SPP1788 Colloquium 2019, Bad Aibling, 3.-6.6.2019

Ensemble Data Assimilation

Algorithms – Applications – Software

Lars Nerger

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

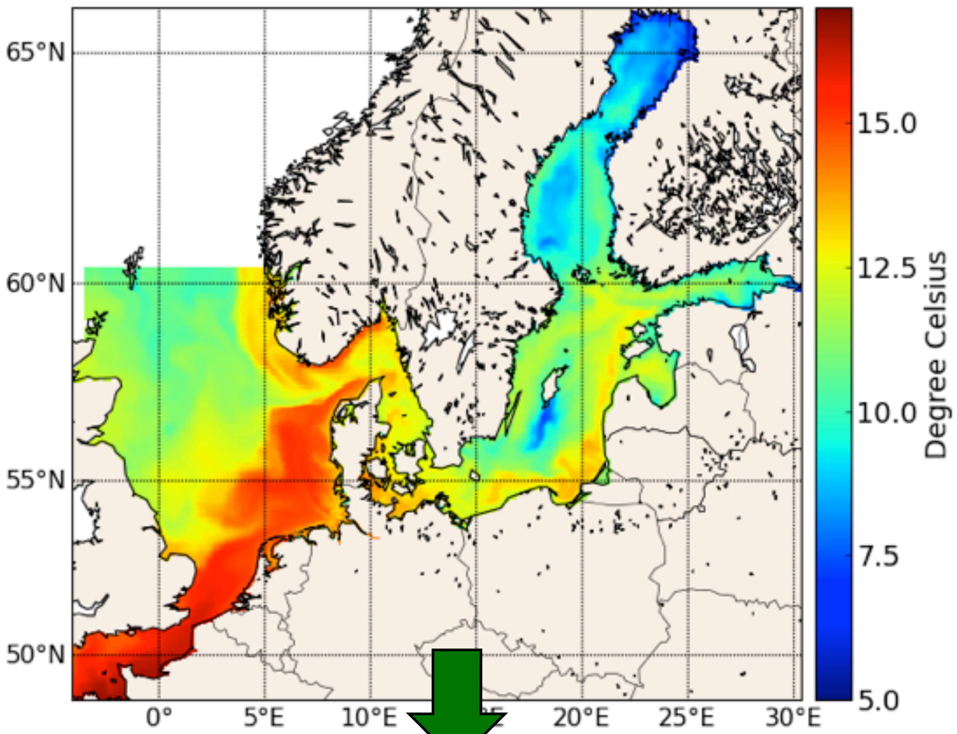
Acknowledgements:
W. Hiller, J. Schröter, S. Losa, A. Androsov,
H. Pradhan, M. Goodliff, Q. Tang, Q. Yang, L. Mu

PDAF Parallel Data Assimilation Framework

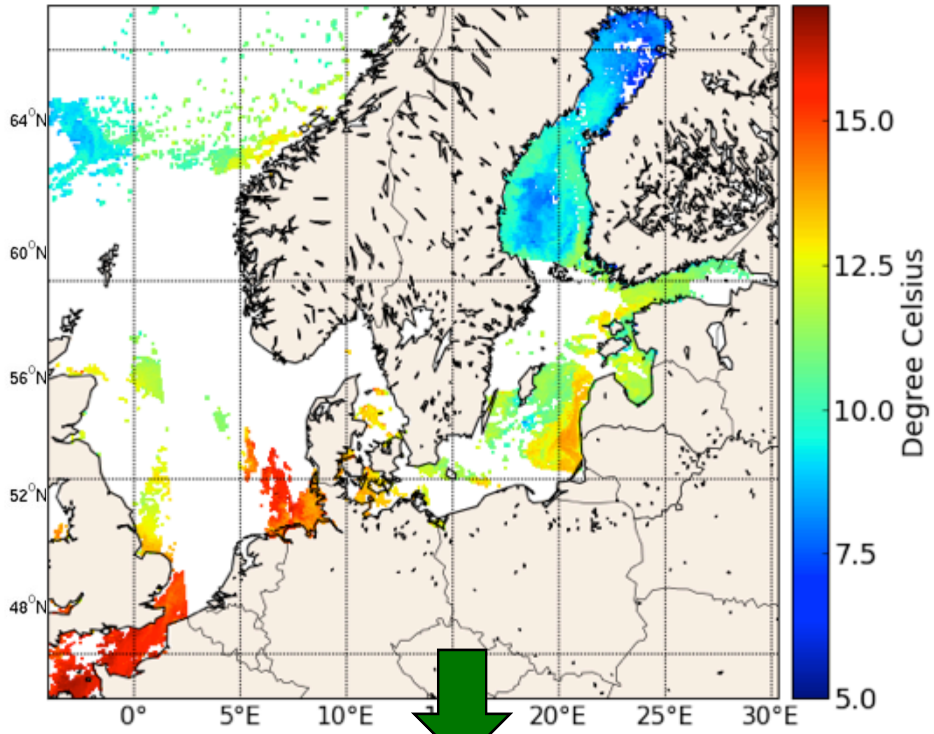
ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR-
UND MEERESFORSCHUNG

Motivation

Model surface temperature



Satellite surface temperature



Combine both sources of information

quantitatively by computer algorithm

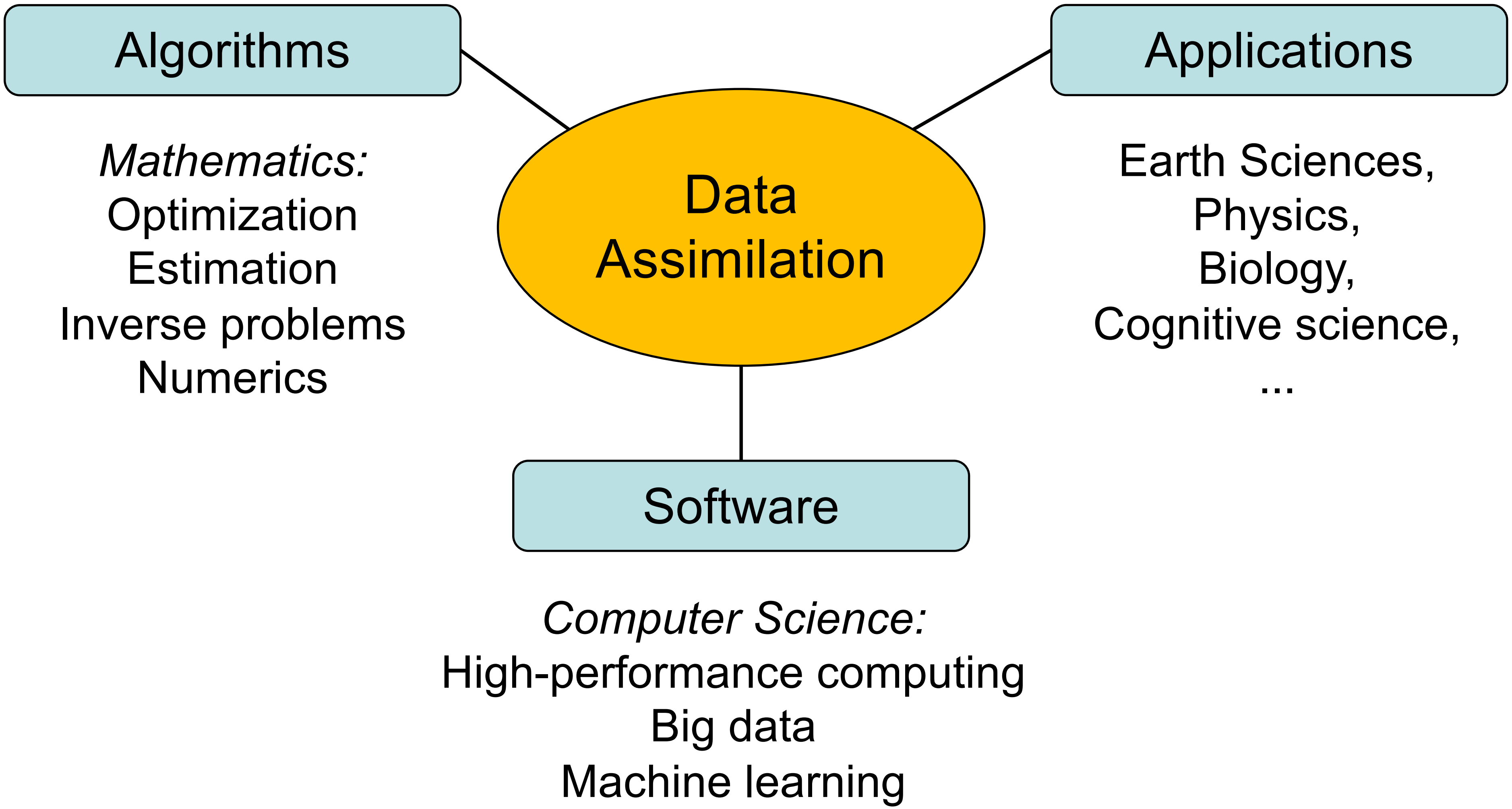
→ Data Assimilation

Data Assimilation

Methodology to combine model with real data

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

Interdisciplinarity of Data Assimilation



Outline

Ensemble Data Assimilation

Algorithms / Methodology

- Efficient methods for high-dimensional nonlinear systems

Applications

- Examples of what one can expect to achieve

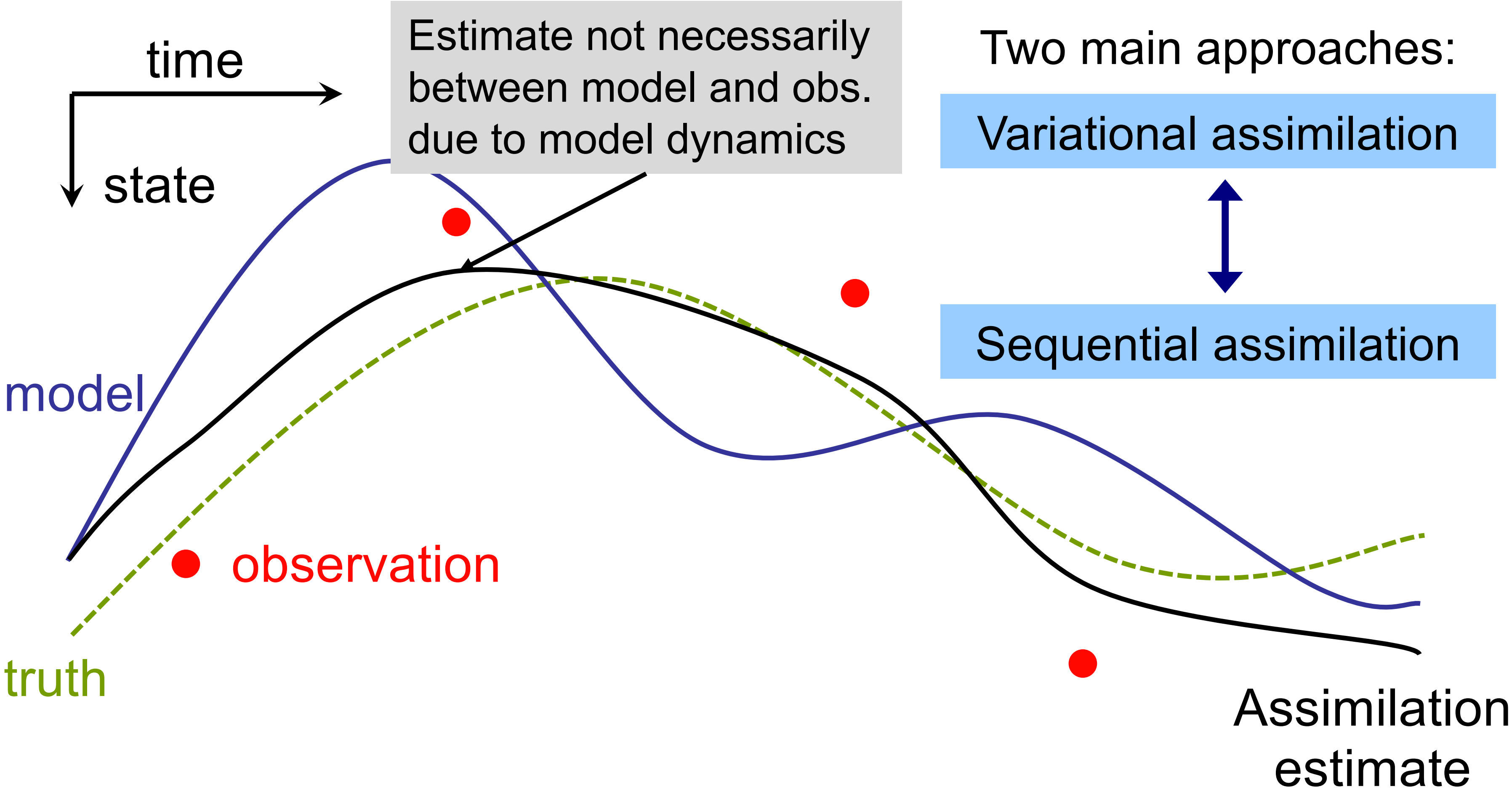
Software

- Make ensemble data assimilation easily usable
  - Parallel Data Assimilation Framework (PDAF)

Methodology

Data Assimilation – a general view

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



Two main approaches:

Variational assimilation

Sequential assimilation

Goal: Obtain optimal estimate of system constrained by model dynamics and observations

Needed for Data assimilation

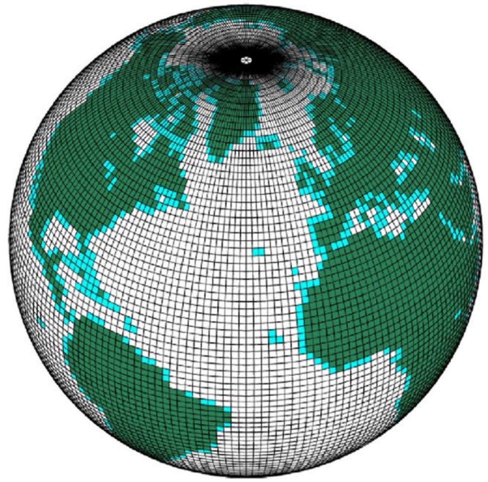
1. Model
   - with some skill
2. Observations
   - with finite errors
   - related to model fields
3. Data assimilation method

Models

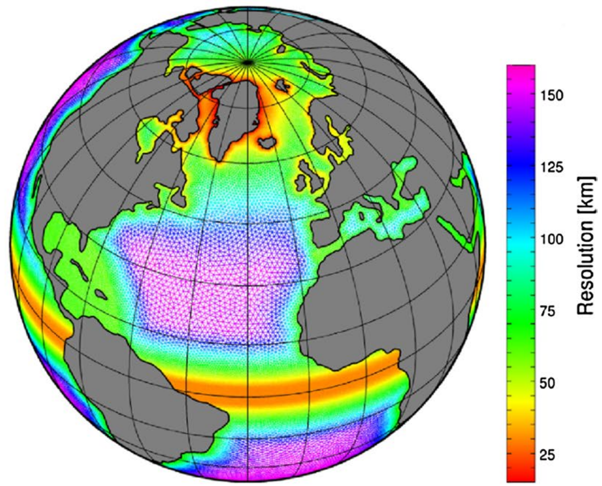
Simulate dynamics, e.g. the ocean

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

- Uncertainties
  - initial model fields
  - external forcing
  - in predictions due to model formulation



Uniform-resolution mesh



Variable-resolution mesh (ocean model FESOM)

Observations

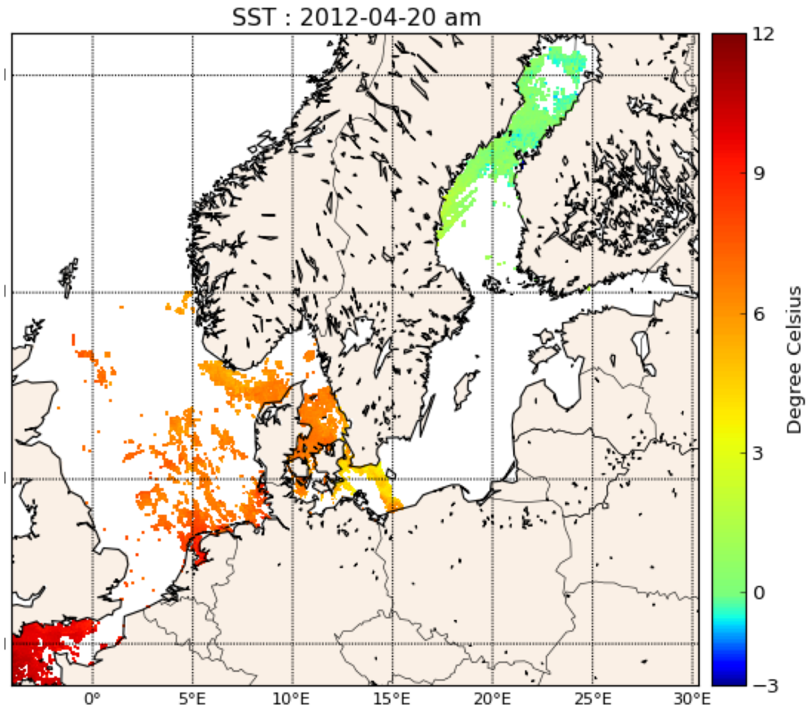
Measure different fields … for example in the Ocean

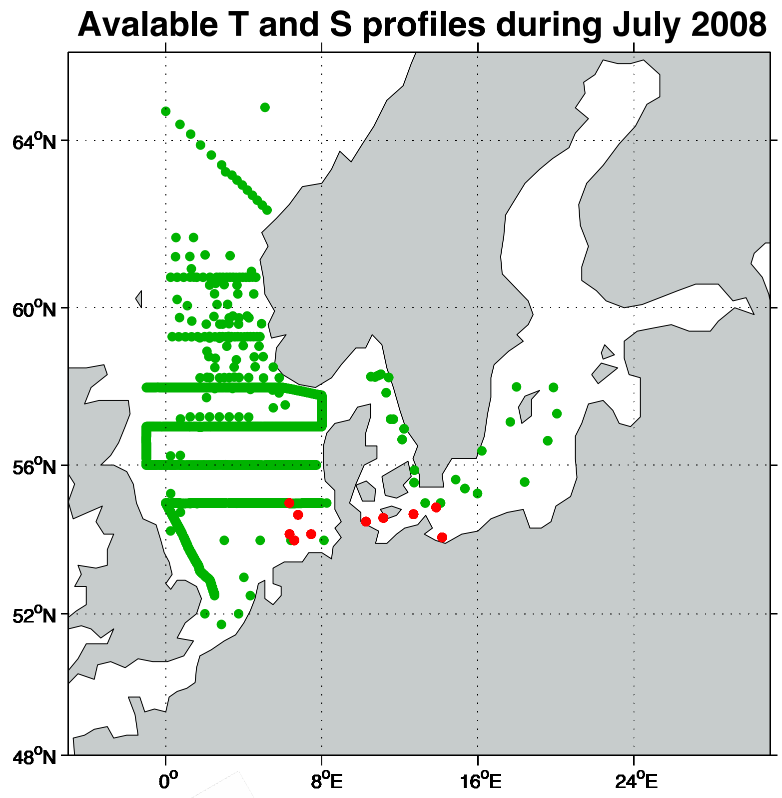
- Remote sensing
  - E.g. surface temperature, salinity, sea surface height, ocean color, sea ice concentrations & thickness
- In situ (ships, autonomous vehicles, …)
  - Argo, CTD, Gliders, …

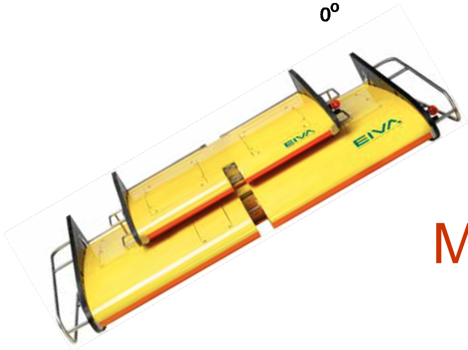
- Data is sparse: some fields, data gaps
- Uncertainties
  - Measurement errors
  - Representation errors:
    Model and data do not represent exactly the same
    (e.g. cause by finite model resolution)

Example: Physical Data in North & Baltic Seas

Satellite surface temperature
(12-hour composite)



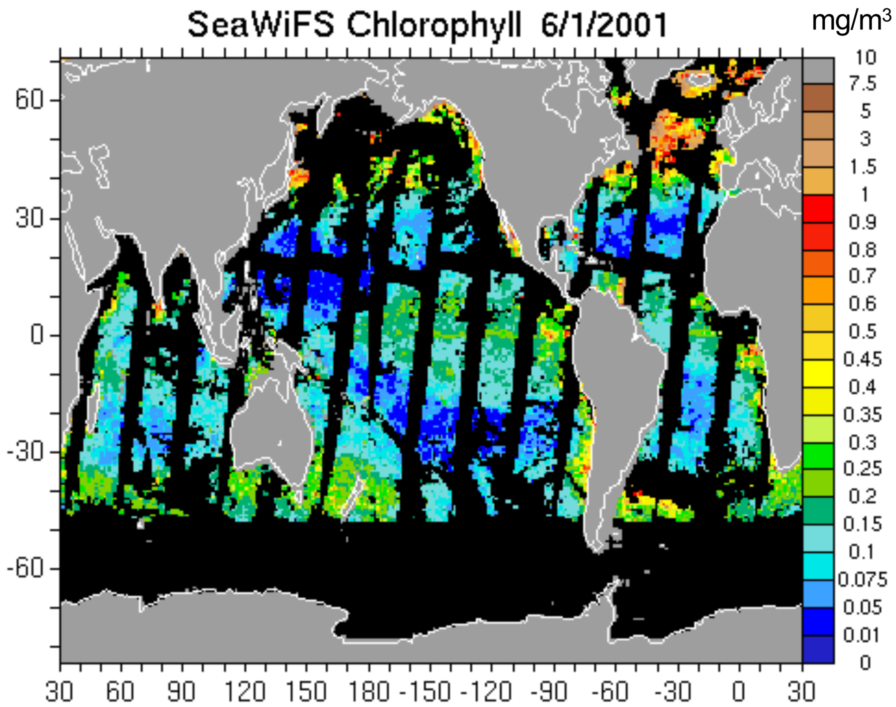


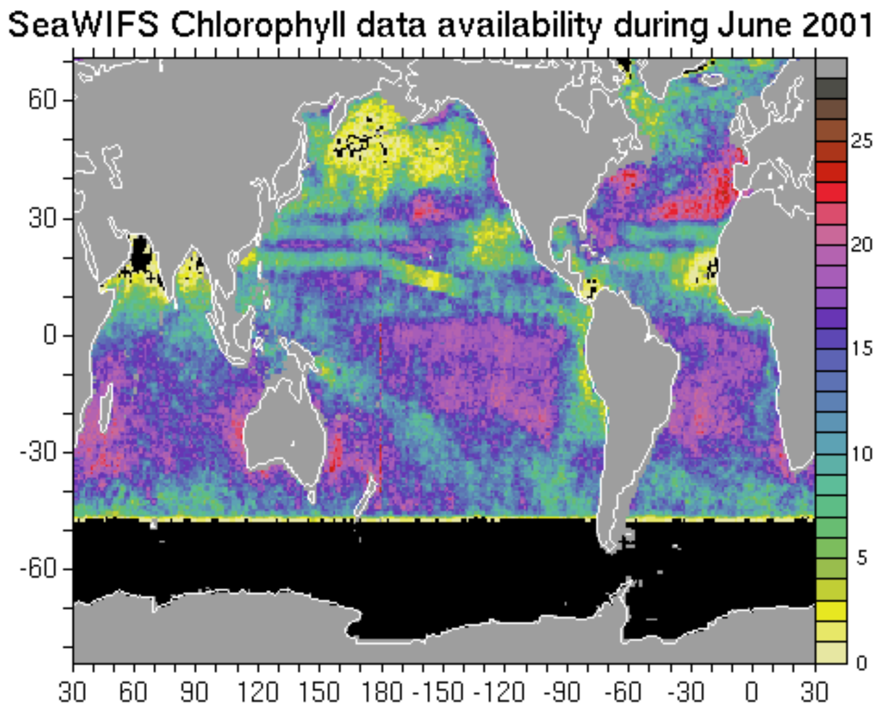


Scanfish and CTD profiles

MARNET stations

Example: Chlorophyll-a observations (SeaWiFS)





Daily gridded SeaWiFS chlorophyll data

- gaps: satellite track, clouds, polar nights
- On model grid: ~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

Observation Error Estimates

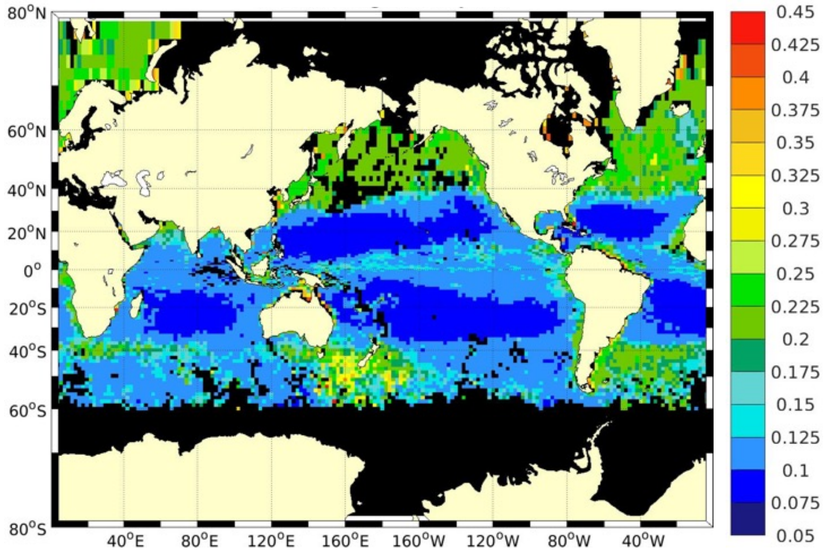
If observation errors available:

- they are typically usable
- usually do not account for representation errors (might be too low)

If no observation errors available:

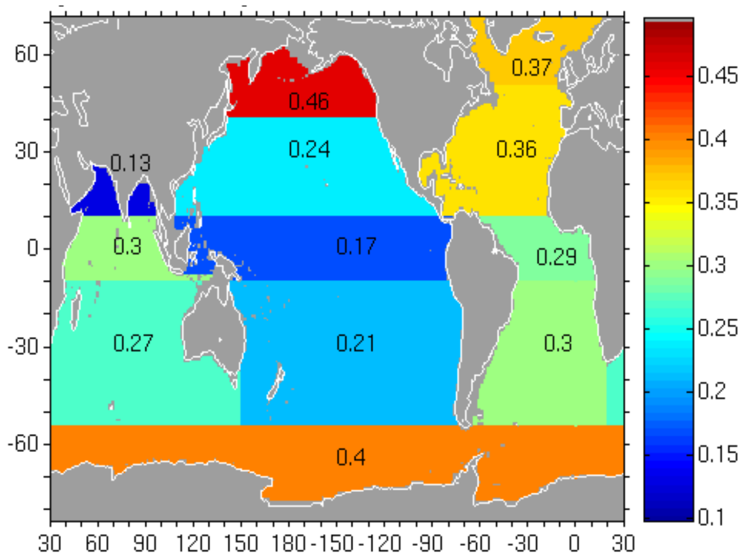
- need to estimate them

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

Data Assimilation Methods

Combine observations and model state estimate

- Account for uncertainty in observations
- Account for uncertainty in model state estimate
- Account for relations (correlations) between observed part of the model state and unobserved parts

Ensemble Data Assimilation

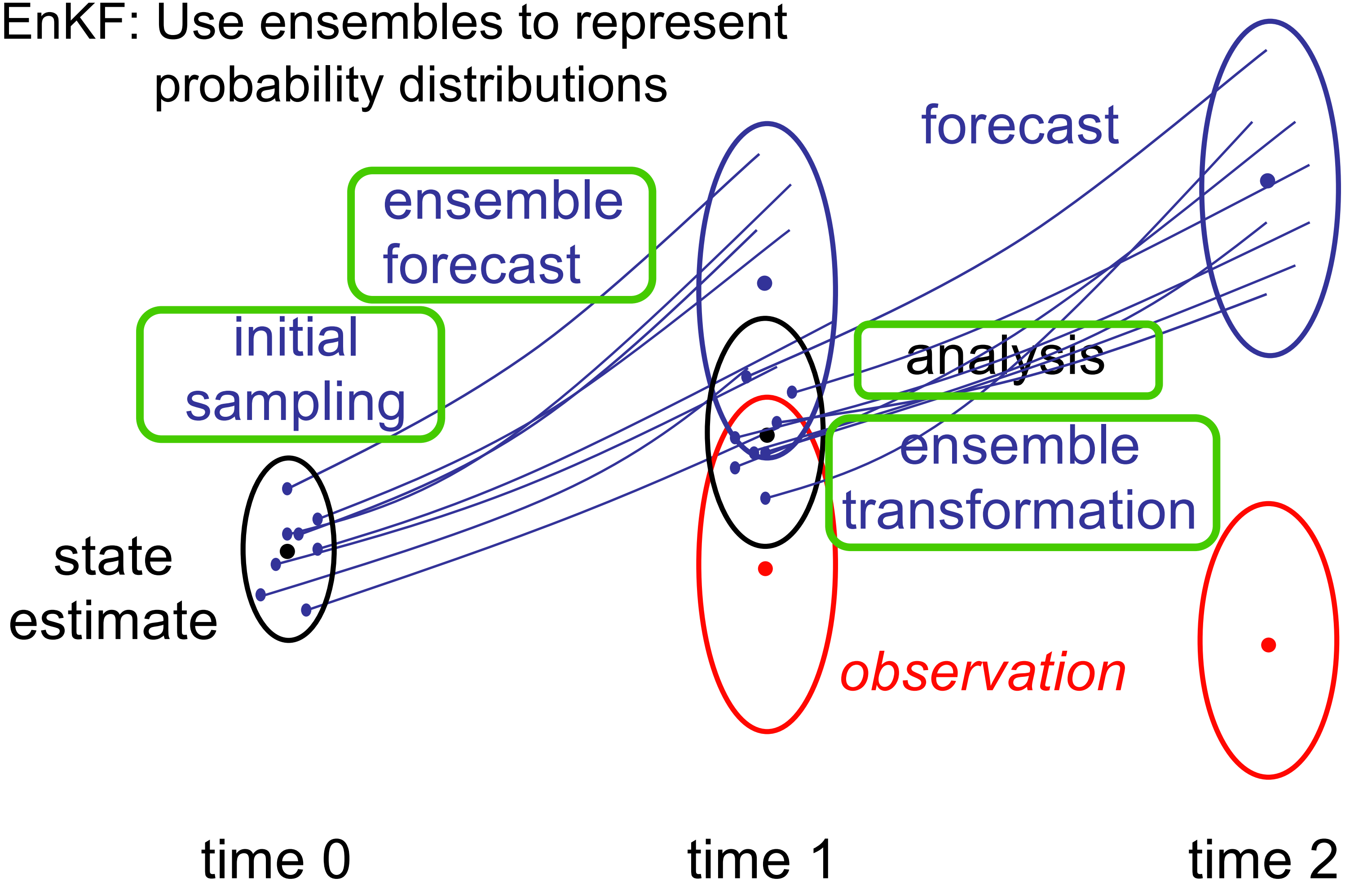
Estimate uncertainty

Ensemble Kalman Filters

First formulated by G. Evensen (EnKF, J. Geophys. Res. 1994)

Kalman filter: express probability distributions by mean and covariance matrix

EnKF: Use ensembles to represent probability distributions



There are many possible choices!

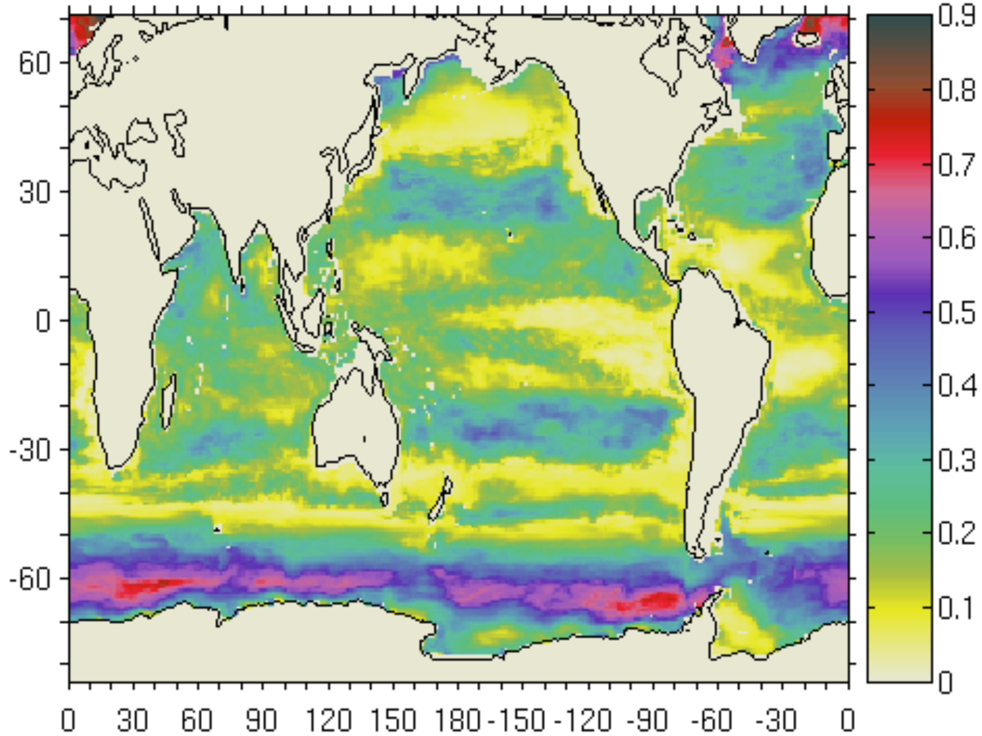
What is optimal is part of our research

Different choices in PDAF

Ensemble Covariance Matrix

- Provide uncertainty information (variances + covariances)
- Generated dynamically
  by propagating ensemble of model states

Uncertainty: Standard deviation of log Chlorophyll

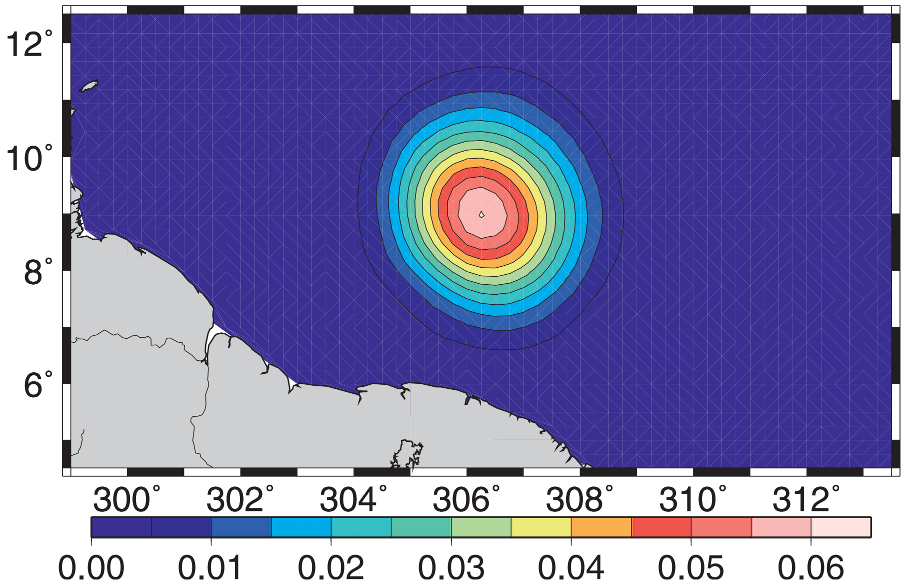


Ensemble Covariance Matrix (II)

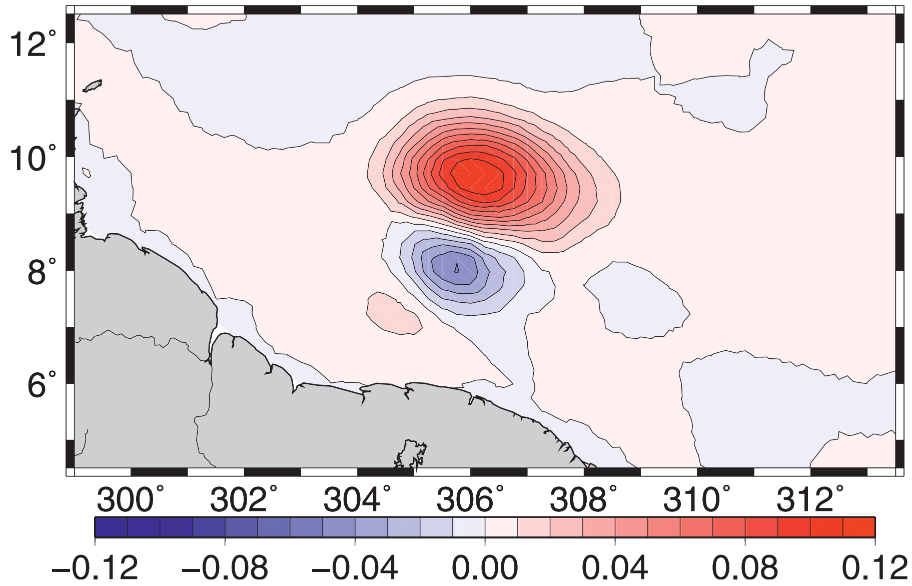
- Also:
  Provide information on error correlations
  (between different locations and different fields)

- Example: Assimilation of sea surface height
  (Brankart et al., Mon. Wea. Rev. 137 (2009) 1908-1927)

Assimilation increment in sea surface height

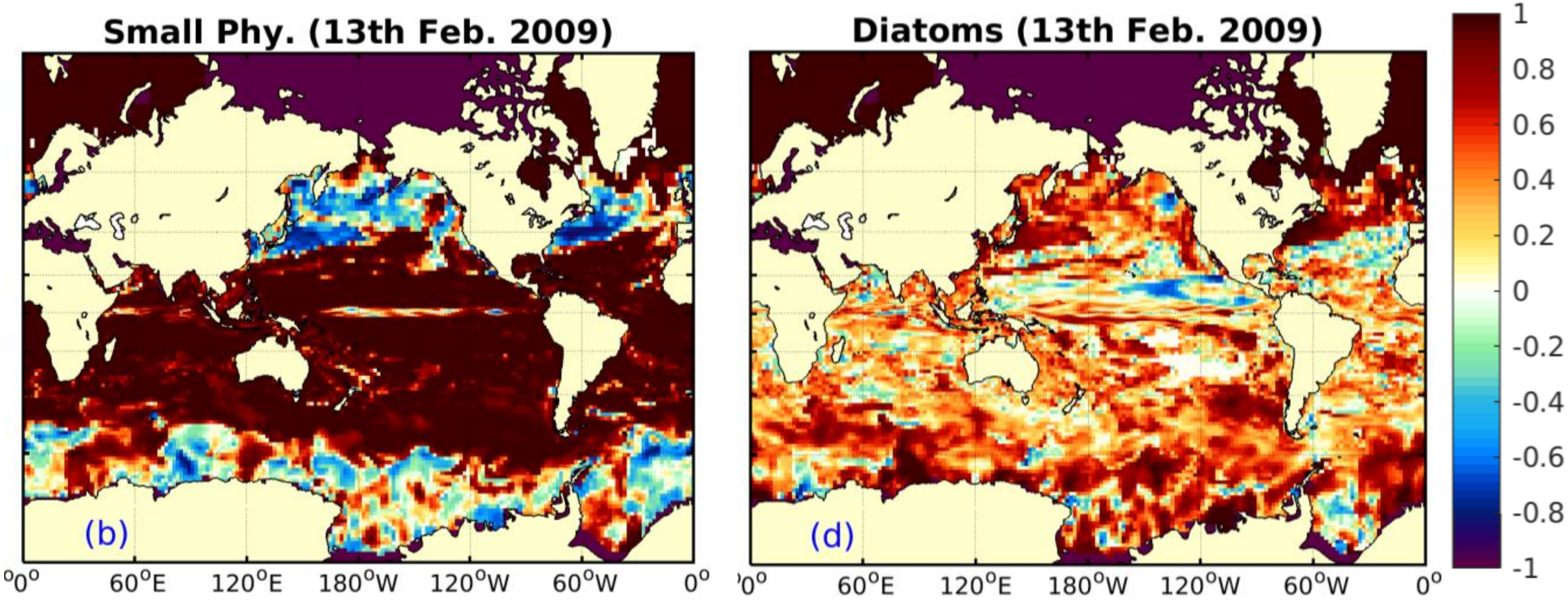


Induced change in zonal velocity



Ensemble-estimated Cross-correlations

Cross correlations between total chlorophyll
and chlorophyll in phytoplankton groups



Cross-correlations are used to correct non-observed quantities from observed ones

Pradhan et al., J. Geophy. Res. Oceans, 124 (2019) 470-490

Ensemble-based/error-subspace Kalman filters

A little "zoo" (not complete):

Which filter should one use?

EnKF(94/98)

Efficiency of SEIK
(Nerger et al. 2005)

RRSQRT

ROEK

SEEK

SEIK

EnKF(2003)

EnKF(2004)

EAKF

EnSRF

DEnKF

ETKF

ESTKF

Filter instability
(Nerger 2015)

MLEF

SPKF

ESSE

RHF

anamorphosis

New filter
formulation
(Nerger et al. 2012)

L. Nerger et al., Tellus 57A (2005) 715-735

L. Nerger et al., Monthly Weather Review 140 (2012) 2335-2345

L. Nerger, Monthly Weather Review 143 (2015) 1554-1567

S. Vetra-Carvalho et al., Tellus A 70 (2018) 1445364

Assessing Ensemble Kalman Filters

Mathematical assessment of ensemble Kalman filters limited by

- optimality only proven for Gaussian error distributions
- convergence properties only clear for large ensemble limit

but

- models are nonlinear -> non-Gaussian distributions
- only small ensemble feasible to run for high-dimensional models

A practical approach

- compare and characterize behavior of different methods
- reach general conclusions from analyzing differences mathematically

Further: Ensemble Kalman filters don't work in 'pure' form

- Need adaptions ('fixes')

Essential “Fixes” for Ensemble Filters

Covariance Inflation

Localization

Covariance inflation

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

  → can lead to filter divergence

- Simple remedy

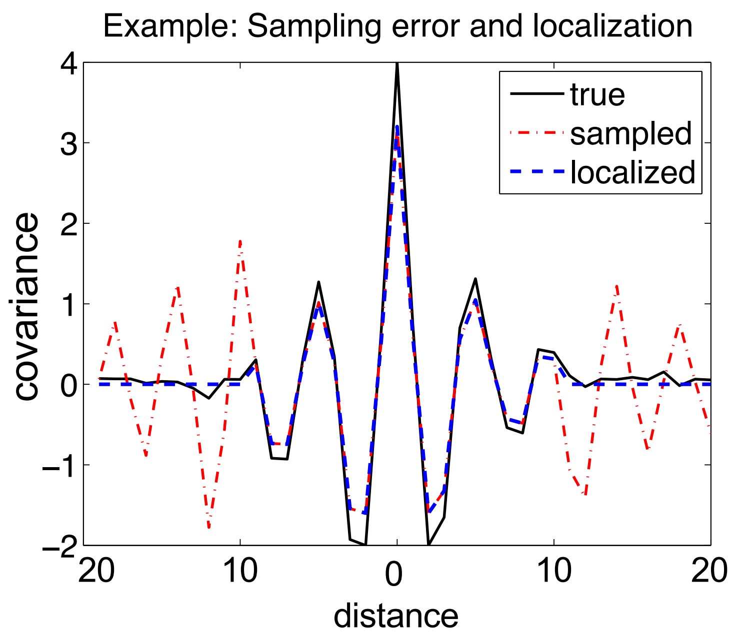
  → Increase error estimate before analysis

- Inflation
  - Increase ensemble spread by constant factor
  - Some filters allow multiplication of a small matrix ("forgetting factor" ≤1; computationally very efficient)
  - Needs to be experimentally tuned

(Mathematically, this is a regularization)

Localization: Why and how?

- Combination of observations and model state based on ensemble estimates of error covariance matrices
- Finite ensemble size leads to significant sampling errors
  - errors in variance estimates
    - usually too small
  - errors in correlation estimates
    - wrong size if correlation exists
    - spurious correlations when true correlation is zero

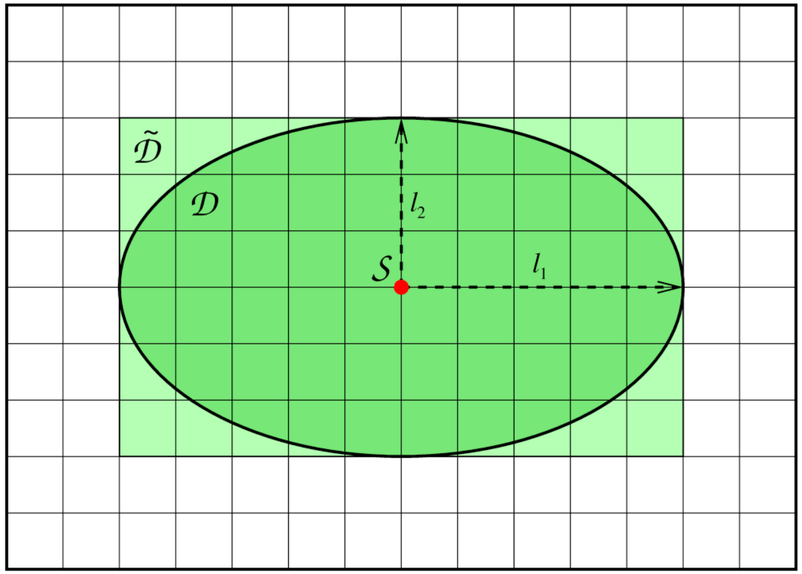


- Assume: long-distance correlations are small in reality
  - Localization: damp or remove estimated long-range correlations (Houtekamer & Mitchell, 1998, 2001)

Observation Localization

Local Analysis:

- Update small regions (like single vertical columns) allows to define distance
- Use only observations within some distance around this region
- State update and ensemble transformation fully local



S: Analysis region
D: Corresponding data region

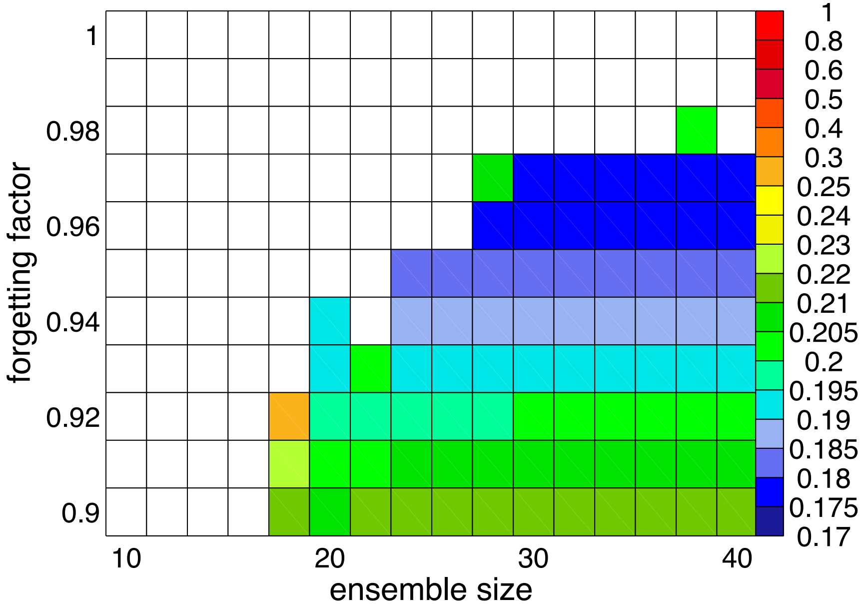
Observation localization:

- Down-weight observations with increasing distance

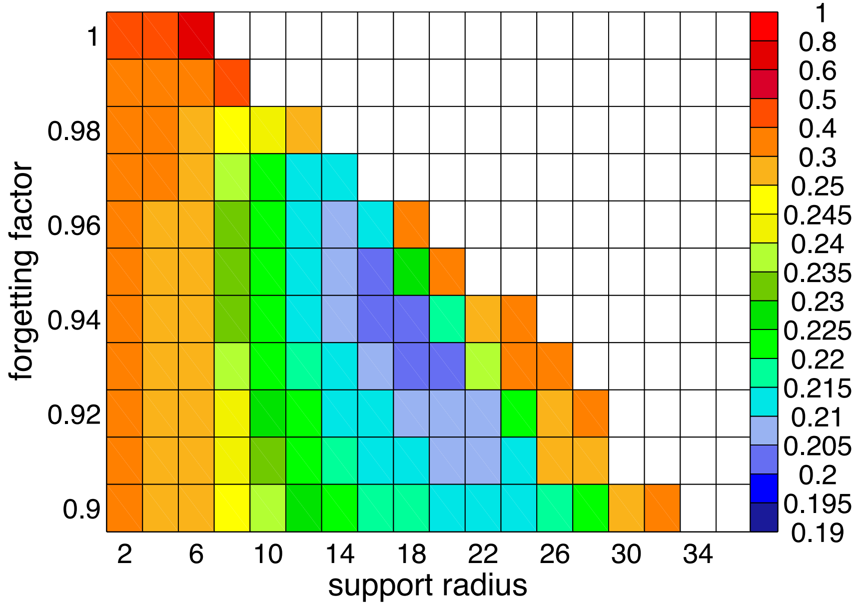
Impact of inflation and localization

Experiments with Lorenz96 model

Global filter



Localized, ensemble size 10



- smaller ensemble usable with localization
- optimal combination of forgetting factor and support radius

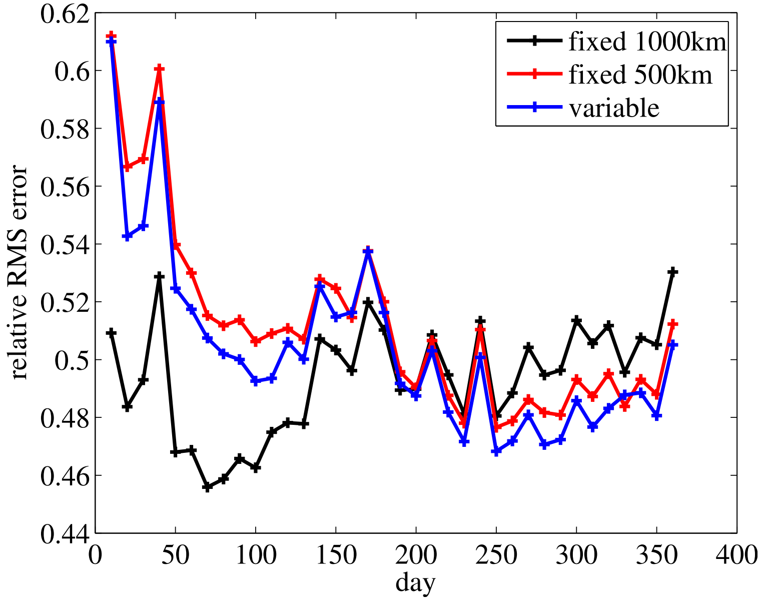
Adaptive localization radius in global ocean model

- Localization radius is usually hand-tuned
- Numerical analysis in small models shows:
  errors minimal when localization radius chosen such that

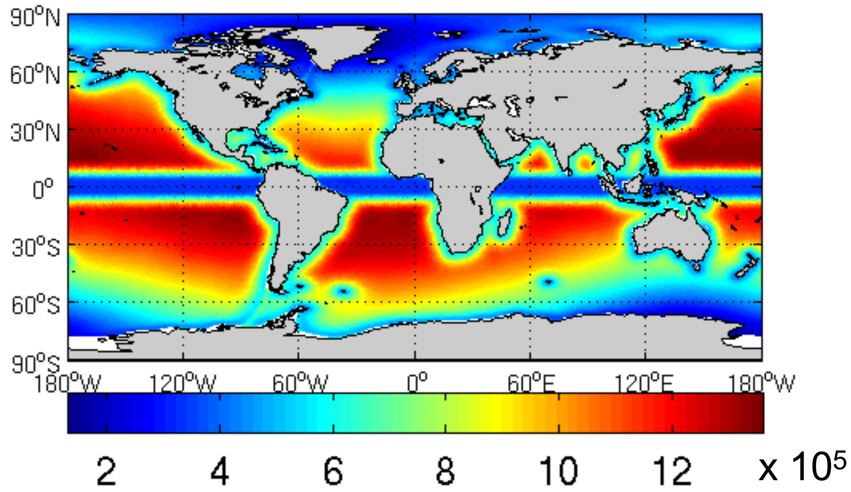
  local sum of observation weights = ensemble size

- Application with FESOM (Finite Element Sea-ice Ocean Model):
  - Fixed 1000km radius leads to increasing errors in 2nd half of year
  - Lower RMS error in sea surface height than fixed 500km radius

Error-reduction of sea surface height



Localization radius [meter]



Kirchgessner, Nerger, Bunse-Gerstner, Mon. Weather Rev., 142 (2012) 2165-2175

Current developements

Current developements

- Ensemble Kalman filters (and standard variational methods) are current ‘work horses’
  - With various ‘fixes’ like localization

- Aim: Better account for nonlinearity

- Fully nonlinear: Particle filters
  - still no established method for high-dim.
- Hybrid methods
  - Hybrid ensemble-variational
  - Hybrid ensemble Kalman – particle filters
- Iterative filters

Linear and Nonlinear Ensemble Filters

- Represent state and its error by ensemble $\mathbf{X}$ of N states
- Forecast:
  - Integrate ensemble with numerical model
- Analysis:
  - update ensemble mean $\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^f + \mathbf{X}'^f \tilde{\mathbf{w}}$
  - update ensemble perturbations $\mathbf{X}'^a = \mathbf{X}'^f \mathbf{W}$

  (both can be combined in a single step)

- Ensemble Kalman & nonlinear filters: Different definitions of
  - weight vector $\tilde{\mathbf{w}}$
  - Transform matrix $\mathbf{W}$

ETKF (Bishop et al., 2001)

- Ensemble Transform Kalman filter
  - Assume Gaussian distributions
  - Transform matrix

$$\mathbf{A}^{-1} = (N-1)\mathbf{I} + (\mathbf{H}\mathbf{X}'^f)^T\mathbf{R}^{-1}\mathbf{H}\mathbf{X}'^f$$

  - Mean update weight vector

$$\tilde{\mathbf{w}} = \mathbf{A}(\mathbf{H}\mathbf{X}'^f)^T\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{H}\overline{\mathbf{x}^f}\right)$$

(depends linearly on $\mathbf{y}$)

  - Transformation of ensemble perturbations

$$\mathbf{W} = \sqrt{(N-1)}\mathbf{A}^{-1/2}\mathbf{\Lambda}$$

(depends only on $\mathbf{R}$, not $\mathbf{y}$)

NETF (Tödter & Ahrens, 2015)

- Nonlinear Ensemble Transform Filter
  - Mean update from Particle Filter weights: for all particles *i*

$$\tilde{w}^i \sim \exp\left(-0.5(\mathbf{y}-\mathbf{H}\mathbf{x}_i^f)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}_i^f)\right)$$

(Nonlinear function of observations $\mathbf{y}$)

  - Ensemble update
    - Transform ensemble to fulfill analysis covariance (like ETKF, but not assuming Gaussianity)
    - Derivation gives

$$\mathbf{W} = \sqrt{N}\left[\mathrm{diag}(\tilde{\mathbf{w}}) - \tilde{\mathbf{w}}\tilde{\mathbf{w}}^T\right]^{1/2}\Lambda$$

($\mathbf{\Lambda}$: mean-preserving random matrix; useful for stability)

Tödter, J. and Ahrens, B. (2015) *Mon. Wea. Rev.* **143**,1347–1367

ETKF-NETF – Hybrid Filter Variants

1-step update (*HSync*)

$$\mathbf{X}^a_{HSync} = \overline{\mathbf{X}}^f + (1-\gamma)\Delta\mathbf{X}_{NETF} + \gamma\Delta\mathbf{X}_{ETKF}$$

- $\Delta\mathbf{X}$: assimilation increment of a filter
- γ: hybrid weight (between 0 and 1; 1 for fully ETKF)

2-step updates

Variant 1 (*HNK*): NETF followed by ETKF

$$\tilde{\mathbf{X}}^a_{HNK} = \mathbf{X}^a_{NETF}[\mathbf{X}^f, (1-\gamma)\mathbf{R}^{-1}]$$

$$\mathbf{X}^a_{HNK} = \mathbf{X}^a_{ETKF}[\tilde{\mathbf{X}}^a_{HNK}, \gamma\mathbf{R}^{-1}]$$

- Both steps computed with increased $\mathbf{R}$ according to γ

Variant 2 (*HKN*): ETKF followed by NETF

Choosing hybrid weight γ

- Hybrid weight shifts filter behavior
- How to choose it?

Possibilities:

- Fixed value
- Adaptive
  - According to which condition?
  - Base on effective sample size $N_{eff} = \sum_i 1/(w^i)^2$

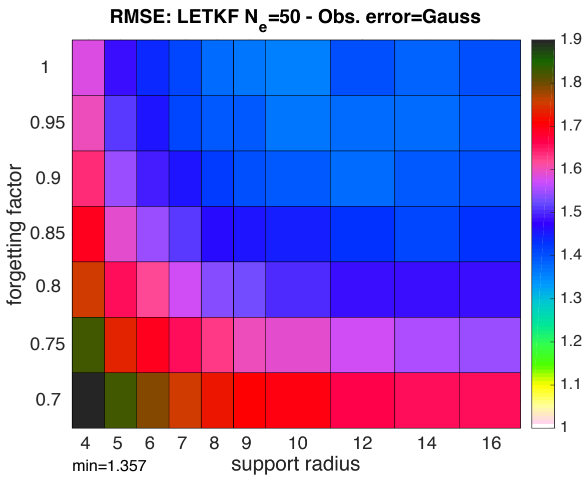
    set

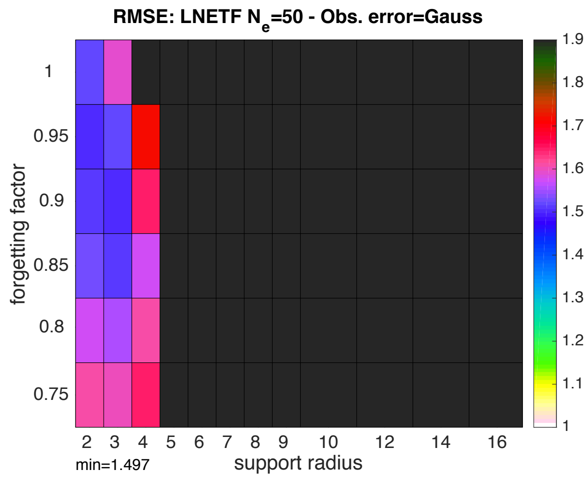
    $$\gamma_{adap} = 1 - N_{eff}/N$$

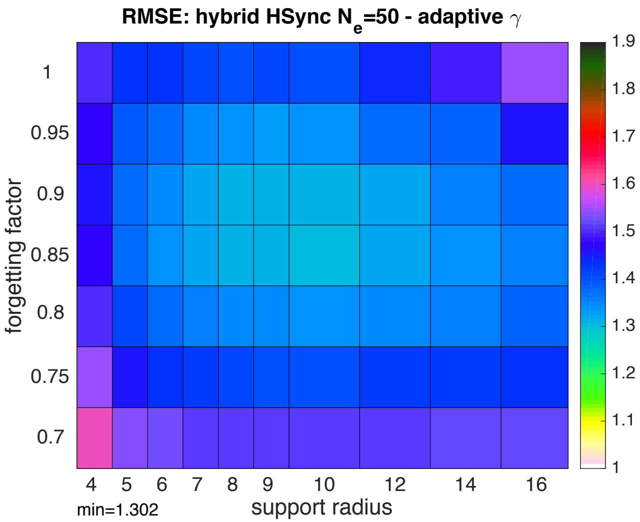
    (close to 1 if N_{eff} small, i.e. small contribution of NETF)

Test with Lorenz-96 Model (ensemble size N=50)

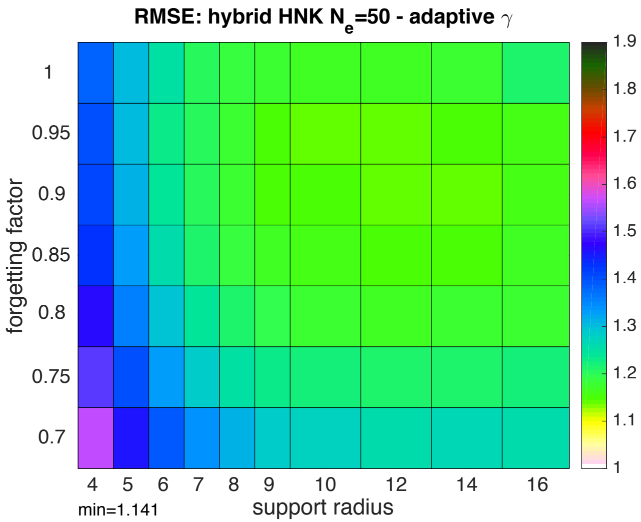
Ensemble size N=50



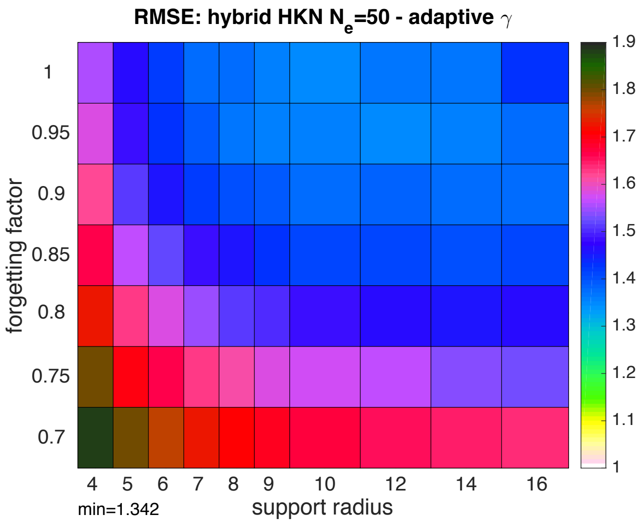




4% improvement



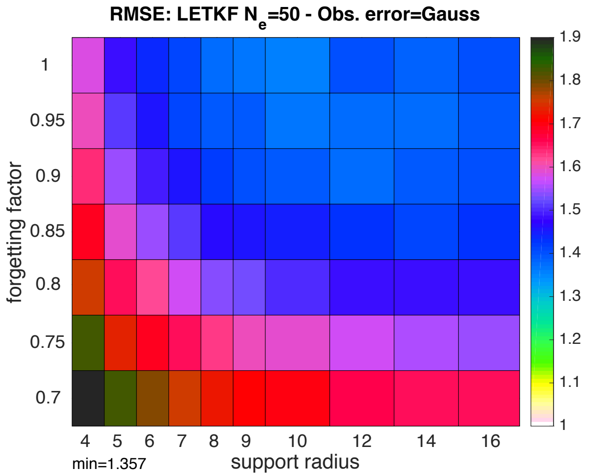
16% improvement



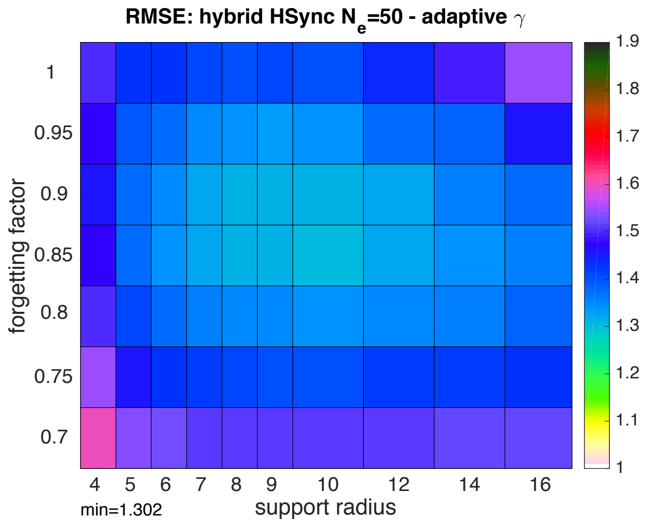
1% improvement

Test with Lorenz-96 Model (ensemble size N=50)

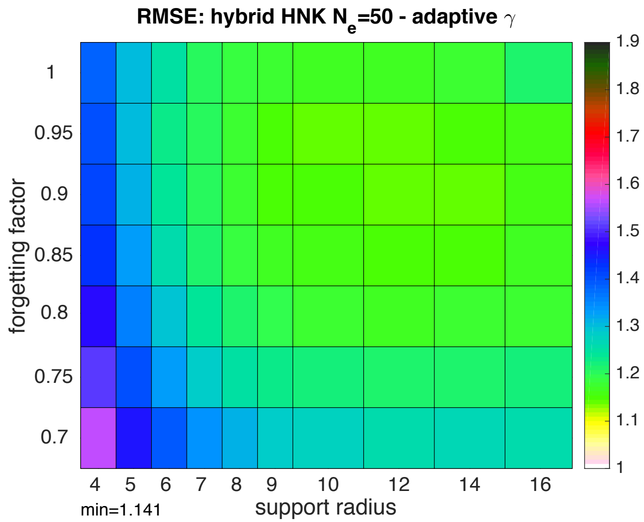
Ensemble size N=50



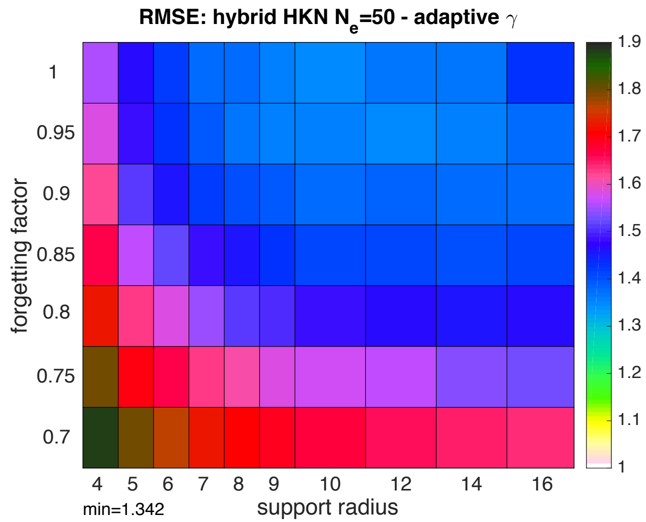
- All hybrid variants improve estimates compared to LETKF & NETF
- Dependence on forgetting factor & localization radius like LETKF
- Similar optimal localization radius
- Largest improvement for variant HNK (NETF before LETKF)
- Currently testing in a larger model ...



4% improvement



16% improvement



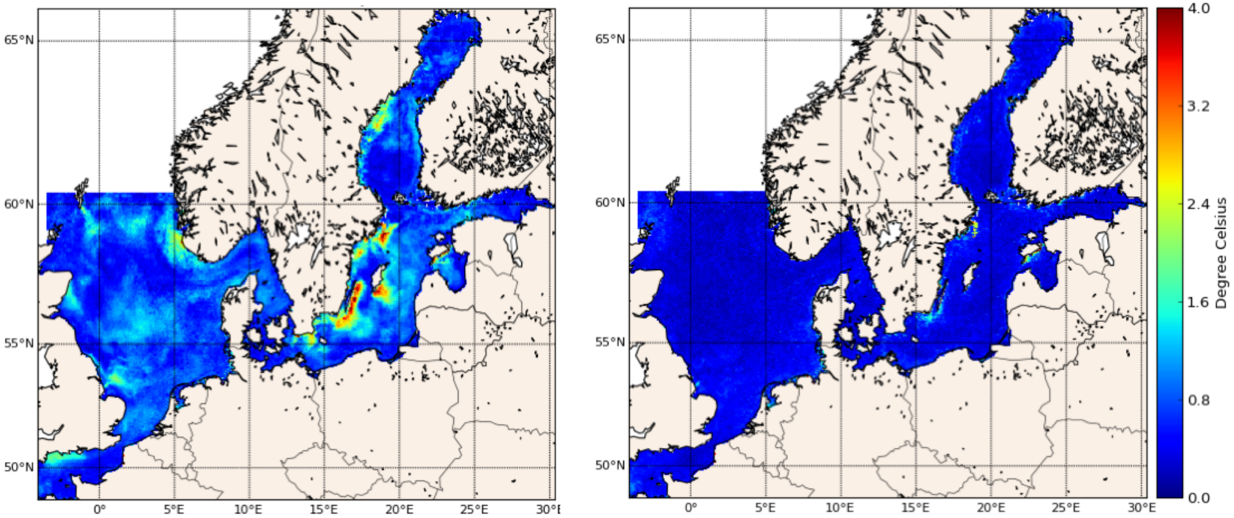
1% improvement

Applications

Assimilation effect on Temperature (September 2012)

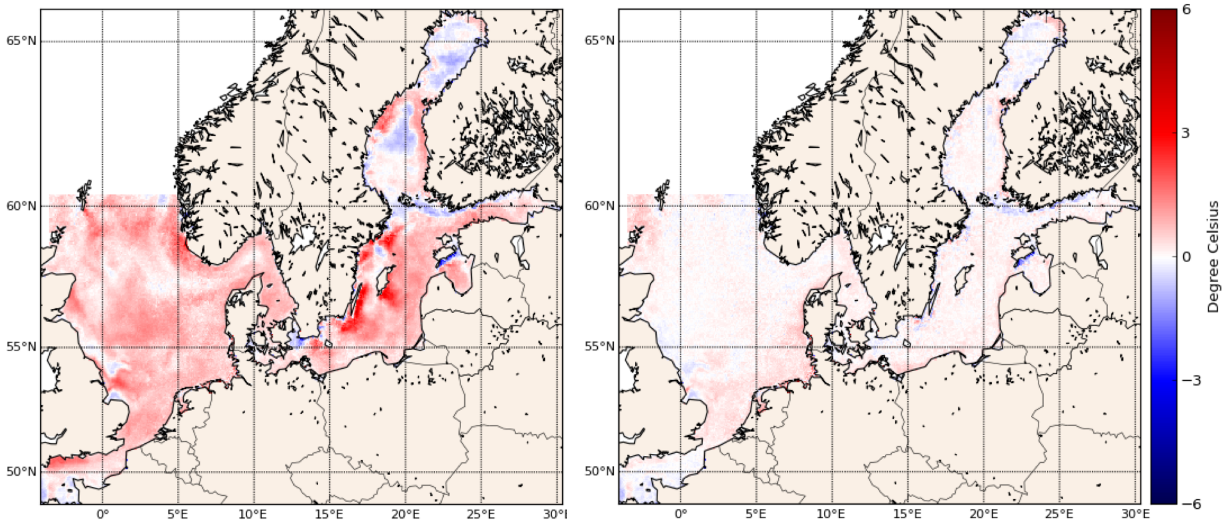
RMS (root-mean-square) deviation

Free run — Assimilation (analysis)



Mean deviation (observation – model)

Free run — Assimilation (analysis)



Assimilate surface temperature each 12 h

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

Reduce RMS deviation and mean deviation (bias)

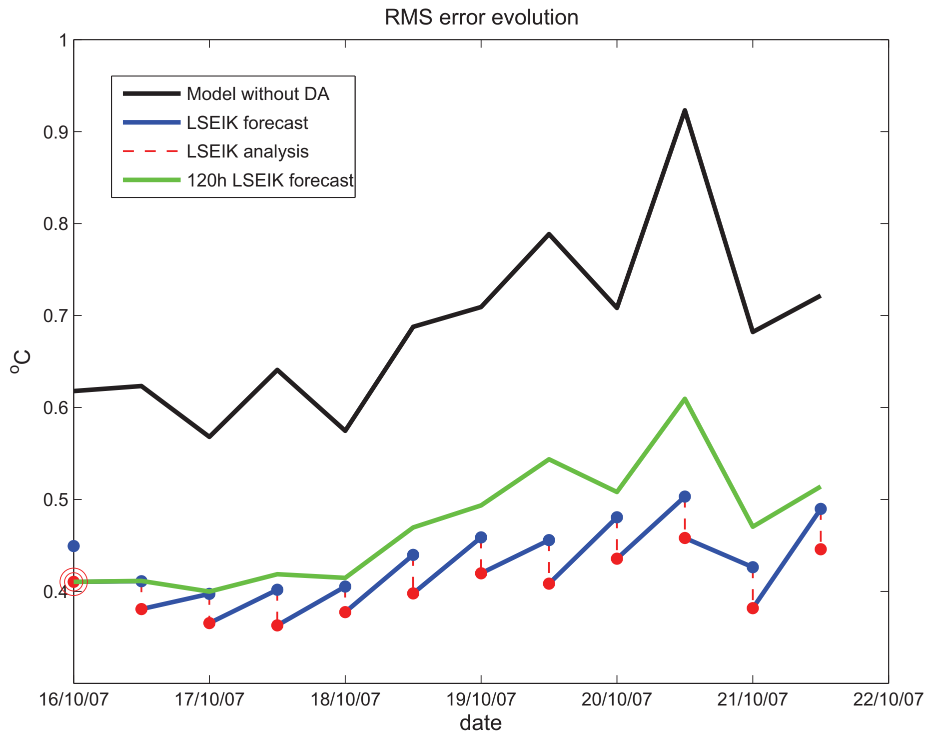
➜ necessary effect

Improving forecasts

DeMarine

Impact of Assimilation for temperature forecasts

(North & Baltic Seas)



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

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

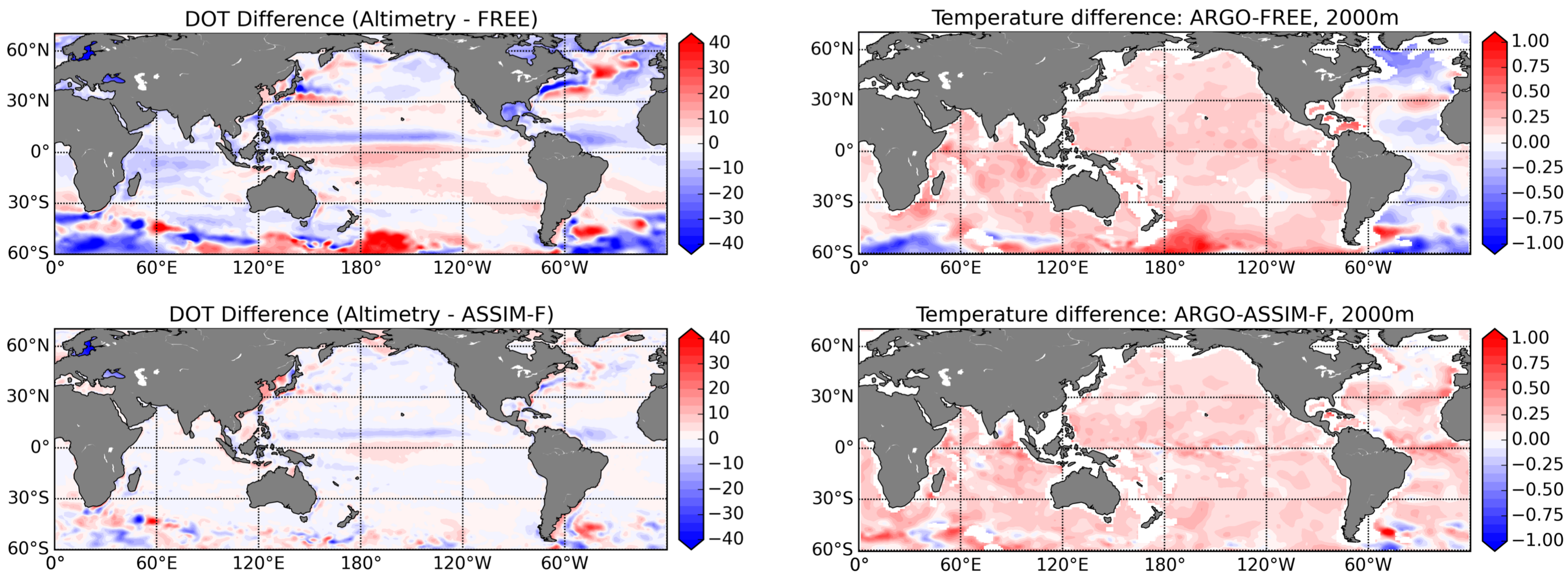
AWI

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

Bias Estimation

Example: Chlorophyll bias of a biogeochemical model

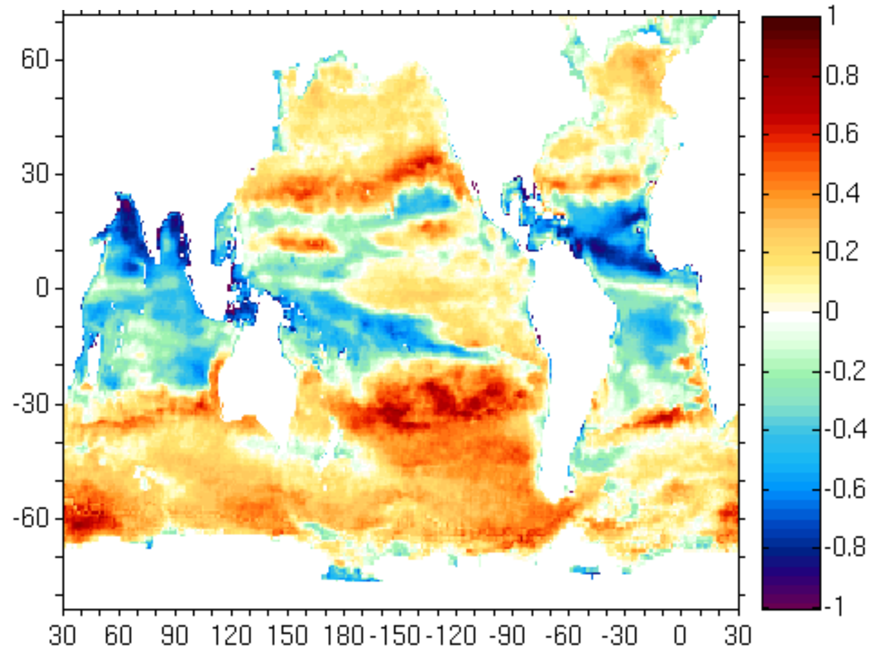
Bias = systematic errors

- *un-biased system*:
  random fluctuation around true state

- *biased system:*
  systematic over- and underestimation
  (common situation with real data)

- *Bias estimation*:
  Separate random from systematic deviations

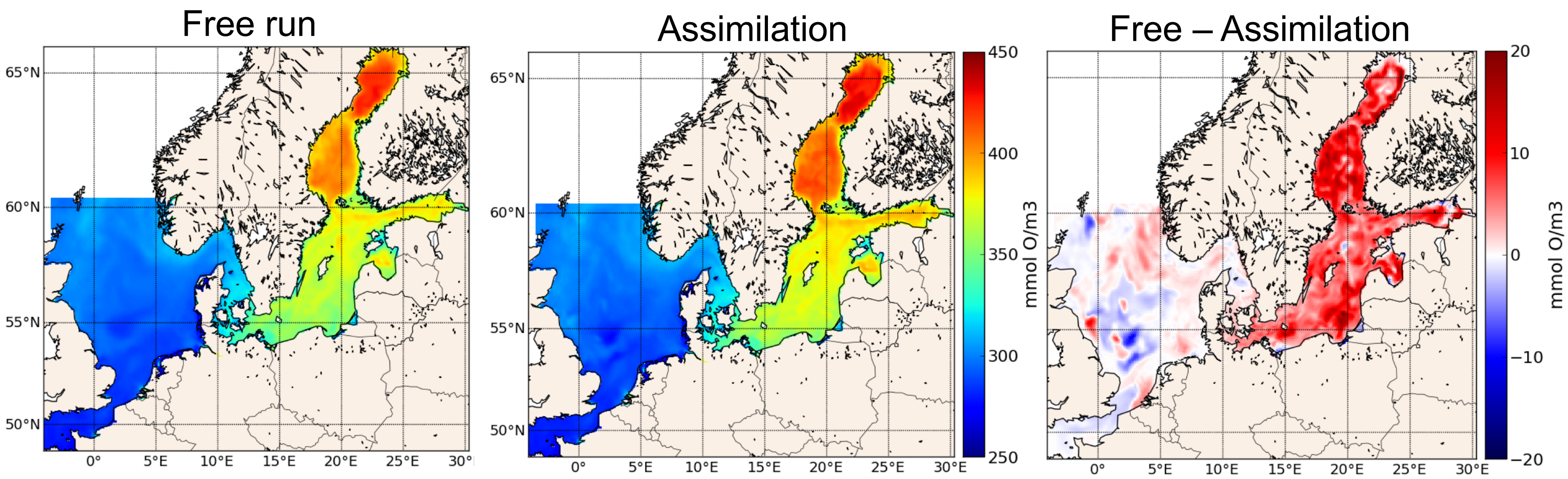
Logarithmic bias estimate
April 15, 2004



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

Biogeochemistry: Coupled data assimilation effect

Surface oxygen mean for May 2012 (as mmol O / m^3)

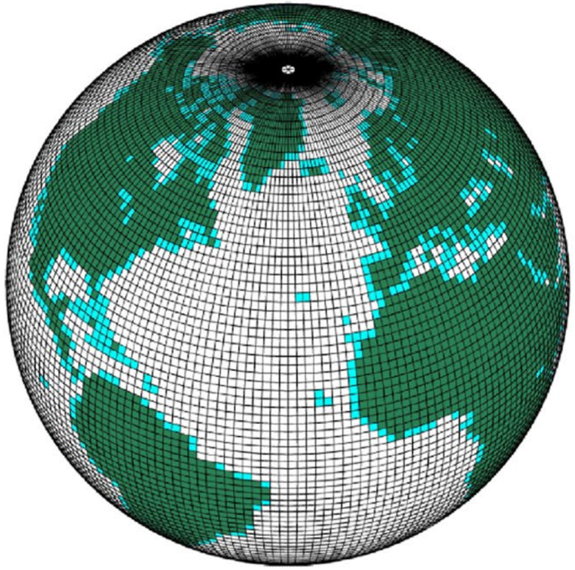


Coupled data assimilation case: physics and biogeochemistry

- Assimilate satellite sea surface temperature observations
- Assimilation directly changes Oxygen and other biogeochemical variables (strongly-coupled assimilation)

Assimilation into coupled model: AWI-CM

Atmosphere

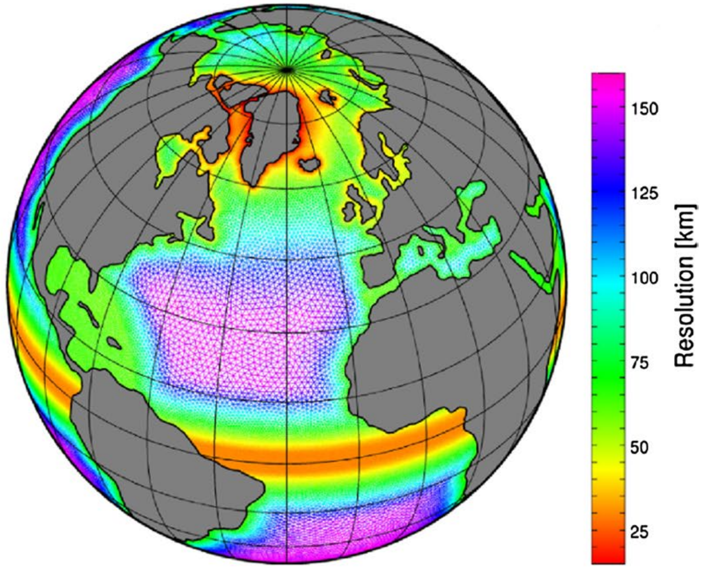


OASIS3-MCT

fluxes

ocean/ice state

Ocean



Atmosphere
- ECHAM6
- JSBACH land

Coupler library
- OASIS3-MCT

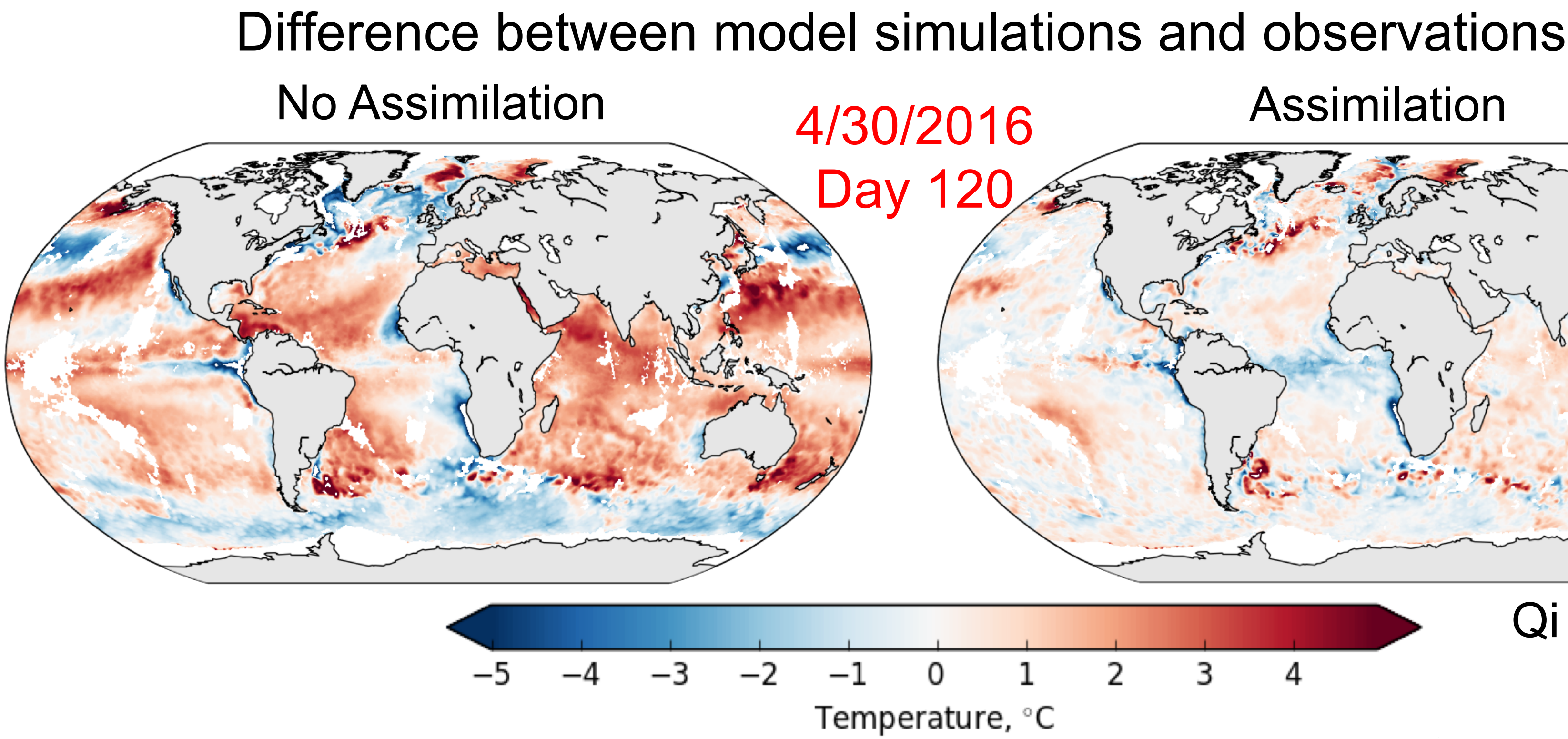
Ocean
- FESOM
- includes sea ice

Two separate executables for atmosphere and ocean

Goal: Develop data assimilation methodology for cross-domain assimilation ("strongly-coupled")

Assimilation Effect on Surface Temperature

Assimilate subsurface temperature profile data



Qi Tang @ AWI

- Also subsurface temperature is improved

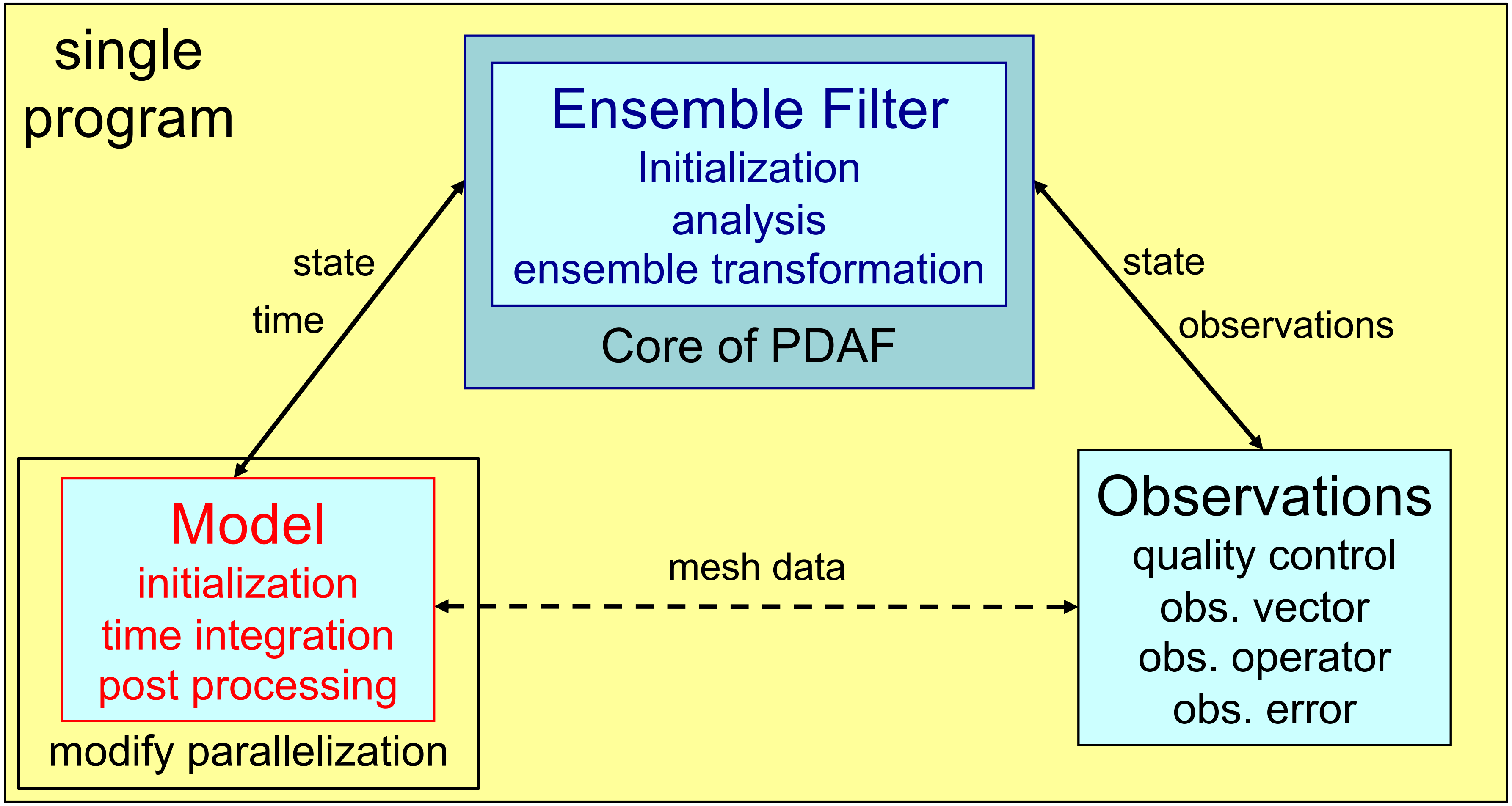
Current work

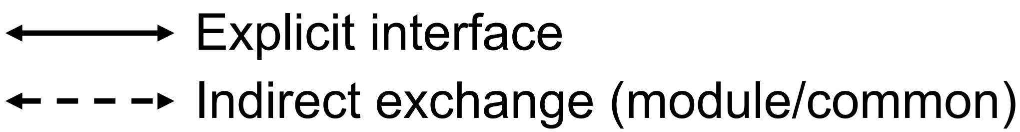
- Assess effect on atmosphere
- Final aim: strongly-coupled assimilation
  (e.g. assimilate oceanic observation into atmosphere)

Software

Components of an Assimilation System

PDAF Parallel Data Assimilation Framework





L. Nerger, W. Hiller, Computers & Geosciences 55 (2013) 110-118

PDAF: A tool for data assimilation

Parallel
Data
Assimilation
Framework

PDAF - Parallel Data Assimilation Framework

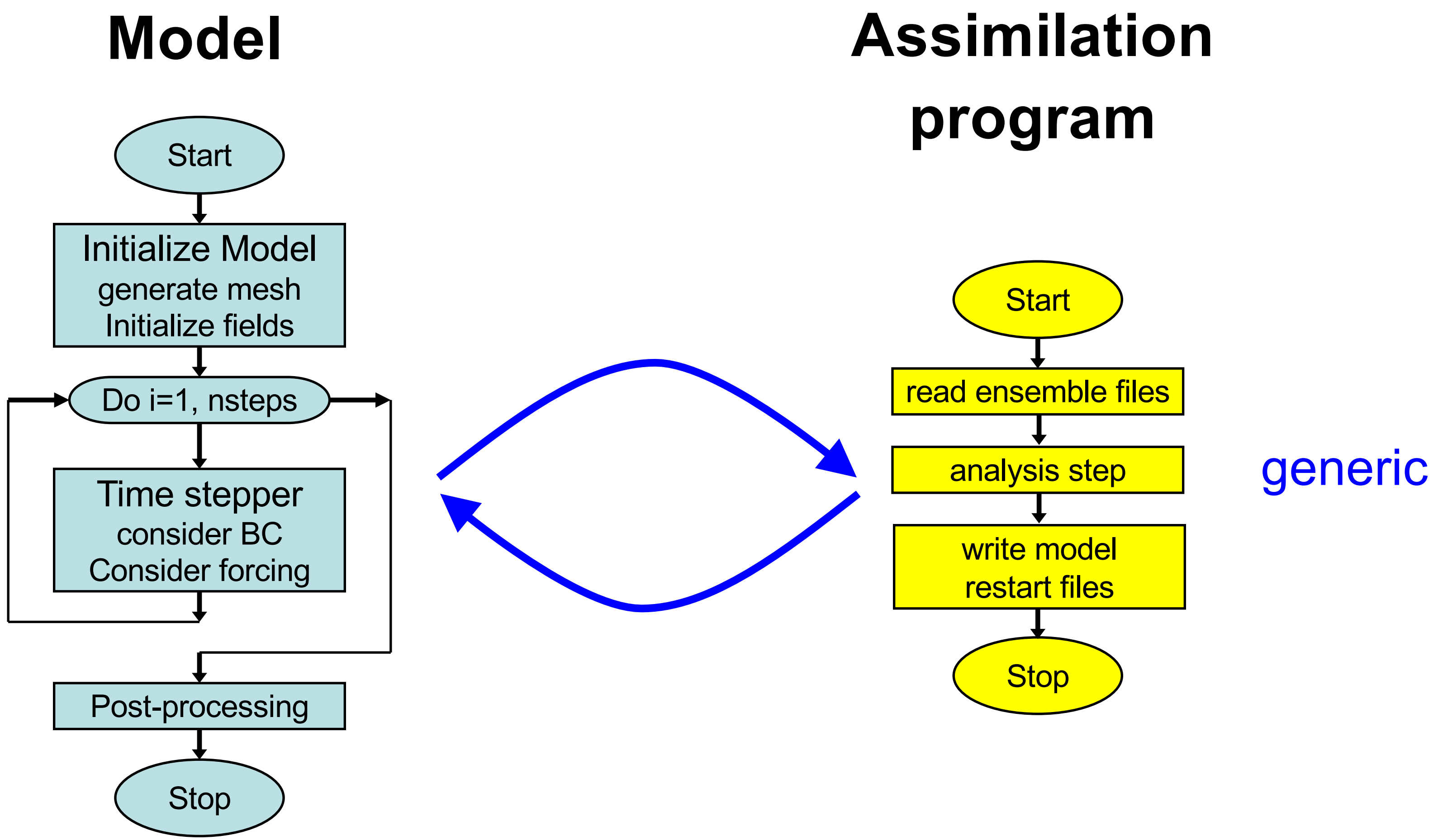
- a program library for ensemble data assimilation
- provide support for parallel ensemble forecasts
- provide fully-implemented & parallelized filters and smoothers (EnKF, LETKF, NETF, EWPF … easy to add more)
- easily useable with (probably) any numerical model (applied with NEMO, MITgcm, FESOM, HBM, TerrSysMP, …)
- run from laptops to supercomputers (Fortran, MPI & OpenMP)
- first public release in 2004; continuous further development
- ~370 registered users; community contributions

Open source:
Code, documentation & tutorials at

http://pdaf.awi.de

L. Nerger, W. Hiller, Computers & Geosciences 55 (2013) 110-118

Offline coupling – separate programs



For each ensemble state

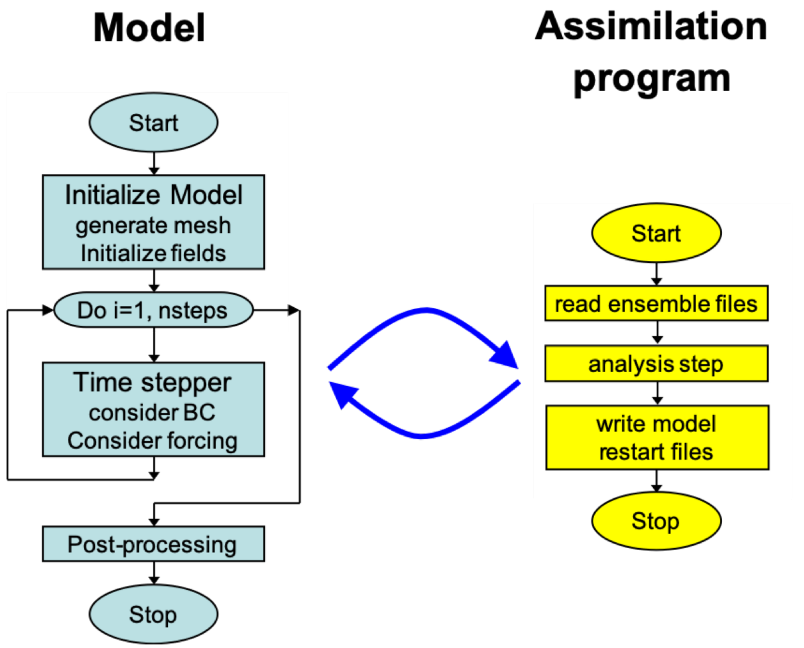
- Initialize from restart files
- Integrate
- Write restart files

- Read restart files (ensemble)
- Compute analysis step
- Write new restart files

Offline coupling - Efficiency

Offline-coupling is simple to implement but can be very ineffcient

Example:
Timing from atmosphere-ocean coupled model (AWI-CM) with daily analysis step:



Model startup:	95 s
Integrate 1 day:	28 s
Model postprocessing:	14 s

overhead

Analysis step: 1 s

Restarting this model is ~3.5 times more expensive than integrating 1 day

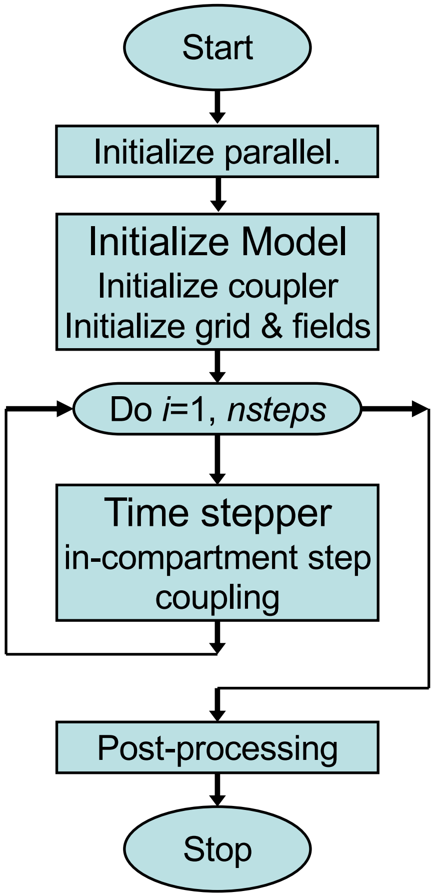
➜ avoid this for data assimilation

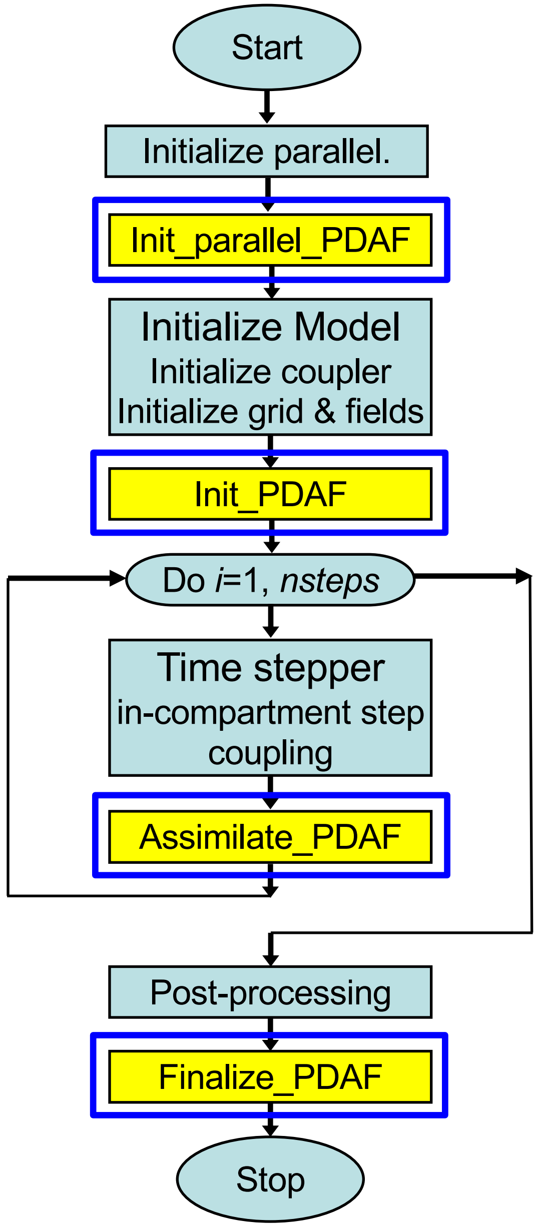
Extending a Model for Data Assimilation

PDAF Parallel Data Assimilation Framework

Model

single or multiple executables





Extension for data assimilation

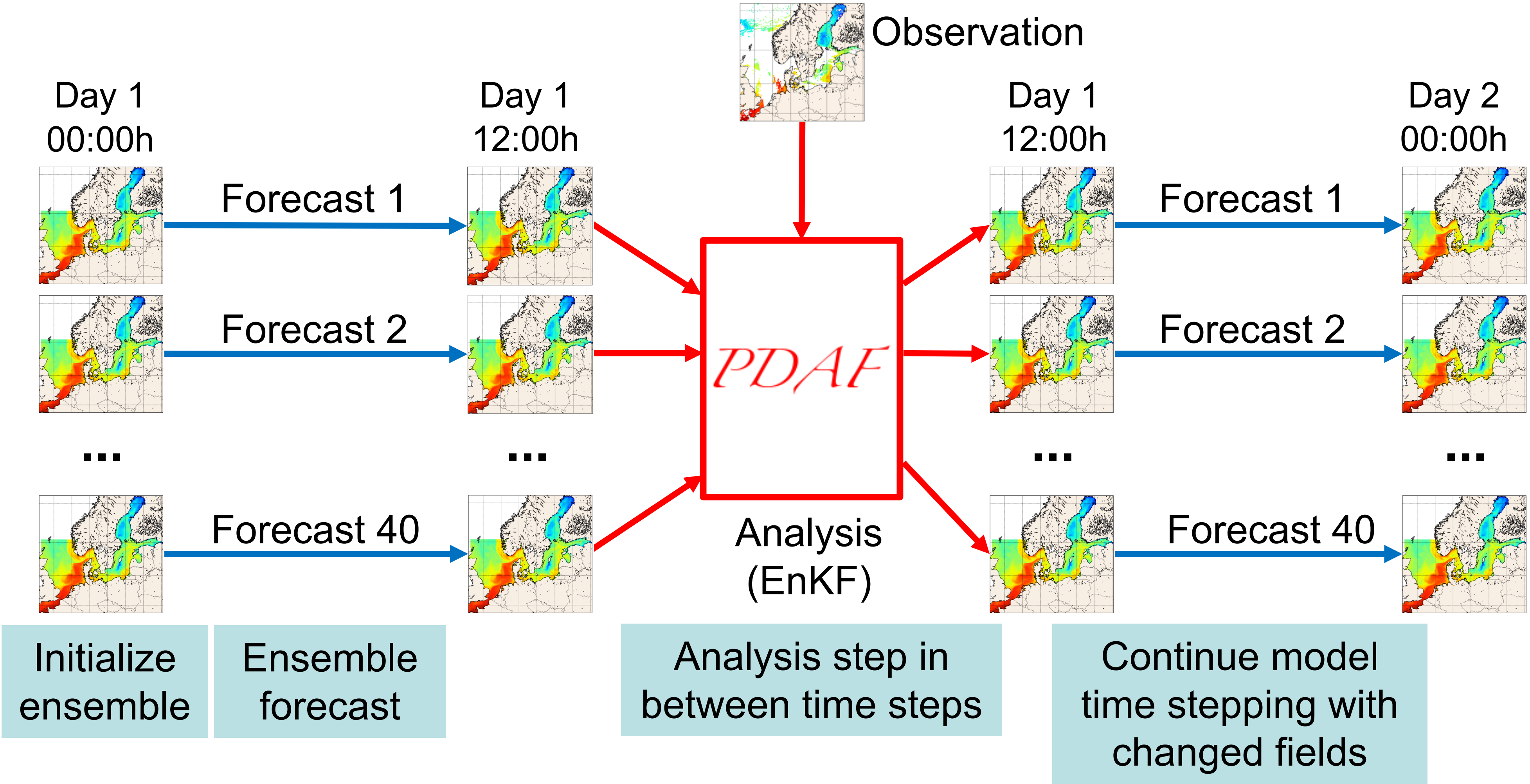
plus:
Possible model-specific adaption

revised parallelization enables ensemble forecast

Augmenting a Model for Data Assimilation

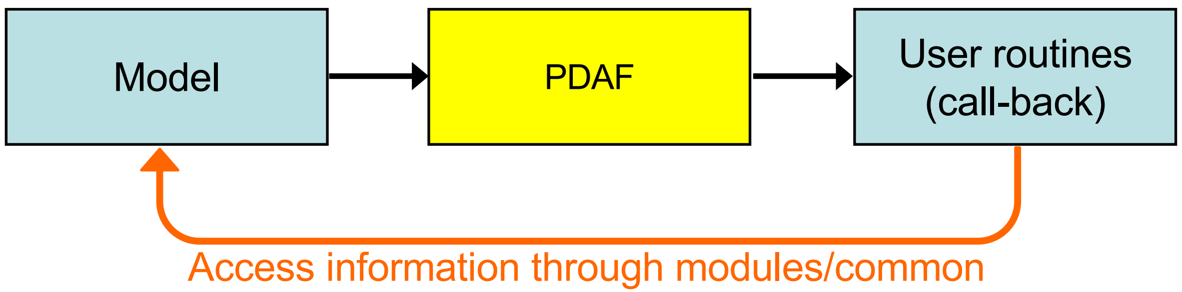
Couple PDAF (Parallel Data Assimilation Framework) with model

- Modify model to simulate ensemble of model states
- Insert correction step (analysis) to be executed at prescribed interval
- Run model as usual, but with more processors and additional options



PDAF interface structure

- Interface routines call PDAF-core routines
- PDAF-core routines call case-specific routines provided by user (included in model binding set)
- User-supplied call-back routines for elementary operations:
  - field transformations between model and filter
  - observation-related operations
- User supplied routines can be implemented as routines of the model (for MITgcm: Fortran-77 fixed-form source code)



PDAF: User-friendliness

Assumption: Users know their model

➜ let users implement assimilation system in model context

For users, model is not just a forward operator

➜ let users extend their model for data assimilation

Keep simple things simple:

- Define subroutine interfaces to separate model and assimilation based on arrays
- No object-oriented programming
(most models don't use it; most model developers don't know it; not many objects would be involved)
- Users directly implement observation-specific routines
(no indirect description of e.g. observation layout)

Example: Value of Efficient Software

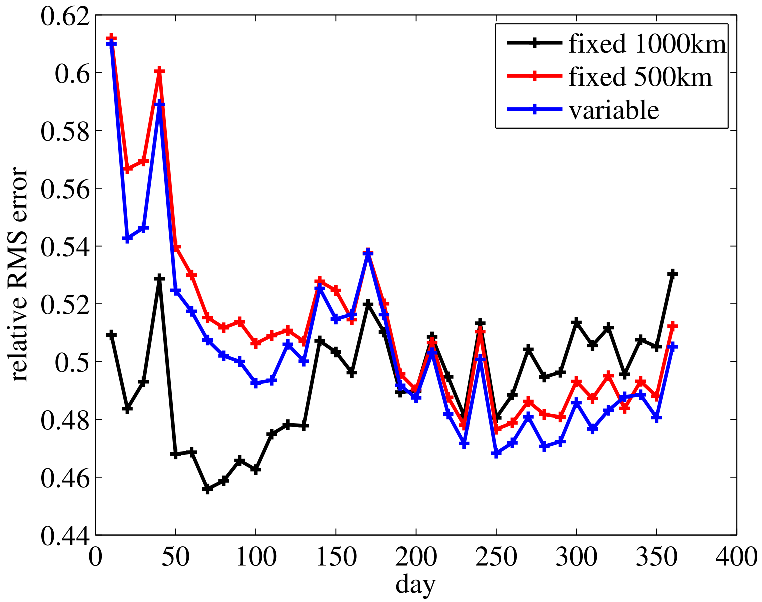
PDAF Parallel Data Assimilation Framework

Adaptive Localization (Kirchgessner et al, 2012)

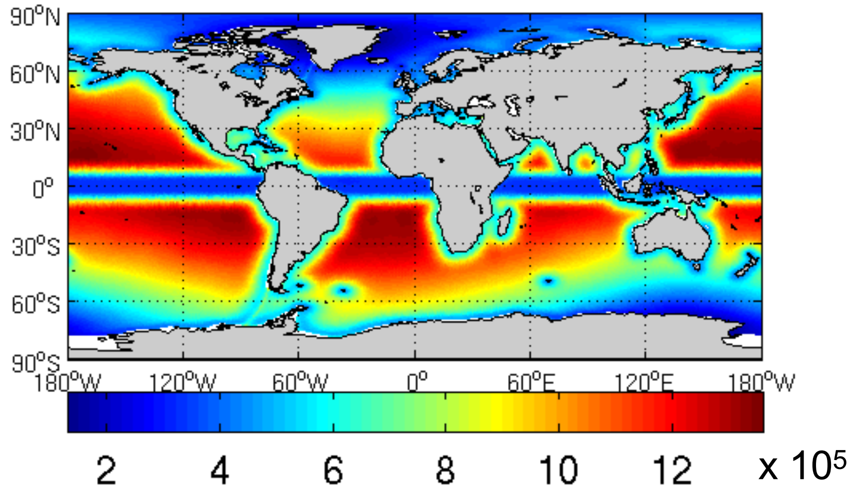
- Original study done with small models (Lorenz-96, shallow water)
- Paper reviewer asked to apply it with full-scale forecast model
- FESOM with PDAF was fully coded without adaptivity
  - ➢ Update PDAF library (just when recompiling)
  - ➢ Adding adaptivity routine and running experiment

$\left.\right\}$ 1 day!

Error-reduction of sea surface height



Localization radius [meter]



Kirchgessner, Nerger, Bunse-Gerstner, Mon. Weather Rev., 142 (2012) 2165-2175

Summary

Ensemble data assimilation

- Quantitative combination of model and observational data
- Improve observed and non-observed fields, fluxes, parameters, and predictions

PDAF simplifies the implementation and application of data assimilation

- Get faster to the application and results

Tomorrow's Tutorial:

- Implementation of PDAF with simple model
- Experiments with an ensemble Kalman filter

References

- http://pdaf.awi.de
- Nerger, L., Hiller, W. *Software for Ensemble-based DA Systems – Implementation and Scalability*. Computers and Geosciences 55 (2013) 110-118
- Nerger, L., Hiller, W., Schröter, J.(2005). *PDAF - The Parallel Data Assimilation Framework: Experiences with Kalman Filtering*, Proceedings of the Eleventh ECMWF Workshop on the Use of High Performance Computing in Meteorology, Reading, UK, 25 - 29 October 2004, pp. 63-83.