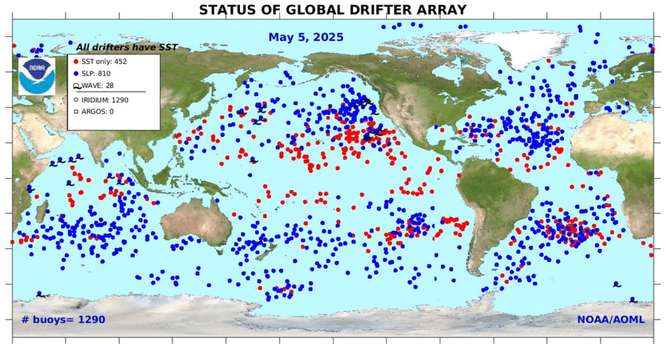
LaCGKN: A Deep Learning Digital Twin for Lagrangian Data Assimilation and Prediction

Zhongrui Wang, Chuanqi Chen, Jin-long Wu , Nan Chen

Department of Mathematics & Department of Mechanical Engineering, University of Wisconsin-Madison

Motivation



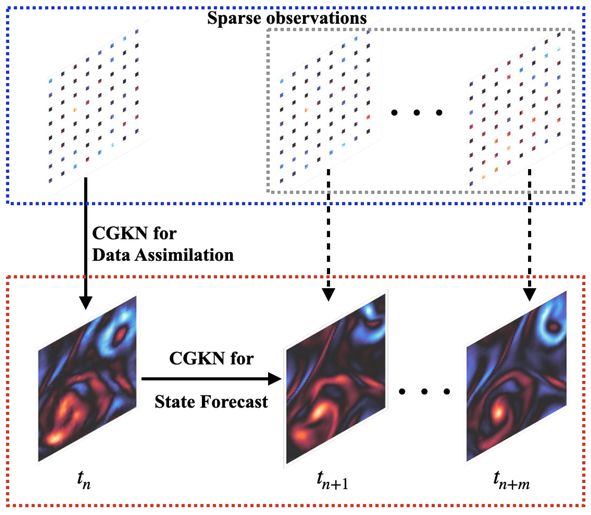
• Lagrangian data assimilation uses **Lagrangian** observations to recover the underlying **Eulerian** flow, e.g., inferring ocean states with surface observations.

• Lagrangian DA is **fundamentally challenging** because of the **intrinsic nonlinear coupling** between tracer trajectories and the underlying flow.

• Traditional deductive (model-based) DA methods often rely on strong assumptions or are computational expensive. **Machine learning** holds great potential to address the inherent nonlinearity in Lagrangian DA and improve **accuracy and efficiency**.

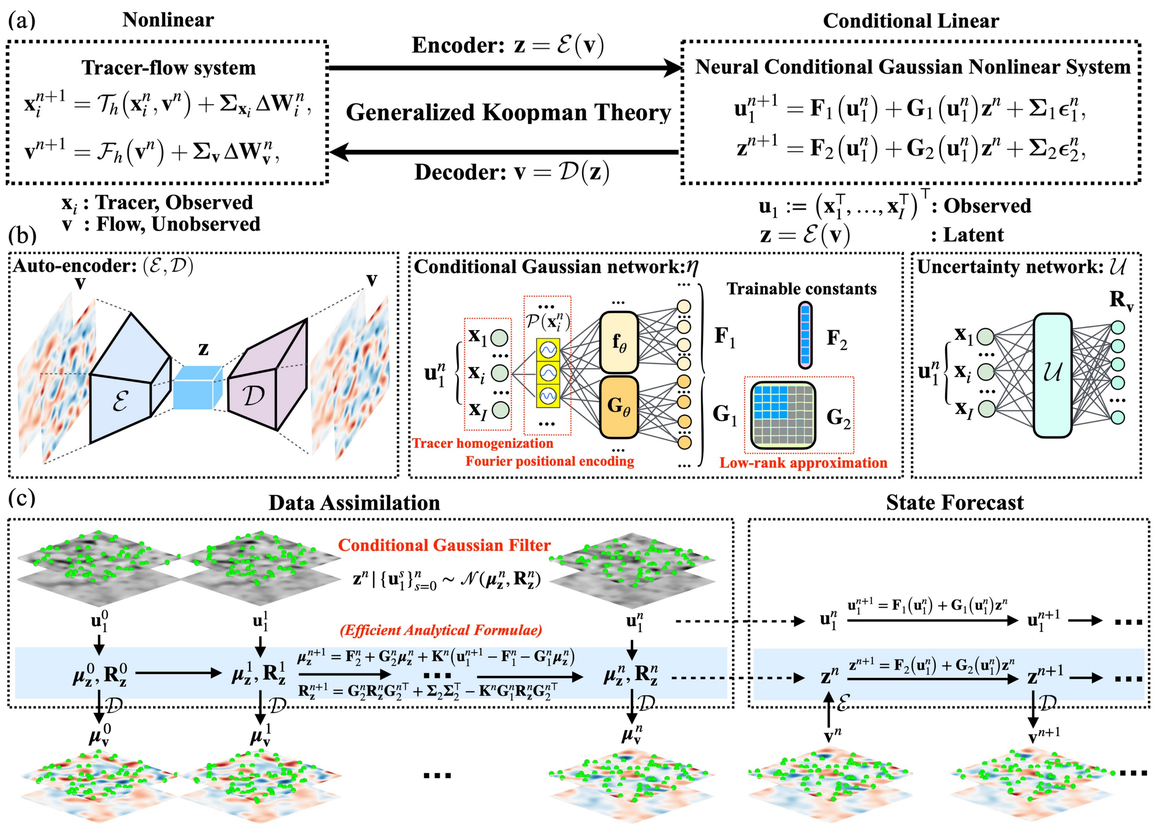
CGKN

Conditional Gaussian Koopman Network (CGKN):
A unified framework of SciML and DA, to learn surrogate models that performs efficient prediction and DA for nonlinear partially observed dynamical systems (Chen et al., 2025a,b) .



Challenge of CGKN for Lagrangian DA: existing CGKN relies on *direct* Eulerian observations, Lagrangian DA uses ***indirect*** Lagrangian observations, which introduces additional constraints on the latent representation*.

LaCGKN



Lagrangian conditional Gaussian Koopman network (LaCGKN) is a structure-preserving and data-driven framework for joint data assimilation and prediction from Lagrangian observations.

(a) The **nonlinear tracer-flow system** is mapped to a **neural conditional Gaussian nonlinear system (Neural CGNS)** by encoding the *unobserved flow* and mapped back by decoding the l*atent variables*.

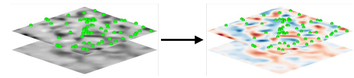
(b) LaCGKN consists of an **auto-encoder** and **conditional Gaussian network (CGN)**. The CGN outputs the coefficients $\mathbf{F}_1, \mathbf{G}_1, \mathbf{F}_2, \mathbf{G}_2$ of the Neural CGNS. The parameterization of $\mathbf{F}_1$ and $\mathbf{G}_1$ incorporates tracer homogenization and Fourier positional encoding, while $\mathbf{G}_2$ is represented by an SVD-inspired low-rank approximation. An **uncertainty network** is used to estimate the posterior standard deviation of the flow.

(c) LaCGKN performs **efficient data assimilation and prediction in latent space** using the CG filter, which admits **analytic posterior updates**.

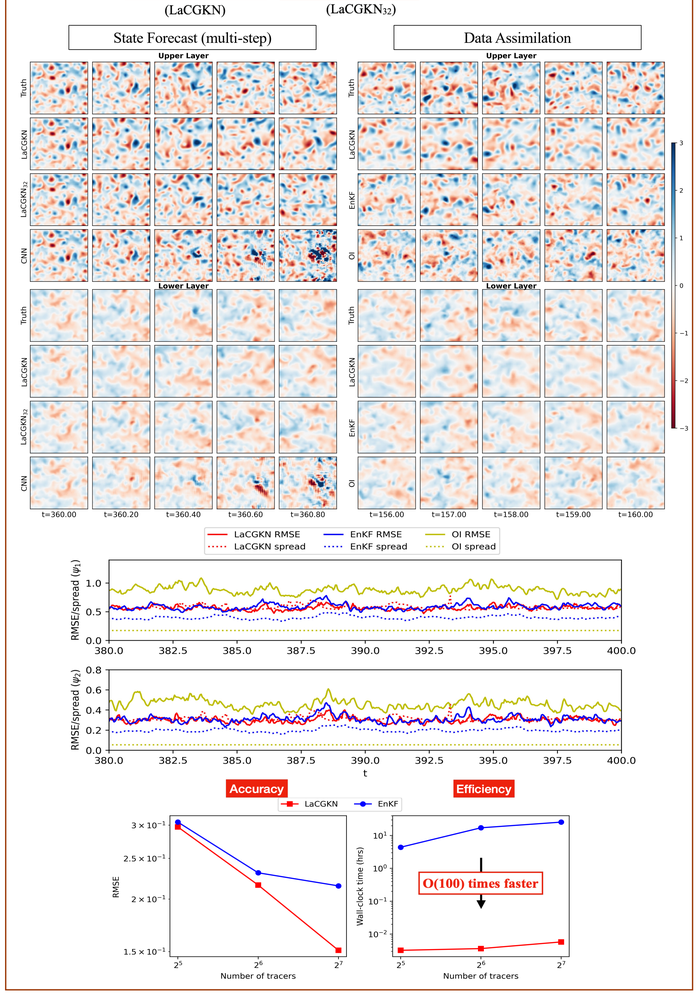
(d) The loss of LaCGKN is a weighted sum of: the **auto-encoder reconstruction loss** for Eulerian flow, the **forecast loss for physical variables**, the **forecast loss for latent variables**, and the **data assimilation loss**.

Numerical Tests

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



- Observation: positions of **64** tracers; recover unobserved states: $\mathbf{64 \times 64 \times 2}$ (two-layer flow)
- Encoder: $\underbrace{\mathbb{R}^{64\times64\times2} \mapsto \mathbb{R}^{16\times16\times2}}_{\text{(LaCGKN)}}$ / $\underbrace{\mathbb{R}^{64\times64\times2} \mapsto \mathbb{R}^{32\times32\times2}}_{\text{(LaCGKN}_{32}\text{)}}$, 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

* The latent representation of LaCGKN must simultaneously encode the **flow dynamics** and **mediate nonlinear tracer-flow interactions**. The latter is crucial to data assimilation, as it captures the information propagation from observed states to unobserved states. The three key architectural innovations address these challenges: (i) tracer homogenization enforces permutation equivariance and enables generalization across varying numbers of tracers; (ii) Fourier-based positional encoding captures rich spatial dependence and reconstructs local flow features at moving tracer locations; and (iii) an SVD-inspired low-rank parameterization of the latent transition operator reduces parameter complexity while preserving expressive capacity.

Zhongrui Wang , Chuanqi Chen, Jin-Long Wu, Nan Chen. A Lagrangian Conditional Gaussian Koopman Network for Data Assimilation and Prediction. Arxiv, 2026