MarDATA Workshop

Ensemble-Based Data Assimilation: Concepts, Methods, and Hands-On Tutorials

Lars Nerger

Alfred Wegener Institute
Helmholtz Center for Polar and Marine Research
Bremerhaven, Germany

Overview

Overall aim of the workshop: Get familiar with concepts of data assimilation and selected methods for high-dimensional models

4 Blocks – lectures plus hands-on tutorials

Today

1. Data Assimilation Basics
   - tutorial 1
2. Ensemble Data Assimilation
   - tutorial 2

Tomorrow

3. Advanced Ensemble Kalman filters
   - tutorial 3
4. Practical aspects

Ensemble-Based Data Assimilation: Concepts, Methods, and Hands-On Tutorials

Lecture 1: Concepts and Basic Data Assimilation

Lars Nerger

Alfred Wegener Institute
Helmholtz Center for Polar and Marine Research
Bremerhaven, Germany

MarDATA Workshop on Data Assimilation, Bremerhaven, July 24-25, 2024

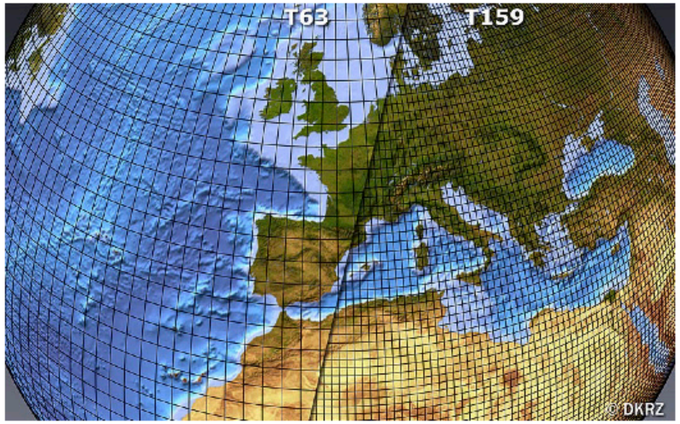
Overview – Lecture 1

Introduction to data assimilation basics: What do we aim to achieve? What do we need? Basic optimization and estimation. Basic assimilation methods

- Application examples
- Data Assimilation fundamentals
- Optimization and estimation
- BLUE
- Basic DA: Nudging and Optimal Interpolation

Data Assimilation – Combining Models and Observations

Models



+

Observations

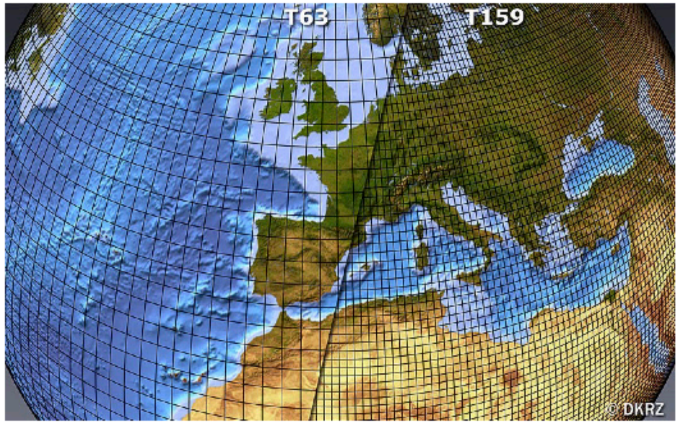
Combine both sources of information

quantitatively and optimally by computer algorithm

→ Data Assimilation

Data Assimilation – Characteristics of Model and Observations

Models



Observations

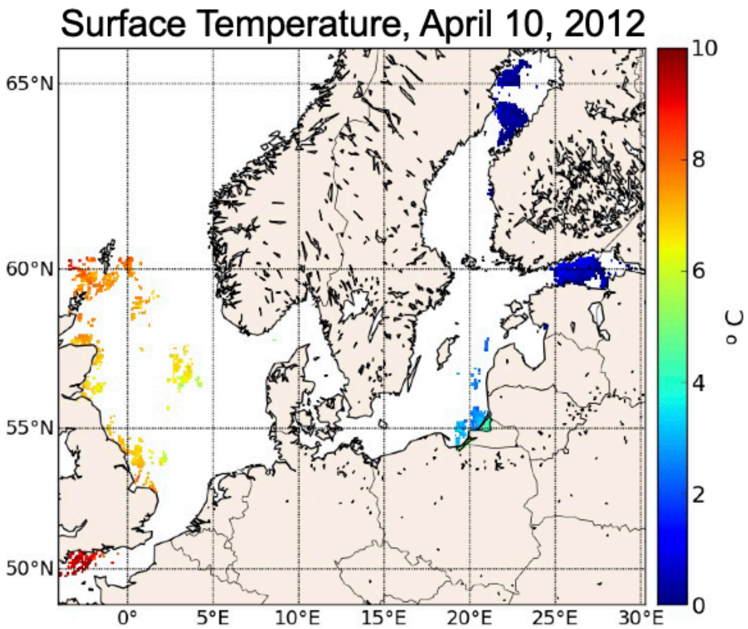
- idealized representation of processes
- all fields, fluxes on model grid
- finite, quantifiable errors

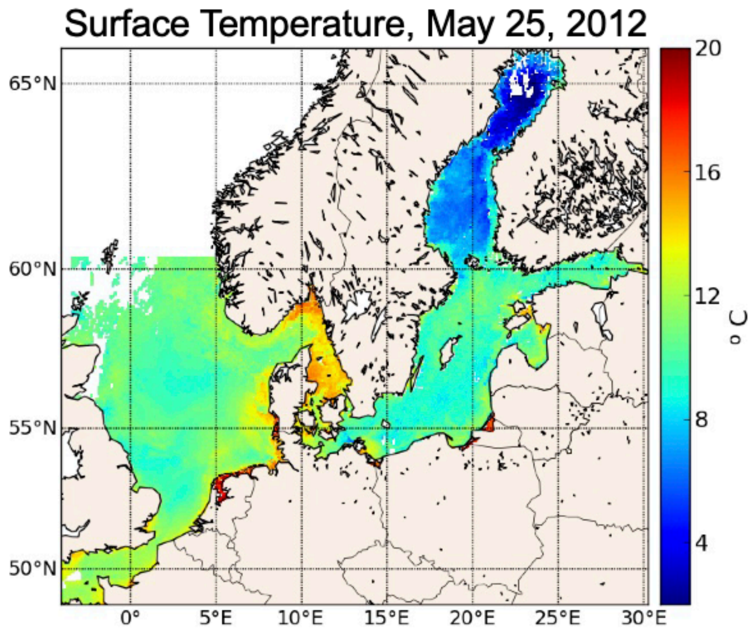
- measurements of 'reality'
- incomplete information:
  only some fields, data gaps
  ocean data: mainly surface (satellite)
- finite, quantifiable errors

Application examples

from Oceanography

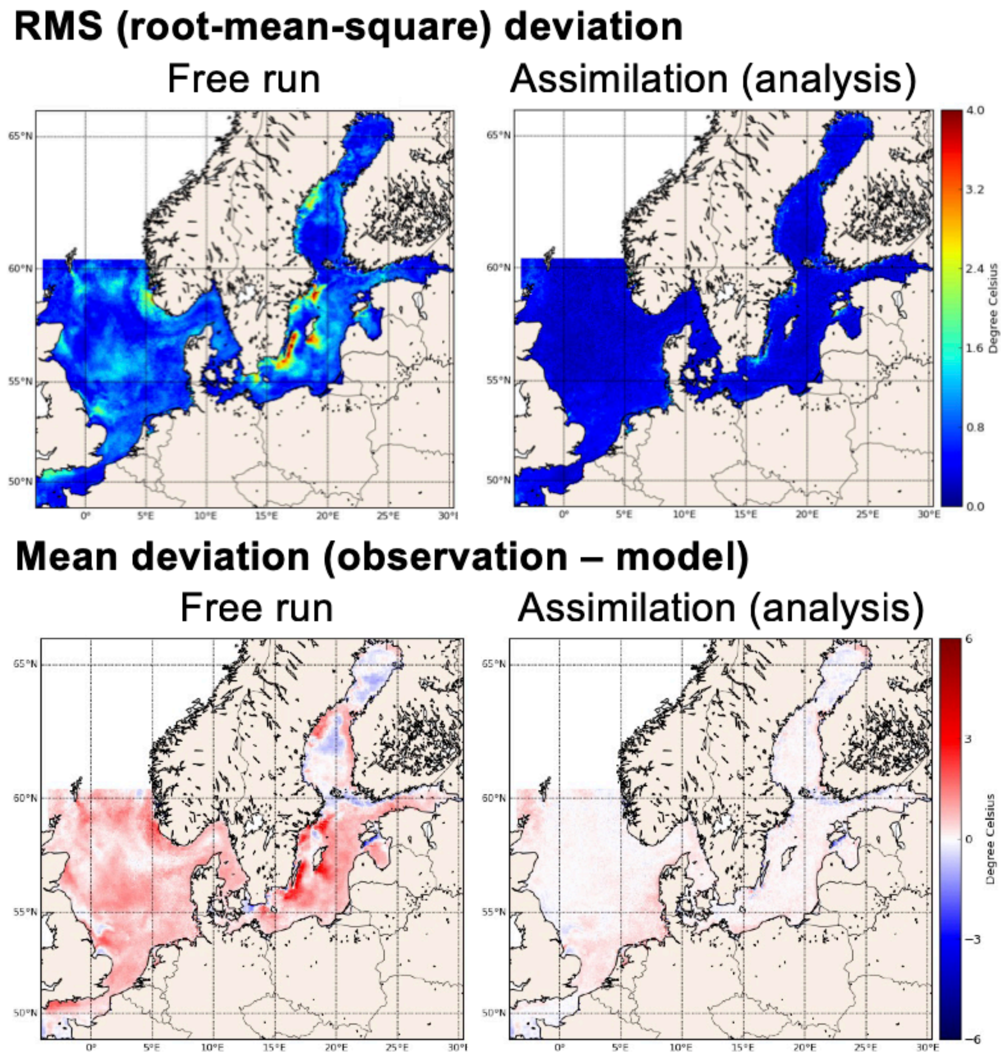
Regional Data Assimilation – Satellite Data





- Surface temperature (from NOAA satellites)
- 12-hour composites
- Strong variation of data coverage (clouds)

DA – effect on Temperature (September 2012)



Regional DA application:
North and Baltic Seas

Assimilate surface temperature each 12 h

Compare assimilated estimate with assimilated surface temperature data (monthly average)

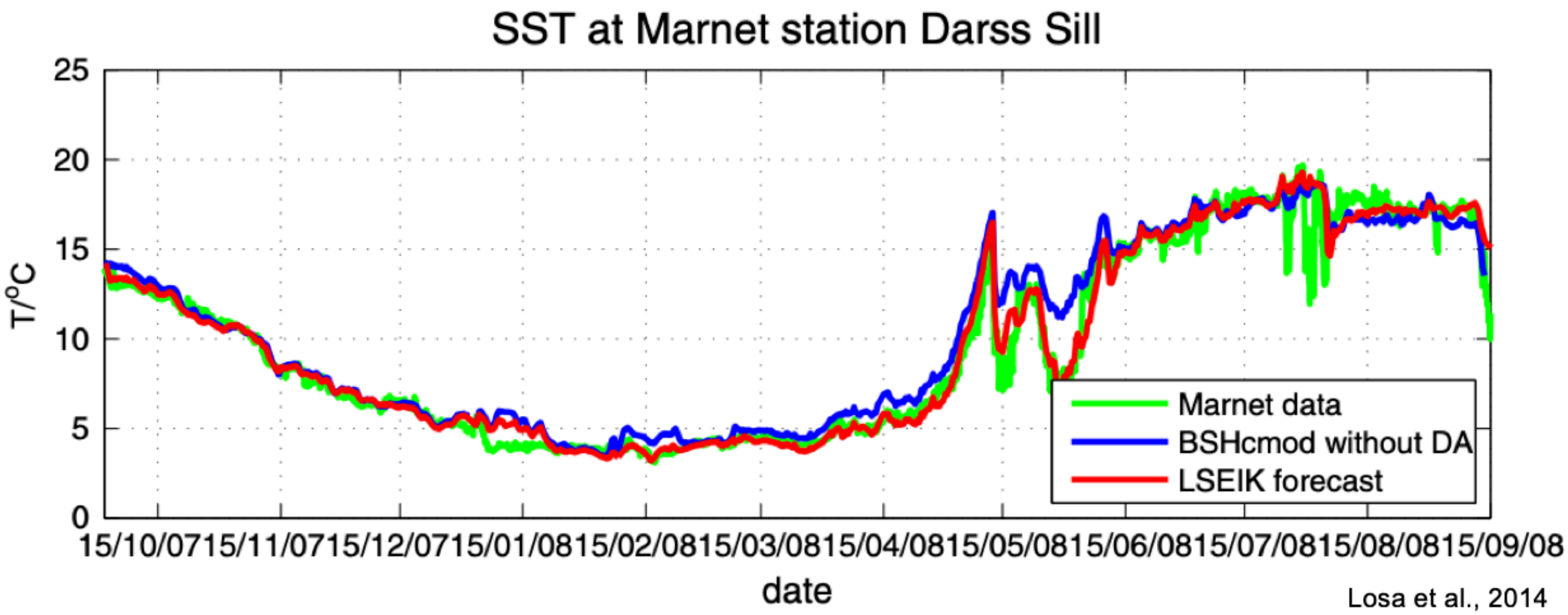
➢ Reduce RMS deviation and mean deviation (bias)

necessary effect

DA does not work if not fulfilled

DA – effect on Temperature

Compare to independent station data



Losa et al., 2014

Further condition for data assimilation:

- reduce deviations from independent observations (sufficient condition: show that the DA does work correctly)

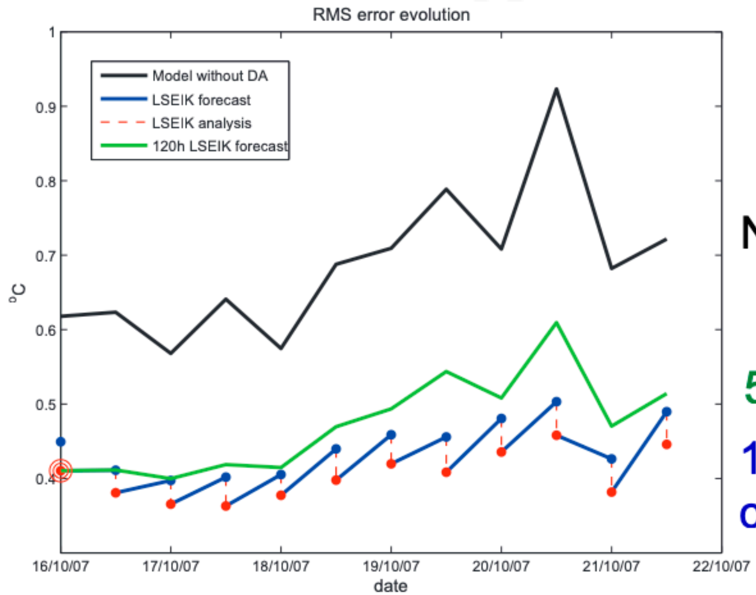
Note: Usually, the observed fields are easy to improve by DA

Improving forecasts

DeMarine

From: S. Losa et al., J. Mar. Syst. 105–108 (2012) 152–162

Impact of Assimilation for temperature forecasts



No assimilation

5-day forecast

12-hour assimilation cycling

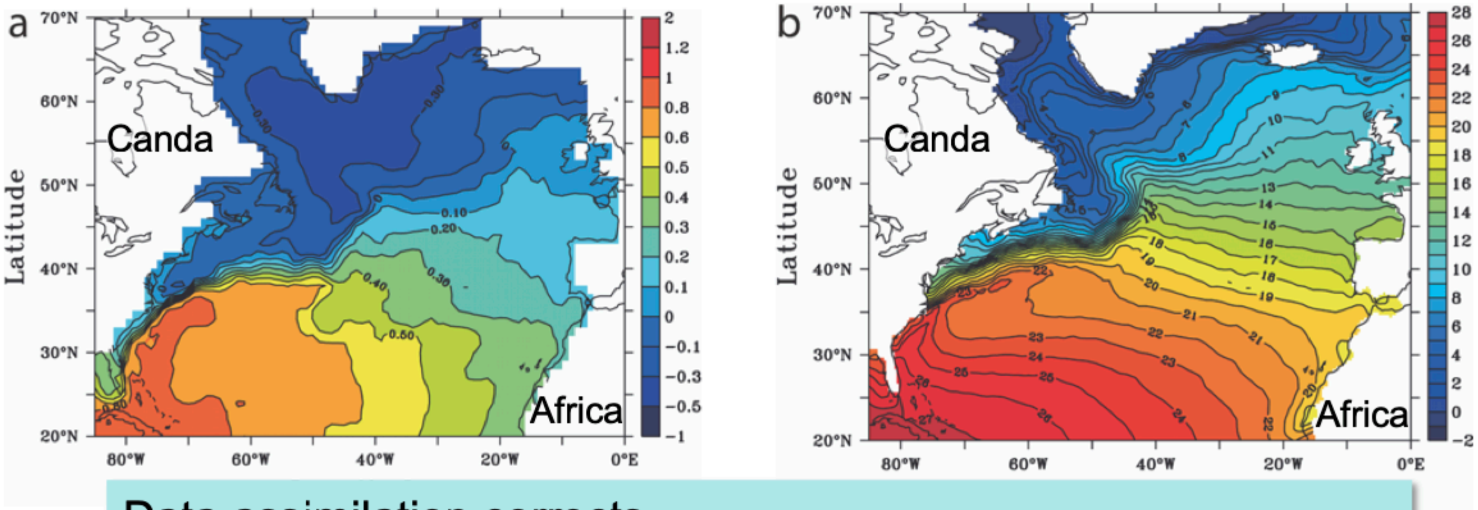
- Very stable 5-days forecasts
- At some point the improvement might break down due to dynamics

Example: Multivariate data assimilation to improve ocean circulation

From: Brankart et al., J. Geophys. Res. 108(C3) 3074, 2003

Assimilation of satellite data in North Atlantic:

sea surface height | sea surface temperature



Data assimilation corrects

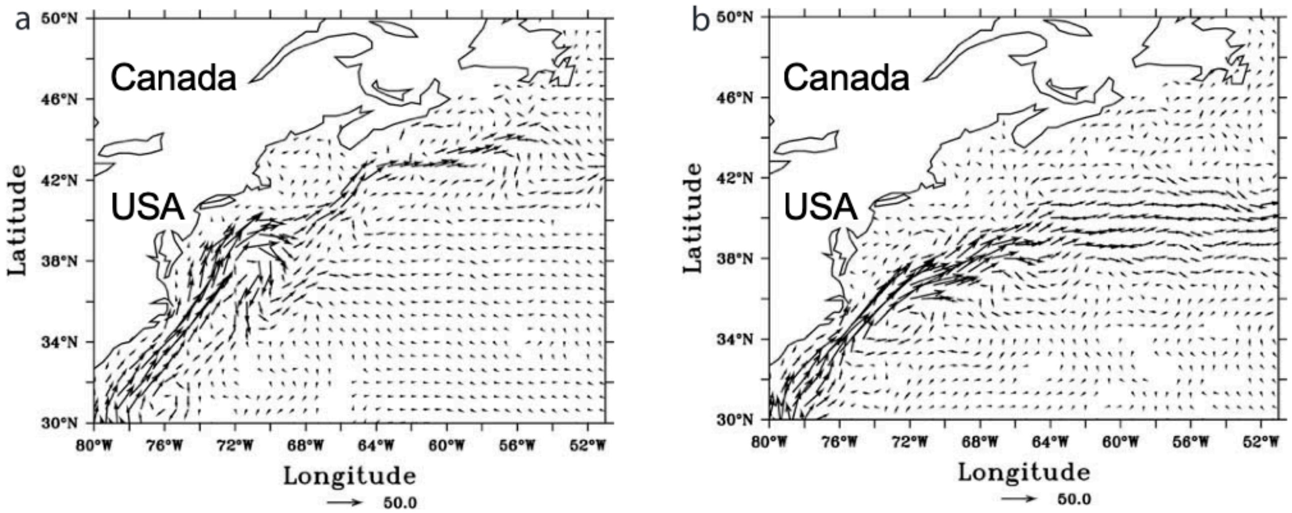
directly: observed sea surface height and temperature

via estimated correlations: subsurface fields and velocities

Example: Improving ocean circulation (2)

From: Brankart et al., J. Geophys. Res. 108(C3) 3074, 2003

Velocity field with assimilation of surface temperature and height data



no assimilation

with assimilation

With data assimilation:

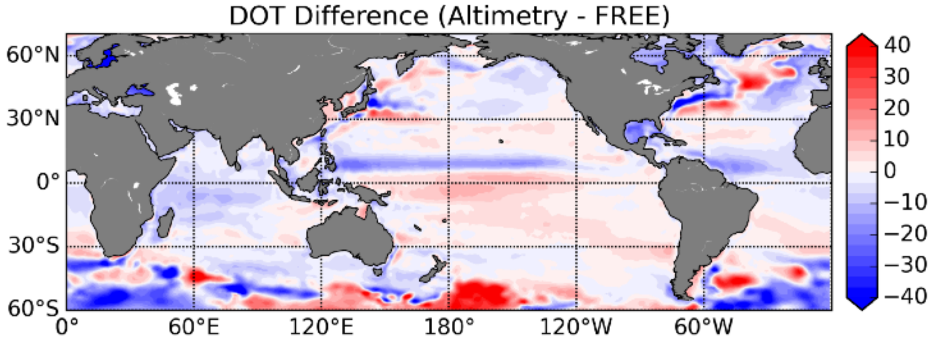
Improvement of mean velocity field in Gulf Stream region

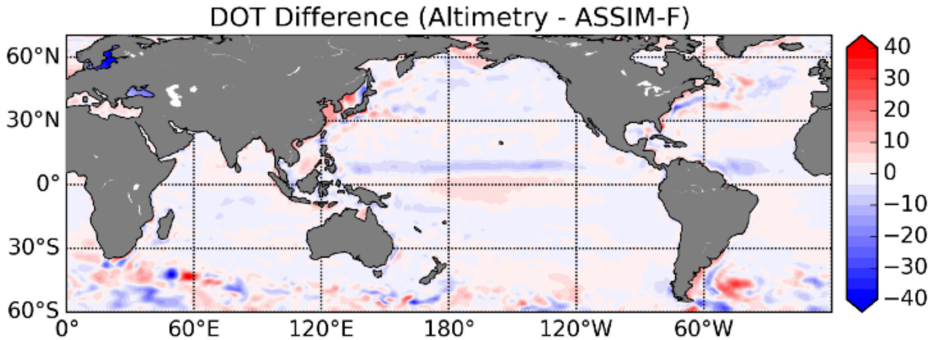
Indirect influence as no velocity data assimilated

Longe-range effect

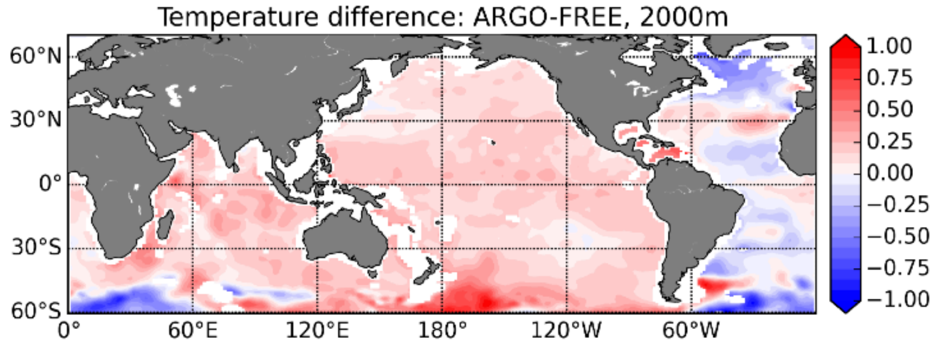
Example: Assimilate satellite sea surface height data (DOT)

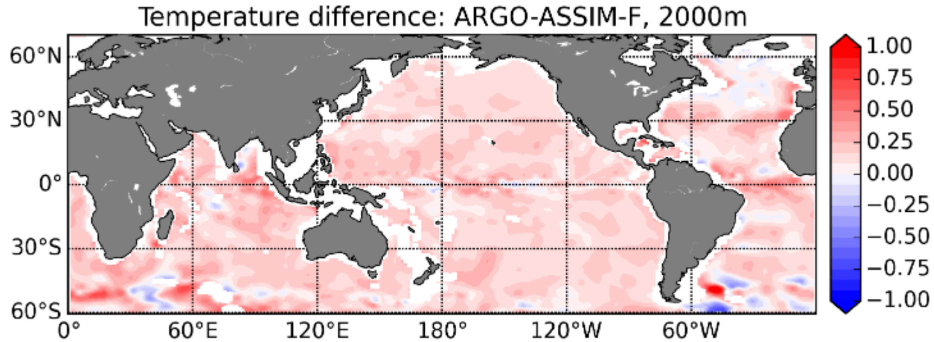
Reduce difference to assimilated data (necessary)





Improve also temperature at 2000m depth





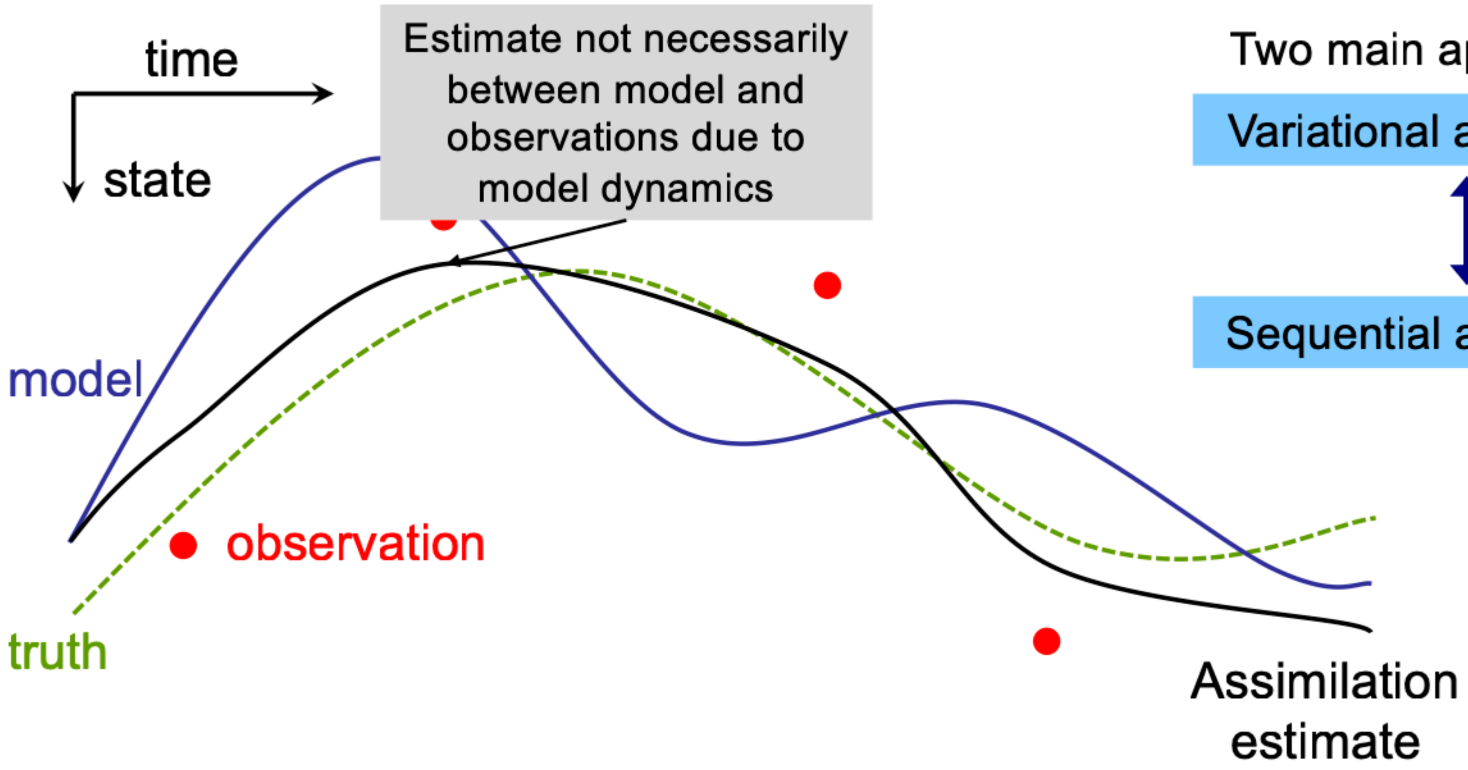
Androsov et al., J. Geodesy, (2019) 93:141–157

Data Assimilation

Combine Models and Observations

Data Assimilation – a general view

Consider some physical system (ocean, atmosphere, land, …)



Two main approaches:

Variational assimilation

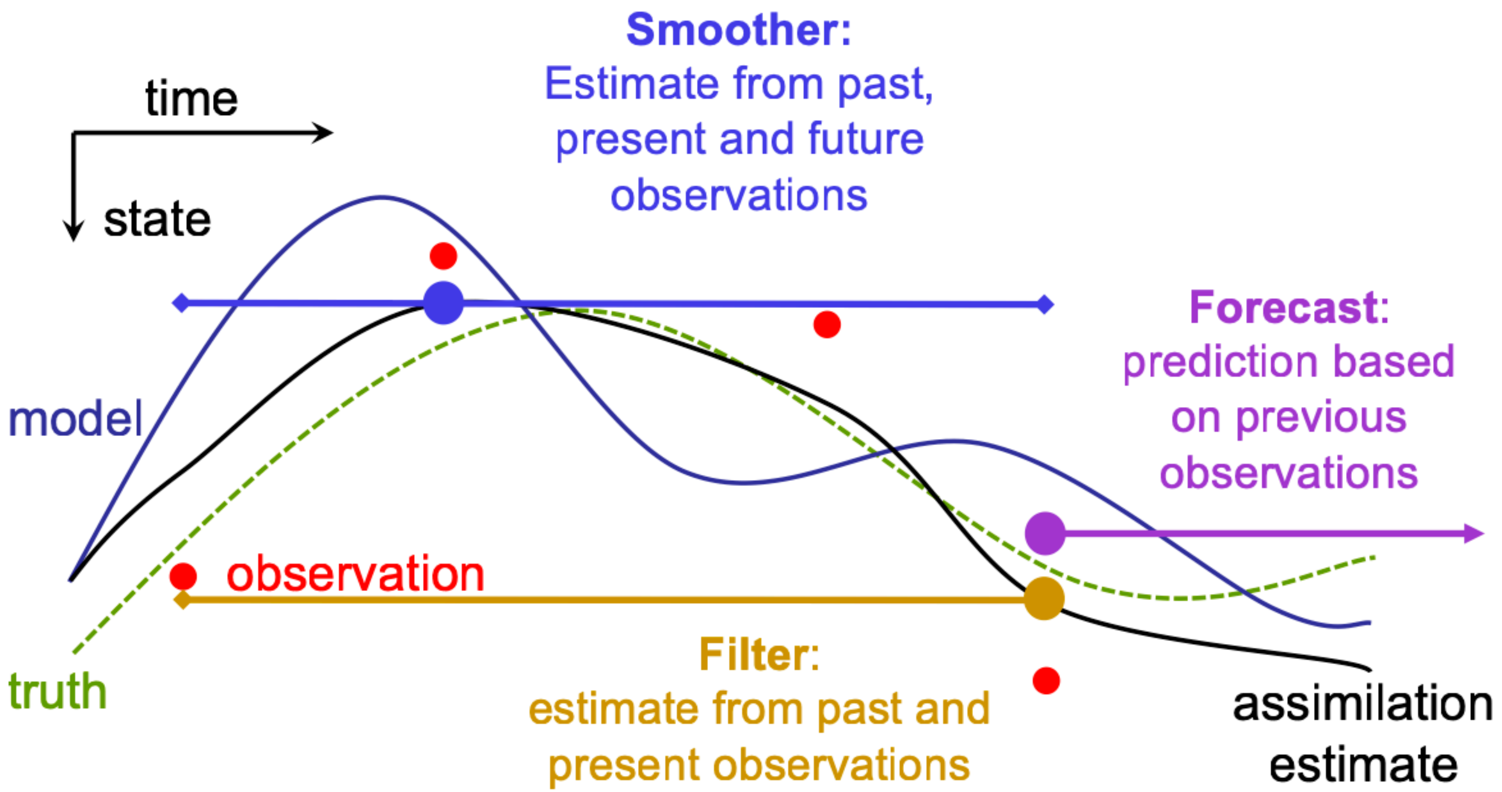
↕

Sequential assimilation

Optimal estimate basically by least-squares fitting
(but constrained by model dynamics)

Terminology: Smoother – Filter – Forecast

Consider some physical system (ocean, atmosphere, land, …)



What is Data Assimilation?

Data Assimilation is the science of combining information from computational models and observations accounting for uncertainties in both information sources.

- Optimal estimation of system state:
  - initial conditions (for weather/ocean forecasts, …)
  - state trajectory (temperature, concentrations, …)
  - parameters (growth of phytoplankton, …)
  - fluxes (heat, primary production, …)
  - boundary conditions and 'forcing' (wind stress, …)
- More advanced: Improvement of model formulation and observation system
  - Detect systematic errors (bias)
  - Revise parameterizations based on parameter estimates
  - Find observations that are relevant

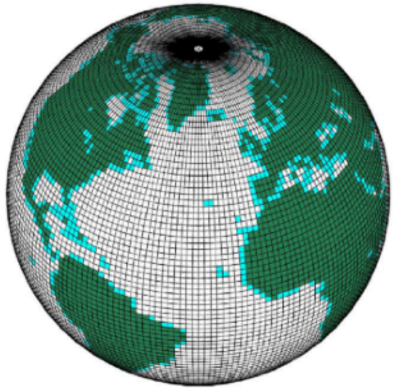
Needed for Data Assimilation

1. Model
   - with some skill
2. Observations
   - with finite errors
   - related to model fields
3. Data assimilation method
   - usable with model and observations (e.g. model dimension or nonlinearity)

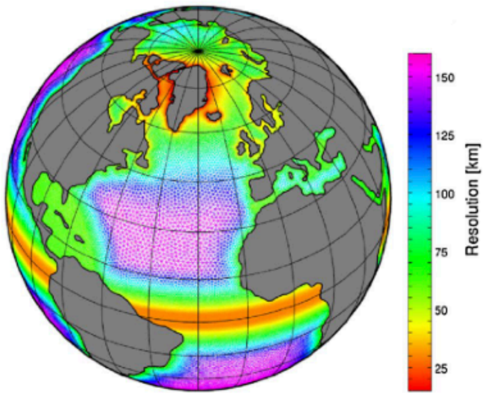
Models

Simulate dynamics of ocean

- Numerical formulation of relevant terms
- Discretization with finite resolution
  in time and space
- “forced” by external sources
  (for ocean: atmosphere, river inflows)



Uniform-resolution mesh



Variable-resolution mesh
(ocean model FESOM)

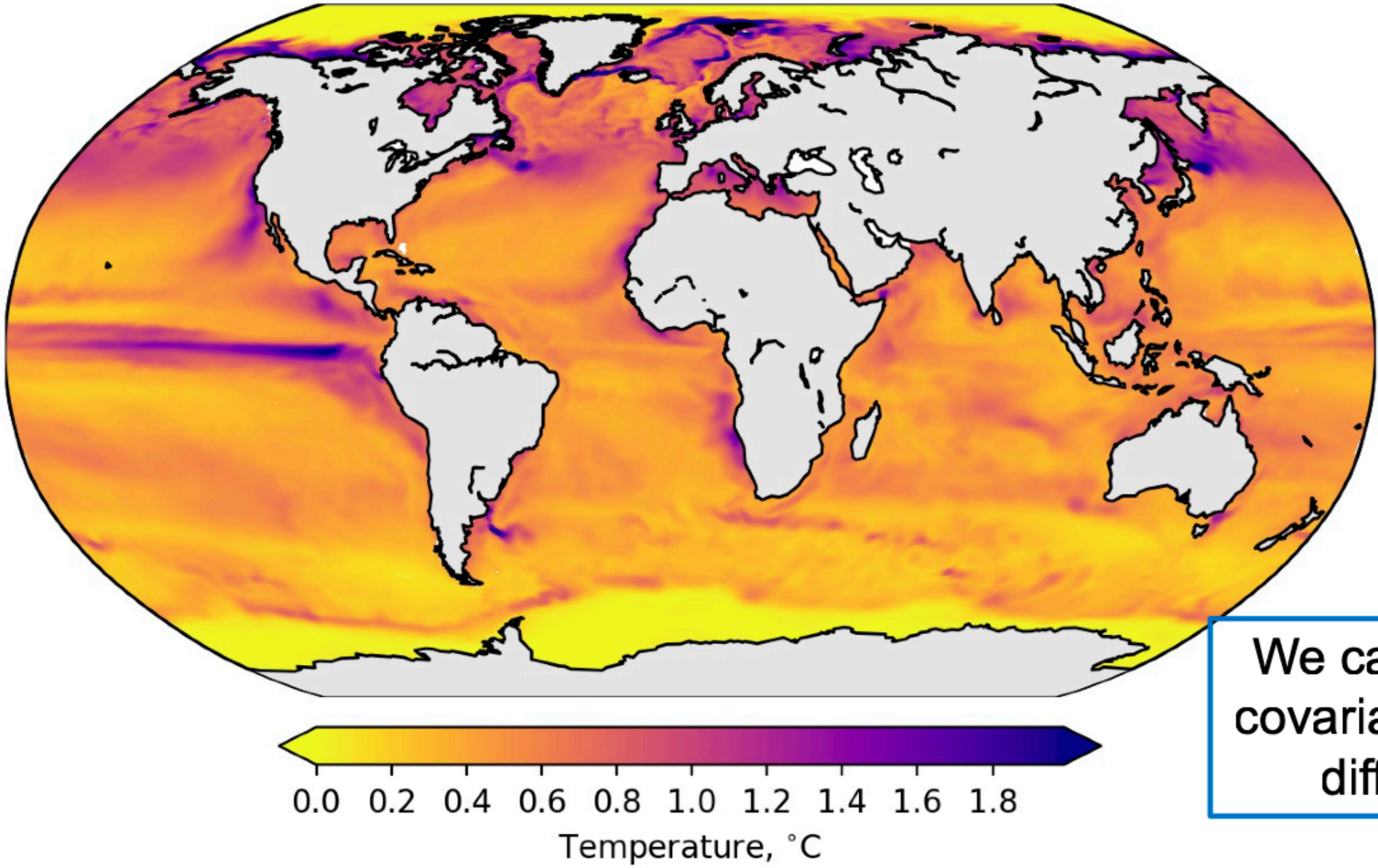
Model errors

- Representation of reality is not exact
  - Insufficient resolution
  - Incomplete equations (e.g. missing processes)
  - Inexact forcing (e.g. wind stress on ocean surface)

- Accounting for model error
  - Inflation (partly)
  - Simulate stochastic part
  - Bias estimation

Model Error Estimate

Error (uncertainty) in sea surface temperature
(estimate from model dynamics of atmosphere-ocean climate model)



We can also estimate error covariances, e.g. in between different model fields

Observations

Measure different fields - e.g. in the Ocean:

- Remote sensing (satellite, aircraft, radar)
  - e.g. surface temperature, salinity, sea surface height, ocean color, sea ice concentrations & thickness
- In situ (ships, fixed stations, buoys, drifters)
  - e.g. temperature, conductivity, pressure, currents
- Data is sparse: some fields, data gaps
- Uncertainties
  - Measurement errors
  - Representation errors:
    Model and data do not represent exactly the same (e.g. caused by finite model resolution)

Observation Error Estimates

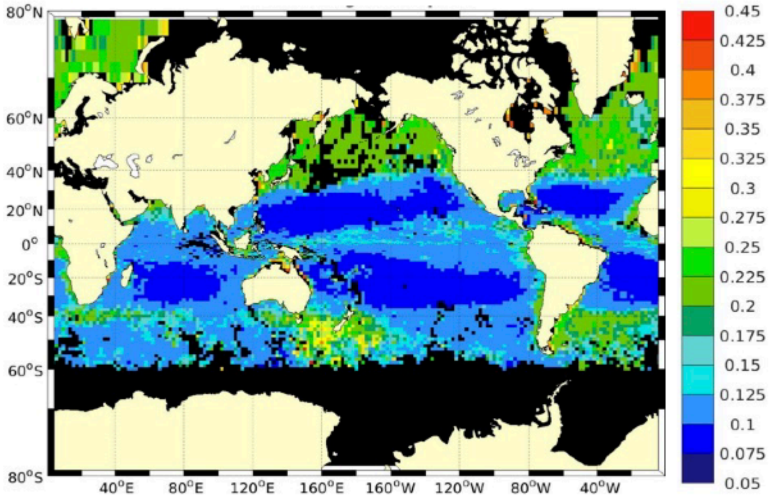
If observation errors available:

- Use as error estimate, but
- Need to account for representation errors (measurement errors are too low)

If no observation errors available:

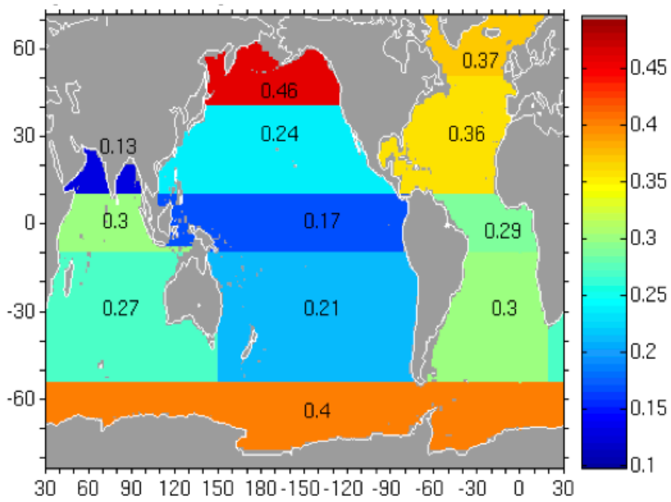
- need to estimate them
- Perhaps need to estimate also representation errors

logarithmic data errors provided with satellite chlorophyll data (OC-CCI)



Pradhan et al, JGR 2019

data errors from comparison with 2186 collocation points of in situ data (SeaWiFS)

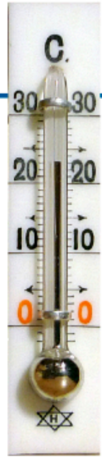


Nerger & Gregg, JMS 2007

Least Squares

and Optimization

Least Squares Approach



Consider two temperature measurements: 19°C and 21°C

Now get estimate for the true value

Measurements: y_1 = 19°C, y_2 = 21°C

State estimate: x

Quadratic mismatch :

$$J(\mathbf{x}) = \frac{1}{2}\left[(y_1 - x)^2 + (y_2 - x)^2\right]$$

Scalar value: cost

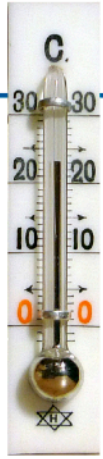
Minimum of *J(x)*:

$$\hat{x} = \frac{y_1 + y_2}{2} = 20^\circ\mathrm{C}$$

General way to get solution:
Minimize *J (optimization)*

$$dJ/dx = 0$$

Least Squares (2)



Two temperature measurements:

y_1=19°C from digital thermometer with 0.1°C scale

y_2=21°C from classical mercury thermometer; 1°C scale

Now get estimate for the true value

How do we account for the different precision of the measurements?

Adapt quadratic mismatch :

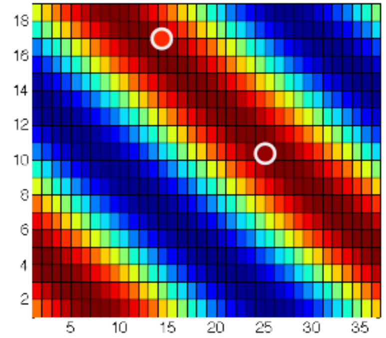
$$J(\mathbf{x}) = \frac{1}{2}\left[\left(\frac{y_1 - x}{0.1}\right)^2 + \left(\frac{y_2 - x}{1}\right)^2\right]$$

Minimum of *J(x)*: $\hat{x} = \frac{10y_1 + y_2}{11} = 19.18°\mathrm{C}$

Least Squares (3)

Now consider estimating a field

We have again 2 observations



Deviation between observation $\mathbf{y}$ and model state $\mathbf{x}$

$$\mathbf{y} - H[\mathbf{x}]$$

Quadratic mismatch :

length=648

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{y} - H[\mathbf{x}])^T \mathbf{R}^{-1}(\mathbf{y} - H[\mathbf{x}])$$

$\mathbf{R}$: weight matrix

length=2

Now find $\hat{\mathbf{x}}$ that minimizes J

- $\mathbf{x}$ is much larger than y ➜ under-constrained problem
- we need more observations or additional information ➜ previous state or model

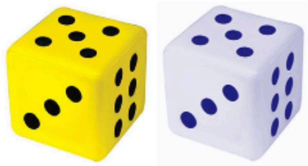
➜ Data assimilation

Some statistics

Variance, Covariance, Correlation – scalar values

Consider two populations of N scalars: $s^k, t^k, 1 \leq k \leq N$

Example:
Throwing dice



Then

Variance: $$\mathrm{var}(s) = \frac{1}{N-1}\sum_{k=1}^{N}\left(s^k - \overline{s}\right)^2$$

Covariance: $$\mathrm{cov}(s,t) = \frac{1}{N-1}\sum_{k=1}^{N}\left(s^k - \overline{s}\right)\left(t^k - \overline{t}\right)$$

Correlation: $$\mathrm{cor}(s,t) = \frac{\mathrm{cov}(s,t)}{\sigma_s \sigma_t}$$

with

$$-1 \leq \mathrm{cor}(s,t) \leq 1,\ \ \mathrm{cor}(s,s) = \mathrm{cor}(t,t) = 1$$

Vector case: Covariance and Correlation Matrices

Consider two populations of N **vectors**: $\mathbf{u}^k, \mathbf{v}^k, 1 \leq k \leq N$

Then

Covariance matrix:

$$\mathrm{cov}\left(\mathbf{u}, \mathbf{v}\right) = \frac{1}{N-1} \sum_{k=1}^{N} \left(\mathbf{u}^k - \overline{\mathbf{u}}\right) \left(\mathbf{v}^k - \overline{\mathbf{v}}\right)^T$$

If $\mathbf{u} = \mathbf{v}$: $\mathrm{cov}\left(\mathbf{v}, \mathbf{v}\right)$ is **auto-covariance matrix** of $\mathbf{v}$

$\mathrm{cov}\left(\mathbf{v}, \mathbf{v}\right)$ is symmetric and positive semi-definite

Correlation matrix:

$$\mathrm{cor}\left(\mathbf{u}, \mathbf{v}\right)_{ij} = \frac{\mathrm{cov}\left(\mathbf{u}, \mathbf{v}\right)_{ij}}{\sigma_{\mathbf{u}_i} \sigma_{\mathbf{v}_j}}$$

Each element is normalized by standard deviations

Auto-correlation matrix:

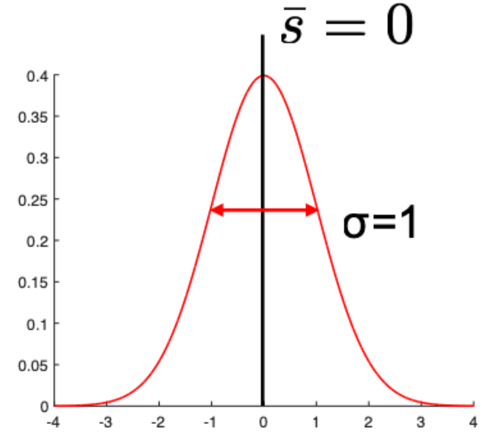
(symmetric)

$$\mathrm{cor}\left(\mathbf{v}, \mathbf{v}\right)_{ij} = \frac{\mathrm{cov}\left(\mathbf{v}, \mathbf{v}\right)_{ij}}{\sigma_{\mathbf{v}_i} \sigma_{\mathbf{v}_j}}$$

Gaussian function / Normal distribution

For scalar values (***n*=1**)

$$N(s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(s-\overline{s})^2}{2\sigma^2}\right]$$



For vectors (**dimension n**):

$$N(\mathbf{u}) = \frac{1}{\sqrt{(2\pi)^n det(\mathbf{P})}} \exp\left[-\frac{1}{2}(\mathbf{u}-\overline{\mathbf{u}})^T\mathbf{P}^{-1}(\mathbf{u}-\overline{\mathbf{u}})\right]$$

With: $\mathbf{P} = \text{cov}(\mathbf{u},\mathbf{u})$

- Gauss function is fully described by mean and variance (or mean vector and covariance matrix)
- Factor before $\exp$ is normalization (ensures that area below curve = 1)

Covariance matrices

E.g. state error covariance matrix $\mathbf{P}$

- Represent error (*variance* σ^2) and dependence of errors in different variables (*covariances*)

Example: 2 variables $\mathbf{x} = \begin{pmatrix} T \\ S \end{pmatrix}$

Covariance matrix: $\mathbf{P} = \begin{pmatrix} \sigma_T^2 & \mathrm{cov}(TS) \\ \mathrm{cov}(TS) & \sigma_S^2 \end{pmatrix}$

Now normalize variances $\sigma_T^2 \rightarrow 1$, $\sigma_S^2 \rightarrow 1$

$$\mathrm{corr}(TS) = \frac{\mathrm{cov}(TS)}{\sigma_T \; \sigma_S} \qquad -1 \leq \mathrm{corr}(TS) \leq 1$$

Correlation matrix: $\mathbf{C} = \begin{pmatrix} 1 & \mathrm{corr}(TS) \\ \mathrm{corr}(TS) & 1 \end{pmatrix}$

$\mathrm{corr}(TS)$ correlation: How much changes S when changing T

BLUE

Statistical Estimate

Consider that our estimates (as realization of random values) have errors:

$$y_1 = y^t + e_1 \qquad y_2 = y^t + e_2$$

y^t is the true value

Statisical assumptions on the errors $e_i, \quad i = 1, 2$

$E(e_i) = \overline{e}_i = 0$ ➜ Measurements are unbiased

$\mathrm{var}(e_i) = \sigma_i^2$ ➜ Errors are known (but random)

$\mathrm{cov}(e_1, e_2) = 0$ ➜ Errors are independent

Now find optimal estimator $\hat{x}$ that is

- Linear $\hat{x} = \alpha_1 y_1 + \alpha_2 y_2$
- Unbiased $E(\hat{x}) = y^t$
- of minimal variance $\mathrm{var}(\hat{x})$ minimum

Yields the best linear unbiased estimate (BLUE)

Statistical Estimate (2)

Now combine $y_1 = y^t + e_1 \qquad y_2 = y^t + e_2$

with $\hat{x} = \alpha_1 y_1 + \alpha_2 y_2$

We get $\hat{x} = (\alpha_1 + \alpha_2) y^t + \alpha_1 e_1 + \alpha_2 e_2$

Expected value:

$$E(\hat{x}) = (\alpha_1 + \alpha_2) y^t + \alpha_1 \underbrace{E(e_1)}_{0} + \alpha_2 \underbrace{E(e_2)}_{0}$$

We require unbiasedness: $E(\hat{x}) = y^t$

This implies: $\boxed{\alpha_1 + \alpha_2 = 1}$

Statistical Estimate (3)

It is

$$\hat{x} = (\alpha_1 + \alpha_2)y^t + \alpha_1 e_1 + \alpha_2 e_2$$
$$= y^t + \alpha_1 e_1 + \alpha_2 e_2$$

Variance:

$$\mathrm{var}(\hat{x}) = E[(\hat{x} - y^t)^2]$$
$$= E[(\alpha_1 e_1 + \alpha_2 e_2)^2]$$
$$= \alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2$$
$$= \alpha_1^2 \sigma_1^2 + (1 - \alpha_1)^2 \sigma_2^2$$

Minimum variance:

$$\frac{\partial}{\partial \alpha_1} \mathrm{var}(\hat{x}) = 0$$

$$0 = 2\alpha_1 \sigma_1^2 + 2\alpha_1 \sigma_2^2 - 2\sigma_2^2$$

$$\boxed{\alpha_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}}$$

Statistical Estimate (4)

Best linear unbiased estimate (BLUE)

$$\hat{x} = \alpha_1 y_1 + \alpha_2 y_2$$

With: $$\alpha_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \qquad \alpha_2 = 1 - \alpha_1 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

Thus: $$\hat{x} = \frac{1}{\sigma_1^2 + \sigma_2^2}\left(\sigma_2^2\ y_1 + \sigma_1^2 y_2\right)$$

Equivalent: $$\hat{x} = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\left(\frac{1}{\sigma_1^2}\ y_1 + \frac{1}{\sigma_2^2} y_2\right)$$

Error in BLUE estimate:

Inverse variance:
(accuracy)

$$\left[\mathrm{var}(\hat{x})\right]^{-1} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

Statistical Estimate (5)

Best linear unbiased estimate (BLUE)

$$\hat{x} = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \left(\frac{1}{\sigma_1^2}\ y_1 + \frac{1}{\sigma_2^2} y_2 \right)$$

Also the solution when minimizing (i.e. optimizing):

$$J(x) = \frac{1}{2} \left(\frac{(x - y_1)^2}{\sigma_1^2} + \frac{(x - y_2)^2}{\sigma_2^2} \right)$$

Statistical reason for temperature example:

$$J(\mathbf{x}) = \frac{1}{2} \left[\left(\frac{y_1 - x}{0.1} \right)^2 + \left(\frac{y_2 - x}{1} \right)^2 \right]$$

Further: $$[\mathrm{var}(\hat{x})]^{-1} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} = J''(\hat{x})$$

Basic Data Assimilation

Data Assimilation – Model and Observations

Two components:

1. State: $\mathbf{x} \in \mathbb{R}^n$ (contains different model variables)

   Dynamical model

   $$\mathbf{x}_i = M_{i-1,i}\left[\mathbf{x}_{i-1}\right]$$

2. Obervations: $\mathbf{y} \in \mathbb{R}^m$ (contains different observed fields)

   Observation equation (relation of observation to state $\mathbf{x}$):

   $$\mathbf{y} = H\left[\mathbf{x}\right]$$

Dimensions:

state	n: $10^6 - 10^9$
observations	m: $10^4 - 10^7$

Direct Insertion

- Part of the state vector is observed
- Now replace observed part of state vector by observations
  - Observed points $\mathbf{x}^a_{k,j'(i)} = \mathbf{y}_{k,i}$
  - Unobserved points $\mathbf{x}^a_{k,j} = \mathbf{x}^f_{k,j}$

  (*j'(i)* is the index of the state vector element corresponding to observation *i)*

- This assumes that model error ≫ observation error
- Problems
  - Inconsistency between updated and preserved parts of state vector
  - Leads to adjustment processes that can degrade model results (e.g. geostrophic adjustments)

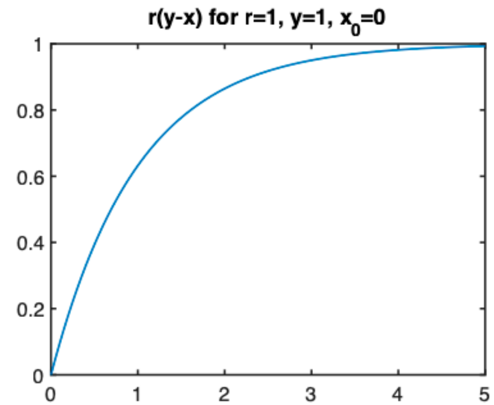
Nudging

- Part of the state vector is observed
- Now replace observed part of state vector by observations
  - Observed points $\mathbf{x}^a_{k,j'(i)} = \mathbf{x}^f_{k,j'(i)} + r_i \left(\mathbf{y}_{k,i} - \mathbf{x}^f_{k,j'(i)}\right)$
  - Unobserved points $\mathbf{x}^a_{k,j} = \mathbf{x}^f_{k,j}$

  relaxation to observation

  (j'(i) is the index of the state vector element corresponding to observation *i)*

- 1/r is relaxation time scale
- Effect is exponential decay



Nudging (2)

- Nudging was common for DA ~30 years ago
  - Superseded by Optimal Interpolation
- Some applications today
  - SST nudging: constrain surface heat flux
  - Nudging towards climatology: prevent drift of model
- Problems
  - introduces small imbalance (initially higher influence)
  - reduces model variability
- practically, nudging modifies the model equations

Optimal Interpolation

BLUE as Statistical Estimate with Model

Best linear unbiased estimate for

Model prediction: $x^b = x^t + e_b$ with: $\mathrm{var}(e_b) = \sigma_b^2$

Observation: $y = x^t + e_o$ with: $\mathrm{var}(e_o) = \sigma_o^2$

Solution: $$\hat{x} = \frac{1}{\sigma_b^2 + \sigma_o^2}\left(\sigma_o^2 x_b + \sigma_b^2 y\right)$$

Equivalent to $$\hat{x} = \underbrace{x_b}_{\text{background}} + \underbrace{\frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2}}_{\text{gain}} \underbrace{(y - x_b)}_{\text{innovation}}$$

background + gain x innovation

Equivalent to minimizing: $$J(x) = \frac{1}{2}\frac{(x - x_b)^2}{\sigma_b^2} + \frac{1}{2}\frac{(x - y)^2}{\sigma_o^2}$$

Error in estimate

$$var(\hat{x}) = \left(\frac{1}{\sigma_o^2} + \frac{1}{\sigma_b^2}\right)^{-1}$$

$$= \left(1 - \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2}\right)\sigma_b^2$$

BLUE in vector form – Optimal Interpolation (OI)

A general time independent linear data assimilation problem

Cost function:

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) + (\mathbf{y} - H[\mathbf{x}])^T \mathbf{R}^{-1} (\mathbf{y} - H[\mathbf{x}])$$

With covariance matrices

$\mathbf{B}$ background error covariance matrix

$\mathbf{R}$ observation error covariance matrix

Minimum of *J*

$$\mathbf{x}^a = \mathbf{x}^b + \underbrace{(\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1}}_{\text{gain matrix}} \underbrace{(\mathbf{y} - \mathbf{H}\mathbf{x}^b)}_{\text{innovation vector}}$$

Equivalent (Sherman-Morrison-Woodbury identity)

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}^b)$$

Practical difficulties

Analysis equation

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}^b)$$

Matrices are huge:

$\mathbf{B}$: *n x n* with *n* between 10^6 and 10^9

$\mathbf{R}$: m x m with m between 10^5 and 10^7

➜ Not feasible to store **B** or to compute $\mathbf{B}^{-1}$

➜ We cannot compute the gain explicitly

➜ **Need to reduce size of problem**

Note:

- Dimensions were smaller ~25 years when OI was common
- But, computers were also much smaller than today

A final note on OI

The method was common ~25 years ago

- It cannot handle nonlinear observation operators
- The local selection of observations sometimes induced discontinuities in $\mathbf{x}^a$

OI was superseded by

  - Variational assimilation
  - Ensemble Kalman filters

- There are still applications for small models (linear or weakly nonlinear)
- Today, some applications use **ensemble OI**, i.e. OI with $\mathbf{B}$ represented by ensemble of model state realizations

Estimation Problem

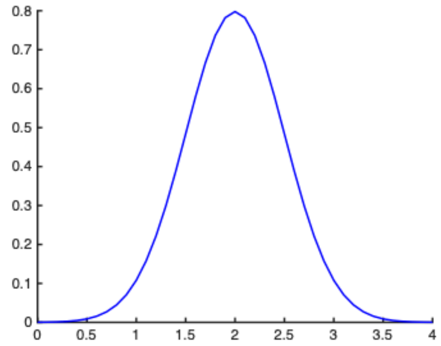
Data Assimilation – An Estimation Problem

Probability density of $\mathbf{x}$: $p\left(\mathbf{x}_i\right)$

Probability density of $\mathbf{y}$: $p\left(\mathbf{y}_i\right)$

Likelihood of $\mathbf{y}$ given $\mathbf{x}$: $p\left(\mathbf{y}_i | \mathbf{x}_i\right)$

(conditional probability)



Area = 1.0

$p \geq 0$

Bayes law: Probability density of $\mathbf{x}$ given $\mathbf{y}$

Likelihood of observations

Prior distribution

$$p\left(\mathbf{x}_i | \mathbf{y}_i\right) = \frac{p\left(\mathbf{y}_i | \mathbf{x}_i\right) p\left(\mathbf{x}_i\right)}{p\left(\mathbf{y}_i\right)}$$

Posterior probability distribution

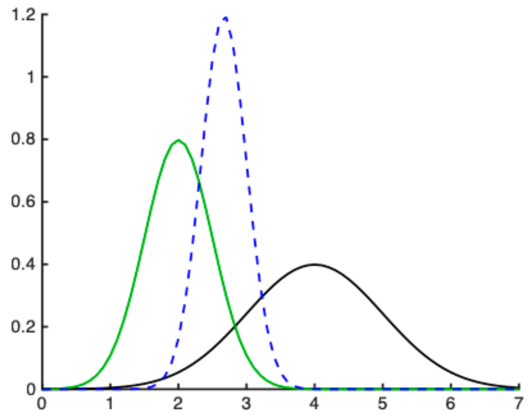
Marginal distribution of observations

Estimate assuming Gaussian error distributions

Assume Gaussian distributions:

$$\mathcal{N}\left(\mu,\sigma^2\right) = a\ e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

(fully described by mean and variance)



Observations: $\mathcal{N}(\mathbf{y},\mathbf{R})$

State: $\mathcal{N}(\mathbf{x},\mathbf{P})$

Estimation problem leads to same assimilation result in case of Gaussian errors

Posterior state distribution

$$p(\mathbf{x}_i|\mathbf{Y}_i) \sim ae^{-J(\mathbf{x})}$$

... but it also yields a solution for non-Gaussian/non-linear cases

With

$$J(\mathbf{x}) = (\mathbf{x}-\mathbf{x}^b)^T\mathbf{P}^{-1}(\mathbf{x}-\mathbf{x}^b) + (\mathbf{y}-H\left[\mathbf{x}\right])^T\mathbf{R}^{-1}(\mathbf{y}-H\left[\mathbf{x}\right])$$

Hands-On Tutorial 1

Nudging and Optimal Interpolation

Hands-On Tutorial 1

Some experiments with basic data assimilation:

- Nudging
- Optimal Interpolation

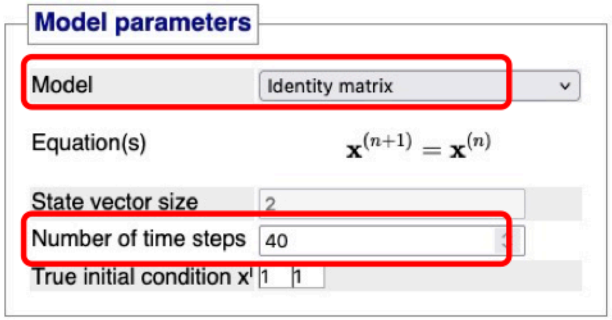
Use the online-tutorial in the browser

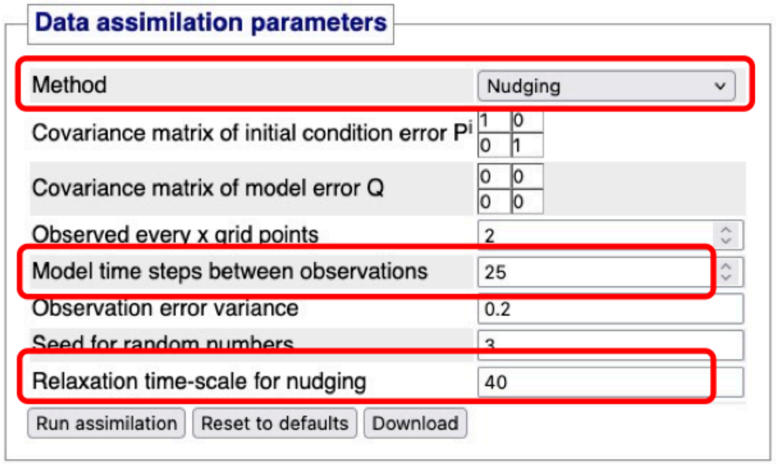
http://pdaf.awi.de/DA_demo/

Hands-On 1: Nudging

Instructions:

- Set ‘Model’ to ‘***Identity matrix***’, i.e. $x_{i+1} = x_i$ or ***‘Oscillation’***
- Set ‘Number of time steps’ = 40 for ‘***Identity matrix***’;
  =200 for ***‘Oscillation’***
- Set ‘Method’ to ***‘Nudging’***
- Keep other values to the default (reload is necessary)





Note: You can select the element of the state vector in the plots:

Hands-On 1: Nudging

Instructions:

In these experiments compare the state estimate from the assimilation (red) with

a) The true state (orange; this is what the DA tries to estimate)

b) The observations (green circles; they are the truth plus noise)

What are the effect of different settings for estimating the true state?

1) Model: *Identity matrix* – Basic behavior of nudging:

a. What is the effect of nudging in case of a single observation?
(Set model time steps between observations = 25)
b. What is the effect of nudging in case of frequent observations?
(Set model time step between observations = 1)
c. What happens in the cases above when you vary the relaxation time scale?
(e.g. test 5, 10, and further in steps of 10 up to 100)
d. What happens for the very small relaxation time scale = 1?

2) Model: *Oscillation*

a. How does the assimilation result change when you vary the relaxation time-scale (5,10,20,…,100)?
b. There is no optimal relaxation time-scale – why?
c. What is the effect of varying 'Model time steps between observations' (range 1-10)?

Hands-On 1: Optimal Interpolation (OI)

Instructions:

- Set model to ***'Identity matrix'*** or '***Oscillation***'
- Set 'Method' to ***'Optimal Interpolation'***

3) Optimal interpolation vs. Nudging

a. For ‚*Identity matrix*', how does the assimilation effect of OI differ from that of nudging?
b. For ‚*Oscillation*', how does the assimilation effect of OI differ from that of nudging?

4) Effect of model dynamics:

a. With the default settings how does the effect of the OI assimilation on the variable x_2 of the model "Identity matrix" and "Oscillation" differ?

5) Influence of correlations:

a. For the model "Identity matrix". What is the effect of Optimal Interpolation on x_1 and x_2 if the non-diagonal values of „Covariance matrix of initial condition error Pi" are set to 0.9?