 F_{r})^\dagger D_F (\hat{S}_1 F_{r}) d\Omega$$

$$+ \int_{\Omega} (\hat{S}_2 u)^\dagger D_u (\hat{S}_2 u) d\Omega.$$

Tracer data *

1,,8

Velocities

Windstress, bottom layer

Imbalances

Smoothness tracer

Smoothness imbalances

Smoothness velocities
how do we calculate the inverse Hessian now?

we want to know the variance of a quantity like the transport of mass through the section, i.e., $L^T x$

the variance of the transport $\text{var}(L^T x)$ is

$$L^T \text{cov} \langle x, x^T \rangle L = L^T H^{-1} L$$

solve $H z = L$

so that $H^{-1} L = z$

then $\text{var}(L^T x) = L^T z$
covariance with say heat transport
$L_2x$ is the product

$$L_2^T \text{cov}(x, x^T) L_1 = L_2^T H^{-1} L_1$$

$$= L_2^T z$$

Codes for calculating Hessian times vector are in packages like TAMC
Data assimilation: general formulation

\[ f(\psi \mid d) = \frac{f(d \mid \psi) f(\psi)}{f(d)} \]

NO INVERSION !!!
Propagation of pdf: Ensemble methods ‘efficient’ propagation for nonlinear models
Sequential Importance Resampling

prior pdf

observation pdf

posterior pdf

resampled pdf
Propagation of pdf: Ensemble methods ‘efficient’ propagation for nonlinear models
SIR-results for ocean around South Africa
SIR-results: errors

Day 0

Day 10

Day 20

Day 30
The Ocean is nonlinear...

Probability density function at 40S 20E
Retrieved bio-parameters(t)
<table>
<thead>
<tr>
<th>Symbol</th>
<th>First guess</th>
<th>Optimized value</th>
<th>Error variance</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g )</td>
<td>1.0</td>
<td>0.40</td>
<td>0.15</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>( \mu_3 )</td>
<td>5 \times 10^{-2}</td>
<td>9 \times 10^{-3}</td>
<td>4.5 \times 10^{-3}</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>( \mu_4 )</td>
<td>5 \times 10^{-2}</td>
<td>1.1 \times 10^{-2}</td>
<td>1.4 \times 10^{-3}</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>( \mu_2 )</td>
<td>0.3</td>
<td>0.24</td>
<td>0.13</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>( k_6 )</td>
<td>0.2</td>
<td>0.20</td>
<td>0.08</td>
<td>mmol N m(^{-3})</td>
</tr>
<tr>
<td>( k_3 )</td>
<td>1.0</td>
<td>2.62</td>
<td>1.4</td>
<td>mmol N m(^{-3})</td>
</tr>
<tr>
<td>( \psi )</td>
<td>1.5</td>
<td>0.99</td>
<td>0.33</td>
<td>m(^3) (mmol N)(^{-1})</td>
</tr>
<tr>
<td>( w_g )</td>
<td>5.0</td>
<td>4.25</td>
<td>2.17</td>
<td>m day(^{-1})</td>
</tr>
</tbody>
</table>
answer:

the photo of the iceberg was taken from here