EGU26 GENERAL ASSEMBLY
6 May 2026

PDAF Parallel Data Assimilation Framework

SC2.7 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 Basel, Department of Environmental Sciences, Switzerland

[2]University of Neuchâtel, 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

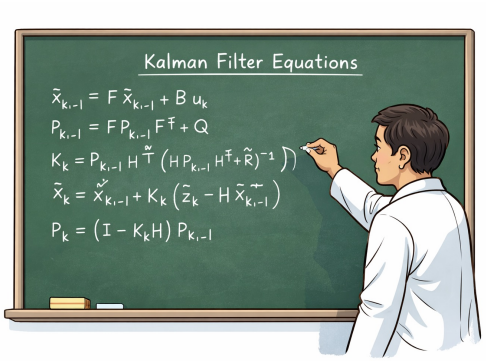
[6]University of Reading, National Centre for Earth Observation, Department of Meteorology, United Kingdom of Great Britain

Universität Basel · unine Université de Neuchâtel · AWI · Universität Bonn · IMPERIAL · University of Reading

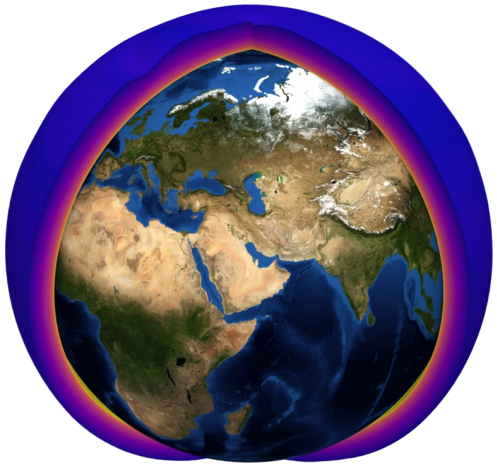
Schedule

PDAF Parallel Data Assimilation Framework

I Theory (45 min)



II Applications (15 min)



III Hands-on (45 min)

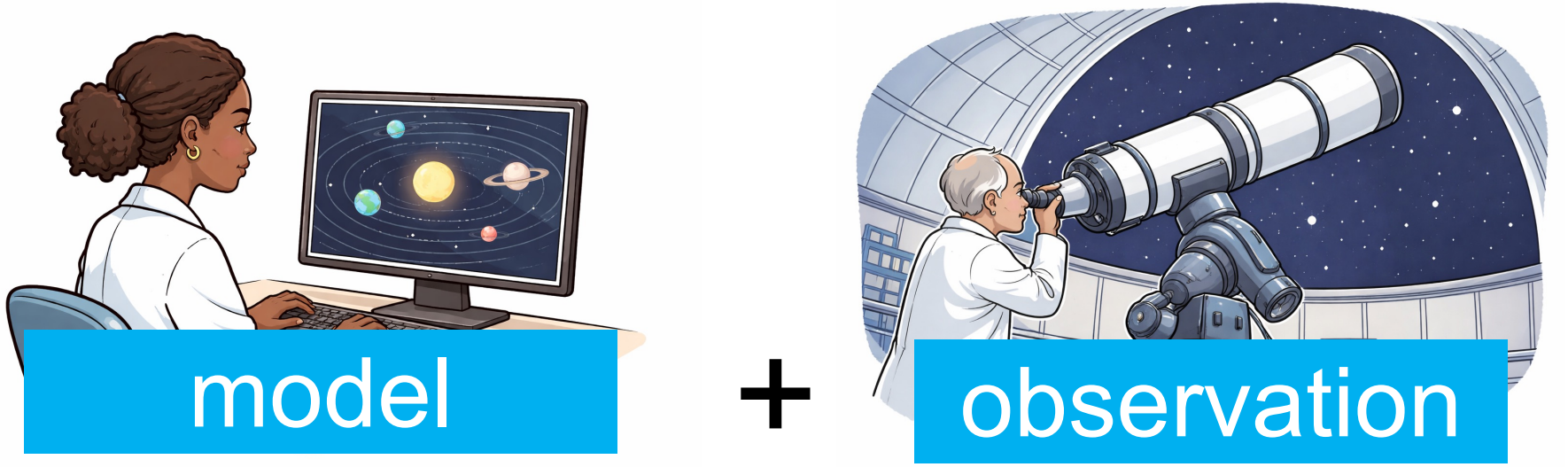
PDAF Parallel Data Assimilation Framework

I Theory

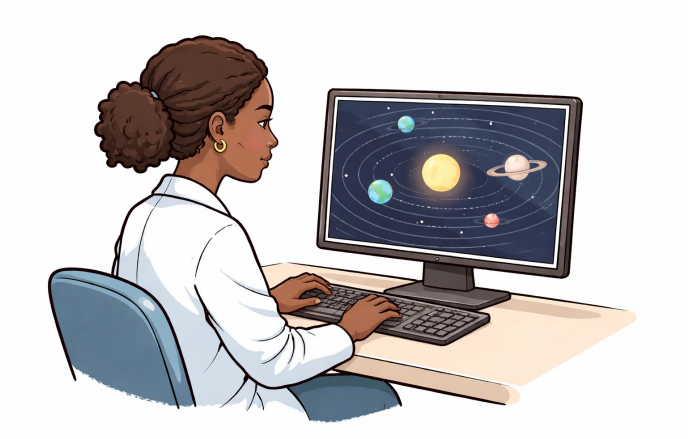
Data Assimilation (DA)

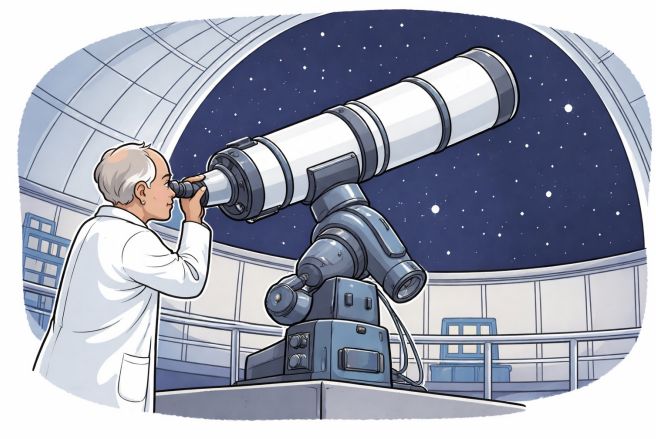
PDAF Parallel Data Assimilation Framework

*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

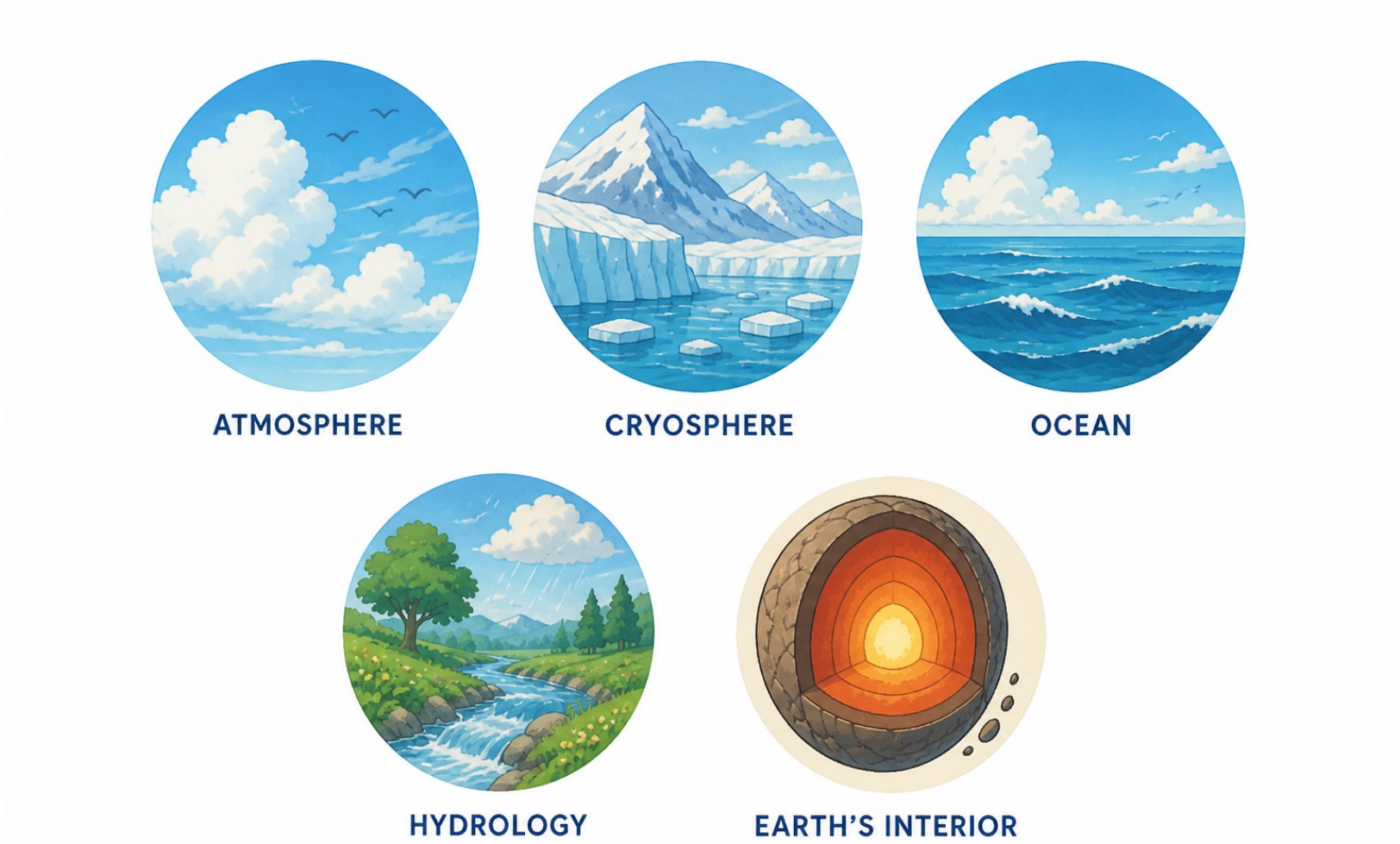




model	observation
- **idealized** representation of a system	+ measurements of “reality”
+ **complete coverage**: often located on a grid or mesh, high temporal and spatial resolution	- **Incomplete:** sparse, discrete, data gaps, irregular sampling, missing state variables
	- outliers
quantifiable systematic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

DA in Geo-science

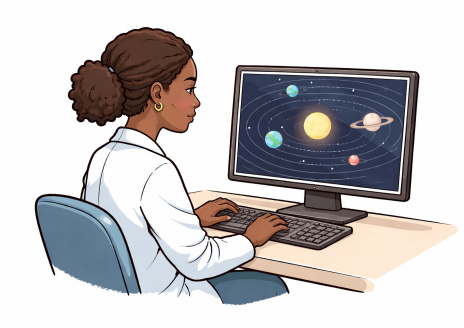
PDAF Parallel Data Assimilation Framework



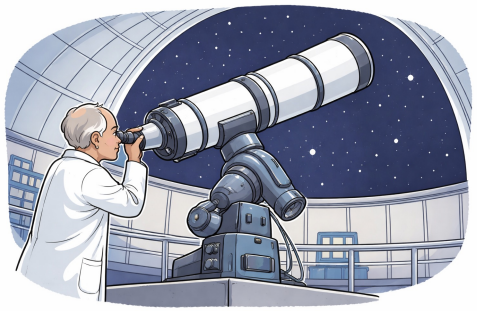
Requirements for DA

PDAF Parallel Data Assimilation Framework

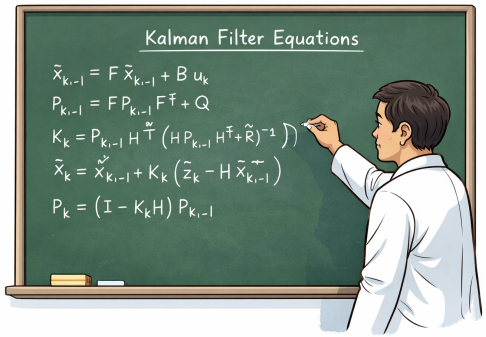
1. Model
   - With some skill



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

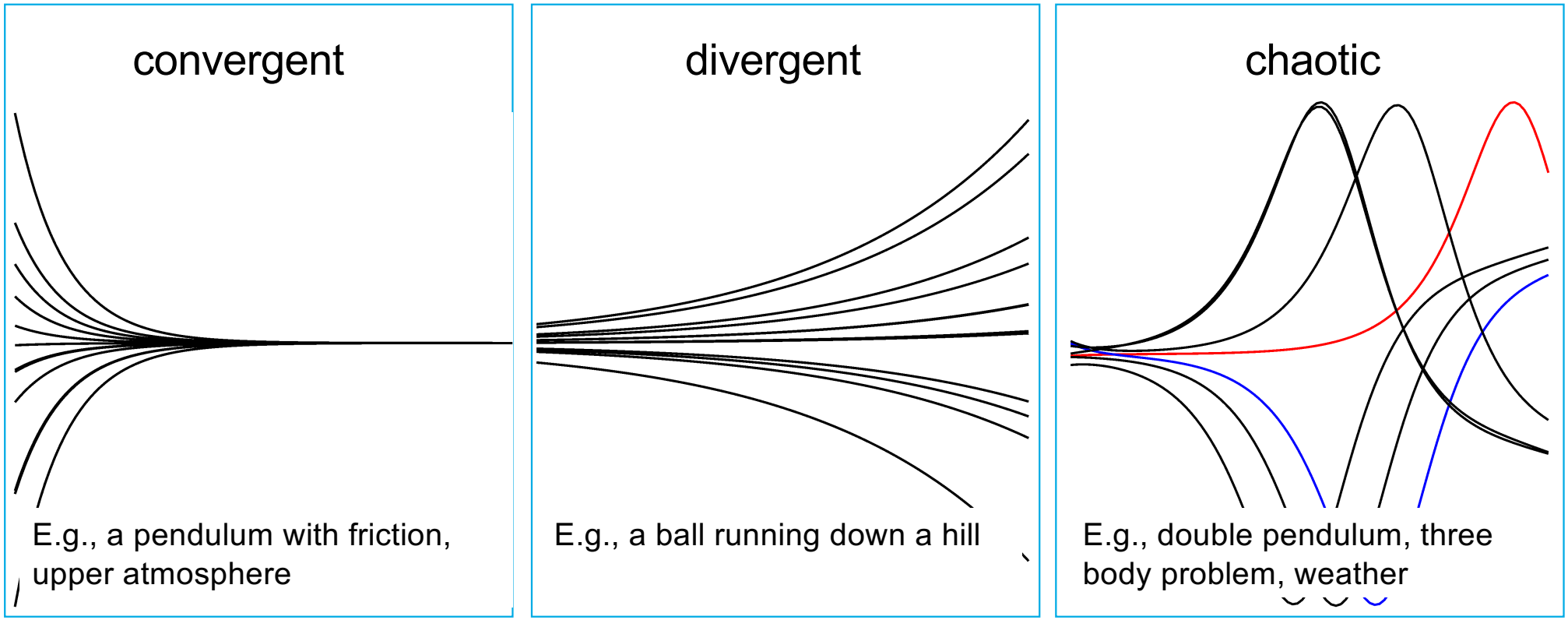


3. Data assimilation method
   - Suitable for application (complexity, nonlinearity)



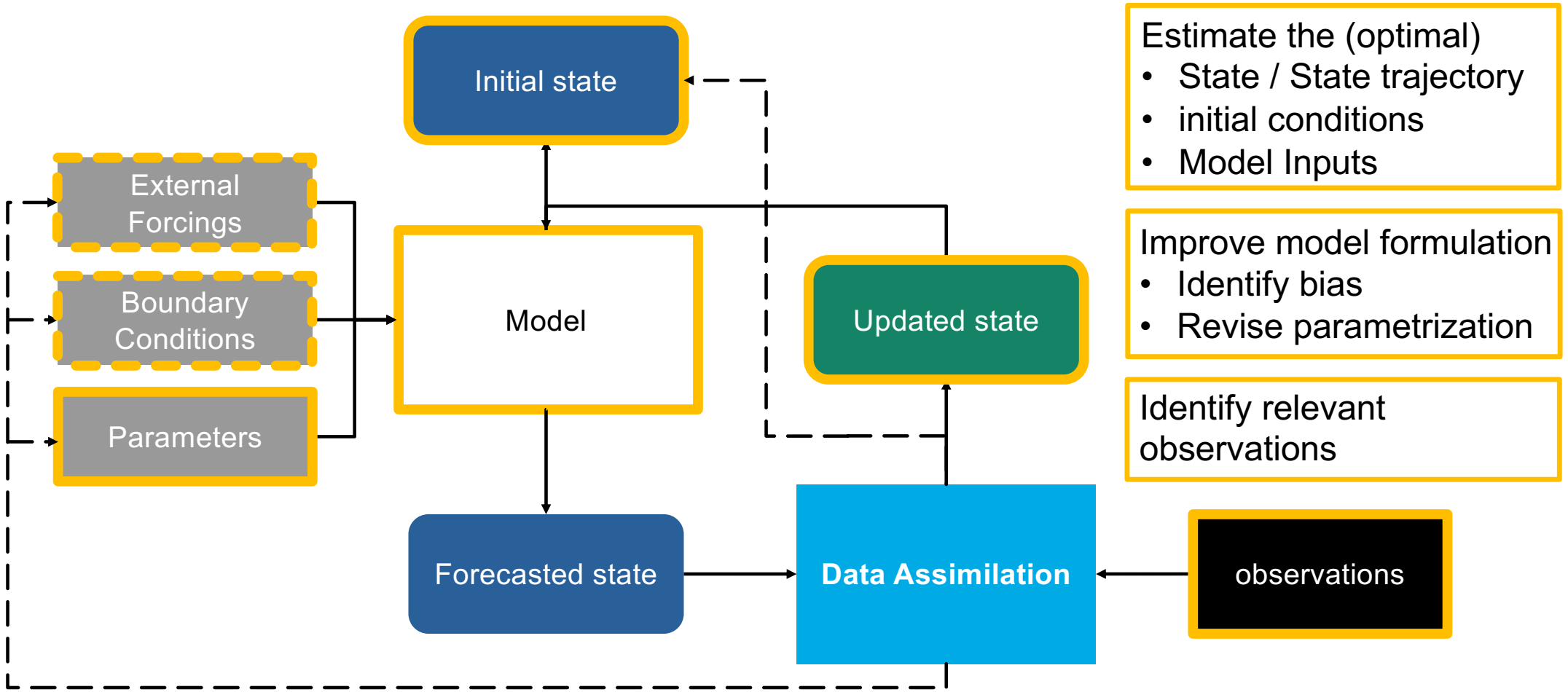
Dynamical System

The future state depends on the present state



What Data Assimilation Can Do

PDAF Parallel Data Assimilation Framework



Estimate the (optimal)
- State / State trajectory
- initial conditions
- Model Inputs

Improve model formulation
- Identify bias
- Revise parametrization

Identify relevant observations

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

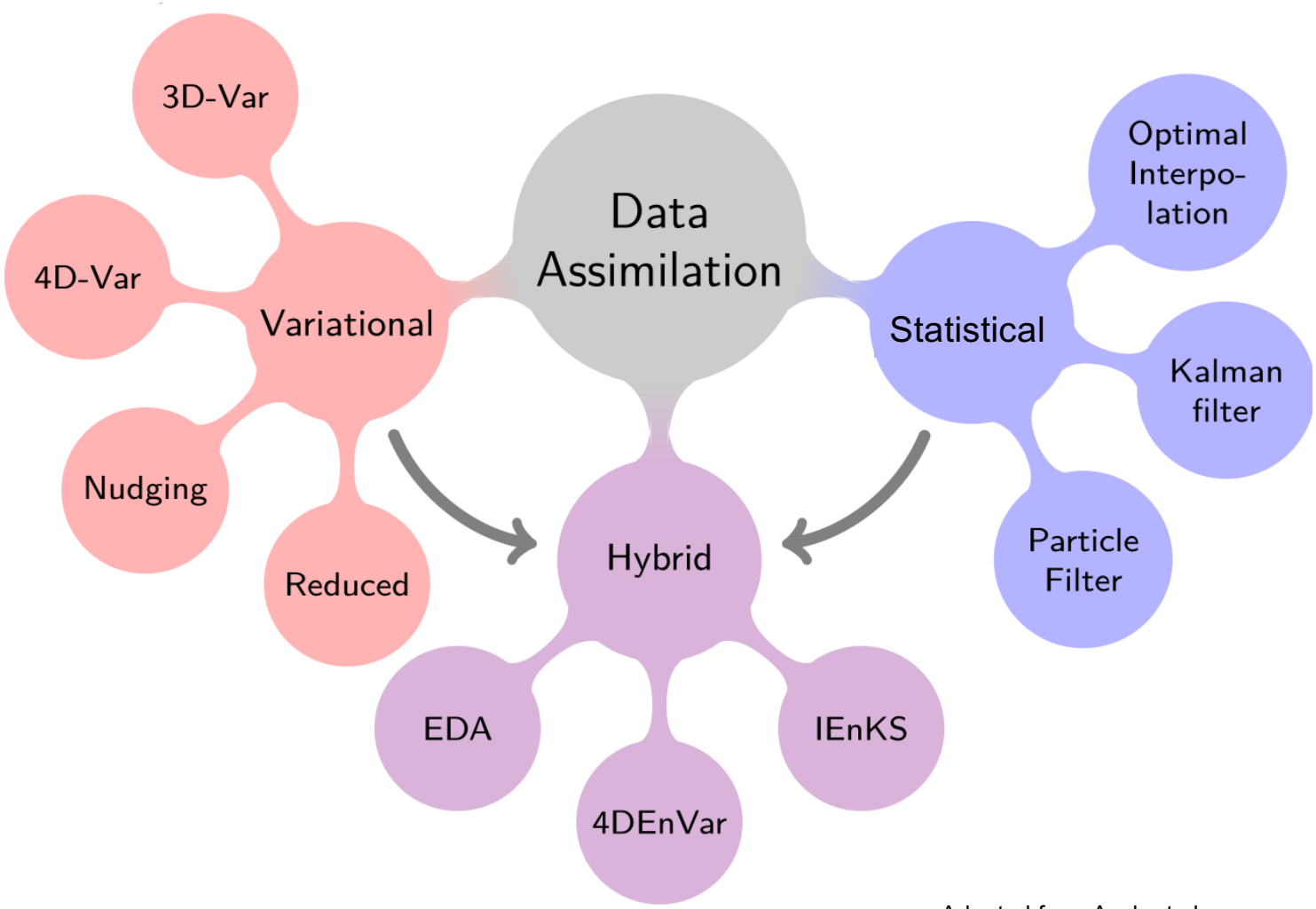
PDAF Parallel Data Assimilation Framework

observations state observation errors

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

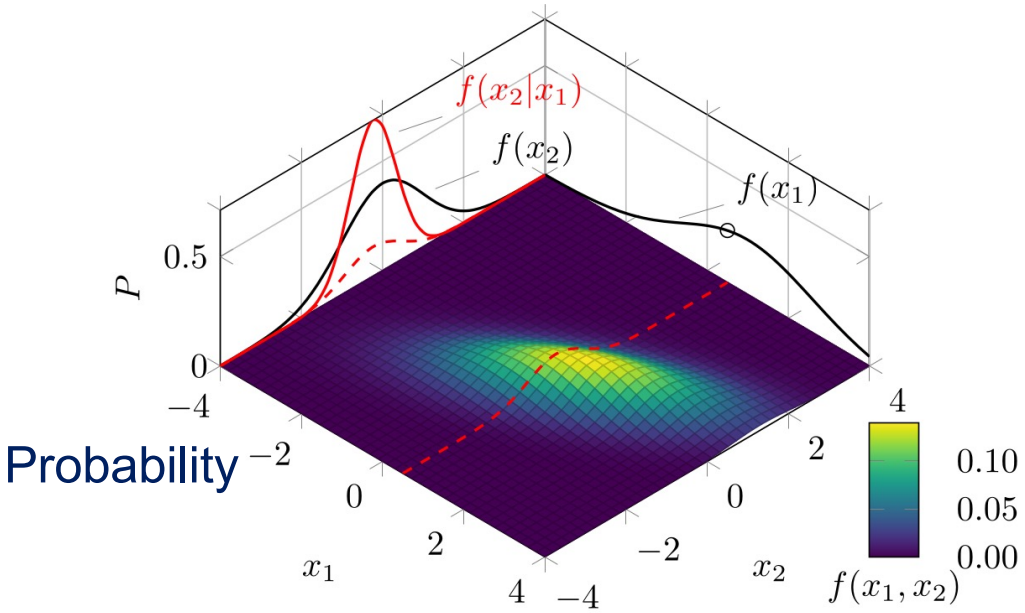
observation operator: maps state to observation

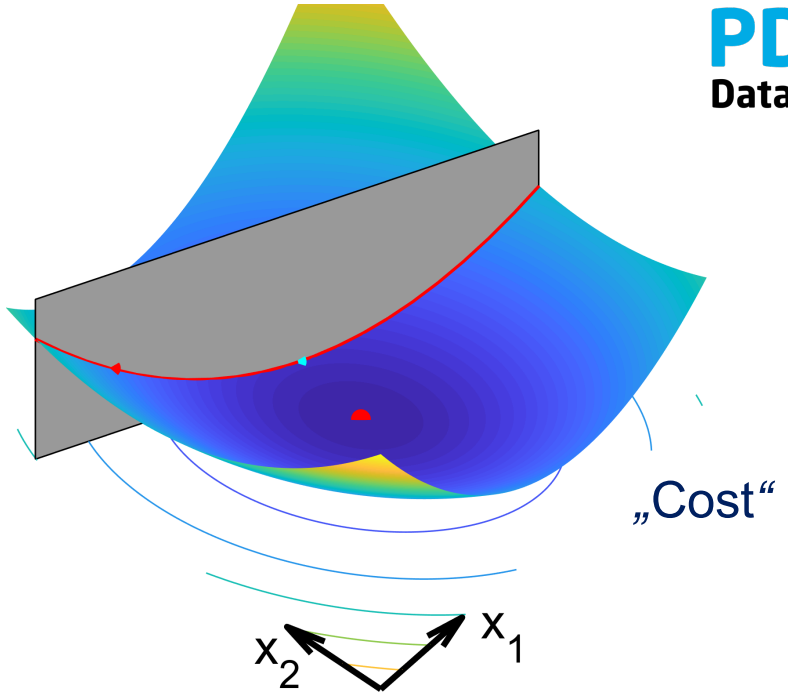
Linearized Operator: $$H = \left.\frac{\partial \mathcal{H}(x)}{\partial x}\right|_{x=x(t)}$$



- Adopted from Asch et al. (2016)

PDAF Parallel Data Assimilation Framework

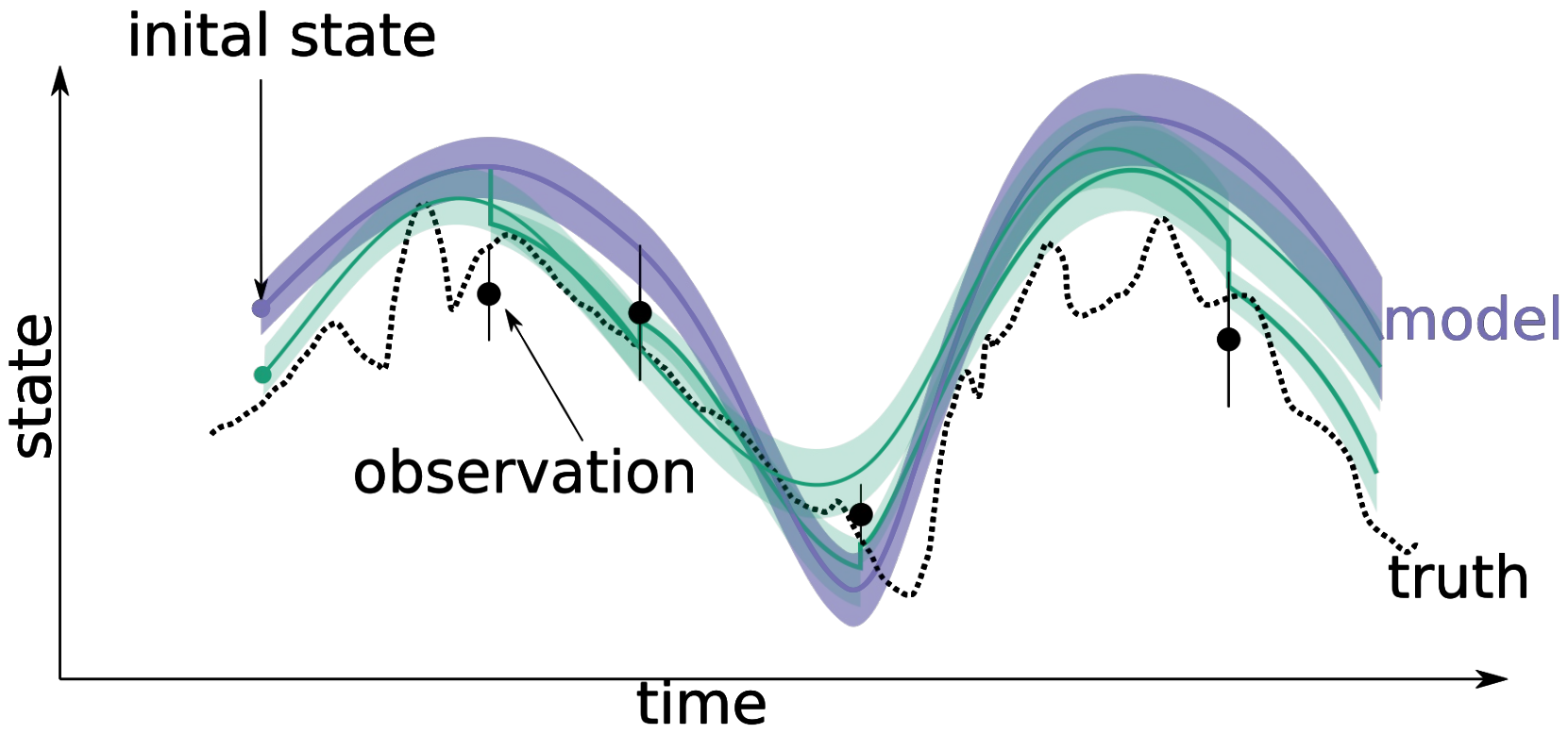




Statistical	Variational
Estimation theory	Optimal control theory
Maximization of probability density (minimization of variance)	Minimization of cost function (e.g. Gauss-Newton, conjugate gradient)

Filter and Smoother

PDAF Parallel Data Assimilation Framework



PDAF Parallel Data Assimilation Framework

	statistical (estimation)	variational (optimization)
filter	Kalman filter Particle filter	3D VAR
smoother	Kalman smoother Particle smoother	4D VAR

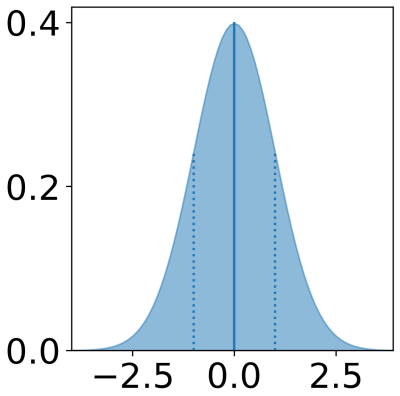
Kalman filter is optimal

PDAF Parallel Data Assimilation Framework

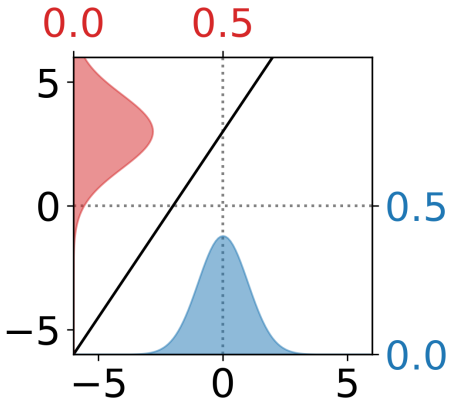
Optimal: state is **unbiased** and has **minimal variance**

Assumptions:

1. everything is Gaussian



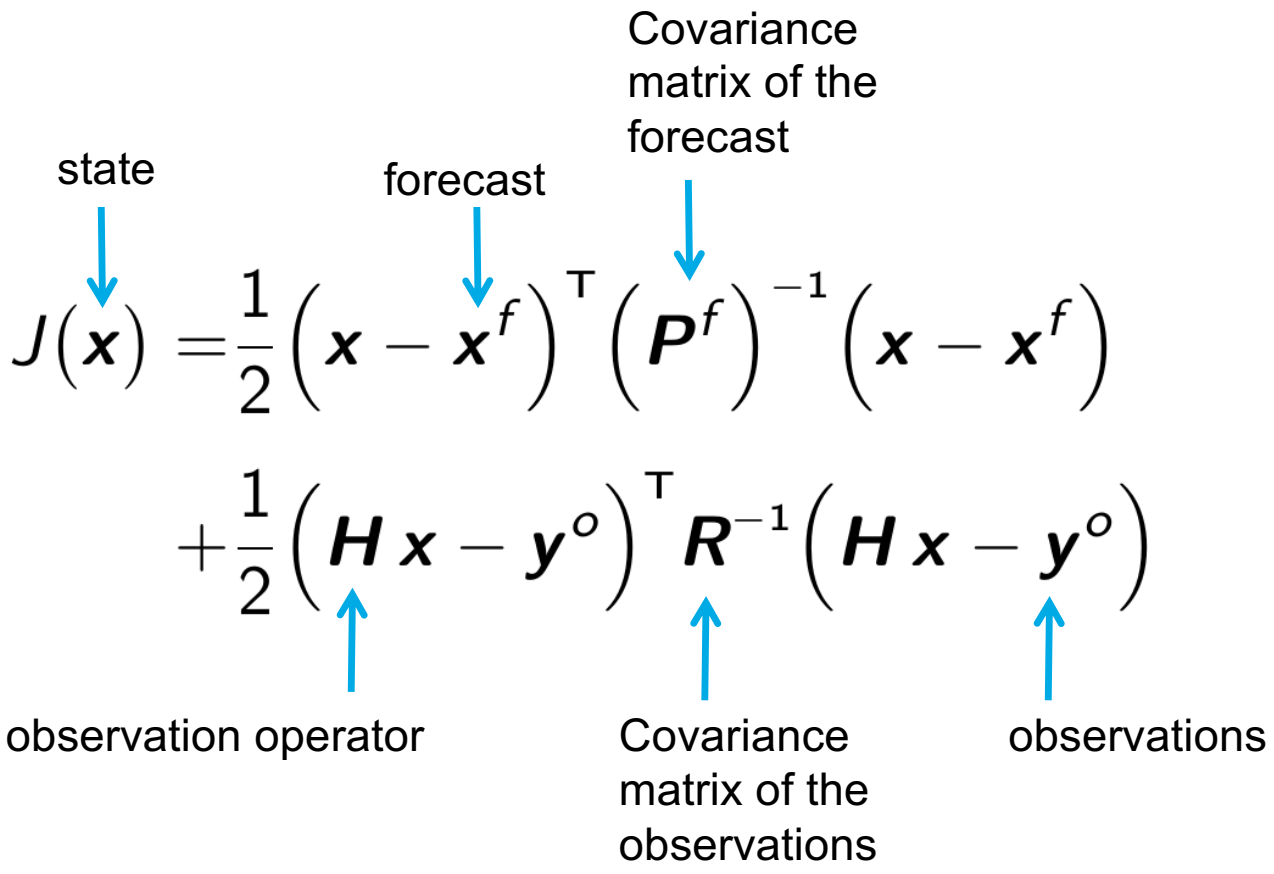
2. model and observation operator are linear



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

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

Link to variational DA



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

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

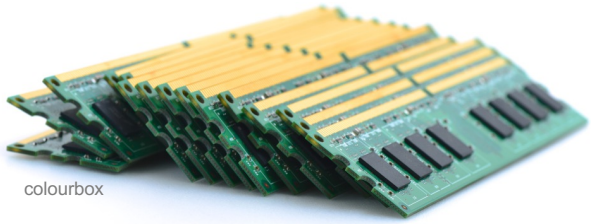
Large Scale Models

PDAF Parallel Data Assimilation Framework

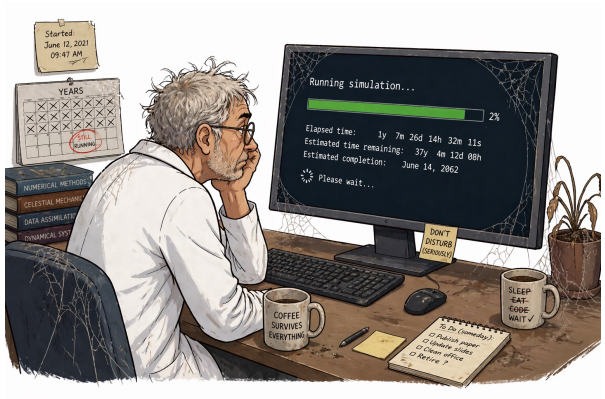
State dimension: 10^6 - 10^9 Observations: 10^5 - 10^7

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

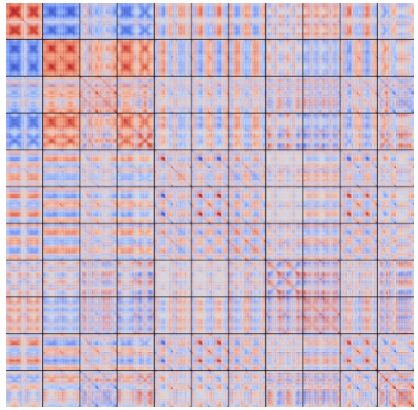
Memory consumption increases quadratically



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



How to get ***P***?

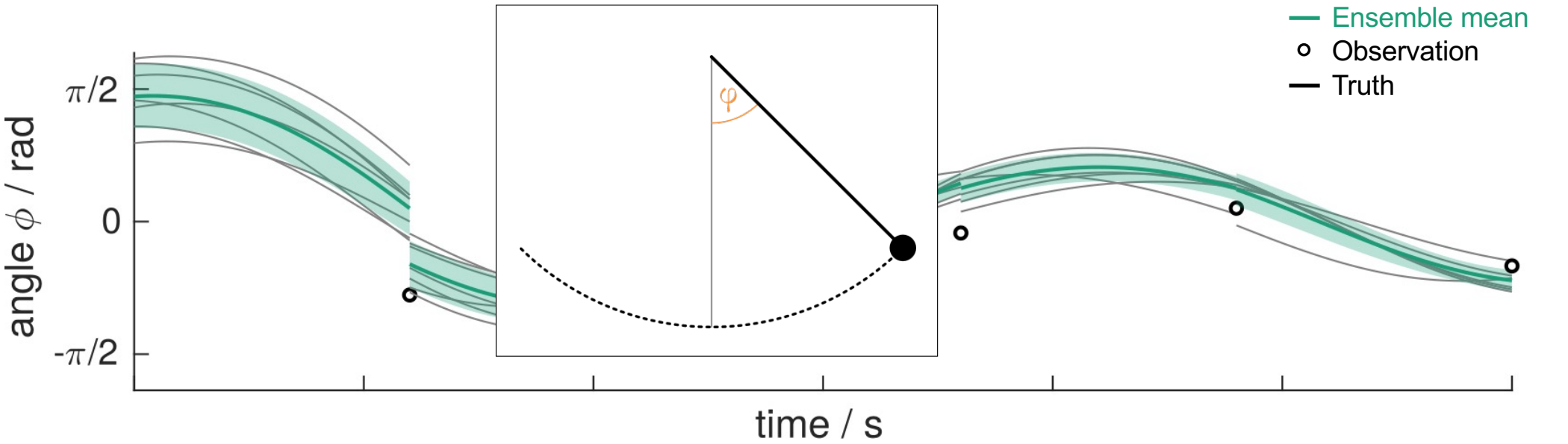


Kalman filter is often infeasible

Ensemble Kalman Filters

PDAF Parallel Data Assimilation Framework

represent state and uncertainty by ensemble of model instances
(Example: Estimate angle of a pendulum)



ensemble matrix

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

ensemble mean

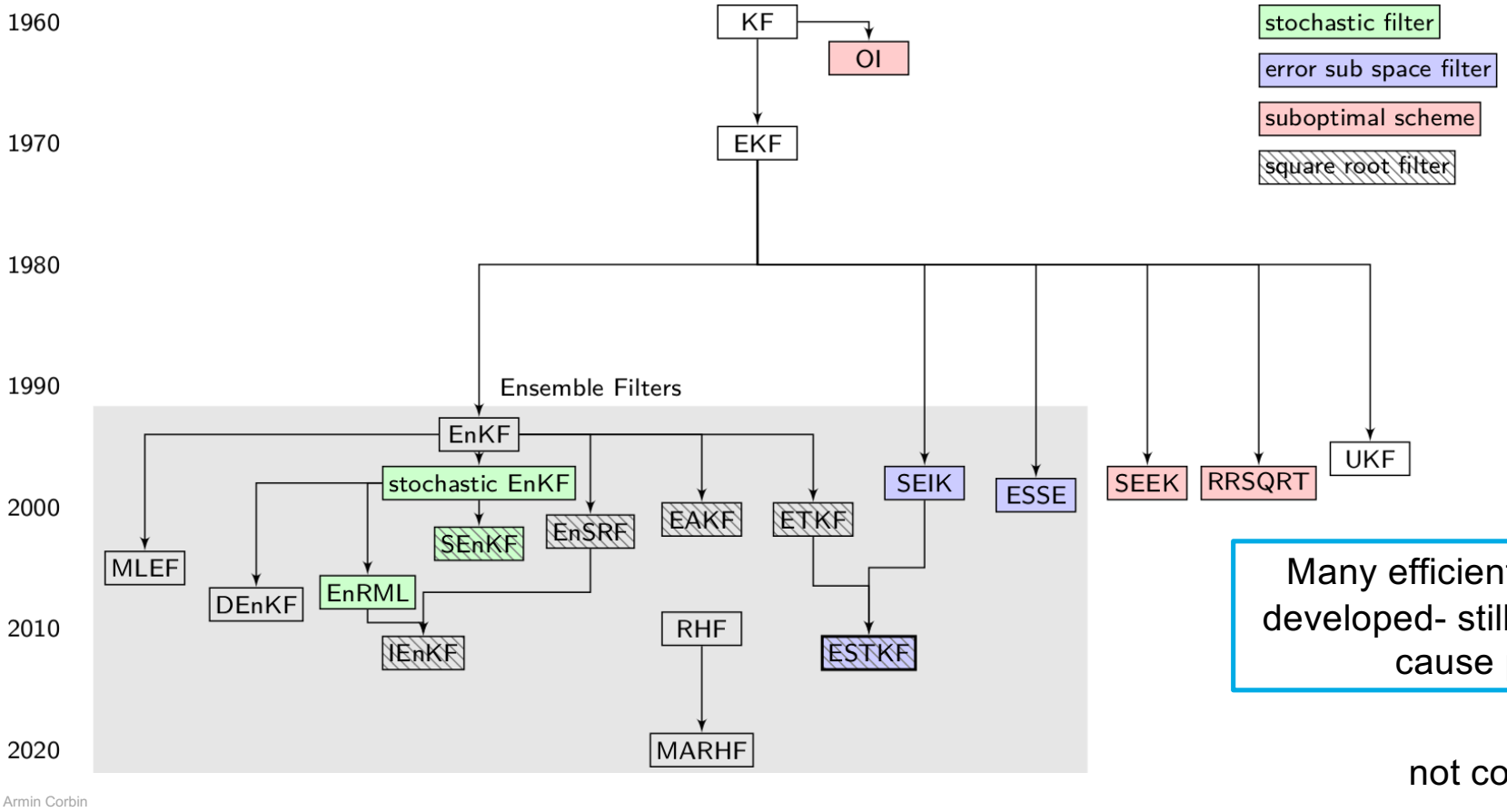
$$\bar{x} = \frac{1}{n} X I$$

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



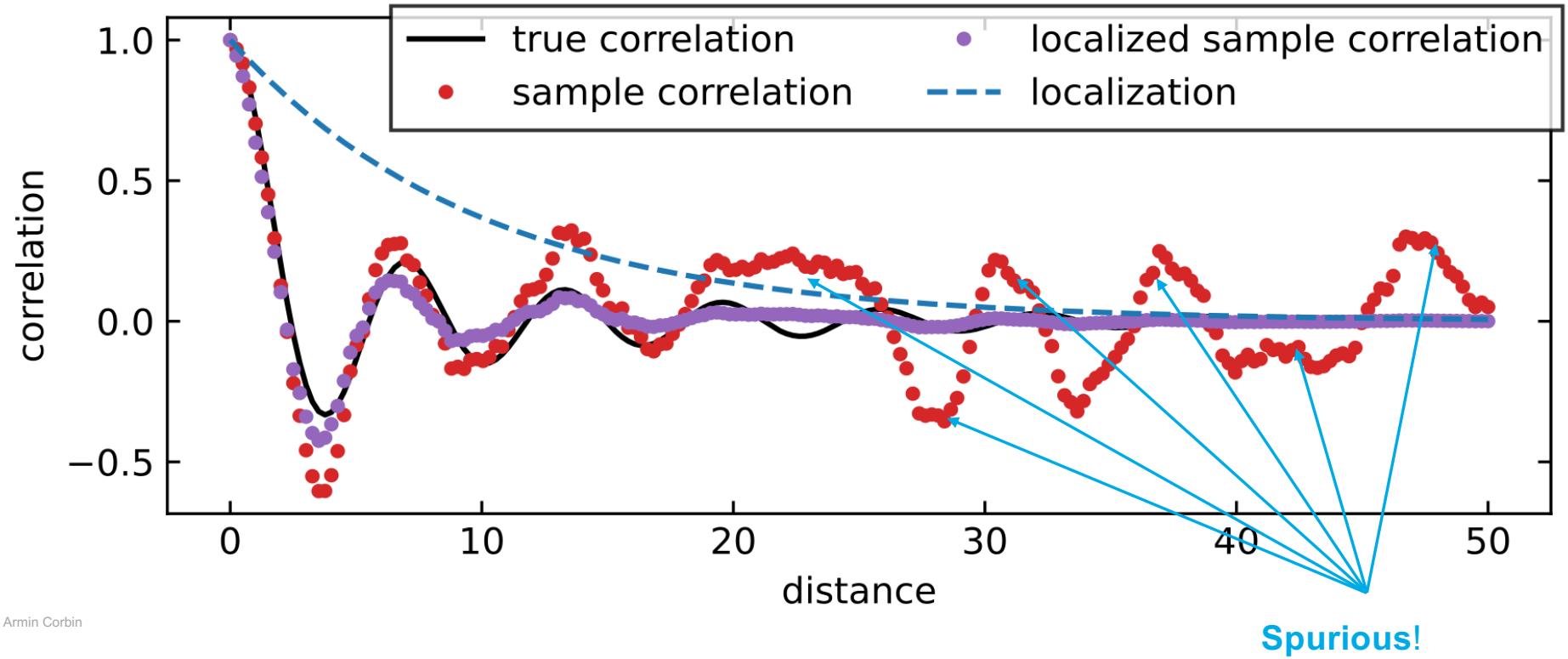
Many efficient schemes were developed- still high dimensions cause problems

not complete

Covariance Localization

PDAF Parallel Data Assimilation Framework

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

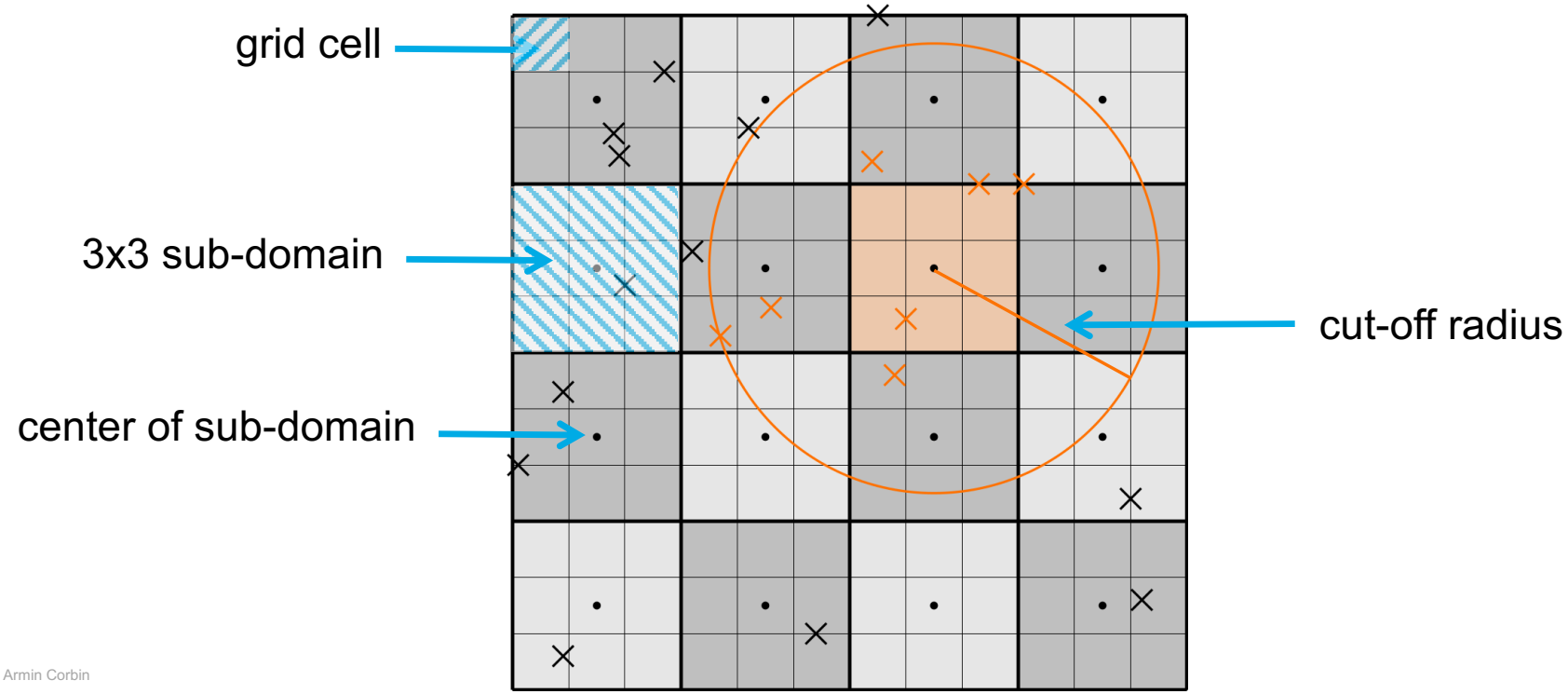


Armin Corbin

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

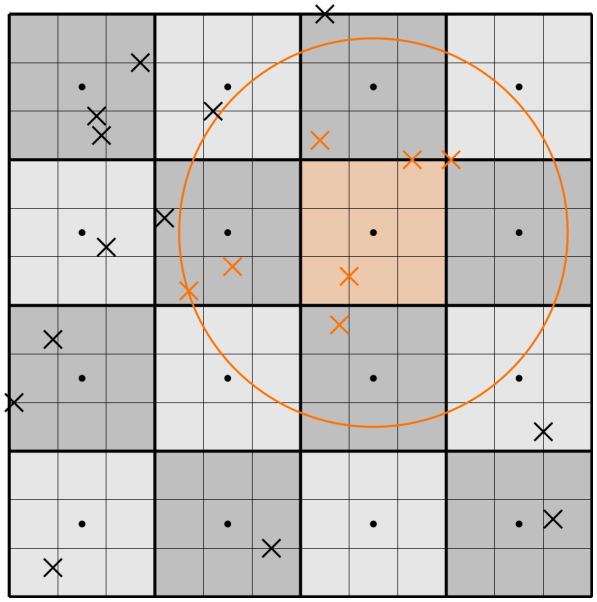


Armin Corbin

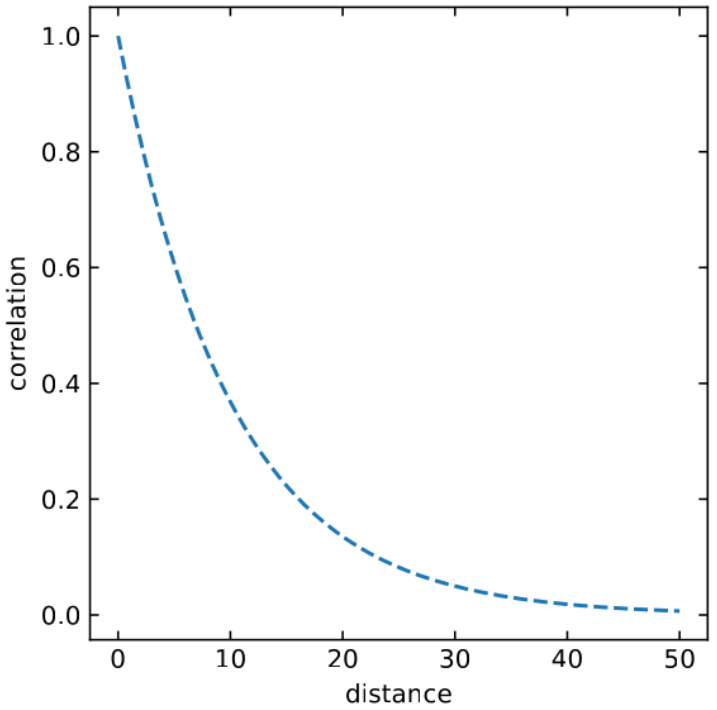
Observation Localization

PDAF Parallel Data Assimilation Framework

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



+



Inflation

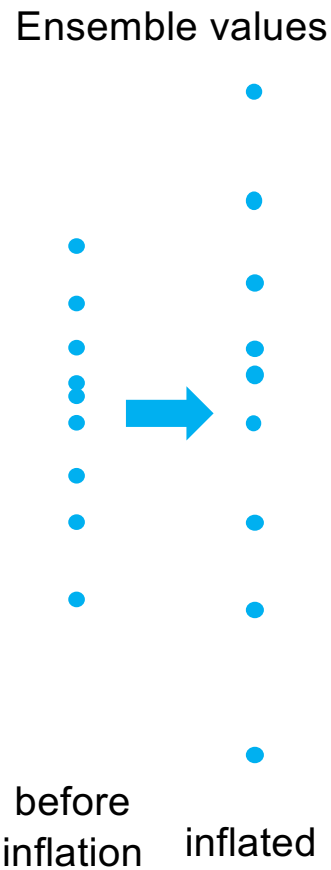
PDAF Parallel Data Assimilation Framework

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

Inflation → Increase error estimate for model

- Possibilities:
  - Increase ensemble spread by constant factor
  - Add some chosen variance
  - Relax spread of analysis ensemble to value before analysis update

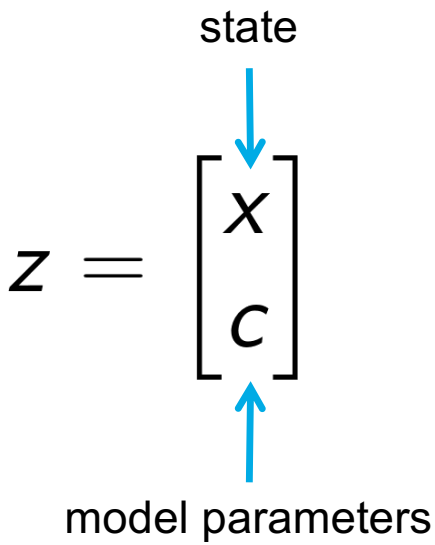
- Needs to be experimentally tuned



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

Soil moisture and water fluxes

PDAF Parallel Data Assimilation Framework

Observations

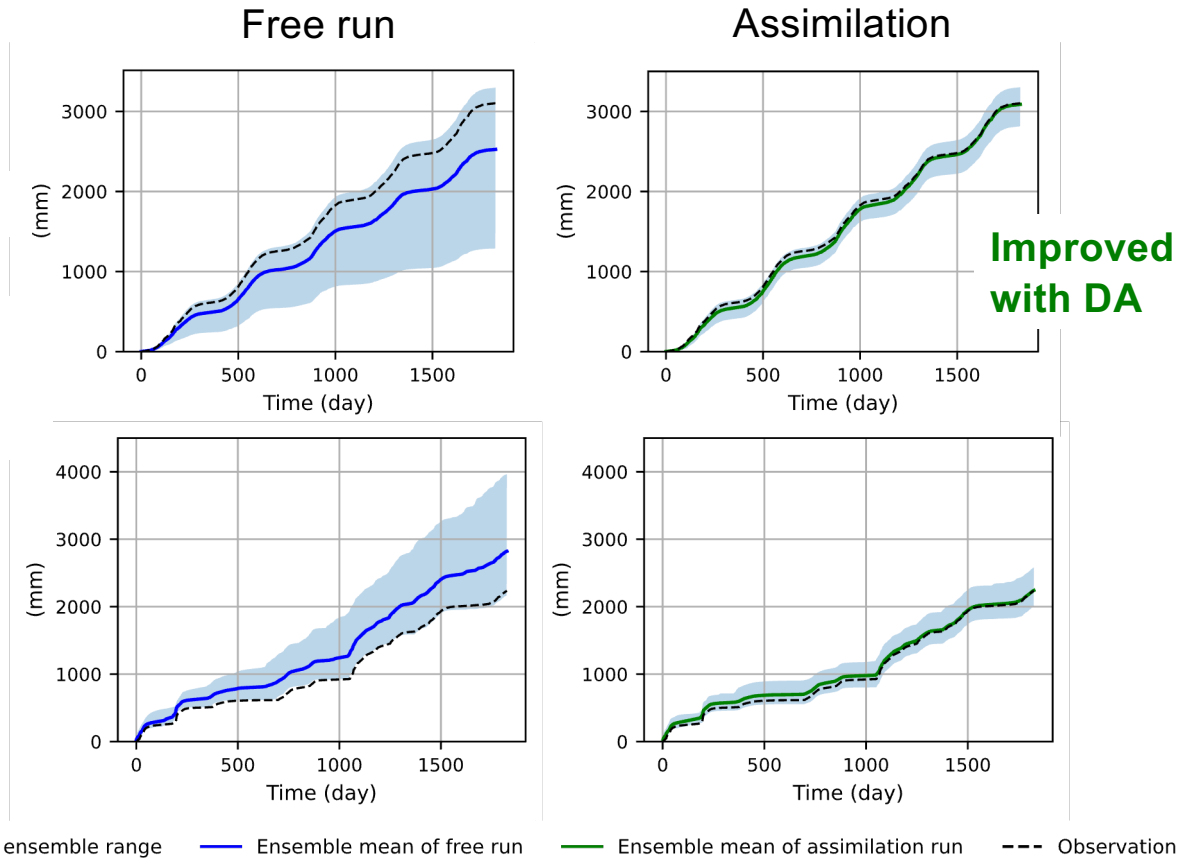
In situ soil moisture sensor measuring volumetric water content

Field scale hydrological model, Daily assimilation over 5 years

Evapotranspiration

Large deviation and uncertainty without DA

Potential groundwater recharge



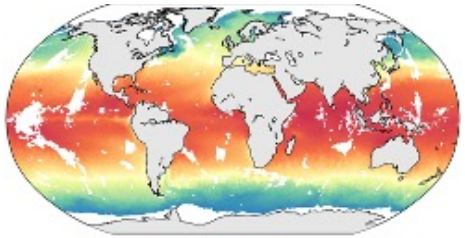
Improved with DA

Coupled **ocean-atmosphere** DA

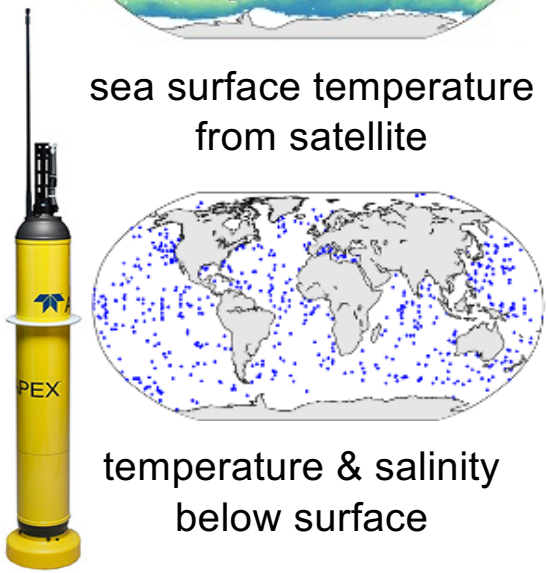
PDAF Parallel Data Assimilation Framework

Observations

Global climate model, Daily assimilation over 1 year



sea surface temperature from satellite



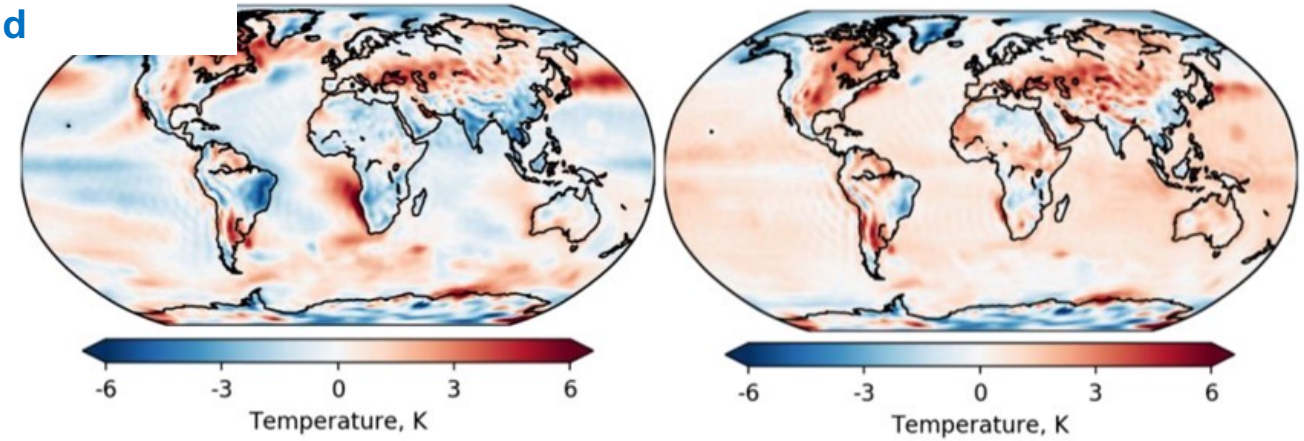
temperature & salinity below surface

Large deviations over the ocean and land

Free run

Assimilation

Lower deviations with DA



Average difference (model simulation - ERA-interim) of temperature at 2 m above sea surface

→ **Better fit to independent data (in atmosphere and ocean)**

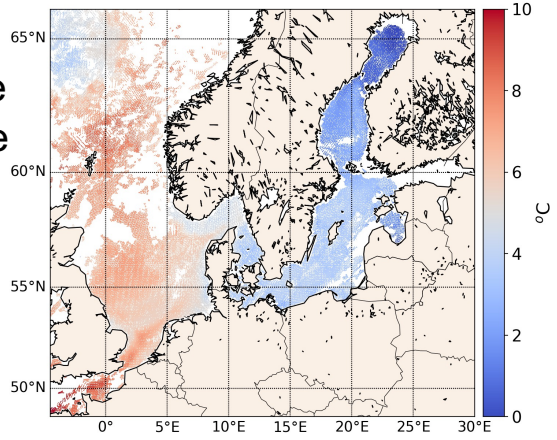
Nerger et al. (2020): Efficient ensemble data assimilation for coupled models with the Parallel Data Assimilation Framework: example of AWI-CM, Geosci. Model Dev.
Tang et al. (2021): Strongly coupled data assimilation of ocean observations into an ocean-atmosphere model. Geophysical Research Letters

Improving regional **ocean** predictions

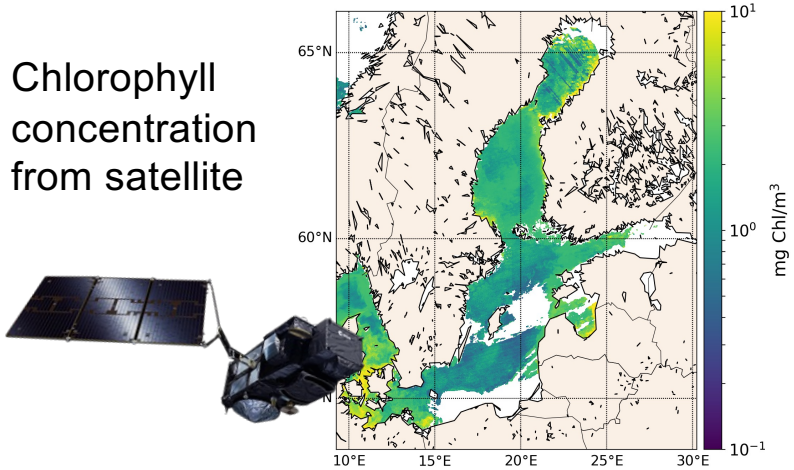
PDAF Parallel Data Assimilation Framework

Observations

Sea surface temperature from satellites



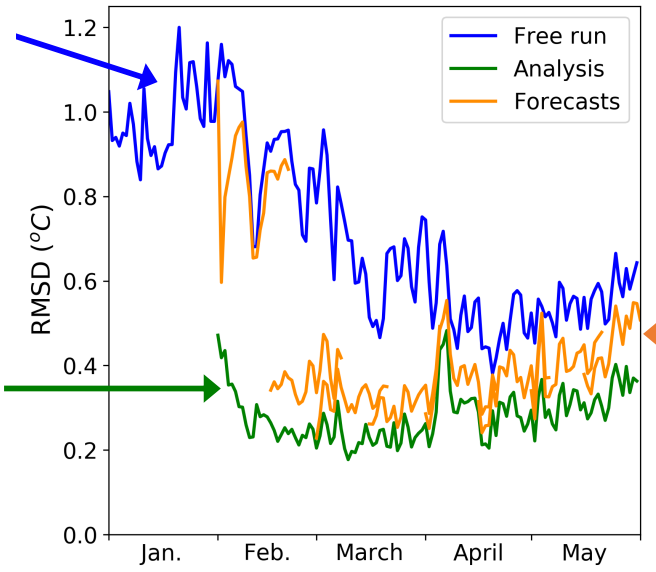
Chlorophyll concentration from satellite



High-resolution model of physics and biology,
Daily assimilation over 4 months

Large deviation without DA

Improved temperature with DA



Improved 3-week forecasts

Root mean square error
of surface temperature over time

SEAMLESS

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

PDAF Parallel Data Assimilation Framework

Post-doc in Ocean Data Assimilation

at AWI

European project with 11 partners

Advance ensemble DA methods
Contribute to advance PDAF
Apply ensemble DA to real ocean model

https://**jobs.awi.de**/Vacancies/2168/Description/2

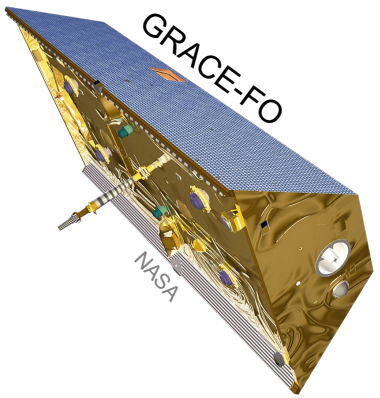
Application deadline: May 22

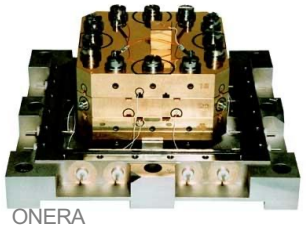
Mass Density in the **Upper Atmosphere**

PDAF Parallel Data Assimilation Framework

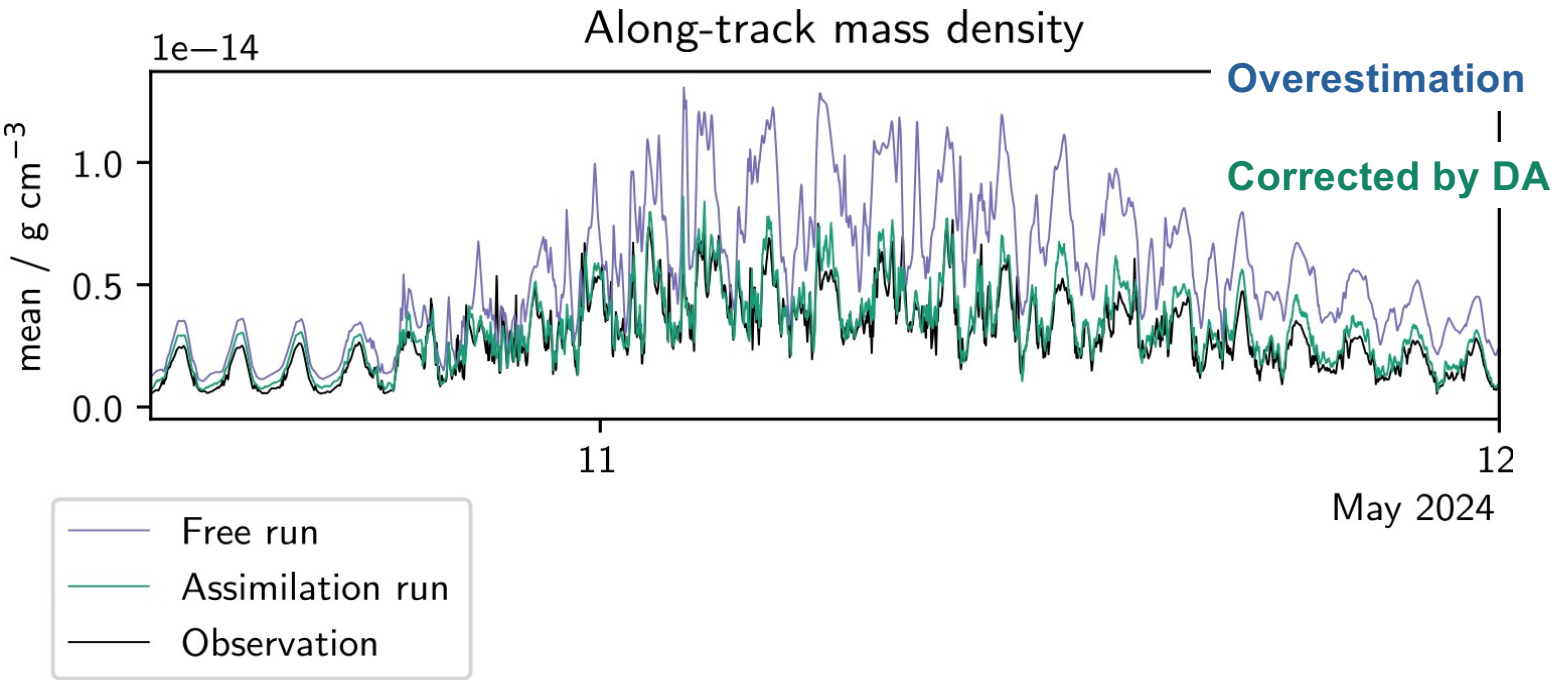
Observations

Global model of the upper atmosphere,
Assimilation every minute over 1 month





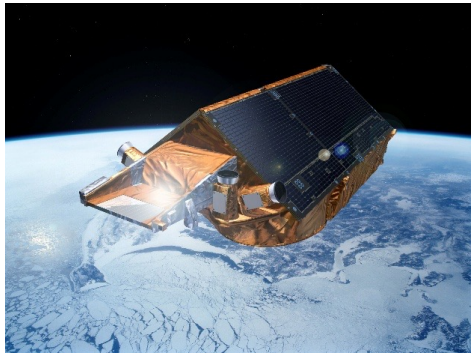
accelerometer



Arctic **Sea Ice**

PDAF Parallel Data Assimilation Framework

Observations

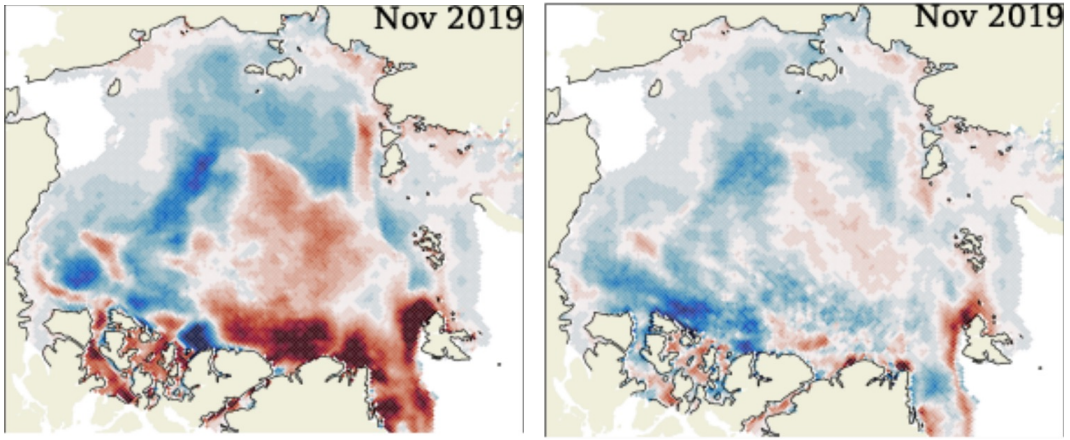


CryoSat-2
Credit: ESA

Advanced Microwave Scanning Radiometer 2 (AMSR2)
Credit: JAXA

Sea ice model, neXtSIM with daily concentration and weekly thickness assimilation

Free Run | **Lower deviations with DA**



Monthly difference between forecast and observations

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

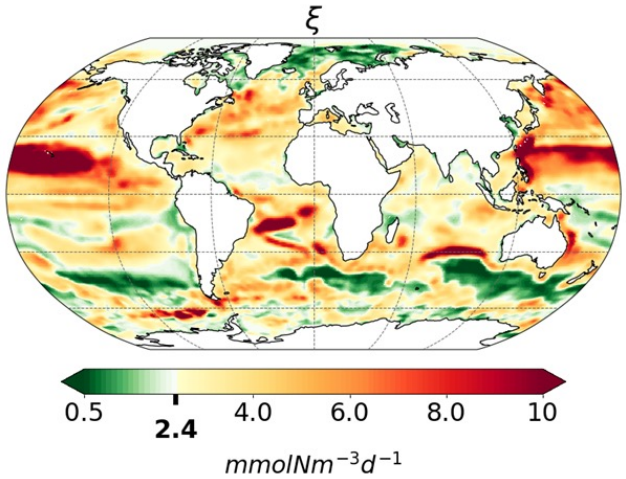
PDAF Parallel Data Assimilation Framework

Observations

Ocean Colour in the North Sea

Image: NASA Earth Observatory

Spatially Varying Parameter Estimates



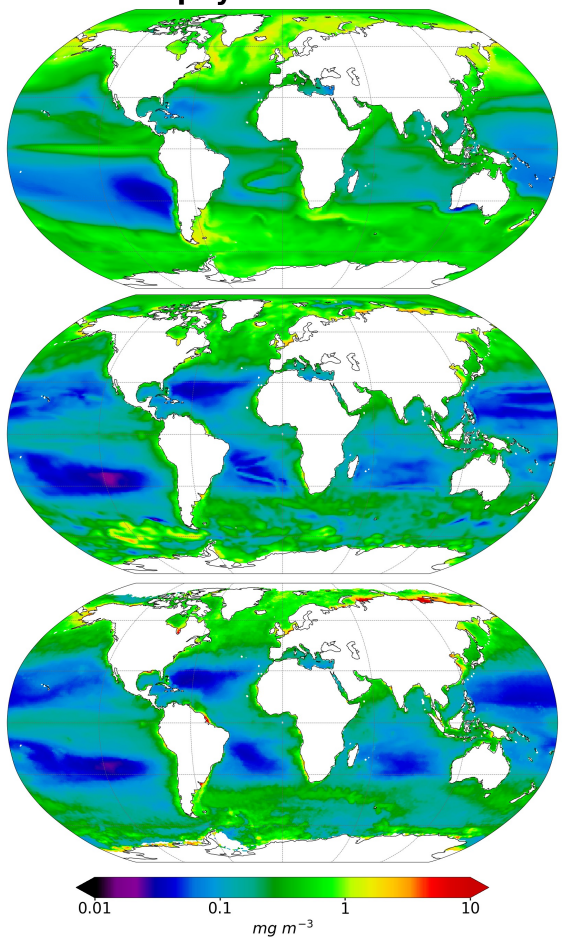
Default parameter

Mean surface chlorophyll concentration for 2019

Simulation with fixed parameter

Simulation with spatially resolved and estimated parameters

Satellite observations

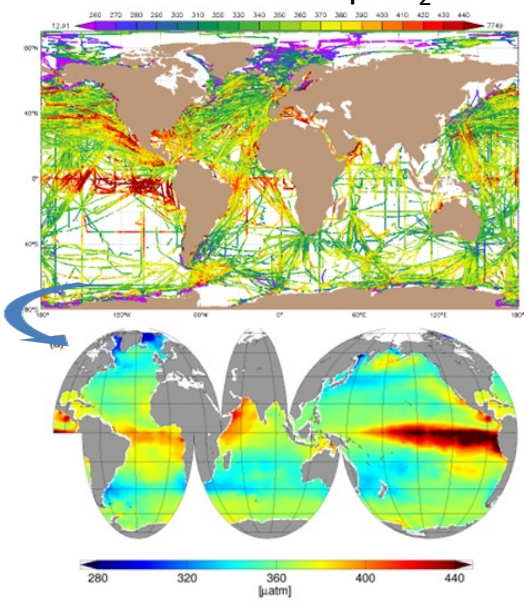


Global **Ocean** Biogeochemical Reanalysis

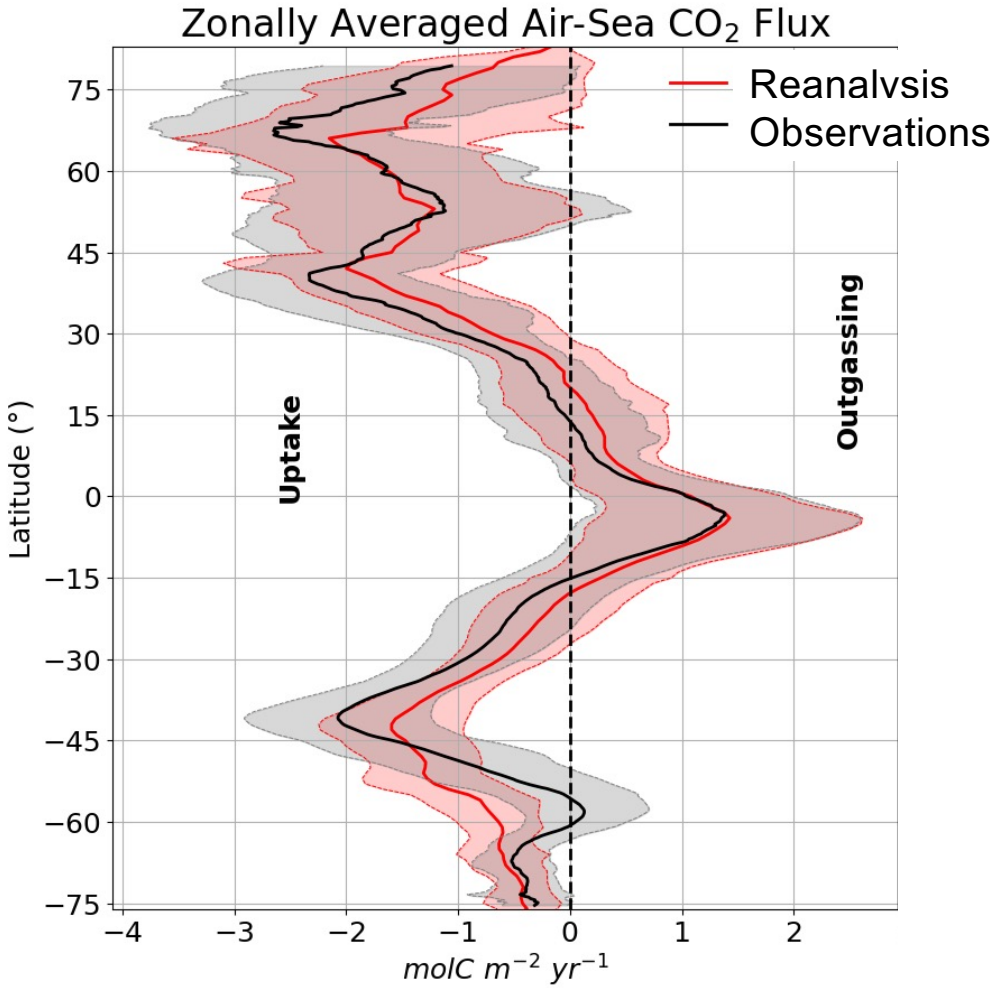
PDAF Parallel Data Assimilation Framework

Observations

Surface ocean pCO_2



Resplandy et al., 2018



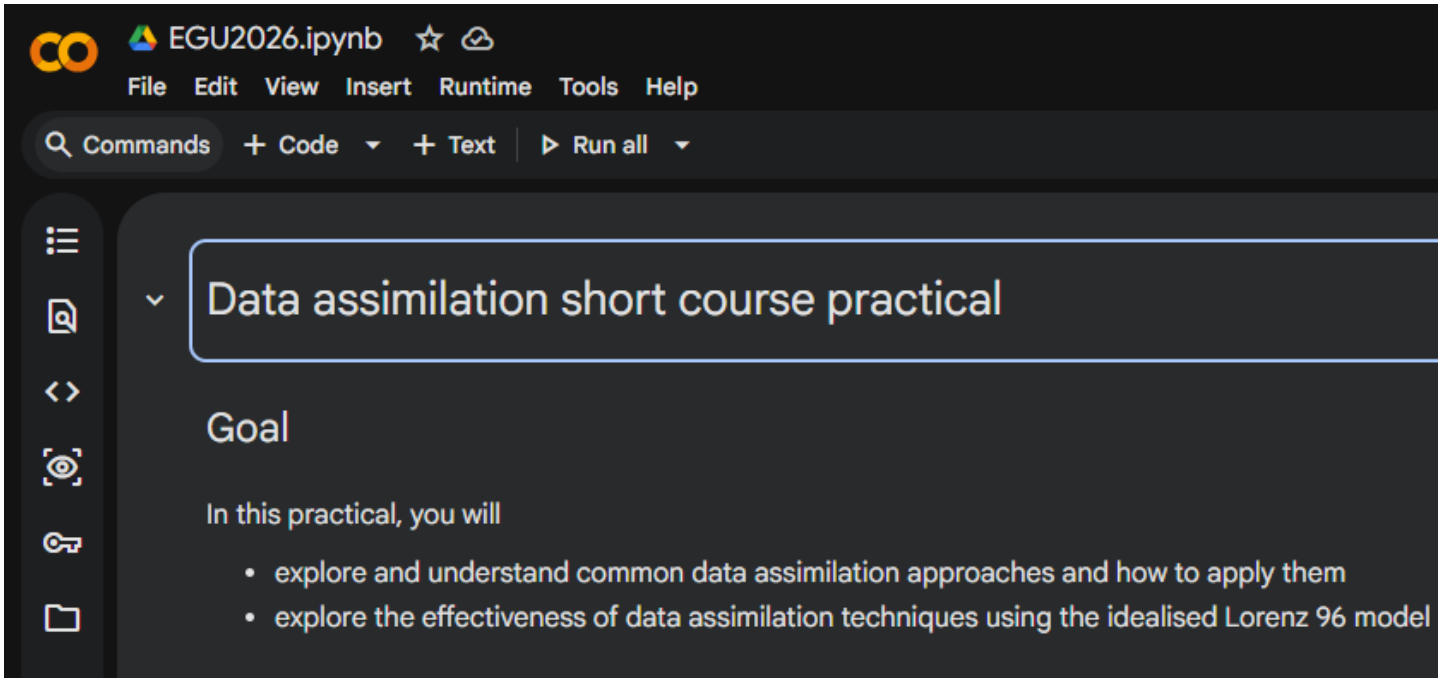
PDAF Parallel Data Assimilation Framework

III Hands On

A taste of DA on Lorenz 96 model

PDAF Parallel Data Assimilation Framework

- Hands on ensemble-based Kalman filter and 3DVar



Available on Google Colab:
https://tinyurl.com/egu26-da-sc

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

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

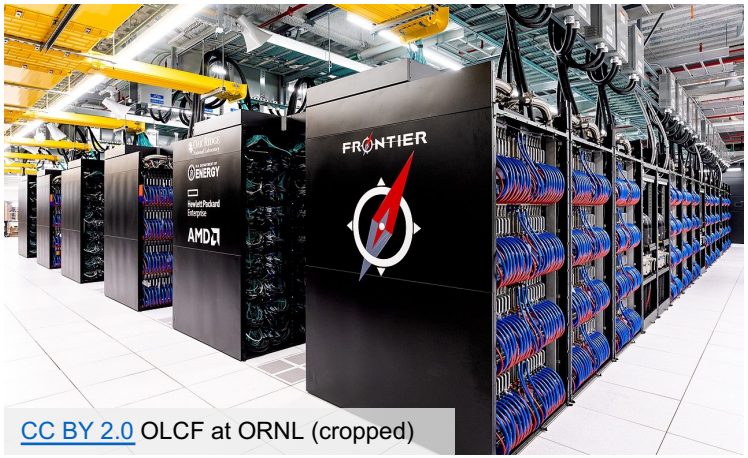
pdaf.awi.de

Memory Requirement

PDAF Parallel Data Assimilation Framework

Dimension of state vector	Required memory (double)	
	x	**P**
1 000 000	8 MB	8 TB
10 000 000	80 MB	800 TB
100 000 000	800 MB	80 000 TB

Combined memory of best super computer[1] accumulates to **9 200 TB**



Storing and multiplication of the covariance matrix of the state becomes too expensive

[1]According to TOP500 List of Nov 2023