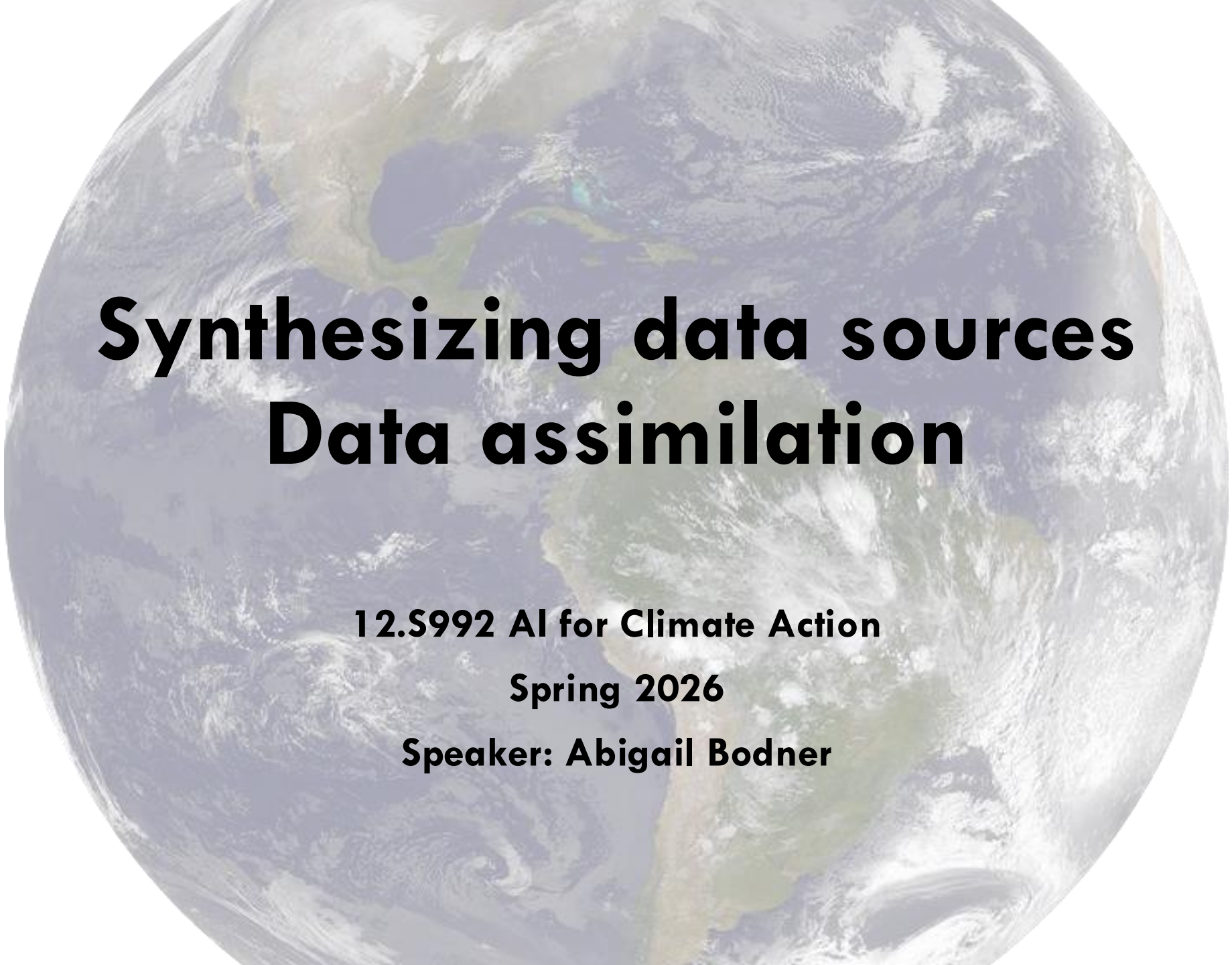
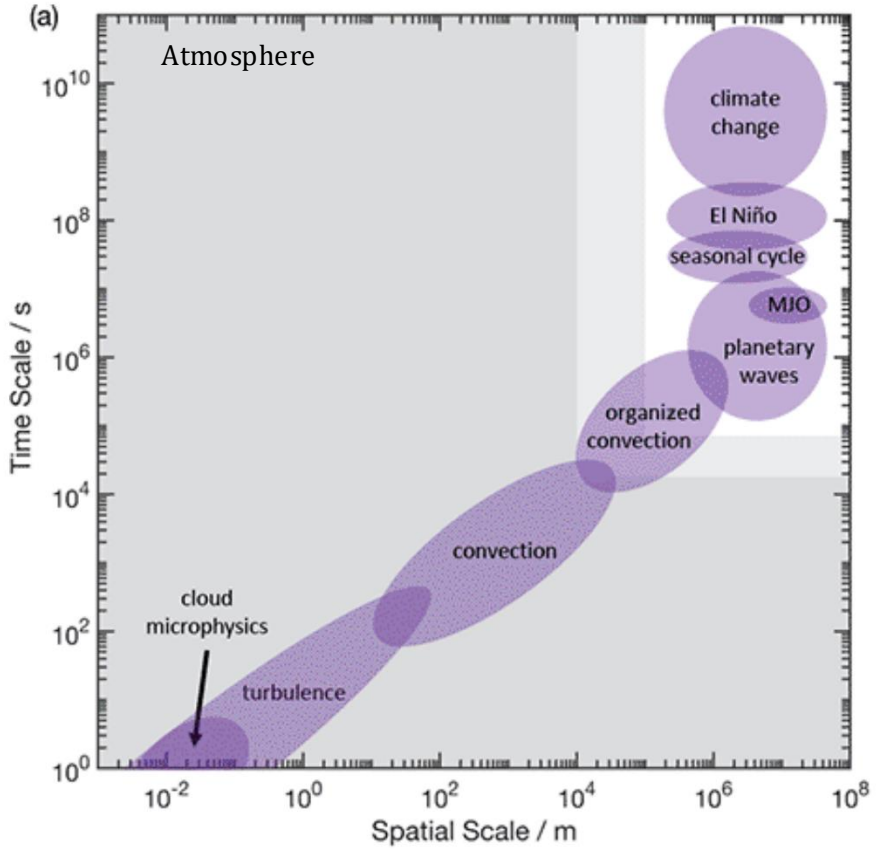


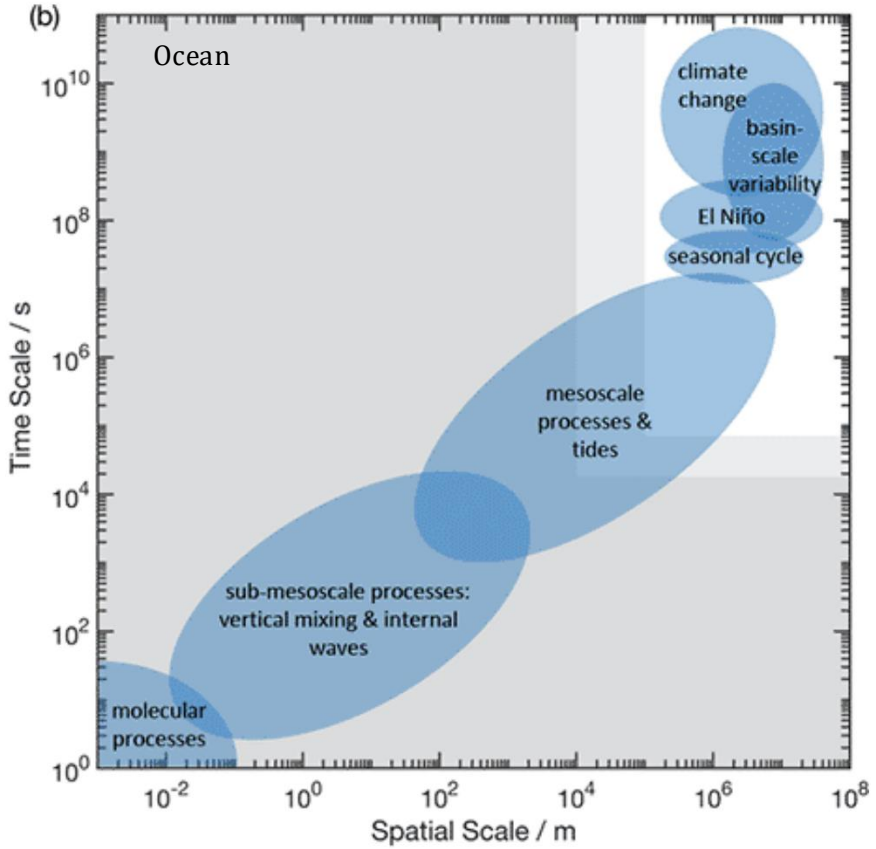
Synthesizing data sources
Data assimilation

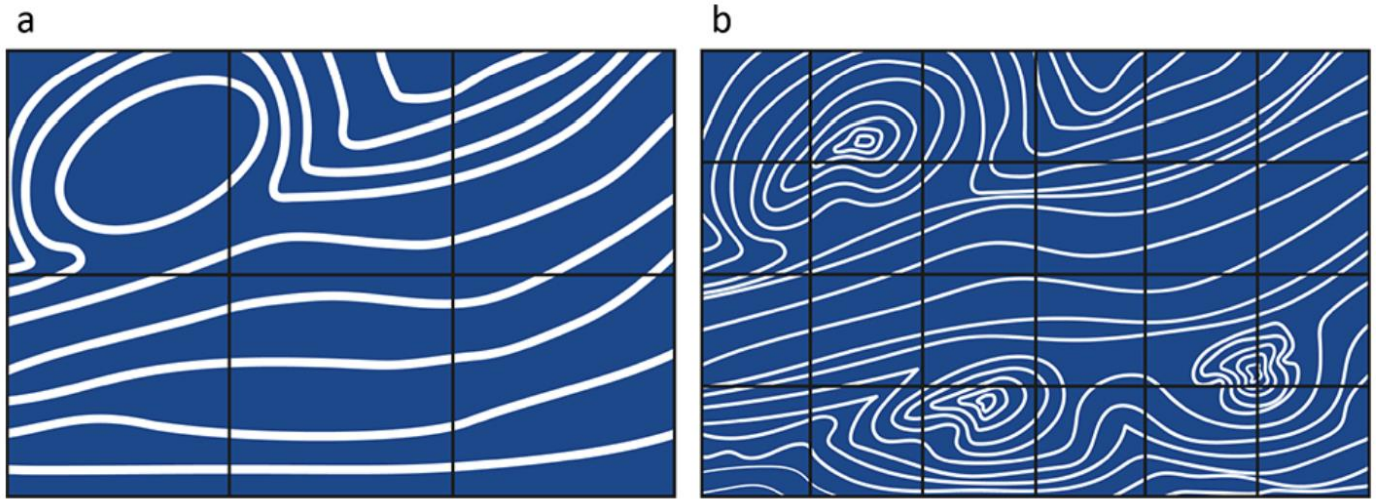
12.S992 AI for Climate Action

Spring 2026

Speaker: Abigail Bodner

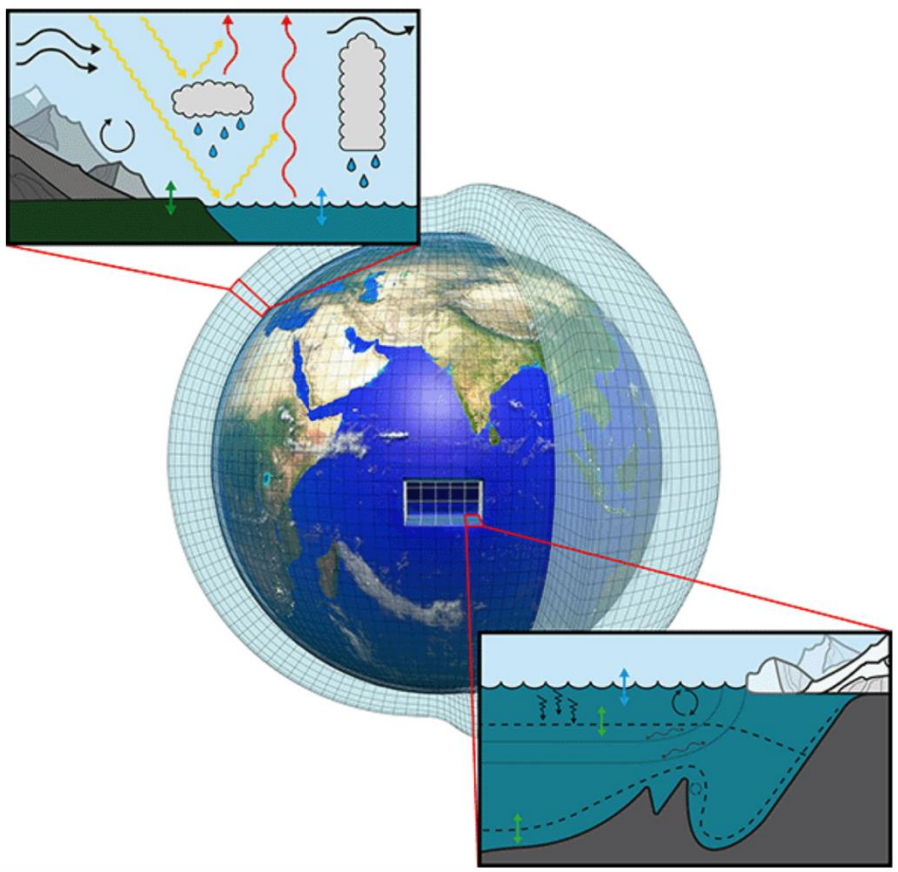


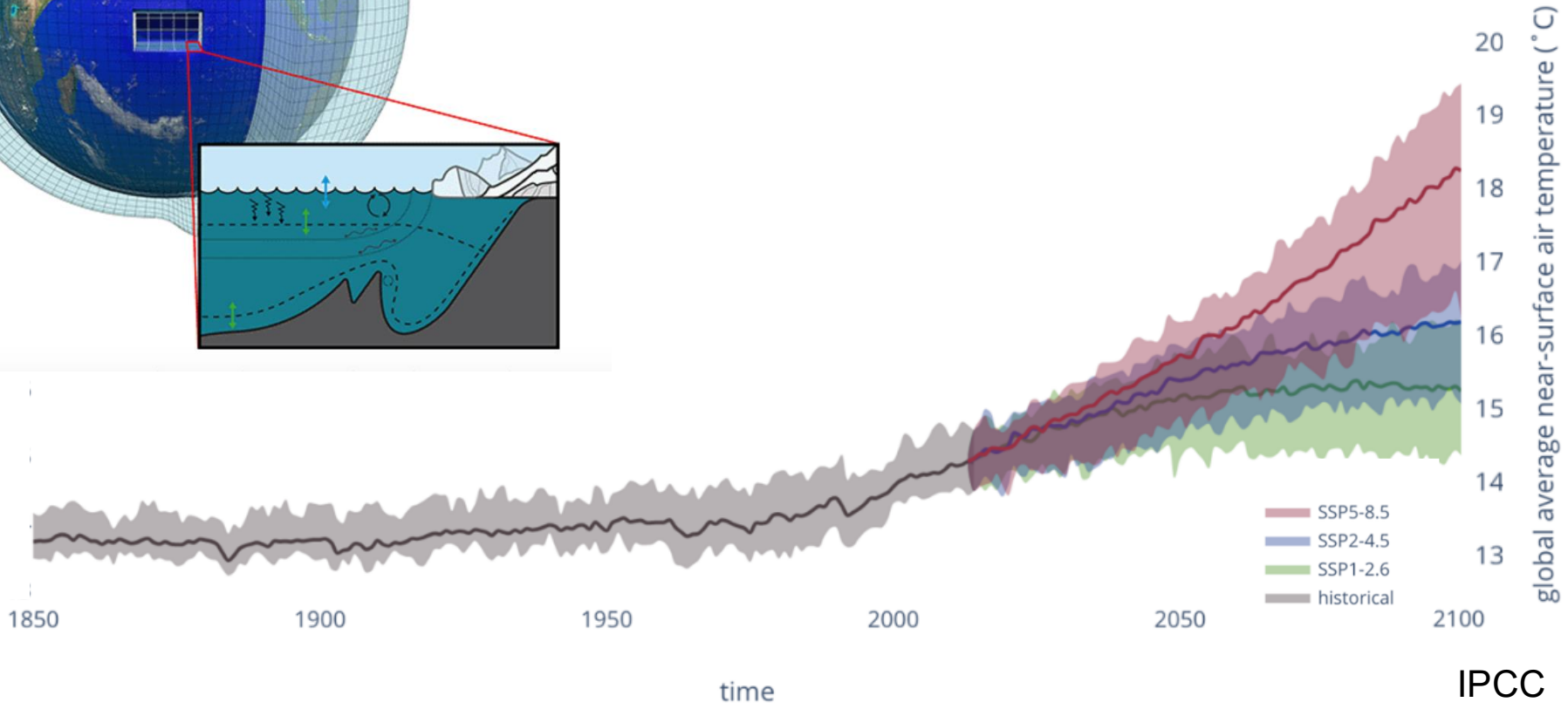




Christensen & Zanna (2022)

Global Climate Change



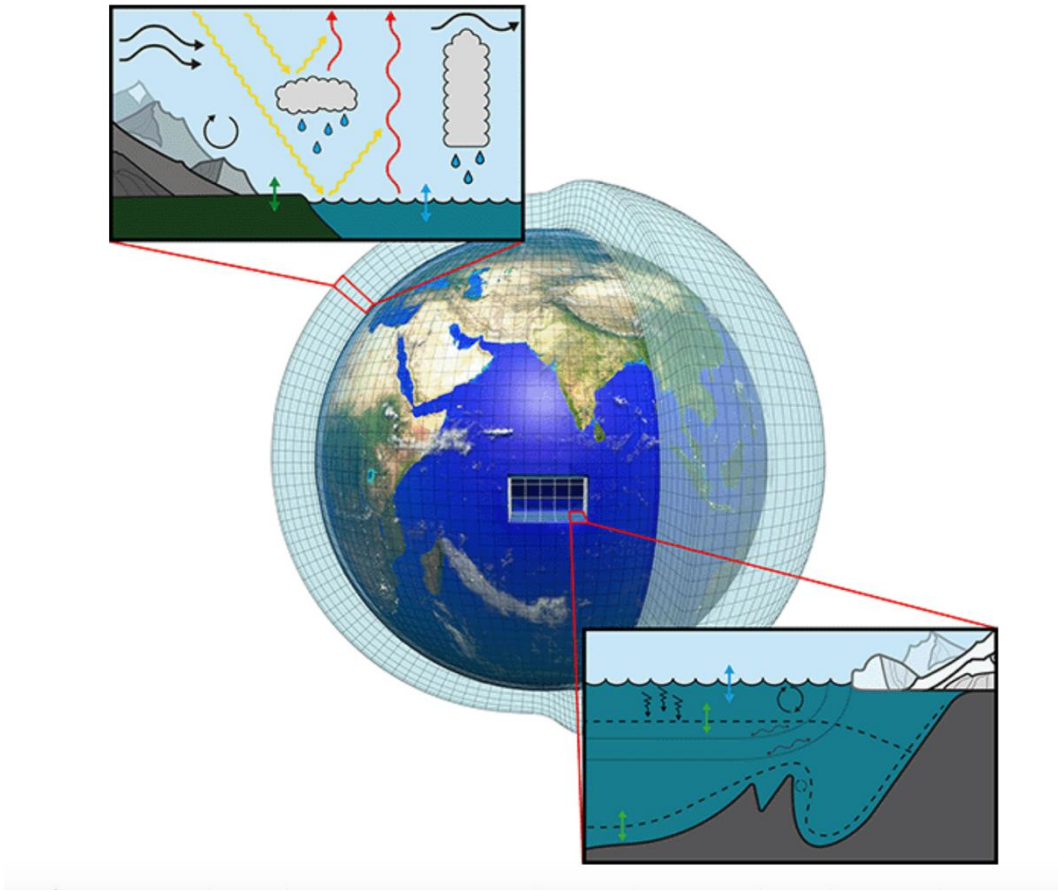


IPCC

Data assimilation (DA)

Also called reanalysis or state estimate

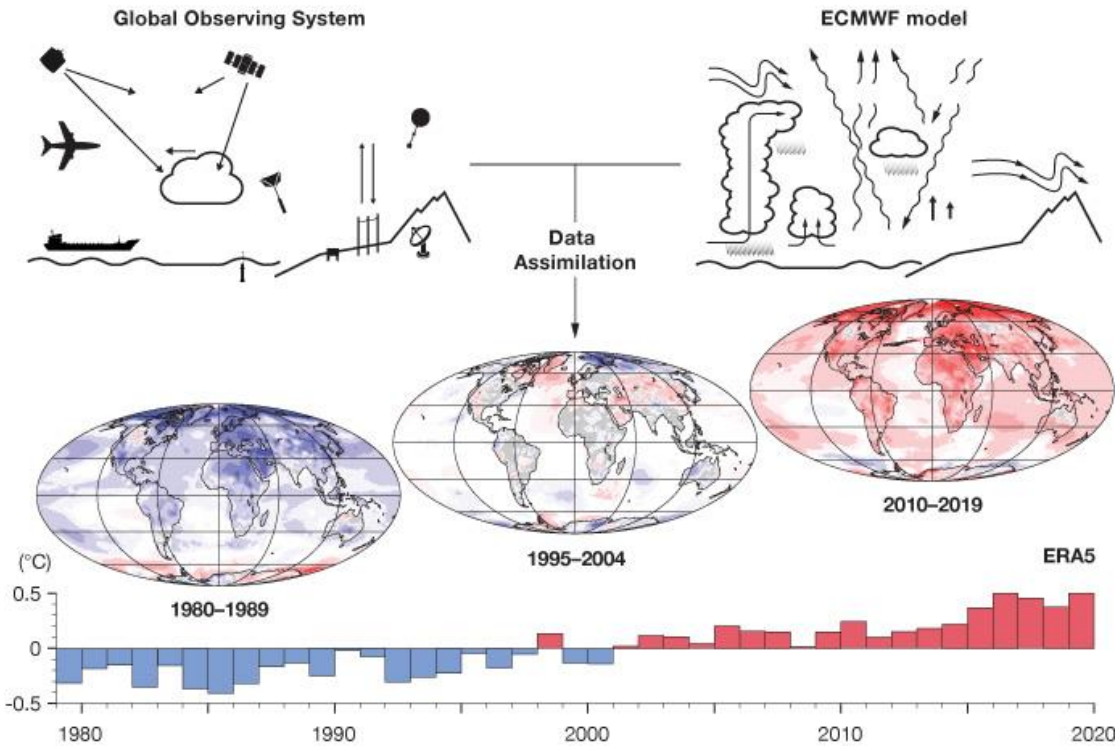
Unlike the parameterization problem, which solves the equations of motion and infers subgrid effects, in DA the goal is to create a gridded product from all available data sources.



Data assimilation (DA)

Also called reanalysis or state estimate

Unlike the parameterization problem, which solves the equations of motion and infers subgrid effects, in DA the goal is to create a gridded product from all available data sources.

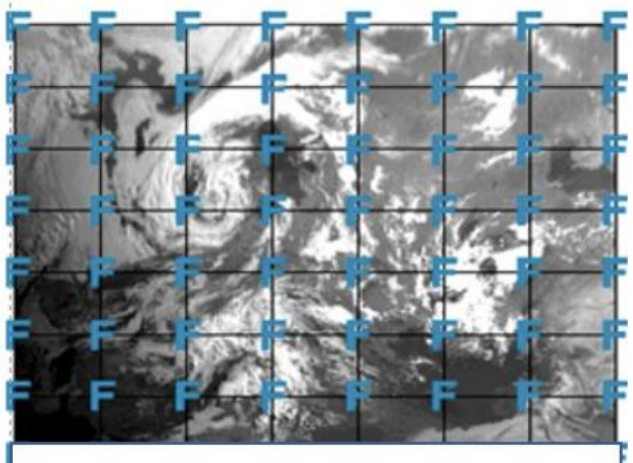


Data assimilation (DA)

Quantify the state of the system (ocean/atmosphere/land/cryosphere) given all available information from:

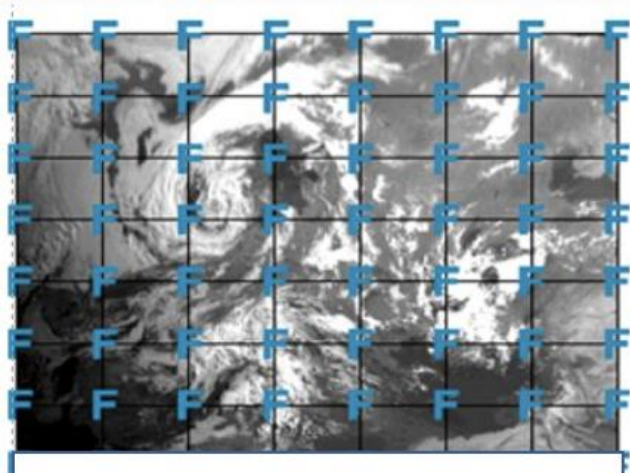
- Models
- Observations

The data assimilation process

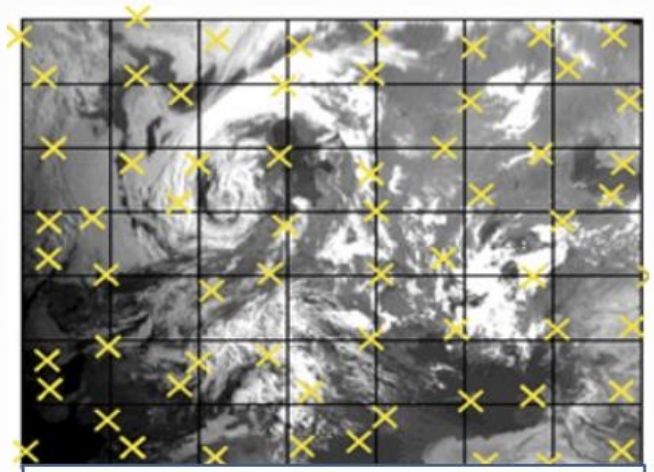


Model forecast (with errors)

The data assimilation process

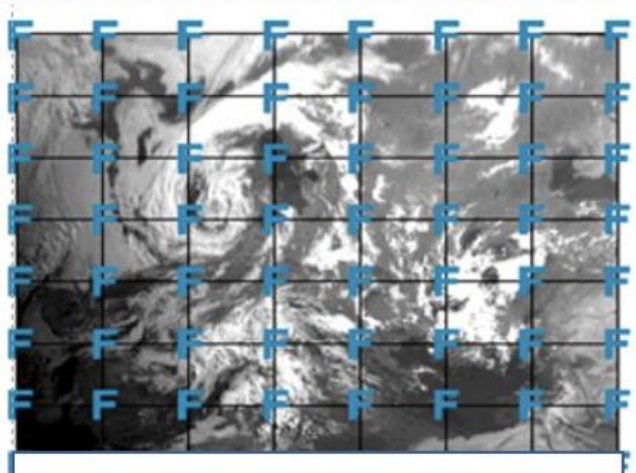


Model forecast (with errors)

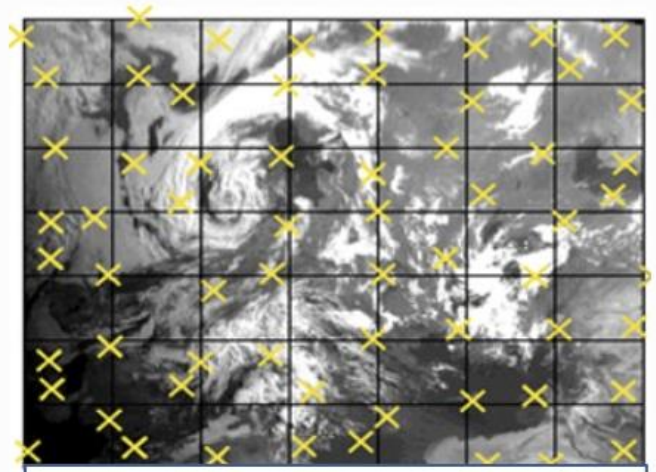


Observations (with errors)

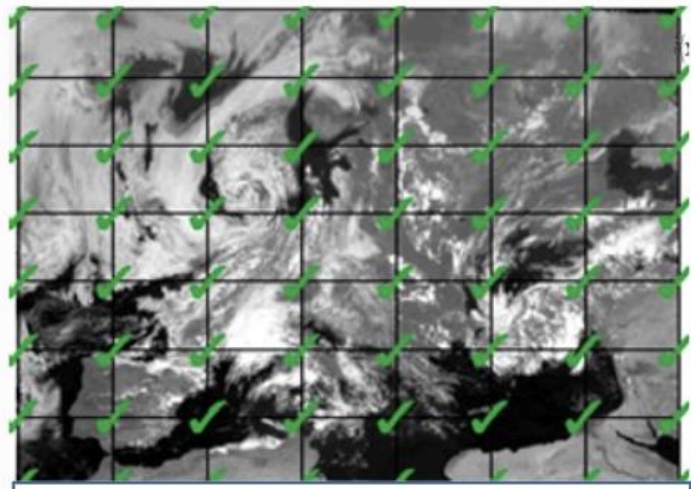
The data assimilation process



Model forecast (with errors)



Observations (with errors)



Analysis (with smaller errors)

Data assimilation (DA)

Quantify the state of the system (ocean/atmosphere/land/cryosphere) given all available information from:

- Models
- Observations

Challenges:

Data assimilation (DA)

Quantify the state of the system (ocean/atmosphere/land/cryosphere) given all available information from:

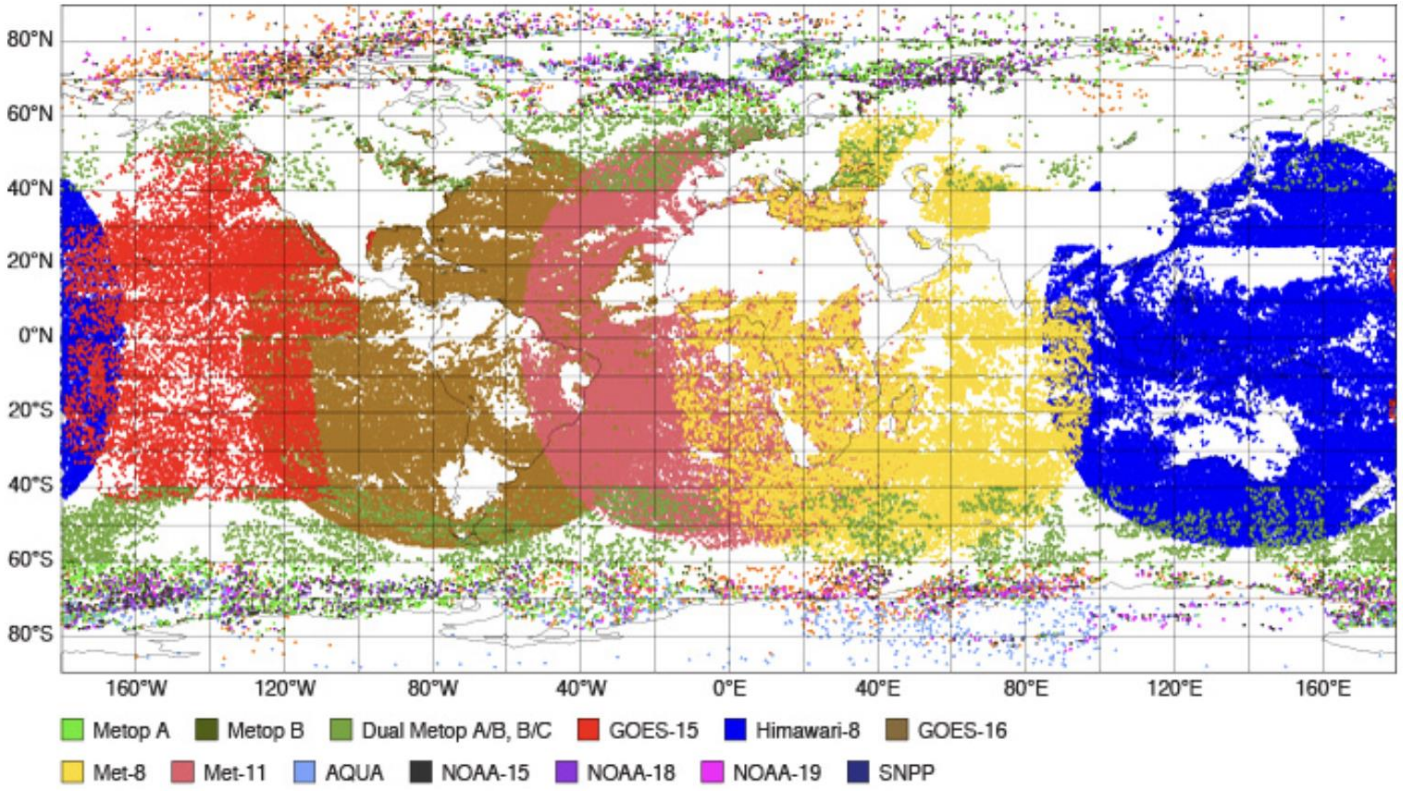
- Models
- Observations

Challenges:

- What is the best data (error estimates)
- How to synthesize data from different sources, i.e. time, space, resolution, etc.

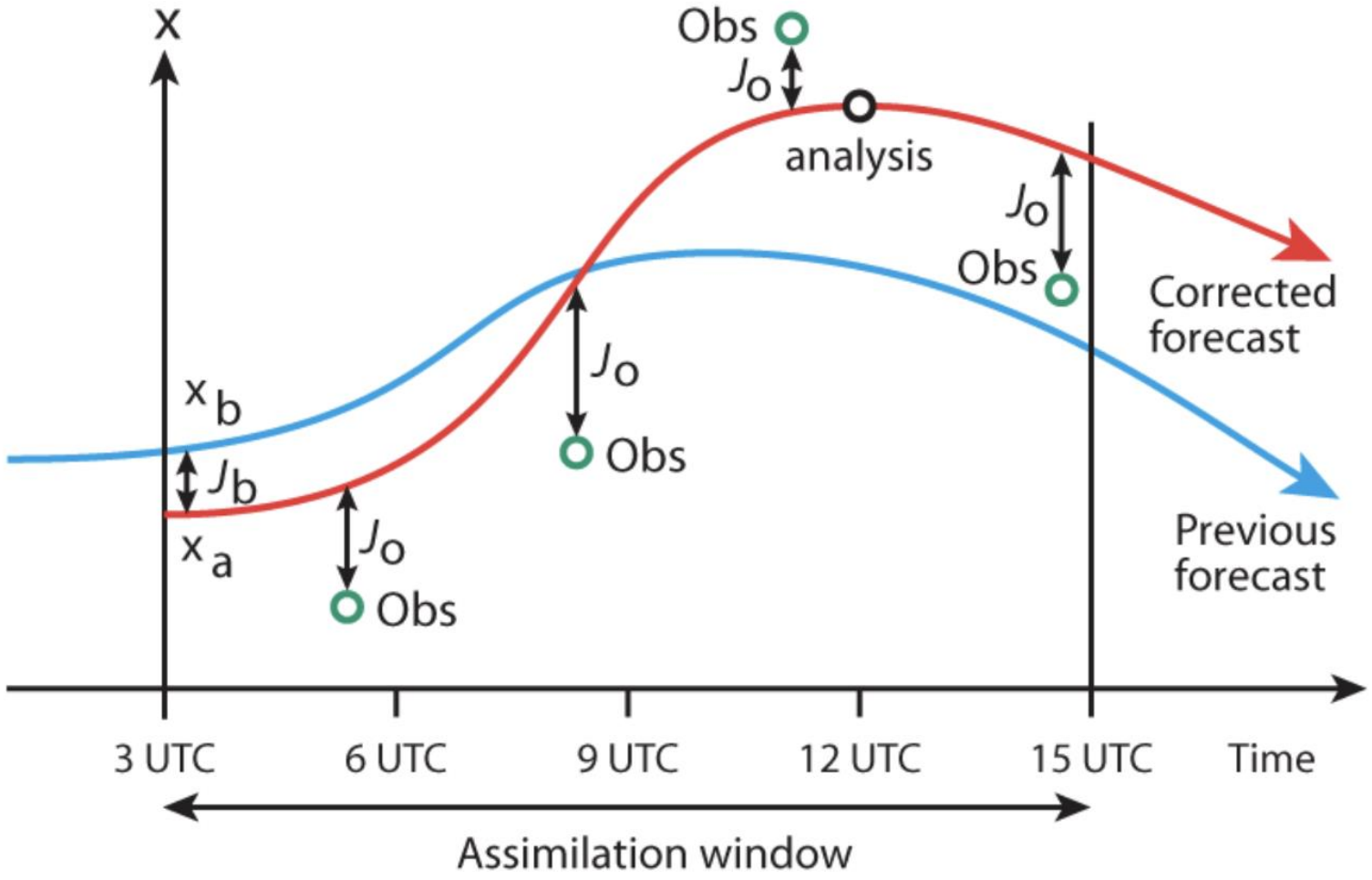
Example: Forecasting (4D-Var)

Weather observations come from many sources, but they cannot provide a complete picture of the state of the Earth system at a given point in time. (Diagram: WMO)



Typical coverage of active atmospheric motion vector (AMV) data for a 12-hour assimilation cycle (00 UTC, 7 March 2019).

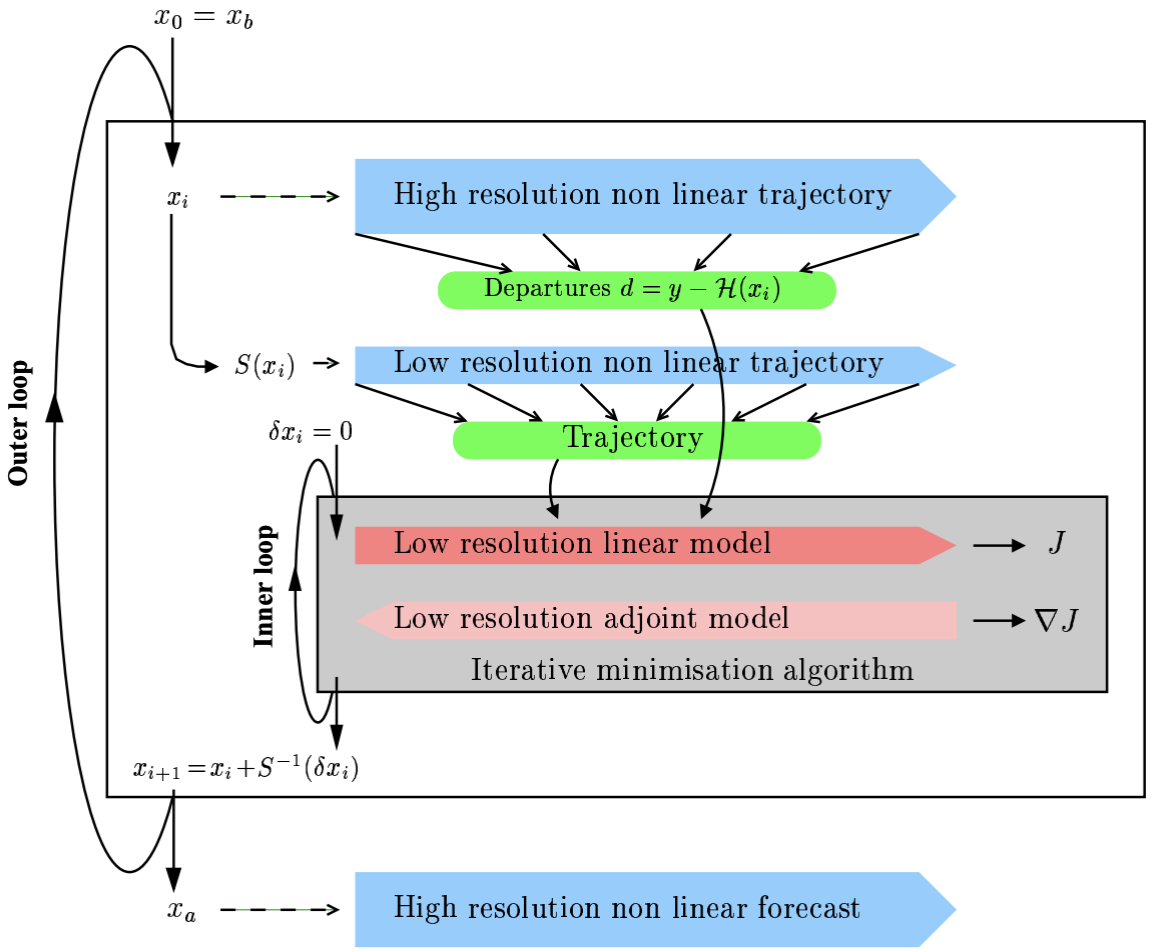
Example: Forecasting (4D-Var)



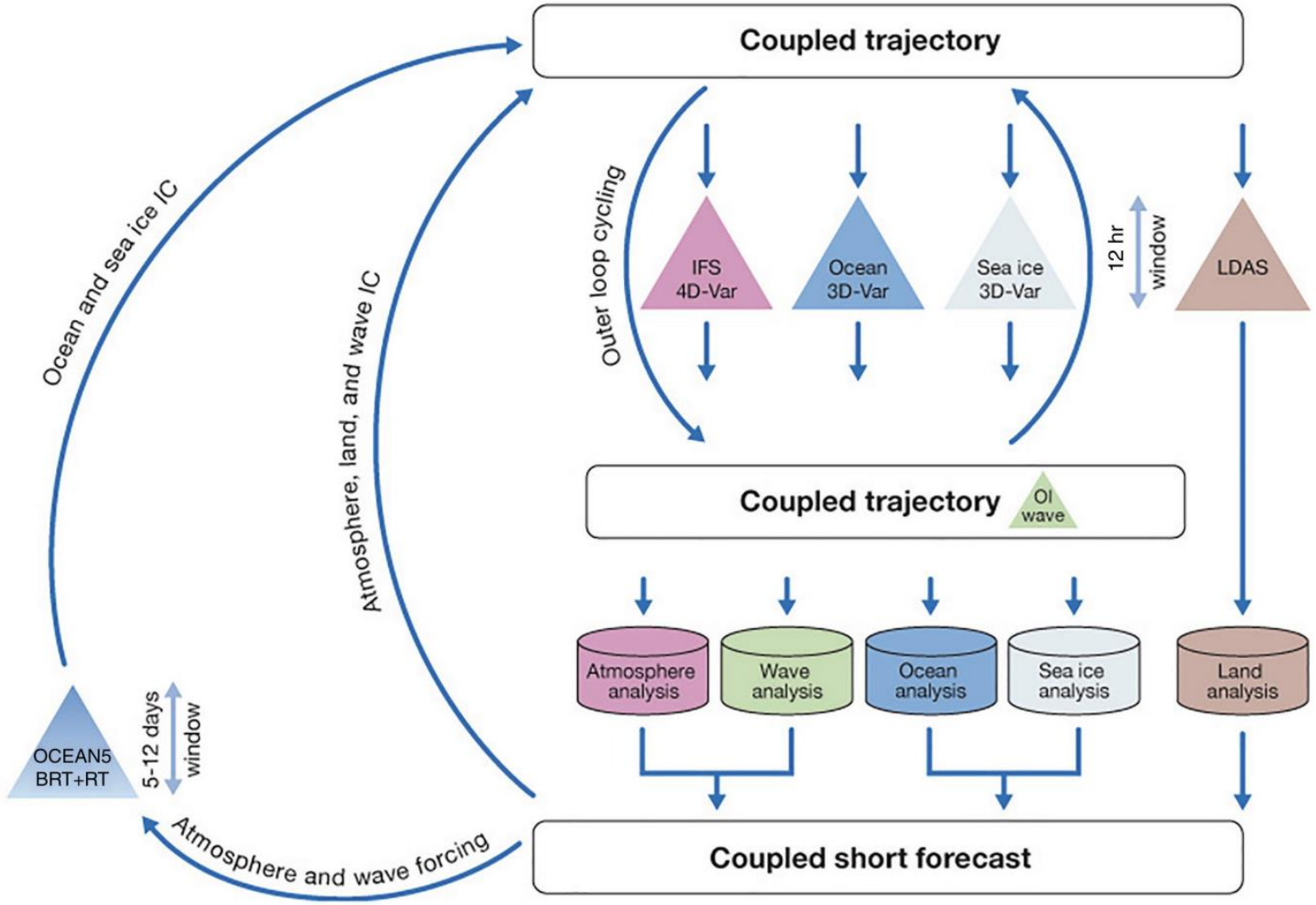
Example: Forecasting (4D-Var)

$$
\begin{aligned}
J(\mathbf{x}) &= \tfrac{1}{2}(\mathbf{x}-\mathbf{x}_\mathrm{b})^\mathrm{T}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_\mathrm{b}) \\
&\quad + \tfrac{1}{2}\sum_k (\mathscr{H}_k\mathscr{M}_k(\mathbf{x})-\mathbf{y}_k)^\mathrm{T}\mathbf{R}_k^{-1}(\mathscr{H}_k\mathscr{M}_k(\mathbf{x})-\mathbf{y}_k)
\end{aligned}
$$

$$
\begin{aligned}
\nabla J(\mathbf{x}) &= \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_\mathrm{b}) \\
&\quad + \sum_k \mathbf{M}_k^\mathrm{T}\mathbf{H}_k^\mathrm{T}\mathbf{R}_k^{-1}(\mathscr{H}_k\mathscr{M}_k(\mathbf{x})-\mathbf{y}_k).
\end{aligned}
$$

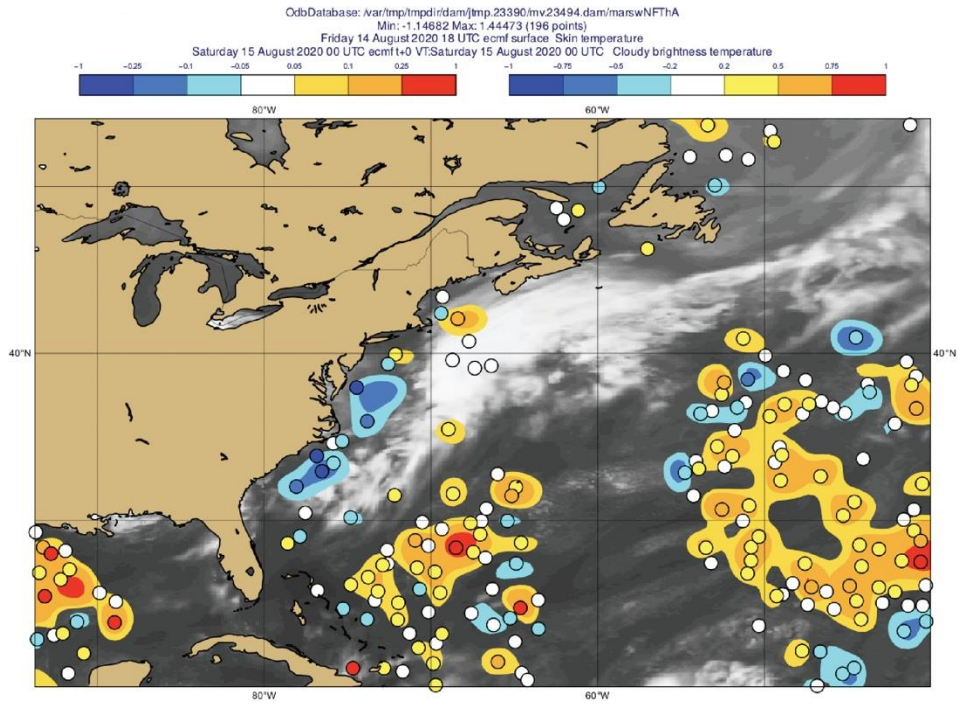


Example: Forecasting (4D-Var)



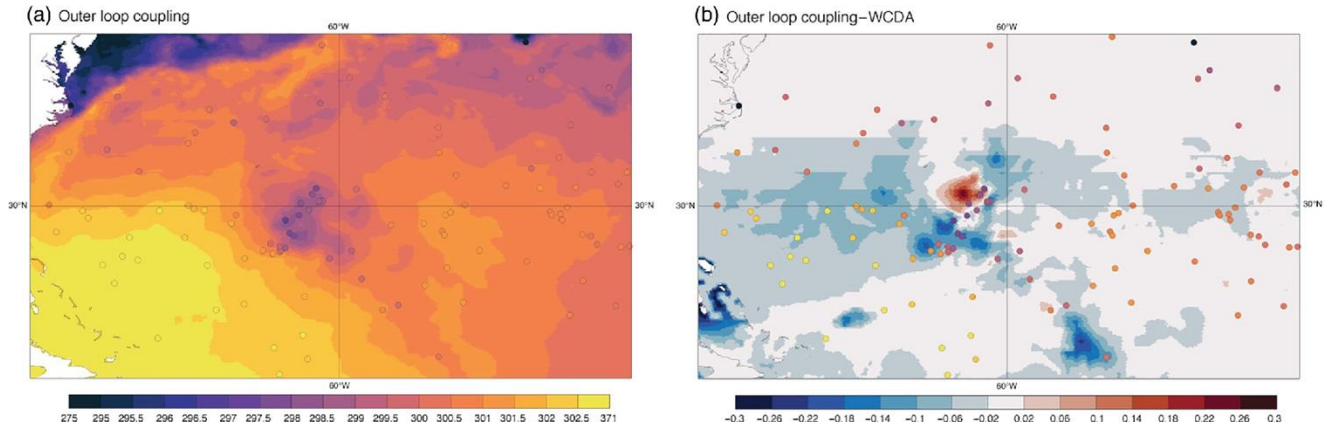
ECMWF
de Rosnay et al. (2022)

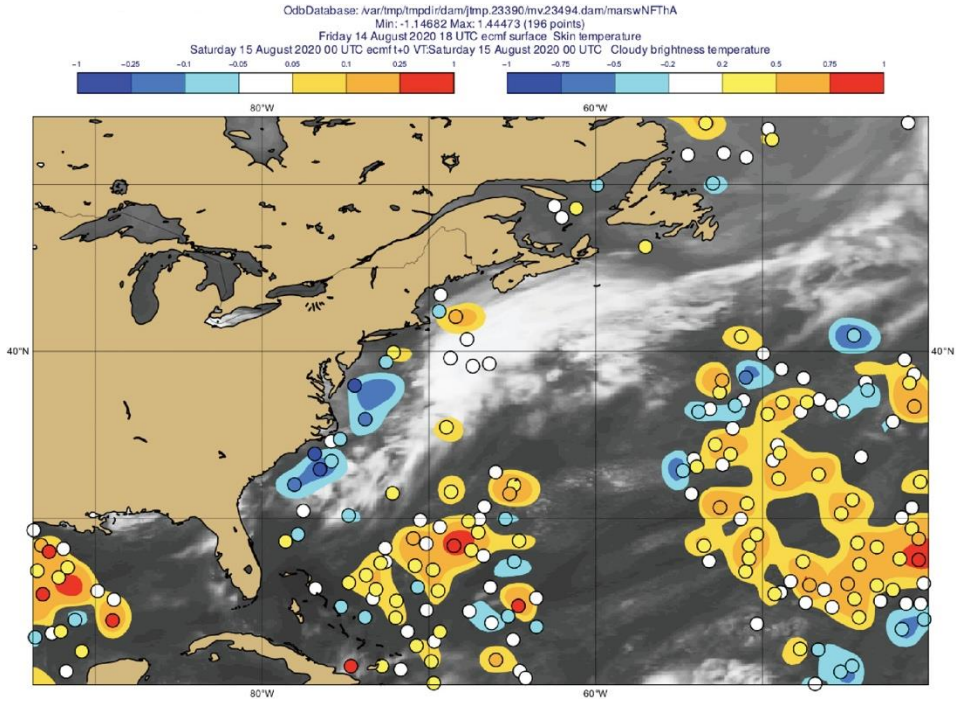
Example: Forecasting (4D-Var)



ECMWF
de Rosnay et al. (2022)

Example: Forecasting (4D-Var)

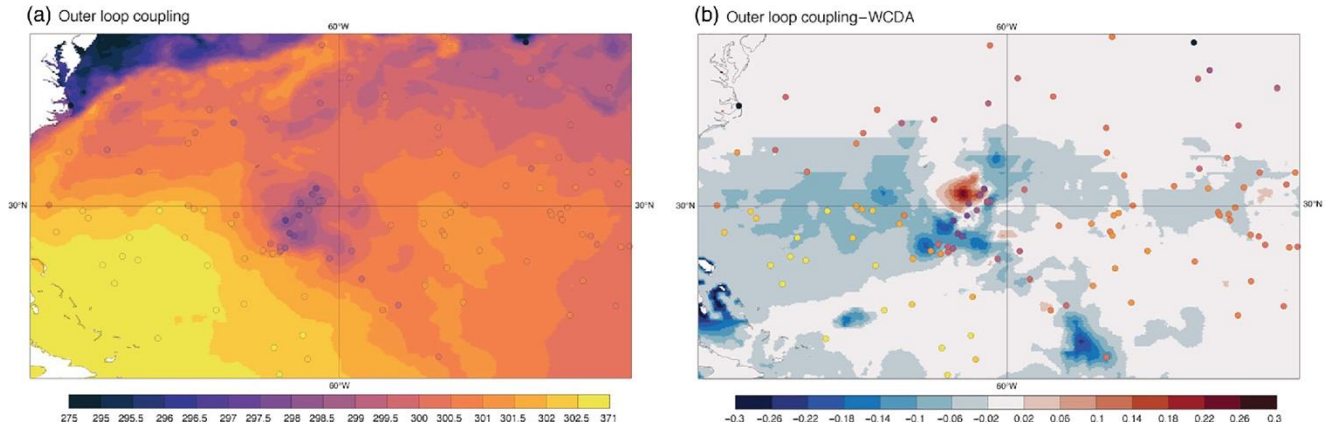


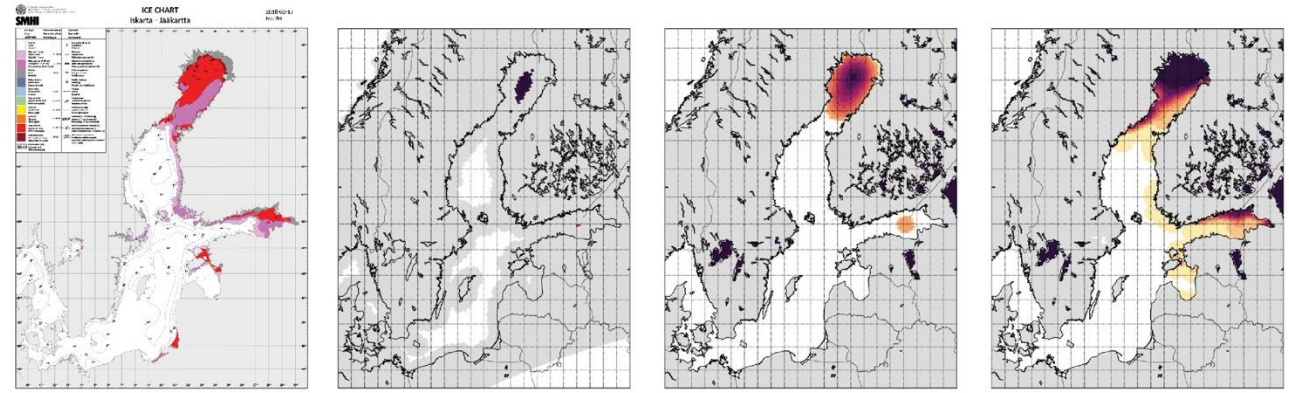


ECMWF

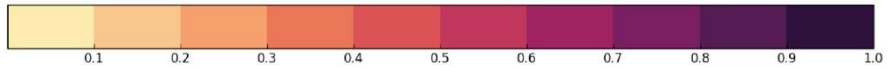
de Rosnay et al. (2022)

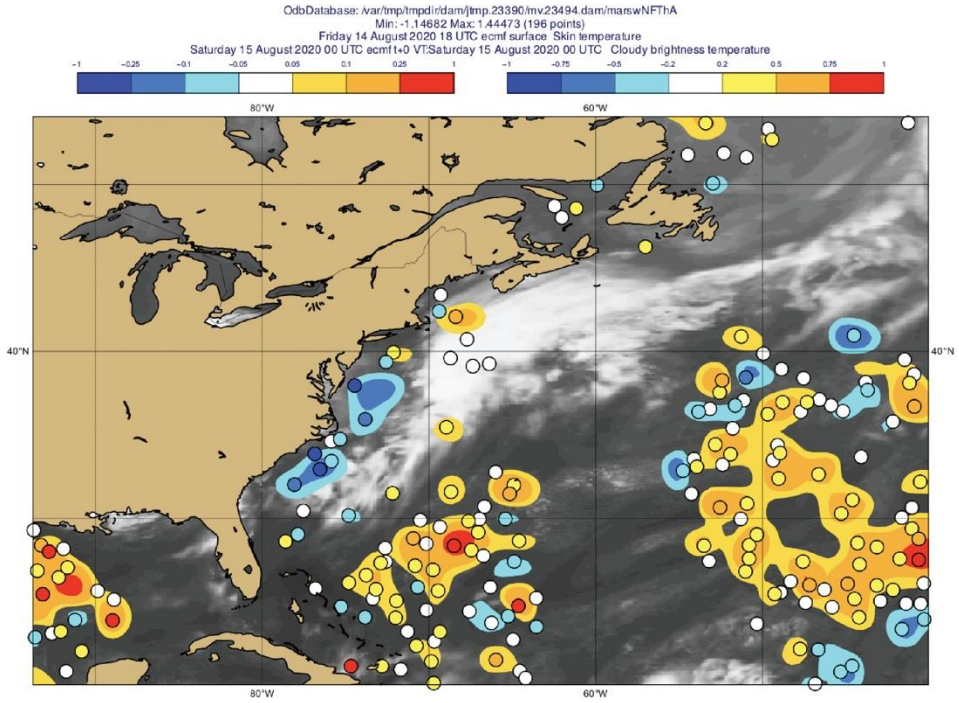
Example: Forecasting (4D-Var)





(a) Manual ice chart (b) L3 from OSI SAF (c) Uncoupled analysis (d) WCDA



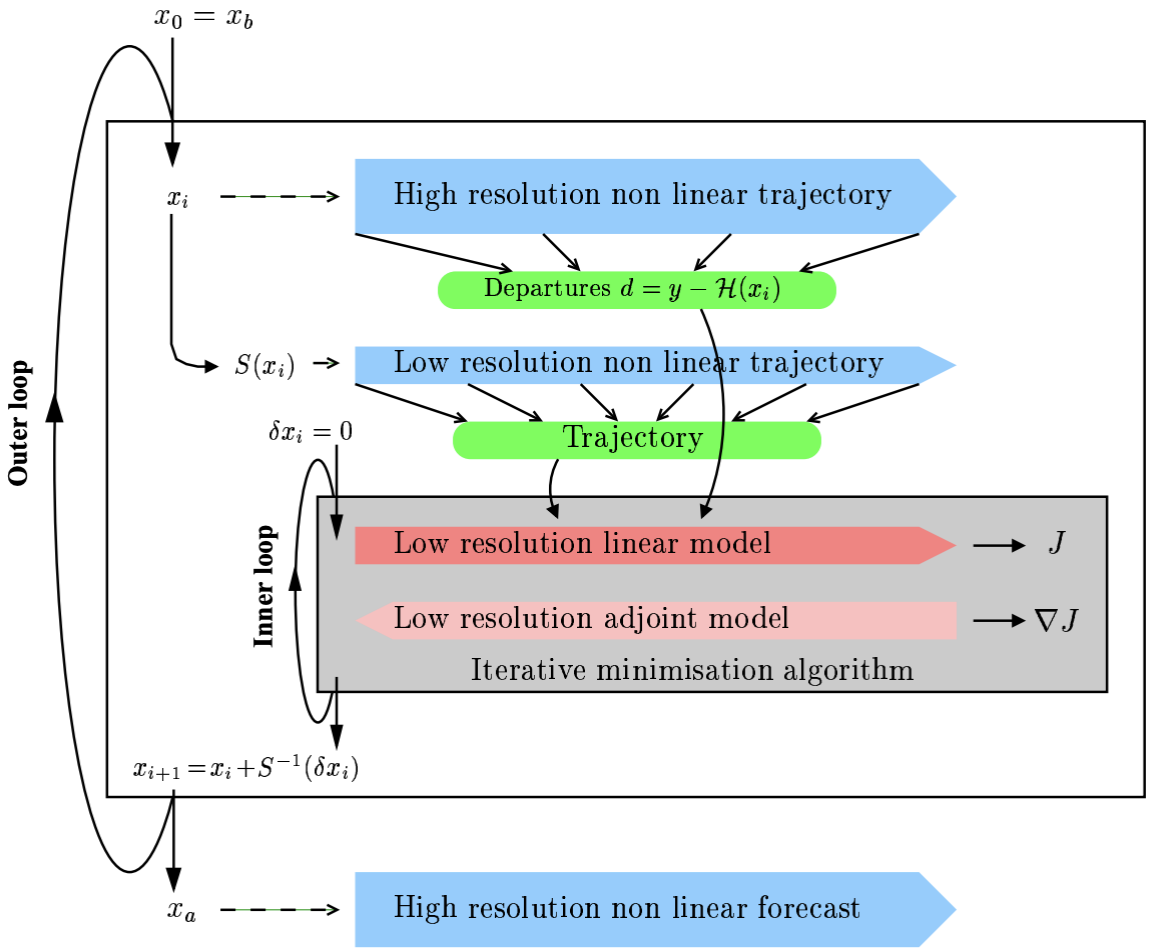


ECMWF
de Rosnay et al. (2022)

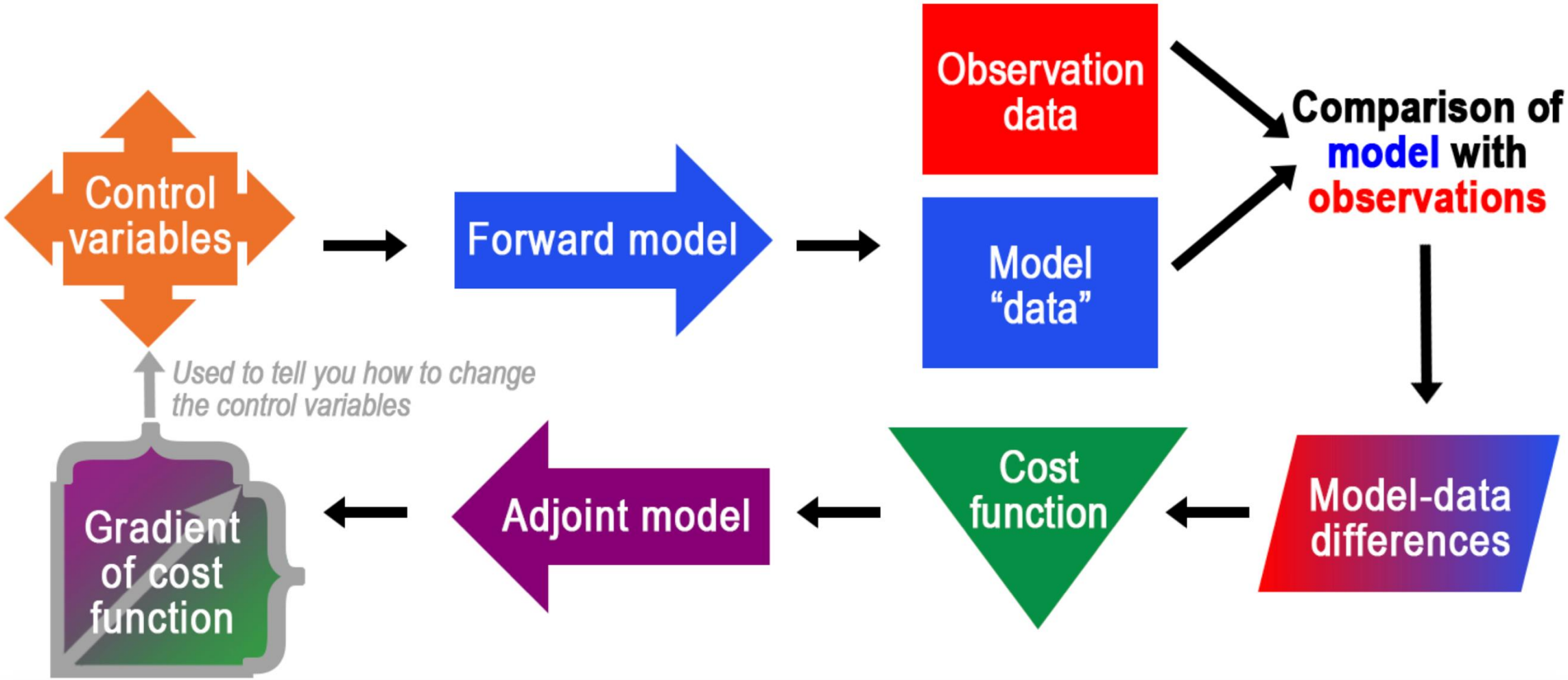
Example: Forecasting (4D-Var)

$$
\begin{aligned}
J(\mathbf{x}) &= \tfrac{1}{2}(\mathbf{x}-\mathbf{x}_\mathrm{b})^\mathrm{T}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_\mathrm{b}) \\
&\quad + \tfrac{1}{2}\sum_k (\mathscr{H}_k\mathscr{M}_k(\mathbf{x})-\mathbf{y}_k)^\mathrm{T}\mathbf{R}_k^{-1}(\mathscr{H}_k\mathscr{M}_k(\mathbf{x})-\mathbf{y}_k)
\end{aligned}
$$

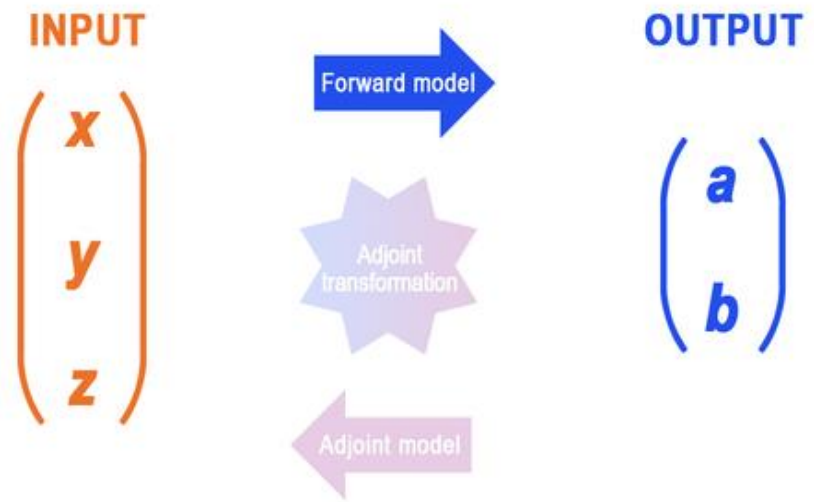
$$
\begin{aligned}
\nabla J(\mathbf{x}) &= \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_\mathrm{b}) \\
&\quad + \sum_k \mathbf{M}_k^\mathrm{T}\mathbf{H}_k^\mathrm{T}\mathbf{R}_k^{-1}(\mathscr{H}_k\mathscr{M}_k(\mathbf{x})-\mathbf{y}_k).
\end{aligned}
$$



Example: State estimate (3D-Var)



Example: State estimate (3D-Var)



$$a = x - 2y + 3z$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & -2 & 3 \\ 4 & 0 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

Here's the other equation... $b = 4x - 5z$

Forward model

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & -2 & 3 \\ 4 & 0 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

Adjoint model

$$\begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ -2 & 0 \\ 3 & -5 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$$

$$\begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ -2 & 0 \\ 3 & -5 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$$

$$\hat{x} = \hat{a} + 4\hat{b}$$

$$\hat{y} = -2\hat{a}$$

$$\hat{z} = 3\hat{a} - 5\hat{b}$$

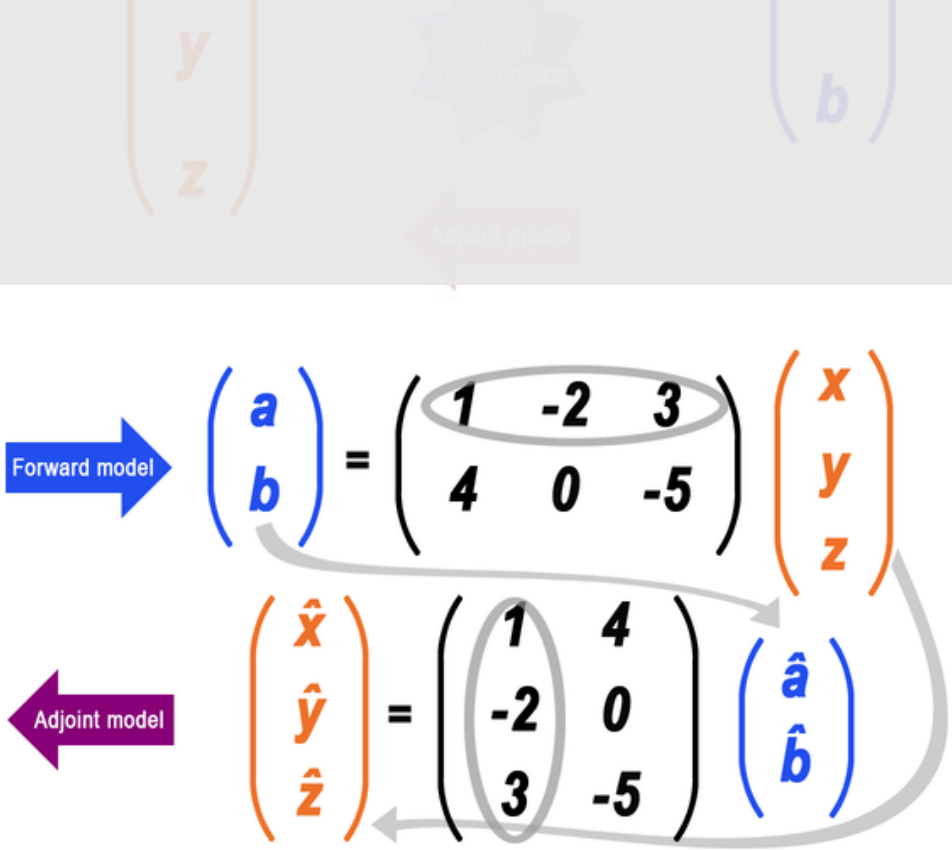
Example: State estimate (3D-Var)

$$a = x - 2y + 3z$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & -2 & 3 \\ 4 & 0 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

Here's the other equation... $b = 4x - 5z$

Example: State estimate (3D-Var)



Example: State estimate (3D-Var)

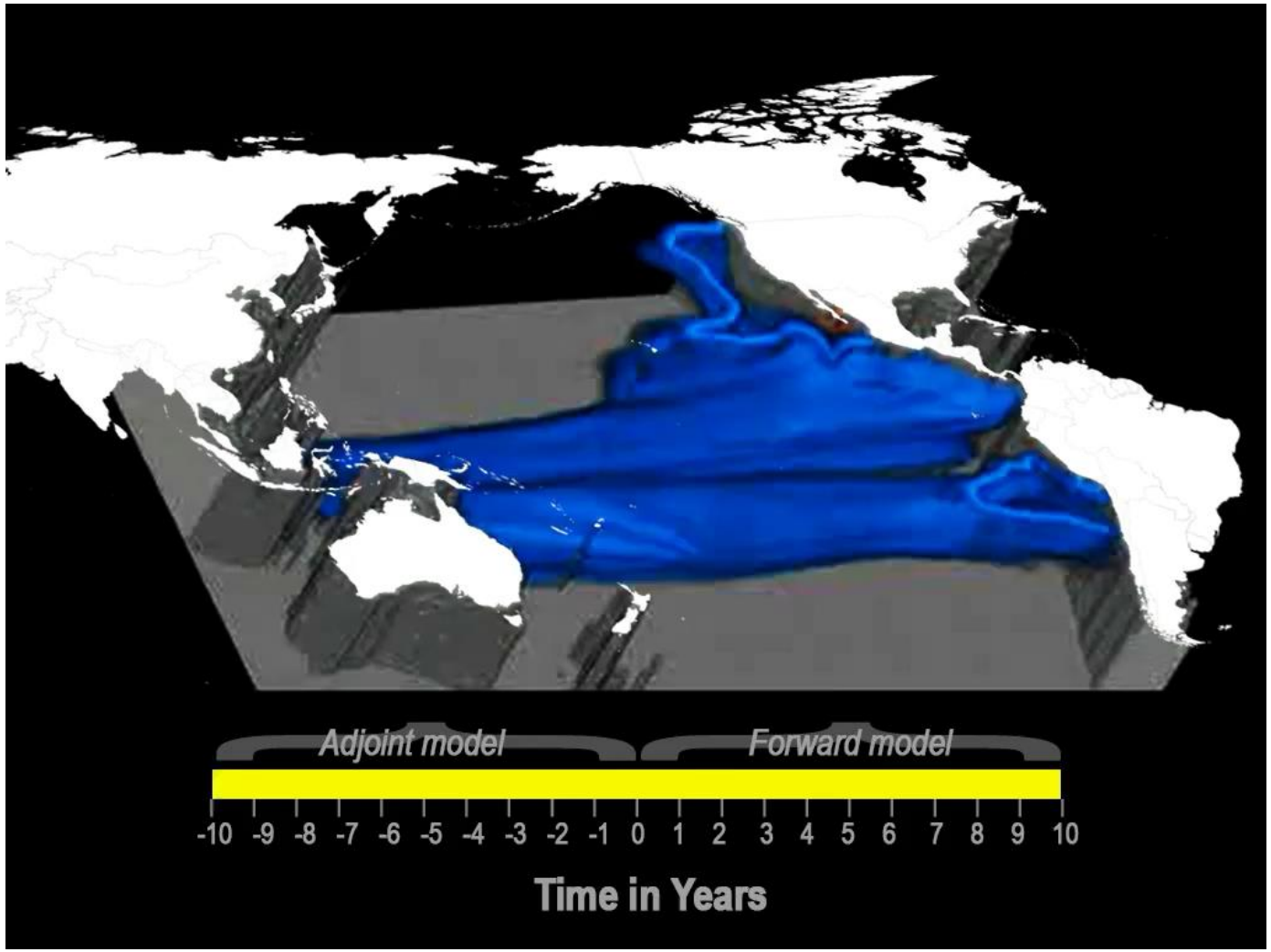
$$\begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ -2 & 0 \\ 3 & -5 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$$

$$\hat{x} = \hat{a} + 4\hat{b}$$

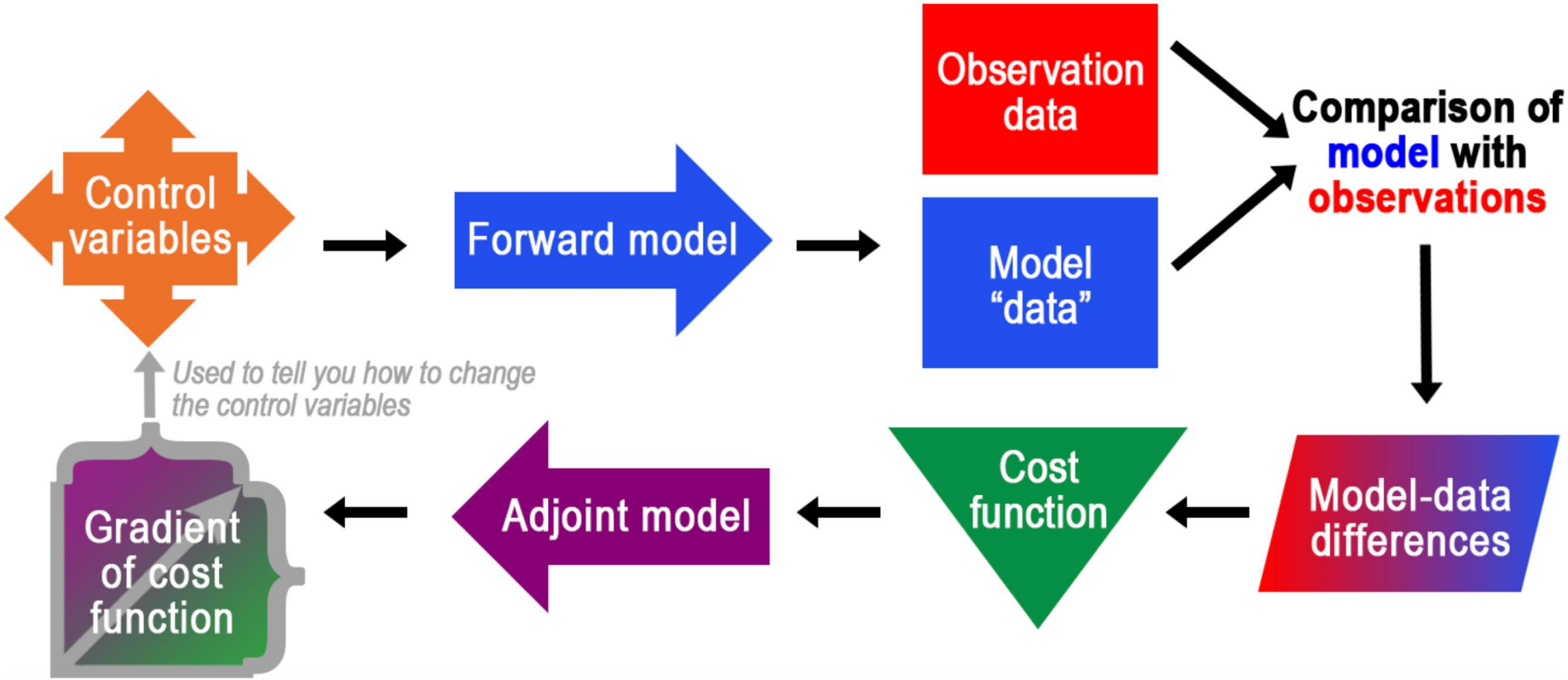
$$\hat{y} = -2\hat{a}$$

$$\hat{z} = 3\hat{a} - 5\hat{b}$$

Example: State estimate (3D-Var)

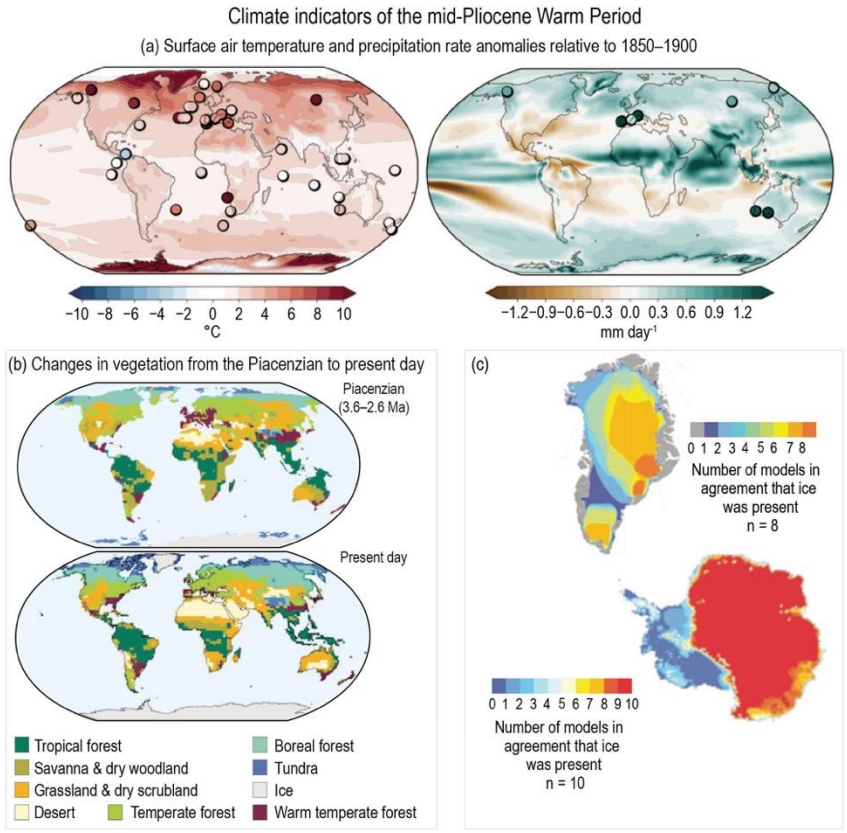


Example: State estimate (3D-Var)



Example: multiple data sources

Paleoclimate Models



Cross-Chapter Box 2.4, Figure 1 in IPCC, 2021: Chapter 2. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Gulev, S.K.,

Proxies and Archives

"Travertine speleothem (Crystal Cave, Main Island, Bermuda) 1" by James St. John is licensed under CC BY 2.0

"Tree rings" by Out of the Fire Blog is licensed under CC BY 2.0

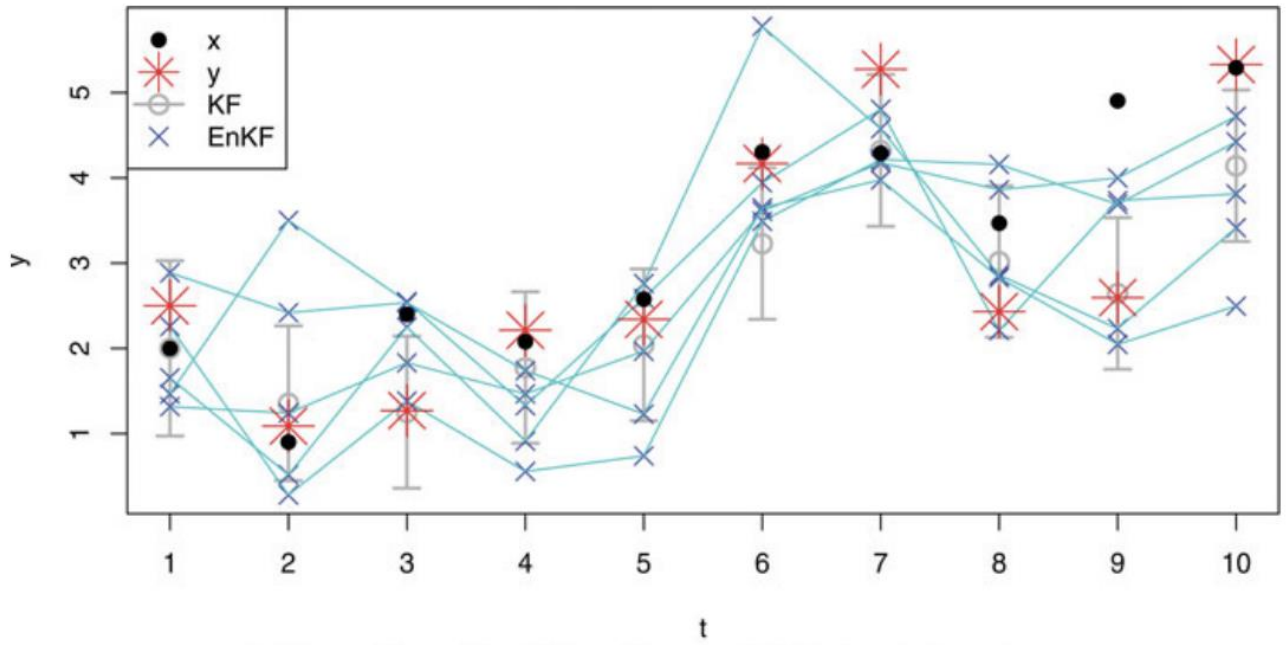
"A volcanic ash layer in the WAIS Divide ice core. Volcanic markers like these were used in the new study to synchronize ice cores from across Antarctica." by Oregon State University is licensed under CC BY-SA 2.0

"2016_Lake_sediment_core_Forlorn_Lakes_Gifford_Pinchot_National_Forest_Washington." by USDA Forest Service is marked with Public Domain Mark 1.0

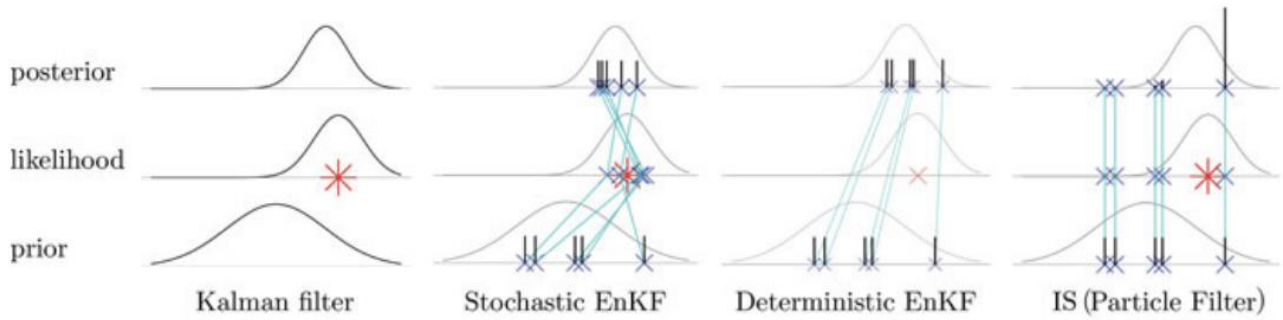
"Diploria fossil brain coral on Devil's Point Hardground (Cockburn Town Member, Grotto Beach Formation, Upper Pleistocene, ~120-123 ka; Cockburn Town Fossil Reef, San Salvador Island, Bahamas) 3" by James St. John is licensed under CC BY 2.0

Example: multiple data sources

Ensemble Kalman Filter



(a) State, observation, Kalman filter, and EnKF for 10 time points

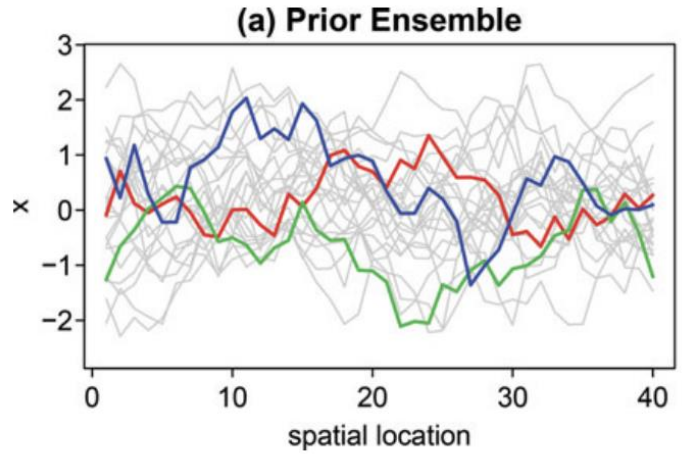


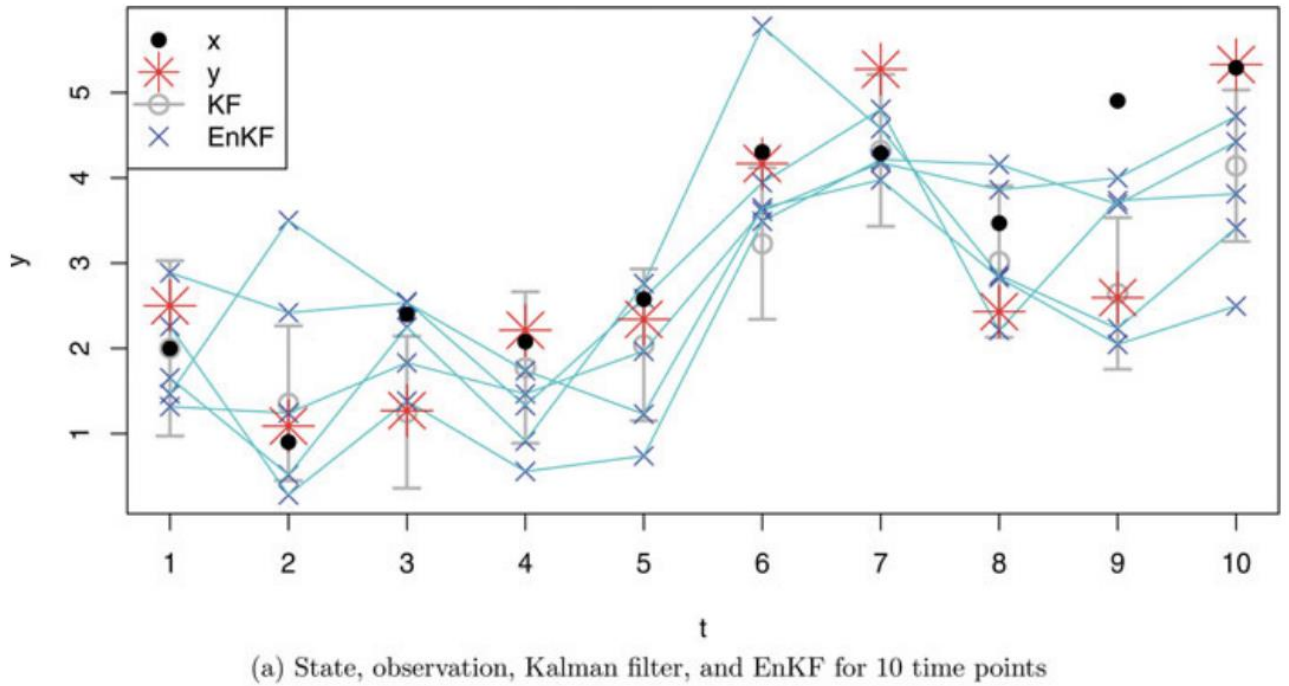
(b) Illustration of different updating schemes at time point $t = 1$

Katzfuss et al (2016)

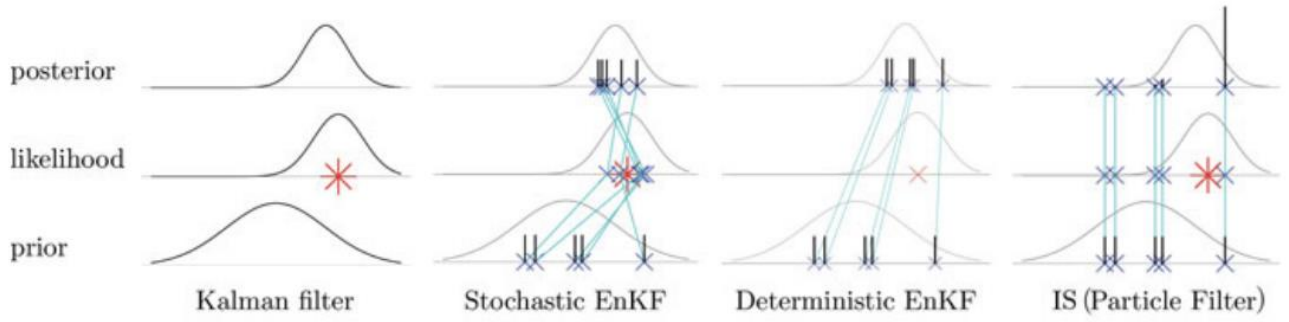
Example: multiple data sources

Ensemble Kalman Filter





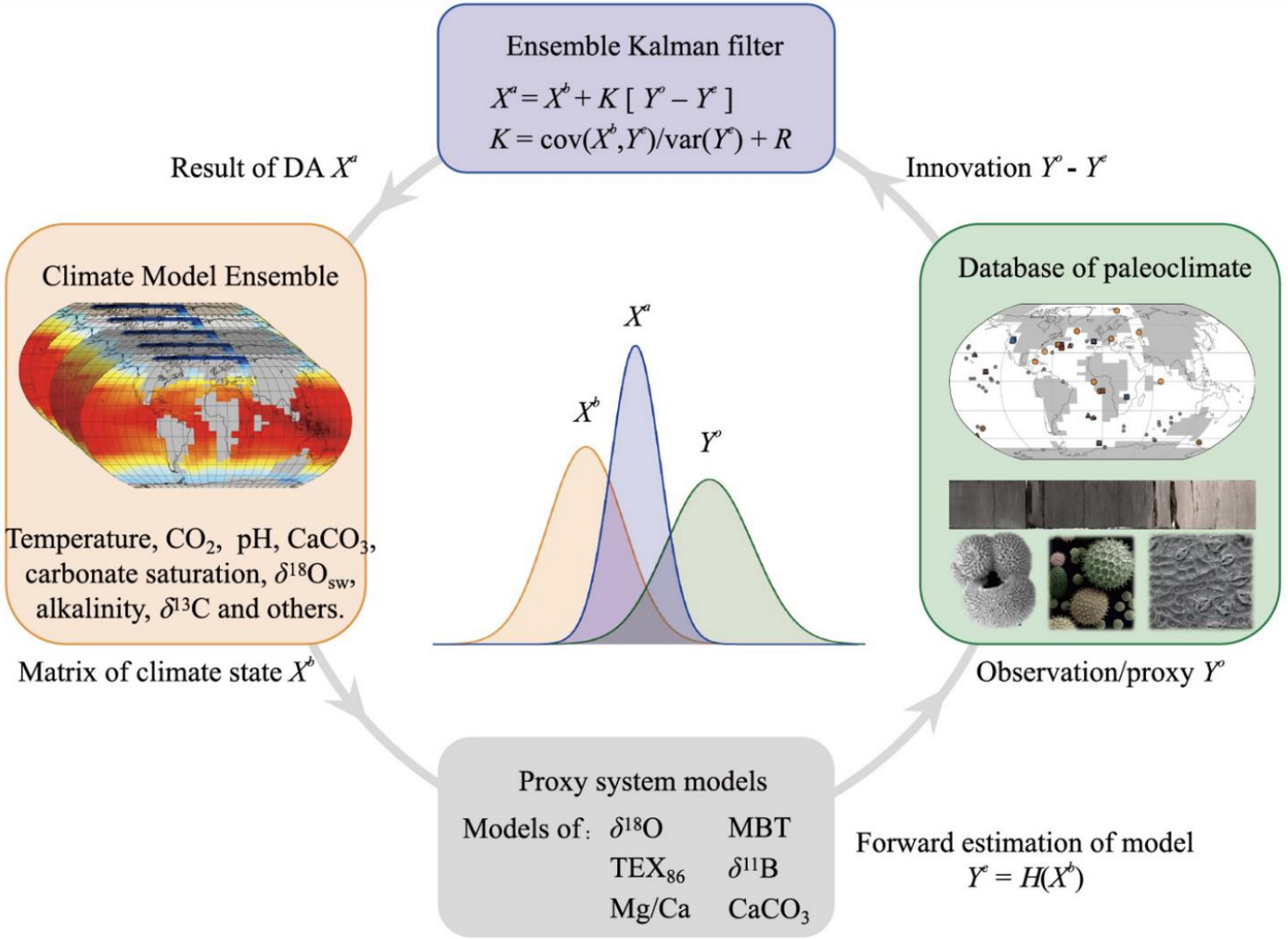
(a) State, observation, Kalman filter, and EnKF for 10 time points



(b) Illustration of different updating schemes at time point $t = 1$

Katzfuss et al (2016)

Example: multiple data sources



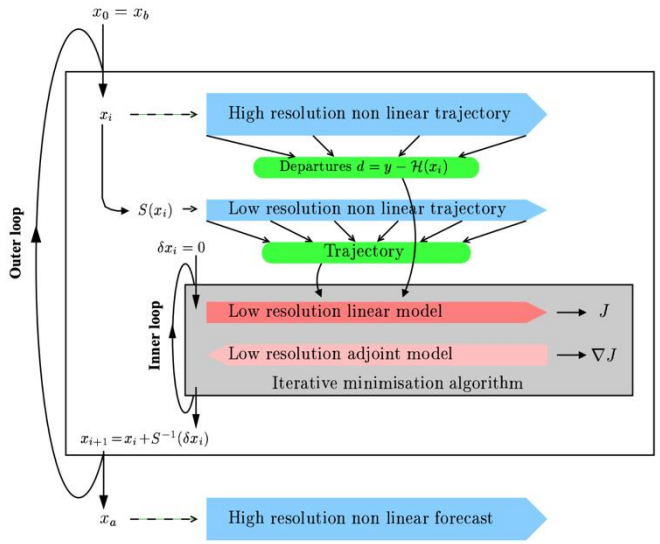
Zhang et al (2025)

Data assimilation with AI

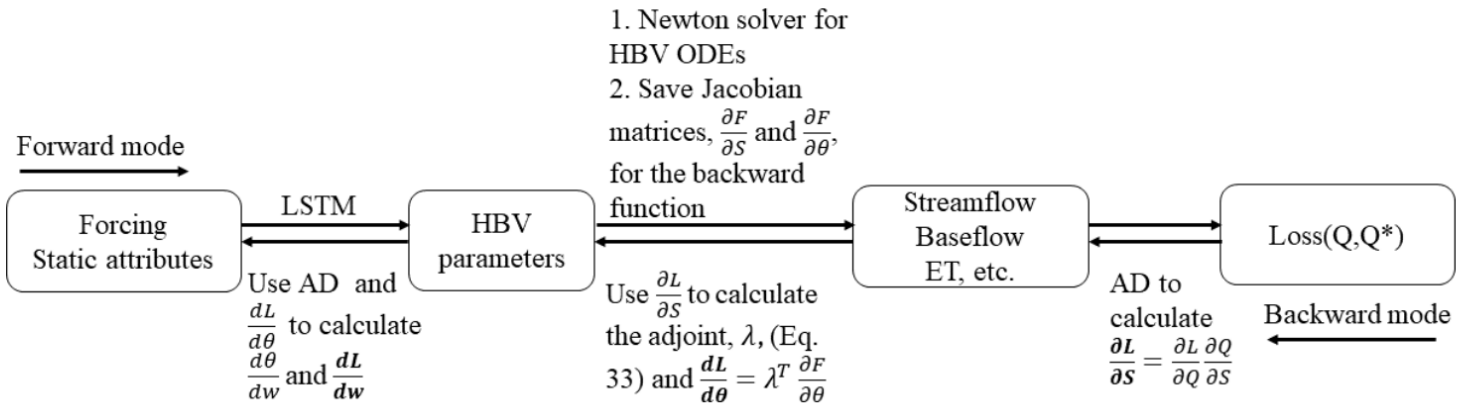
Data assimilation with AI

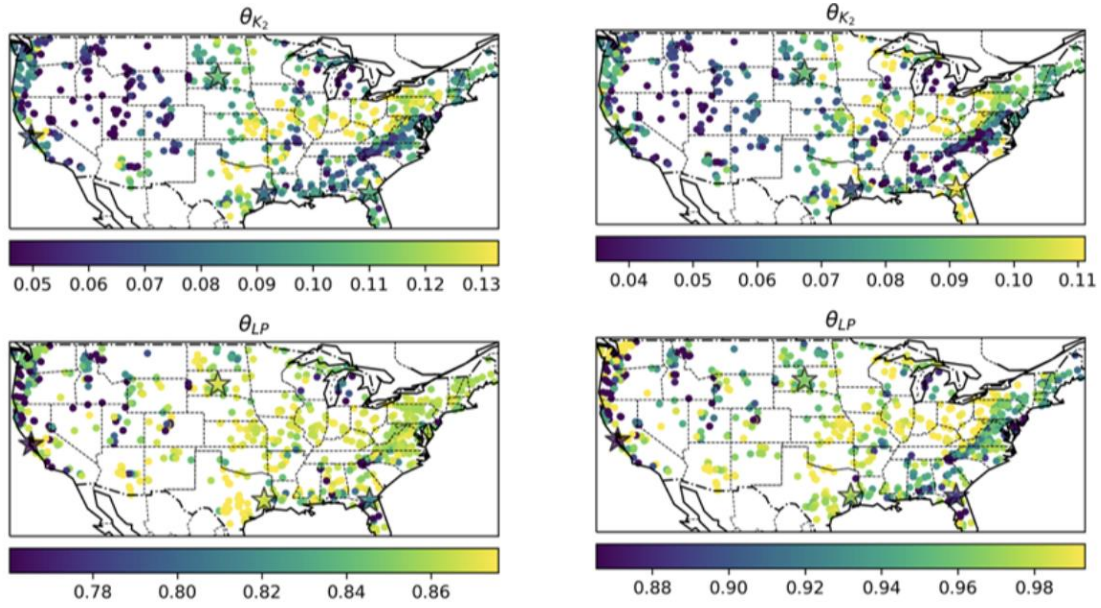
Data assimilation is naturally suited for AI integration

- It is already an optimization problem!
- Includes backprop in adjoint methods
- Data-driven task i.e., mostly agnostic to underlying physics
- Potential for big impact in climate: inference is always challenge due to limited sampling



Example: deep learning with adjoint methods





(a) Sequential model (b) Adjoint model

Song et al (2024)

Example: Combined with EnKF

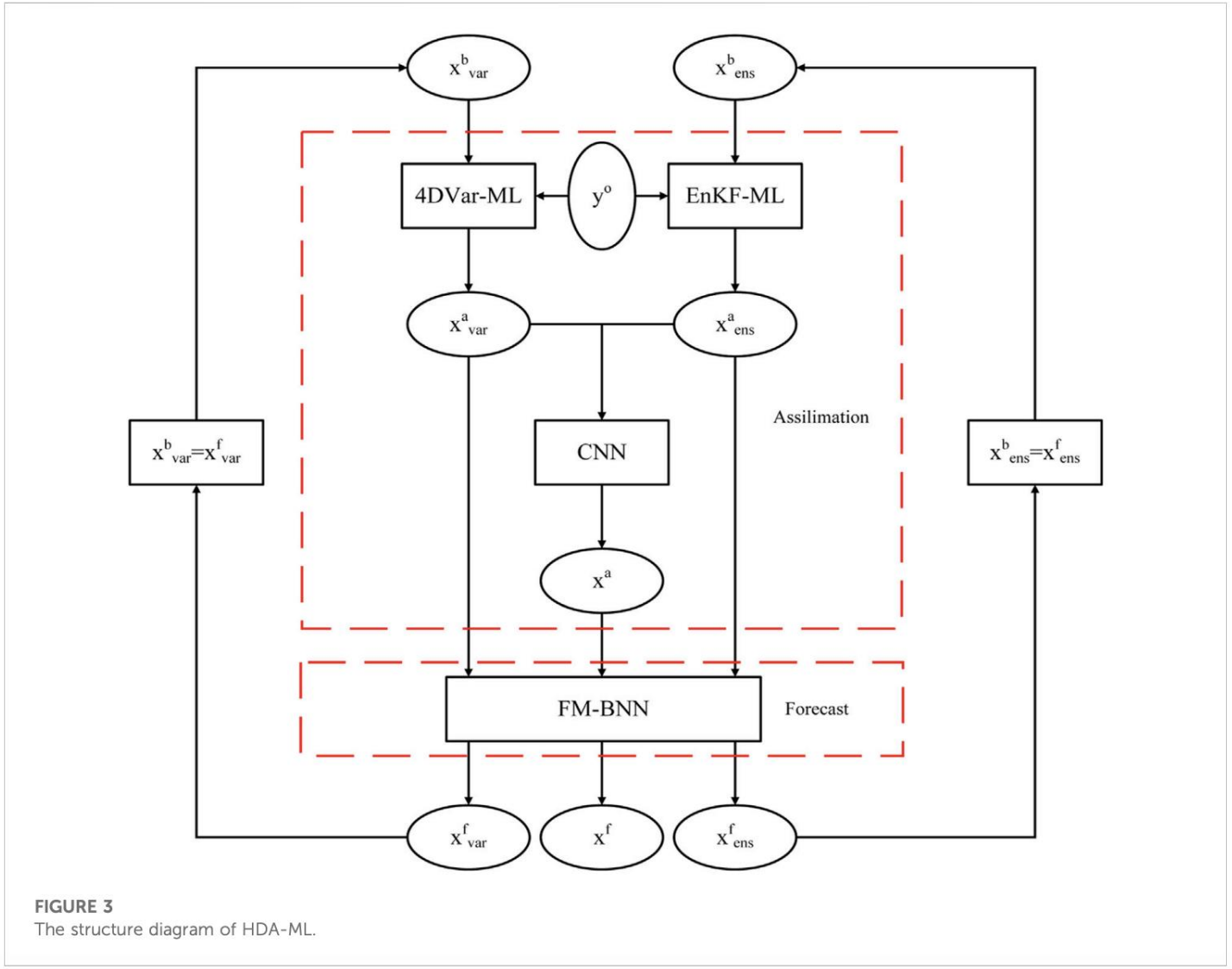


FIGURE 3
The structure diagram of HDA-ML.

Dong et al (2022)

Example: reduced order modeling for 3D/4D-Var

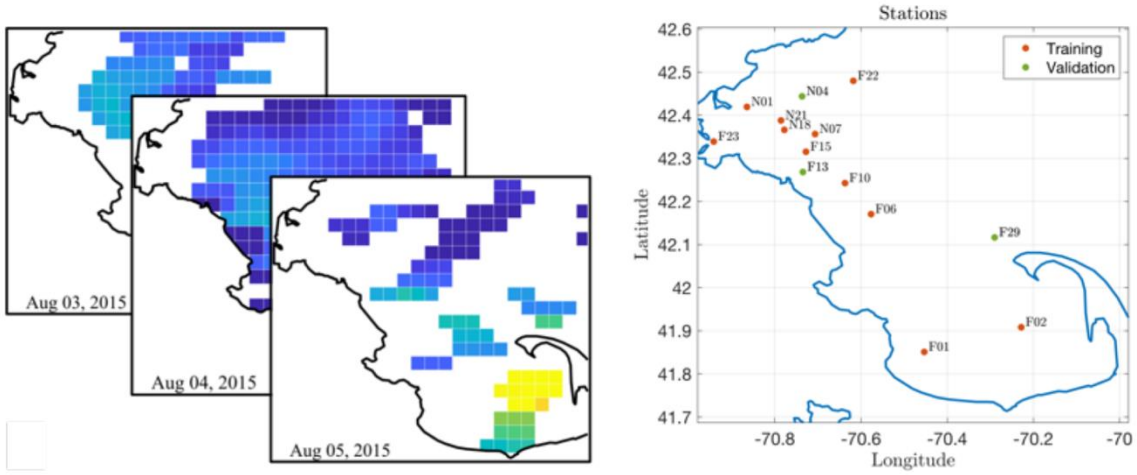


Figure 2: **Sensor Data.** The low fidelity data (satellite (left)) is only available on days with low cloud coverage. The high fidelity data (buoys (right)) is local in space and sparse.

Champenois et al (2024)

Example: reduced order modeling for 3D/4D-Var

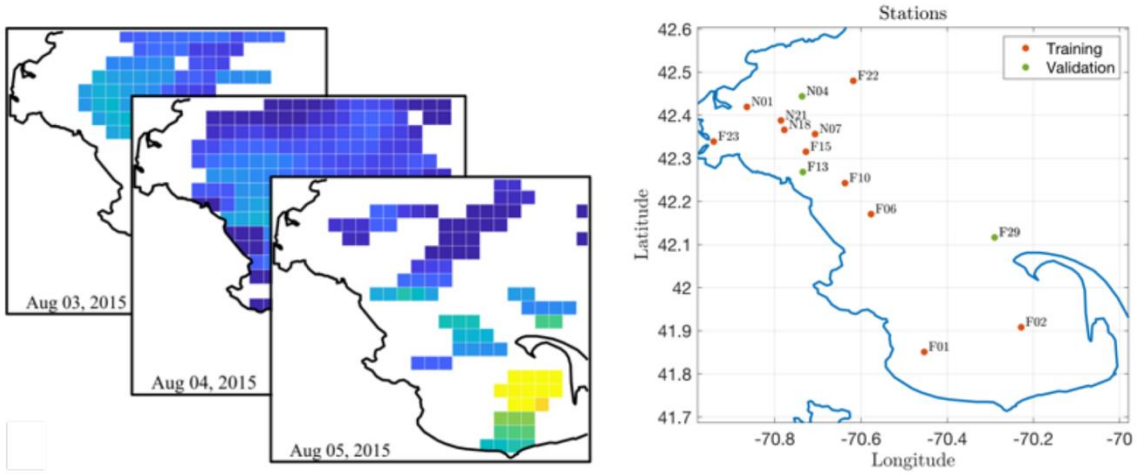
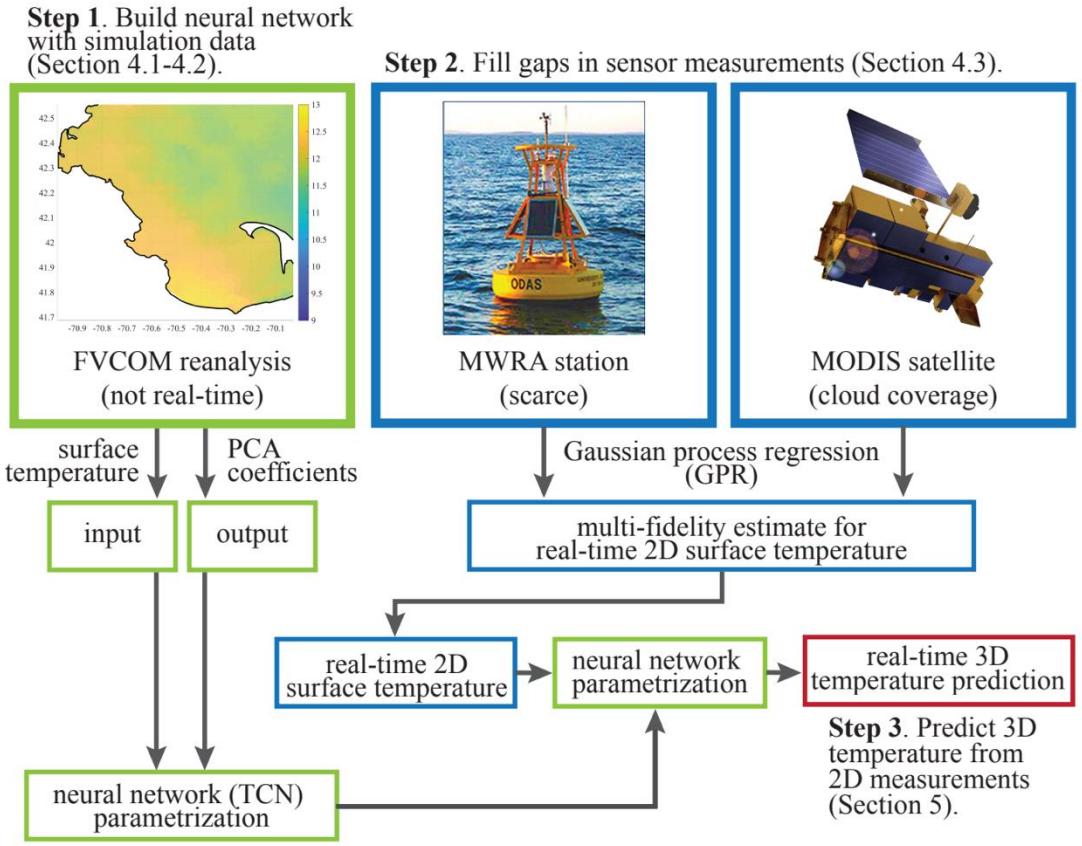
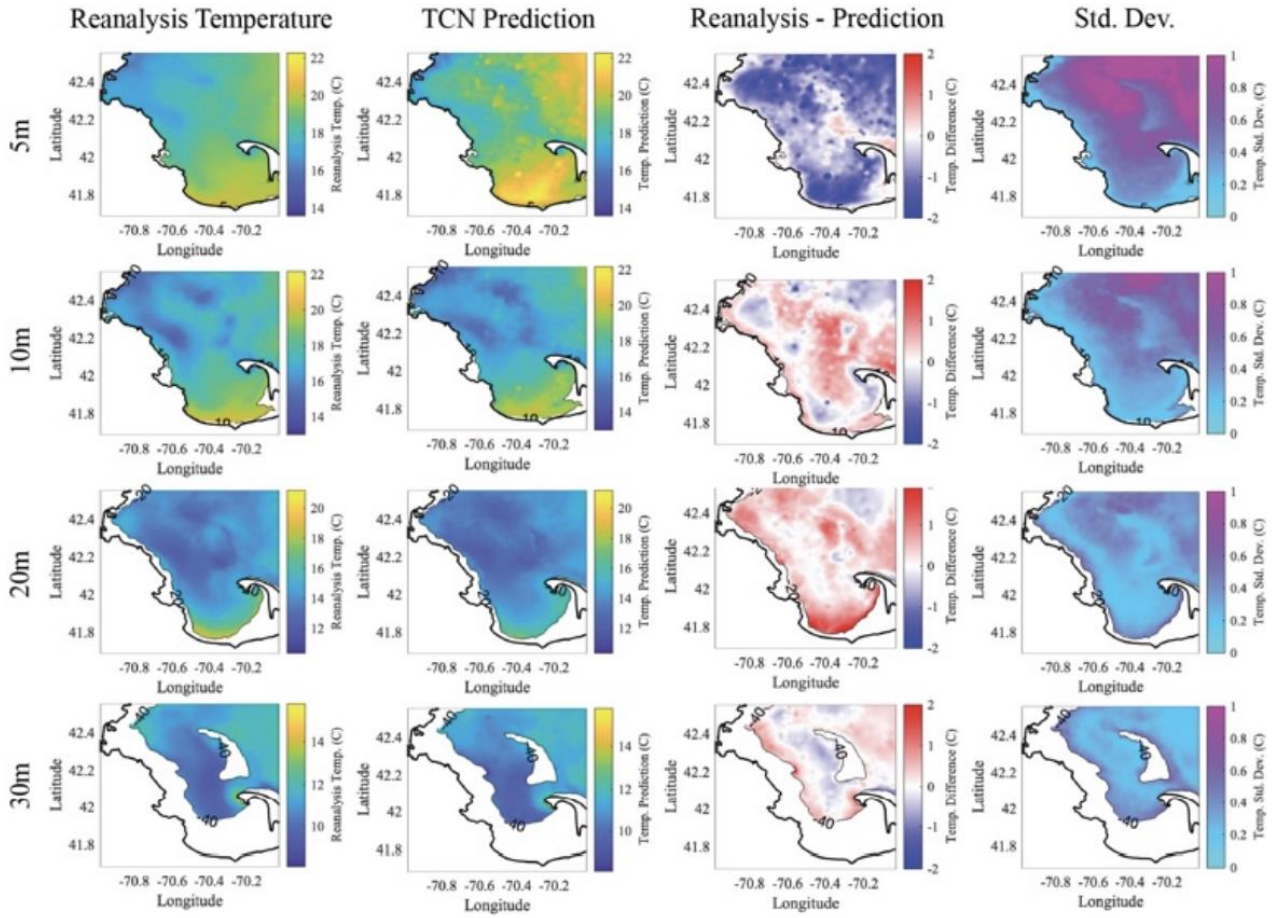


Figure 2: **Sensor Data.** The low fidelity data (satellite (left)) is only available on days with low cloud coverage. The high fidelity data (buoys (right)) is local in space and sparse.



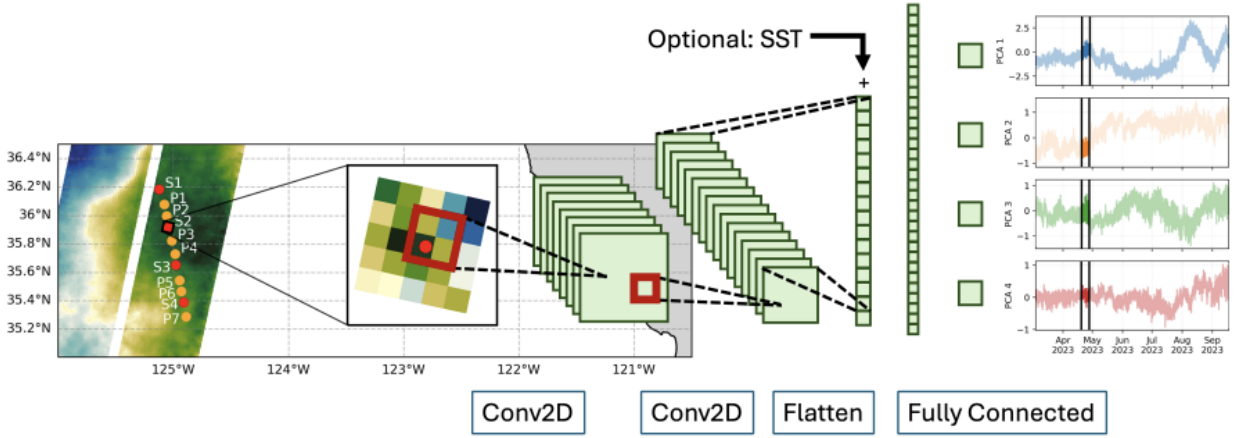
Champenois et al (2024)

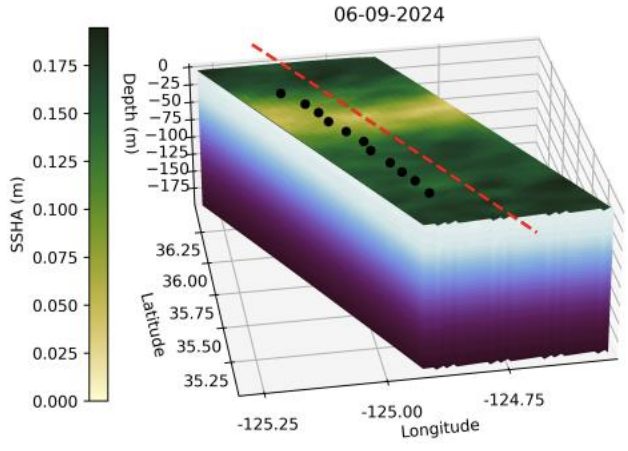
Example: reduced order modeling for 3D/4D-Var

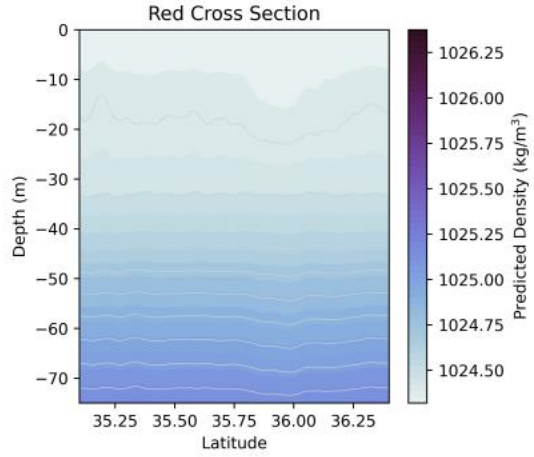


Champenois et al (2024)

Example: reduced order modeling for 3D/4D-Var



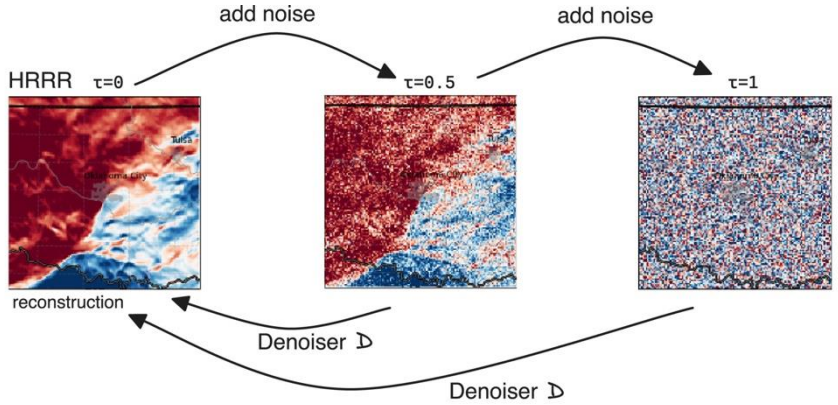




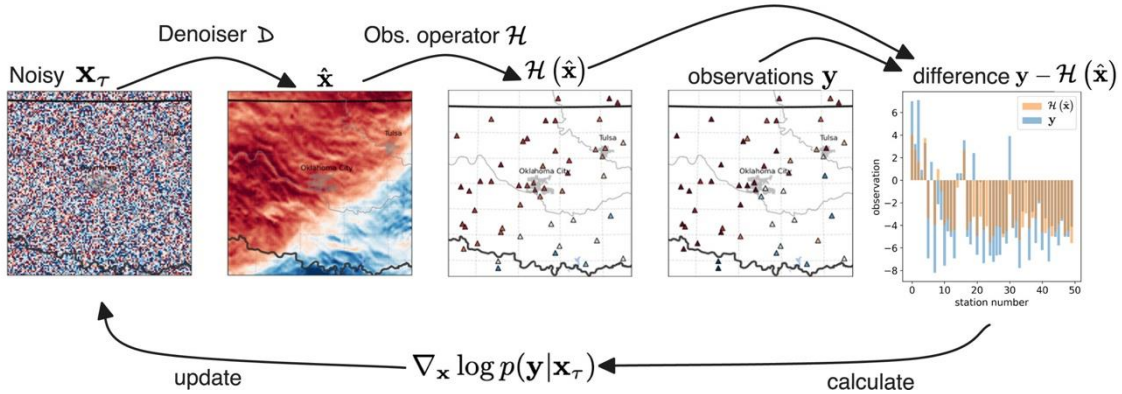
Champenois et al (under review)

Example: generative modeling

Denoiser training

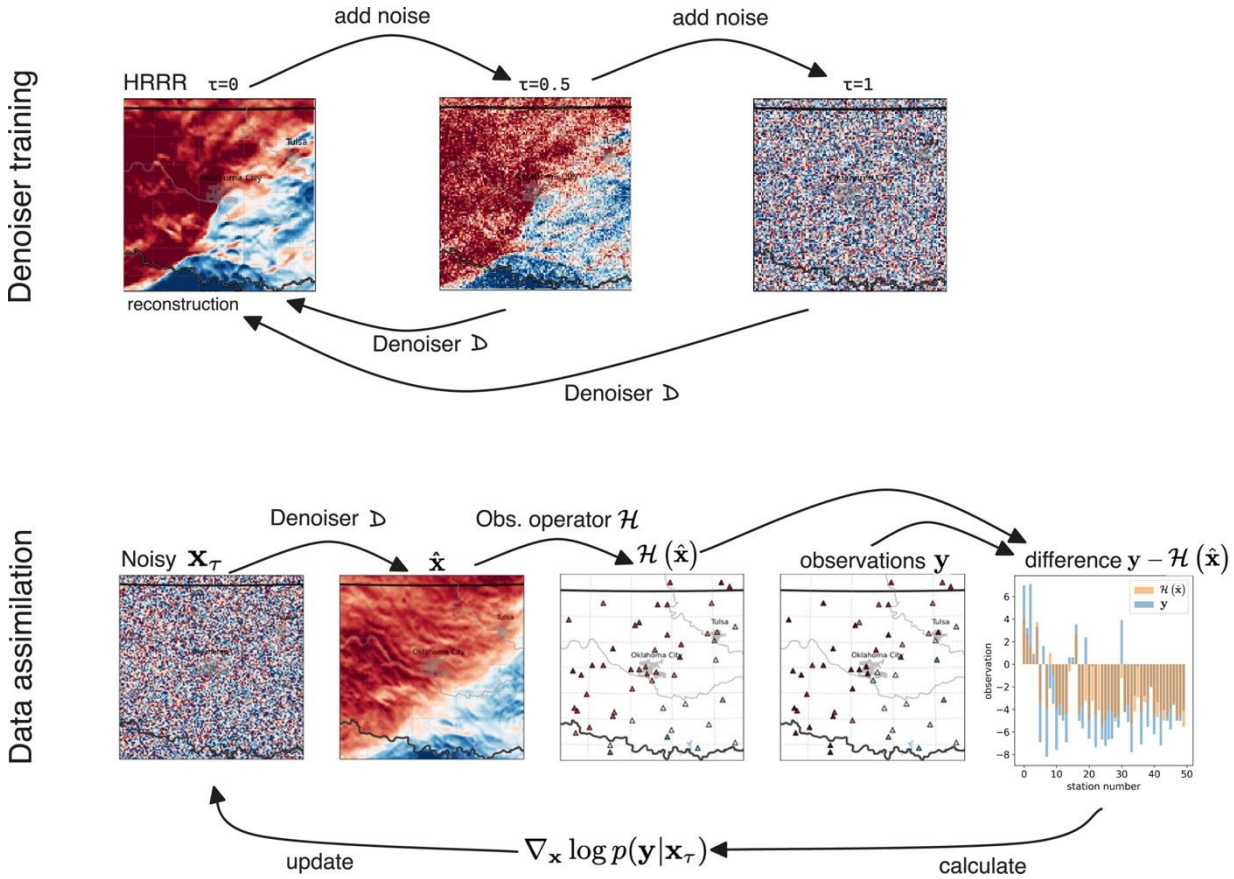


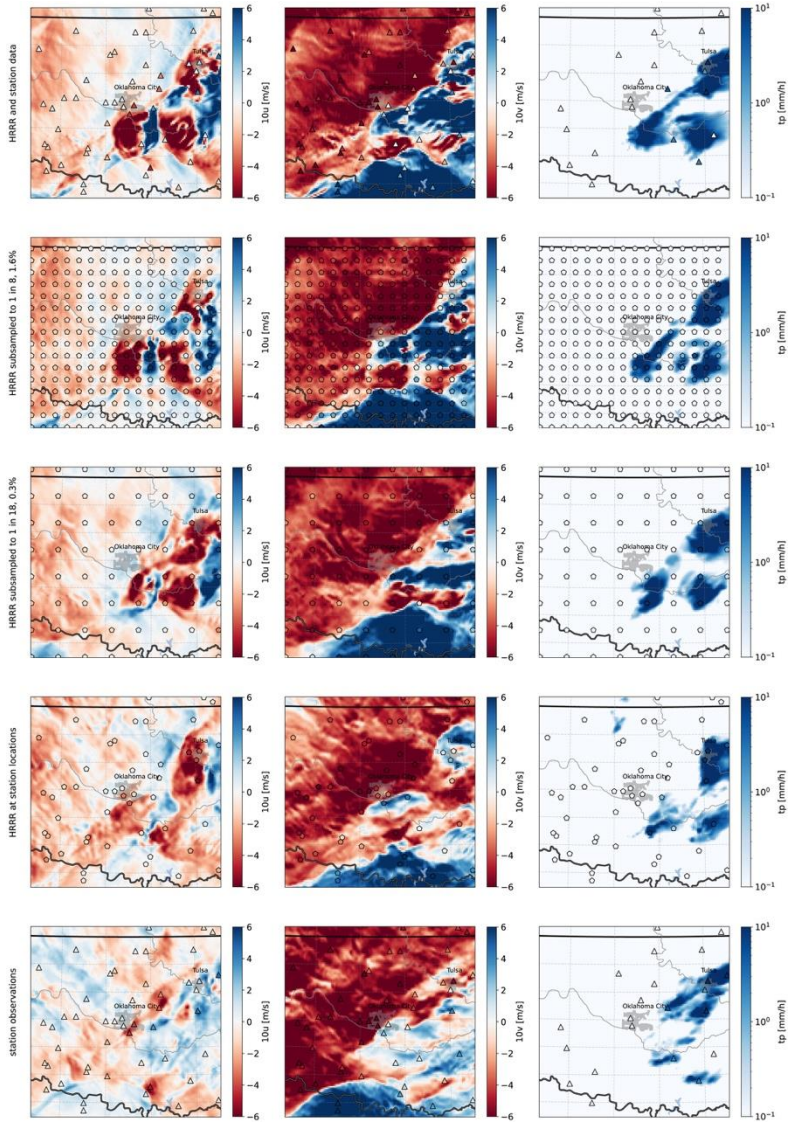
Data assimilation



Manshausen et al (2025)

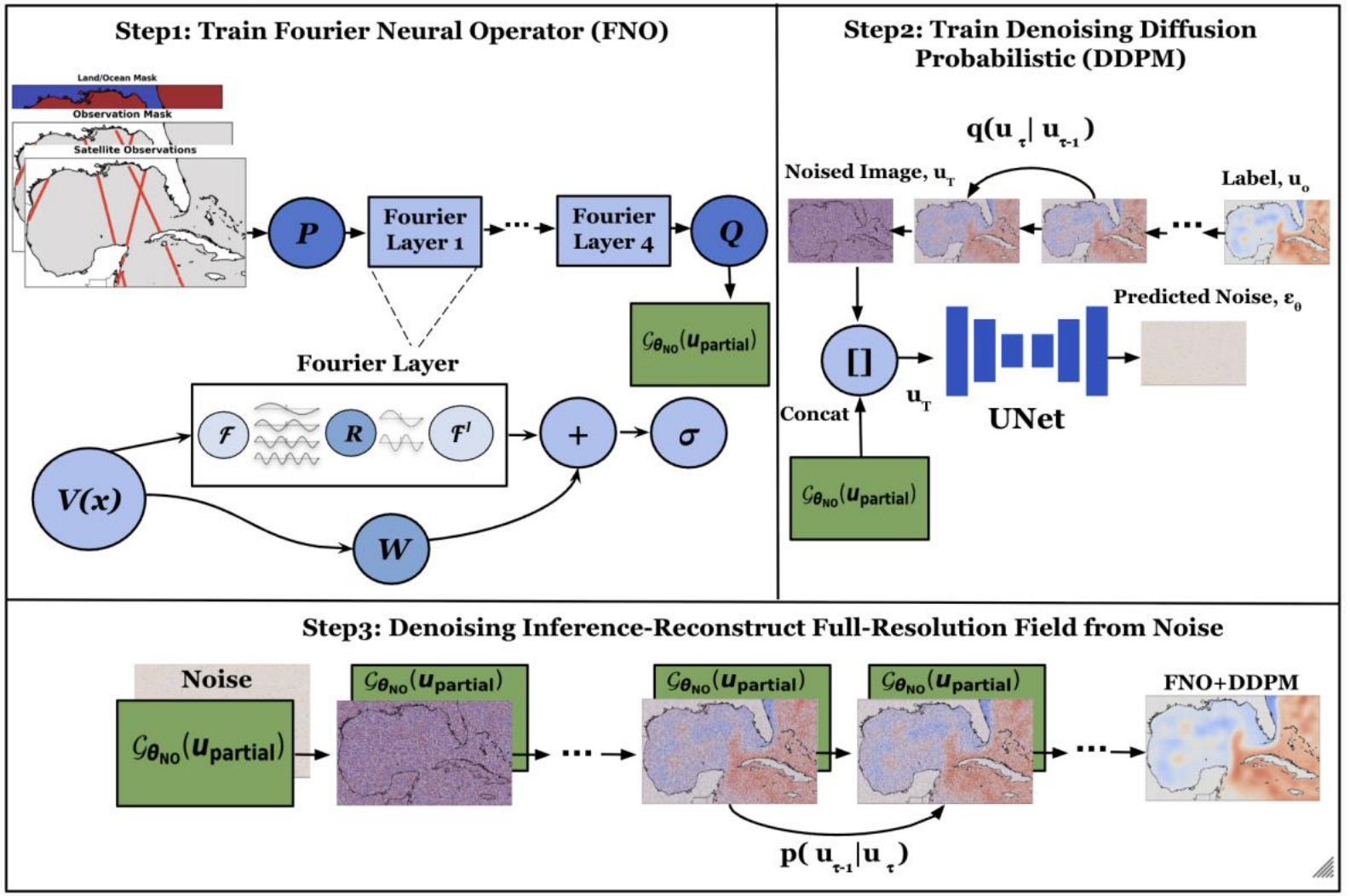
Example: generative modeling





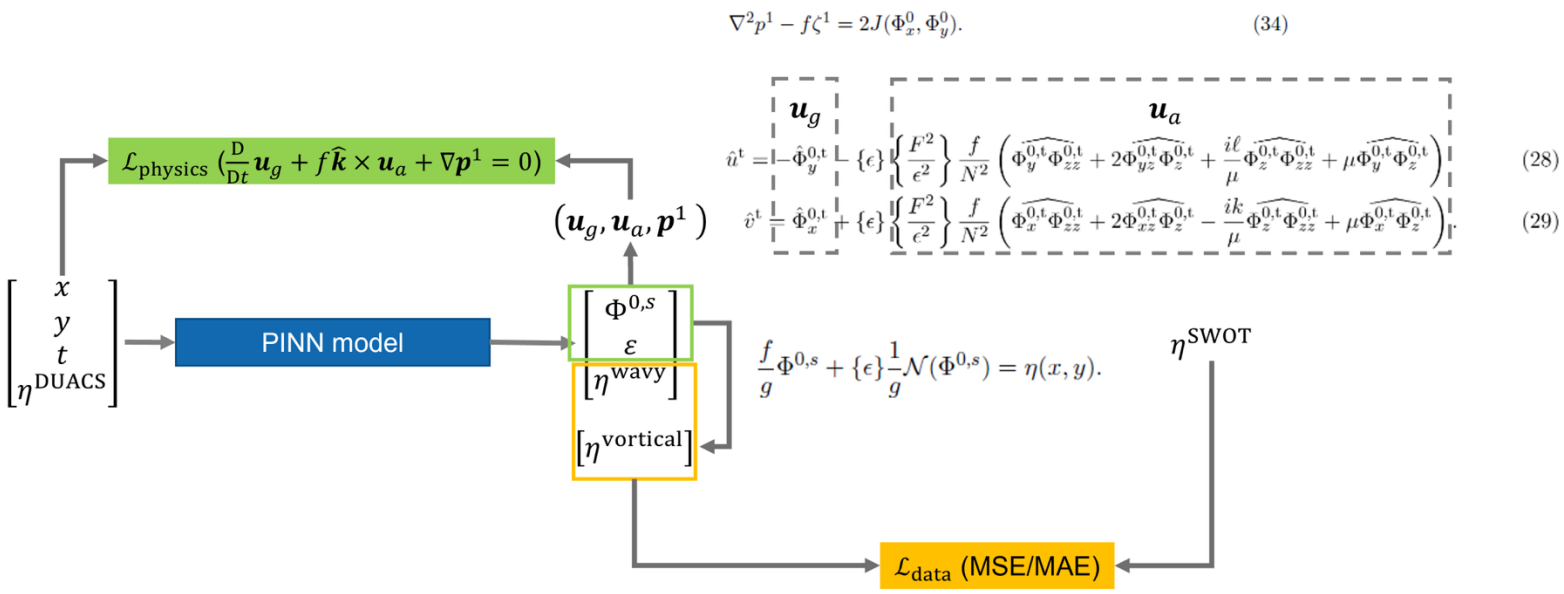
Manshausen et al (2025)

Example: generative modeling



Asefi et al (2025)

Example: physics informed neural network (PINN)



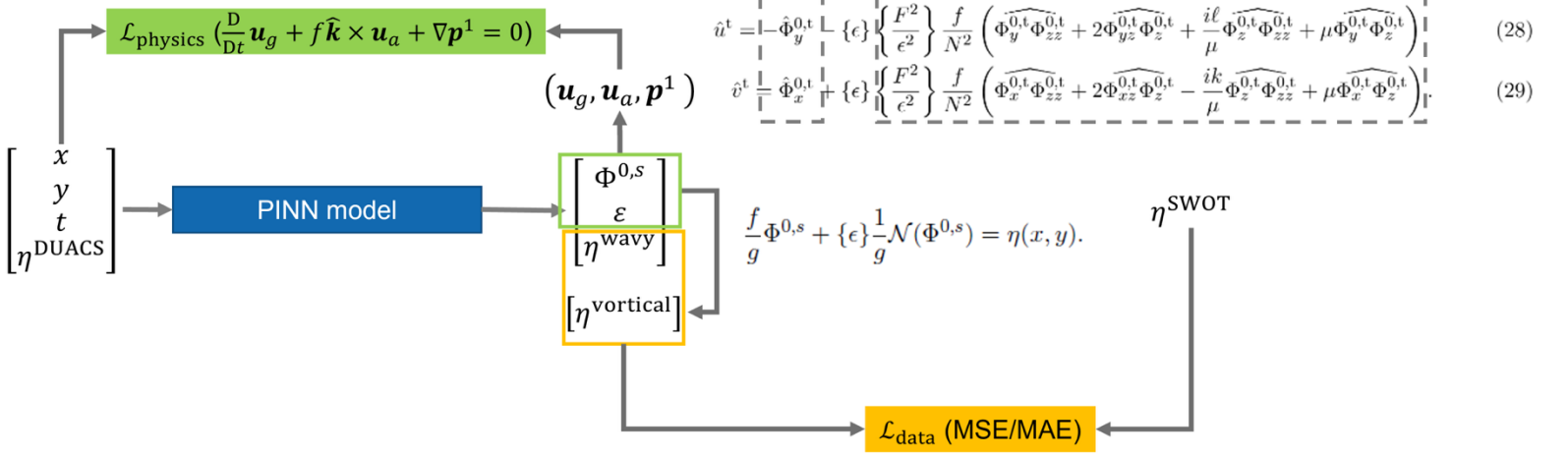
Guo et al (in prep)

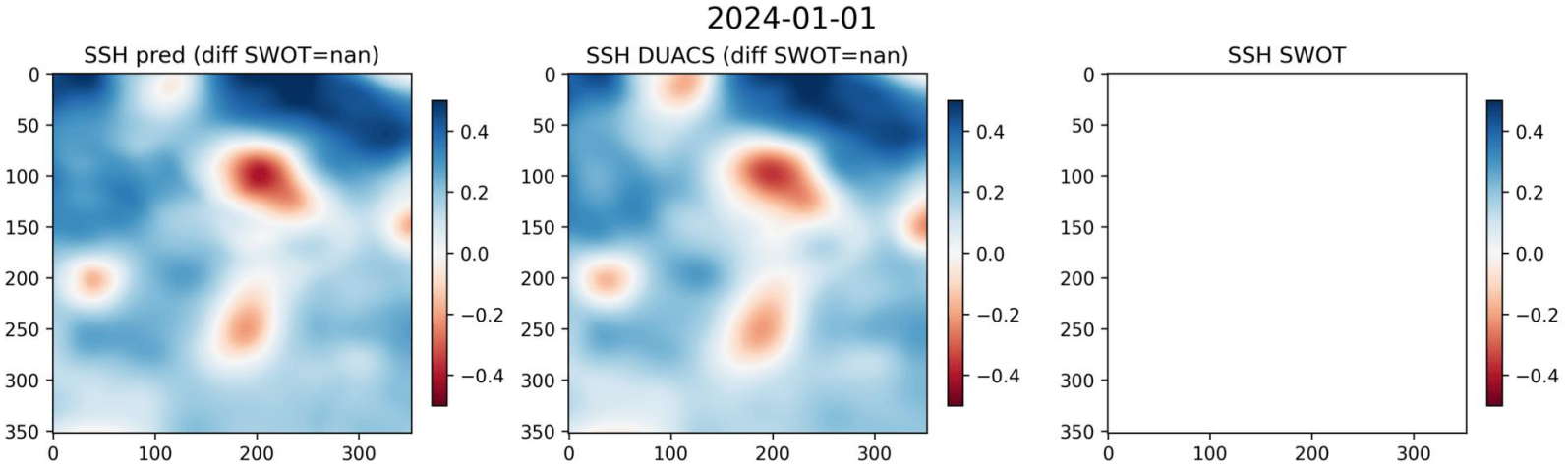
Example: physics informed neural network (PINN)

$$\nabla^2 p^1 - f\zeta^1 = 2J(\Phi_x^0, \Phi_y^0). \tag{34}$$

$$\hat{u}^{\mathrm{t}} = -\hat{\Phi}_y^{0,\mathrm{t}} + \{\epsilon\}\left\{\frac{F^2}{\epsilon^2}\right\}\frac{f}{N^2}\left(\widehat{\Phi_y^{0,\mathrm{t}}\Phi_{zz}^{0,\mathrm{t}}} + 2\widehat{\Phi_{yz}^{0,\mathrm{t}}\Phi_z^{0,\mathrm{t}}} + \frac{i\ell}{\mu}\widehat{\Phi_z^{0,\mathrm{t}}\Phi_{zz}^{0,\mathrm{t}}} + \mu\widehat{\Phi_y^{0,\mathrm{t}}\Phi_z^{0,\mathrm{t}}}\right) \tag{28}$$

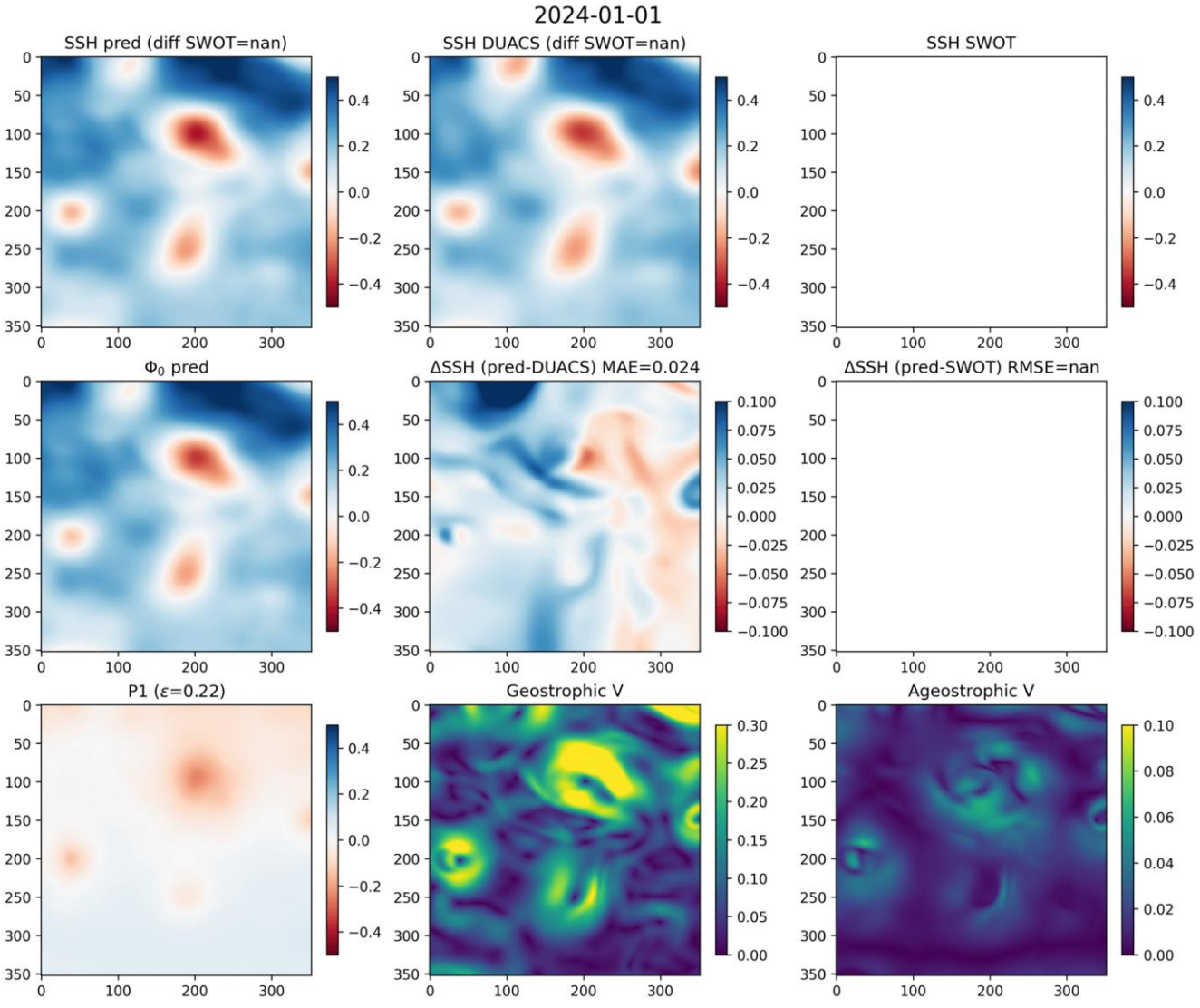
$$\hat{v}^{\mathrm{t}} = \hat{\Phi}_x^{0,\mathrm{t}} + \{\epsilon\}\left\{\frac{F^2}{\epsilon^2}\right\}\frac{f}{N^2}\left(\widehat{\Phi_x^{0,\mathrm{t}}\Phi_{zz}^{0,\mathrm{t}}} + 2\widehat{\Phi_{xz}^{0,\mathrm{t}}\Phi_z^{0,\mathrm{t}}} - \frac{ik}{\mu}\widehat{\Phi_z^{0,\mathrm{t}}\Phi_{zz}^{0,\mathrm{t}}} + \mu\widehat{\Phi_x^{0,\mathrm{t}}\Phi_z^{0,\mathrm{t}}}\right). \tag{29}$$





Guo et al (in prep)

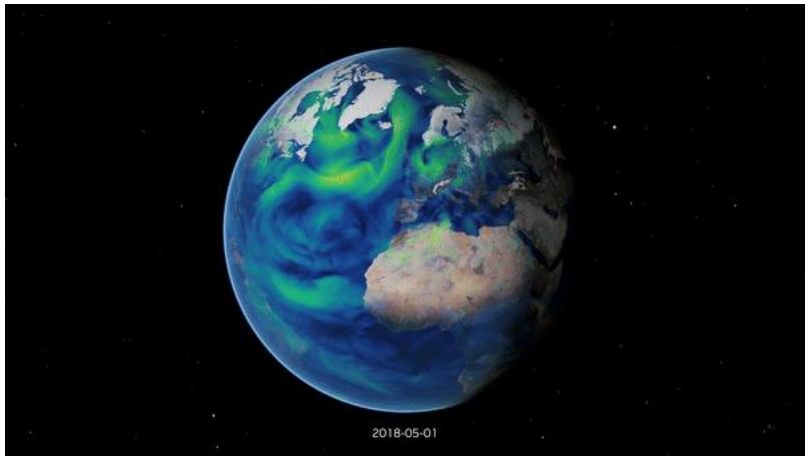
Example: physics informed neural network (PINN)



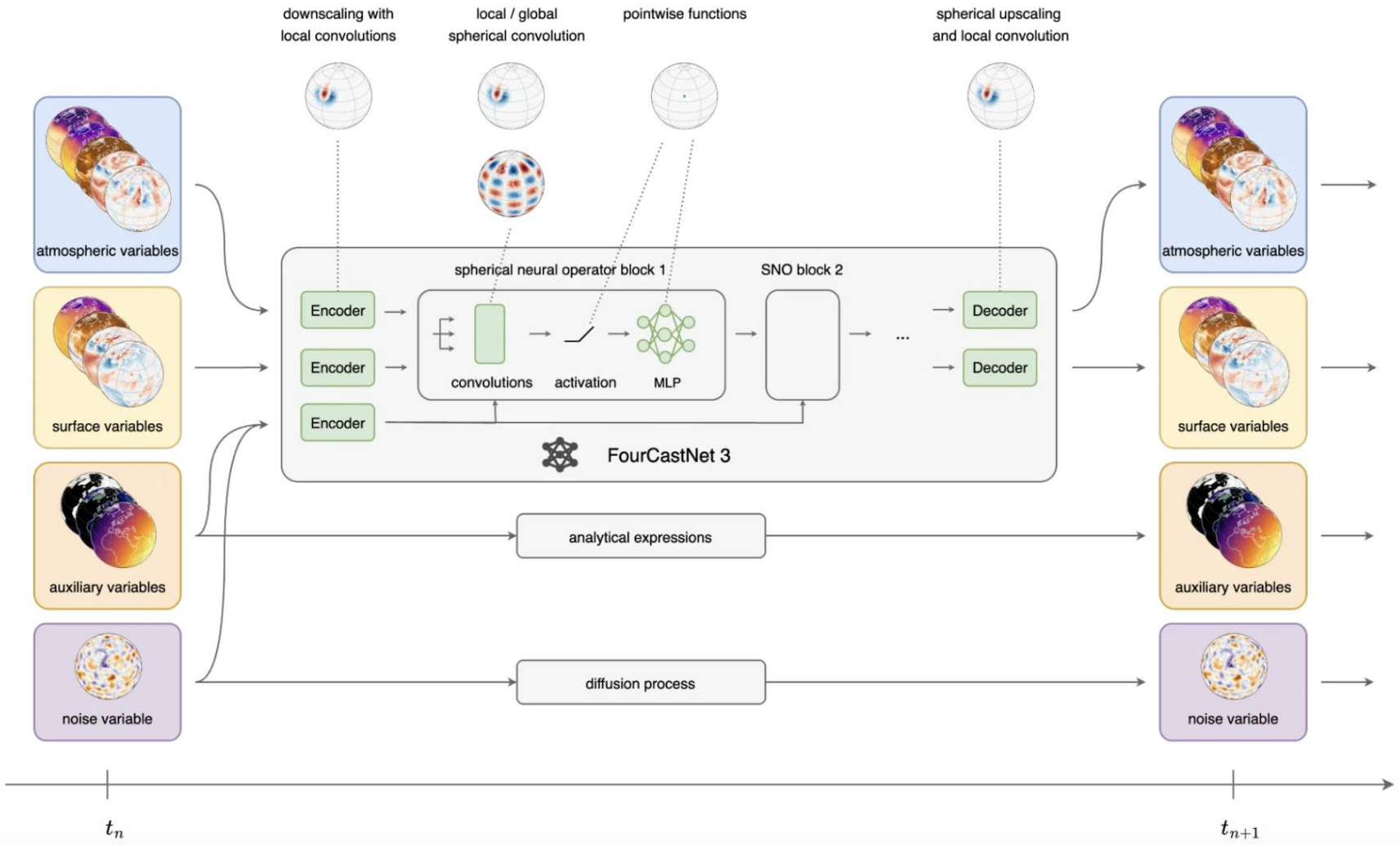
Guo et al (in prep)

Example: Forecasting

FourCastNet3

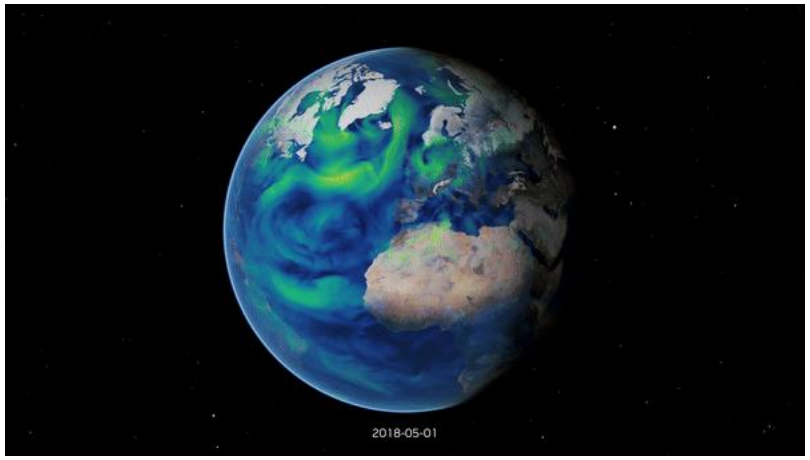


Example: Forecasting

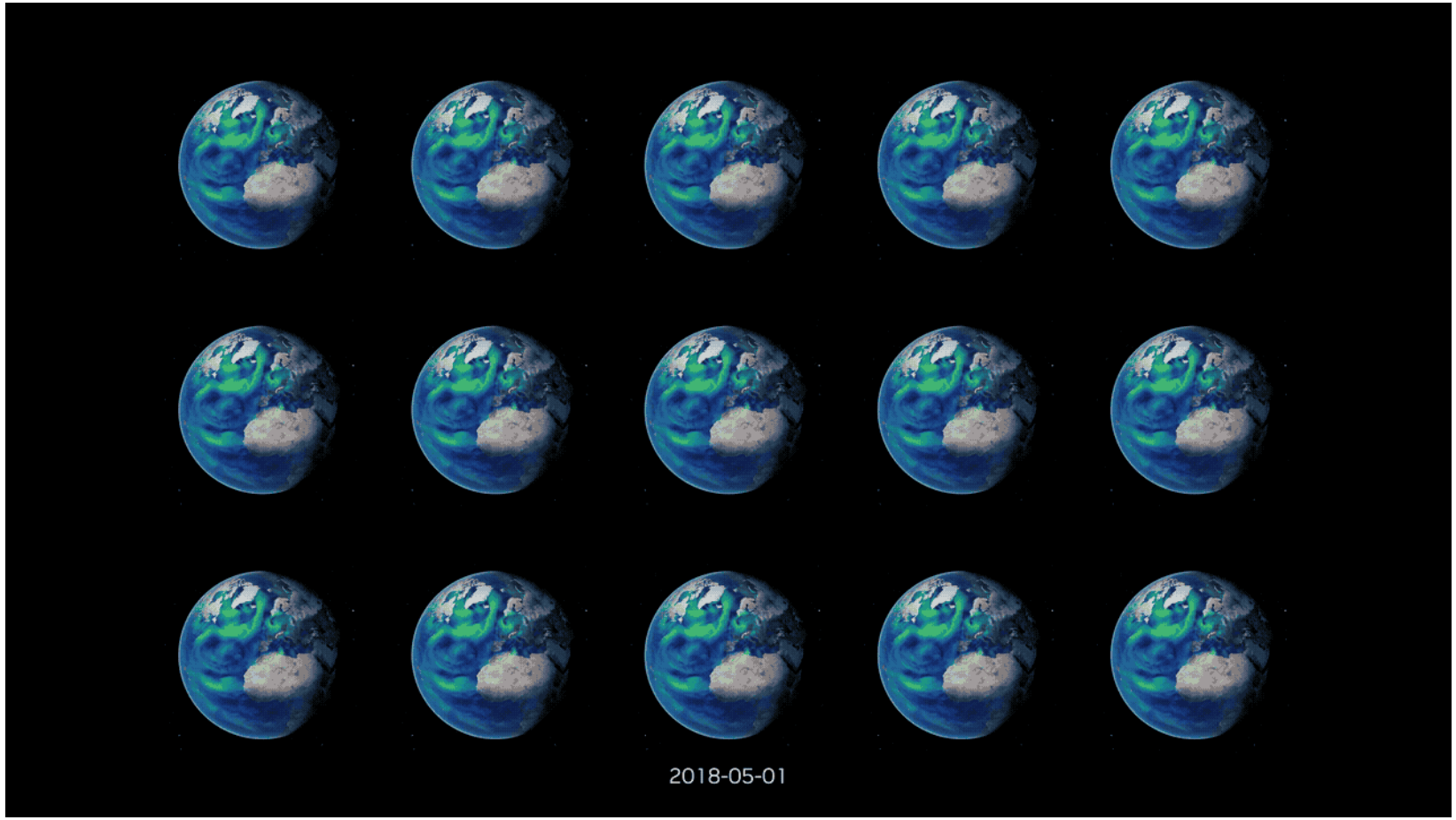


Example: Forecasting

FourCastNet3



Ensemble



Summary

- Data assimilation combines numerical models and observations to estimate the most accurate state of the Earth system (atmosphere, ocean, land, cryosphere).
- Forecast systems use methods like 3D-Var, 4D-Var, and Ensemble Kalman Filters to update model states based on observational data.
- These methods minimize the mismatch between model predictions and observations while accounting for uncertainties in both.
- AI is increasingly used in data assimilation to improve computational efficiency, integrate diverse data sources, and handle sparse or incomplete observations.