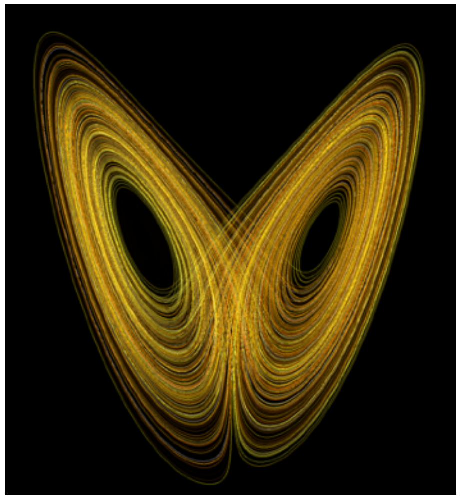
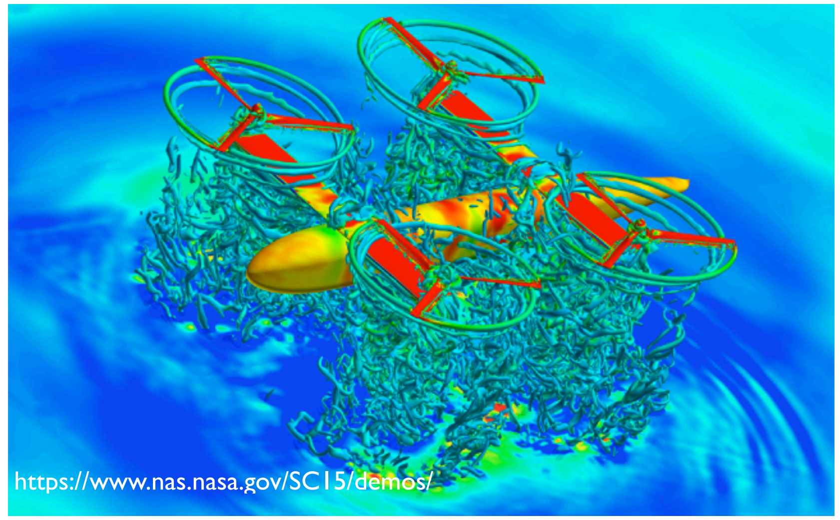
A Lagrangian Conditional Gaussian Koopman Network (LaCGKN) for Data Assimilation and Prediction

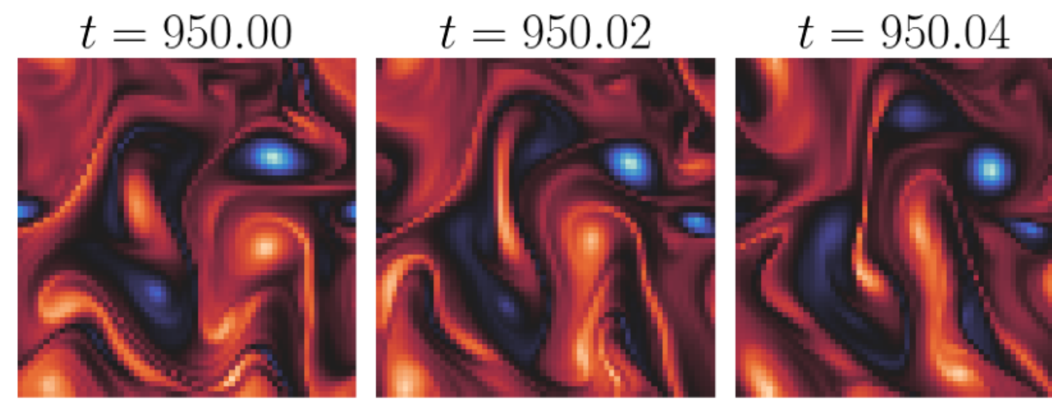
Zhongrui Wang[a], Chuanqi Chen[b], Jin-Long Wu[b], Nan Chen[a]

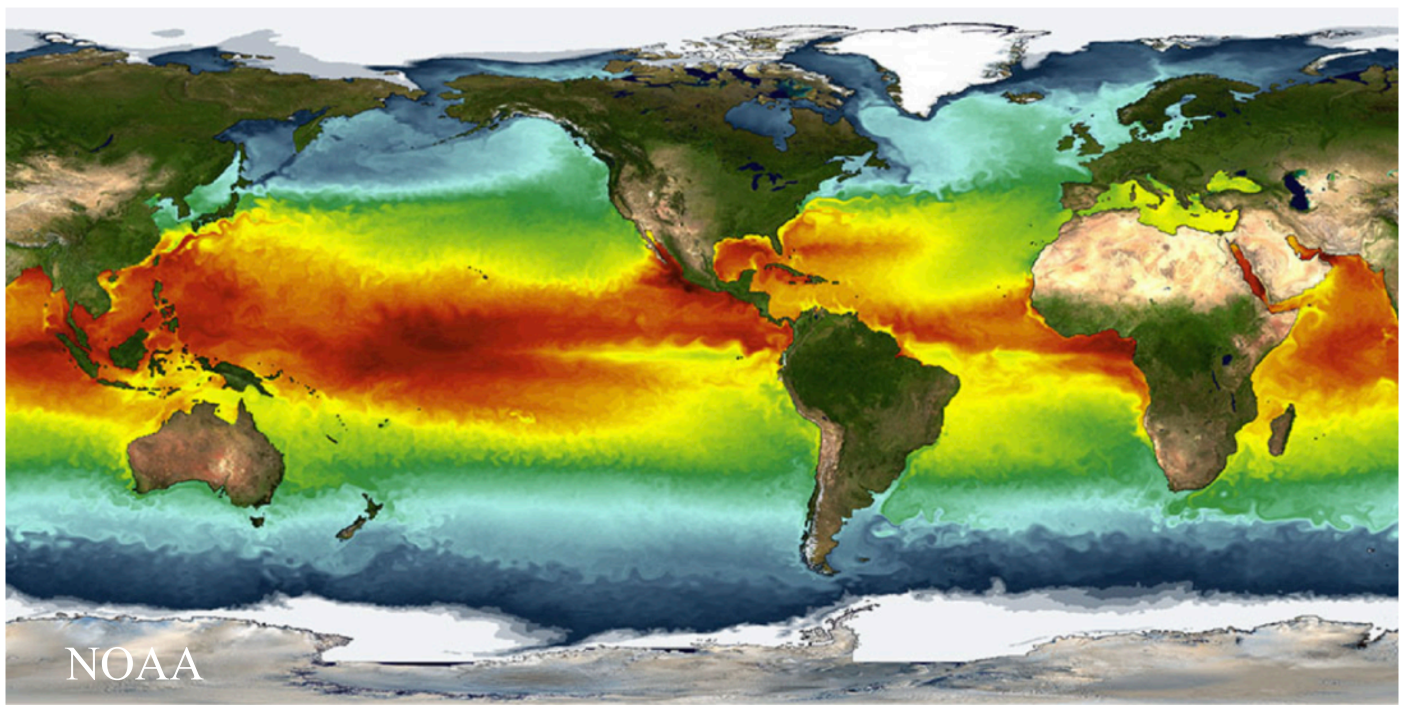
[a]*Department of Mathematics, University of Wisconsin–Madison*

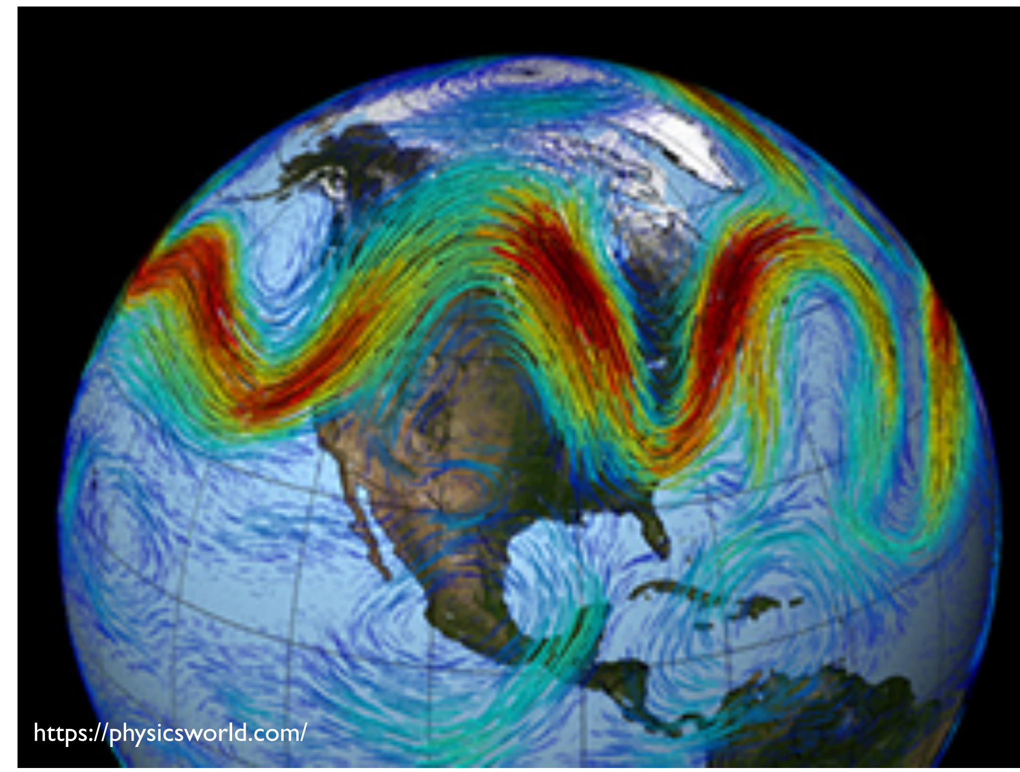
[b]*Department of Mechanical Engineering, University of Wisconsin–Madison*





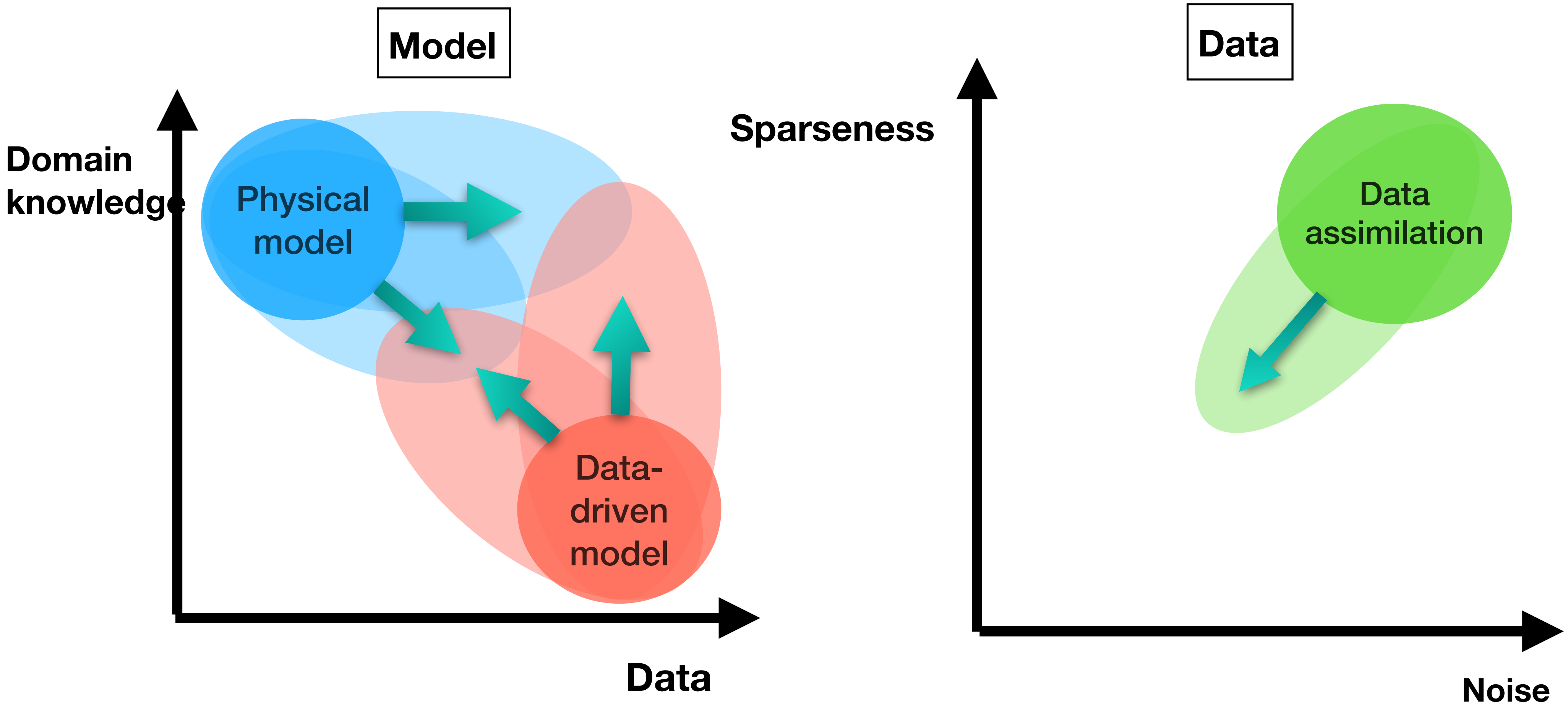






- **Complex dynamical systems**: nonlinear, chaotic, multi-scale, turbulent, intermittent, non-Gaussian; common in fluid dynamics, geophysics, neuroscience, material science…
- **Goal**:
  - Make predictions —> models
  - Use observations to improve predictions —> data assimilation (DA)

DA is widely used in real-time forecasting, parameter estimation, and optimal control.



- Physical model:
  - Based on governing equations derived from first principles (interpretable)
  - May require strong assumptions; Usually computationally expensive (e.g., Numerical Weather Prediction)
- Data-driven model:
  - Computationally efficient; Works with governing equations unknown
  - Lack of interpretability; May require a large amount of data (can be sparse and noisy in reality)
- Data assimilation is especially useful when data is sparse and noisy
  - Combing data with existing models, DA can recover complete data, with less uncertainty
  - As new observations become available, DA can utilize this information to improve real-time predictions

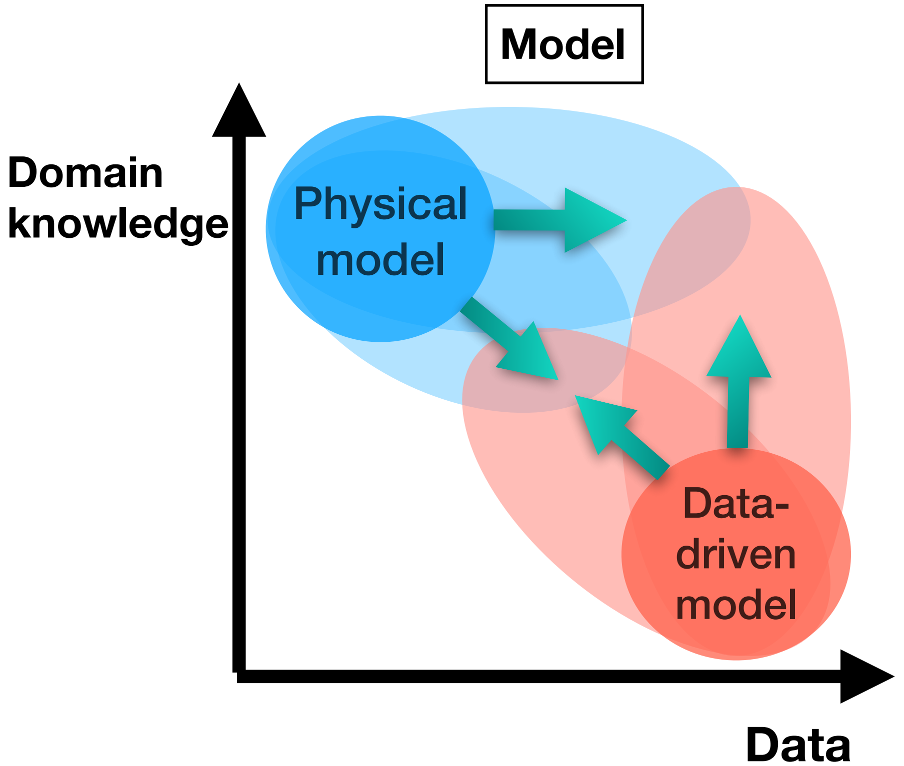
- **Review of Scientific Machine Learning (SciML) and DA**
- **Conditional Gaussian Koopman Network (CGKN):** a deep-learning framework that unifies SciML and DA
- **Lagrangian Data Assimilation**
- **Lagrangian Conditional Gaussian Koopman Network (LaCGKN)** for Lagrangian Data Assimilation and Prediction

Data-driven models for dynamical systems (SciML)

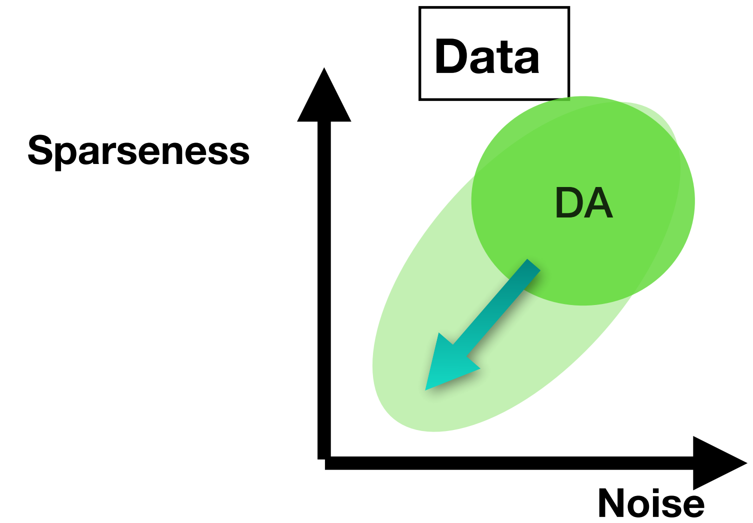
- *Reduced-order models*: linear stochastic models; dimension reduction;
- *Dynamics identification*: Sparse dynamics approximation (Schaeffer, 2012); SINDy via sparse regression (Brunton, 2016), causation entropy-based identification (Chen, 2023)
- *Recurrent Neural Networks* (Schuster and Paliwal, 1997; Gauthier et al., 2021);
- *Neural ODE (Chen, 2019)*;
- *Operator learning* (Lu, 2019; Li, 2020);
- *Gaussian processes* (Chen, 2021);

Combination with physical models:

- *Residual learning*: closure models (Levine and Stuart, 2022)
- *Physics-informed machine learning*: PINNs (Raissi, 2019)



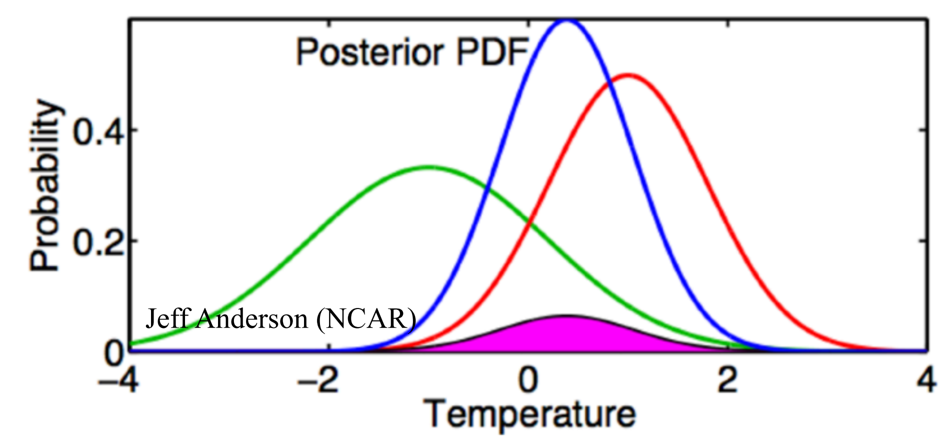
DA for using observations to improve predictions



- DA is based on **Bayes' rule**. It combines model predictions and observations to get a better state estimate; especially useful when observations are sparse and noisy

- **Traditional model-based DA methods:**
  - Solve the posterior using Bayes' rule **explicitly**, often relying on linear or Gaussian assumptions to make it parametric and computationally tractable.
  - e.g., Classical Kalman Filter (linear, Gaussian); Ensemble Kalman Filter (EnKF); Variational methods (3D/4D-Var); Particle Filter (PF)

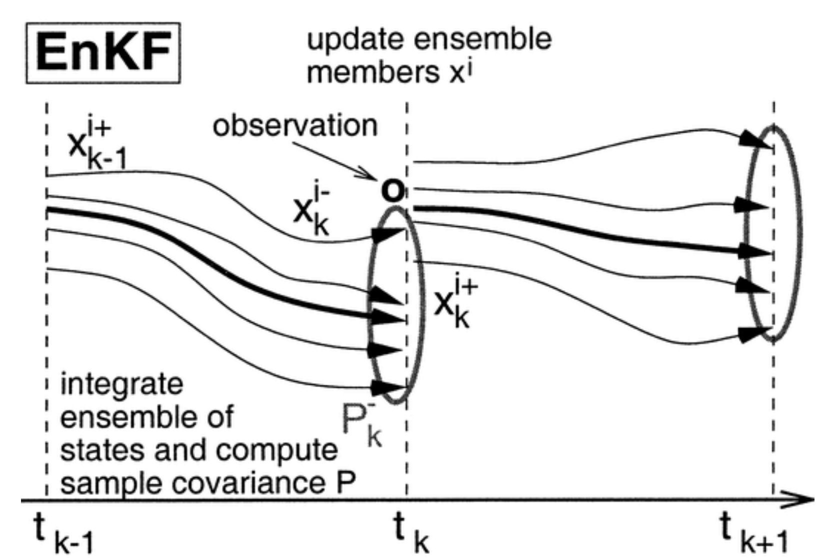
$$P(\mathbf{x}_{t_k}|\mathbf{Y}_{t_k}) = \frac{P(\mathbf{y}_k|\mathbf{x})P(\mathbf{x}_{t_k}|\mathbf{Y}_{t_{k-1}})}{Normalization}$$



Bayes rule (1D Gaussian case).

Approximations ↓↑ Costs

- **Challenges:**
  - nonlinearity, non-Gaussianity
  - high computational costs due to high dimensionality



[Reichle, R. H., 2002]

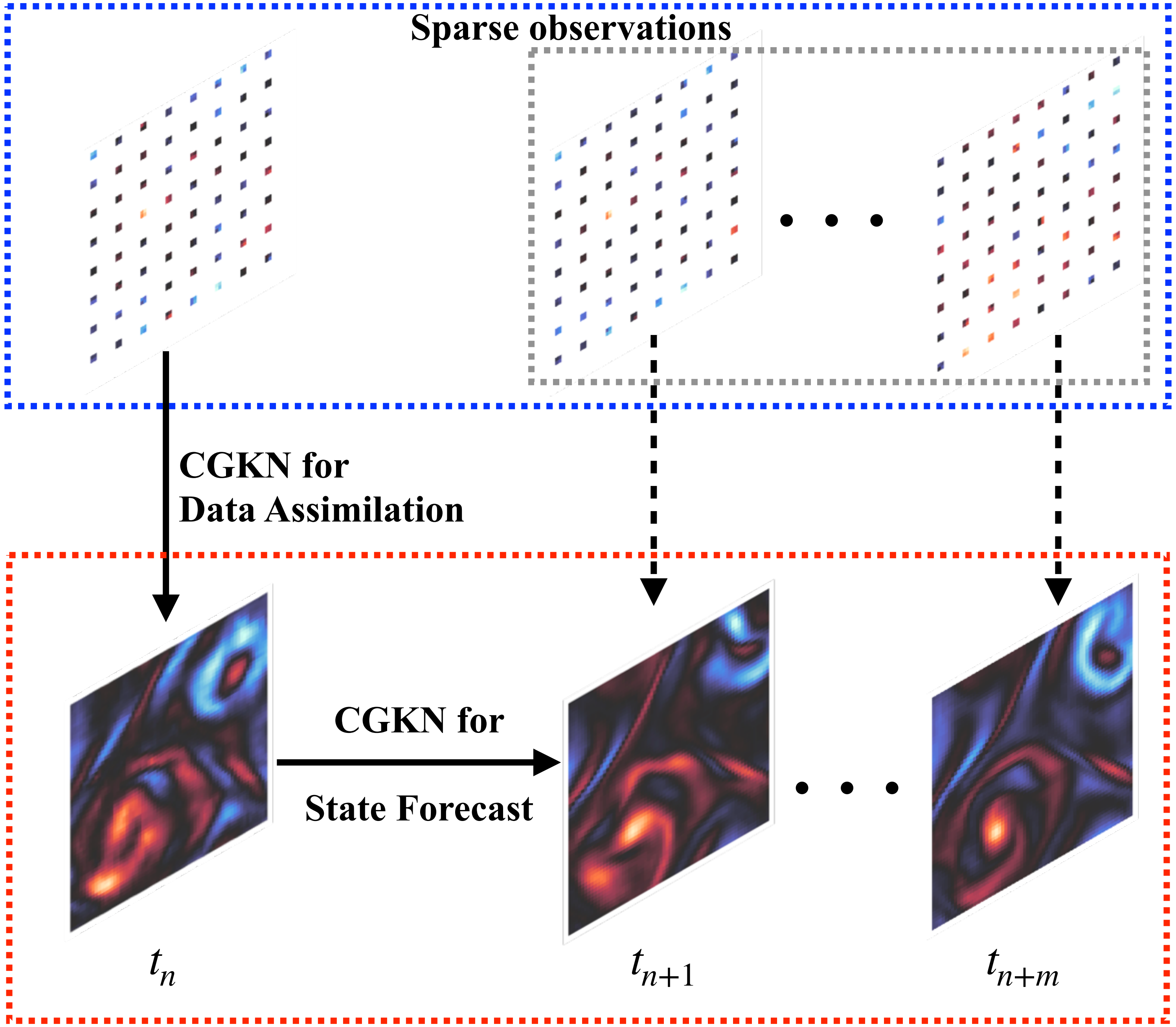
Combining SciML with DA

- **Data-enhanced/driven DA methods** (ML for DA)
  - ML models as cheap surrogate models to generate ensembles; model error correction; learning parameters of an (parameterized) analysis map;
  - Pure data-driven DA, e.g., generative DA based on conditional sampling, transport maps
  - Hybrid approaches that preserve analytically tractable nonlinear structures, e.g., CGNSDE (Chen et al., 2024), CGKN (Chen et al. 2025)
- **DA for ML**
  - DA analysis as better training data
  - Online correction of ML model predictions
  - Derivative-free optimization (EKI; Iglesias, 2013)

[Review: Cheng et al., 2023; Bach et al., 2025]

- **Reviewing Scientific Machine Learning (SciML) and DA**
- **Conditional Gaussian Koopman Network (CGKN):** a deep-learning framework that unifies SciML and DA
- **Lagrangian Data Assimilation**
- **Lagrangian Conditional Gaussian Koopman Network (LaCGKN)** for Lagrangian Data Assimilation and Prediction

Conditional Gaussian Koopman Network (CGKN)



- A **deep learning digital twins framework** that unifies SciML and DA, to learn surrogate models that performs **efficient DA** and **prediction (with UQ)** simultaneously for **nonlinear partially observed** dynamical systems.

[CGKN: A deep learning framework for modeling complex dynamical systems and efficient data assimilation. Chen et al. 2025]
[Modeling partially observed nonlinear dynamical systems and efficient data assimilation via discrete-time conditional Gaussian Koopman network, Chen et al. , 2025]

Motivation from DA (CG filter):

$\mathbf{u}_1$: observed states
$\mathbf{v}$: unobserved states

Conditional Gaussian Nonlinear System

$$\frac{\mathrm{d}\mathbf{u}_1}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{u}_1) + \mathbf{g}_1(\mathbf{u}_1)\mathbf{v} + \boldsymbol{\sigma}_1 \dot{\mathbf{W}}_1$$

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \mathbf{f}_2(\mathbf{u}_1) + \mathbf{g}_2(\mathbf{u}_1)\mathbf{v} + \boldsymbol{\sigma}_2 \dot{\mathbf{W}}_2$$

CGNS gallery
- The noisy Lorenz model
- Boussinesq equation
- Rotating shallow water equations
- The predator-prey model

…

[Chen and Majda, 2018]

CG filter

- The posterior distribution is Gaussian given the past observations up to time t.

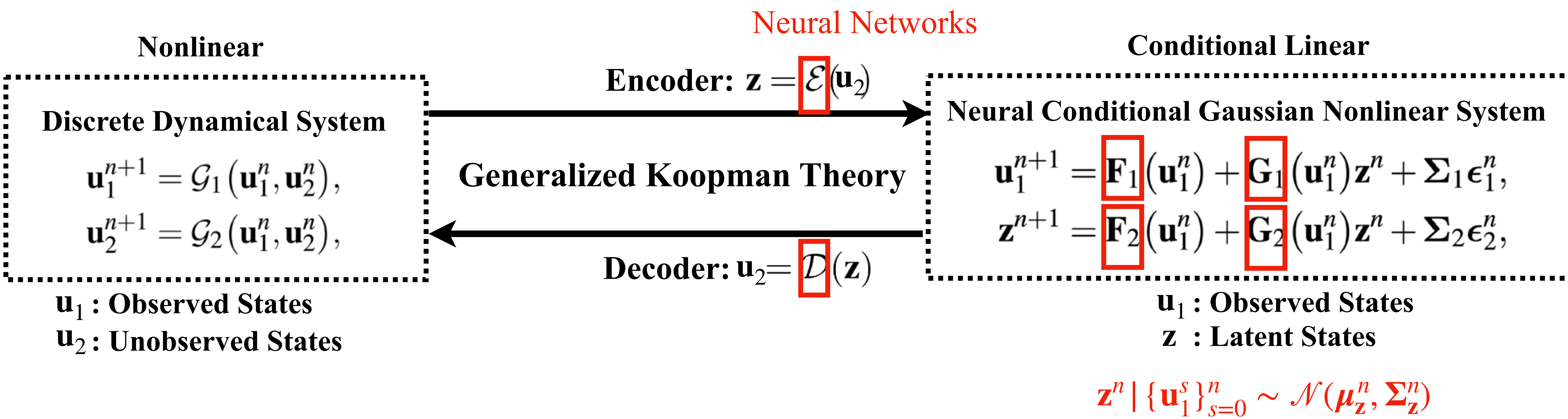
$$\mathbf{v}(t)|\mathbf{u}_1(s), s \leq t \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{v}}(t)), \mathbf{R}_{\mathbf{v}}(t))$$

- The posterior mean and covariance can be solved via analytic formulae.

$$\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{v}}}{\mathrm{d}t} = (\mathbf{f}_2 + \mathbf{g}_2\boldsymbol{\mu}_{\mathbf{v}}) + (\mathbf{R}_{\mathbf{v}}\mathbf{g}_1^{\mathrm{T}})(\boldsymbol{\sigma}_1\boldsymbol{\sigma}_1^{\mathrm{T}})^{-1}\left(\frac{\mathrm{d}\mathbf{u}_1}{\mathrm{d}t} - (\mathbf{f}_1 + \mathbf{g}_1\boldsymbol{\mu}_{\mathbf{v}})\right)$$

$$\frac{\mathrm{d}\mathbf{R}_{\mathbf{v}}}{\mathrm{d}t} = \mathbf{g}_2\mathbf{R}_{\mathbf{v}} + \mathbf{R}_{\mathbf{v}}\mathbf{g}_2^{\mathrm{T}} + \boldsymbol{\sigma}_2\boldsymbol{\sigma}_2^{\mathrm{T}} - \mathbf{R}_{\mathbf{v}}\mathbf{g}_1^{\mathrm{T}}(\boldsymbol{\sigma}_1\boldsymbol{\sigma}_1^{\mathrm{T}})^{-1}(\mathbf{g}_1\mathbf{R}_{\mathbf{v}})$$

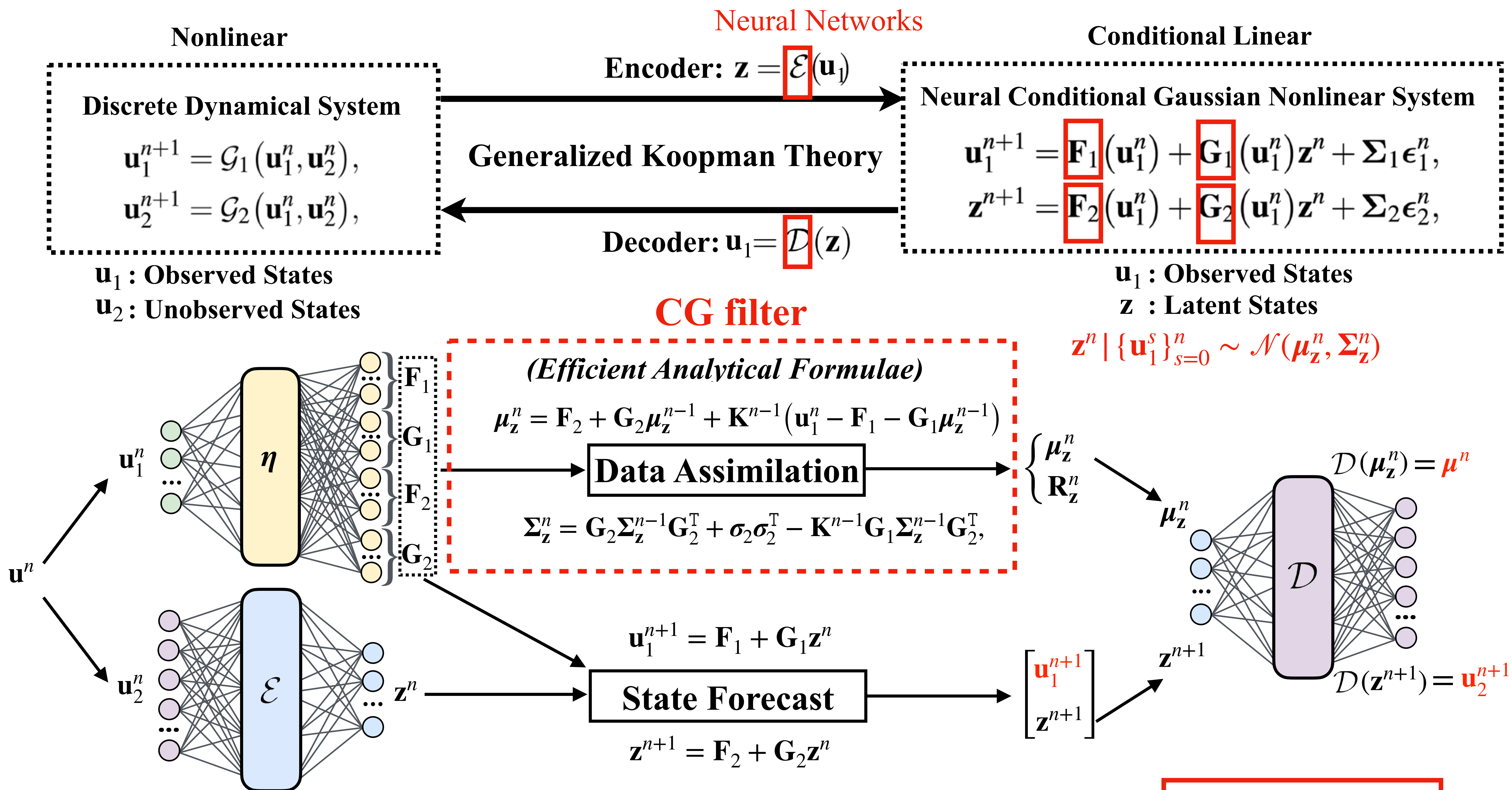
Motivation from Koopman theory:



- The latent representation $\mathbf{z}$ is conditional linear given $\mathbf{u}_1$ being observed, which leads to a neural conditional Gaussian nonlinear system (CGNS) that has analytic DA formulae.
- Instead of directly applying Koopman theory, CGKN only seeks for embeddings of the unobserved states $\mathbf{u}_2$.

[Koopman, 1931; Brunton, 2022]

CGKN (Discrete-time)

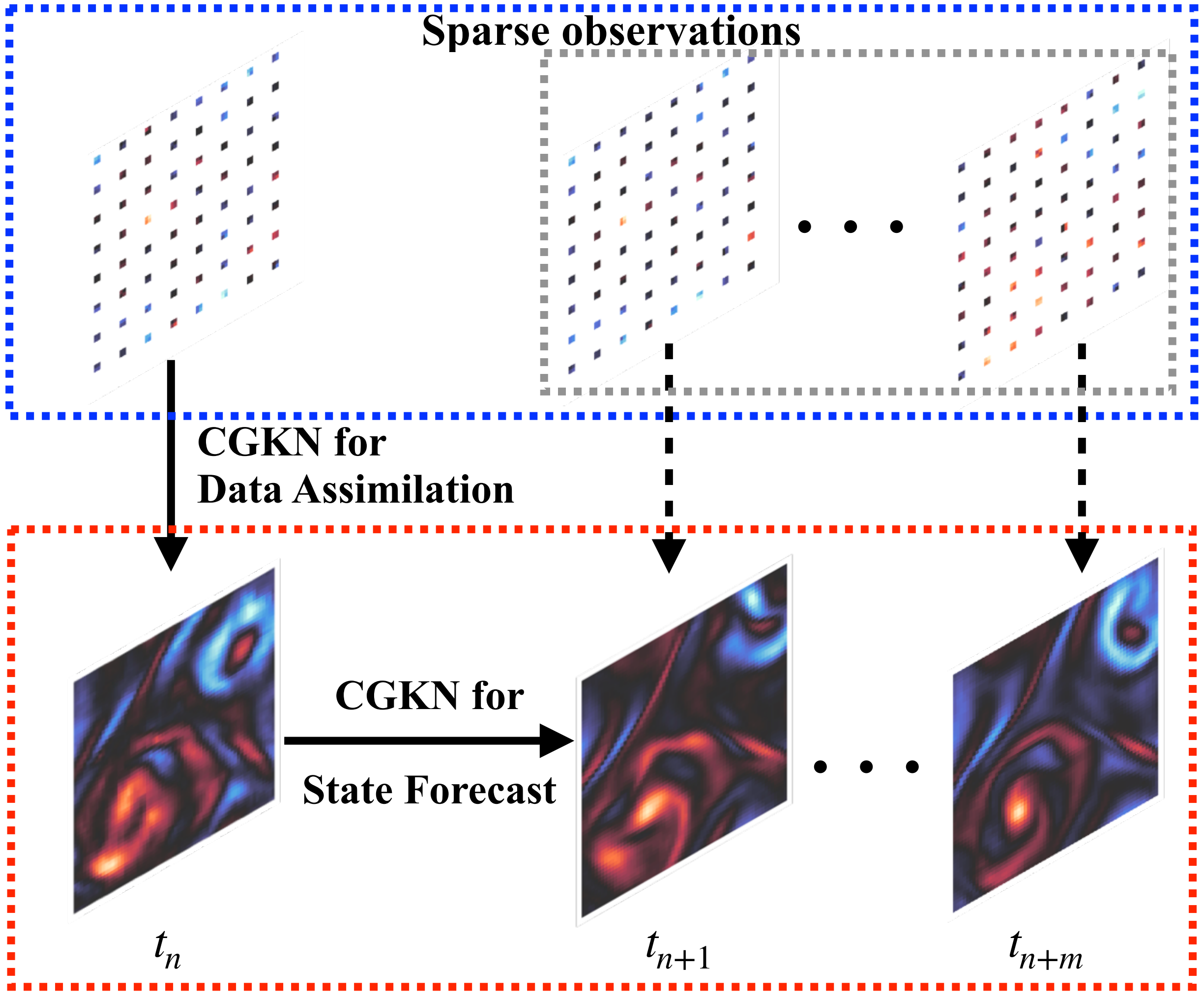


$$L(\boldsymbol{\theta}_\mathcal{E}, \boldsymbol{\theta}_\mathcal{D}, \boldsymbol{\theta}_\boldsymbol{\eta}) := \lambda_{\mathrm{AE}} L_{\mathrm{AE}}(\boldsymbol{\theta}_\mathcal{E}, \boldsymbol{\theta}_\mathcal{D}) + \lambda_\mathbf{u} L_\mathbf{u}(\boldsymbol{\theta}_\mathcal{E}, \boldsymbol{\theta}_\mathcal{D}, \boldsymbol{\theta}_\boldsymbol{\eta}) + \lambda_\mathbf{z} L_\mathbf{z}(\boldsymbol{\theta}_\mathcal{E}, \boldsymbol{\theta}_\boldsymbol{\eta}) + \lambda_{\mathrm{DA}} L_{\mathrm{DA}}(\boldsymbol{\theta}_\mathcal{D}, \boldsymbol{\theta}_\boldsymbol{\eta})$$

Total loss | Auto-encoder loss | Forecast loss of physical variables | Forecast loss of latent variables | Data assimilation loss

- State forecast and DA are performed in the latent space (with reduced dimension)
- DA formulae is part of the model structure as inductive bias, and learning is constrained by DA loss.

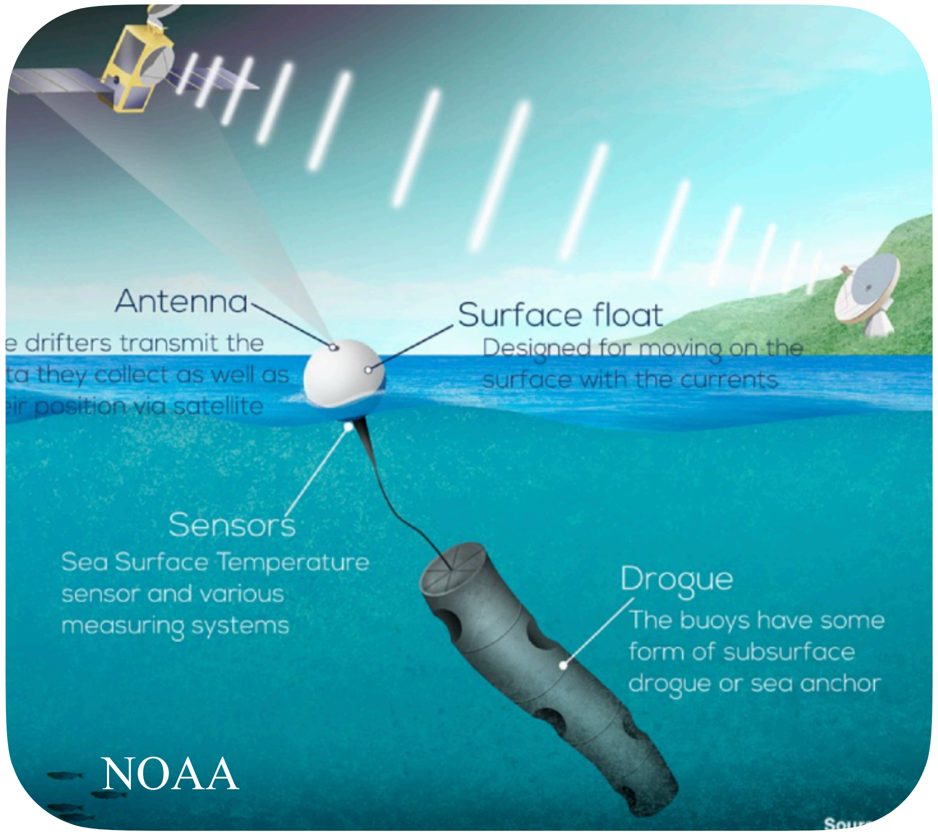
CGKN (Discrete-time)

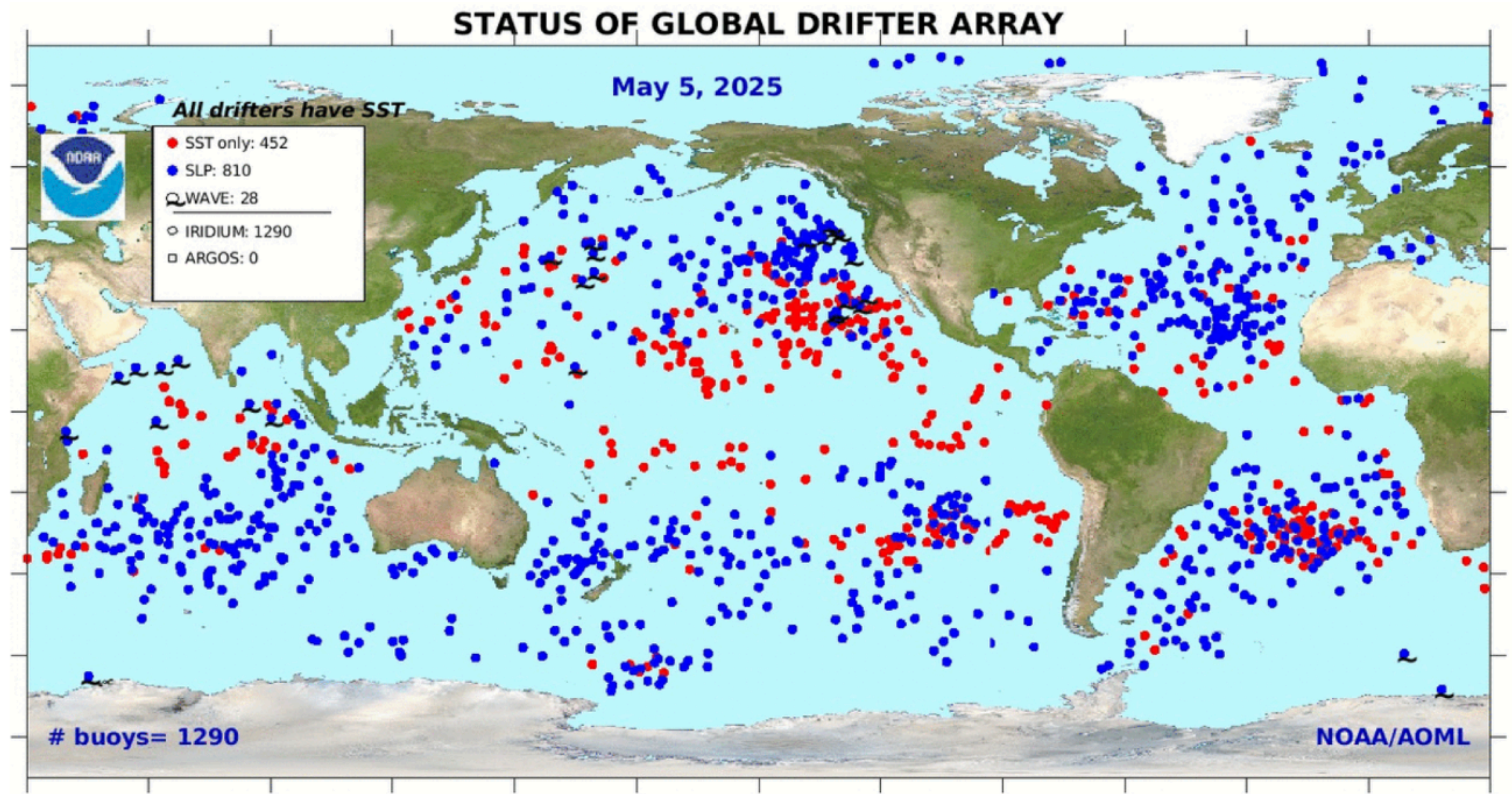


- Make predictions with only **partially and noisy observed initial conditions**, and **improve predictions as new observations come (DA)**
- *Analytic DA* formulae of CG filter (model-based DA):
  - Ensures **accuracy** and **efficiency** of DA (avoid using ensembles)
  - Introduces **inductive bias** to model structure, and constrains the learning
  - Also facilitates the downstream tasks like **uncertainty quantification (UQ)**
- Bridges model-based DA and data-driven DA, leading to a **unified framework for SciML and DA**.

- **Reviewing Scientific Machine Learning (SciML) and DA**
- **Conditional Gaussian Koopman Network (CGKN):** a deep-learning framework that unifies SciML and DA
- **Lagrangian Data Assimilation**
- **Lagrangian Conditional Gaussian Koopman Network (LaCGKN)** for Lagrangian Data Assimilation and Prediction

Lagrangian data assimilation





- Lagrangian DA uses **Lagrangian** observations to recover the underlying **Eulerian** flow.
  - e.g., inferring deep ocean states with surface observations; inferring upper atmosphere with ground-surface observations
- Lagrangian DA has **increasing practical importance** due to a growing amount of Lagrangian observations in operational oceanic and atmospheric observing systems.

Lagrangian DA: a canonical example

- Consider Lagrangian observations of I passive tracer positions driven by an hidden Eulerian flow:

$$\text{Tracer:} \qquad \mathbf{x}_i^{n+1} = \mathcal{T}_h\big(\mathbf{x}_i^n, \mathbf{v}^n\big) + \boldsymbol{\Sigma}_{\mathbf{x}_i} \Delta\mathbf{W}_i^n, \quad i = 1, \ldots, I,$$

$$\text{Flow:} \qquad \mathbf{v}^{n+1} = \mathcal{F}_h\big(\mathbf{v}^n\big) + \boldsymbol{\Sigma}_{\mathbf{v}} \Delta\mathbf{W}_{\mathbf{v}}^n,$$

- The Lagrangian filtering problem aims to find the posterior of the unobserved flow given past tracer observations:

$$p\big(\mathbf{v}^n \big| \{\mathbf{x}_i^s\}_{i=1,s=0}^{I,n}\big)$$

- The tracer operator $\mathcal{T}_h$ involves an interpolation or integral operator that evaluates flow $\mathbf{v}$ at a local tracer position $\mathbf{x}_i$, which is typically **nonlinear**.

- The **observation process** of Lagrangian DA is **intrinsically nonlinear**, making it more **challenging** than Eulerian DA.

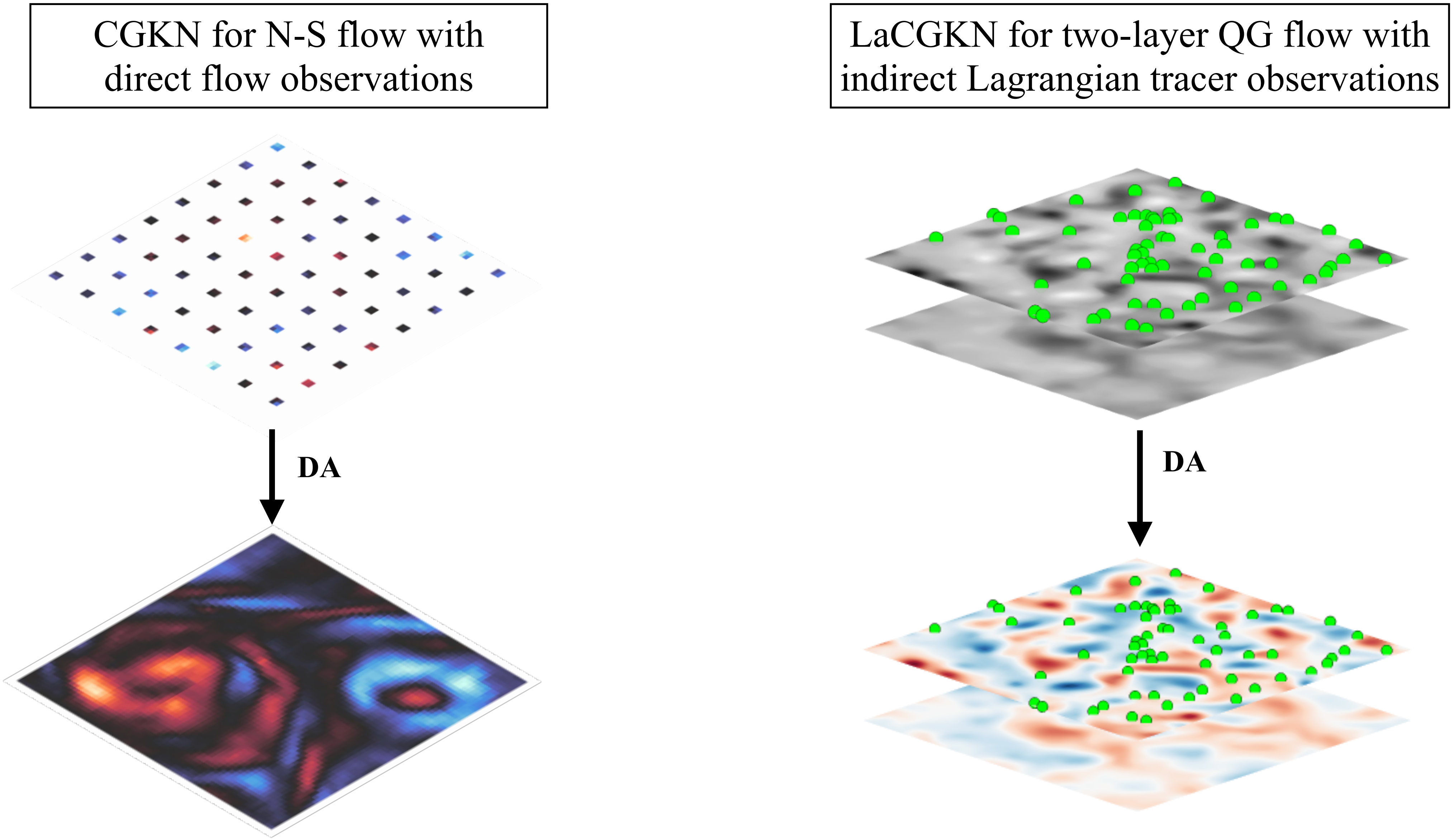
Lagrangian DA: approaches

- **Traditional deductive (model-based) methods**
  - Kalman-based methods: extended Kalman filter (EKF; Ide et al. 2002); optimal interpolation (OI; Molcard et al. 2003); local ensemble transform Kalman filter (LETKF; Sun & Penny, 2019)
  - Partical filters (PF), Markov chain Monte Carlo (MCMC) smoothers (Apte et al., 2008)
  - Hybrid approaches, e.g., combining PF for nonlinear tracer dynamics with EnKF for near-Gaussian flow estimation
- **ML inductive (data-driven) methods**
  - Combining neural operators with generative models (Asefi et al., 2025)
  - Multimodal contrastive learning (Baptista et al., 2025)

- ML for Lagrangian DA remains relatively limited.
- ML is particularly well suited to addressing the strong nonlinearity inherent in Lagrangian DA, holding substantial potential to improve accuracy and efficiency.

- **Reviewing Scientific Machine Learning (SciML) and DA**
- **Conditional Gaussian Koopman Network (CGKN):** a deep-learning framework that unifies SciML and DA
- **Lagrangian Data Assimilation**
- **Lagrangian Conditional Gaussian Koopman Network (LaCGKN)** for Lagrangian Data Assimilation and Prediction

CGKN for Lagrangian DA: Challenges



- Another challenge: compared to the earlier application of CGKN to N-S equations with *direct* observations of the flow, Lagrangian DA uses **sparse and *indirect*** Lagrangian observations.

CGKN for Lagrangian DA: Challenges

Nonlinear

Tracer-flow system

$$\mathbf{x}_i^{n+1} = \mathcal{T}_h(\mathbf{x}_i^n, \mathbf{v}^n) + \boldsymbol{\Sigma}_{\mathbf{x}_i} \Delta \mathbf{W}_i^n,$$

$$\mathbf{v}^{n+1} = \mathcal{F}_h(\mathbf{v}^n) + \boldsymbol{\Sigma}_{\mathbf{v}} \Delta \mathbf{W}_{\mathbf{v}}^n,$$

$\mathbf{x}_i$: Tracer, Observed
$\mathbf{v}$: Flow, Unobserved

Encoder: $\mathbf{z} = \mathcal{E}(\mathbf{v})$ →

Generalized Koopman Theory

← **Decoder:** $\mathbf{v} = \mathcal{D}(\mathbf{z})$

Conditional Linear

Neural Conditional Gaussian Nonlinear System

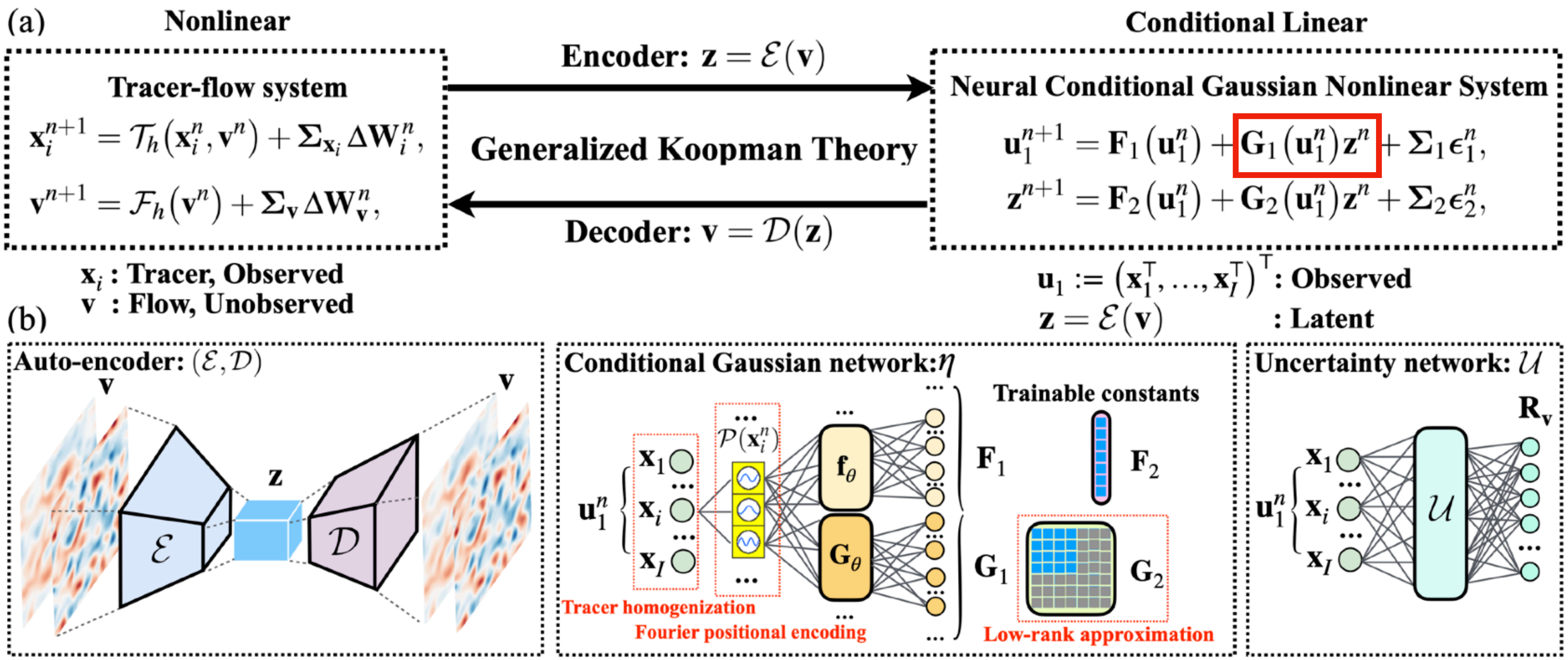
$$\mathbf{u}_1^{n+1} = \mathbf{F}_1(\mathbf{u}_1^n) + \mathbf{G}_1(\mathbf{u}_1^n)\mathbf{z}^n + \boldsymbol{\Sigma}_1 \boldsymbol{\epsilon}_1^n,$$

$$\mathbf{z}^{n+1} = \mathbf{F}_2(\mathbf{u}_1^n) + \mathbf{G}_2(\mathbf{u}_1^n)\mathbf{z}^n + \boldsymbol{\Sigma}_2 \boldsymbol{\epsilon}_2^n,$$

$\mathbf{u}_1 := (\mathbf{x}_1^\top, \ldots, \mathbf{x}_I^\top)^\top$: Observed
$\mathbf{z} = \mathcal{E}(\mathbf{v})$: Latent

- Lagrangian DA is thus a more **challenging yet practical** problem **for CGKN**:
  - not only the nonlinear flow dynamics of $\mathbf{v}$ needs to be well approximated by a latent linear dynamics of $\mathbf{z}$ (as in standard Koopman theory)
  - but also the nonlinear tracer dynamics should be well approximated by the neural tracer dynamics driven by latent flow. → $\mathbf{G}_1(\mathbf{u}_1^n)\mathbf{z}^n$
- **The latter is crucial to data assimilation**, as it captures the information propagation from observed states to unobserved states. This imposes additional requirements and regularizations for **latent embedding z**.

LaCGKN structure design



1. **Homogenization over tracers.** Lagrangian tracers can often be assumed to be homogeneous. We therefore construct $\mathbf{F}_1$ and $\mathbf{G}_1$ by applying the same neural networks to each tracer position independently:

$$\mathbf{F}_1(\mathbf{u}_1^n) := \left(\mathbf{f}_\theta(\mathbf{x}_1^n)^\top, \ \ldots, \ \mathbf{f}_\theta(\mathbf{x}_I^n)^\top\right)^\top,$$

$$\mathbf{G}_1(\mathbf{u}_1^n) := \left(\mathbf{G}_\theta(\mathbf{x}_1^n)^\top, \ \ldots, \ \mathbf{G}_\theta(\mathbf{x}_I^n)^\top\right)^\top,$$

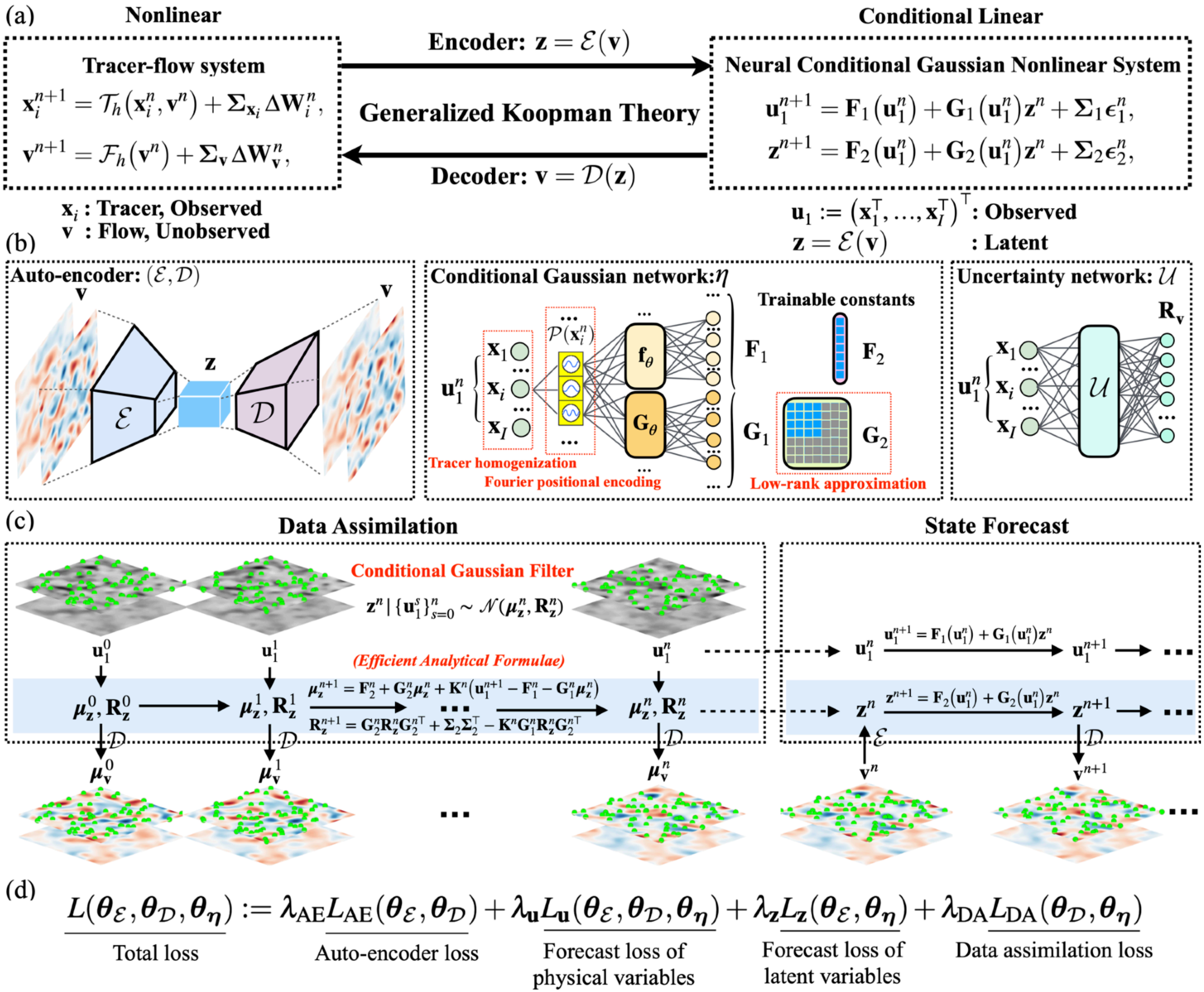
2. **Fourier positional encoding.** To accurately reconstruct the local value of the flow field at moving tracer locations, $\mathbf{G}_1$ must learn a rich nonlinear dependence on position $\mathbf{x}$. We therefore adopt the Fourier positional encoding:

$$\mathcal{P}(\mathbf{x}_i^n) = \left[x_{i,1}^n, \ldots, x_{i,d}^n, \ \left\{\sin(2^k \pi x_{i,j}^n), \cos(2^k \pi x_{i,j}^n)\right\}_{j=1,k=0}^{d,K-1}\right]$$

3. **Low-rank approximation of $\mathbf{G}_2$.** To control the computational complexity while the latent dimension scales up, a SVD-inspired low-rank approximation of $\mathbf{G}_2$ is adopted:

$$\mathbf{G}_2 = \mathbf{U}\,\mathrm{diag}(\mathbf{s})\,\mathbf{V}^\top + \mathrm{diag}(\boldsymbol{\delta})$$

LaCGKN overview



Numerical tests

- **Tested case:** A **two-layer quasi-geostrophic (QG) flow** with passive tracer position observations (Tracers are advected by the upper-layer flow).

Tracer: $$\left\{ \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = \mathbf{v}(\mathbf{x}_i, t) + \boldsymbol{\Sigma}_{\mathbf{x}_i}\dot{\mathbf{W}}_i, \quad i = 1, \ldots, I \right.$$

Flow: $$\begin{cases} \dfrac{\partial q_1}{\partial t} + J(\psi_1, q_1) + \beta \dfrac{\partial \psi_1}{\partial x} + U_1 \dfrac{\partial}{\partial x}\nabla^2 \psi_1 + \dfrac{k_d^2}{2}\left(U_1 \dfrac{\partial \psi_2}{\partial x} - U_2 \dfrac{\partial \psi_1}{\partial x}\right) = -\nu \Delta^s q_1, \\ \dfrac{\partial q_2}{\partial t} + J(\psi_2, q_2) + \beta \dfrac{\partial \psi_2}{\partial x} + U_2 \dfrac{\partial}{\partial x}\nabla^2 \psi_2 + \dfrac{k_d^2}{2}\left(U_2 \dfrac{\partial \psi_1}{\partial x} - U_1 \dfrac{\partial \psi_2}{\partial x}\right) = -\left(U_2 \dfrac{\partial h}{\partial x} + \kappa \nabla^2 \psi_2\right) - \nu \Delta^s q_2, \\ q_1 = \nabla^2 \psi_1 + \dfrac{k_d^2}{2}(\psi_2 - \psi_1), \quad q_2 = \nabla^2 \psi_2 + \dfrac{k_d^2}{2}(\psi_1 - \psi_2) + h, \quad \mathbf{v}_\ell = \left(\dfrac{\partial \psi_\ell}{\partial y}, -\dfrac{\partial \psi_\ell}{\partial x}\right)^\top \end{cases}$$

Quantity of interest (to be recovered): stream function $\{\psi_1, \psi_2\}$

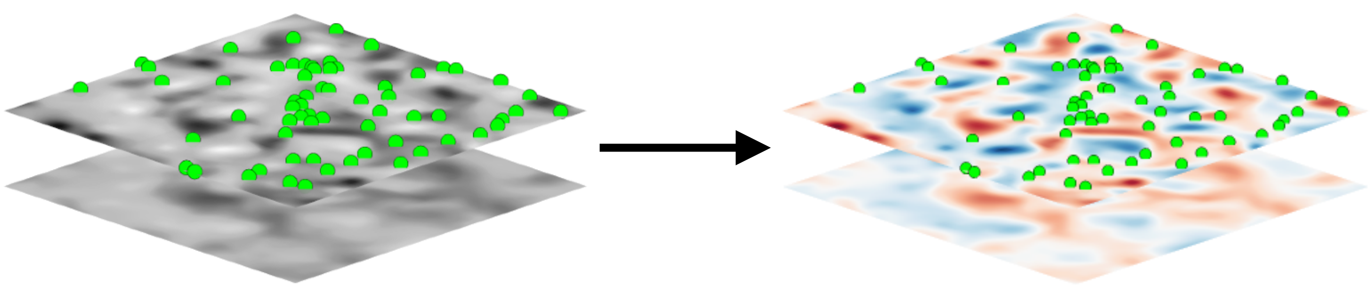
- **Compared benchmarks:**
  - LaCGKN
  - DNN for tracer + CNN for flow. (Predication)
  - Persistence
  - Ensemble Kalman Filter (EnKF) (Data assimilation)
  - Optimal interpolation (OI)
  - Climatology.

Two-layer QG flow with surface tracer observations

Experimental Settings

- Parmeters: $k_d = 10$, $\beta = 22$, $U = 1$, $\kappa = 9$, $\nu = 10^{-12}$, and $h(x,y) = 40\cos x + 80\cos(2y)$
- Data generated on 128×128 pseudo-spectral grid over $\Omega = [0,2\pi)^2$, $N_t = 2 \times 10^6$ time steps, $\Delta t = 2 \times 10^{-3}$
- Data are then sub-sampled to 64×64 grids, $\Delta t_{\text{obs}} = 4 \times 10^{-2}$
- Training/validation/test data: 80,000/10,000/10,000 steps
- Fourier positional encoding with $K = 6$ frequencies. Low-rank approximation of $\mathbf{G}_2$ with effective rank $r = 64$
- CGKN hyperparameters: $N_s = 1$, $N_l = 100$, $N_b = 20$, $\lambda_{\text{AE}} = \lambda_{\mathbf{u}} = \lambda_{\mathbf{v}} = \lambda_{\text{DA}} = 1$.
- EnKF ensemble size=40; Both EnKF and OI use the perfect QG model for flow forecast

- Observation: positions of **64** tracers with measurement noise $\sim \mathcal{N}(0, 0.01^2)$; unobserved states: $\mathbf{64 \times 64 \times 2}$ (two-layer flow)



- Encoder $\mathbb{R}^{64\times64\times2} \mapsto \mathbb{R}^{16\times16\times2}$ (LaCGKN) / $\mathbb{R}^{64\times64\times2} \mapsto \mathbb{R}^{32\times32\times2}$ (LaCGKN_{32}), decoder mirrors encoder

Summary of numerical results

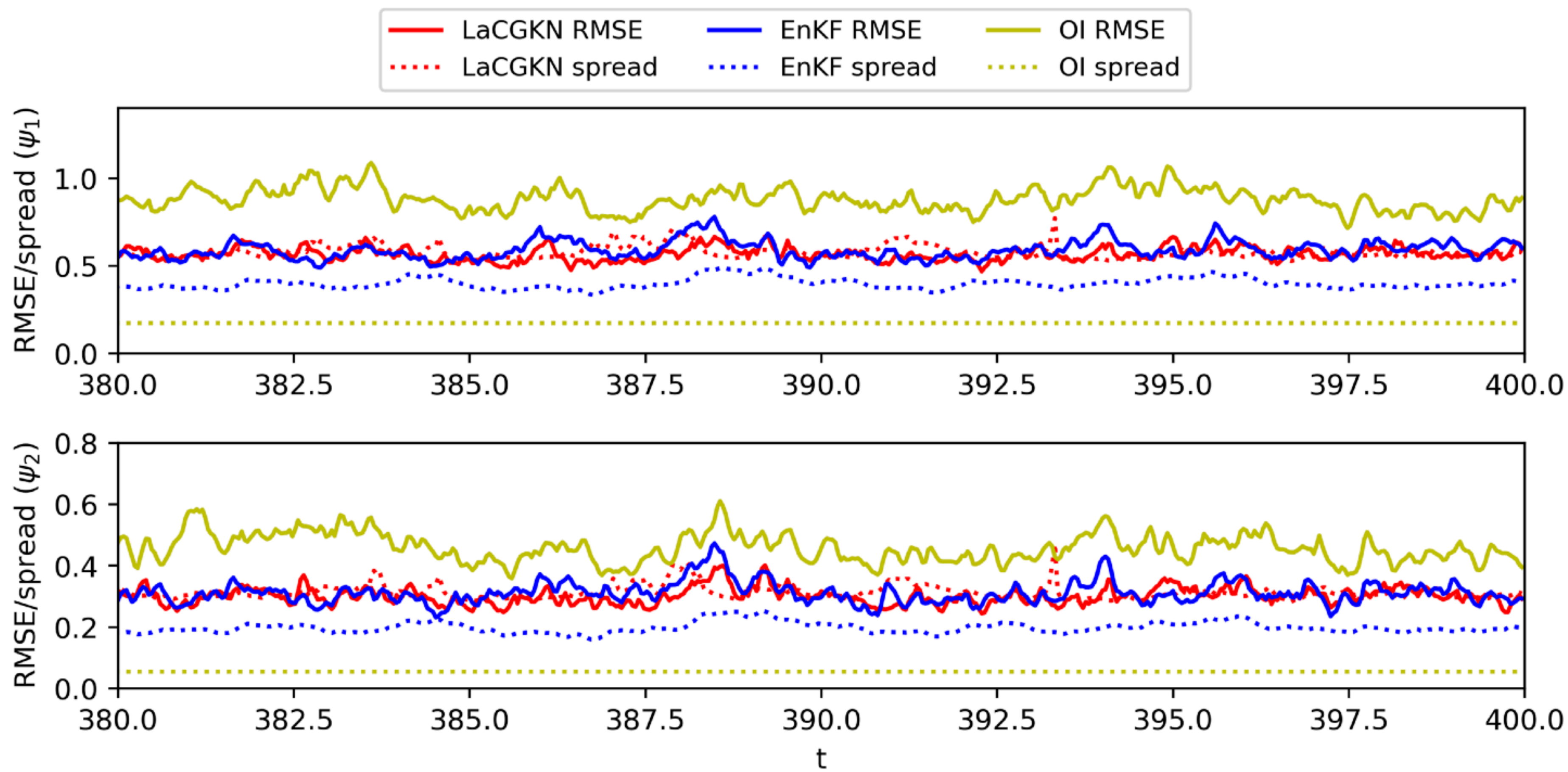
Table 2: Relative RMSEs of state forecast (one-step prediction)

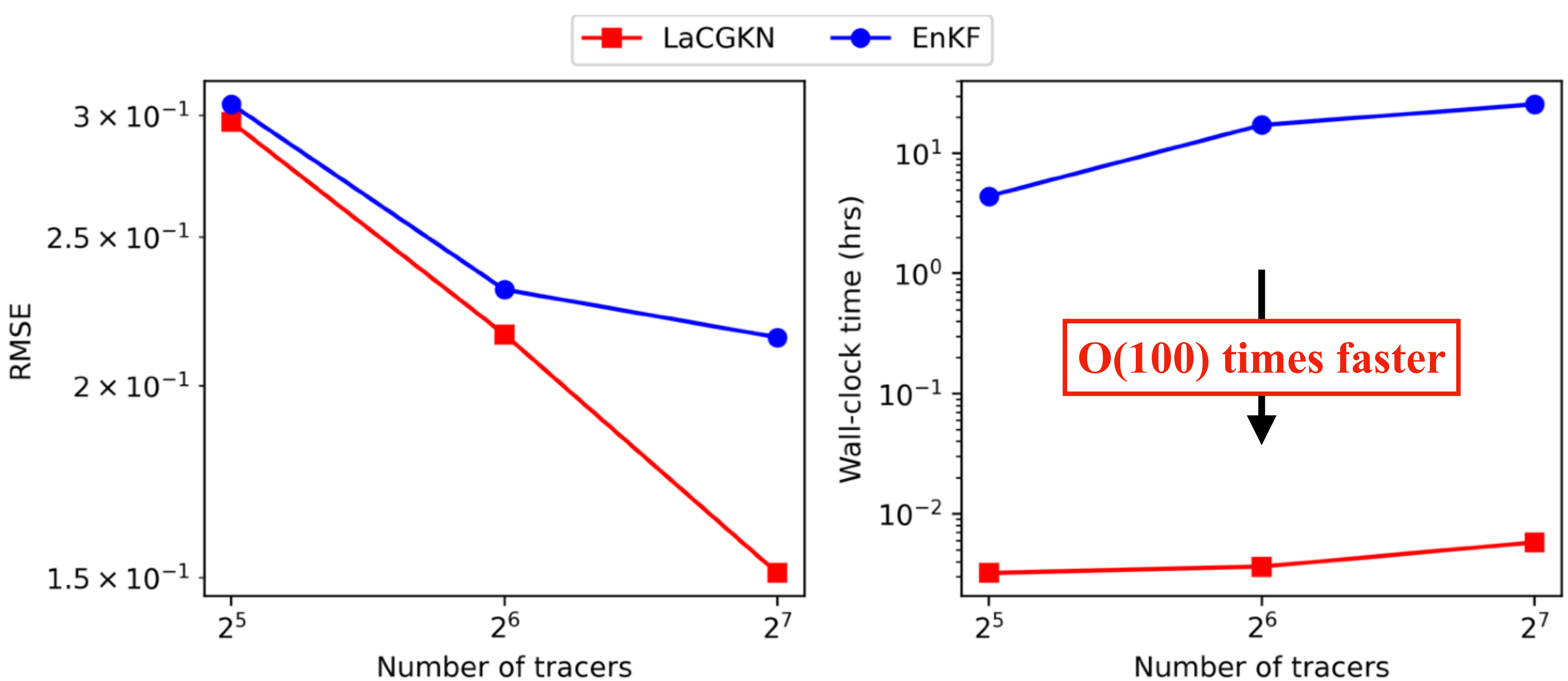
Method	Tracer	Upper Layer	Lower Layer	Two Layers
LaCGKN	0.099	0.125	0.079	0.104
LaCGKN_{32}	0.094	0.042	0.032	0.037
DNN+CNN	0.064	0.071	0.069	0.070
Persistence	0.136	0.294	0.177	0.243

Table 4: Relative RMSEs of data assimilation posterior estimates.

Method	Upper Layer	Lower Layer	Two Layers
LaCGKN	0.579	0.310	0.464
EnKF	0.599	0.321	0.481
OI	0.890	0.467	0.710
Climatology	0.870	0.414	0.681

DA Performance (with UQ): Accuracy and Efficiency

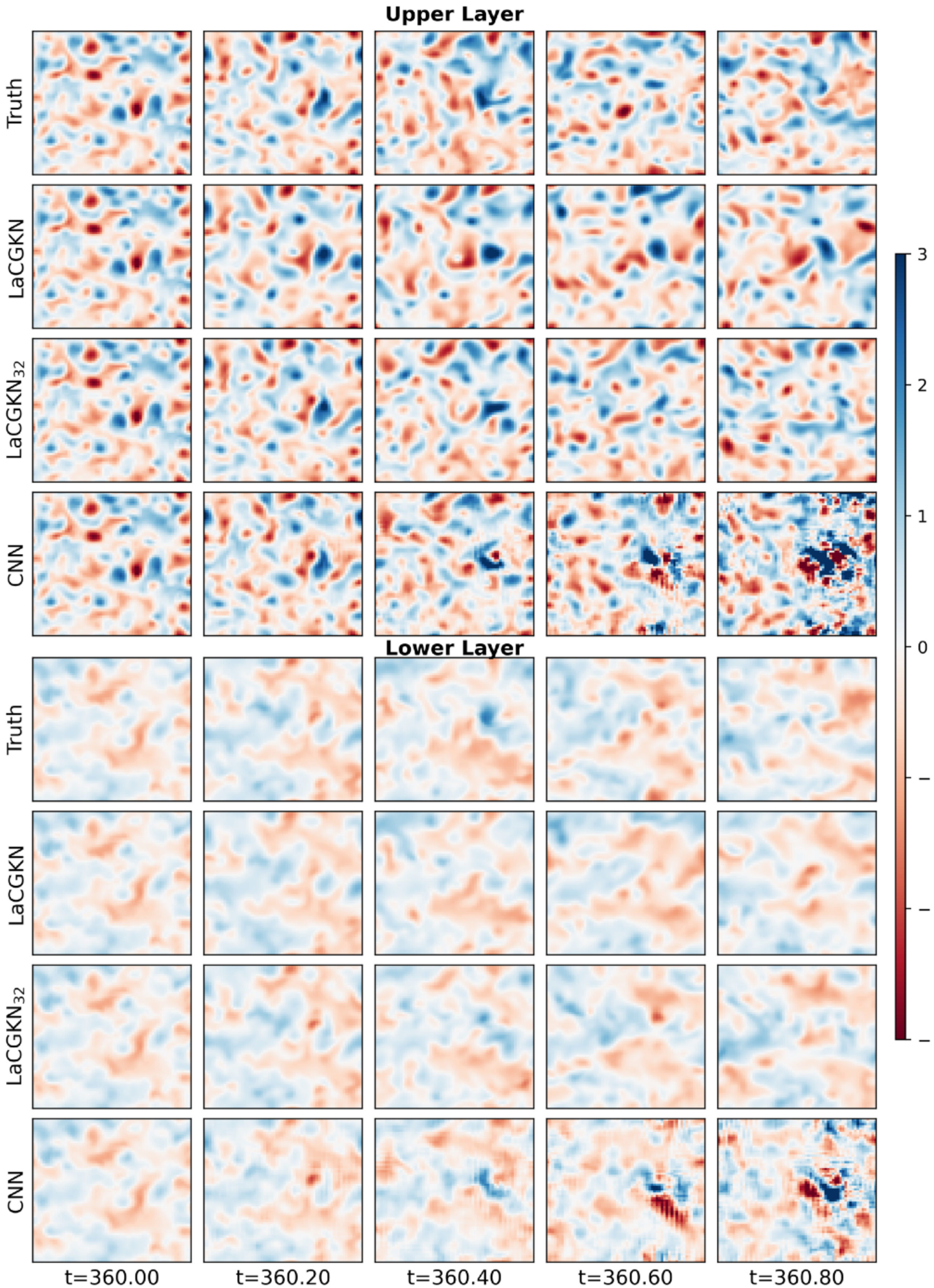




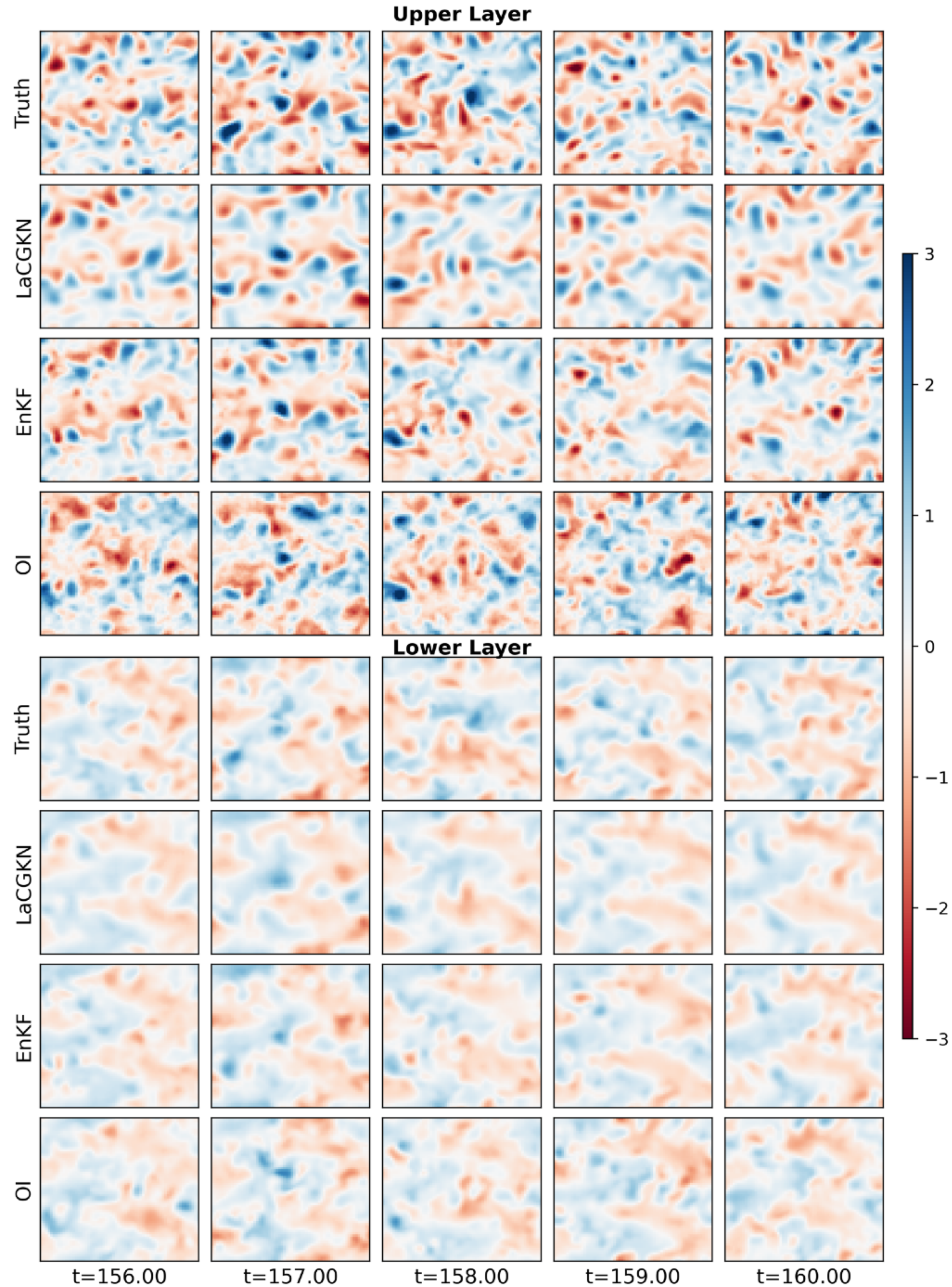
Accuracy

Efficiency

Two-layer QG flow with surface tracer observations



State forecast (multi-step)



Data assimilation

Summary

- **Lagrangian DA is a practically important while fundamentally challenging** task due to the nonlinear coupling between tracer observations and the underlying flow, making analytic posterior estimation computationally intractable for high-dimensional systems.

- The recently proposed discrete-time CGKN is a **unified deep learning framework** for **efficient state forecast** and **DA**.

- Existing applications of CGKN limit to *direct* (though partial) observations of the hidden system. **LaCGKN** applies to ***indirect* partial observations** that are **nonlinearly coupled** with hidden states.

- Several innovative design to address the nonlinear tracer–flow coupling and high dimensionality, including (1) **tracer homogenization,** (2) **Fourier positional encoding,** (3) **low-rank SVD-inspired parameterization**.

- An application to the two-layer QG flow with surface tracer observations demonstrates that LaCGKN **performs efficient and accurate Lagrangian DA** and prediction **without reliance on ensembles or the physical model**.

Arxiv preprint: "A Lagrangian Conditional Gaussian Koopman Network for Data Assimilation and Prediction"

Chuanqi will present CGKN on Wednesday, 2:30-2:55 (MS153)
Room: Lakeshore B - Main Level
"Modeling Partially Observed Nonlinear Dynamical Systems Via Conditional Gaussian Koopman Network"