EGU General Assembly 2025

30 Apr 2025

PDAF Parallel Data Assimilation Framework

SC3.14 Getting Started with Data Assimilation: Theory and Application

Qi Tang[1,2], Lars Nerger[3], Armin Corbin[4], Nabir Mamnun[5], Yumeng Chen[6]

[1]University of Neuchâtel, Centre for Hydrogeology and Geothermics (CHYN), Switzerland

[2]University of Basel, Centre for Hydrogeology and Geothermics (CHYN), Switzerland

[3]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Germany

[4]University of Bonn, Institute for Geodesy and Geoinformation, Astronomical, Physical and Mathematical Geodesy Group, Germany

[5]Imperial College London, Physics, United Kingdom of Great Britain – England, Scotland, Wales

[6]University of Reading, National Centre for Earth Observation, Department of Meteorology, United Kingdom of Great Britain

unine Université de Neuchâtel

Universität Basel

AWI

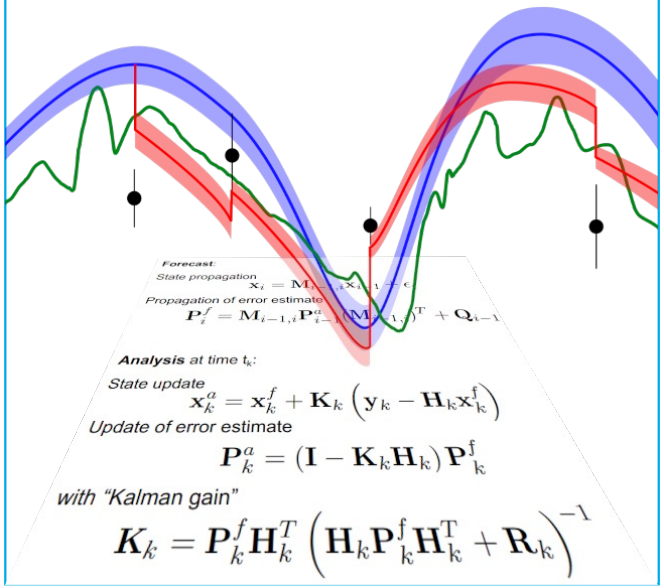
UNIVERSITÄT BONN

University of Reading

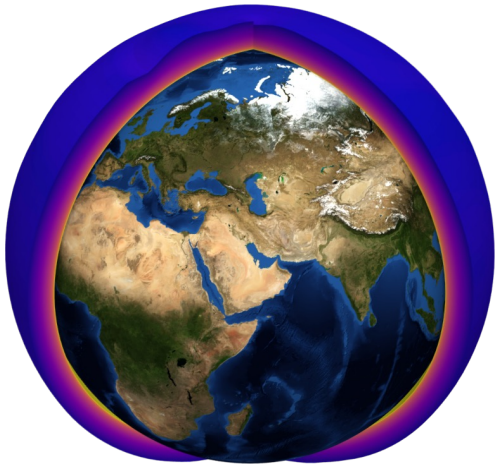
Schedule

PDAF Parallel Data Assimilation Framework

I Theory (45 min)



II Applications (15 min)



III Hands-on (45 min)

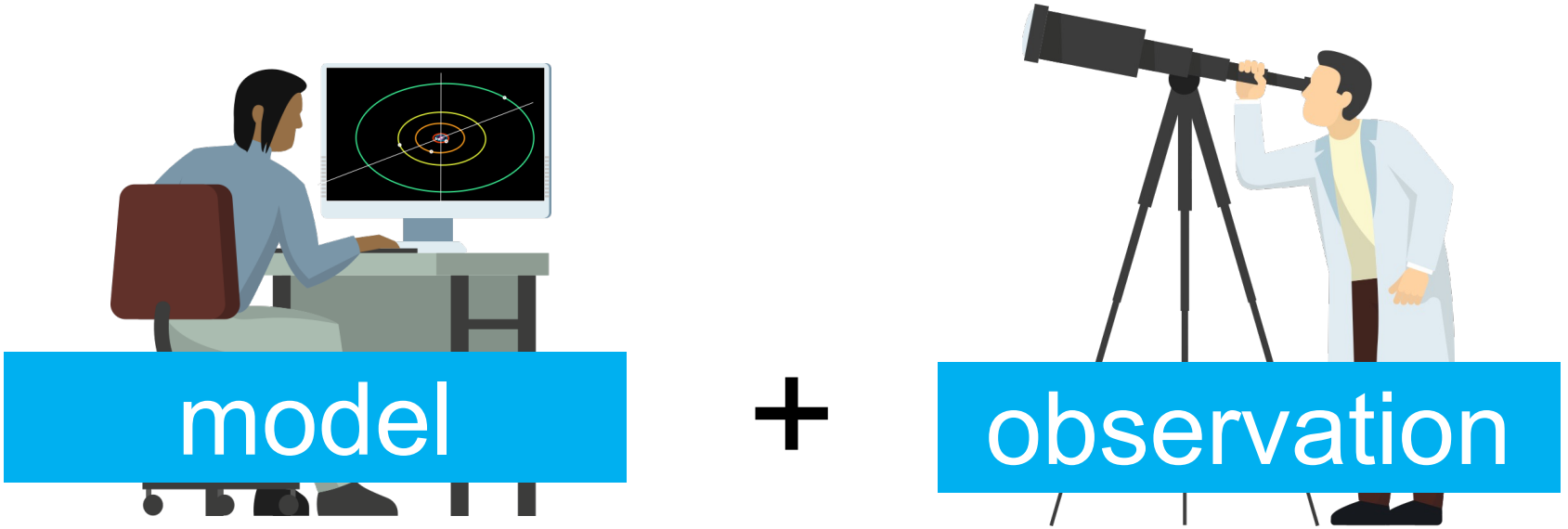
PDAF Parallel Data Assimilation Framework

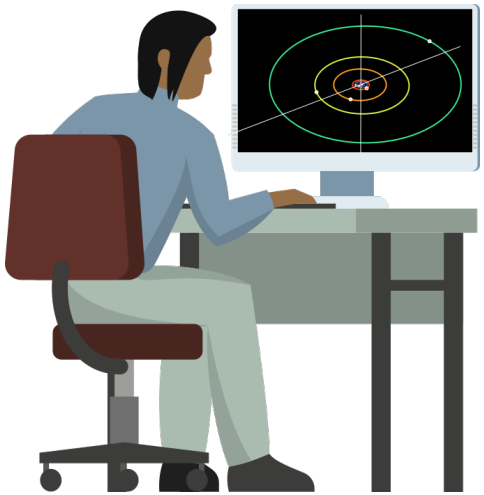
I Theory

Data Assimilation (DA)

PDAF Parallel Data Assimilation Framework

*Data assimilation (DA) is the science of **combining observations** of a system, **including their uncertainty**, with estimates of that system from a dynamical **model**, including its **uncertainty**, to obtain a new and more accurate description of the system including an uncertainty estimate of that description.* Vetra-Carvalho et al. (2018)



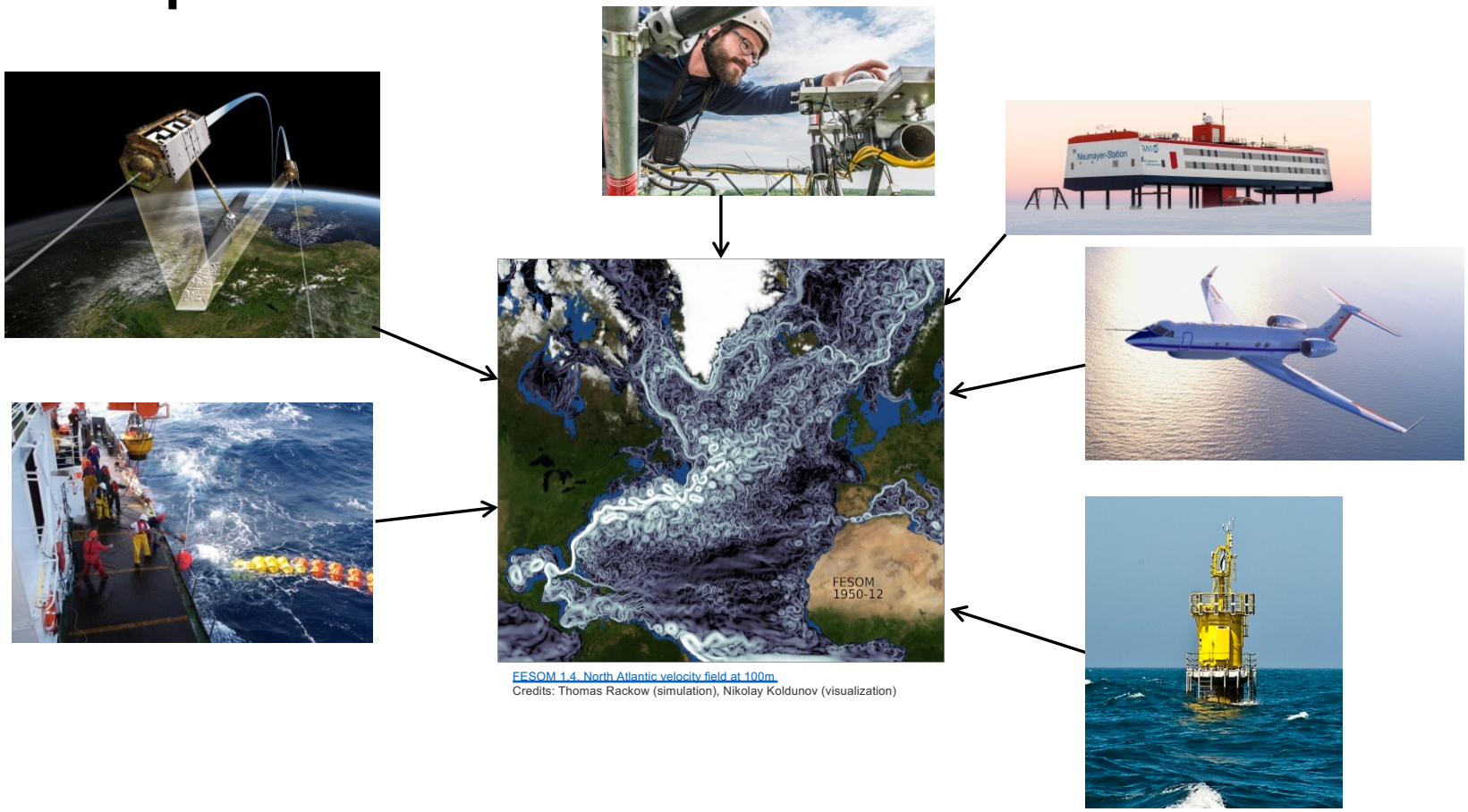


PDAF Parallel Data Assimilation Framework

model	observation
- **idealized** representation of a system	+ measurements of “reality”
+ **complete coverage**: often located on a grid or mesh, high temporal and spatial resolution	- **Incomplete:** sparse, discrete, data gaps, irregular sampling, missing state variables
	- outliers
quantifiable systematic and random errors	

Example: Ocean

PDAF Parallel Data Assimilation Framework



FESOM 1.4, North Atlantic velocity field at 100m.
Credits: Thomas Rackow (simulation), Nikolay Koldunov (visualization)

Example: Climate

PDAF Parallel Data Assimilation Framework

Image: DWD's Climate Services Observing – Modelling – Consulting

Requirements for DA

PDAF Parallel Data Assimilation Framework

1. Model
   - With some skill

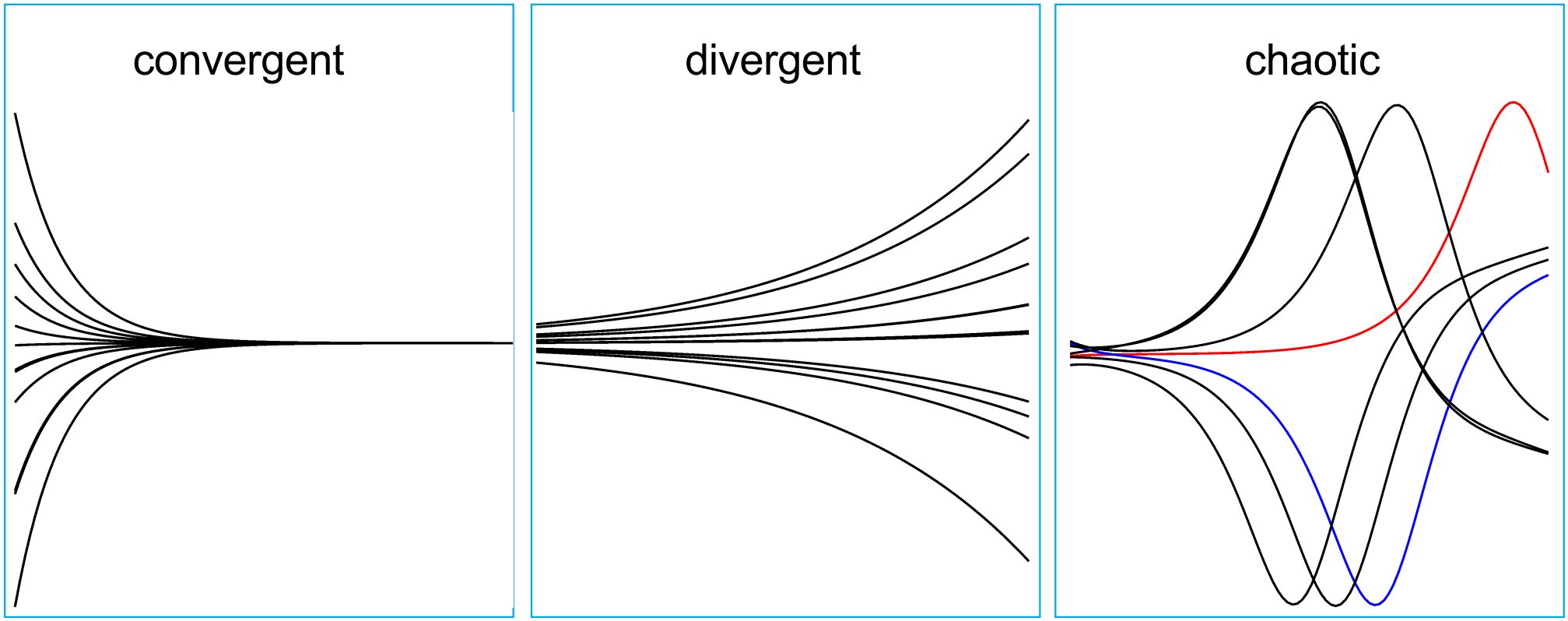
2. Observations
   - With finite errors
   - Related to model fields

3. Data assimilation method

Dynamical System

PDAF Parallel Data Assimilation Framework

The future state depends on the present state



Model Operator

PDAF Parallel Data Assimilation Framework

state

model errors

$$x(t) = \mathcal{M}_{s,t}(x(s)) + \eta(t)$$

model/forward operator: propagates state from time s to t

Linearized Operator: $$M_{s,t} = \left.\frac{\partial \mathcal{M}_{s,t}(x)}{\partial x}\right|_{x=x(s)}$$

Observation Operator

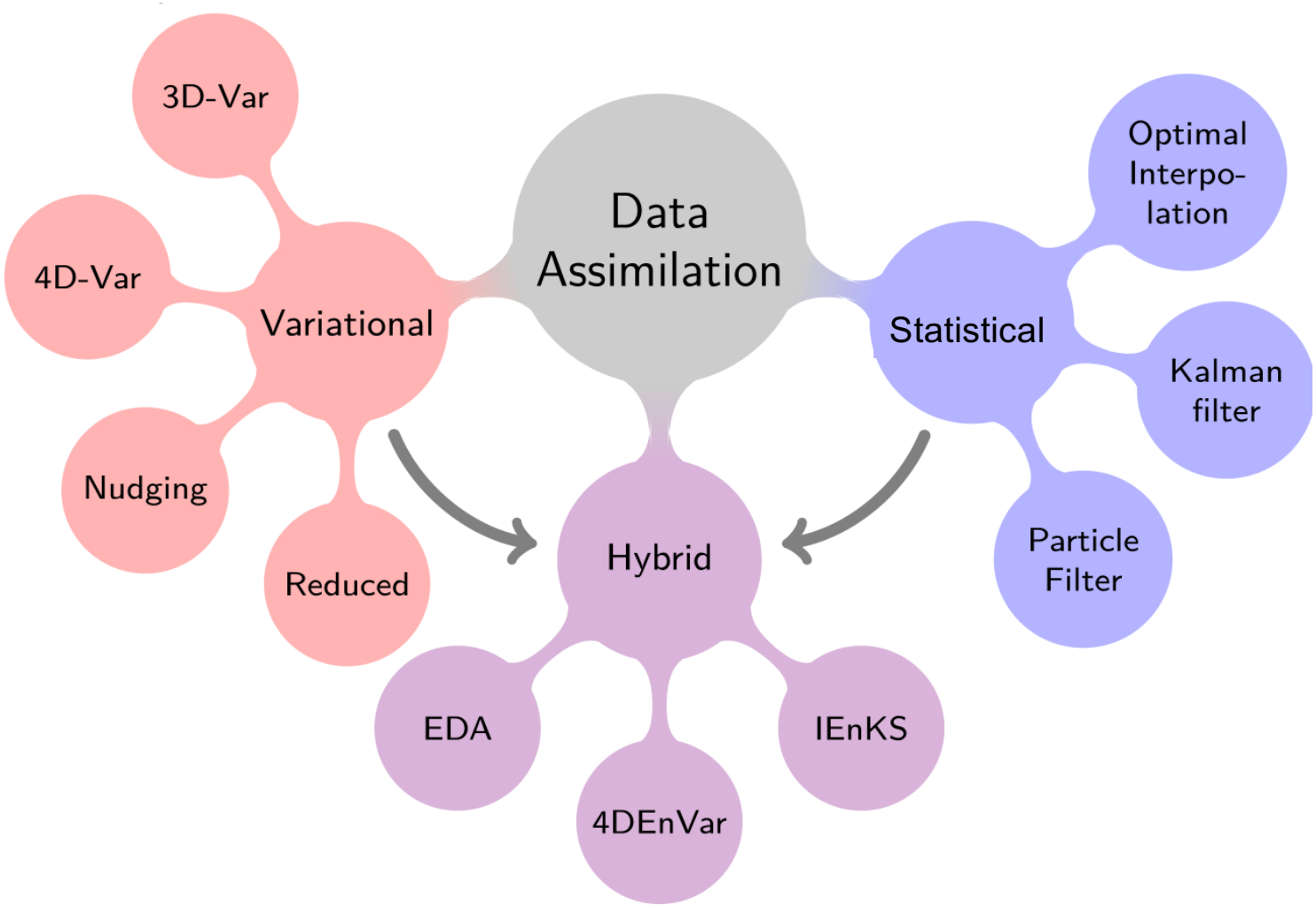
PDAF Parallel Data Assimilation Framework

observations state observation errors

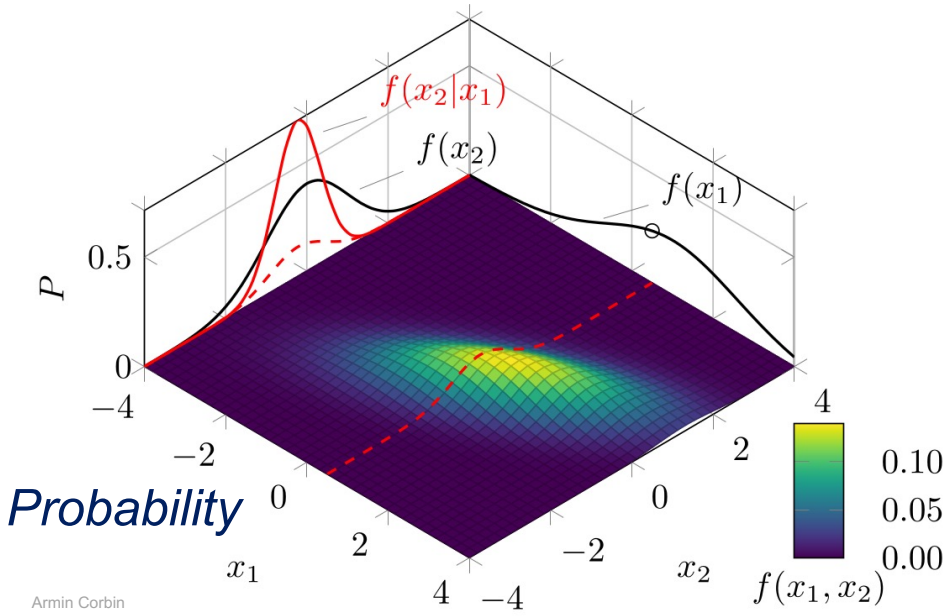
$$y(t) = \mathcal{H}\left(x(t)\right) + \varepsilon(t)$$

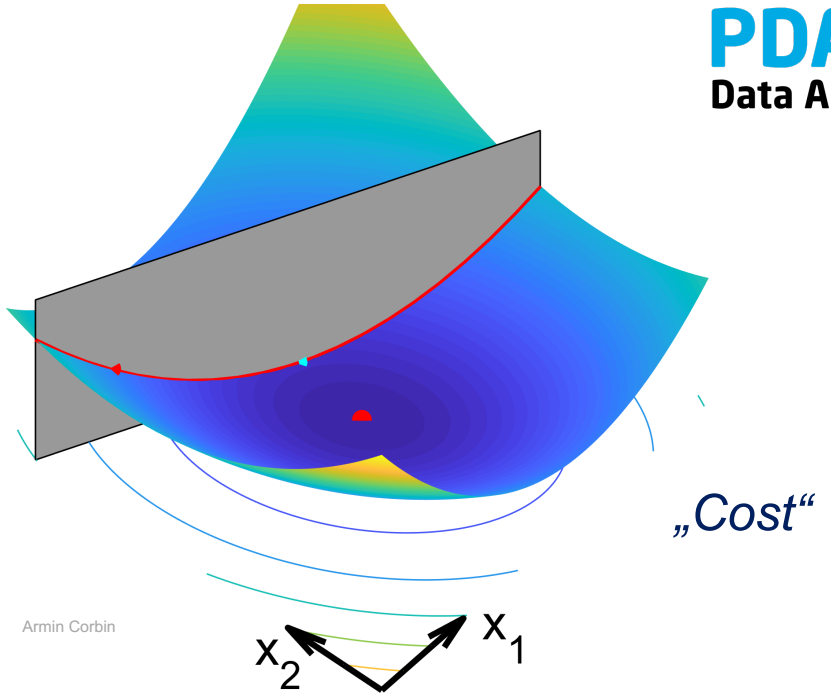
observation operator: maps state to observation

Linearized Operator: $$H = \left.\frac{\partial \mathcal{H}(x)}{\partial x}\right|_{x=x(t)}$$



Adopted from Asch et al. (2016)



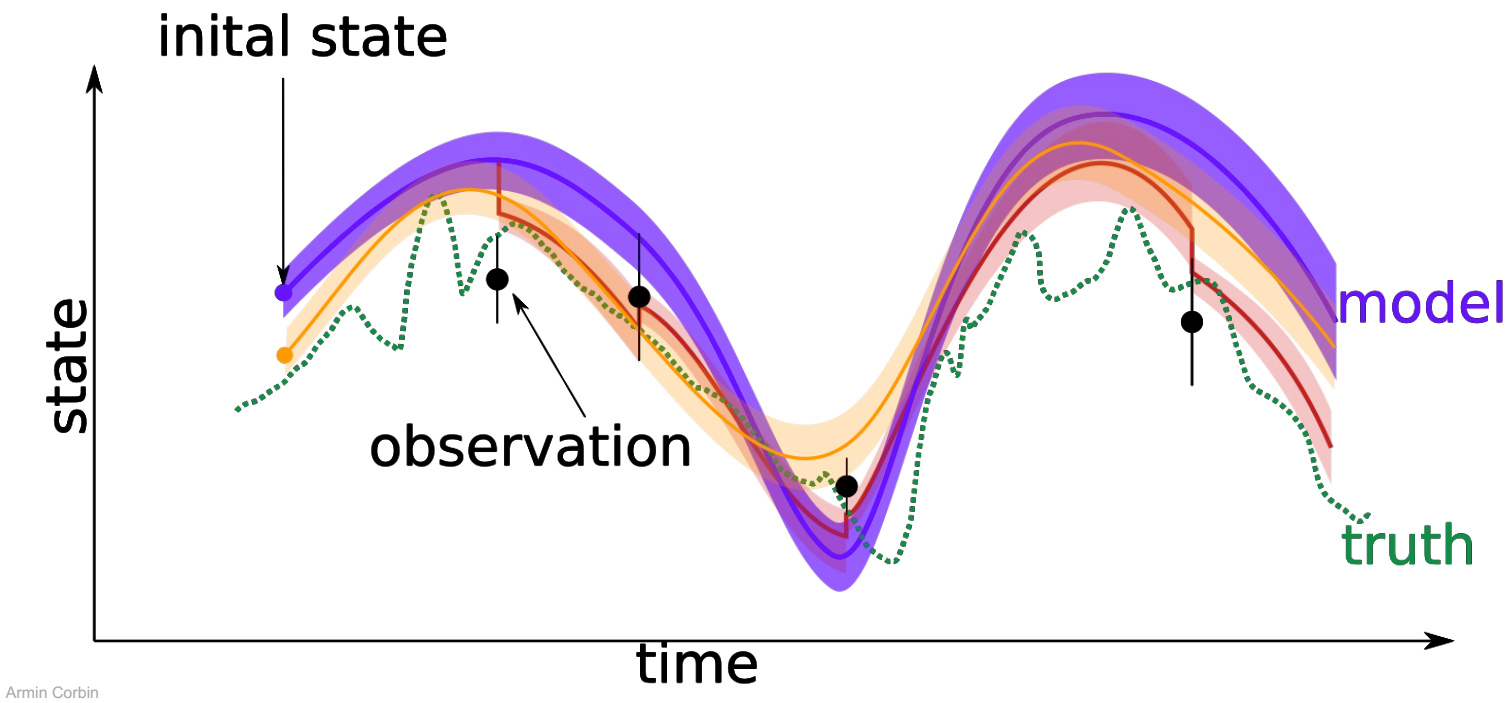


PDAF Parallel Data Assimilation Framework

Statistical	Variational
Estimation theory	Optimal control theory
Maximization of probability density (minimization of variance)	Minimization of cost function (e.g. Gauss-Newton, conjugate gradient)

Filter and Smoother

PDAF Parallel Data Assimilation Framework



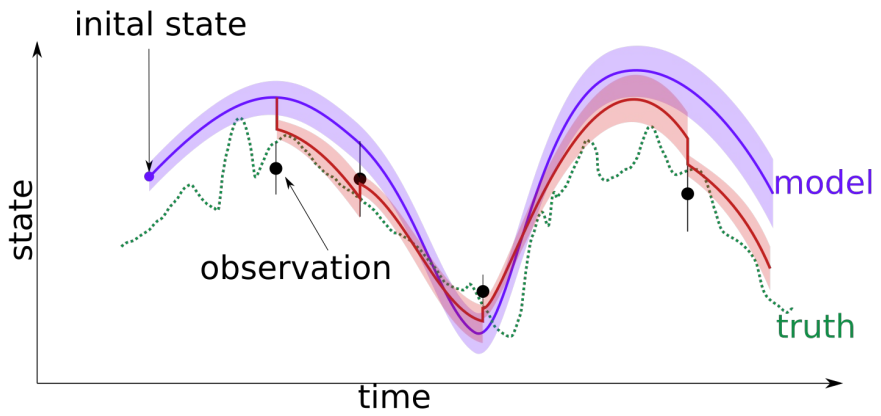
PDAF Parallel Data Assimilation Framework

	statistical	variational
filter	Kalman filter Particle filter	3D VAR
smoother	Kalman smoother Particle smoother	4D VAR

Optimization

PDAF Parallel Data Assimilation Framework

Kalman Filter



forecast

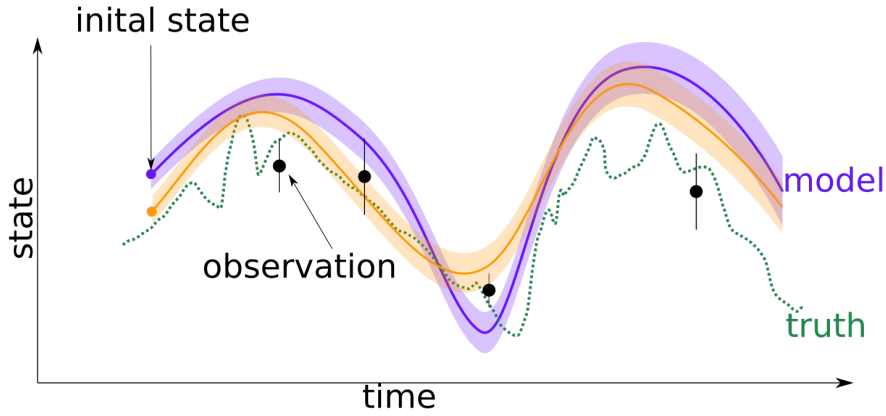
observations

$$J(x) = \left| x - x_i^f \right|^2_{\left(P_i^f\right)^{-1}} + \left| \mathcal{H}(x) - y_i^o \right|^2_{(R_i)^{-1}}$$

state

observation operator

4D Var



$$J(x_0) = \left| x_0 - x_0^b \right|^2_{(P_0)^{-1}} + \sum_{i=0}^{n} \left| \mathcal{H}(x_i) - y_i^o \right|^2_{(R_i)^{-1}}$$

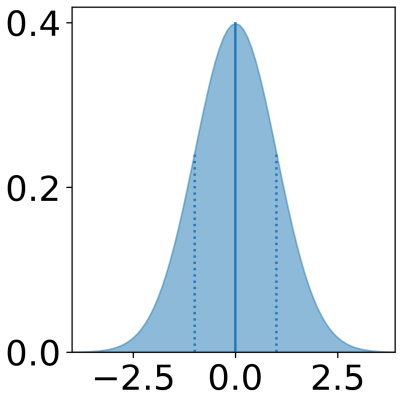
Kalman filter is optimal

PDAF Parallel Data Assimilation Framework

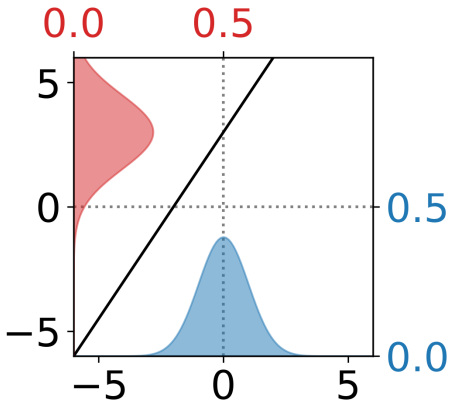
Optimal: state is **unbiased** and has **minimal variance**

Assumptions:

1. everything is Gaussian



2. model and observation operator are linear



3. model errors are not correlated with state or observation errors

$$\begin{bmatrix} R & 0 \\ 0 & P \end{bmatrix}$$

Kalman Filter

PDAF Parallel Data Assimilation Framework

1. Forecast/Prediction

State propagation

$$x_i = M_{i-1,i} x_{i-1} + \varepsilon_i$$

Propagation of error estimate

$$P_i^f = M_{i-1,i} P_{i-1}^a M_{i-1,i}^T + Q_{i-1}$$

1. M and P explicitly required
2. Linear Transformation
3. Scales poorly with the size of the problem

2. Analysis/Update at time t_k

State update

$$x_k^a = x_k^f + K_k (y_k^o - H_k x_k^f)$$

Propagation of error estimate

$$P_k^a = (I - K_k H_k) P_k^f$$

with Kalman gain

$$K_k = P_k^f H_k^T \left(H_k P_k^f H_k^T + R_k \right)^{-1}$$

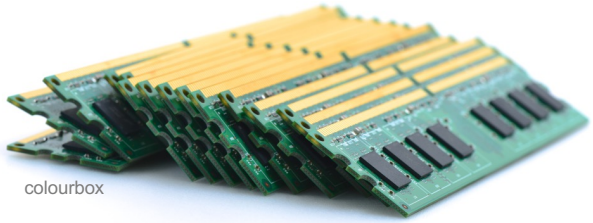
Large Scale Models

PDAF Parallel Data Assimilation Framework

State dimension: 10^6 - 10^9 Observations: 10^5 - 10^7

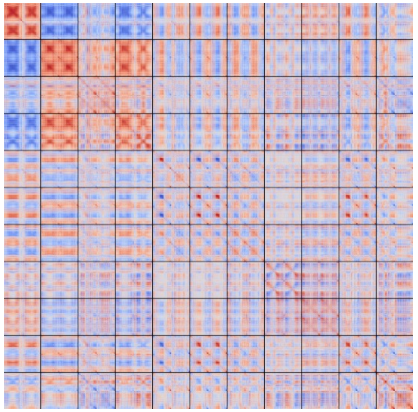
The covariance matrix of the model errors ***P*** is the limiting factor.

Memory consumption increases quadratically



Matrix multiplication has complexity of $\mathcal{O}(n^3)$

How to get ***P***?

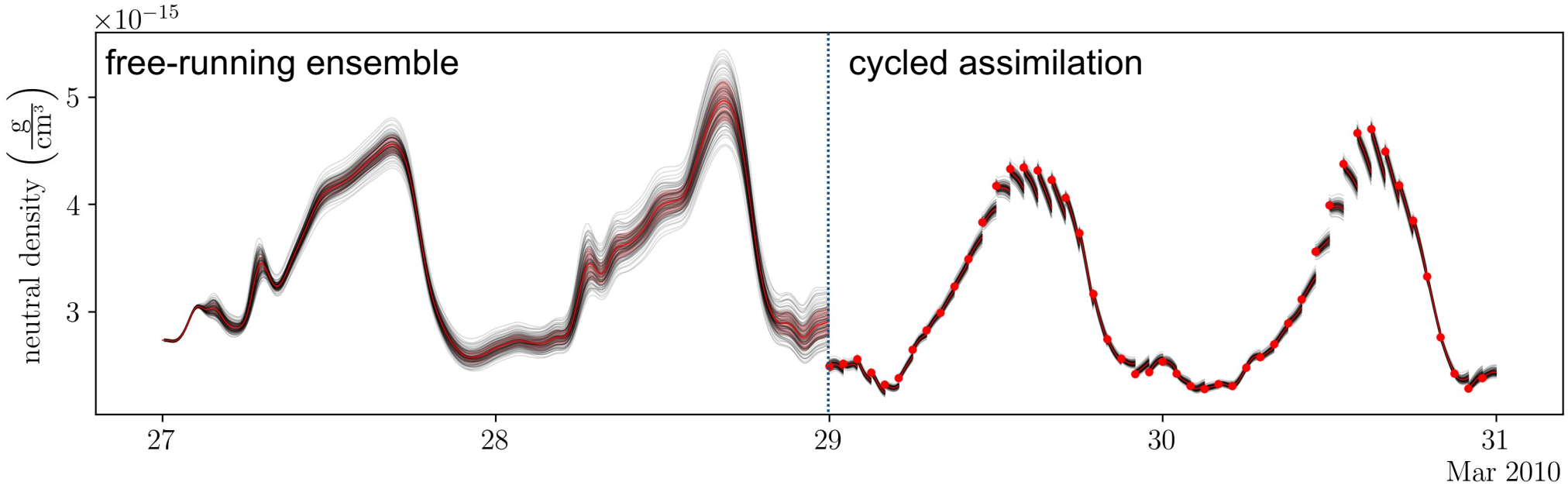


Kalman filter is often infeasible

Ensemble Kalman Filters

PDAF Parallel Data Assimilation Framework

represent state and uncertainty by ensemble of model instances



ensemble matrix

$$X = [x_1\, x_2\, \cdots\, x_n]$$

ensemble mean

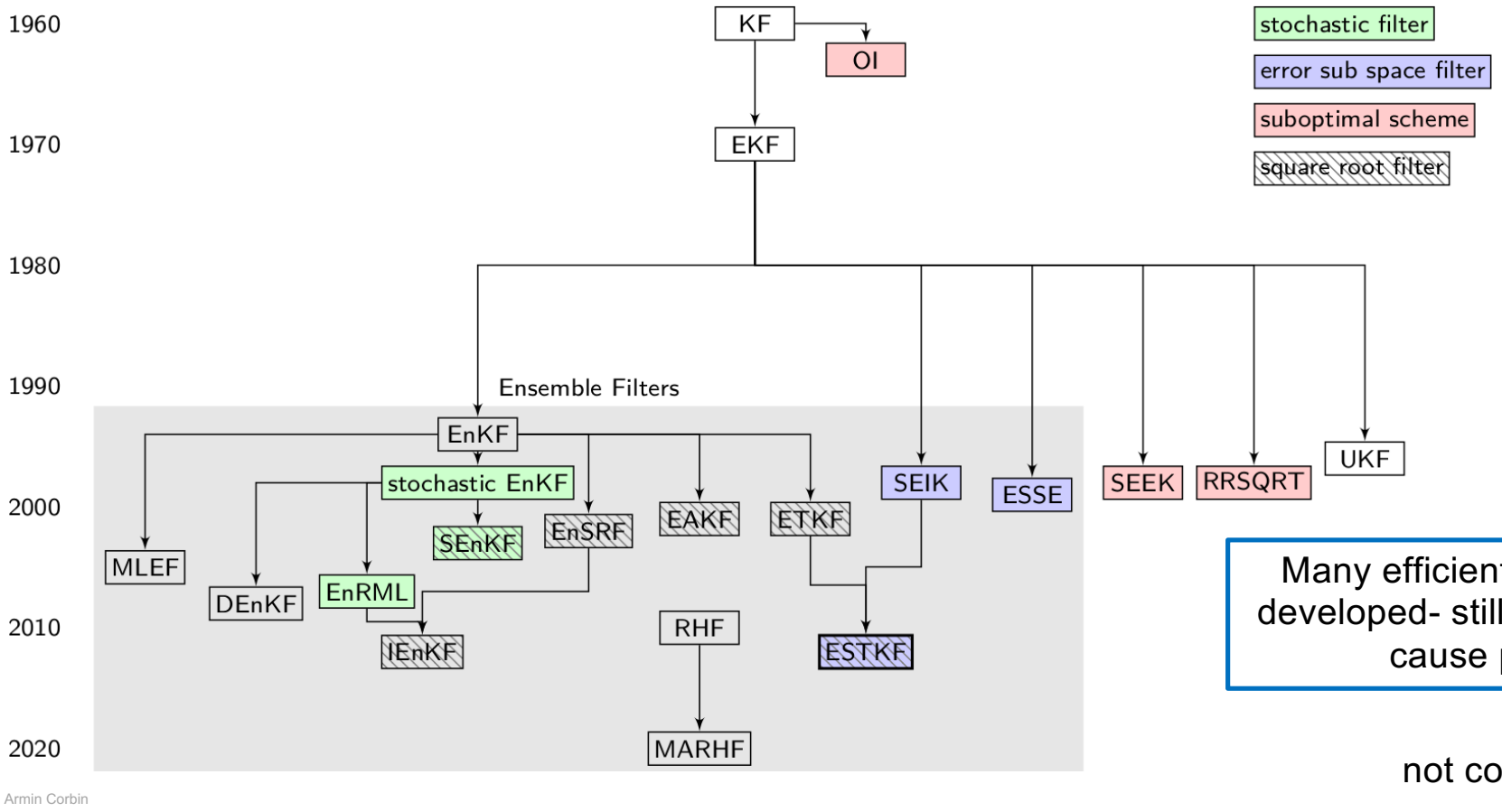
$$\bar{x} = \frac{1}{n} X I$$

ensemble variance

$$P^f \approx \frac{1}{n-1}(X - \overline{X})(X - \overline{X})^T$$

The Zoo of Kalman Filters

PDAF Parallel Data Assimilation Framework



Many efficient schemes were developed- still high dimensions cause problems

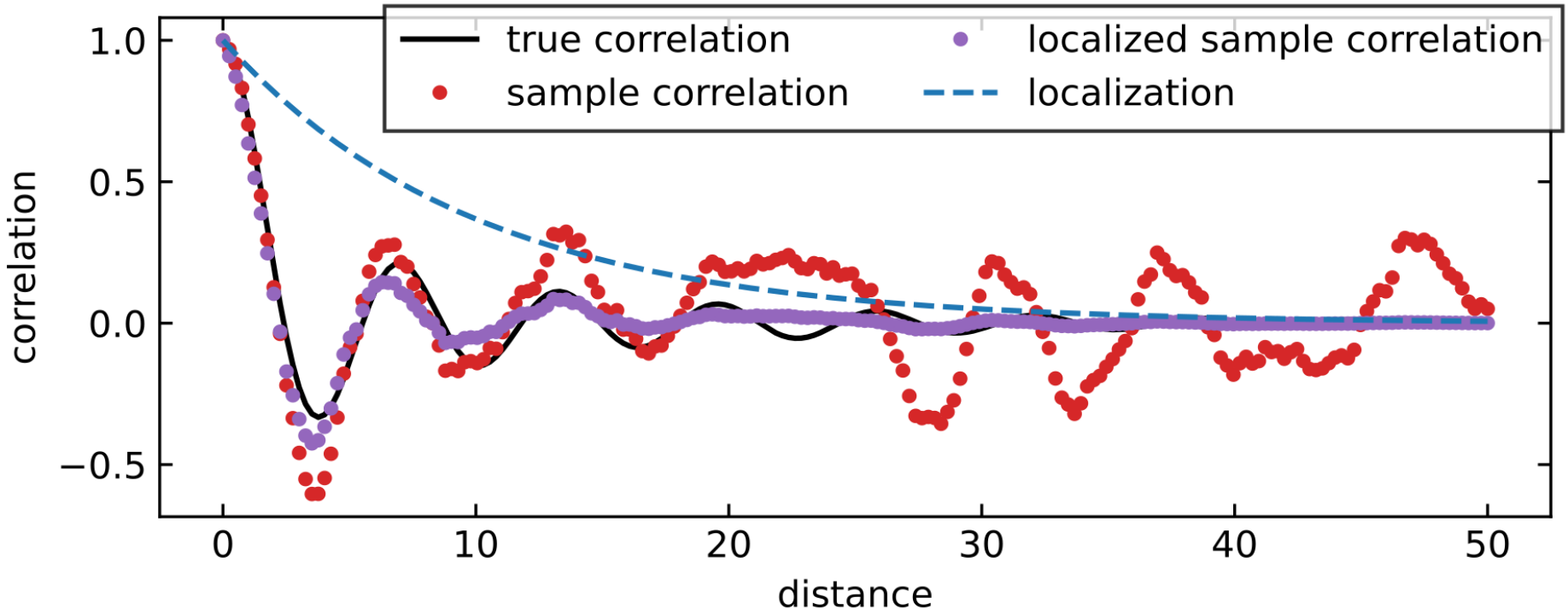
not complete

Armin Corbin

Covariance Localization

PDAF Parallel Data Assimilation Framework

- Multiply covariance matrix of forecasted ensemble point wise with finite covariance function or exponential decay

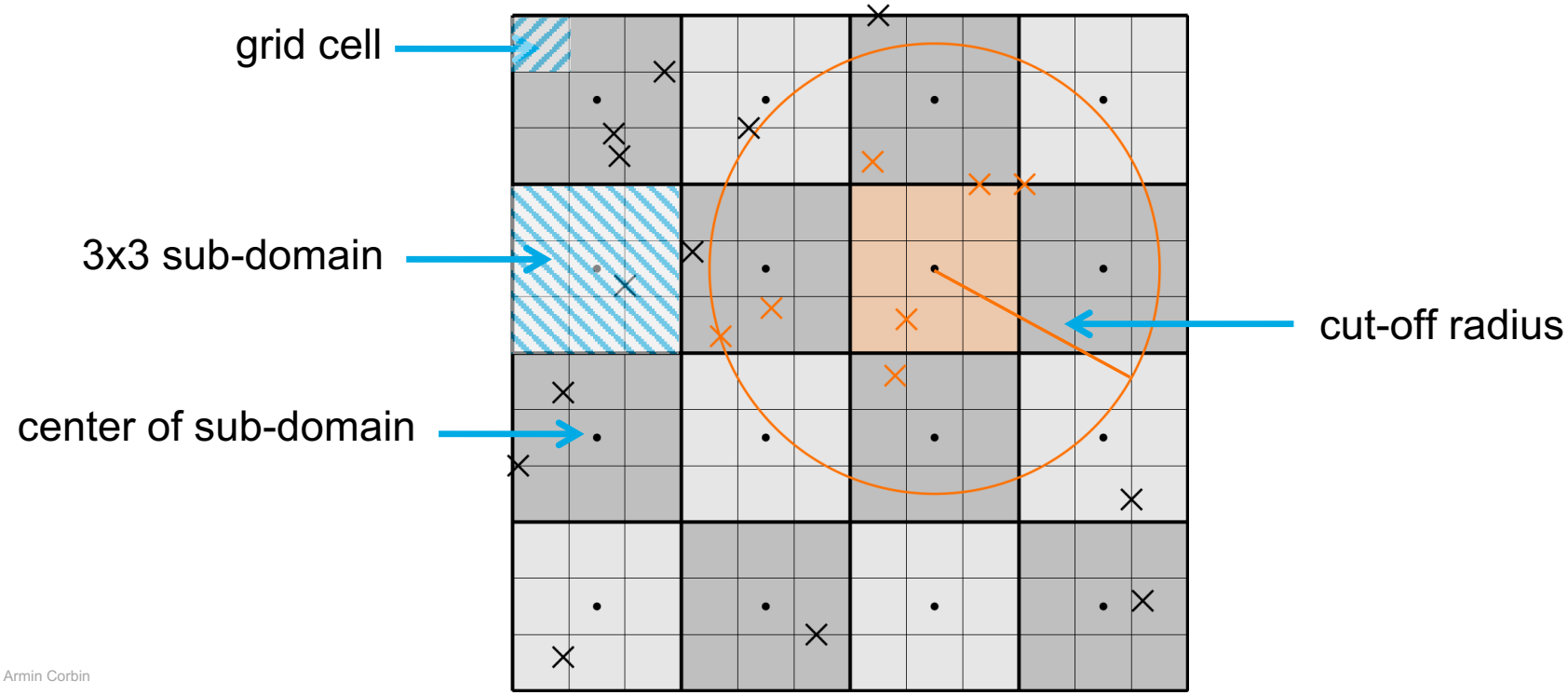


Armin Corbin

Domain Localization

PDAF Parallel Data Assimilation Framework

- subdivide model into disjoint sub-domains
- update each sub-domain individually taking only observations within specific distance into account

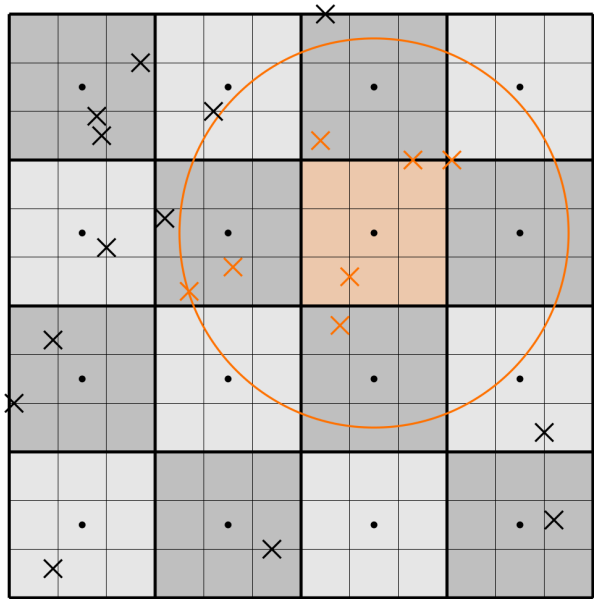


Armin Corbin

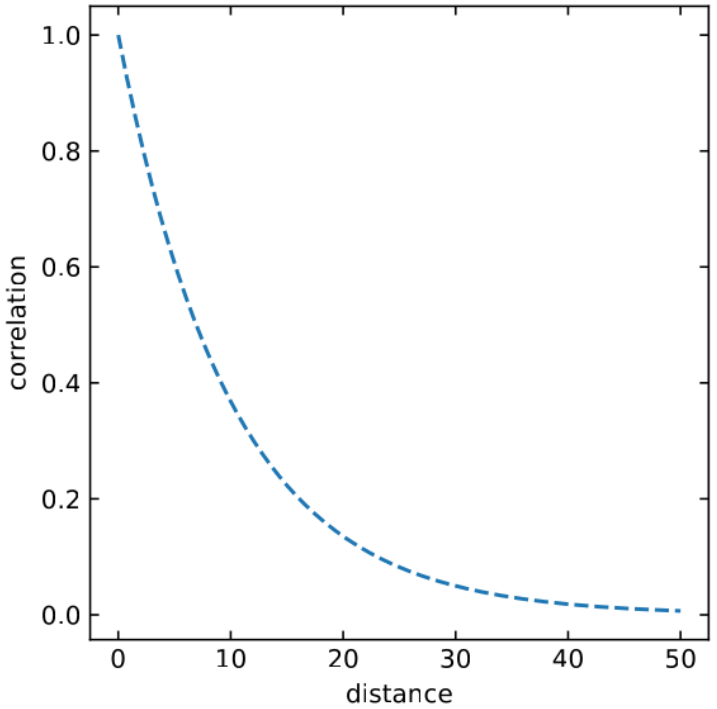
Observation Localization

PDAF Parallel Data Assimilation Framework

- implies domain localization
- weigh observations of each subdomain with a (finite) covariance function dependence on distance



+



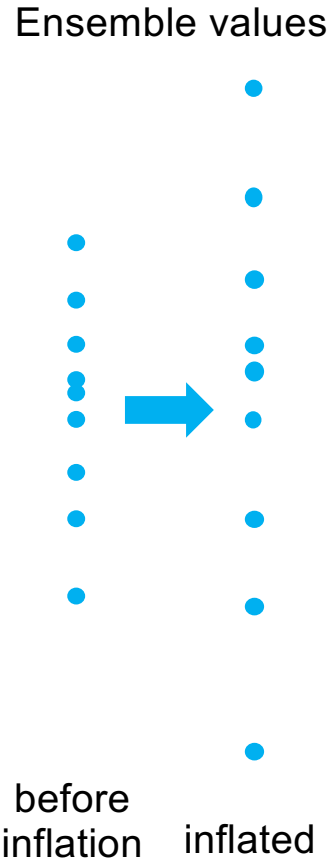
Inflation

PDAF Parallel Data Assimilation Framework

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

Inflation → Increase error estimate before analysis

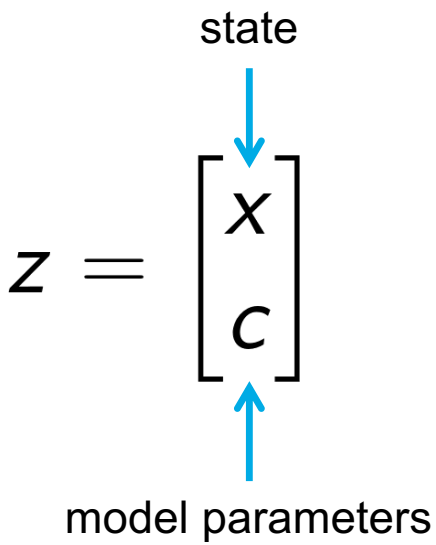
- Increase ensemble spread by constant factor
- Needs to be experimentally tuned



Co-Estimation of Model Parameters (Model Calibration)

PDAF Parallel Data Assimilation Framework

1. augment state vector with model parameters
2. estimate parameters using observations of model fields



PDAF Parallel Data Assimilation Framework

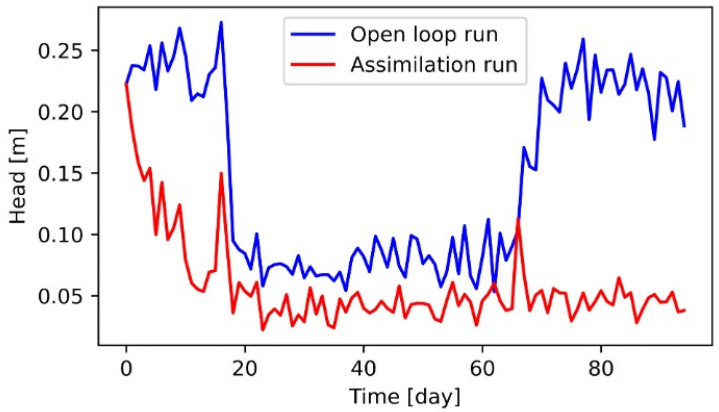
II Applications

Improve real-time groundwater level

PDAF Parallel Data Assimilation Framework

piezometers & soil moisture

Absolute head difference



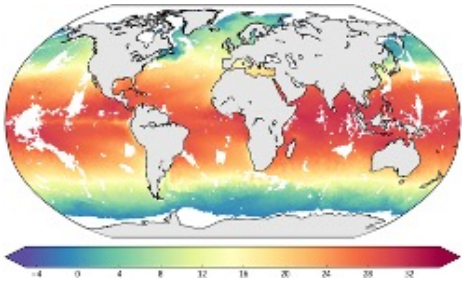
- **Without DA:** the simulation can lead to an error of 25cm during the pumping period
- **With DA:** errors are reduced by 60%

Dossier	
Domain	Hydrogeology
Model	HydroGeoSphere
N state	316 240
N obs.	8
N ens.	128
ΔT	1 day
duration	96 days
Filter	EnKF

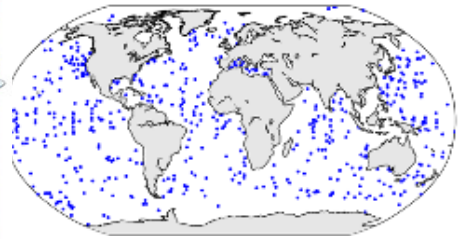
Tang, Q., Delottier, H., Kurtz, W., Nerger, L., Schilling, O. S., and Brunner, P.: HGS-PDAF (version 1.0): a modular data assimilation framework for an integrated surface and subsurface hydrological model, Geosci. Model Dev., 17, 3559–3578, , 2024.

Coupled ocean-atmosphere DA

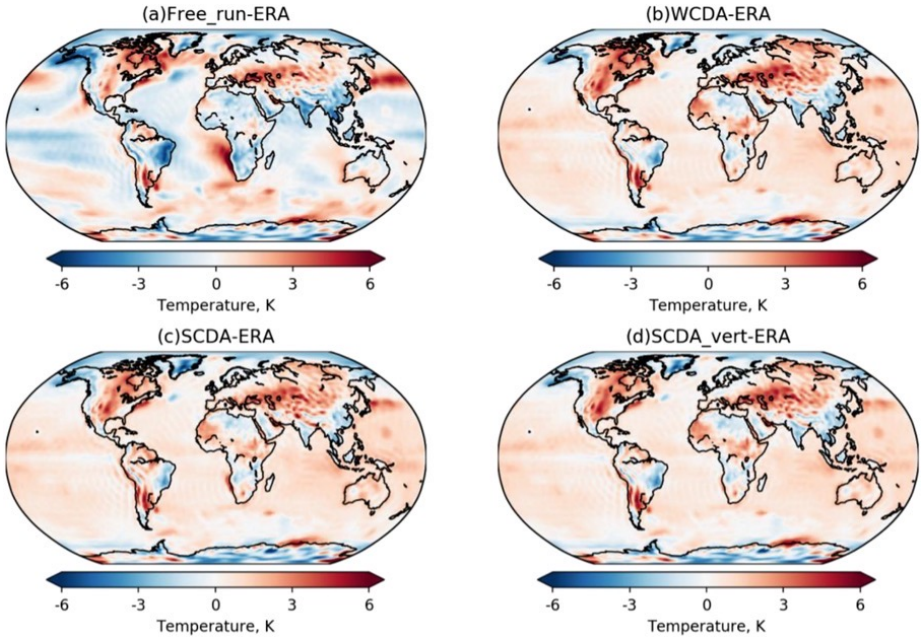
PDAF Parallel Data Assimilation Framework



Satellite surface temperature



Profiles



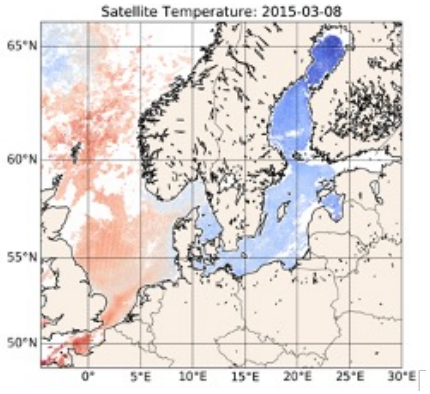
Average bias (model simulation - ERA-interim) of temperature at 2 m above surface for the free run and assimilation runs;

Dossier	
Domain	Ocean & atmosphere
Model	AWI-CM
N state	Global scale
N obs.	Global
N ens.	46
ΔT	1 day
duration	1 year
Filter	LESTKF

Nerger et al. (2020): Efficient ensemble data assimilation for coupled models with the Parallel Data Assimilation Framework: example of AWI-CM (AWI-CM-PDAF 1.0), Geosci. Model Dev.
Tang et al. (2021): Strongly coupled data assimilation of ocean observations into an ocean-atmosphere model. Geophysical Research Letters

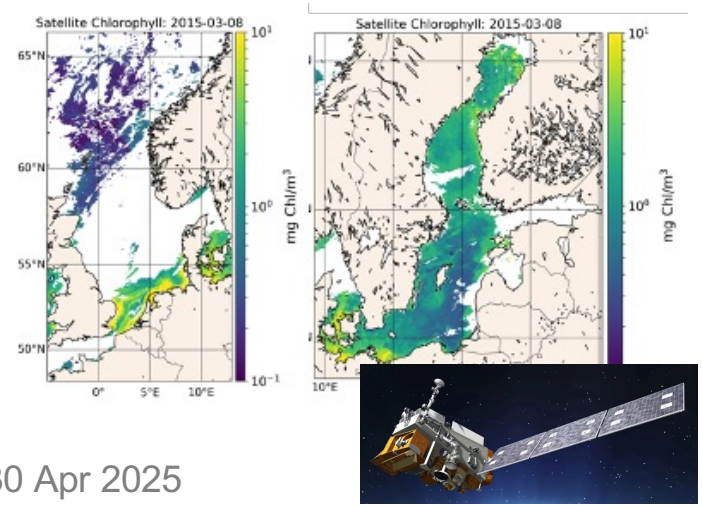
Improving regional ocean physics and biogeochemistry

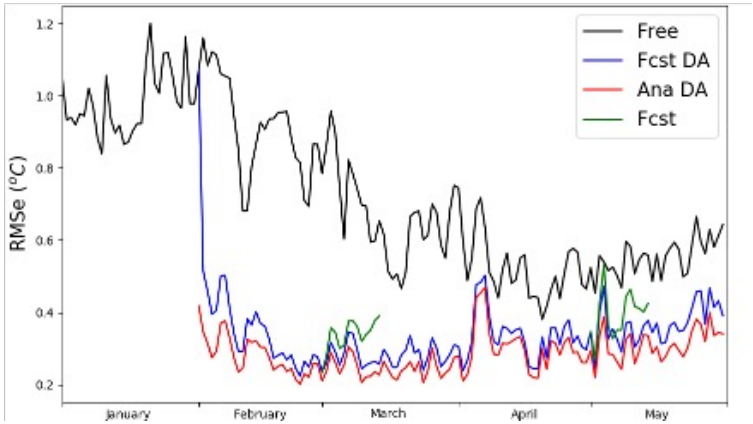
PDAF Parallel Data Assimilation Framework



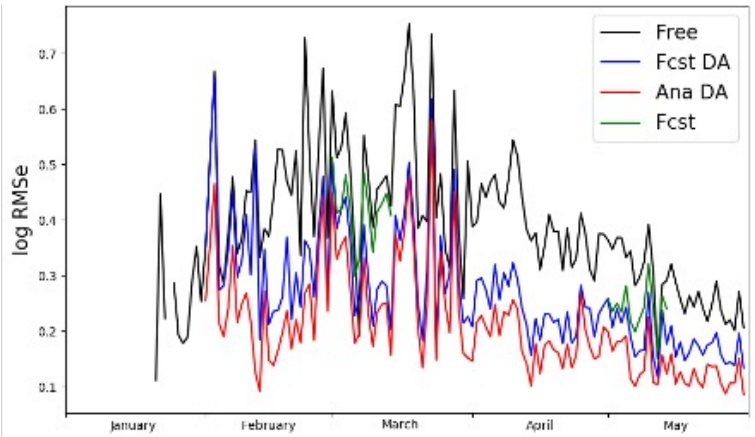
Satellite surface temperature

Satellite chlorophyll concentration





Assimilation effect (RMS error) on surface tempeature (top) and chlorophyll (bottom), with 14-day forecasts (green)



Dossier	
Domain	Ocean
Model	NEMO-ERGOM
N state	~700,000,000
N obs.	200,000
N ens.	30
ΔT	1 day
duration	5 months
Filter	LESTKF

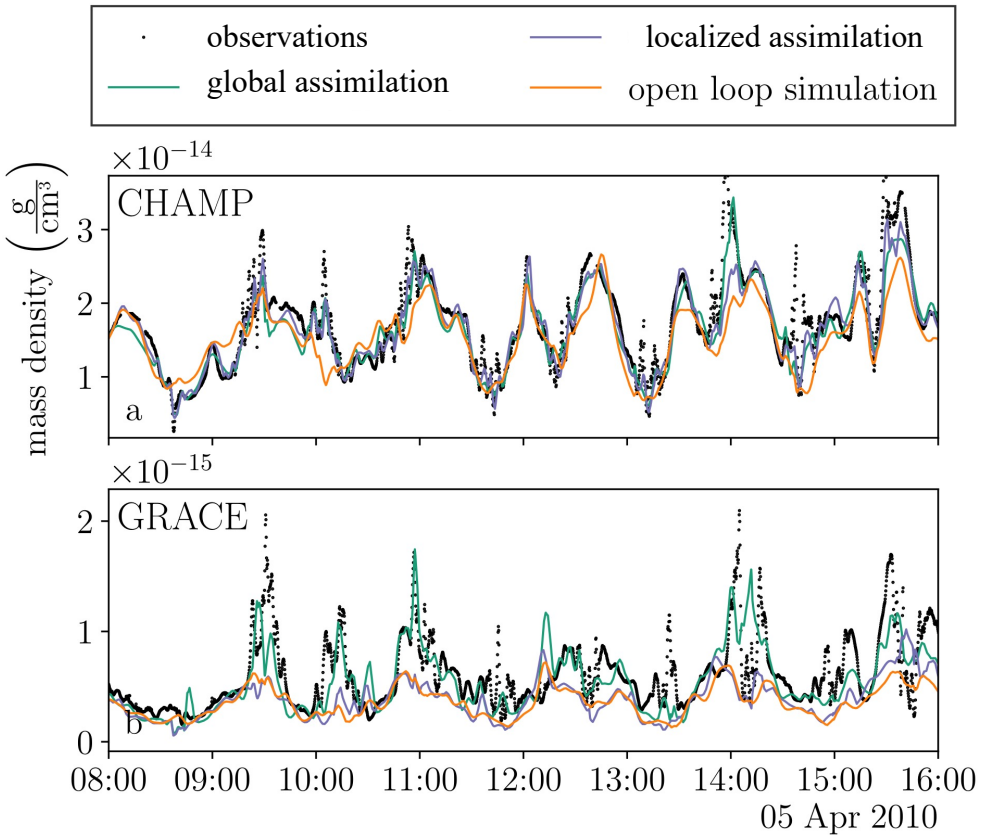
SEAMLESS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

Upper Atmosphere Density

PDAF Parallel Data Assimilation Framework

accelerometer

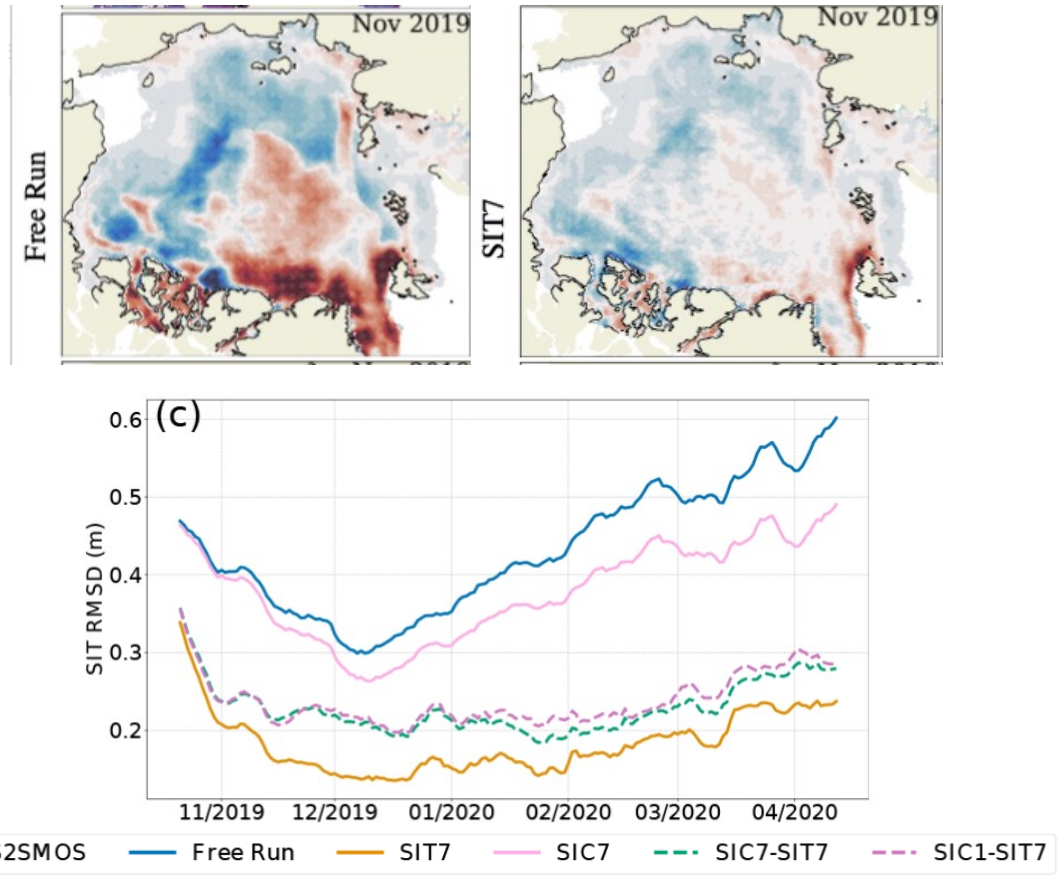


Dossier	
Domain	Atmosphere
Model	TIE-GCM
N state	870 912
N obs.	1
N ens.	96
ΔT	1 min
duration	2 weeks
Filter	(L)ESTKF

Corbin, A. & Kusche, J. Improving the estimation of thermospheric neutral density via two-step assimilation of in situ neutral density into a numerical model. Earth, Planets and Space 74, 183 (2022).

Arctic sea ice

PDAF Parallel Data Assimilation Framework



Dossier	
Domain	Cryosphere
Model	neXtSIM
Model resolution	12 km
Obs.	Sea ice thickness and concentration
N ens.	40
ΔT	1 min
duration	6 months
Filter	DEnKF

Cheng, S., Chen, Y., Aydoğdu, A., Bertino, L., Carrassi, A., Rampal, P., and Jones, C. K. R. T.: Arctic sea ice data assimilation combining an ensemble Kalman filter with a novel Lagrangian sea ice model for the winter 2019–2020, The Cryosphere, 17, 1735–1754, https://doi.org/10.5194/tc-17-1735-2023, 2023.

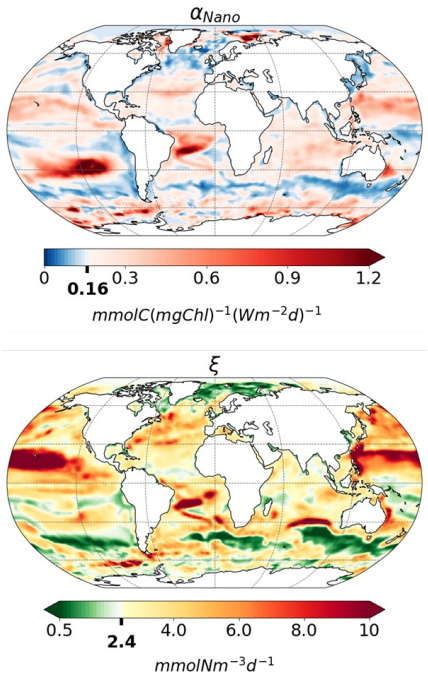
Estimating Spatially Varying Ocean Biogeochemical Process Parameters

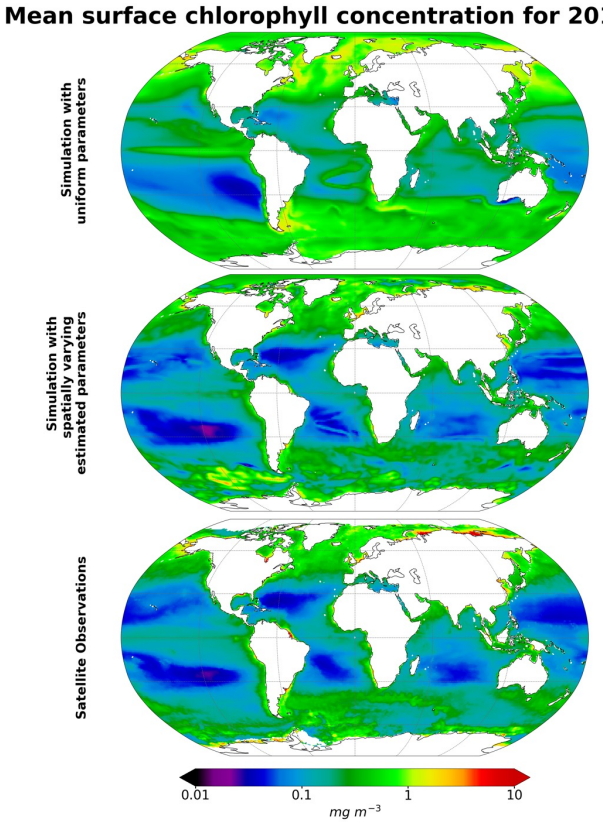
PDAF Parallel Data Assimilation Framework

Ocean Colour in the North Sea

Image: NASA Earth Observatory

Spatially Varying Parameter Estimates

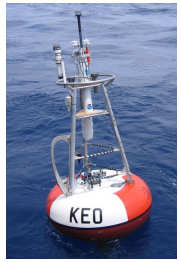




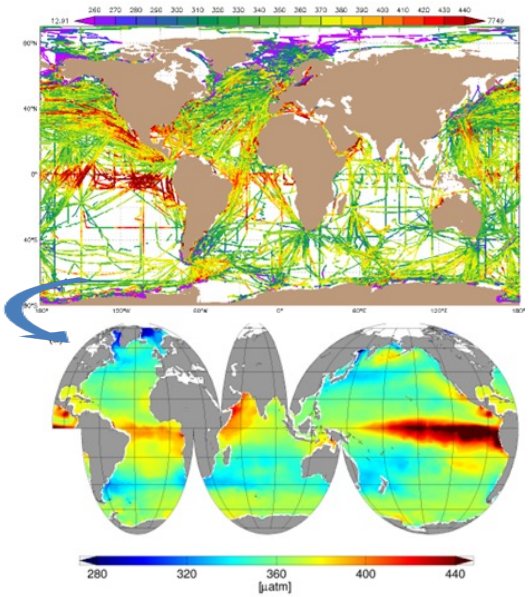
Dossier	
Domain	Ocean
Model	MITgcm-REcoM2
Model resolution	20-110km
Obs.	Ocean Colour Chlorophyll-a
N ens.	40
ΔT	5 day
duration	1 year
Filter	LESTKF

Global Ocean Biogeochemical Reanalysis

PDAF Parallel Data Assimilation Framework

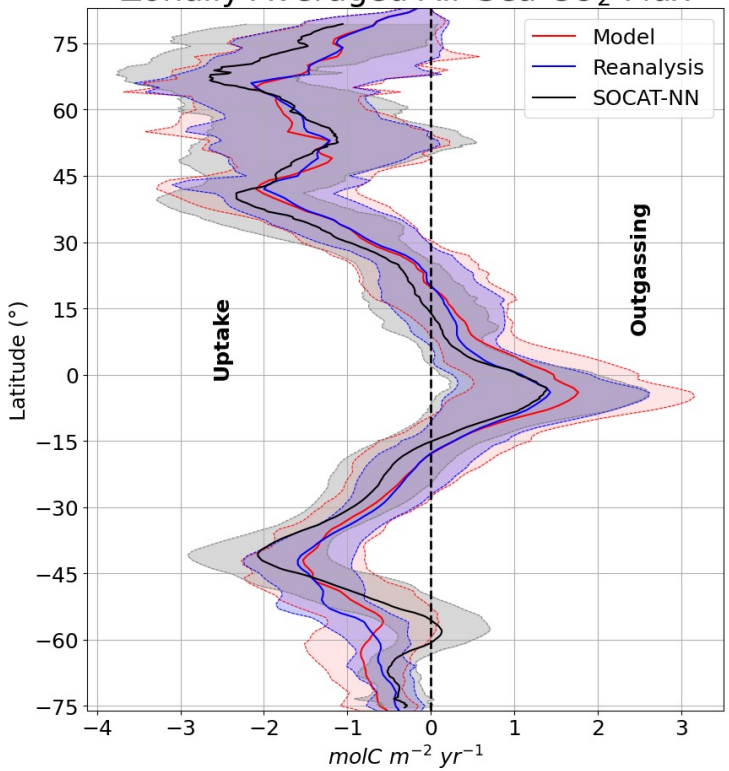


Surface ocean pCO_2



Resplandy et al., 2018

Zonally Averaged Air-Sea CO_2 Flux



Dossier	
Domain	Ocean
Model	NEMO-PISCES
Model resolution	1/4 degree
Obs.	Daily Sat. Chl. SOCAT-NN
ΔT	7 day
duration	30 year
Filter	SEEK

PDAF Parallel Data Assimilation Framework

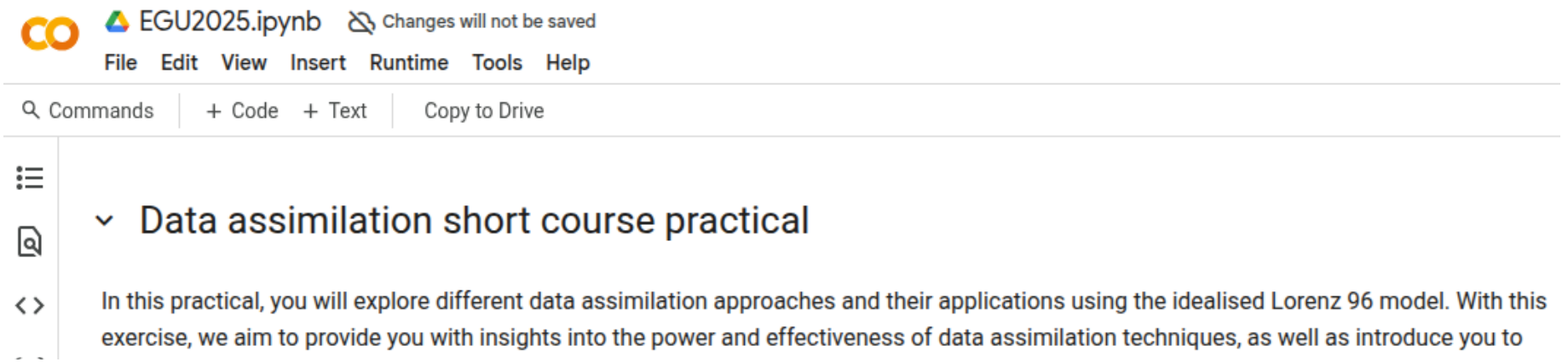
III Hands On

A taste of DA on Lorenz 96 model

PDAF Parallel Data Assimilation Framework

- Hands on ensemble-based Kalman filter and 3DVar
- Twin experiments
  - One model run is deemed truth
  - Observations are sampled synthetically from the truth
  - Assimilation is performed with a different model initial condition
- Feel free to ask questions if you run into any problems!

Available on Google Colab:
https://tinyurl.com/2kmkfr74



References

PDAF Parallel Data Assimilation Framework

Asch, M, M. Bocquet, M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications*, SIAM, 2017

Evensen, G., F. Vossepoel, P. J. van Leeuwen, *Data Assimilation Fundamentals*, Springer, 2022

Moreaux, G., *Compactly Supported Radial Covariance Functions*, Journal of Geodesy 82.7, 2008

Vetra-Carvalho, S., Van Leeuwen, P.J., Nerger, L., Barth, A., Altaf, M.U., Brasseur, P., Kirchgessner, P. and Beckers, J.-M., *State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems,* Tellus A: Dynamic Meteorology and Oceanography, 70(1), 2018

pdaf.awi.de