Aspects of the practical application of ensemble-based Kalman filters

Lars Nerger

Alfred Wegener Institute for Polar and Marine Research
Bremerhaven, Germany

and

Bremen Supercomputing Competence Center BremHLR

Lars.Nerger@awi.de

BremHLR
Kompetenzzentrum für Höchstleistungsrechnen Bremen

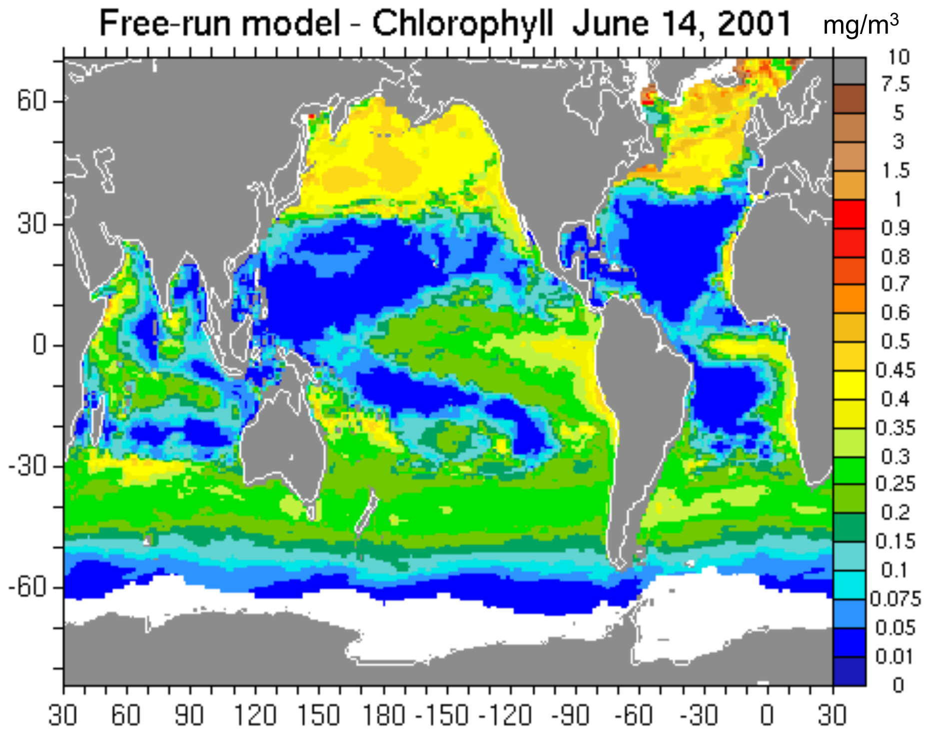
AWI

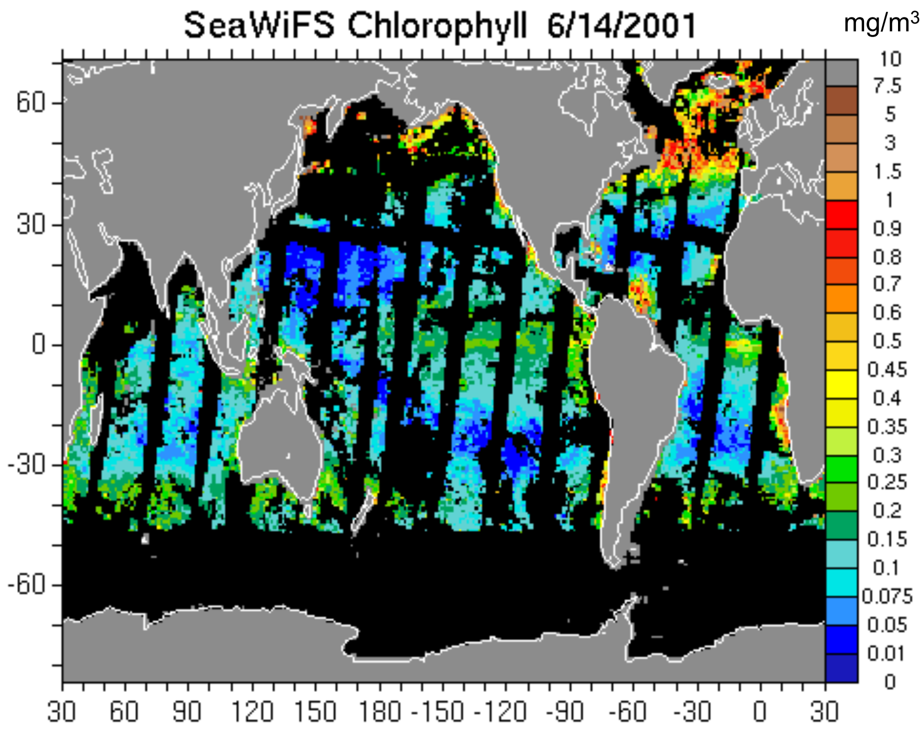
Overview

- Ensemble generation
- Localization
- Covariance inflation
- Observations and their errors
- Model errors
- Bias correction
- Validation data

Data Assimilation - in short

System Information: Chlorophyll in the ocean





Information: Model

- Generally correct, but has errors
- all fields, fluxes, …

Information: Observation

- Generally correct, but has errors
- sparse information
  (only surface, data gaps, one field)

Combine both sources of information by data assimilation

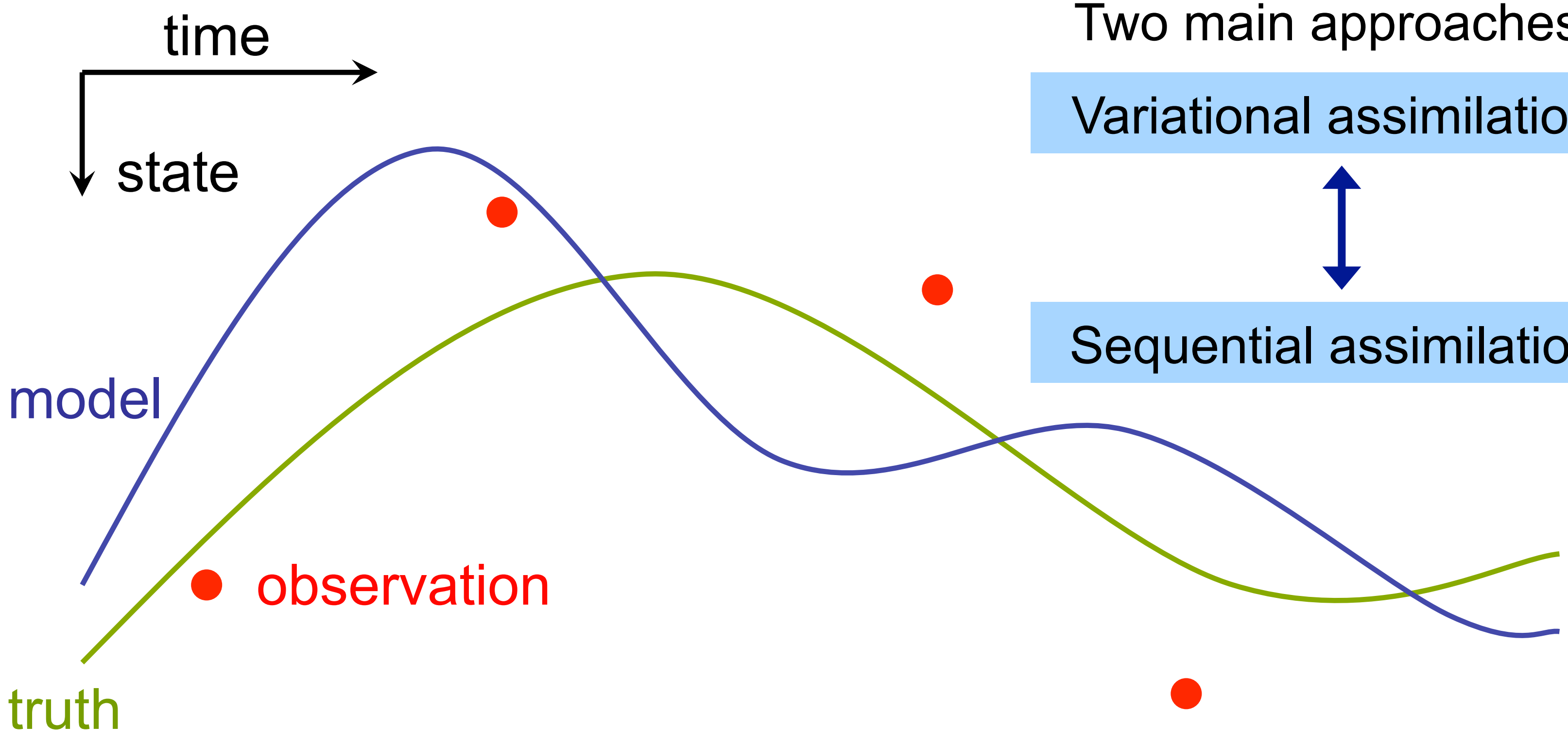
Data Assimilation

- Optimal estimation of system state:
  - initial conditions (for weather/ocean forecasts, …)
  - trajectory (temperature, concentrations, …)
  - parameters (growth of phytoplankton, …)
  - fluxes (heat, primary production, …)
  - boundary conditions and 'forcing' (wind stress, …)

- Characteristics of system:
  - high-dimensional numerical model - $\mathcal{O}(10^7)$
  - sparse observations
  - non-linear

Data Assimilation

Consider some physical system (ocean, atmosphere,…)



Two main approaches:

Variational assimilation

↕

Sequential assimilation

Optimal estimate basically by least-squares fitting

Zoo of ensemble-based/error-subspace Kalman filters

- A little "zoo" (not complete):

EnKF(94/98)

SEIK

EAKF ETKF EnSQRTKF EnKF(2003) EnKF(2004)

SEEK RRSQRT ROEK

MLEF SPKF ESSE

(Properties and differences are hardly understood)

Issues of the practical application

- No filter works without tuning
  - ⇨ Covariance inflation (forgetting factor)
  - ⇨ Localization
- Other issues
  - ⇨ Optimal initialization unknown (is it important?)
  - ⇨ Ensemble integration still costly
  - ⇨ Simulating model error
  - ⇨ Bias (model and observations)
  - ⇨ Observation errors are often unknown
  - ⇨ Nonlinearity
  - ⇨ Non-Gaussian fields or observations
  - ⇨ …

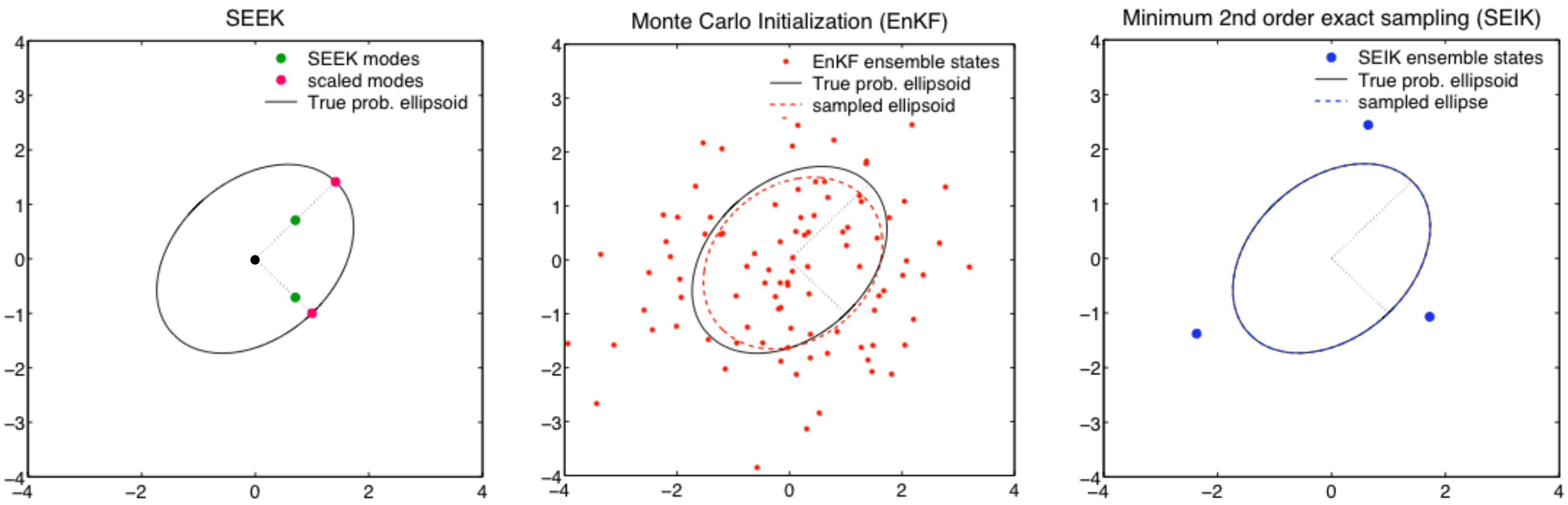
Ensemble generation

What is the “right” ensemble?

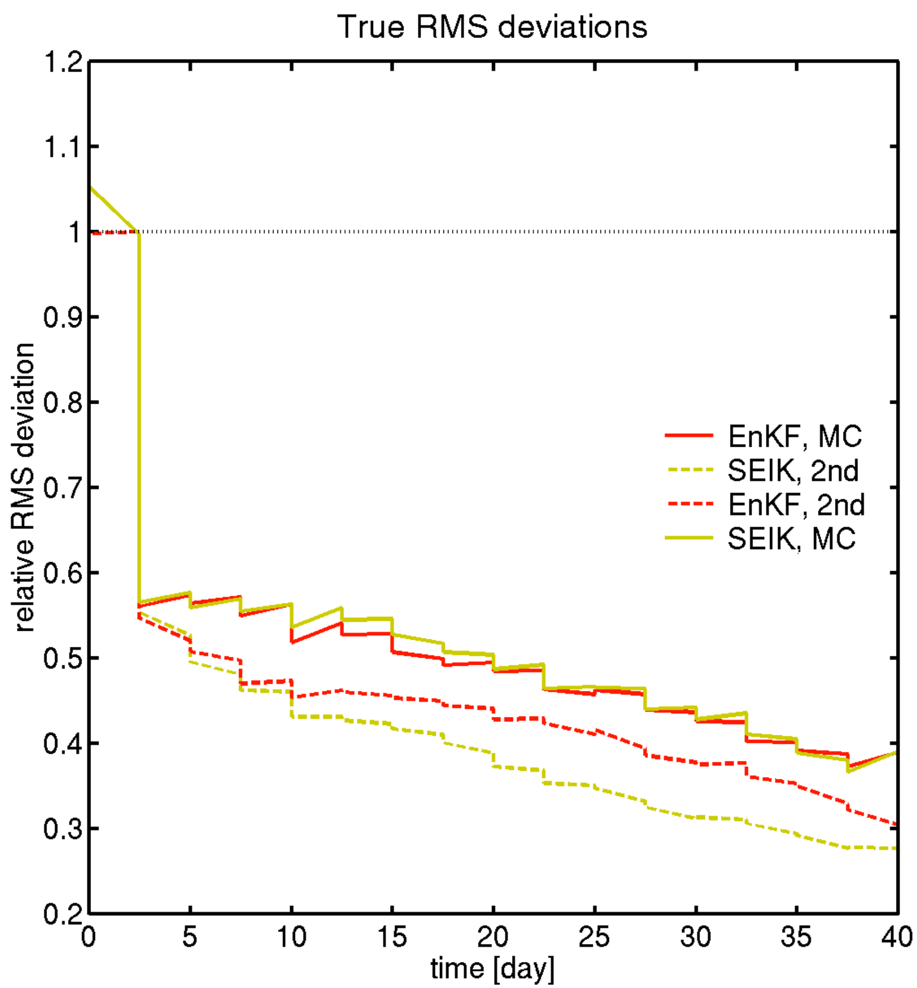
- Ensemble represents
  - ⇨ state estimate and error covariance matrix
  - ⇨ uncertainty of (initial) state estimate
  - ⇨ correlations between observed and unobserved variables
- Methods (just a selection)
  - ⇨ Deviations between model and observations (not all variables/locations observed)
  - ⇨ Variability from long model integration (self-consistent; correct timing required; related to eigenvalues)
    - ⇨ random drawing vs. SVD-based selection
  - ⇨ Set of short-term model integrations
  - ⇨ “Breeding”

Sampling Example

$$\mathbf{P}_t = \begin{pmatrix} 3.0 & 1.0 & 0.0 \\ 1.0 & 3.0 & 0.0 \\ 0.0 & 0.0 & 0.01 \end{pmatrix}; \quad \mathbf{x}_t = \begin{pmatrix} 0.0 \\ 0.0 \end{pmatrix}$$



3D Box - interchanged intializations



Ensemble size=10

Covariance matrix **P** from long model simulation

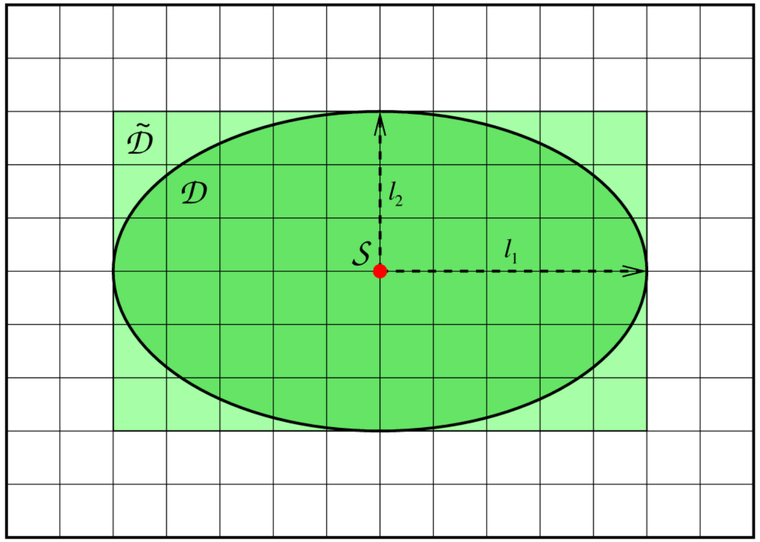
MC: random sampling of **P**

2nd: sample low-rank approximation of **P**

Localization

Domain localization - Local SEIK filter

- Analysis:
  - Update small regions (e.g. single vertical columns)
  - Consider only observations within cut-off distance
    - ➢ neglects long-range correlations



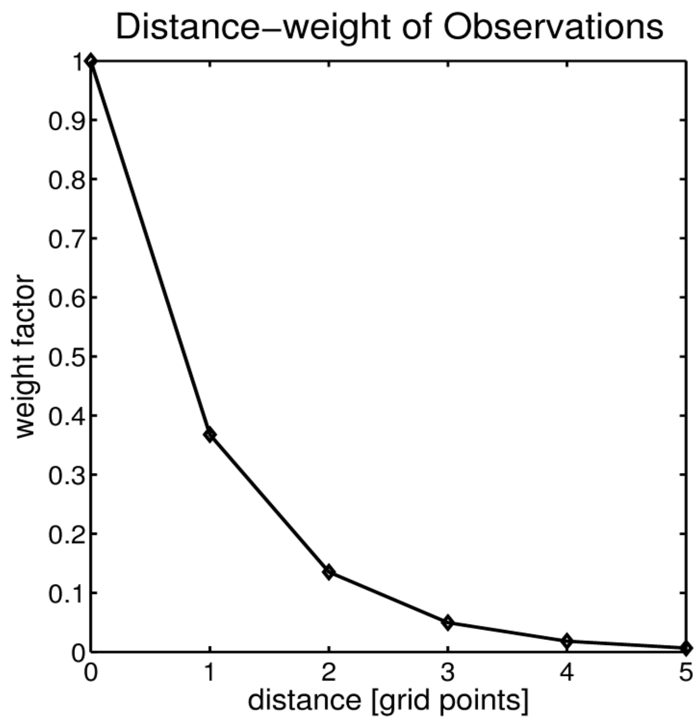
- Re-Initialization:
  - Transform local ensemble
  - Use same transformation matrix in each local domain

Nerger, L., S. Danilov, W. Hiller, and J. Schröter. Ocean Dynamics 56 (2006) 634

Local SEIK filter II – Observation localization

Localizing weight

- reduce weight for remote observations by increasing variance estimates
- use e.g. exponential decrease or polynomial representing correlation function of compact support
- similar, sometimes equivalent, to *covariance localization* used in other ensemble-based KFs



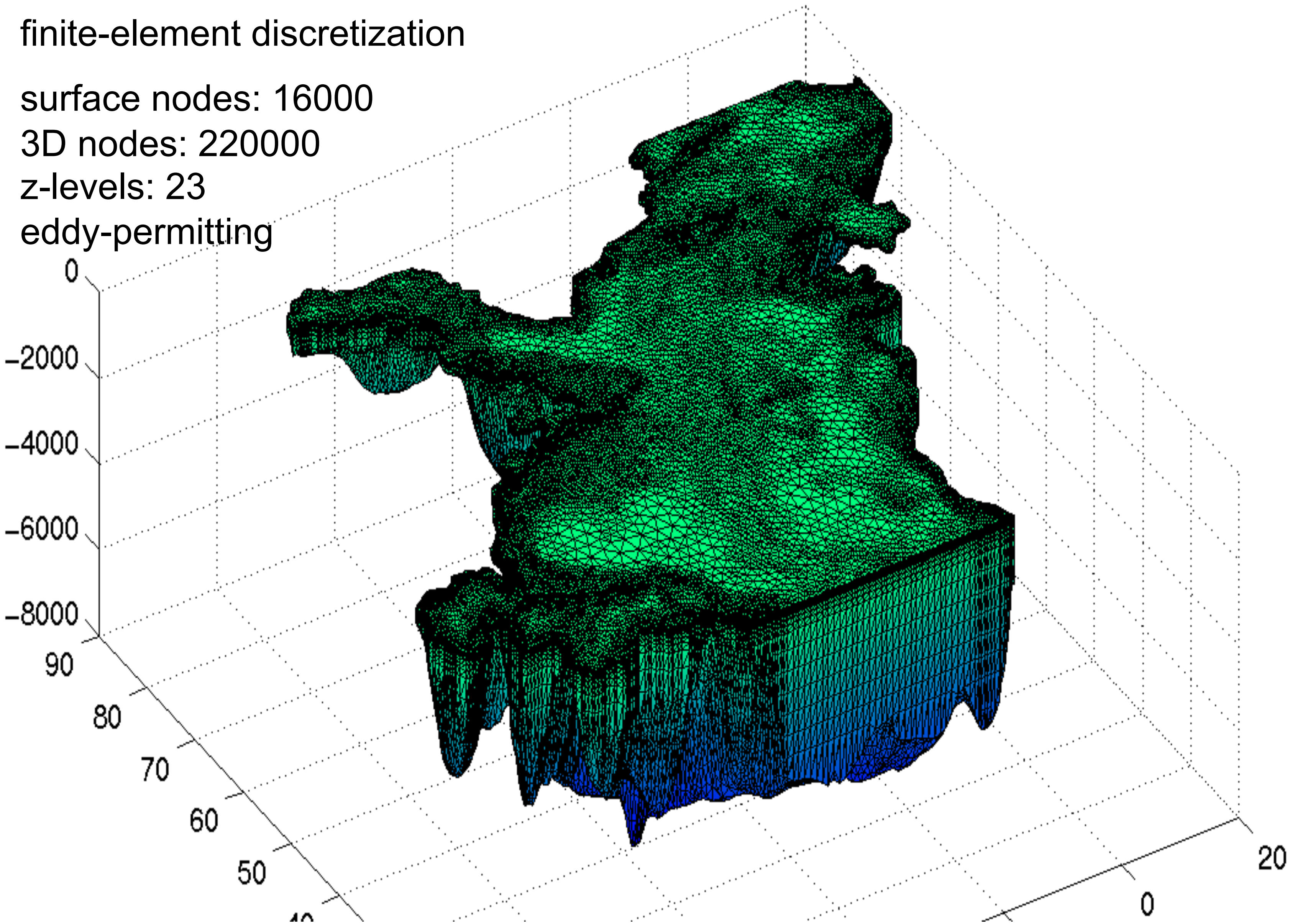
Example:

Assimilation of pseudo sea surface height observations in the North Atlantic (twin experiment)

FEOM – Mesh for North Atlantic

finite-element discretization

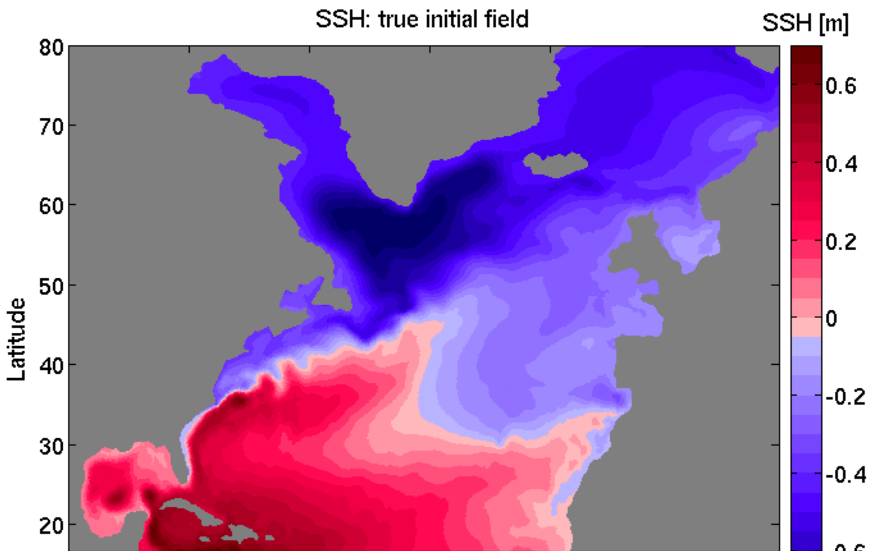
surface nodes: 16000
3D nodes: 220000
z-levels: 23
eddy-permitting

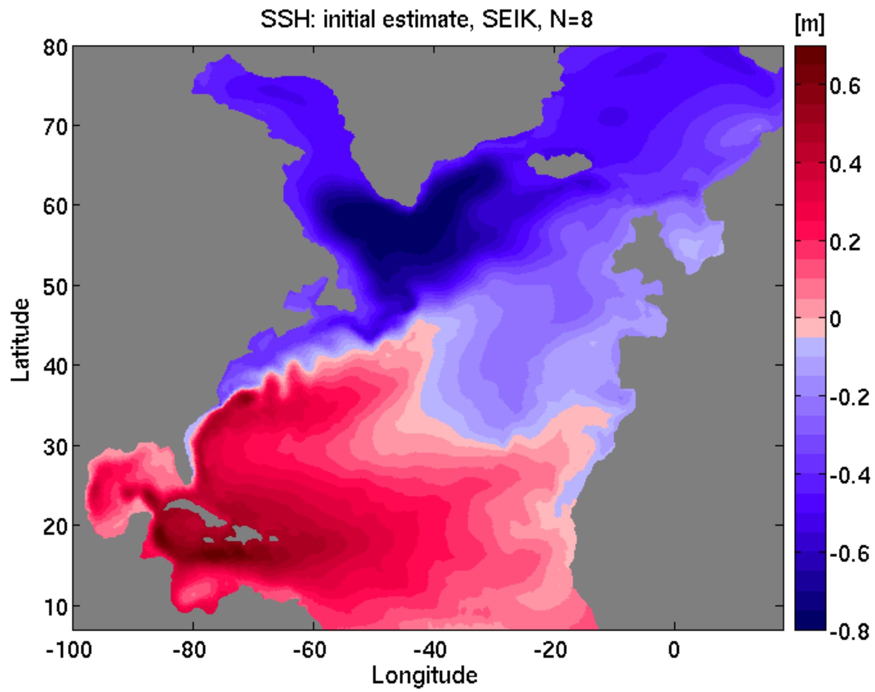


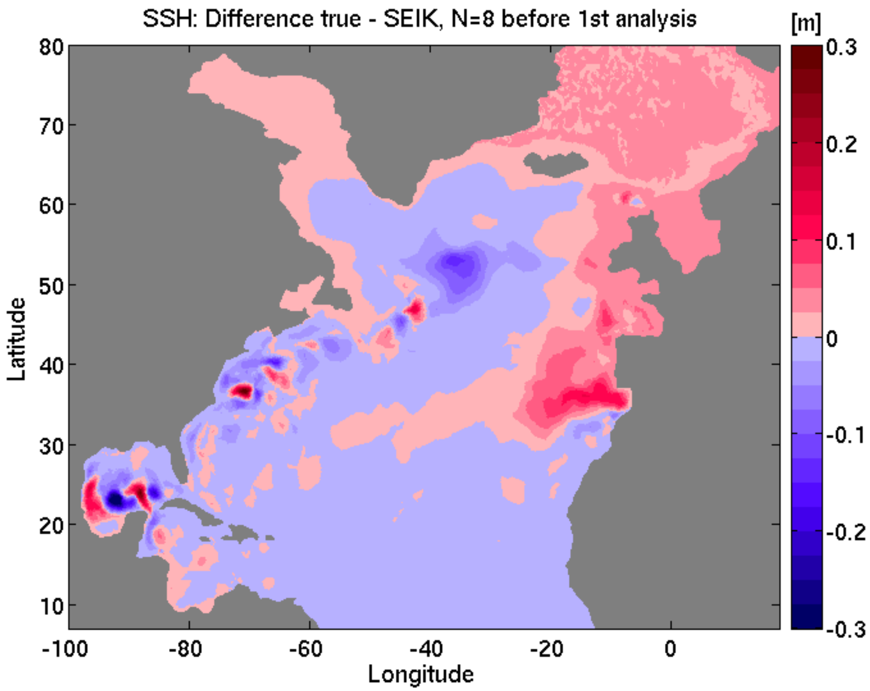
Configuration of twin experiments

- Generate true state trajectory for 12/1992 - 3/1993
- Assimilate synthetic observations of sea surface height (generated by adding uncorrelated Gaussian noise with std. deviation 5cm to true state)
- Covariance matrix estimated from variability of 9-year model trajectory (1991-1999) initialized from climatology
- Initial state estimate from perpetual 1990 model spin-up
- Monthly analysis updates (at initial time and after each month of model integration)
- No model error; forgetting factor 0.8 for both filters

Modeled Sea Surface Height (Dec. 1992)



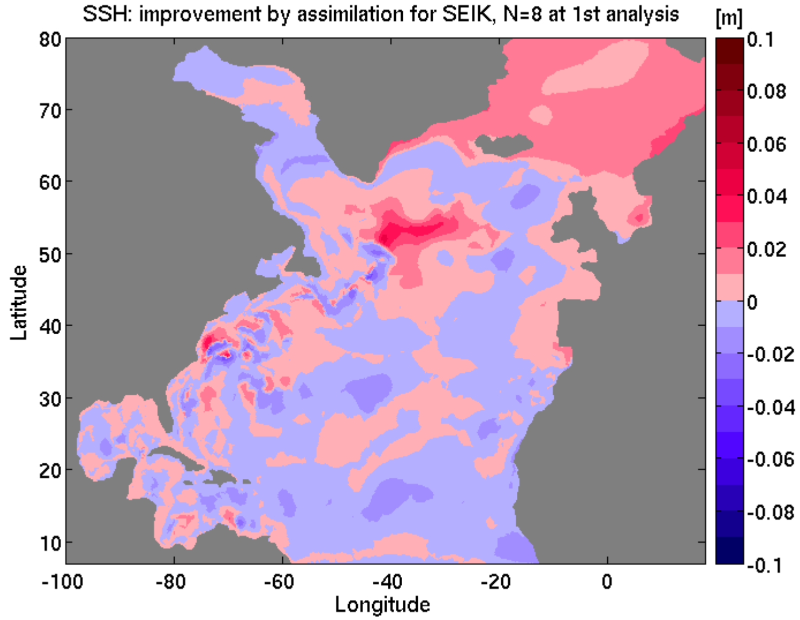




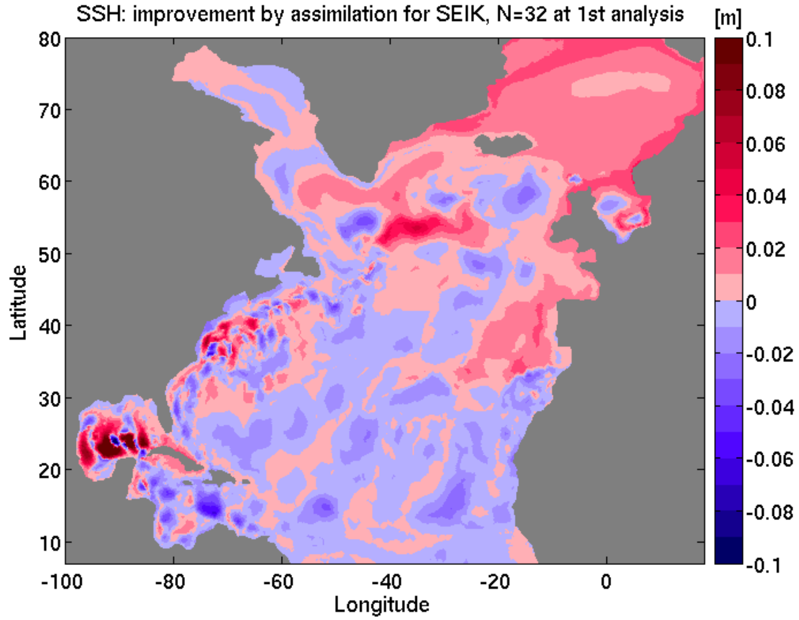
- large-scale deviations of small amplitude
- small-scale deviations up to 40 cm

Improvement of Sea Surface Height (Dec. 1992)

N=8



N=32

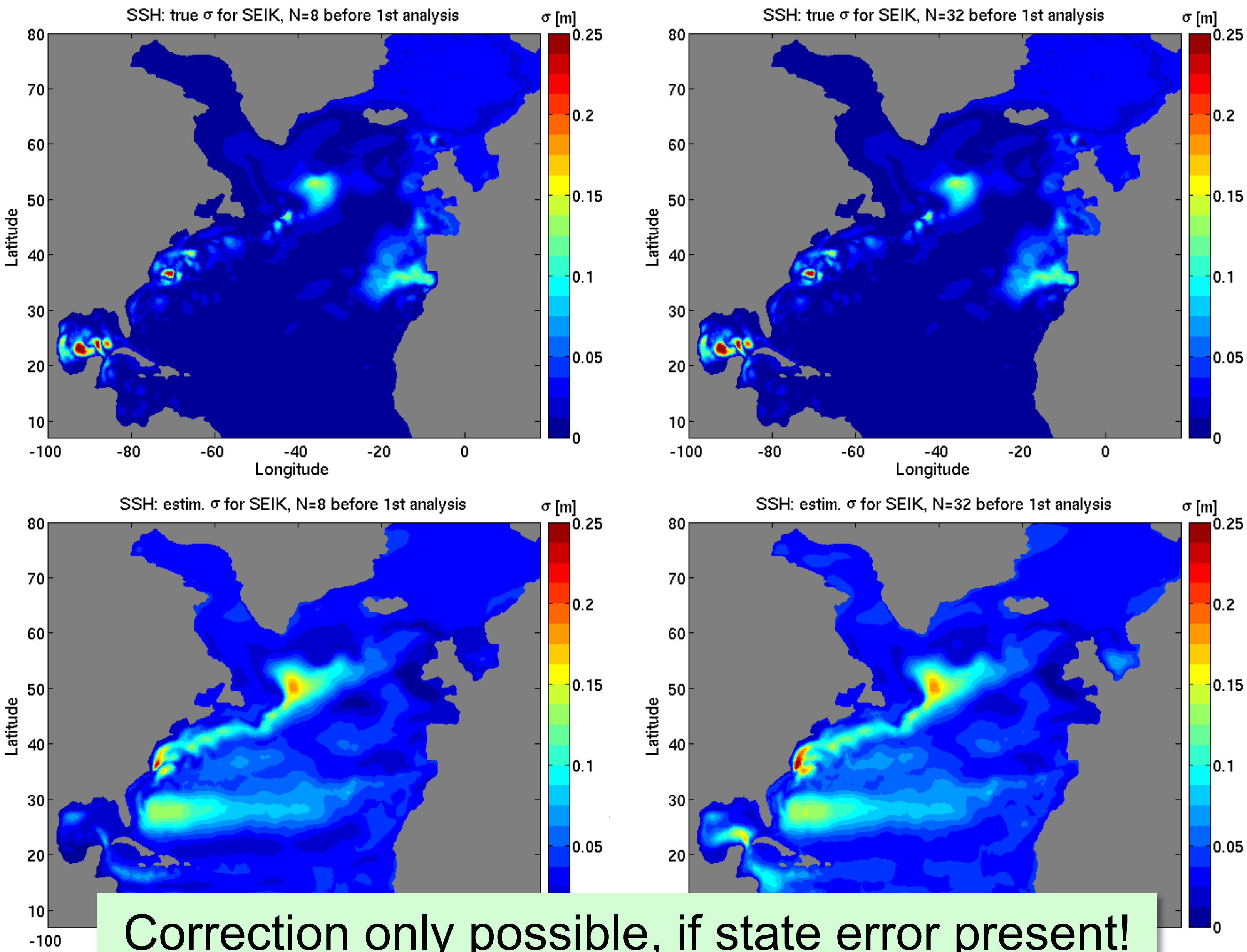


- Improvement: red - deterioration: blue

⇒ For N=8 rather coarse-scale corrections

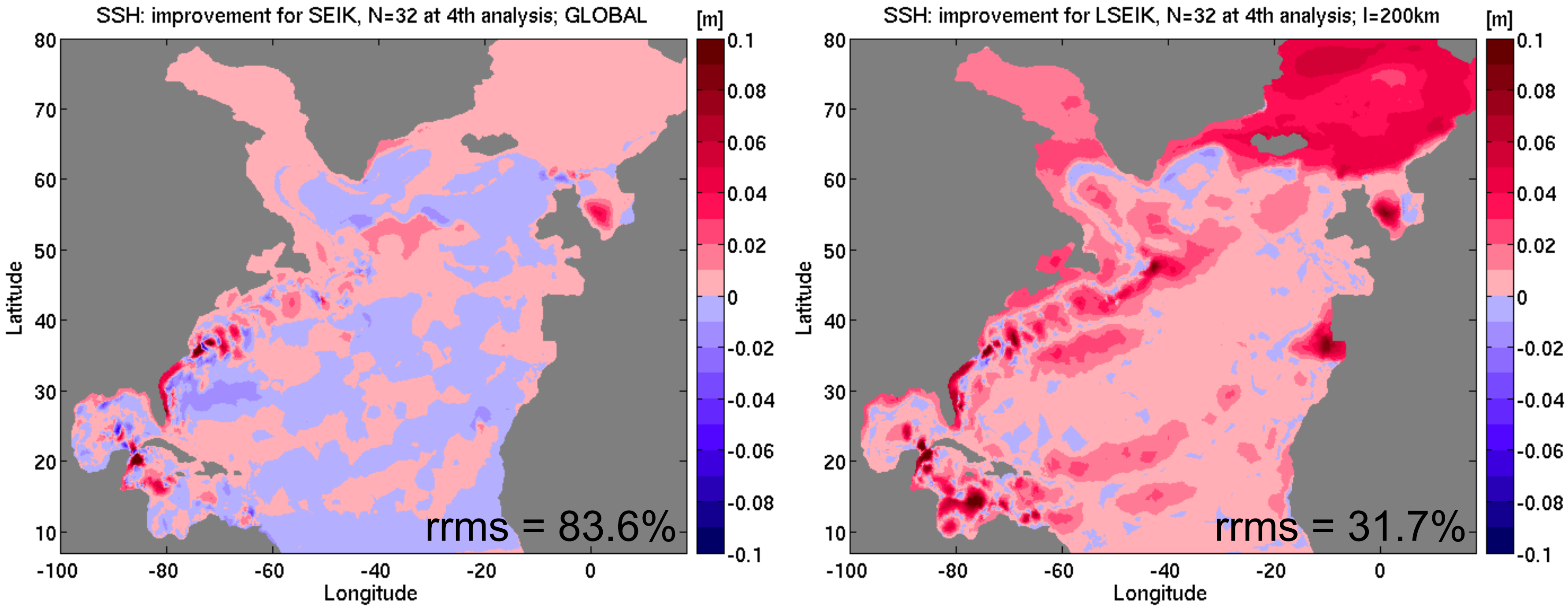
⇒ Increased ensemble size adds finer scales (systematically)

True and estimated errors (Dec. 1992)



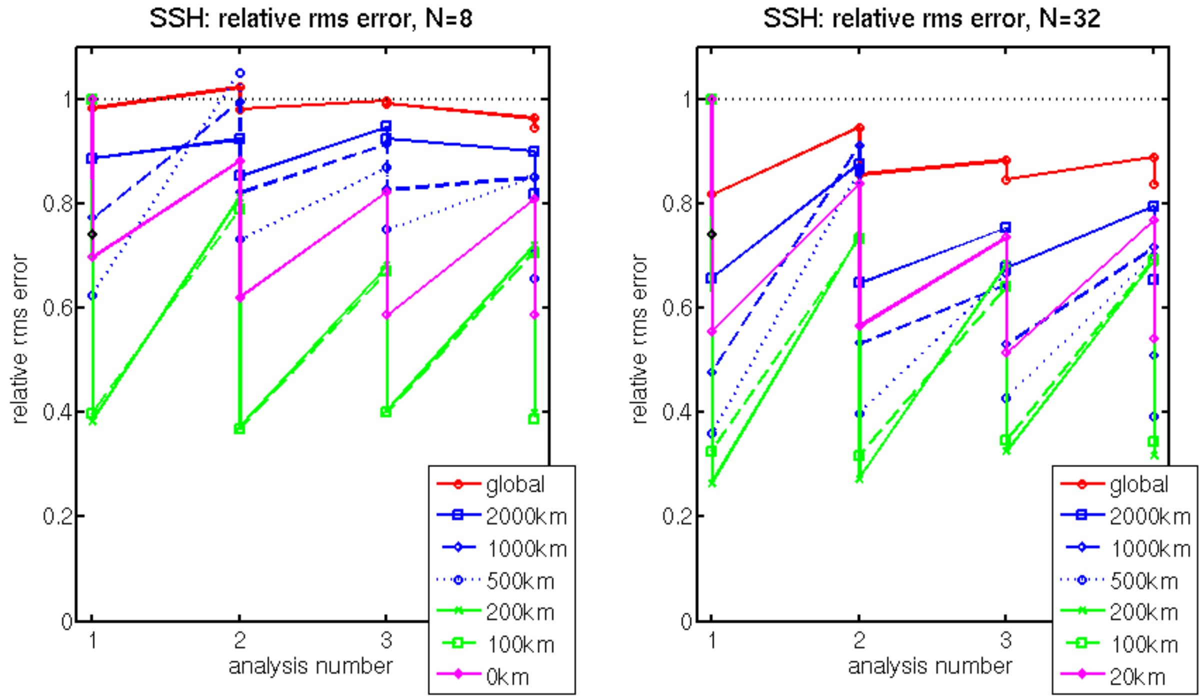
Correction only possible, if state error present!

Global vs. Local SEIK, N=32 (Mar. 1993)



- Improvement regions of global SEIK also improved by local SEIK
- localization provides improvements in regions not improved by global SEIK
- regions with error increase diminished for local SEIK

Relative rms errors for SSH



- global filter: significant improvement for larger ensemble
- global filter with N=100: relative rms error 0.74
- localization strongly improves estimate
    - larger error-reduction at each analysis update
    - but: stronger error increase during forecast
- very small radius results in over-fitting to noise

Covariance inflation

Covariance inflation

- True variance is always underestimated
  - finite ensemble size
  - sampling errors (unknown structure of P)
  - model errors

  ➜ can lead to filter divergence

- Simple remedy

  ➜ Increase error estimate before analysis

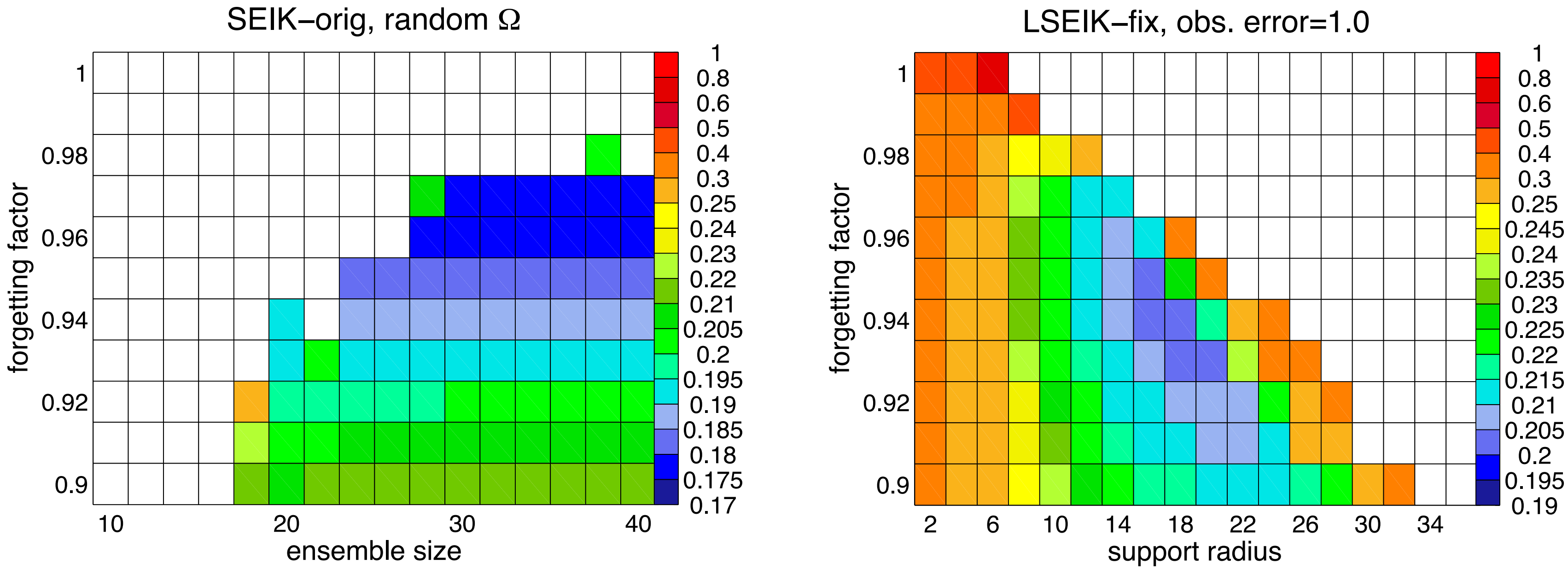
- Possibilities
  - Multiply covariance matrix by a factor (inflation factor, 1/forgetting factor)
  - Additive error (e.g. on diagonal)

Impact of inflation on stability & performance

Experiments with Lorenz96 model

Global filter

Localized, ensemble size 10



- Increased stability with stronger inflation (smaller forgetting factor)
- Optimal choice for inflation factor

Observations and their errors

Real observations

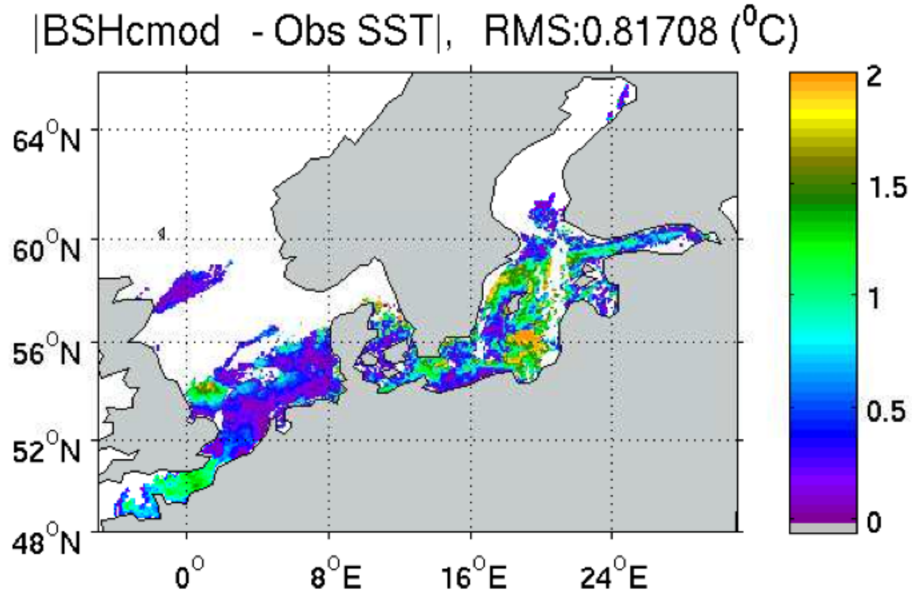
- They are not ideal
  - Incomplete (space, time)
  - Errors only estimated
  - Errors can be correlated
  - Can be biased

➜ Usual way of handling: pragmatism

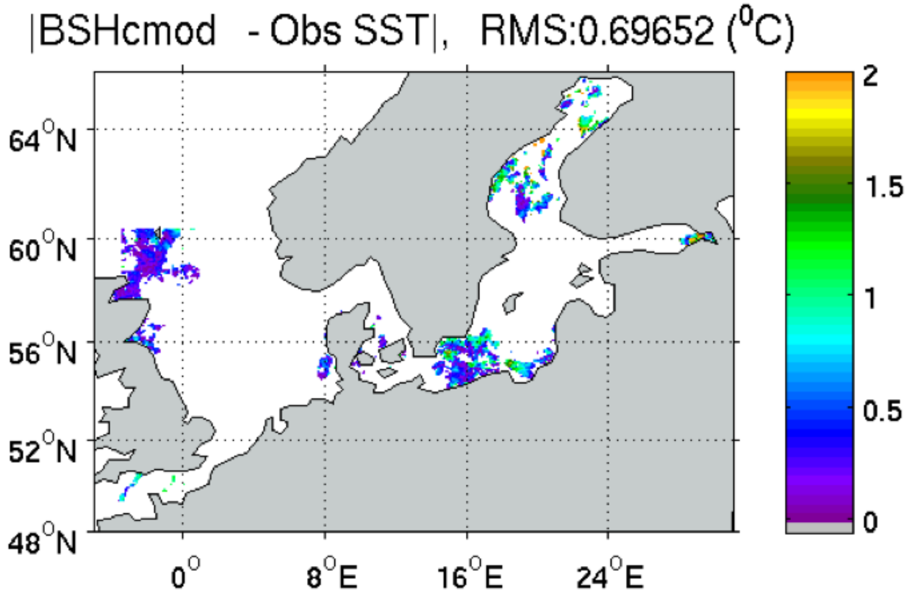
Observation availability

Surface temperature

14.10.2007 00:00±6h



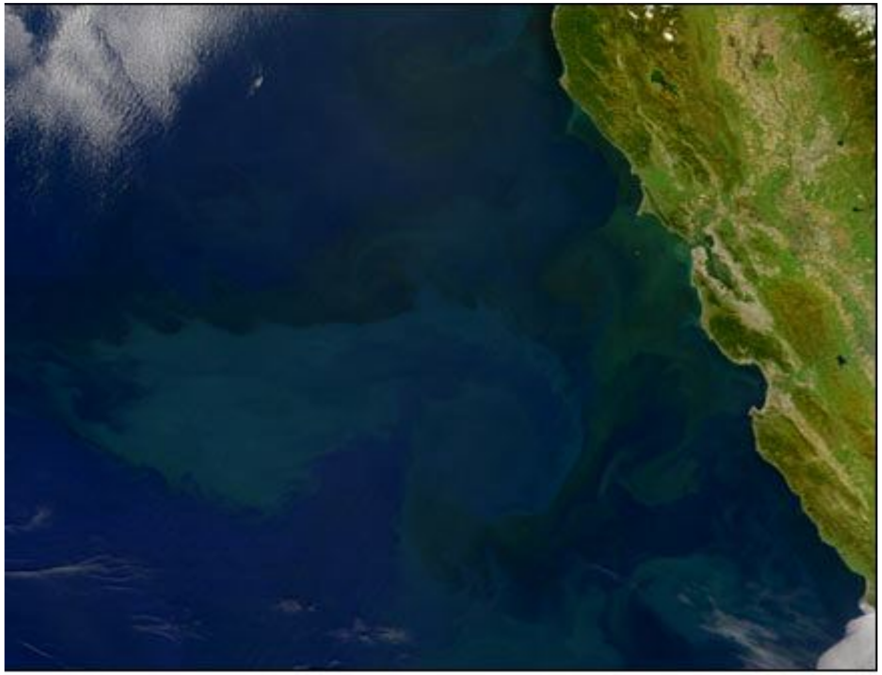
27.10.2007 00:00±6h



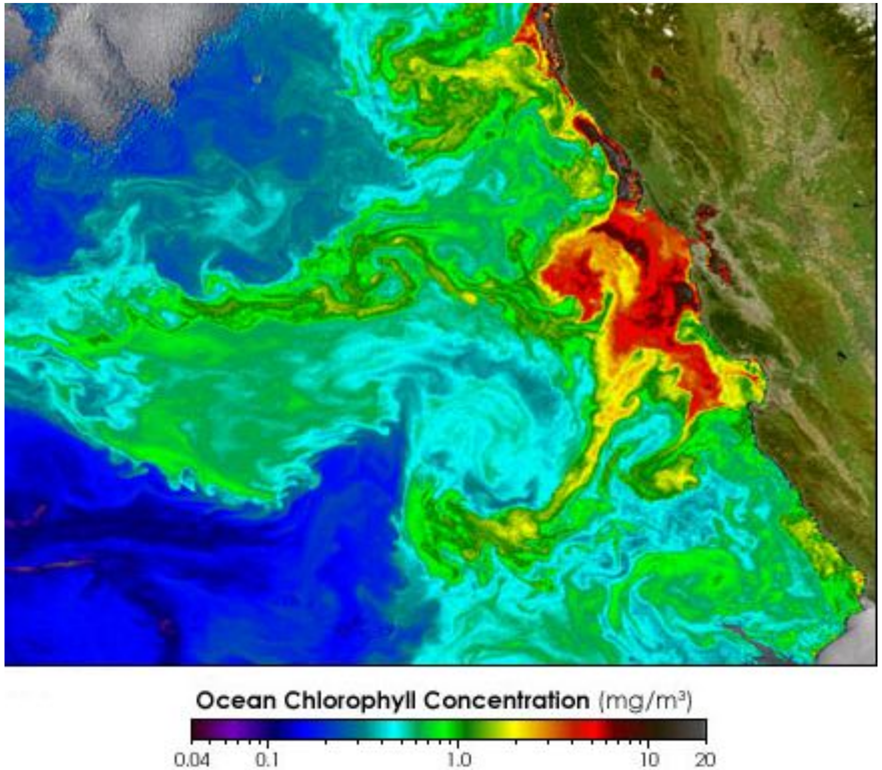
- Strongly irregular data availability
- Frequent data gaps
- Assume constant error and homogeneous spatial influence

Satellite Ocean Color (Chlorophyll) Observations

Natural Color 3/16/2004

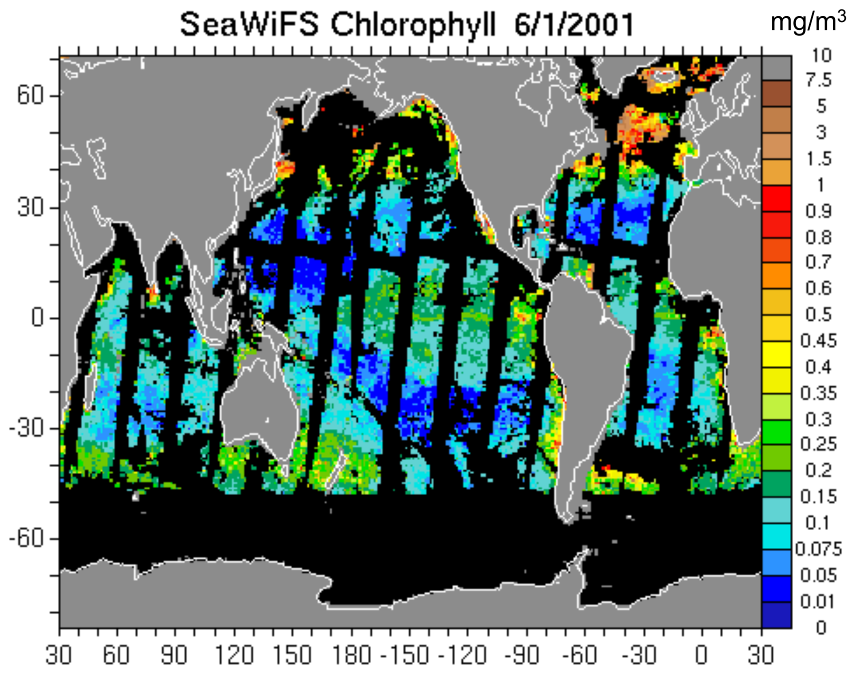


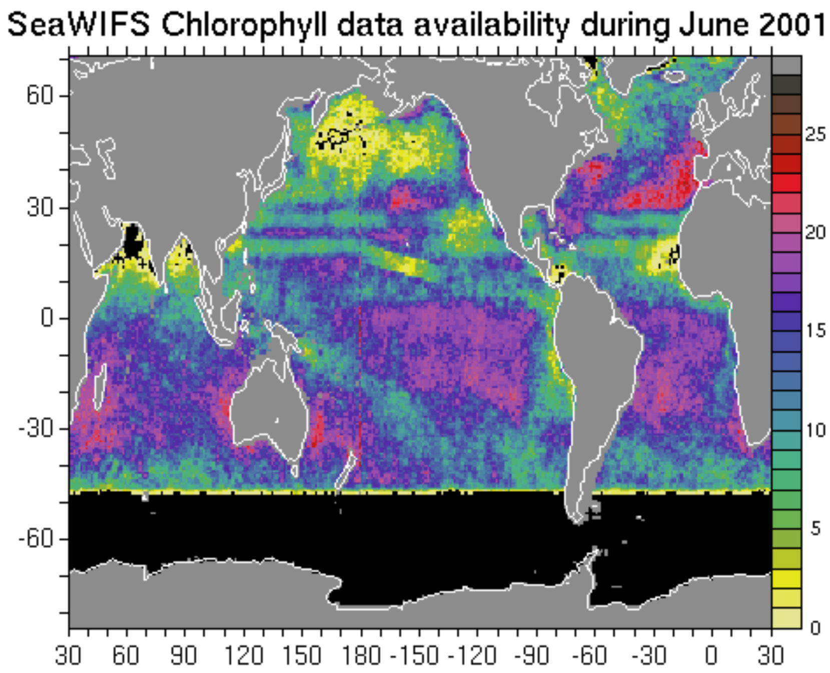
Chlorophyll Concentrations



Source: NASA "Visible Earth", Image courtesy the SeaWiFS Project, NASA/GSFC, and Orbimage

Assimilated Observations

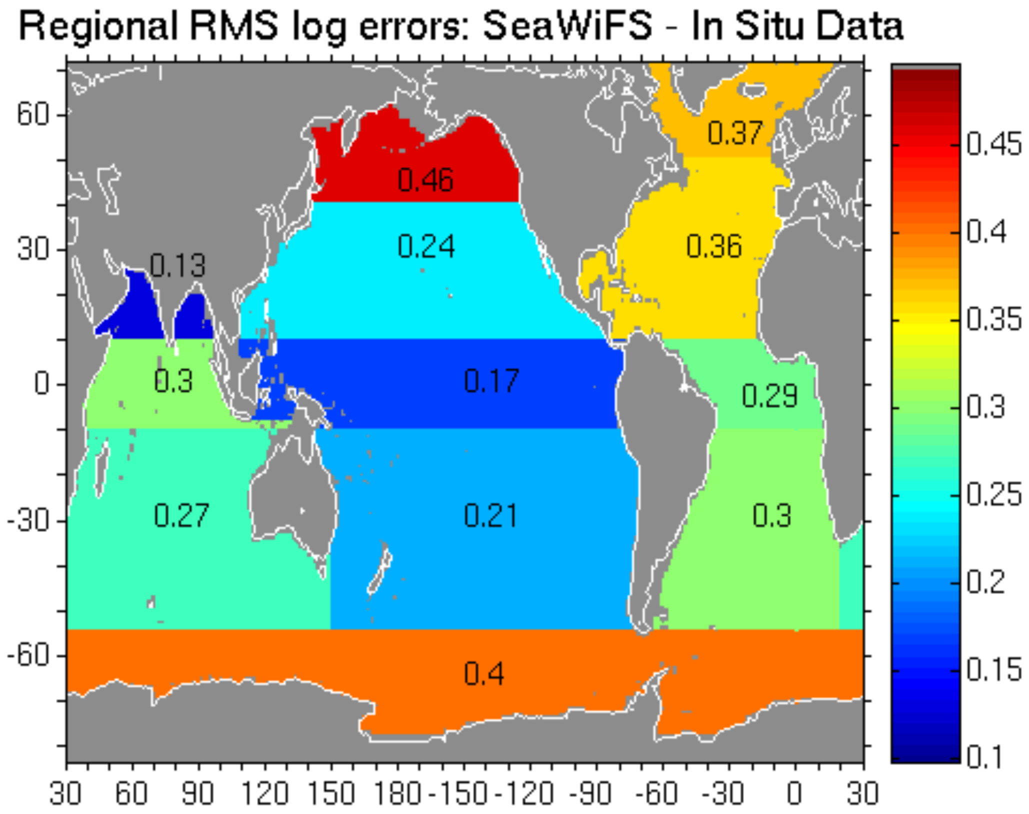




- Daily gridded SeaWiFS chlorophyll data
  - ➢ gaps: satellite track, clouds, polar nights
  - ➢ ~13,000-18,000 data points daily
    (of 41,000 wet grid points)
  - ➢ irregular data availability

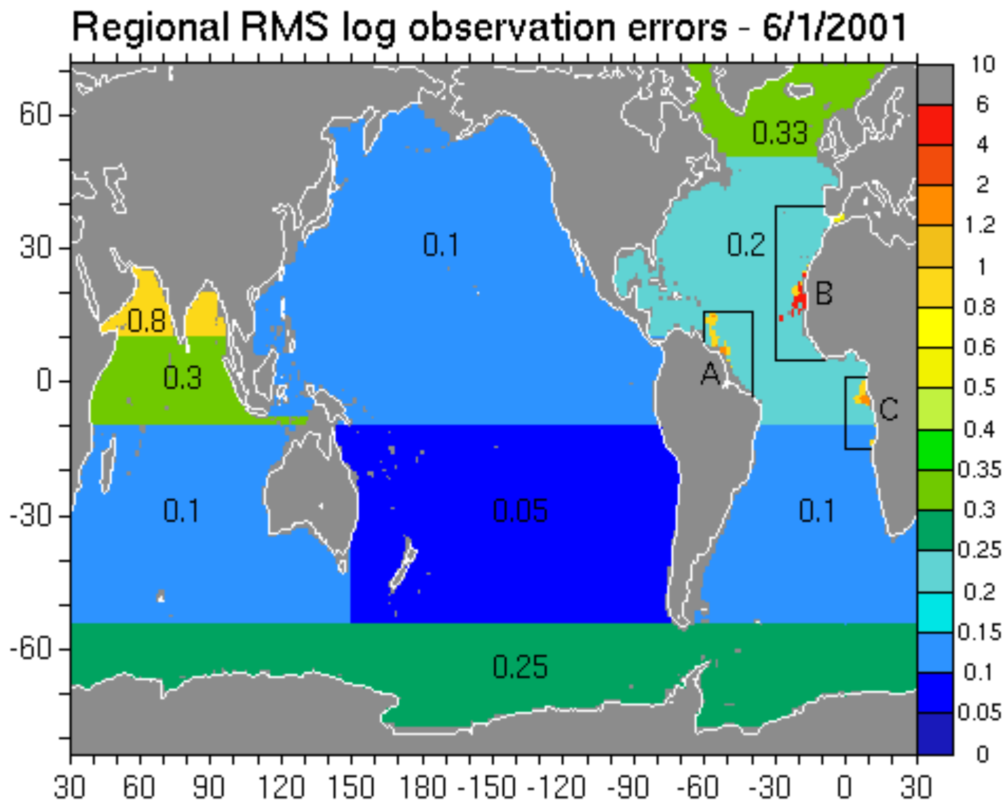
Nerger, L., and W.W. Gregg. J. Marine Systems **68** (2007) 237

Error Estimates



Regional data errors from comparison with 2186 collocation points of in situ data

Observation errors II



- Account regionally for larger errors caused by
  - aerosols (North Indian Ocean, tropical Atlantic)
  - CDOM (Congo and Amazon)
- Error estimates adjusted for filter performance and stability

Model Errors

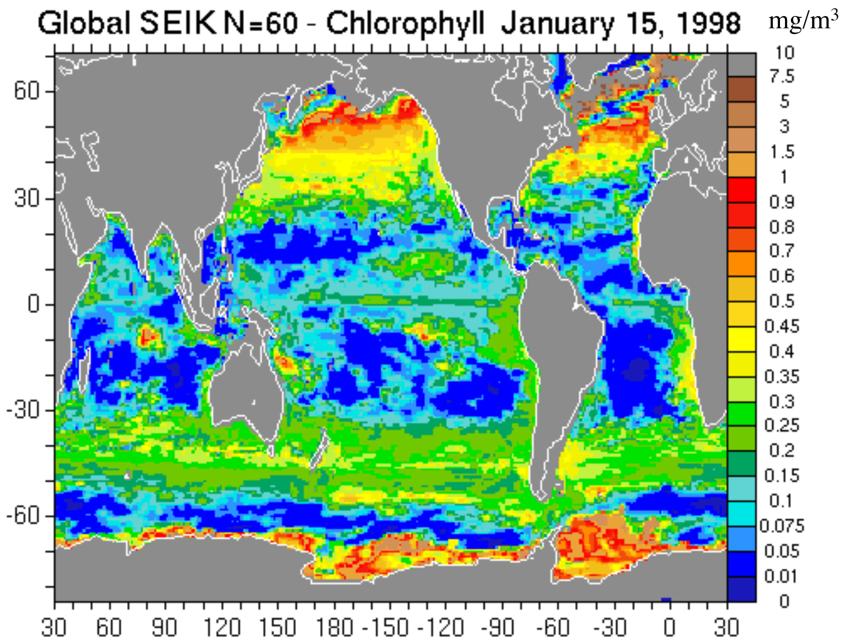
Model errors

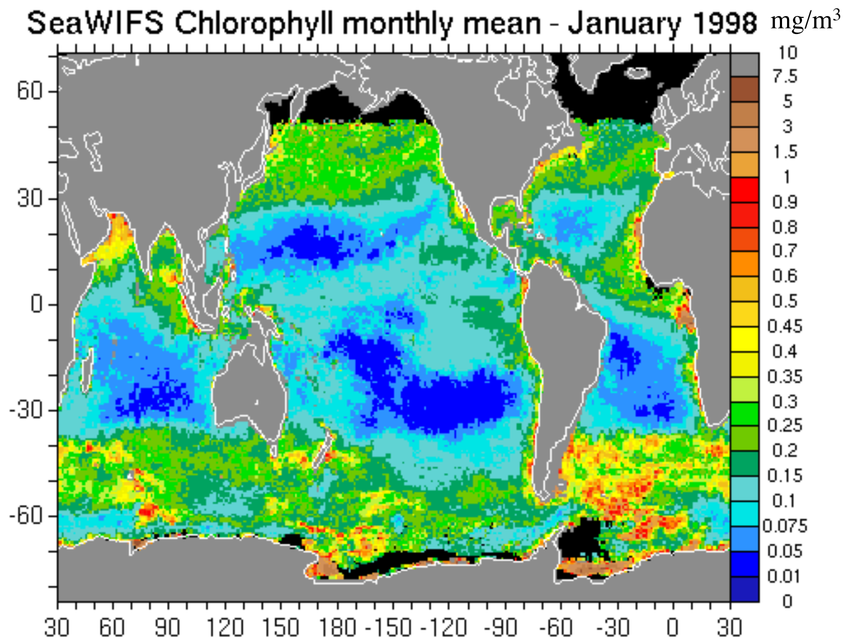
- Representation of reality is not exact
  - Incomplete equations (e.g. missing processes)
  - Inexact forcing (e.g. wind stress on ocean surface)

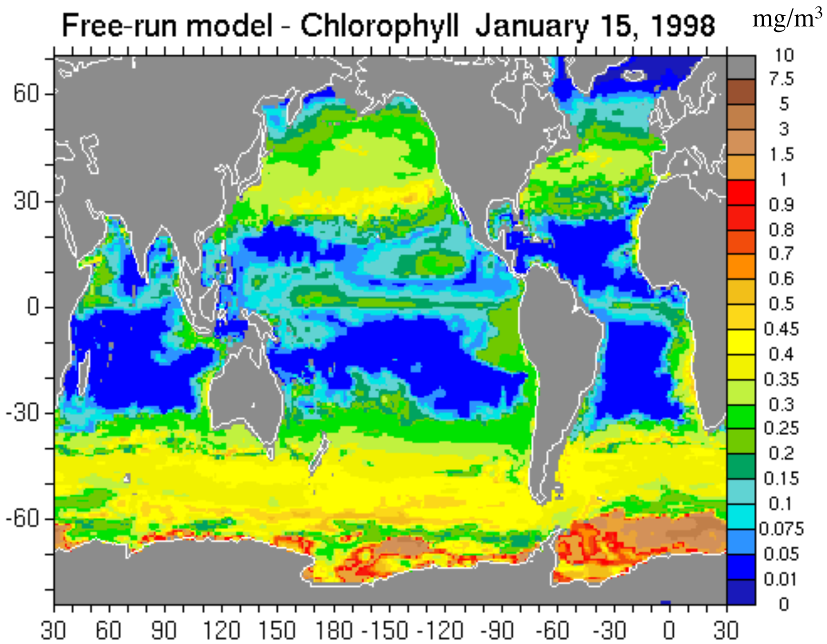
- Accounting for model error
  - Inflation (partly)
  - Simulate stochastic part
  - Bias estimation

Bias correction

Assimilation with global SEIK filter



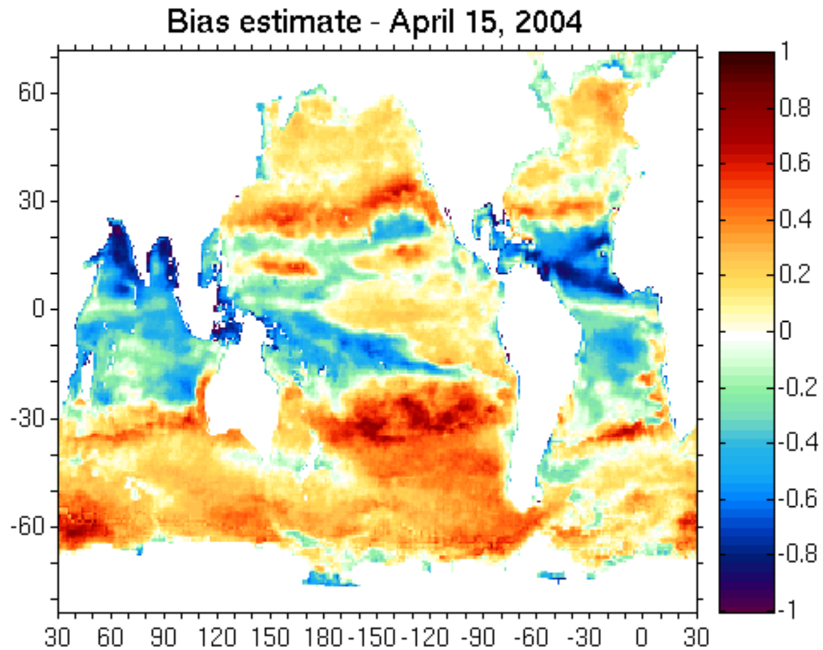




- some improvements of estimated total Chlorophyll
- Increased estimation errors in region with polar night
- SEIK assimilation crashes (earlier for larger ensemble sizes)

Bias Estimation

- un-biased system:
  fluctuation around true state
- biased system:
  systematic over- and underestimation
  (common situation with real data)



- 2-stage bias online bias correction

  1. Estimate bias
     (using fraction of covariance matrix used in 2.)

  2. Estimate de-biased state

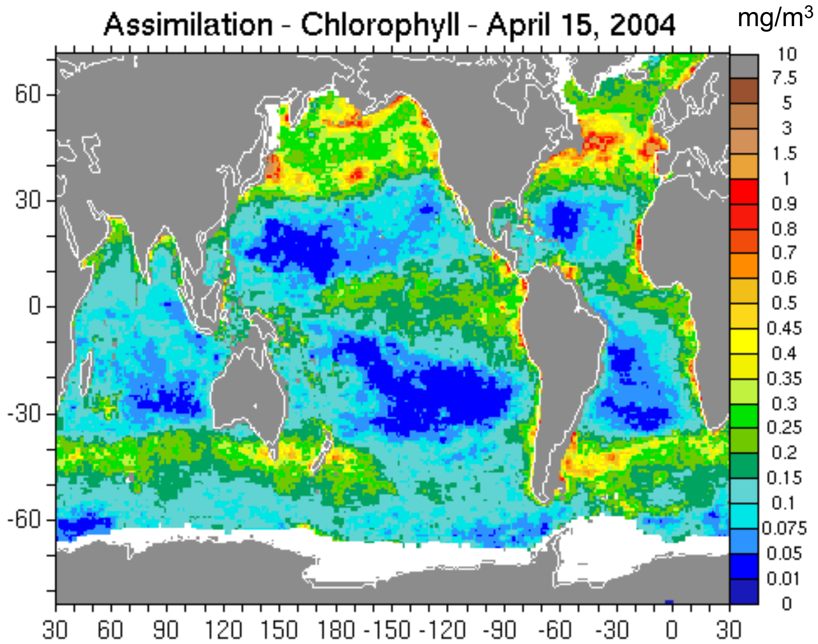
- Forecast

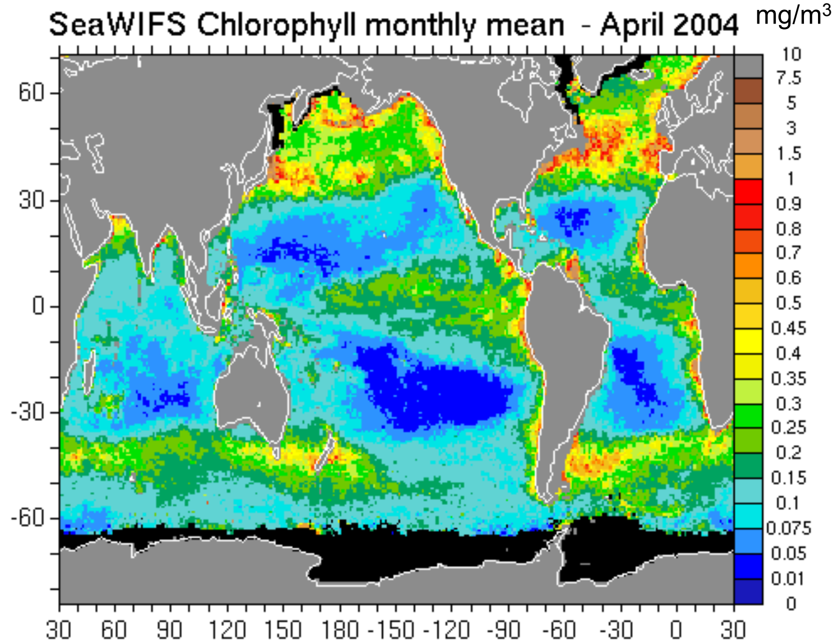
  1. forecast ensemble of biased states

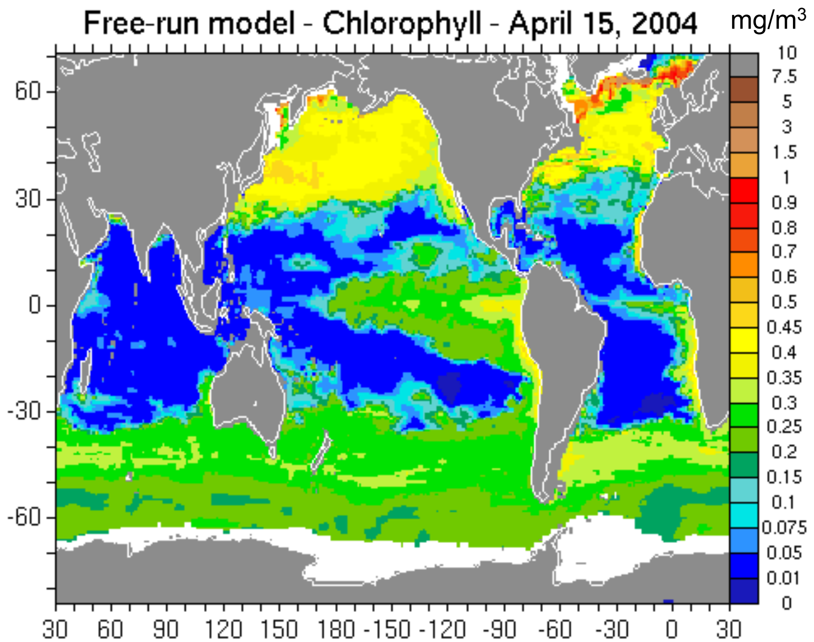
  2. no propagation of bias vector

Nerger, L., and W.W. Gregg. J. Marine Systems, 73 (2008) 87-102

Estimated Chlorophyll - April 15, 2004

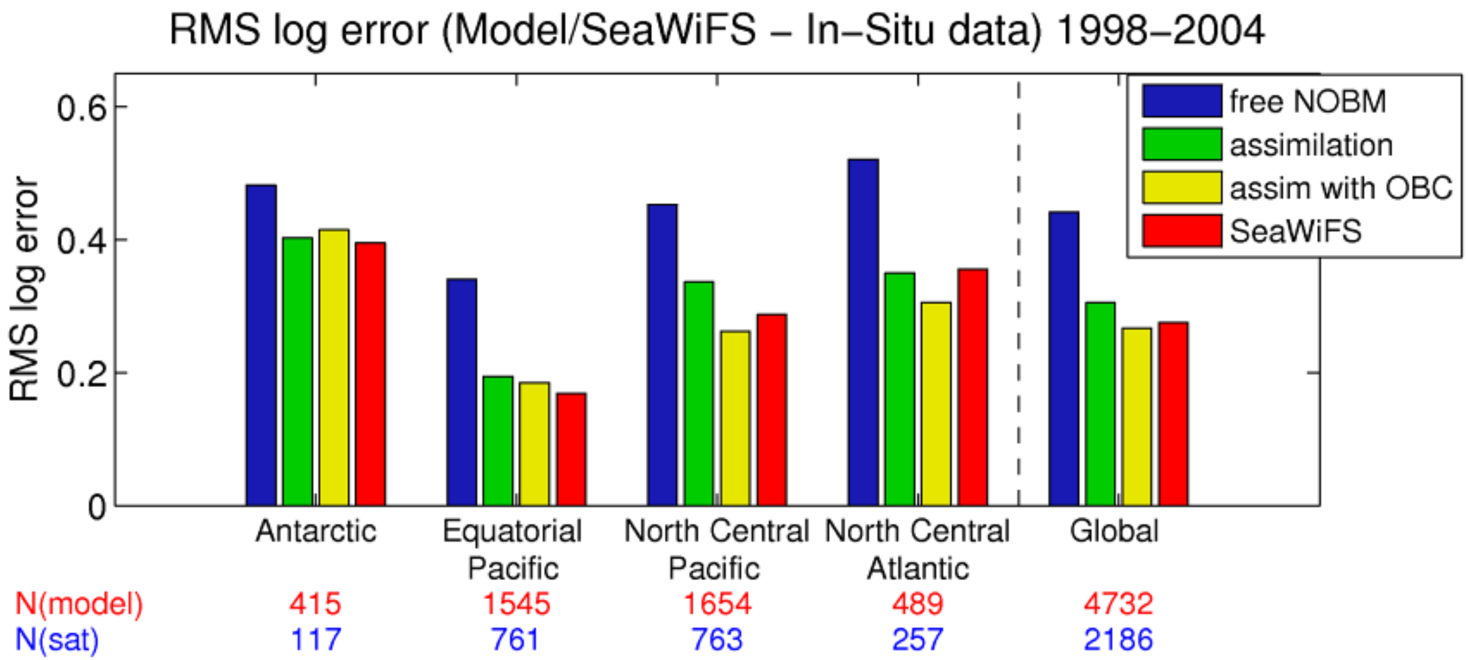






- strongly improved surface Chlorophyll estimate
- intended deviations (Arabian Sea, Congo, Amazon)
- other deviations in high-Chlorophyll regions

Comparison with independent data



- In situ data from SeaBASS/NODC over 1998-2004 (shown basins include about 87% of data)
- Independent from SeaWiFS data (only used for verification of algorithms)
- Compare daily co-located data points

⇒ Assimilation in most regions below SeaWiFS error

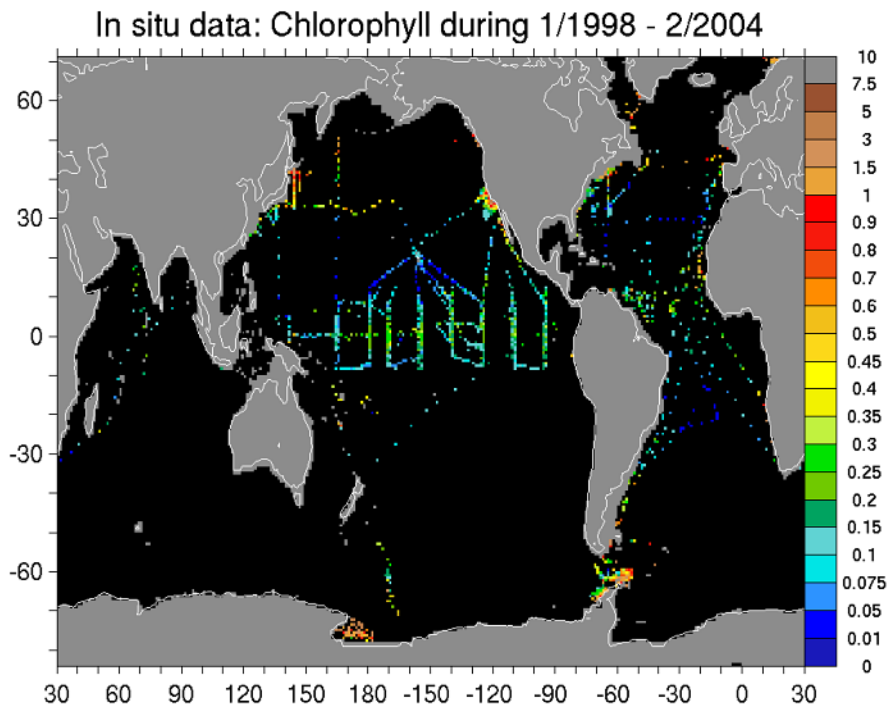
⇒ Bias correction improves almost all basins

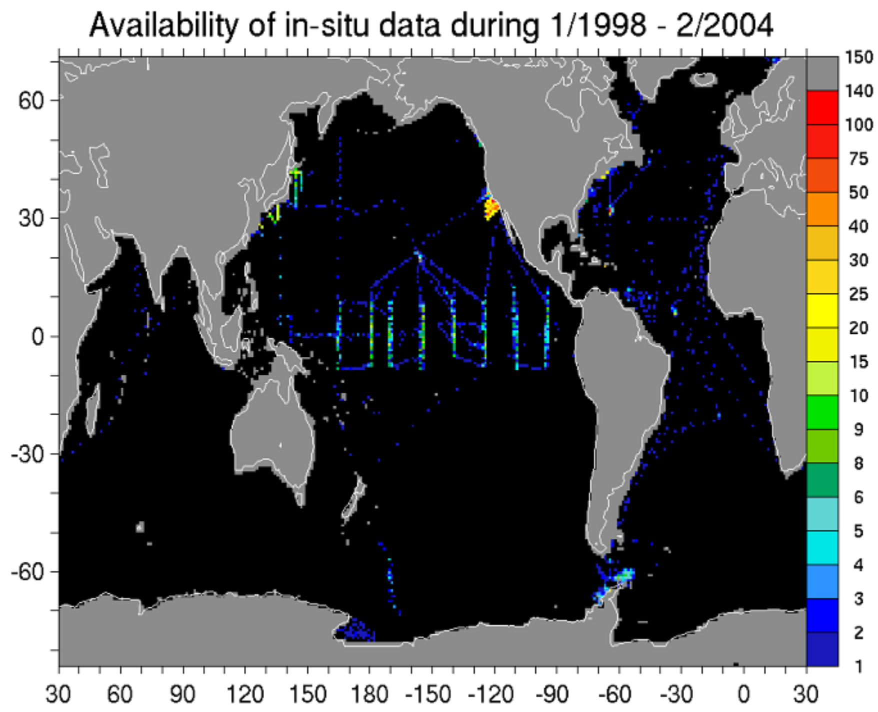
Validation data

Validating a data assimilation system

- Need independent data for validation
  - Necessary, but not sufficient:
    Reduction of deviation from assimilated data
  - Required:
    - Reduction of deviation from independent data
    - Reduction of errors for unobserved variables
- Want to assimilate all available data
  - Data-withholding experiments
  - Twin experiments
  - Validate with data of small influence

In-Situ chlorophyll data





- In situ data from SeaBASS/NODC over 1/1998-2/2004
- Independent from SeaWiFS data
  (only used for verification of algorithms)
- North Central Pacific dominated by CalCOFI data
- North Central Atlantic dominated by BATS data

Summary

- Practical assimilation with ensemble-based Kalman filters
  - ➜ Care and pragmatism required
  - ➜ “pure” filter works suboptimal or not at all
- Theoretical foundation is incomplete
  - ➜ Advancements in between

Thank you!

Acknowledgements:

W. Hiller, J. Schröter, T. Janjic, S. Losa (AWI)
Watson Gregg, Nancy Casey (NASA/GSFC)