Principles of data assimilation

From linear regression to 4DVAR

Examples of
inverse modelling and
state estimation

AWI
Alfred Wegener Institute
for Polar and Marine Research

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

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

γ^{-1} is a relaxation time

Inverse modelling

or:

given the answer, what was the question?

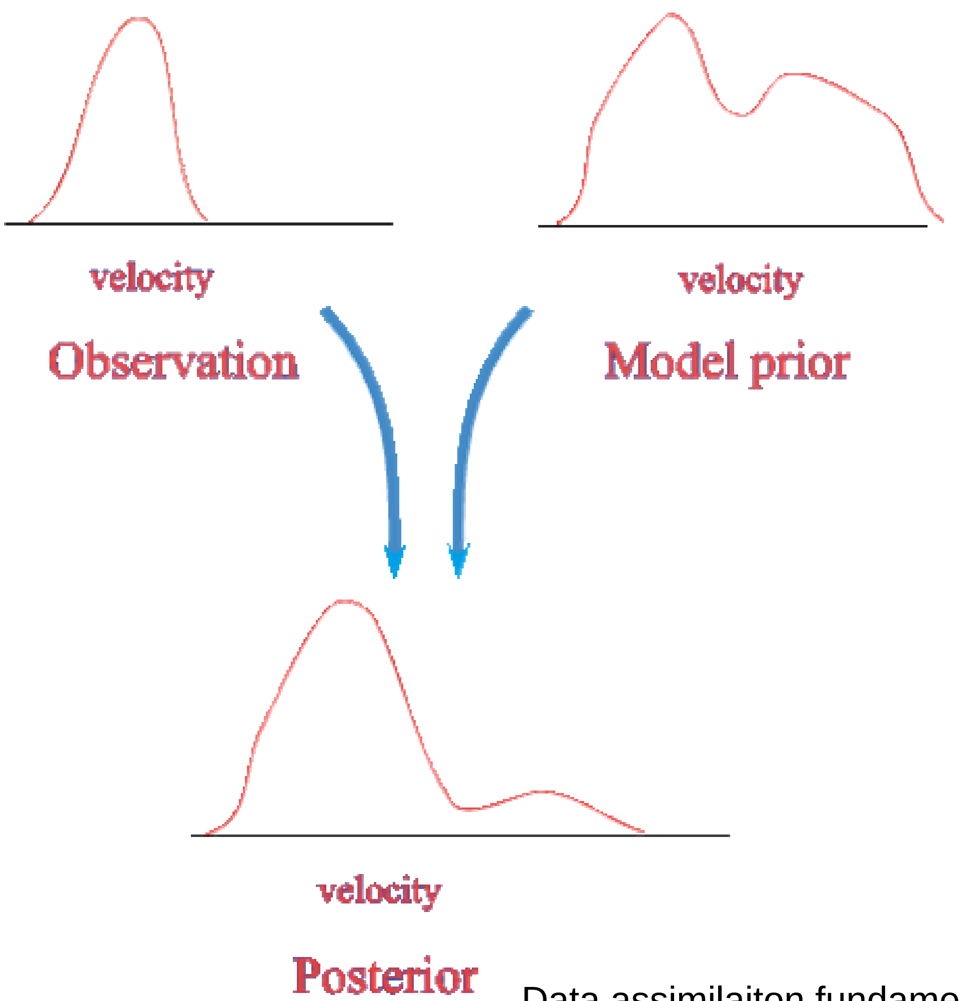
Inverse modelling

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 statemens:

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

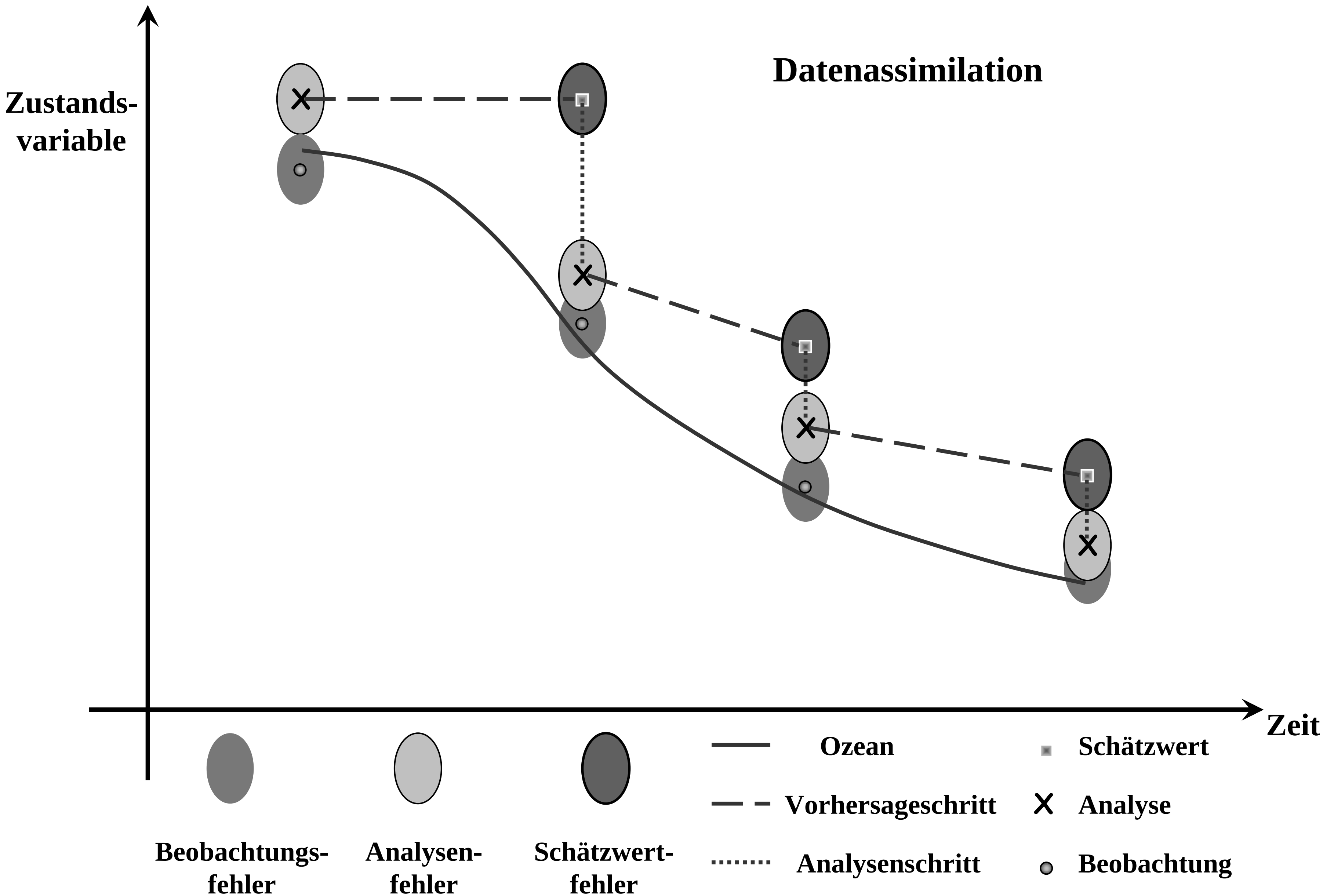
cost function, penalty function,
merit function, objective function,

least squares sum, generalised distance,

negative exponent of pdf,

beauty principle,

which must be minimized


Datenassimilation
Zustands-
variable
Zeit
Ozean
Vorhersageschritt
Analysenschritt
Schätzwert
Analyse
Beobachtung
Beobachtungs-
fehler
Analysen-
fehler
Schätzwert-
fehler

The best estimate is a weighted mean between model and data

$$\psi^{opt} = a\,\psi^{model} + b\,d^{measured}$$

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

.

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

$$Nx = \sum_{i=1,N} y_i \qquad x = \frac{1}{N} \sum_{i=1,N} y_i$$

$$\frac{dx}{dy_i} = \frac{1}{N} \qquad \mathrm{var}(x) = \frac{1}{N} \mathrm{var}(y)$$

Update of a mean value

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

.

$$\mathrm{x}_2 = \frac{1}{M}\sum_{i=N+1}^{N+M} y_i$$

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:.

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

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

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

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

$$= \frac{N}{N+M} x^{model} + \frac{M}{N+M} x^{data}$$

update of a mean value

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

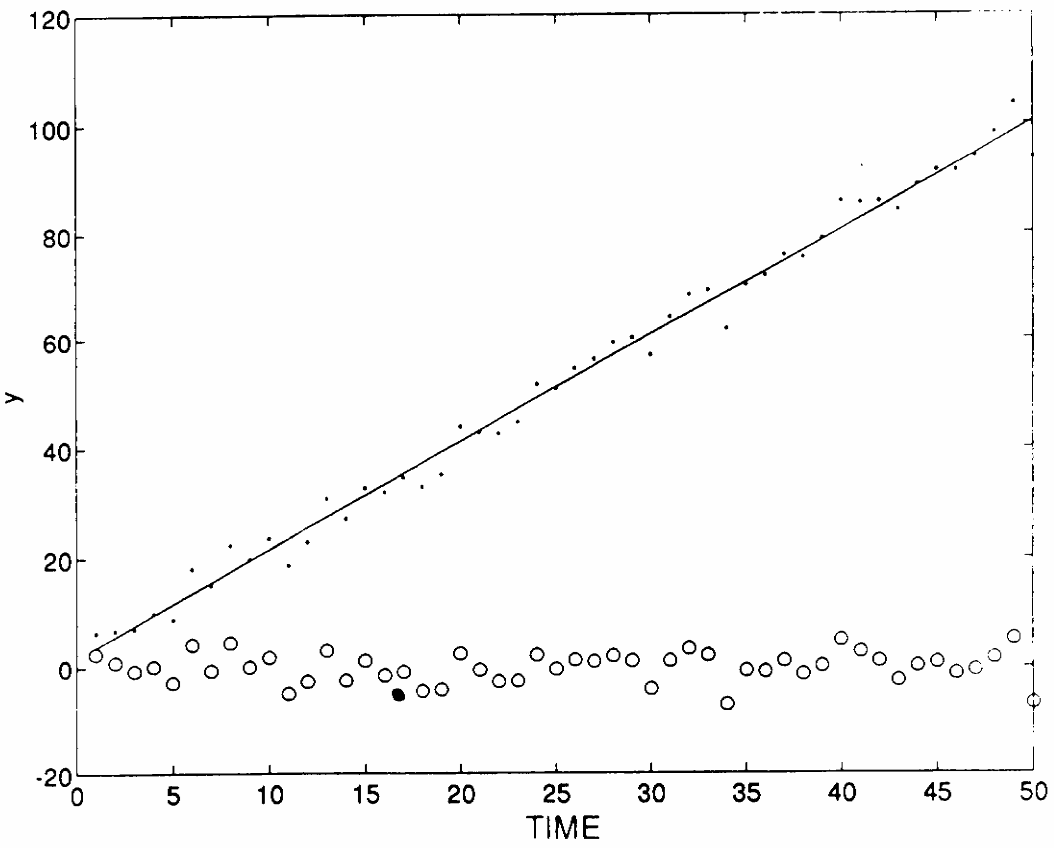
$$x = \frac{N}{N+M}\frac{1}{N}\sum_{i=1}^{N} y_i + \frac{M}{N+M}\frac{1}{M}\sum_{i=N+1}^{N+M} y_i$$

$$\frac{dx}{dy_i} = \frac{1}{N+M}$$

$$\mathrm{var}(x) = \frac{1}{N+M}\mathrm{var}(y) \le \mathrm{var}(model)$$

$$\mathrm{var}(x) \le \mathrm{var}(data)$$

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 yi is weighted with the inverse of its estimated variance σ_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

$$a = \frac{\sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i \, y_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - (\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2})^2}$$

fitting a straight line through data

with the solution

$$b = \frac{\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i\, y_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - (\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2})^2}$$

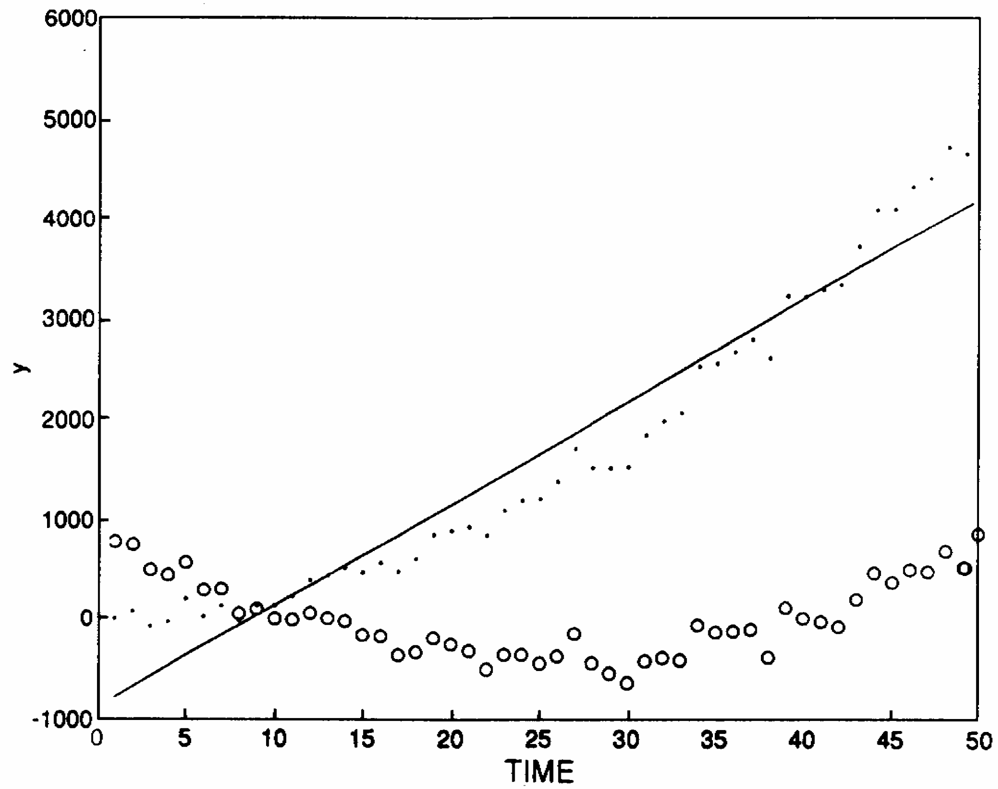
fitting a straight line through data

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

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_i^{\,2}}}{\left(\sum_{i=1}^{N} \frac{(x_i - \bar{x})}{\sigma_i}\right)^2} = \frac{Sxy}{Sx^2}$$

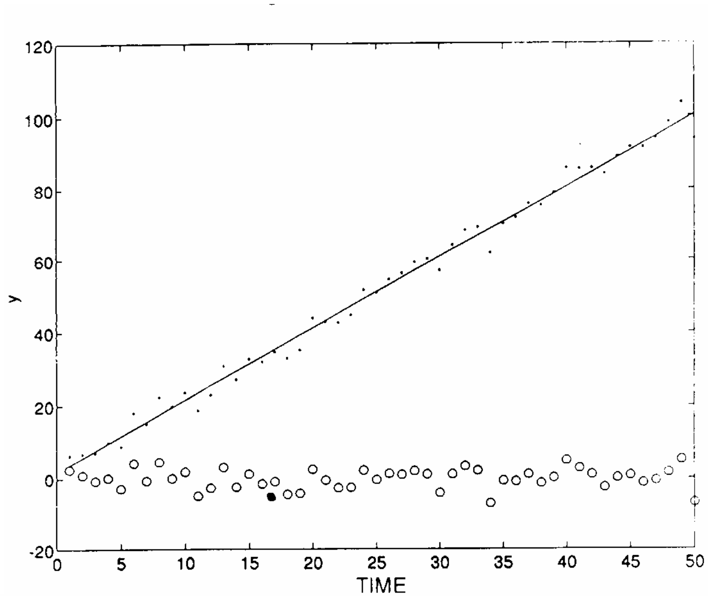
what if the data is not linear?

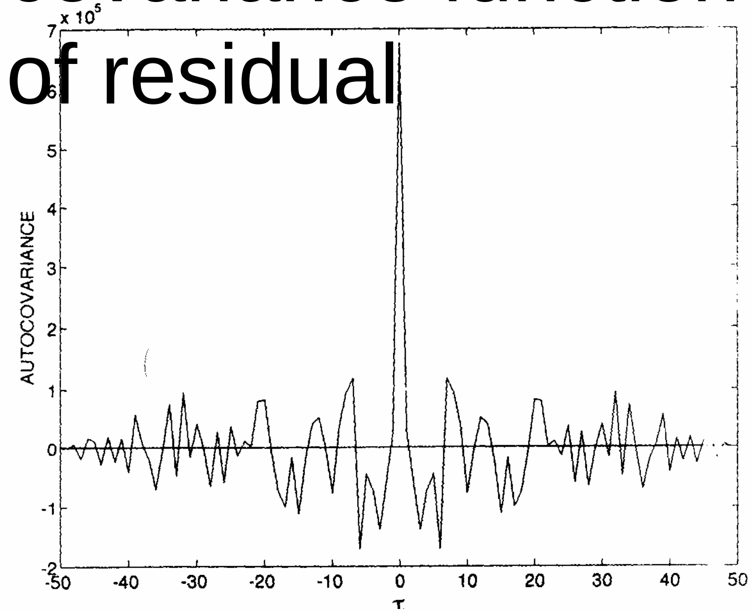


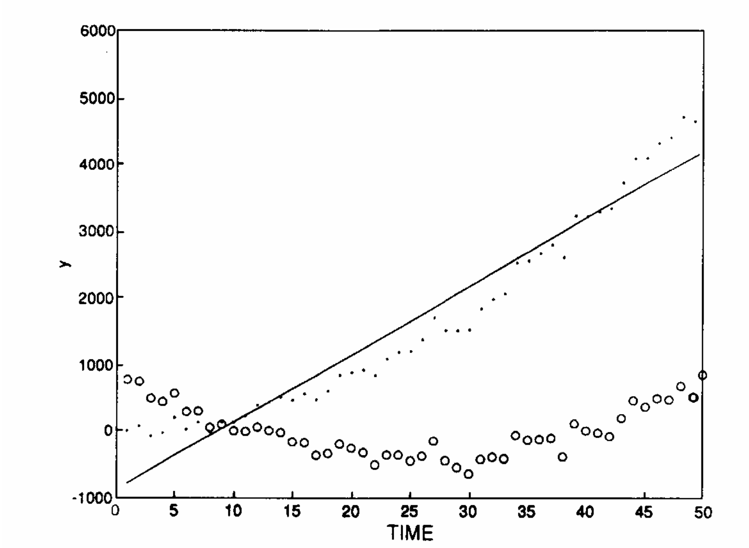
the residual tells

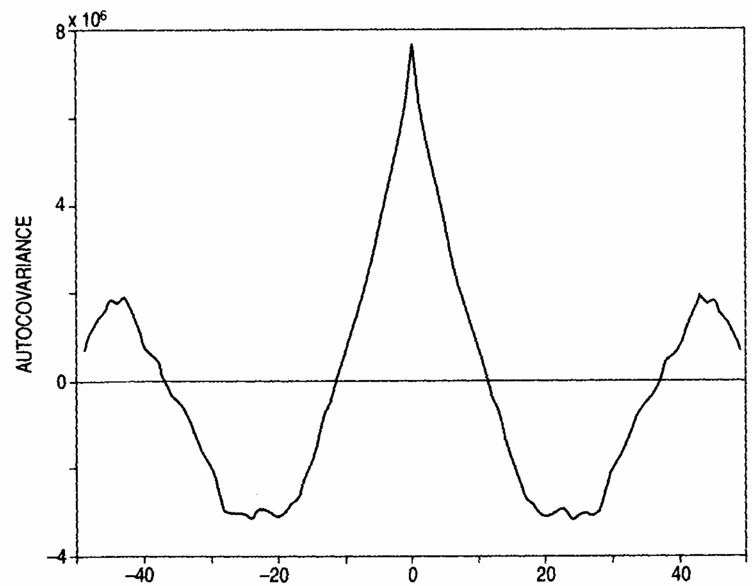
Fit to data

covariance function of residual









Wunsch

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

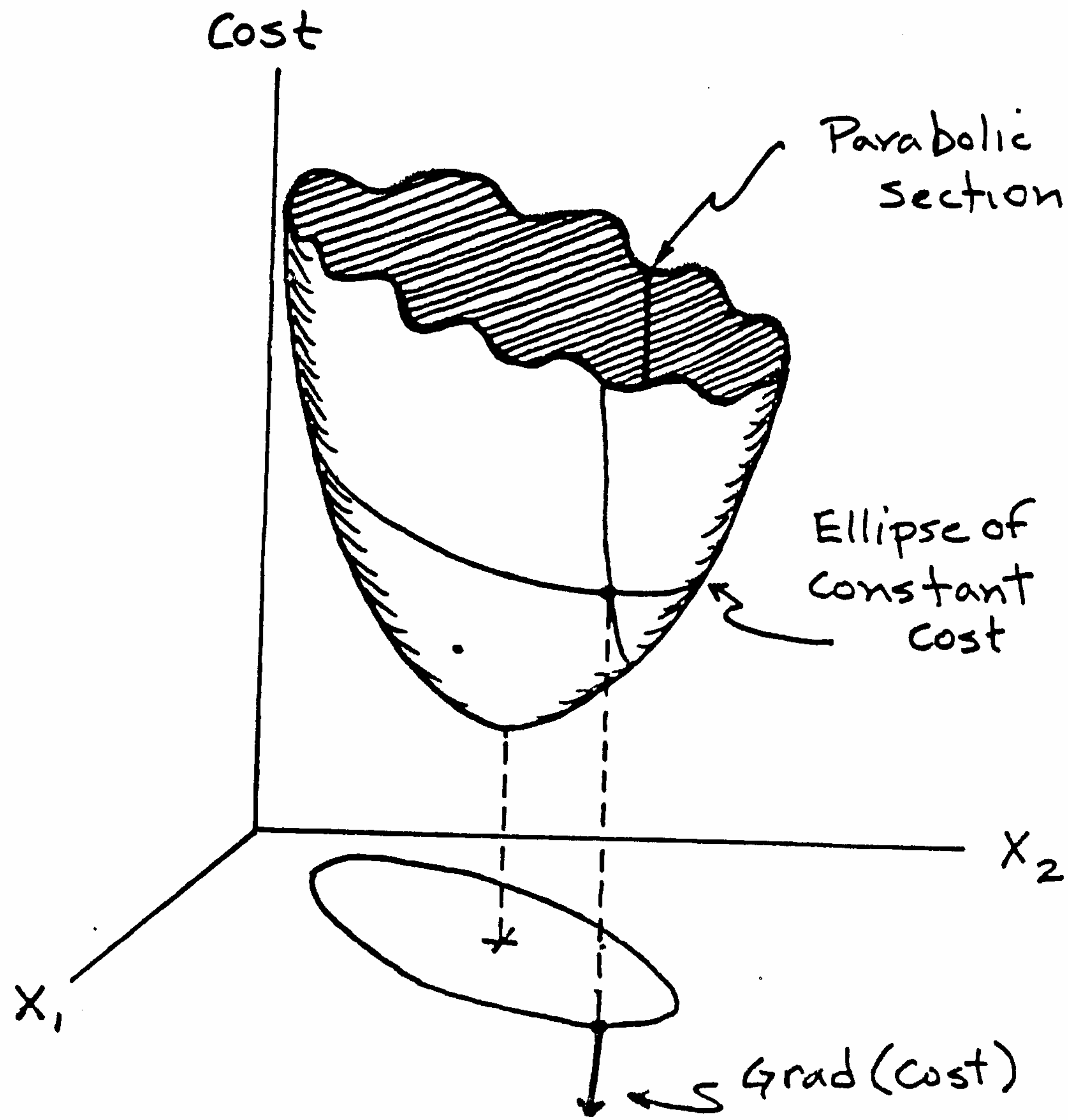
$$cov\langle a,b\rangle = \mathrm{H}^{-1}$$

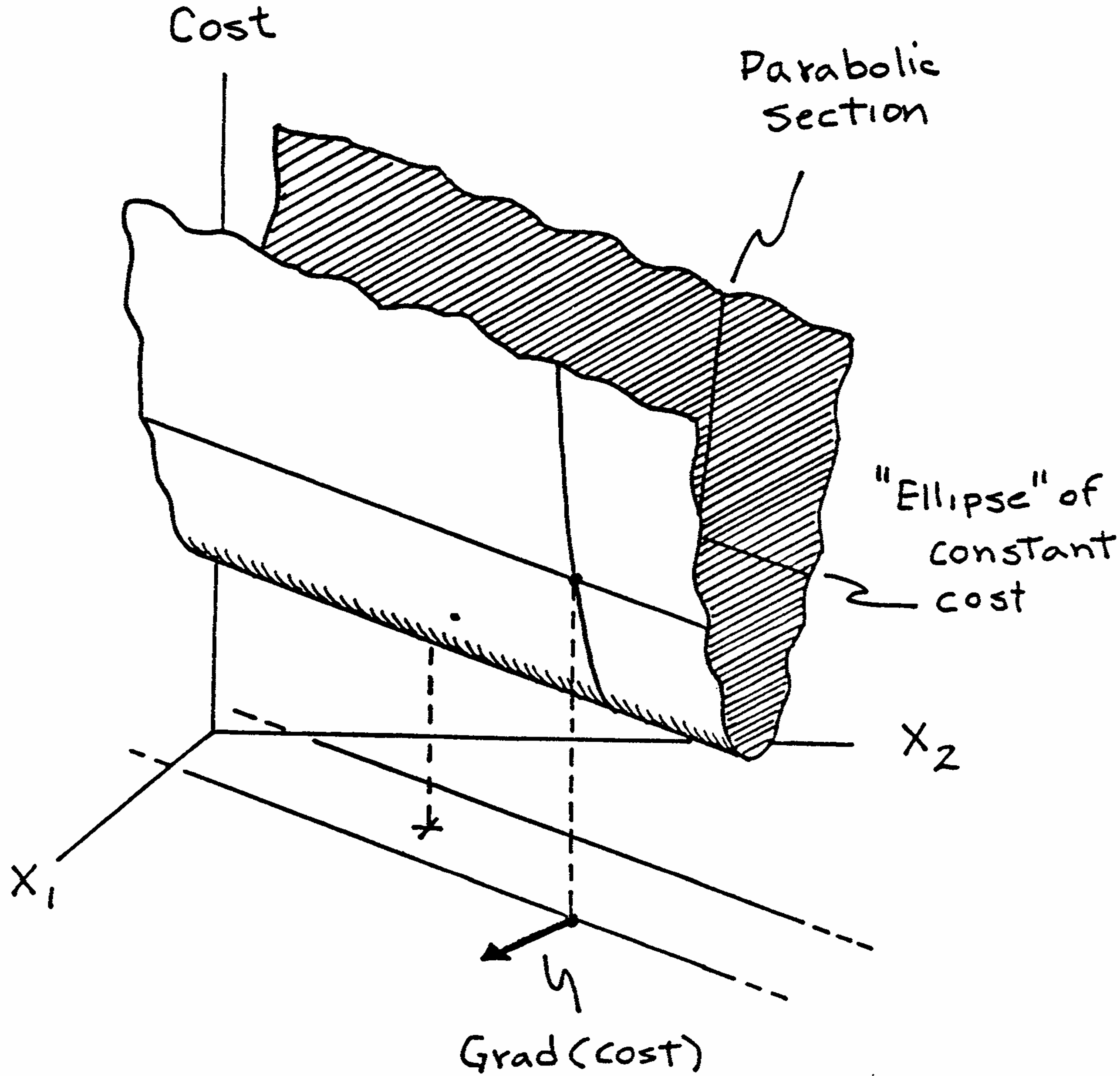
$$\mathrm{H} = \begin{pmatrix} \dfrac{\partial^2 j}{\partial a \partial a} & \dfrac{\partial^2 j}{\partial a \partial b} \\ \dfrac{\partial^2 j}{\partial b \partial a} & \dfrac{\partial^2 j}{\partial b \partial b} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} \dfrac{1}{\sigma_i^2} & \sum_{i=1}^{N} \dfrac{x_i}{\sigma_i^2} \\ \sum_{i=1}^{N} \dfrac{x_i}{\sigma_i^2} & \sum_{i=1}^{N} \dfrac{x_i^2}{\sigma_i^2} \end{pmatrix}$$

fitting a straight line through data

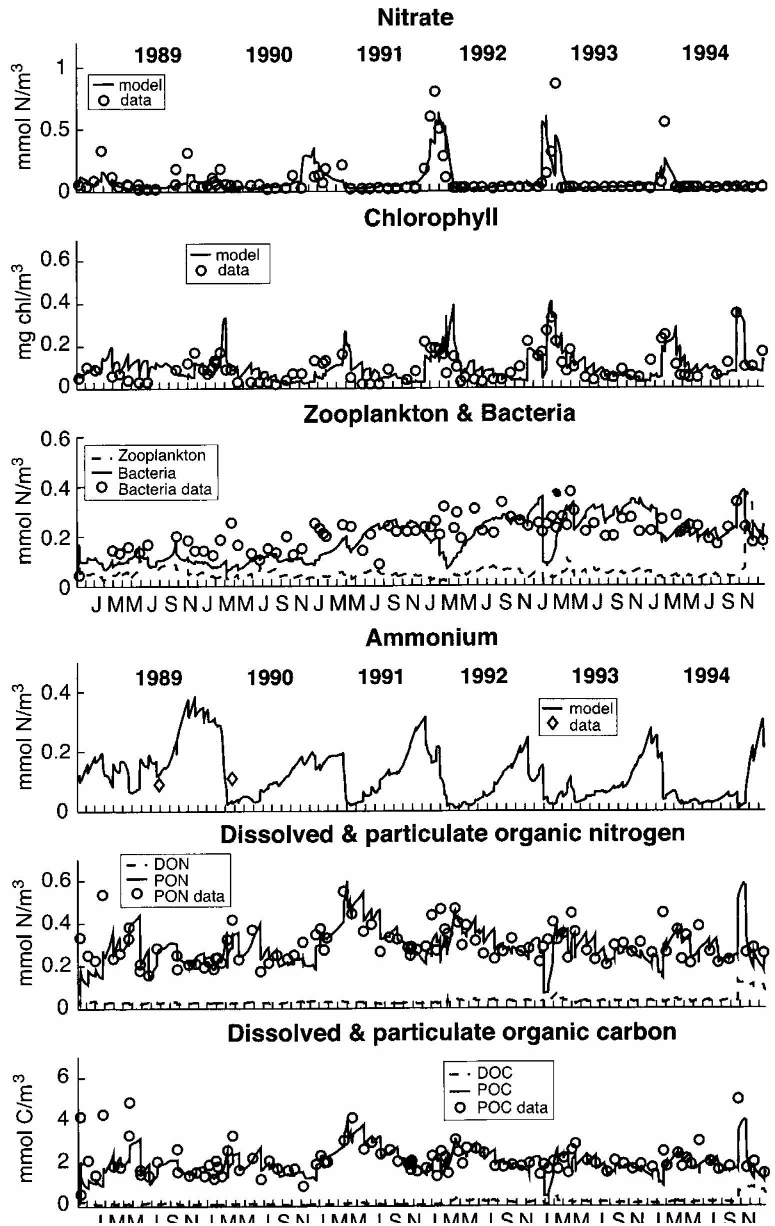
student exercise (for volunteers)

1. generate data yi at fixed xi,
2. use regression software to determine a and b,
3. generate k realizations for the data yi, 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.


Cost
Parabolic section
Ellipse of constant cost
X_2
X_1
Grad (cost)


Cost
Parabolic
Section
"Ellipse" of
constant
cost
X_2
X_1
Grad(cost)

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 2
Fixed model parameters

Symbol	Parameter	Value	Unit
k_w	attenuation coefficient of downwelling irradiance	0.04	m^{-1}
k_c	light attenuation due to phytoplankton	0.03	$m^2\ mmol^{-1}$
β_1, β_2, β_3	zooplankton assimilation efficiency	0.75	
r_1	zooplankton feeding preferency	0.5	
r_2, r_3	zooplankton feeding preferency	0.25	
δ	fraction of zooplankton losses going to DON	0.2	
ϵ	fraction of zooplankton losses going to ammonium	0.7	
η	ammonium:DON uptake ratio	0.6	
R_P	carbon to nitrogen ratio for phytoplankton	7	
R_Z	carbon to nitrogen ratio for zooplankton	5.5	
R_B	carbon to nitrogen ratio for bacteria	5	

Table 1
Adjusted model parameters

Symbol	Parameter	First guess	Unit
μ_1	phytoplankton maximum specific mortality rate	0.05	day^{-1}
k_1, k_2	half-saturation constants for nutrient and ammonium uptake	0.5	mmol N m^{-3}
k_5	phytoplankton mortality half-saturation constant	0.2	mmol N m^{-3}
ψ	nitrate uptake ammonium inhibition parameter	1.5	m^3 (mmol N)$^{-1}$
α	initial slope of the $P-I$ curve	0.025	m^2 W^{-1} day^{-1}
g	zooplankton maximum ingestion	1	day^{-1}
μ_2	zooplankton maximum loss rate	0.3	day^{-1}
k_3	zooplankton ingestion half-saturation constant	1	mmol N m^{-3}
k_6	zooplankton loss rate half-saturation constant	0.2	mmol N m^{-3}
μ_3	the bacterial excretion rate	0.05	day^{-1}
V_b	bacterial maximum uptake rate	2	day^{-1}
k_4	bacterial half-saturation constant for uptake	0.5	mmol N m^{-3}
μ_4	detrital breakdown rate	0.05	day^{-1}
w_g	detrital sinking rate	5.000	m day^{-1}

Analyse the possibility to determine p via the Hessian matrix **H**

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

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

$$\mathrm{H}_{\mathrm{l,m}} = \frac{\partial^2 j}{\partial p_l \partial p_m}$$

Analyse the possibility to determine p via the Hessian matrix H

A singular value analysis of H reveals

$$H = U S V^T$$

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

where matrix U contains the singular vectors and the diagonal matrix S the singular values

Singular vectors of **H**

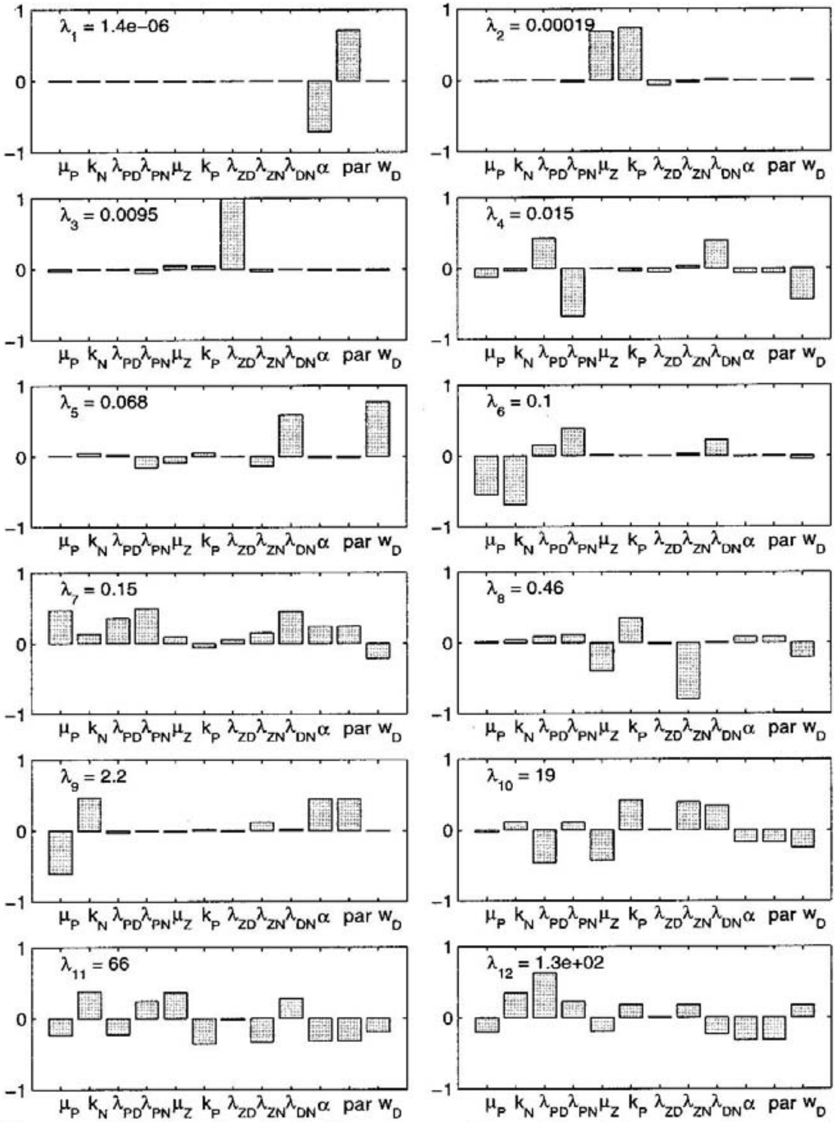


Fig. 3. Parameter resolution for experiment E1. Monthly measurements of nitrate and phytoplankton concentrations were employed.

Singular vectors of **H** (2)

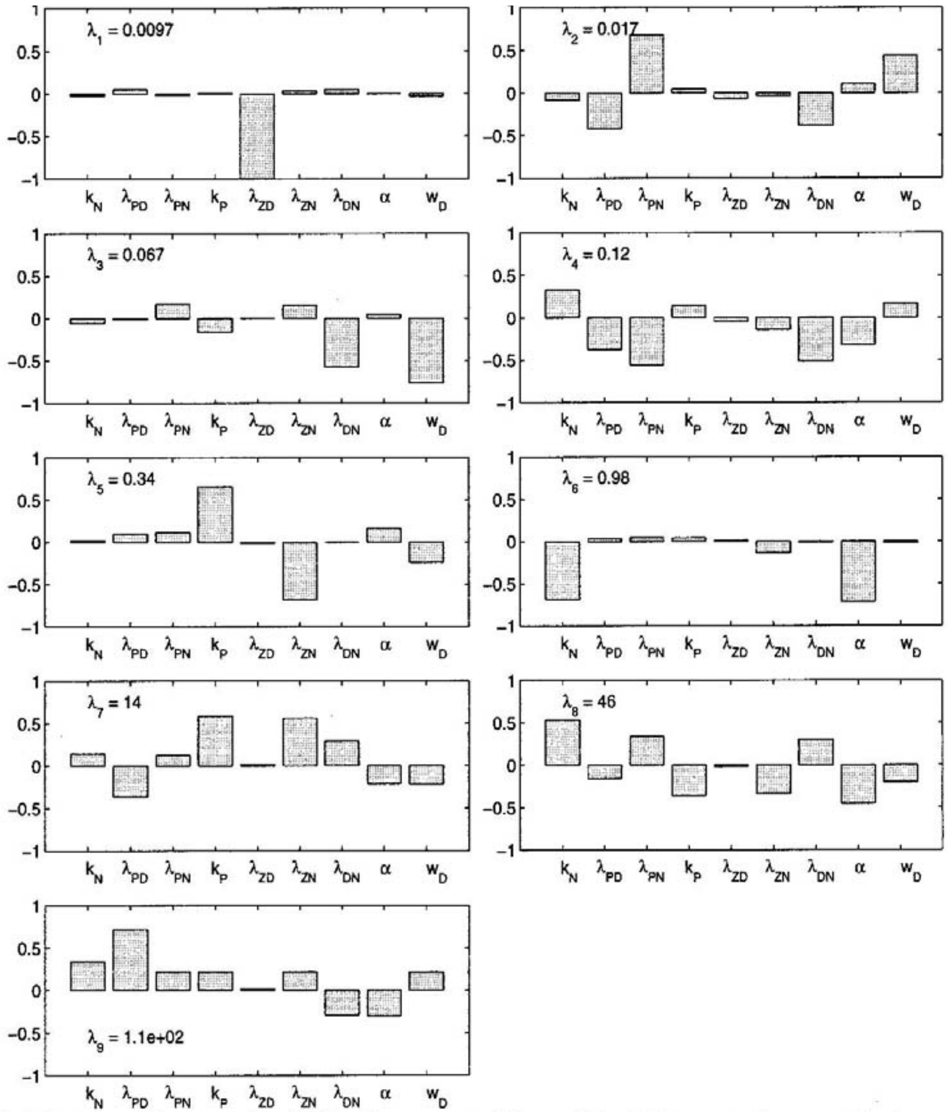
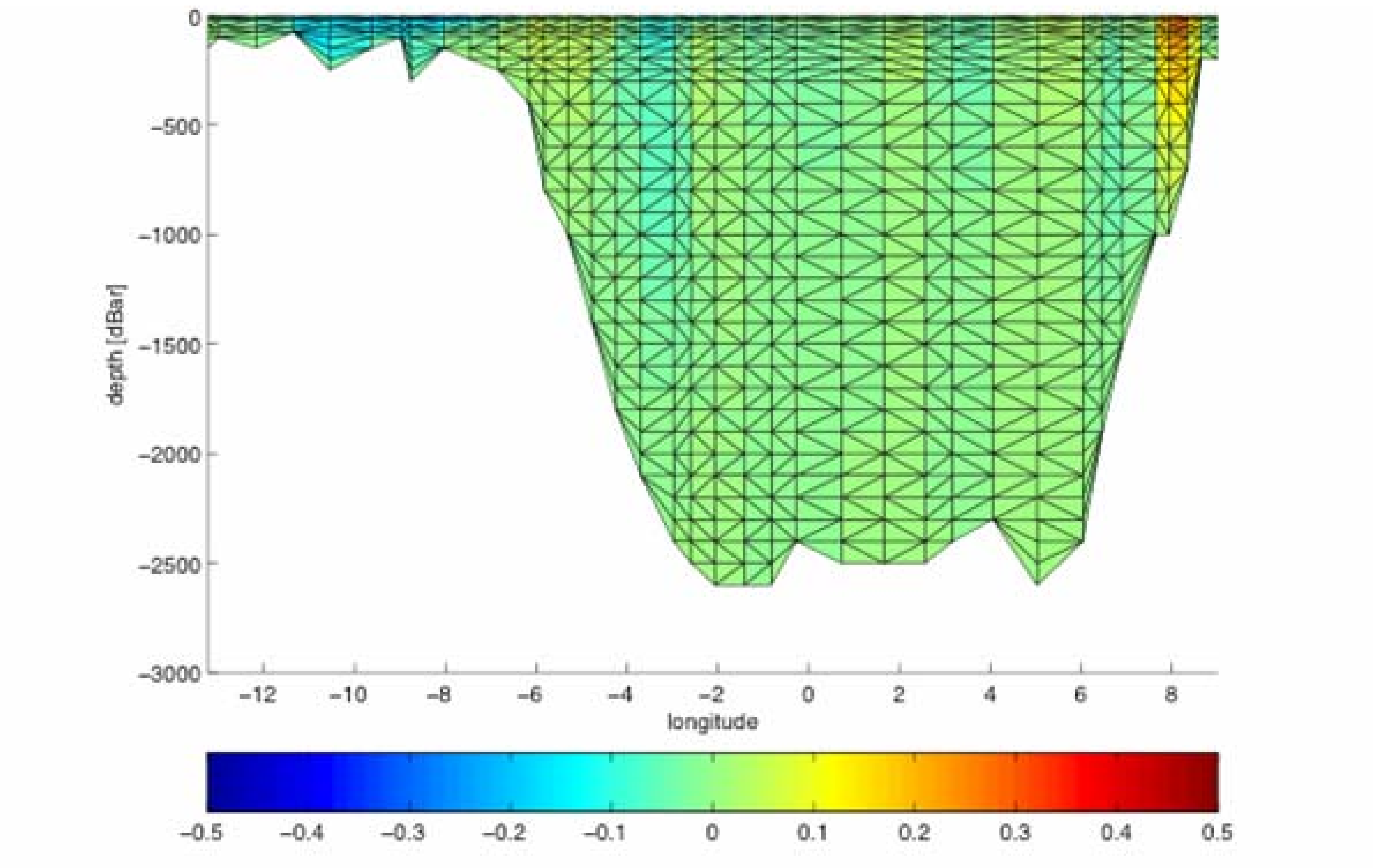


Fig. 5. Parameter resolution for experiment E3. Monthly measurements of nitrate and phytoplankton concentrations were employed. par, μ_Z

FEMSECT:

inverse model to determine velocities and transports through a hydrographic section



equations of section inverse model

1. Geostrophy and hydrostatics:

$$\mathbf{u} - \frac{g}{\rho_0 f} \int_{-H}^{z} (\mathbf{k} \times \nabla \rho) dz - \mathbf{u}_b = 0, \qquad (1)$$

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

$$\rho - \mathcal{R}(p, T, S) = 0, \qquad (2)$$

$$\frac{\partial^2 w}{\partial z^2} - \frac{g}{\rho_0 f^2} (\nabla \rho \times \mathbf{k}) \cdot \nabla f = 0, \qquad (3)$$

with boundary conditions

$$w(0) - (\frac{1}{\rho} \mathrm{curl}\, \tau + F^{w}) = 0,$$

$$w(-H) - [(\mathbf{u}_b \cdot \nabla H) + F_w] = 0,$$

Tracer advective diffusive equations (N=8)

$$(\mathbf{u} \cdot \nabla C_n) + w\frac{\partial C_n}{\partial z} - F_n = 0,$$

$$1 \le n \le N : \theta, S, C_3, \ldots, C_N,$$

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

$$J_0 = \frac{1}{2}[\sum_{m,n}{}^* \sum_{z} \sum_{z'} (\hat{I} C_m(z) - C_m^*(z))^\dagger$$

$$W_{m,n}(z,z')(\hat{I} C_n(z') - C_n^*(z'))]$$

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

$$+ W^w \int_{z=0} (F^w)^2 dx + W_w \int_{z=-H} (F_w)^2 dx$$

$$+ \int_\Omega \sum_n F_n^\dagger W_F^n F_n d\Omega$$

$$+ \int_\Omega \sum_n (\hat{S}_1 C_n)^\dagger D^n (\hat{S}_1 C_n) d\Omega$$

$$+ \int_\Omega \sum_n (\hat{S}_1 F_n)^\dagger D_F^n (\hat{S}_1 F_n) d\Omega$$

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

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 var ($L^T x$) is

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

$$solve \qquad \mathrm{H}z = L$$

$$so\,that \;\; \mathrm{H}^{-1} L = z$$

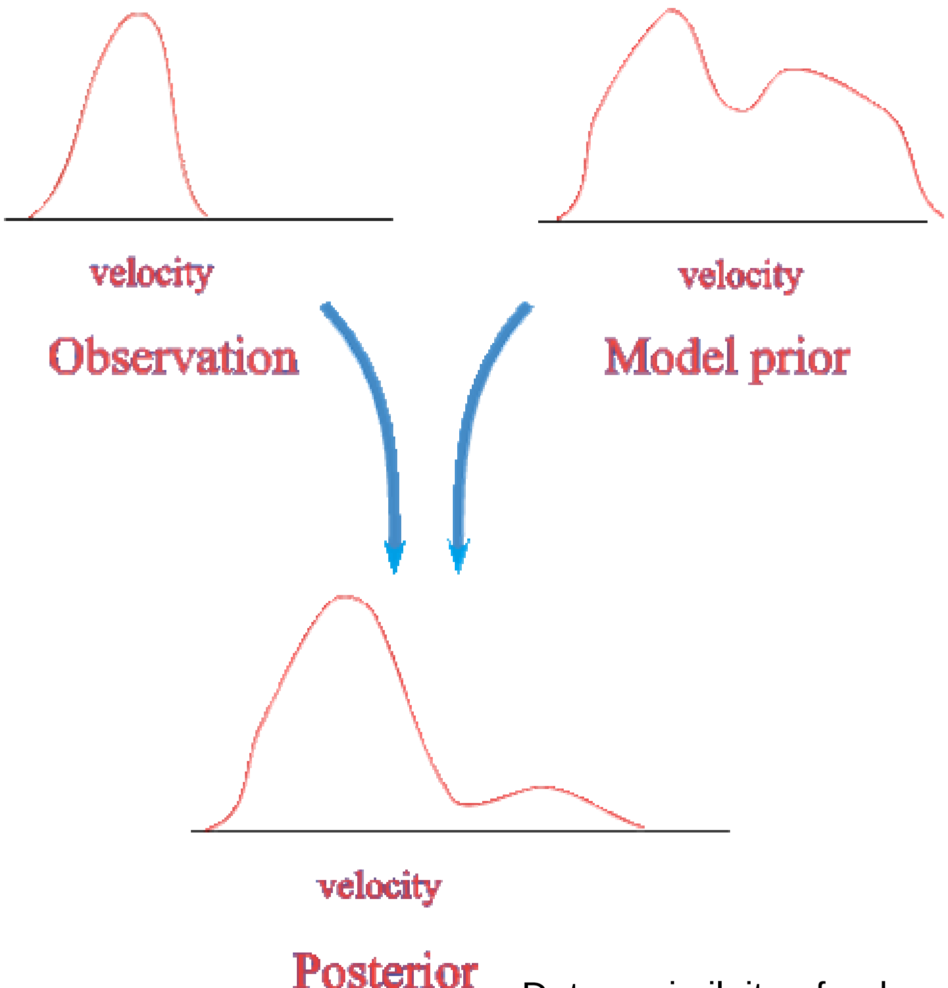
$$then \;\; \mathrm{var}(L^T x) = L^T z$$

covariance with say heat transport
L2x is the product

$$L_2^T\, cov\langle x, x^T \rangle L_1 \;= L_2^T\, \mathrm{H}^{-1} L_1$$
$$= L_2^T z$$

Codes for calculating Hessian times vector are in packages like TAMC

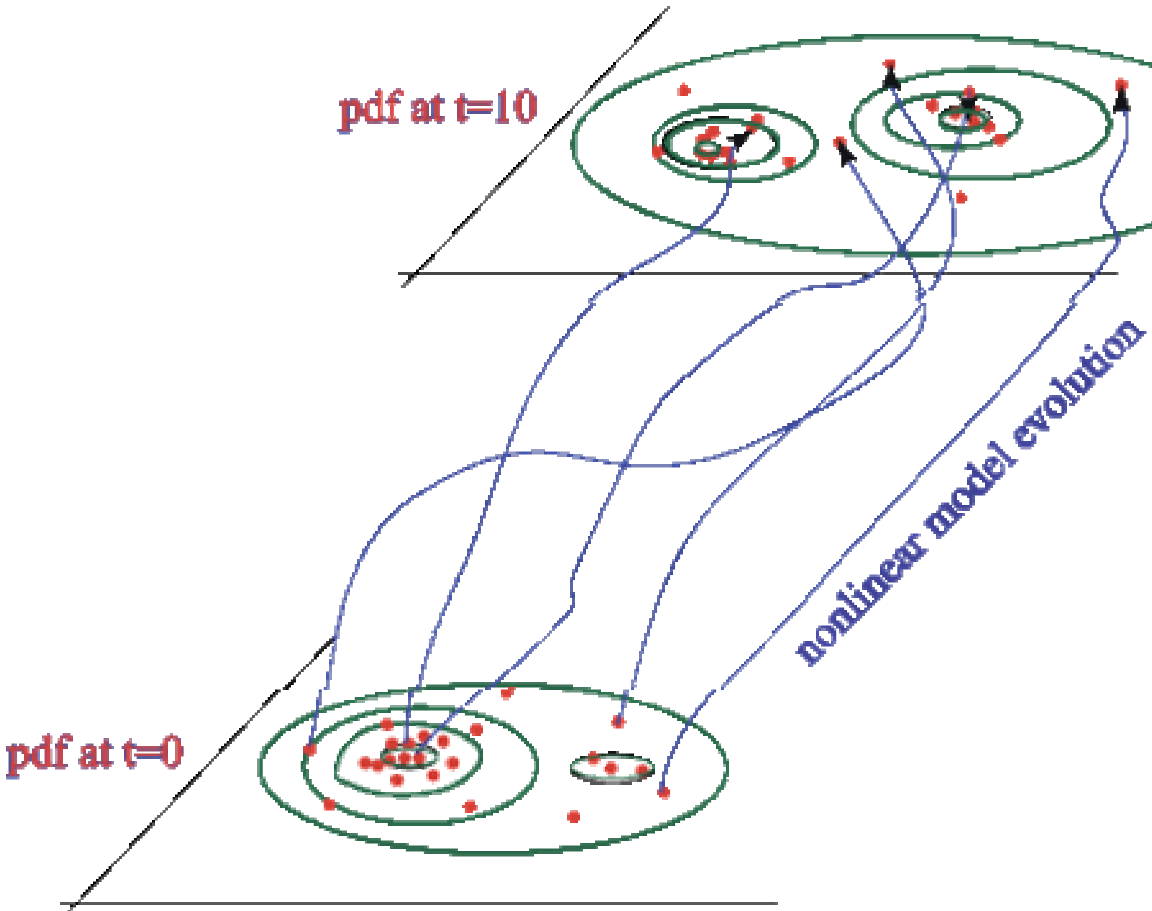
Data assimilation: general formulation



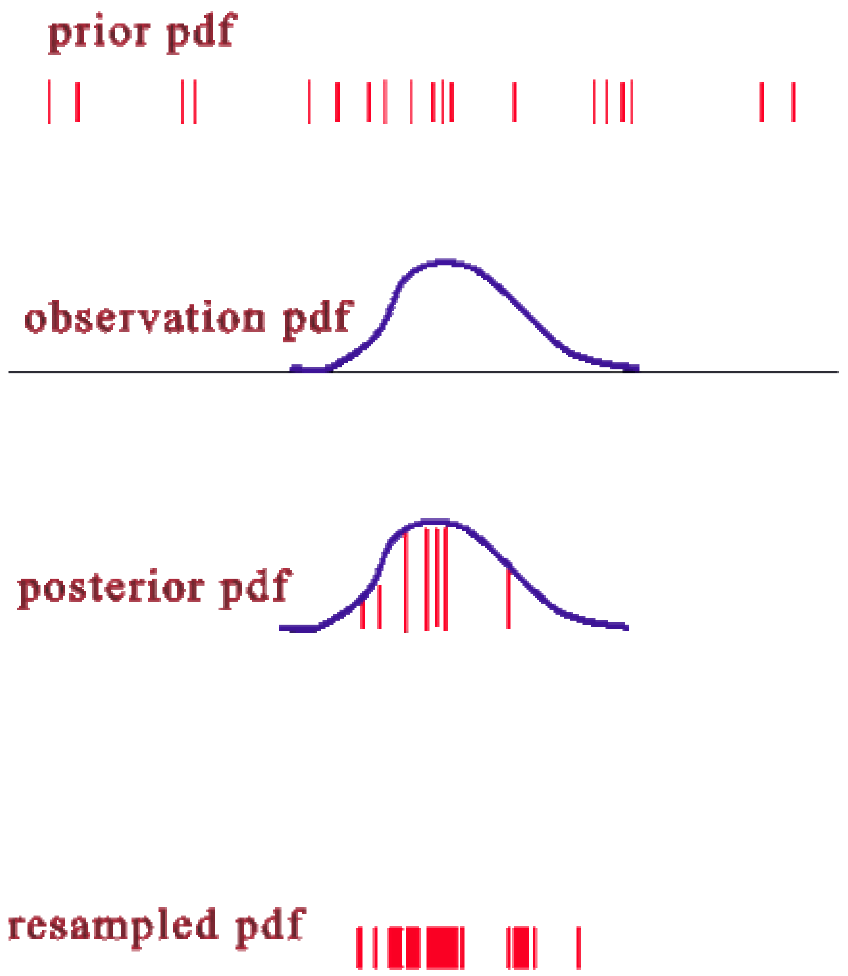
$$f(\psi|d) = \frac{f(d|\psi)\, f(\psi)}{f(d)}$$

NO INVERSION !!!

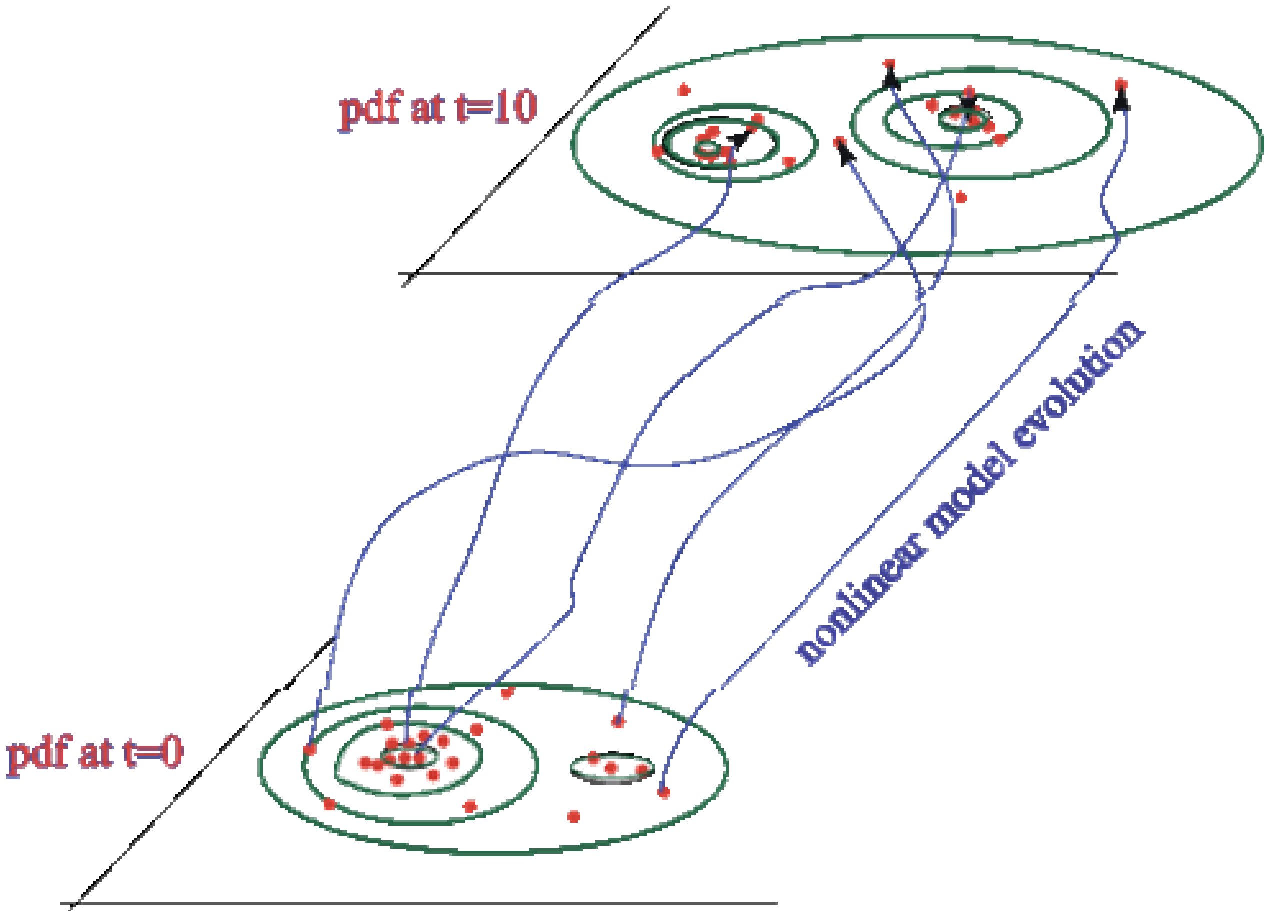
Propagation of pdf: Ensemble methods 'efficient' propagation for nonlinear models



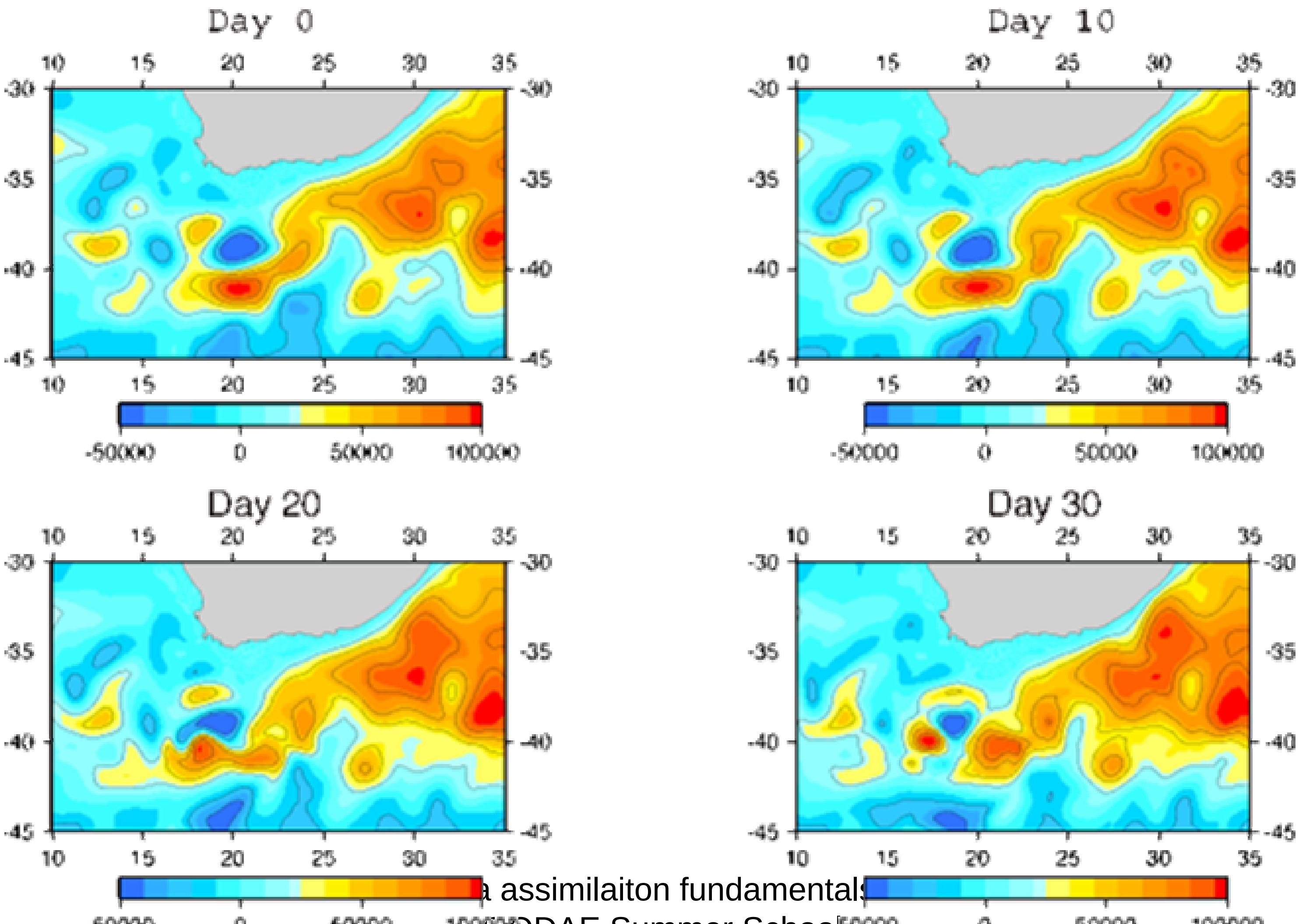
Sequential Importance Resampling



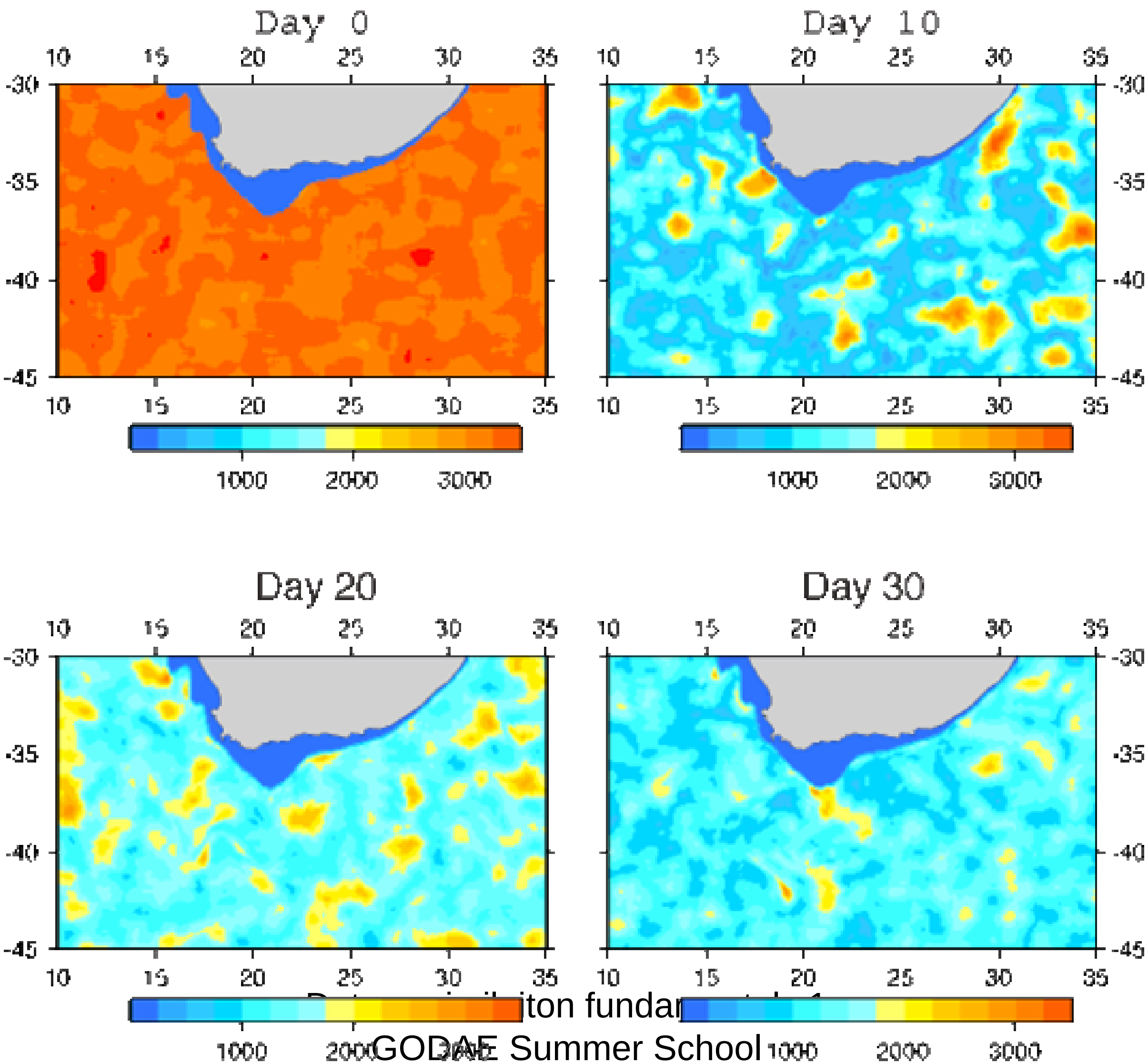
Propagation of pdf: Ensemble methods
‘efficient’ propagation for nonlinear models



SIR-results for ocean around South Africa



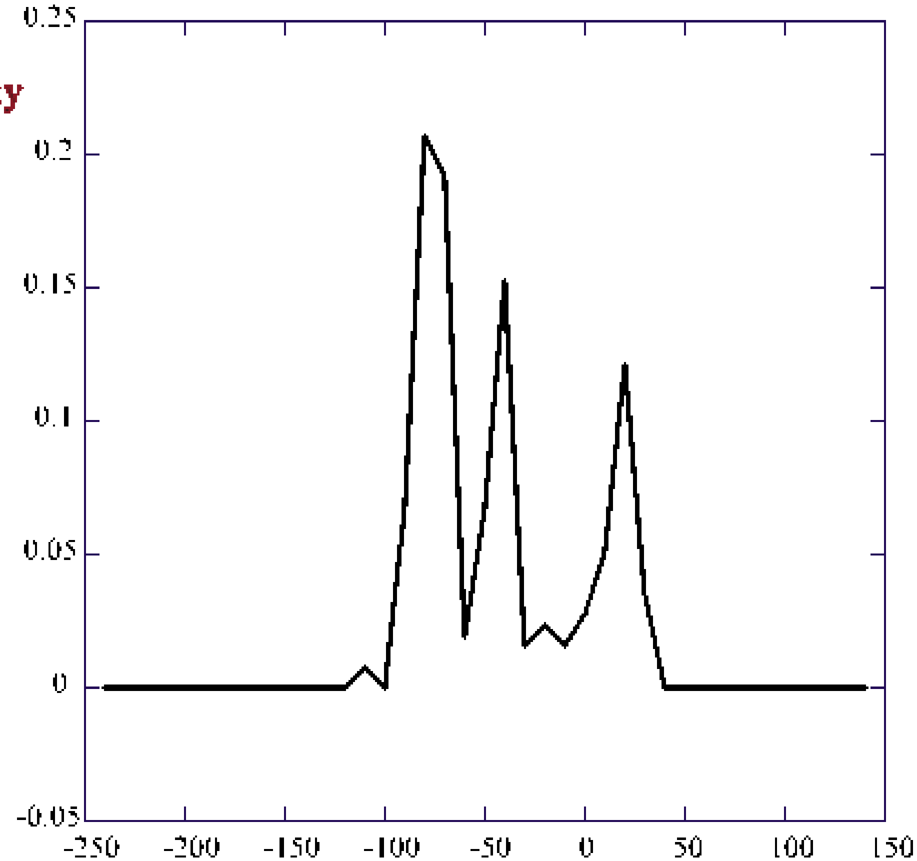
SIR-results: errors



The Ocean is nonlinear…

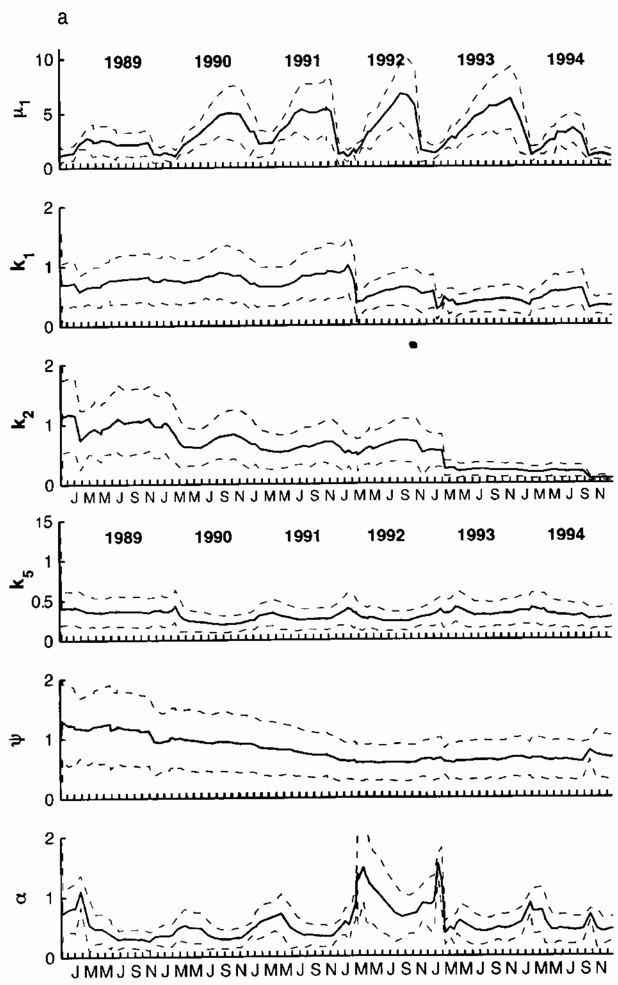
Probability density function at 40S 20E

Probability density



Sea-surface height at 40S 20E

Retrieved bio-parameters(t)



e evolution of the estimates for the model parameters. Values are normalized with respect to the model parameter initia

Table 3
Stable estimates for model parameters

Symbol	First guess	Optimized value	Error variance	Unit
g	1.0	0.40	0.15	day^{-1}
μ_3	5×10^{-2}	9×10^{-3}	4.5×10^{-3}	day^{-1}
μ_4	5×10^{-2}	1.1×10^{-2}	1.4×10^{-3}	day^{-1}
μ_2	0.3	0.24	0.13	day^{-1}
k_6	0.2	0.20	0.08	mmol N m^{-3}
k_3	1.0	2.62	1.4	mmol N m^{-3}
ψ	1.5	0.99	0.33	$\text{m}^3 (\text{mmol N})^{-1}$
w_g	5.0	4.25	2.17	m day^{-1}

answer:

the photo of the iceberg was taken from here


END