arXiv:2603.14115v1 [cs.CE] 14 Mar 2026

用于数据同化与预测的拉格朗日条件高斯 Koopman 网络

Zhongrui Wang[1], Chuanqi Chen[2],Jin-Long Wu[2],和 Nan Chen[1],*

[1]威斯康星大学麦迪逊分校数学系，美国威斯康星州麦迪逊市 53706[2]威斯康星大学麦迪逊分校机械工程系，美国威斯康星州麦迪逊市 53706

*通讯作者：Nan Chen，chennan@math.wisc.edu

2026年3月17日

摘要

拉格朗日数据同化旨在从运动示踪剂的稀疏间接观测中恢复隐藏的欧拉流场。由于示踪粒子轨迹与底层流场之间存在非线性耦合，这一问题在本质上极具挑战性，使得针对现实高维系统的后验推断在计算上难以实现。在本研究中，我们开发了拉格朗日条件高斯 Koopman 网络（LaCGKN），这是一种结构保持的数据驱动框架，可基于拉格朗日观测联合进行数据同化与预测。LaCGKN 将欧拉流体动力学嵌入低维潜在空间，该空间由具有条件高斯结构的非线性随机系统支配，从而无需进行集合预报即可实现解析后验更新。与依赖直接欧拉观测的现有条件高斯 Koopman 表述不同，拉格朗日框架对潜在表示提出了额外约束：它必须同时编码流动动力学，并协调非线性的示踪物-流场相互作用。为应对这些挑战，LaCGKN 纳入了三个关键架构组件：（i）示踪粒子齐次化，以强制实现置换等变性，并支持对不同示踪粒子数量的泛化；（ii）基于傅里叶的位置编码，以捕捉空间依赖性，并重建运动示踪粒子位置处的局部流场特征；（iii）受奇异值分解启发的潜在转移算子低秩参数化，在降低参数复杂度的同时保持表达能力。将该方法应用于具有表面示踪剂观测的双层准地转流，结果表明，LaCGKN 能够准确、高效地进行拉格朗日数据同化与预测，且无需依赖集合方法或控制物理模型。这些结果确立了 LaCGKN 的地位：它是传统基于模型的方法与纯黑箱数据驱动方法的统一且计算上可行的替代方案。

1 引言

拉格朗日数据由运动示踪粒子的轨迹构成，例如随流场中粒子运动的漂流浮标和漂流浮标。通过记录这些示踪粒子随时间变化的位置，拉格朗日观测为了解底层流场的输运特性提供了直接窗口，广泛用于刻画大尺度环流模式、相干结构和混合过程（Businger 等，1996；Gould 等，2004；Lumpkin 等，2017；Centurioni 等，2017）。虽然通过直接分析拉格朗日轨迹可以提取有价值的信息，但一种更系统且有效的方法是通过拉格朗日数据同化，将这些观测与动力学模型或统计模型相结合。在拉格朗日框架下，数据同化的目标是利用示踪粒子轨迹观测来估计底层流场状态（Mariano 等，2002年；Ide 等，2002年）。随着全球拉格朗日观测系统迅速扩展，拉格朗日数据同化在地球物理研究和业务预报中的重要性不断提升（Legler 等，

2015年；McPhaden 等，2023）。然而，观测过程的性质带来了一个根本性挑战：拉格朗日测量是通过在示踪粒子随时间变化的位置处评估局部流速获得的，因此会形成高度非线性的观测算子（Ide 等，2002年；Kuznetsov 等，2003年）。当这一内在非线性与地球物理流强非线性且往往湍流的动力学特征相结合时，拉格朗日数据同化便尤具挑战性。因此，对于现实中的高维系统，后验分布的解析表达式通常难以求得。

早期的拉格朗日数据同化方法通过在欧拉流模型中加入显式示踪剂动力学，并采用基于卡尔曼滤波的方法来处理非线性。代表性方法包括使用扩展卡尔曼滤波器（EKF）同化示踪粒子轨迹（Ide 等人，2002年）、将有限差分拉格朗日速度视为具有固定增益的替代观测的最优插值方案（Mol- card 等人，2003年），以及将局地集合变换卡尔曼滤波器（LETKF）应用于漂流浮标（Sun 和 Penny，2019年；Chen 等人，2022年）。尽管这些方法在业务应用中仍颇具吸引力，但它们依赖高斯假设和线性校正步骤，在强非线性情形下有时会限制精度。为更好地表征非线性和非高斯性，研究者探索了粒子滤波器（PF；van Leeuwen 等人，2019年）和平滑器马尔可夫链蒙特卡罗（MCMC）方法（Apte 等人，2008），但其计算成本给高维系统带来了严峻挑战。混合策略则试图在近似精度与计算成本之间取得平衡，例如，将粒子滤波用于非线性示踪剂动力学，将集合卡尔曼滤波用于近似高斯的流场估计（Slivinski 等人，2015年）；或者利用特殊结构来处理流场非线性（Wang 等人，2025）。尽管取得了这些进展，如何设计兼具高精度与高计算效率的拉格朗日数据同化方法，仍是一项尚待解决的挑战。

传统数据同化方法主要是演绎式的（基于模型），通过贝叶斯定理从第一性原理推导而来。后验推断通过变分优化显式进行，例如三维/四维变分方法（Lorenc，1986，2003），或基于蒙特卡罗的滤波方法，例如集合卡尔曼滤波器（EnKF；Evensen，1994）。为了使高维系统的计算保持可行，这些方法通常会施加严格的结构性假设，尤其是假设误差统计服从高斯分布，且动力学或观测算子是线性化的。受机器学习进展推动，归纳式（数据驱动）数据同化方法日益受到关注（Cheng 等，2023；Bach 等，2025）。科学机器学习（SciML）的发展为复杂动力系统提供了灵活的代理模型，包括循环神经网络（Schuster 和 Paliwal，1997年；Hochreiter 和 Schmidhuber，1997年；Gauthier 等，2021）、神经常微分方程（Chen 等，2019年）、物理信息神经网络（Raissi 等，2019年）、神经算子（Lu 等，2021；Li 等，2021；Chen 和 Wu，2025），以及用于随机系统的生成模型（Du 等，2024；Dong 等，2025；Stamatelopoulos 和 Sapsis，2025）。与传统滤波器和平滑器结合时，SciML 可通过提供高效的预报模型（Penny 等，2022）、校正模型误差（Farchi 等，2023）或学习分析映射的组成部分（Bach 等，2025），提升数据同化能力。近期，人们提出了混合方法，以保留解析上易于处理的非线性结构，例如条件高斯神经随机微分方程，从而兼顾表达能力与高效的后验推断（Chen 等，2024）。此外，完全数据驱动方法则尝试直接学习分析映射或后验映射（Bocquet 等，2024）。其他示例包括集合输运（Spantini 等，2022）、生成式数据同化（Qu 等，2024；Li 等，2025；Martin 等，2025；Yang 等，2025），以及联合学习预报映射与分析映射的方法（Fablet 等，2021；Boudier 等，2023）。总的来说，这些进展涵盖了从演绎式到归纳式的一系列方法，展现出人们正持续探索物理可解释性、统计严谨性与计算效率之间的权衡。

然而，大多数数据驱动数据同化方法都是针对固定空间网格上的欧拉观测而开发的。相比之下，机器学习在拉格朗日数据同化中的应用仍相对有限，近期只有少数研究将神经算子与生成模型（Asefi 等，2025）或多模态对比学习（Baptista 等，2025）相结合。这种差异值得关注，因为机器学习特别适合处理拉格朗日观测中固有的强非线性，因此在提高准确性和计算效率方面具有巨大潜力。与此相关的一项进展是条件高斯 Koopman 网络（CGKN；Chen 等，2025a，b），它引入了一种嵌入条件高斯非线性随机系统中的数据驱动动力学模型，该系统能够进行解析后验更新，并避免集合方法通常需要的各种调参。通过将有原则的滤波公式直接纳入

模型架构作为归纳偏置，CGKN 为 SciML 与数据同化提供了一个结构化接口。然而，现有的 CGKN 表述仅适用于具有直接、部分欧拉观测的系统，并未解决拉格朗日数据带来的根本性挑战。

本文开发了一种拉格朗日条件高斯 Koopman 网络（LaCGKN），用于根据示踪粒子轨迹估计潜在流场。除了拉格朗日数据同化固有的非线性外，这一设定还带来了与采用欧拉观测的 CGKN 所面临问题在性质上截然不同的挑战。具体而言，潜在表示必须同时支持对欧拉流体动力学进行高效、准确的编码，并刻画支配示踪粒子运动的非线性但具有结构的观测机制。后者在拉格朗日数据同化中发挥着核心作用，因为它决定了运动示踪粒子携带的信息如何传递至不可观测的流场，因而对潜在嵌入施加了欧拉框架中不存在的额外约束。为应对这些挑战，我们构建了一种纯数据驱动的代理模型，将欧拉流嵌入低维潜在空间，同时将示踪剂动力学建模为由潜在状态驱动、关于示踪粒子位置的非线性函数。这一设计保留了非线性但具有条件高斯结构的特性，使得无需对物理模型进行集合预报，即可在潜在空间中实现解析滤波。由此得到的 LaCGKN 包含三个关键架构组件：(i) 示踪粒子齐次化，确保示踪粒子置换等变性，并实现对不同示踪粒子数量的泛化；(ii) 基于傅里叶的位置编码，用于捕捉丰富的空间依赖性，并便于重建运动示踪粒子位置处的局部流场特征；(iii) 受奇异值分解启发的低秩参数化，在控制计算成本的同时，于高潜在维度下保持表达能力。这些设计选择相结合，实现了准确、高效的拉格朗日数据同化与预测，为传统的基于模型的方法和纯黑箱数据驱动方法提供了一种统一的替代方案。

本文其余部分安排如下。第 2 节介绍拉格朗日数据同化和离散时间 CGKN 的形式化描述，随后对用于数据同化和预测的 LaCGKN 进行形式化定义，并重点说明其关键架构设计选择。第 3 节通过表层示踪粒子观测下的双层准地转流应用，展示所提出的方法。在预测方面，将 LaCGKN 与深度学习基线进行比较；该基线使用深度神经网络（DNN）预测示踪剂，并使用卷积神经网络（CNN）进行流场预测。此外，还将其与简单的持续性基线进行比较。在数据同化方面，将 LaCGKN 与集合卡尔曼滤波、最优插值和朴素气候态基线进行比较。第 4 节通过讨论总结全文。

2 方法

2.1 拉格朗日数据同化

拉格朗日数据同化旨在同化由通常以欧拉坐标描述的流场驱动的拉格朗日观测。为阐明这一思想，我们聚焦于一种典型设定：由欧拉流驱动的被动示踪粒子位置的拉格朗日观测，其描述如下：

$$\frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = \mathbf{v}(\mathbf{x}_i, t) + \mathbf{\Sigma}_{\mathbf{x}_i}\dot{\mathbf{W}}_i, \quad i = 1, \ldots, I \tag{1a}$$

$$\frac{\partial \mathbf{v}}{\partial t} = \mathcal{F}[\mathbf{v}] + \mathbf{\Sigma}_{\mathbf{v}}\dot{\mathbf{W}}_{\mathbf{v}}, \tag{1b}$$

其中，$\mathbf{x}_i(t) \in \Omega \subset \mathbb{R}^d$ 是第 i 个示踪剂的观测位置，$\mathbf{v}(\mathbf{x},t) \in \mathbb{R}^d$ 是流速场，且 $(\mathbf{x},t) \in \Omega \times \in (0,T\ ]$。流算子 $\mathcal{F}$ 通常是非线性的。该系统是随机的，包含高斯白噪声 $\dot{W}$ 和噪声强度矩阵 $\mathbf{\Sigma}$，用于表示未解析细节所带来的不确定性。通过修改描述拉格朗日观测的 (1a)，使其依赖于示踪剂位置，可将这一表述推广到任意拉格朗日观测。这一典型设定已揭示出拉格朗日数据同化的一项关键挑战：拉格朗日观测过程是非线性的，因为复合函数 $\mathbf{v}(\mathbf{x}_i(t),t)$ 通常关于 $\mathbf{x}_i(t)$ 是非线性的。

实际应用中，考虑（1）的离散形式如下：

$$\mathbf{x}_i^{n+1} = \mathcal{T}_h\left(\mathbf{x}_i^n, \mathbf{v}^n\right) + \boldsymbol{\Sigma}_{\mathbf{x}_i} \Delta \mathbf{W}_i^n, \quad i = 1, \ldots, I, \tag{2a}$$

$$\mathbf{v}^{n+1} = \mathcal{F}_h\left(\mathbf{v}^n\right) + \boldsymbol{\Sigma}_{\mathbf{v}} \Delta \mathbf{W}_{\mathbf{v}}^n, \tag{2b}$$

其中，$\mathrm{x}_i^n := \mathrm{x}_i(\mathrm{t}_n) \in \Omega$ 和 $\mathrm{v}^n := \mathrm{v}(\cdot, \mathrm{t}_n)|_{\Omega_{\mathbf{x}}^{(M)}} \in \mathbb{R}^{dM}$ 分别是在离散时刻 t_n 取值的相应状态变量，时间间隔为 $\mathrm{t}_{n+1} - \mathrm{t}_n = \Delta \mathrm{t}$，离散空间域 $\Omega_{\mathbf{x}}^{(M)}$ 包含 M 个点。$\mathcal{T}_h$ 和 $\mathcal{F}_h$ 分别是示踪剂和流场从 t_n 到 t_{n+1} 的解映射。示踪剂算子 $\mathcal{T}_h$ 通常涉及插值算子或积分算子（例如傅里叶变换），用于计算局部示踪剂位置 x_i^n 处的 $\mathrm{v}(\mathrm{x}_i^n, \mathrm{t}_n)$。

离散设定中定义的拉格朗日滤波问题（2）**旨在根据过去的示踪剂观测数据** $\mathbf{x}_i^n\}_{,n\ i=1,s=0}^{I}$，求取未观测流场状态 $\mathbf{v}^n$ 的后验分布，即

$$p(\mathbf{v}^n | \{\mathbf{x}_i^s\}_{i=1,s=0}^{I,n}). \tag{3}$$

由于解映射 $\mathcal{T}_h$ 和 $\mathcal{F}_h$ 通常是非线性的，因此在实践中，高维系统的后验分布解析解几乎难以求得，这使得拉格朗日数据同化成为一项具有挑战性的问题。

2.2 离散时间 CGKN

条件高斯 Koopman 网络（CGKN；Chen 等，2025a，b）是一种统一的深度学习框架，可实现高效的数据同化和状态预报。它将原始动力系统转换为具有条件高斯结构的系统，其中嵌入的未观测变量的后验分布可通过解析公式求解。转换后的动力系统可用作代理模型，从而实现高效的数据同化和状态预报。随后，将同化和预报后的状态映射回原始状态空间，以获得最终结果。

考虑如下形式的一般部分观测离散时间动力系统：

$$\mathbf{u}_1^{n+1} = \mathcal{G}_1\left(\mathbf{u}_1^n, \mathbf{u}_2^n\right), \tag{4a}$$

$$\mathbf{u}_2^{n+1} = \mathcal{G}_2\left(\mathbf{u}_1^n, \mathbf{u}_2^n\right), \tag{4b}$$

其中 $\mathbf{u}_1 \in \mathbb{R}^{d_1}$ 和 $\mathbf{u}_2 \in \mathbb{R}^{d_2}$ 分别是表示观测状态和未观测状态的向量。状态转移算子 $\mathcal{G}_1$ 和 $\mathcal{G}_2$ 将解从 t_n 映射到 t_{n+1}，时间间隔为 $\Delta t = t_{n+1} - t_n$。

对于 he d离散动力系统（4），CGKN 将其转换为如下形式的代理模型 t :

$$\mathbf{u}_1^{n+1} = \mathbf{F}_1\left(\mathbf{u}_1^n\right) + \mathbf{G}_1\left(\mathbf{u}_1^n\right)\mathbf{z}^n + \boldsymbol{\Sigma}_1 \boldsymbol{\epsilon}_1^n, \tag{5a}$$

$$\mathbf{z}^{n+1} = \mathbf{F}_2\left(\mathbf{u}_1^n\right) + \mathbf{G}_2\left(\mathbf{u}_1^n\right)\mathbf{z}^n + \boldsymbol{\Sigma}_2 \boldsymbol{\epsilon}_2^n, \tag{5b}$$

其中 $\mathrm{z} \in \mathbb{R}^{d_\mathrm{z}}$ 是原始未观测状态 u_2 的潜在嵌入。潜在状态与原始状态之间的转换通过 $\mathrm{z} = \mathcal{E}(\mathrm{u}_2)$ 和 $\mathrm{u}_2 = \mathcal{D}(\mathrm{z})$ 实现，其中 $\mathcal{E}: \mathbb{R}^{d_2} \mapsto \mathbb{R}^{d_\mathrm{z}}$ 是编码器，$\mathcal{D}: \mathbb{R}^{d_\mathrm{z}} \mapsto \mathbb{R}^{d_2}$ 是解码器。系数 F_1, G_1, F_2, 和 G_2 是观测状态 u_1 的函数。项 $\boldsymbol{\Sigma}_1\boldsymbol{\epsilon}_1$ 和 $\boldsymbol{\Sigma}_2\boldsymbol{\epsilon}_2$ 是乘以噪声强度矩阵的高斯白噪声。实际应用中，编码器 $\mathcal{E}$、解码器 $\mathcal{D}$ 以及系数函数 F_1, G_1, F_2, G_2 均由神经网络参数化。噪声强度矩阵 $\boldsymbol{\Sigma}_1$ 通过使用不含数据同化损失的已训练 CGKN，根据观测状态 u_1 的单步残差误差进行估计。噪声强度矩阵 $\boldsymbol{\Sigma}_2$ 可手动设置，也可作为训练参数。

代理模型 (5) 具有显著的条件高斯结构，可实现高效的数据同化。具体而言，由于 $\mathbf{F}_1$, $\mathbf{G}_1$, $\mathbf{F}_2$, 和 $\mathbf{G}_2$ 仅是 $\mathbf{u}_1$, **的函数，因此在给定** u_1 的观测轨迹时，$\mathbf{z}$ 的条件分布 $(\mathbf{p}^n | \{\mathbf{z}_1^s\}_{s=0}^n \mathrm{u} = \mathcal{N})$ $\boldsymbol{\mu}_\mathbf{z}^n, \mathbf{R}_\mathbf{z}^n)$ 服从高斯分布，其均值和协方差可通过解析公式求解：

$$\begin{aligned} \boldsymbol{\mu}_\mathbf{z}^{n+1} &= \mathbf{F}_2^n + \mathbf{G}_2^n \boldsymbol{\mu}_\mathbf{z}^n + \mathbf{K}^n\left(\mathbf{u}_1^{n+1} - \mathbf{F}_1^n - \mathbf{G}_1^n \boldsymbol{\mu}_\mathbf{z}^n\right) \\ \mathbf{R}_\mathbf{z}^{n+1} &= \mathbf{G}_2^n \mathbf{R}_\mathbf{z}^n \mathbf{G}_2^{n\top} + \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_2^\top - \mathbf{K}^n \mathbf{G}_1^n \mathbf{R}_\mathbf{z}^n \mathbf{G}_2^{n\top} \end{aligned} \tag{6}$$

其中

$$\mathbf{K}^n = \mathbf{G}_2^n \mathbf{R}_{\mathbf{z}}^n \mathbf{G}_1^{n\top} \left(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_1^\top + \mathbf{G}_1^n \mathbf{R}_{\mathbf{z}}^n \mathbf{G}_1^{n\top} \right)^{-1} \tag{7}$$

以及$\mathbf{F}_i^n := \mathbf{F}_i(\mathbf{u}_1^n)$，$\mathbf{G}_i^n := \mathbf{G}_i(\mathbf{u}_1^n)$，其中$i = 1, 2$。闭合解析公式能够提供精确且准确的数据同化解，避免了许多集合方法中的经验性调参。潜在状态的后验均值通过解码器 $\boldsymbol{\mu}_2^{n+1} = \mathcal{D}(\boldsymbol{\mu}_{\mathbf{z}}^{n+1})$ 转换回原始状态。未观测状态的后验协方差$\mathbf{R}_2^n$ 可通过残差分析计算；残差分析是 CGKN 的后处理步骤，用于量化后验均值的不确定性。

2.3 LaCGKN

LaCGKN 代理模型。在离散示踪粒子流动系统（2）中，观测拉格朗日示踪粒子位置，并恢复欧拉流场。这自然地引出了离散时间 CGKN 中观测变量和未观测变量的定义：

$$\mathbf{u}_1^n := \left(\mathbf{x}_1^{n\top}, \ldots, \mathbf{x}_I^{n\top} \right)^\top \in \mathbb{R}^{dI}, \qquad \mathbf{u}_2^n := \left(\mathbf{v}_1^{n\top}, \ldots, \mathbf{v}_M^{n\top} \right)^\top \in \mathbb{R}^{dM}, \tag{8}$$

其中$\mathbf{u}_1$ 汇集了所有拉格朗日示踪粒子位置，而$\mathbf{u}_2$ 包含离散空间域 $\Omega_{\mathbf{x}}^{(M)}$上的欧拉流场状态。由此得到的 LaCGKN 代理模型形式如下：

$$\mathbf{u}_1^{n+1} = \mathbf{F}_1(\mathbf{u}_1^n) + \mathbf{G}_1(\mathbf{u}_1^n)\mathbf{z}^n + \boldsymbol{\Sigma}_1 \boldsymbol{\epsilon}_1^n, \tag{9a}$$

$$\mathbf{z}^{n+1} = \mathbf{F}_2 + \mathbf{G}_2 \mathbf{z}^n + \boldsymbol{\Sigma}_2 \boldsymbol{\epsilon}_2^n, \tag{9b}$$

这与一般的离散时间 CGKN 表述（5）一致，不同之处在于，$\mathbf{F}_2$ 和$\mathbf{G}_2$ 现在是可训练常数，而不是关于$\mathbf{u}_1^n$的函数。这一简化利用了被动示踪剂不会影响底层流场这一物理假设。欧拉流$\mathbf{u}_2^n$ 通过编码器映射到低维潜在嵌入$\mathbf{z}^n$，再通过解码器映射回物理状态 $\mathcal{D}$。这种流场嵌入与基于深度学习的现代应用 Koop man 理论相一致（Lusch 等人，2018），但有一个关键区别：嵌入后的流场$\mathbf{z}$ 通过（9a）与非线性示踪剂动力学相耦合。线性潜在流动力学（9b）与变换后的示踪剂动力学（9a）共同构成 LaCGKN 代理模型。图1展示了 LaCGKN 的概览。

用于数据同化和状态预报的 LaCGKN。LaCGKN 在潜在空间中高效地执行数据同化和状态预测。在数据同化过程中，利用条件高斯滤波器（6）对潜在状态的后验均值和协方差进行解析更新，从而避免集合传播和经验调参。随后，通过解码器对潜在状态的后验均值进行解码，得到物理流场的后验均值$\mathcal{D}$。由于非线性解码，物理流场的后验不确定性无法用封闭形式表示，但可以将辅助不确定性网络 $\mathcal{U}$ 用作后处理步骤，以估计物理空间中的后验协方差。网络 $\mathcal{U}$ 以观测状态$\mathbf{u}_1$ 为输入，并预测残差 $\|\mathbf{u}_2^\star - \boldsymbol{\mu}_2\|$，其中 $\mathbf{u}_2^\star$ 表示真实值，而 $\boldsymbol{\mu}_2 = \mathcal{D}(\boldsymbol{\mu}_{\mathbf{z}})$ 是解码得到的后验均值。在高斯假设下，该残差可提供物理流场后验标准差的最大似然估计。在状态预测中，使用 LaCGKN 代理模型（9）推进示踪剂状态和潜在流场，再对预测得到的潜在流场进行解码以恢复物理流场。由于滤波和预报均在通常维度较低的潜在空间中进行，LaCGKN 对高维湍流系统而言效率很高。

值得强调的是，拉格朗日观测（即示踪粒子位置）是对隐藏系统状态（即流场）的间接观测，并且通过非线性动力学过程与隐藏状态相耦合。这不同于 CGKN 以往的应用，在以往的应用中，观测变量是对同一底层系统的直接但不完整的观测。因此，对于 CGKN 而言，拉格朗日数据同化是一个更具挑战性但也更具实际意义的问题：不仅需要像标准 Koopman理论中那样，用潜在线性动力学对非线性流体动力学进行良好近似，还需要用5a对非线性示踪粒子动力学进行良好近似。后者对于数据同化至关重要，因为它捕捉了从观测状态到未观测状态的信息传播。这对潜在嵌入 $\mathbf{z}$提出了额外的要求和正则化。此外，

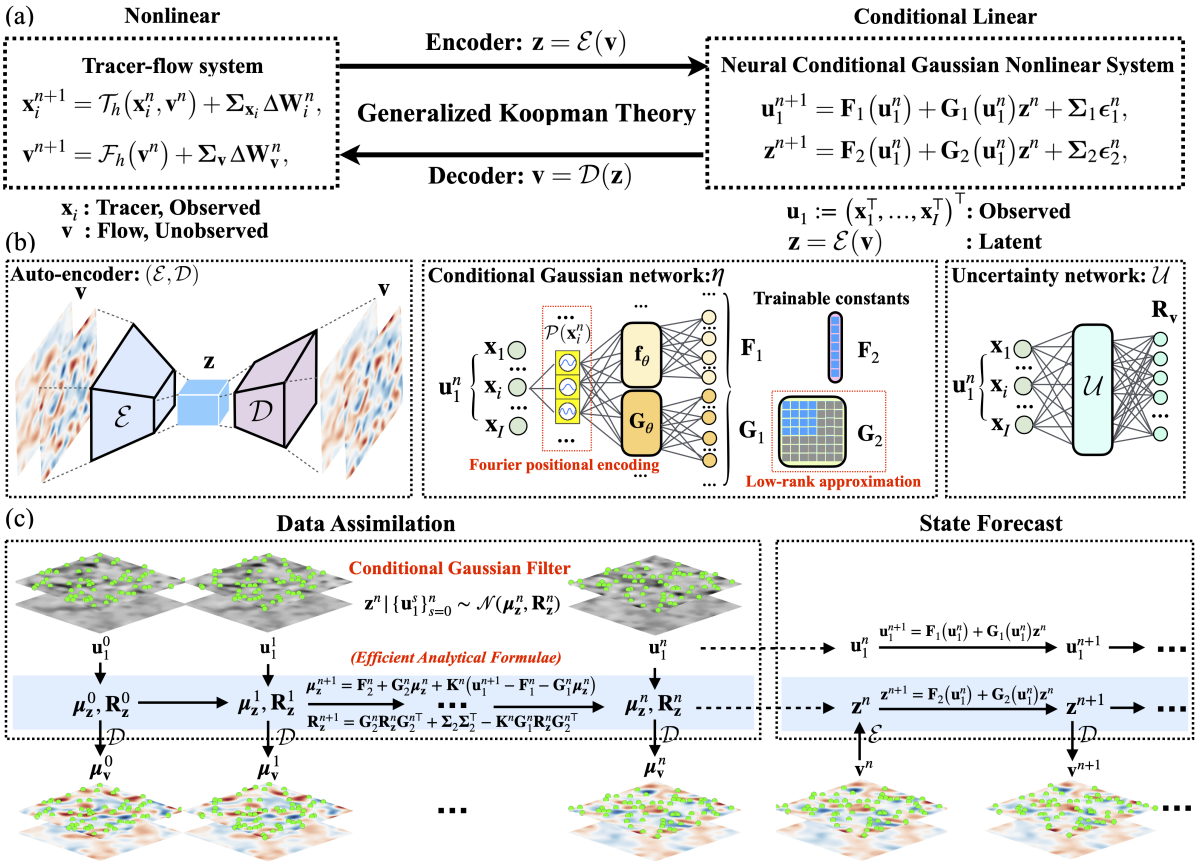


图 1：拉格朗日条件高斯 Koopman 网络（LaCGKN）概览。(a) 非线性示踪剂流动系统通过对未观测流场进行编码，映射为神经条件高斯非线性系统（Neural CGNS），再通过解码潜变量映射回原系统。(b) LaCGKN 由一个自编码器（$\mathcal{E}$、$\mathcal{D}$）和条件高斯网络（CGN） η组成。CGN 输出 Neural CGNS 的系数 $\mathbf{F}_1$,$\mathbf{G}_1$,$\mathbf{F}_2$,$\mathbf{G}_2$。对 $\mathbf{F}_1$ 和 $\mathbf{G}_1$ 的参数化纳入了示踪粒子齐次化和傅里叶位置编码，而 $\mathbf{G}_2$ 则采用受奇异值分解启发的低秩近似表示。辅助不确定性网络 $\mathcal{U}$ 用于估计流场的后验标准差。(c) LaCGKN 利用条件高斯滤波器在潜在空间中高效地进行数据同化和预测。

当建模系统具有高维度和强湍流特性时（这在地球物理应用中十分常见），可能需要相对较大的潜在维度，从而增加 CGKN 的计算成本。在本工作中，我们通过引入以下创新设计来应对这些挑战。

（a）示踪粒子齐次化。拉格朗日示踪粒子通常可以被假设为具有齐次性且彼此可交换。这一点已在示踪粒子方程（1a）中有所体现，其中不同示踪粒子遵循相同的动力学定律。因此，我们通过对每个示踪粒子位置分别应用相同的神经网络来构造 LaCGKN 观测算子$\mathbf{F}_1$ 和$\mathbf{G}_1$，即：

$$\mathbf{F}_1(\mathbf{u}_1^n) := \left(\mathbf{f}_\theta(\mathbf{x}_1^n)^\top, \ldots, \mathbf{f}_\theta(\mathbf{x}_I^n)^\top\right)^\top, \tag{10a}$$

$$\mathbf{G}_1(\mathbf{u}_1^n) := \left(\mathbf{G}_\theta(\mathbf{x}_1^n)^\top, \ldots, \mathbf{G}_\theta(\mathbf{x}_I^n)^\top\right)^\top, \tag{10b}$$

其中，$\mathbf{f}_\theta$： $\mathbb{R}^d \to \mathbb{R}^d$ 和 $\mathbf{G}_\theta$： $\mathbb{R}^d \to \mathbb{R}^{d\times d_\mathbf{z}}$ 是在所有示踪剂之间共享参数的神经网络。随后，示踪粒子的更新形式为：

$$\mathbf{x}_i^{n+1} = \mathbf{f}_\theta(\mathbf{x}_i^n) + \mathbf{G}_\theta(\mathbf{x}_i^n)\mathbf{z}^n + \text{noise}, \qquad i = 1, \ldots, I. \tag{11}$$

这种设计确保代理动力学关于示踪剂具有置换等变性，并使同一个 LaCGKN 只需训练一次，便可在测试阶段应用于任意数量的示踪剂。在实现中，通过在训练期间随机子采样不同的示踪粒子子集，进一步促进了对示踪粒子置换的不变性；而预测和滤波始终使用同一组共享网络 $\mathbf{f}_\theta$ 和 $\mathbf{G}_\theta$。与使用接收所有示踪粒子位置 $\mathbf{u}_1^n$ 作为输入、以参数化 $\mathbf{F}_1$ 和 $\mathbf{G}_1$, 的大型神经网络相比，示踪粒子齐次化能够显著增强模型的正则化效果，简化模型开发，并提高计算效率。

(b) 基于傅里叶的位置编码。为了准确重构移动示踪剂位置处欧拉流场的局部值，$\mathbf{G}_\theta(\mathbf{x}_i^n)$ 必须学习对位置丰富且非线性的依赖关系。因此，我们对示踪剂位置采用基于傅里叶的位置编码，研究表明，这能显著提升神经网络表示高频空间变化的能力（Tancik 等人,2020年; Mildenhall 等人,2020年）。对于周期域上位置为$\mathbf{x}_i^n = (x_{i,1}^n, \ldots, x_{i,d}^n)$的示踪剂，可通过用一系列三角函数扩展$\mathbf{x}_i^n$来构造傅里叶位置特征：

$$\mathcal{P}(\mathbf{x}_i^n) = \left[x_{i,1}^n, \ldots, x_{i,d}^n, \left\{\sin(2^k\pi x_{i,j}^n), \cos(2^k\pi x_{i,j}^n)\right\}_{j=1,k=0}^{d,K-1}\right], \tag{12}$$

然后将其输入网络 f_θ 和 G_θ。这种编码还融入了示踪粒子方程中的物理启发式。回想一下，式（2a）中的解算子 $\mathcal{T}_h(\mathrm{x}_i^n, \mathrm{v}^n)$ 通常需要通过傅里叶变换，在示踪剂位置 $\mathrm{v}(\mathrm{x}_i^n, \mathrm{t}_n)$处计算欧拉速度场，其中不同频率处的傅里叶系数与相应的傅里叶基函数相乘，再进行线性求和。基于傅里叶的位置编码在此处发挥了类似傅里叶基函数的作用。它极大地增强了 $\mathrm{G}_1(\mathrm{u}_1^n)\mathrm{z}^n$ 逼近非线性复合函数 $\mathrm{v}(\mathrm{x}_i^n, \mathrm{t}_n)$的表达能力，并有助于发现紧凑的潜在嵌入 z。

（c）对 G 进行低秩近似$_2$ 以提升可扩展性。当嵌入高维复杂流动系统需要相对较高的潜在维度时，潜在转移矩阵 $\mathbf{G}_2 \in \mathbb{R}^{d_\mathbf{z}\times d_\mathbf{z}}$ 可能成为 CGKN 的计算瓶颈，因为其参数数量随潜在维度 $d_\mathbf{z}$的平方增长。为确保可扩展性并保留表达能力，我们采用对 $\mathbf{G}_2$ 的低秩近似：

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

其中，$\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d_\mathbf{z}\times r}$ 为正交归一矩阵，$\mathbf{s} \in \mathbb{R}^r$ 包含非负奇异值，而 $\boldsymbol{\delta} \in \mathbb{R}^{d_\mathbf{z}}$ 是可学习的对角校正。有效秩 $r \ll d_\mathbf{z}$ 控制模型容量与效率之间的权衡。在训练期间，通过对无约束矩阵 $\mathbf{U}_{\text{raw}}$ 和 $\mathbf{V}_{\text{raw}}$进行 QR正交化，得到 $\mathbf{U}$ 和$\mathbf{V}$，从而确保

使 $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$ 在每次前向传播时均满足正交条件；同时，$\mathbf{s}$ 通过 Softplus变换进行参数化，以确保其为正。额外的对角项 $\mathrm{diag}(\boldsymbol{\delta})$作为残差校正，可改善条件性，并使潜在模态能够独立缩放。这种参数化将可训练参数的数量从 $\mathcal{O}(d_{\mathbf{z}}^2)$ 降至 $\mathcal{O}(d_{\mathbf{z}}r)$，大幅降低了潜在变量更新中的内存和计算成本（9b）。实践中我们发现，适中的秩（例如 $r = 64$ 至 128）已经能够捕捉潜在流动力学的基本耦合结构，同时保持数值稳定性。这种对$\mathbf{G}_2$ 进行的受奇异值分解启发的低秩构造，使 LaCGKN 能够高效扩展至高维湍流，同时几乎不损失表征能力。

从数据中学习 LaCGKN。LaCGKN 通过最小化复合损失函数进行训练：

$$L(\boldsymbol{\theta}_{\mathcal{E}},\boldsymbol{\theta}_{\mathcal{D}},\boldsymbol{\theta}_{\eta}) := \lambda_{\mathrm{AE}}L_{\mathrm{AE}}(\boldsymbol{\theta}_{\mathcal{E}},\boldsymbol{\theta}_{\mathcal{D}}) + \lambda_{\mathbf{u}}L_{\mathbf{u}}(\boldsymbol{\theta}_{\mathcal{E}},\boldsymbol{\theta}_{\mathcal{D}},\boldsymbol{\theta}_{\eta}) + \lambda_{\mathbf{z}}L_{\mathbf{z}}(\boldsymbol{\theta}_{\mathcal{E}},\boldsymbol{\theta}_{\eta}) + \lambda_{\mathrm{DA}}L_{\mathrm{DA}}(\boldsymbol{\theta}_{\mathcal{D}},\boldsymbol{\theta}_{\eta}), \tag{13}$$

其中

$$L_{\mathrm{AE}} := \mathbb{E}_{\mathbf{u}_2^\star}\left[\frac{1}{d_2}\left\|\mathbf{u}_2^\star - \mathcal{D}(\mathcal{E}(\mathbf{u}_2^\star))\right\|_2^2\right], \tag{14a}$$

$$L_{\mathbf{u}} := \mathbb{E}_{\mathbf{u}^{0\star}}\left[\frac{1}{N_s}\sum_{n=1}^{N_s}\frac{1}{d_{\mathbf{u}}}\left\|\mathbf{u}^{n\star} - \mathbf{u}^{n}\right\|_2^2\right], \tag{14b}$$

$$L_{\mathbf{z}} := \mathbb{E}_{\mathbf{u}^{0\star}}\left[\frac{1}{N_s}\sum_{n=1}^{N_s}\frac{1}{d_{\mathbf{z}}}\left\|\mathbf{z}^{n\star} - \mathbf{z}^{n}\right\|_2^2\right], \tag{14c}$$

$$L_{\mathrm{DA}} := \mathbb{E}_{\mathbf{u}^{0\star}}\left[\frac{1}{N_l - N_b}\sum_{n=N_b+1}^{N_l}\frac{1}{d_2}\left\|\mathbf{u}_2^{n\star} - \boldsymbol{\mu}_2^{n}\right\|_2^2\right]. \tag{14d}$$

损失函数（14a）–（14d）分别表示欧拉流的自编码器重构损失、物理变量的预报损失 $\mathbf{u} = \{\mathbf{u}_1,\mathbf{u}_2\} \in \mathbb{R}^{d_{\mathbf{u}}}$、潜变量的预报损失 $\mathbf{z}$，以及基于后验均值的数据同化损失 $\boldsymbol{\mu}_2^n$，其中后验均值分别由条件高斯滤波器和解码器得到。下标 * 表示参考（真实值）解，而 $\|\cdot\|_2$ 表示欧几里得 ℓ^2-范数。预报步数N_s、数据同化步数N_l、数据同化预热步数N_b以及损失权重 λ_{AE}、 $\lambda_{\mathbf{u}}$、 $\lambda_{\mathbf{z}}$和 λ_{DA}均为可调超参数。

为了从数据中学习 LaCGKN（9），我们遵循 Chen等人（2025b）针对离散时间 CGKN 所提出的两阶段训练策略。在阶段1中，通过最小化复合损失（13）来联合学习自编码器（$\mathcal{E}$、$\mathcal{D}$）和系数（F_1,G_1,F_2,G_2），其中不包含数据同化损失（$\lambda_{\mathrm{DA}} = 0$）。完成阶段1后，通过观测状态 u_1 的单步预测均方根误差（RMSE）估计噪声强度矩阵 $\boldsymbol{\Sigma}_1$。在本文中，噪声强度矩阵 $\boldsymbol{\Sigma}_2$ 通过手动设定获得，但也可以通过潜在状态 z 的单步预测 RMSE 进行估计，或将其设为可训练参数。在阶段2中，通过最小化包含 L_{DA} 的完整复合损失（13），对 LaCGKN代理模型进行数据同化优化。由此得到的 LaCGKN 同时学习了示踪粒子—流场预测和拉格朗日数据同化模型。

算法1 从数据中学习 LaCGKN

输入： $\{\mathbf{u}_1^{n\star},\ \mathbf{u}_2^{n\star}\}_{n=0}^{N}$ ▷训练数据：示踪粒子位置和流场

$\boldsymbol{\theta}_{\mathcal{E}}^{(1)},\ \boldsymbol{\theta}_{\mathcal{D}}^{(1)},\ \boldsymbol{\theta}_{\eta}^{(1)} \leftarrow \arg\min\ \{\lambda_{\mathrm{AE}}L_{\mathrm{AE}} + \lambda_{\mathbf{u}}L_{\mathbf{u}} + \lambda_{\mathbf{z}}L_{\mathbf{z}}\}$ ▷ 训练 LaCGKN，不使用数据同化损失

$\mathrm{diag}(\boldsymbol{\Sigma}_1) \leftarrow$ 均方根误差$(\mathbf{u}_1^\star,\ \mathbf{u}_1)$, $\mathrm{diag}(\boldsymbol{\Sigma}_2) \leftarrow$ 手动设置 ▷ 状态预报的不确定性量化

$\boldsymbol{\theta}_{\mathcal{E}}^{(2)},\ \boldsymbol{\theta}_{\mathcal{D}}^{(2)},\ \boldsymbol{\theta}_{\eta}^{(2)} \leftarrow \arg\min\ \{\lambda_{\mathrm{AE}}L_{\mathrm{AE}} + \lambda_{\mathbf{u}}L_{\mathbf{u}} + \lambda_{\mathbf{z}}L_{\mathbf{z}} + \lambda_{\mathrm{DA}}L_{\mathrm{DA}}\}$ ▷ 使用数据同化损失训练 LaCGKN

$\boldsymbol{\theta}_{\mathcal{U}}^{*} =$ MSE 的最小化$(\|\mathbf{u}_2^\star - \boldsymbol{\mu}_2\|,\ \mathcal{U}(\mathbf{u}_1;\ \boldsymbol{\theta}_{\mathcal{U}}))$ 数据同化的不确定性量化

输出： $\boldsymbol{\theta}_{\mathcal{E}}^{(2)},\ \boldsymbol{\theta}_{\mathcal{D}}^{(2)},\ \boldsymbol{\theta}_{\eta}^{(2)}, \boldsymbol{\theta}_{\mathcal{U}}^{*}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ ▷训练参数和噪声强度

3 数值实验

3.1 应用于具有表层示踪剂观测的两层准地转流

准地转（QG）方程提供了一种非线性但在分析上简洁的模型，能够捕捉中纬度大气和海洋动力学的基本物理机制，并被广泛用于研究大尺度地球物理流（Vallis，2017年）。在本文中，LaCGKN 被应用于一个示踪剂–流场系统，其中流场由具有斜压不稳定性的两层准地转模型控制（Qi 和 Majda，2016年）：

$$\frac{\partial q_1}{\partial t}+J(\psi_1,q_1)+\beta\frac{\partial \psi_1}{\partial x}+U_1\frac{\partial}{\partial x}\nabla^2\psi_1+\frac{k_d^2}{2}\left(U_1\frac{\partial \psi_2}{\partial x}-U_2\frac{\partial \psi_1}{\partial x}\right)=-\nu\Delta^s q_1, \tag{15a}$$

$$\frac{\partial q_2}{\partial t}+J(\psi_2,q_2)+\beta\frac{\partial \psi_2}{\partial x}+U_2\frac{\partial}{\partial x}\nabla^2\psi_2+\frac{k_d^2}{2}\left(U_2\frac{\partial \psi_1}{\partial x}-U_1\frac{\partial \psi_2}{\partial x}\right)=-\left(U_2\frac{\partial h}{\partial x}+\kappa\nabla^2\psi_2\right)-\nu\Delta^s q_2, \tag{15b}$$

其中，$\psi_\ell(\mathbf{x},t)\in\mathbb{R},(\ell=1,2)$ 表示定义在 $(\mathbf{x},t)\in\Omega\times(0,T]\subset\mathbb{R}^2\times\mathbb{R}$ 上的第 ℓ 层流函数，是主要的感兴趣量。上层和下层的位涡扰动分别定义为

$$q_1=\nabla^2\psi_1+\frac{k_d^2}{2}(\psi_2-\psi_1),\quad q_2=\nabla^2\psi_2+\frac{k_d^2}{2}(\psi_1-\psi_2)+h, \tag{16}$$

。这里，h 表示海底地形，$J(A,B)=A_xB_y-A_yB_x$ 表示雅可比，(U_1,U_2) 表示两层的平均流，k_d 表示与罗斯贝变形半径相关的变形波数，β表示罗斯贝参数。项 $\kappa\nabla^2\psi_2$ 表示下层的埃克曼阻尼（底摩擦），而$\nu\Delta^s q_i$ 表示各层的超黏性。在本研究中，我们设定 $(U_1,U_2)=(U\ -U)$，以施加对称的垂直切变，从而引发斜压不稳定性，并设定两层深度相等：$H_1=H_2=H/2$。第 ℓ 层的不可压缩速度场与流函数之间的关系为

$$\mathbf{v}_\ell=\left(\frac{\partial \psi_\ell}{\partial y},-\frac{\partial \psi_\ell}{\partial x}\right)^\top. \tag{17}$$

示踪剂被视为被动粒子，并根据（2a ）.

具有表层示踪粒子观测的双层准地转流，因其固有的多尺度湍流动力学、强非线性以及存在未观测的海底地形，为 LaCGKN 提供了一个具有挑战性的测试平台。与（8）中一样，LaCGKN 将示踪粒子位置 $(\mathrm{x}_1^n,\ldots,\mathrm{x}_I^n)$ 作为观测状态 u_1^n，而将流函数场（ψ_1^n,ψ_2^n）（而非速度场）作为未观测状态 u_2^n。未观测状态 u_2^n通过卷积自编码器编码为具有空间结构 $(\mathrm{z}_h,\mathrm{z}_w,\mathrm{n}_c)$ 的潜在状态 $\mathrm{z}^n\in\mathbb{R}^{d_\mathbf{z}}$。我们发现，将通道数 n_c 设为与层数一致，通常能得到更好的潜在表示。由于该区域为双周期的，示踪粒子位置通过角度嵌入 $(\mathrm{x},\mathrm{y})\mapsto(\cos\mathrm{x},\sin\mathrm{x},\cos\mathrm{y},\sin\mathrm{y})$ 进行编码，且卷积自编码器采用循环填充以保持周期性（详情见附录A）。如第 2.3 节所述，LaCGKN 的观测算子 $(\mathrm{F}_1,\mathrm{G}_1)$ 使用傅里叶位置编码，并在所有示踪剂间共享参数。当潜在维度较高时（例如 $\mathrm{z}_h=\mathrm{z}_w=32$），潜在演化矩阵 G_2 采用受奇异值分解启发的低秩近似进行参数化，以减少可训练参数。

3.2 实验设计

在带有表层示踪粒子观测的双层准地转流上对所提出的 LaCGKN 进行评估。将其在状态预测和数据同化方面的性能与以下列出的几种基线方法进行比较。

状态预报。 以下方法用于对流场和示踪粒子位置进行单步预测。

(i) *LaCGKN*。基于 LaCGKN代理模型（9）提出的深度学习方法，可同时预测示踪剂和流动动力学。潜在维度默认为 $d_\mathbf{z}=16\times16\times2$。

(ii) *LaCGKN*$_{32}$。 LaCGKN 的高容量变体，通过增大潜在维度 $d_{\mathbf{z}} = 32 \times 32 \times 2$，以增强模型的表达能力。

(iii) *DNN*（示踪剂） + 卷积神经网络（流场）。一种模块化深度学习基线，使用深度神经网络（DNN）预测示踪粒子位置，并使用卷积神经网络预测流场。对于示踪剂预测，首先通过傅里叶位置编码（12）对每个示踪粒子位置进行嵌入。随后应用特征级线性调制（FiLM；Perez 等人，2017年），以增强对示踪粒子位置处局部流场信息的提取：

$$\mathbf{u}_1^{n+1} = \mathbf{u}_1^n + \mathrm{DNN}(\beta(\tilde{\mathbf{u}}_1^n) + \gamma(\tilde{\mathbf{u}}_1^n) \odot \mathbf{u}_2^n), \tag{18}$$

其中 $\tilde{\mathbf{u}}_1^n = \mathcal{P}(\mathbf{u}_1^n)$ 表示经过位置编码的示踪剂状态，$\odot$ 表示逐元素乘积。调制参数 β 和 γ 由两个浅层全连接网络给出。对于流场预测，采用带循环填充的卷积神经网络，根据 $\mathbf{u}_2^n$预测 $\mathbf{u}_2^{n+1}$。这种模块化架构为纯状态预报提供了强有力的深度学习基线。

(iv) 持续性。一种简单的基线，假设系统不随时间演化，即 $\mathbf{u}^{n+1} = \mathbf{u}^n$。持续性为预测技巧提供了下限，并揭示了相较于简单复制当前状态所取得的改进。

数据同化。 以下方法用于评估数据同化，旨在根据表面示踪剂观测推断欧拉流场状态。

(i) *LaCGKN*。 所提出的方法使用条件高斯滤波器（6）在潜在空间中进行解析滤波，然后解码到物理空间。同化过程中不使用物理模型。

(ii) 集合卡尔曼滤波器（*EnKF*）。 一种用于非线性系统的经典序贯数据同化方法（Evensen，1994）。我们采用一种确定性 EnKF 变体——集合调整卡尔曼滤波器（EAKF；Anderson，2001），并使用增广状态公式进行拉格朗日数据同化，方法参照Wang 等人 (2025)。同时采用恒定乘性协方差膨胀（Anderson and Anderson，1999）和 Gaspari–Cohn 局地化（Gaspari and Cohn，1999）。EAKF 使用完美QG模型运行集合预报。它作为强有力的基于物理的基线，凸显了 LaCGKN 的准确性和效率。

(iii) 最优插值（*OI*）。一种简化的基于卡尔曼滤波的方法，采用预先设定且随时间不变的背景误差协方差和观测误差协方差。根据Molcard 等人（2003年、2005年），通过对 Δt_{obs} 时间段内的示踪粒子位置进行有限差分来构造拉格朗日速度（从而将位置观测转换为速度观测），并通过线性回归将表层校正投影到下层。采用高斯分布局地化。OI 使用完美的准地转模型进行流场预报。它提供了一个简化的基线，并凸显了捕捉示踪粒子–流场动力学中强非线性的重要性。

(iv) 气候态。一种忽略观测、将长期平均流场作为固定背景状态的朴素基线方法。它提供了一个不进行同化时的参考误差水平。

数据生成。通过对示踪剂流动系统进行数值积分，生成成对的拉格朗日示踪剂观测和欧拉QG流场。QG流在湍流状态下演化，其无量纲参数为 $\mathrm{k}_d = 10$, $\beta = 22$、$\mathrm{U} = 1$, $\kappa = 9$, $\nu = 10^{-12}$、$\mathrm{s} = 4$，以及 $\mathrm{h}(\mathrm{x}, \mathrm{y}) = 40\cos\mathrm{x} + 80\cos(2\mathrm{y})$。QG方程在双周期域 $\Omega = [0, 2\pi)^2$上的 128×128 伪谱网格上求解。通过将位涡扰动采样为 $\mathrm{q}_1^0 = \mathrm{q}_2^0 \sim \mathcal{N}(0, 10^2)$ 对流场进行初始化，并在预热阶段进行1000步后，以时间步长 $\Delta\mathrm{t} = 2 \times 10^{-3}$ 积分 $\mathrm{N}_t = 2 \times 10^6$ 步。共初始化1024个示踪剂，位置为 $\mathrm{x}_i^0 \sim \mathcal{N}(\pi, 0.12)$，并在表面流作用下以强度为 $\sigma_x = \sigma_y = 0.1$ 的随机噪声进行平流。通过添加独立的高斯测量噪声 $\mathcal{N}(0, 0.012)$ 构造合成示踪粒子观测。在训练期间，向流函数场中加入相同水平的噪声，以模拟真实部署场景并评估鲁棒性。随后将数据下采样至 64×64 网格，观测间隔为 $\Delta\mathrm{t}_{\mathrm{obs}} = 4 \times 10^{-2}$。所得数据集由示踪粒子位置数据 $\{\mathrm{x}_i^n\}_{\mathrm{i}=1,\mathrm{n}=0}^{I,N_{t_{obs}}}$ 和欧拉QG流场 $\{\psi_{1,m}^n, \psi_{2,m}^n\}_{\mathrm{m}=1,\mathrm{n}=0}^{M,N_{t_{obs}}}$ 组成，其中 $\mathrm{M} = 64 \times 64$ 和 $\mathrm{N}_{t_{obs}} = 10^5$。

数据集被划分为训练段、验证段和测试段，分别包含80，000/10，000/10*000*个同化步数。除非另有说明，预测和同化过程中仅使用$I = 64$ 个示踪剂。在训练期间，每个批次从总计1024个示踪剂中随机抽取64个，以促进（$\mathbf{F}_1$,$\mathbf{G}_1$）的示踪粒子齐次化并提高泛化能力。

LaCGKN 训练设置。LaCGKN 由一个用于欧拉流的循环卷积神经网络自编码器（$\mathcal{E},\mathcal{D}$）和一个带有系数（F_1,G_1,F_2, G_2）的条件高斯网络（9）组成。默认情况下，编码器 $\mathcal{E}$：$\mathbb{R}^{64\times64\times2} \mapsto \mathbb{R}^{16\times16\times2}$ 将流场 u_2^n 映射到潜在状态 z^n ($\mathrm{d}_{\mathbf{z}} = 512$)。解码器 $\mathcal{D}$：$\mathbb{R}^{16\times16\times2} \mapsto \mathbb{R}^{64\times64\times2}$ 与编码器相对应，将 z^n 映射回 u_2^n。给定角度嵌入的示踪粒子位置 $\mathrm{u}_1^n \in \mathbb{R}^{4I}$，每个角度表示（$\cos \mathrm{x}_i$, $\sin \mathrm{x}_i$, $\cos \mathrm{y}_i$, $\sin \mathrm{y}_i$）首先转换回（x_i, y_i），然后通过具有 $\mathrm{K} = 6$ 个频率的傅里叶位置编码进行嵌入，从而为每个示踪粒子生成 $\tilde{\mathbf{x}}_i^n = \mathcal{P}(\mathrm{x}_i^n) \in \mathbb{R}^{26}$。子网络 f_θ：$\mathbb{R}^{26} \to \mathbb{R}^4$ 和 G_θ：$\mathbb{R}^{26} \to \mathbb{R}^{4\times512}$均为 DNN，以 $\tilde{\mathbf{x}}_i^n$ $\in \mathbb{R}^{512}$ $_1 \in \mathbb{R}^{4I}$ $_1 \in \mathbb{R}^{4I\times512}$ $_2 \in \mathbb{R}^{512\times512}$ $_2 \in$ $\mathbf{z} =$ x 作为输入，并构造 F 和 G。潜在转移参数 F 和 G 是学习得到的常数。为探索更高容量的模型，还考虑了更大的潜在维度 $\mathrm{d}32\times32\times2$，称为 LaCGKN_{32}；在这种情况下，对 G_2（第 2.3 节）采用受奇异值分解启发的低秩参数化，其有效秩为 $\mathrm{r} = 64$。LaCGKN 的训练设置包括：状态预报步数 $\mathrm{N}_s = 1$（即单步预测）、数据同化步数 $\mathrm{N}_l = 100$、数据同化预热步数 $\mathrm{N}_b = 20$以及损失权重 $\lambda_{\mathrm{AE}} = \lambda_{\mathbf{u}} = \lambda_{\mathbf{z}} = \lambda_{\mathrm{DA}} = 1$。

测试与评估。在包含 10，000 步的测试数据集中，评估状态预报的单步预测误差，以及数据同化的后验误差，并将前 50 步作为自旋调整阶段排除在外。LaCGKN 执行解析潜在空间滤波，初始化为 $\mathbf{z}^0 \sim \mathcal{N}(\mathbf{0}, 0.12\mathbf{I})$。集合卡尔曼滤波和最优插值使用完美QG模型进行序贯滤波；在每个网格点和每一层，流场初始化为 $\Psi_{\ell,m}^0 \sim \mathcal{N}(0, 0.12)$。对于集合卡尔曼滤波，每个模拟示踪剂均初始化为围绕初始示踪剂观测值 $\mathbf{x}^0 \sim \mathcal{N}(\mathbf{x}_{\mathrm{obs}}^0, 0.012\mathbf{I})$。集合卡尔曼滤波采用集合规模 $N_e = 40$、1.025 的膨胀因子和 16 个网格点的局地化半径。最优插值使用预设的速度背景误差标准差 $\sigma_b = 1.0$，以及 1 个网格点的高斯局地化半径。用于将表层修正投影到下层的回归系数，根据两层之间逐网格点的速度计算得出。所有可调参数均通过经验优化确定。

评估指标。均方根误差（RMSE）用作点估计指标。对于感兴趣量$\mathbf{a} \in \mathbb{R}^{d_{\mathbf{a}}}$（例如堆叠示踪剂位置或展平流场）及其参考值$\mathbf{a}^\star$，RMSE 为

$$\mathrm{RMSE} := \left(\frac{1}{d_{\mathbf{a}}} \left\| \mathbf{a} - \mathbf{a}^\star \right\|_2^2 \right)^{1/2}. \tag{19}$$

对于数据同化，集合卡尔曼滤波和最优插值使用后验均值作为点估计，而 LaCGKN 使用解码后的潜在后验均值 $\boldsymbol{\mu}_2 = \mathcal{D}(\boldsymbol{\mu}_{\mathbf{z}})$。对于所有基于深度学习的方法，为确保比较公平，可训练参数的数量设置为相近。默认 LaCGKN 的可训练参数数量为 1,021,512，LaCGKN_{32},为 927,932，DNN+卷积神经网络为 1,089,894。

3.3 数值结果

LaCGKN 的性能从单步预测和数据同化两方面进行评估。一个关键区别在于，LaCGKN 在统一框架内执行预报和数据同化，而下文所讨论的基准方法仅针对这两项任务中的其中一项。主要发现总结如下：

- **准确且稳定的流场预报**。默认的 LaCGKN 显著优于持续性方法，展现了其学习非平凡动力学结构的能力。增大潜在维度（LaCGKN_{32}）后，其均方根误差在所有对比方法中最低，并能保持稳定的长期滚动预测；相比之下，CNN基线在约 10 步后变得不稳定（表1 和图 2）。

- **准确的数据同化与可靠的不确定性量化**。在无法使用物理模型的情况下，与采用完美QG模型的基于物理的EnKF、最优插值和气候态相比，LaCGKN 的后验均方根误差最低，且流场结构相干（表 2 和图 3）。LaCGKN 还提供了校准良好的后验不确定性量化，其后验离散度随时间变化与后验均方根误差高度吻合（图 4）。

- **计算效率高且可扩展性强**。 LaCGKN 的速度约为并行化 EnKF 的百倍，同时后验均方根误差更低；随着示踪剂数量增加，其效率优势也愈发明显。此外，LaCGKN 训练完成后，无需重新训练即可泛化到不同数量的示踪剂（图5）。

详细结果如下所示。

状态预报。 表 1 总结了不同方法的单步预测均方根误差。默认 LaCGKN 和 LaCGKN_{32} 在所有变量上的均方根误差都显著低于持续性方法，这表明学习得到的替代模型捕捉到了有意义的动力学结构，而不只是记忆数据。对于上层流场，这种改进尤为显著，因为强烈的湍流变化会导致持续性方法产生较大的误差。尽管默认 LaCGKN 的均方根误差高于 DNN+CNN 基线，但增大 LaCGKN 的潜在维度可显著提升流场预测精度。具体而言，LaCGKN_{32} 将两层均方根误差从 0.104 降至 0.037，取得所有方法中最低的均方根误差。这一显著提升凸显了潜在嵌入容量对于表示复杂湍流的重要性。提供足够大的潜在维度后，条件高斯替代模型能够很好地近似底层非线性动力系统。

在示踪物位置预测方面，LaCGKN 的表现不如 DNN+卷积神经网络。这是意料之中的，因为 DNN 基线直接利用物理空间中的流场来输运示踪剂，而 LaCGKN 依赖潜在嵌入，这些嵌入需要同时捕捉示踪物-流场相互作用和流动动力学。尽管如此，本工作的重点是实现准确的流场预测，而 LaCGKN_{32} 在这方面表现出色。这一优势在流场的自回归多步预测中也得到进一步体现（图2），其中 LaCGKN_{32} 在长时间滚动预测中保持了稳定且符合物理规律的流场结构，而卷积神经网络模型在约 10 步后变得不稳定。这些结果表明，LaCGKN 不仅仅是一个预测器，还是一种动力替代模型，能够约束潜在演化的结构，从而提升长时域稳定性。此外，LaCGKN 除了独立预报之外，还能高效地进行数据同化，这是其相较于纯预测型深度学习模型的一项关键优势。

表1：不同方法的状态预报（单步预测）均方根误差，包括默认潜在维度的 LaCGKN、$_{32}$潜在维度增大的 LaCGKN$d_{\mathbf{z}} = 32 \times 32 \times 2$、DNN（示踪剂）+卷积神经网络（流场）和持续性方法。误差针对示踪粒子位置、上层流函数 ψ_1、下层流函数 ψ_2以及两层流函数总和计算。

方法	示踪剂	上层	下层	两层
LaCGKN	0.099	0.125	0.079	0.104
LaCGKN_{32}	0.094	0.042	0.032	0.037
DNN+卷积神经网络	0.064	0.071	0.069	0.070
持续性	0.136	0.294	0.177	0.243

数据同化。表2 汇总了不同方法的流场估计后验均方根误差。在不使用物理准地转模型的情况下，默认的 LaCGKN 在两层中的均方根误差均低于集合卡尔曼滤波、最优插值和气候态。LaCGKN 相较于最优插值的优势尤为明显。由于最优插值仅在示踪剂位置附近引入局部修正，并通过层间线性投影传播修正，因此在这种强非线性湍流状态下表现不佳。作为更强的基于物理模型的基准方法，集合卡尔曼滤波通过使用完整物理模型传播集合预报，并估计依赖于流场的误差协方差，能够更有效地捕捉非线性

。尽管如此，LaCGKN 的后验均方根误差仍低于集合卡尔曼滤波，而且无需进行完整物理模型的集合传播，也无需对膨胀和局地化等超参数进行经验调优。图3 中的流场快照进一步证实，LaCGKN 能够准确恢复两层中的相干流动结构。这些结果表明，训练得到的 LaCGKN 代理模型成功捕捉了实现准确数据同化所需的非线性示踪剂-流场耦合和多层相互作用。此外，后验离散度与后验均方根误差随时间高度吻合，说明 LaCGKN 在提供准确均值估计的同时，也能实现可靠的不确定性量化（图4）。

敏感性与计算成本。图5比较了在示踪剂数量不同的情况下，LaCGKN 和集合卡尔曼滤波的后验均方根误差及计算成本（CPU 上的墙钟时间）。结果揭示了几个关键现象。首先，在不同示踪剂数量下，LaCGKN 的后验均方根误差始终低于集合卡尔曼滤波。随着示踪剂数量增加，两者的性能差距进一步扩大，这表明学习得到的代理模型能够有效利用额外的观测信息。其次，LaCGKN 的速度约为并行化 EnKF 的$O(100)$倍。这种效率提升源于在低维潜在空间中使用闭式更新进行滤波，而非在完整物理空间中传播和更新集合。计算复杂度主要随潜在维度而非集合规模变化。第三，LaCGKN 在推理阶段不受示踪剂数量影响。针对某一配置完成训练后，同一模型无需重新训练即可直接用于不同数量的示踪剂。相比之下，随着示踪剂数量变化，集合卡尔曼滤波需要重新调整超参数，才能保持最佳性能。在真实部署场景中，漂移观测仪器的数量可能随时间变化，这种对示踪剂数量的不变性大幅降低了运行成本。总体而言，这些结果表明，LaCGKN 不仅能够实现准确的数据同化并提供可靠的不确定性量化，而且在训练和推理阶段都保持了很高的效率。

表 2：不同方法的数据同化后验估计的均方根误差，包括采用默认潜在维度的 LaCGKN、集合卡尔曼滤波、最优插值和气候态。误差分别针对上层流函数 ψ_1、下层流函数 ψ_2 以及两层流函数的总误差计算。

方法	上层	下层	两层
LaCGKN	0.579	0.310	0.464
EnKF	0.599	0.321	0.481
OI	0.890	0.467	0.710
气候态	0.870	0.414	0.681

4 结论与讨论

我们开发了拉格朗日条件高斯 Koopman 网络（LaCGKN），以应对拉格朗日数据同化与预测中的挑战性任务，重点是根据拉格朗日示踪剂观测推断高维欧拉流场。主要难点在于，在移动的示踪剂位置对流场进行评估会引发内在非线性，同时流动动力学本身也具有很强的非线性和湍流特性。LaCGKN 通过学习具有条件高斯结构的代理模型来应对这一挑战：将欧拉流编码为具有近似线性动力学的低维潜在状态，同时以由潜在流场驱动的非线性但结构化的模型来表示。该结构支持解析的潜在空间滤波，并可通过闭式形式进行后验更新。为提高拉格朗日框架下的鲁棒性和可扩展性，我们引入了：（i）示踪粒子齐次化，以实现置换等变性，并支持模型在不同示踪剂数量下的泛化；（ii）基于傅里叶的位置编码，以捕捉示踪剂位置处流场丰富的非线性空间依赖性；以及（iii）受奇异值分解启发的潜在转移矩阵低秩参数化 $\mathbf{G}_2$，以减少可训练参数的数量，同时保留表达能力和效率。最终的 LaCGKN 通过复合损失进行训练，从而形成一个统一的深度学习模型，可同时用于预测和数据同化。

LaCGKN在具有表面示踪剂位置观测的湍流双层准地转（QG）系统上进行了评估。对于状态预报，默认的 LaCGKN已经能够通过显著

优于持续性方法，提供实用的预报模型；增加潜在维度（LaCGKN_{32}）后，其流场预测准确率在所比较的方法中达到最高。LaCGKN还能产生稳定且准确的多步滚动预测，而CNN基线在约10步后便变得不稳定。对于数据同化，LaCGKN获得的后验RMSE与采用完美QG模型的基于物理的EnKF相当，并且显著低于最优插值和气候态，同时还能提供可靠的不确定性量化。此外，LaCGKN具有很高的计算效率，在保持较高准确率的同时，其速度约为并行化 EnKF的$O(100)$倍；随着示踪剂数量增加，这些优势会更加明显。最后，由于LaCGKN的设计对示踪剂数量具有不变性，因此只需训练一次即可应用于不同数量的示踪剂，无需重新训练，从而降低开发和部署成本，并为拉格朗日数据同化与预测提供一种高效、完全数据驱动的替代方案。

条件高斯（CG）滤波器将 Kalman-Bucy滤波器推广到具有非线性观测过程的系统，其源自滤波的连续时间表述（Kalman 和 Bucy，1961；Jazwinski，1970；Chen 和 Majda，2018）。以 Kalman滤波器为代表的滤波离散时间表述，与其连续时间对应形式有着根本区别（Wang 等人，2025）。这种区别并非仅仅源于时间离散化，因为条件高斯滤波器也存在离散时间版本，而是源自观测本身的表述方式。在标准离散时间滤波中，观测噪声与系统动力学相互独立；而在连续时间滤波（及其离散化代理模型）中，观测噪声直接进入动力学过程。在当前的 CGKN 表述中，示踪粒子位置的测量误差未被显式建模，而是通过动力学噪声进行隐式近似。尽管数值结果表明，该方法对小幅测量噪声具有鲁棒性，但仍值得将测量噪声显式纳入 CGKN框架，从而有望衔接连续时间与离散时间滤波表述，并进一步提升含噪观测条件下的数据同化精度。

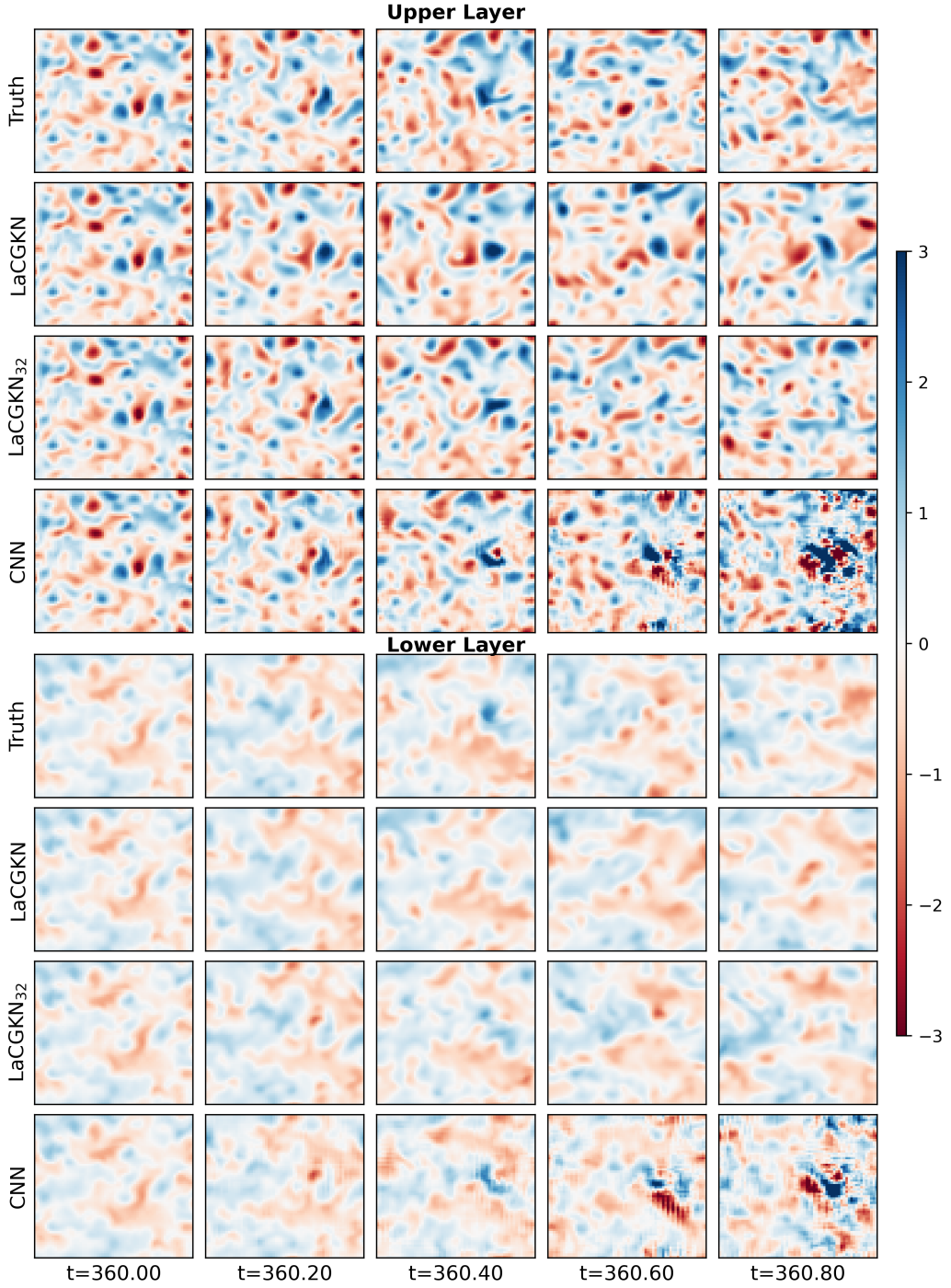


图 2：用于状态预测比较的流函数流场快照。预报以时刻 $t = 360$ 为初始时刻，并以自回归方式滚动预测多个时间步。比较真实值、LaCGKN（默认潜在维度 $d_{\mathbf{z}} = 16 \times 16 \times 2$）、$\text{LaCGKN}_{32}$（潜在维度 $d_{\mathbf{z}} = 32 \times 32 \times 2$）和卷积神经网络的结果。

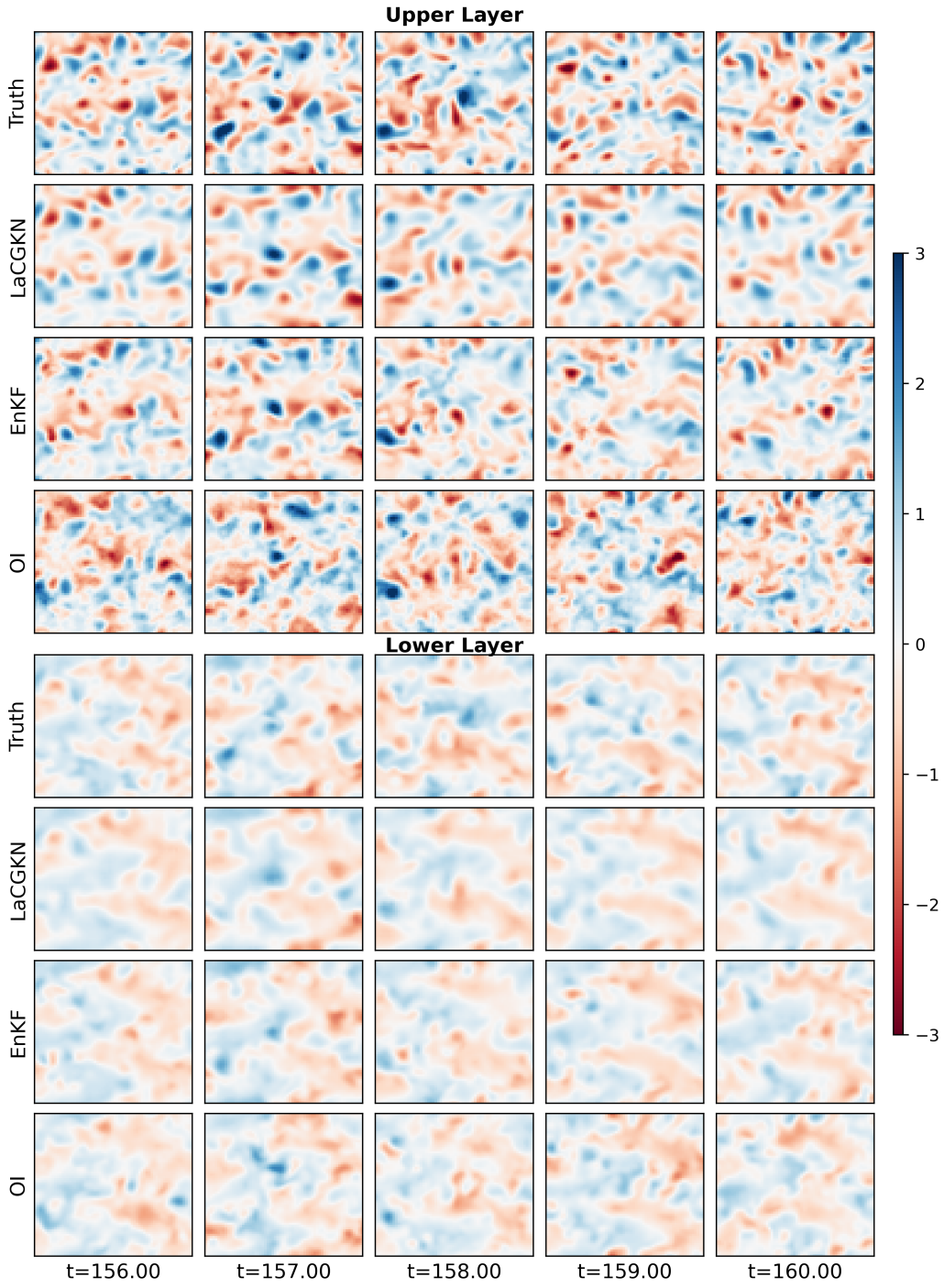


图3：用于数据同化对比的流函数流场快照。将真实值与 LaCGKN、集合卡尔曼滤波和最优插值的后验估计进行比较。

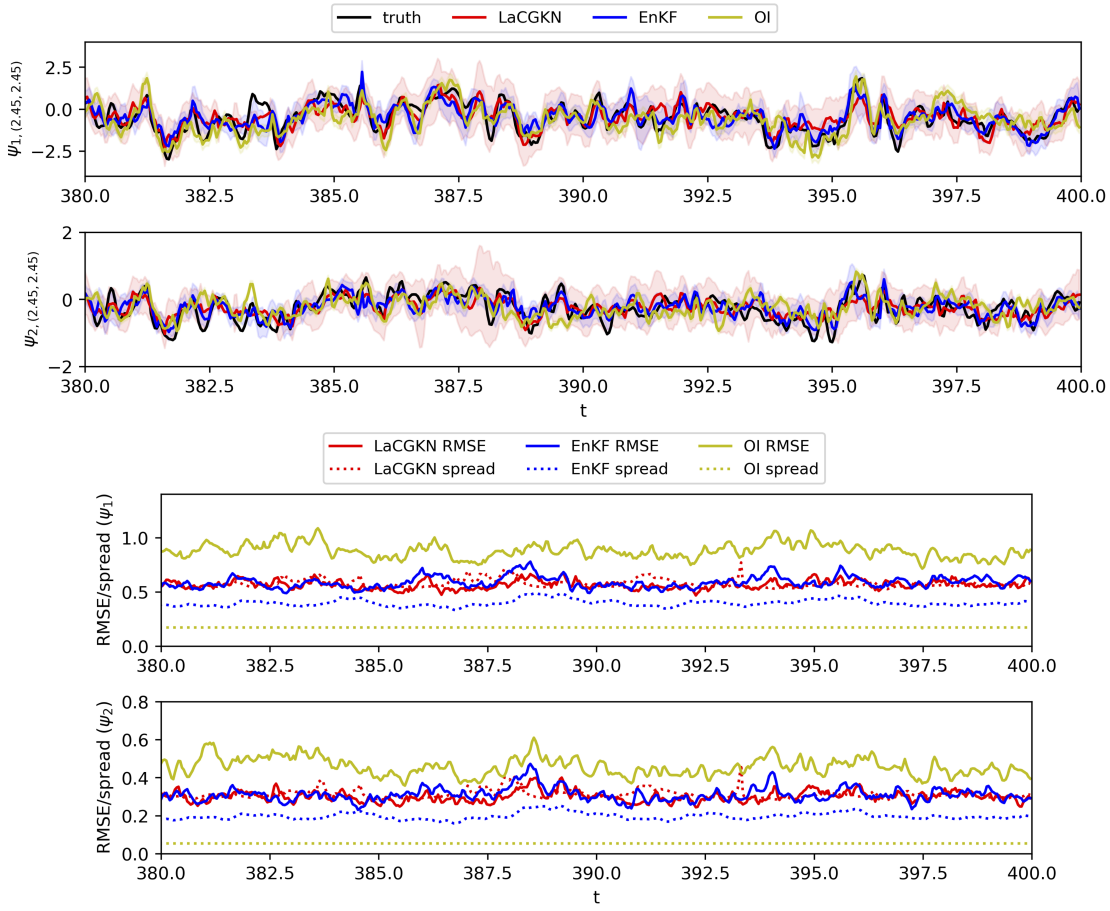


图 4: （a）位置（x，y）$=(2.45,\ 2.45)$ 处数据同化后验估计的时间序列。真值以黑线表示。比较方法包括 LaCGKN（红线）、集合卡尔曼滤波（蓝线）和最优插值（黄线）。不确定性以相应阴影区域表示的 ± 2 后验标准差展示。（b）LaCGKN（红线）、集合卡尔曼滤波（蓝线）和最优插值（黄线）的数据同化后验均方根误差（实线）和后验离散度（标准差；虚线）的时间序列。

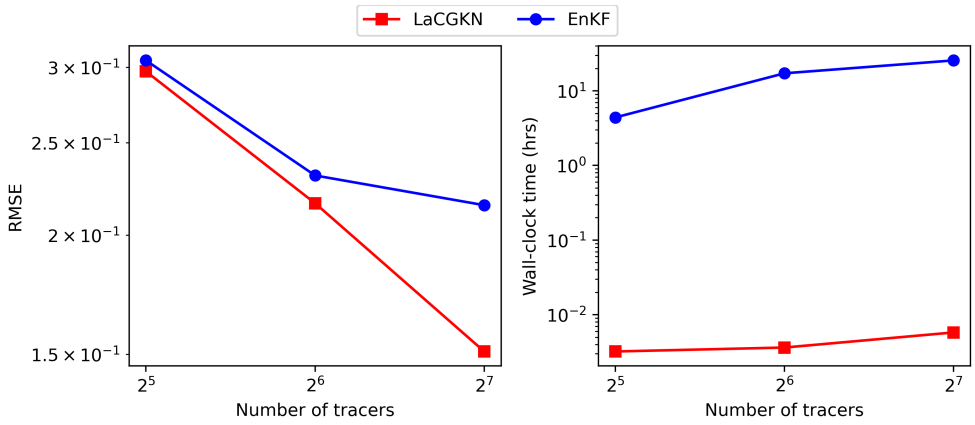


图 5：LaCGKN 和集合卡尔曼滤波的后验均方根误差及计算成本（实际运行时间）随示踪剂数量的变化。实际运行时间是在同一 CPU 上运行程序测得的。集合卡尔曼滤波的模型预报通过集合成员并行执行。

致谢。 N.C. 的研究由美国海军研究办公室（资助编号 N00014-24-1-2244）和美国陆军研究办公室（资助编号 W911NF-23-1-0118）资助。Z.W. 的研究部分由这些资助项目支持。C.C. 和 J.W. 的研究由威斯康星大学麦迪逊分校副校长研究与研究生教育办公室资助，资金来自威斯康星校友研究基金会。

数据可用性声明。 代码可在此处获取： https://github.com/zhongruiw/LaCGKN .

附录A. 处理周期性条件的技术细节。

A.1. 示踪粒子位置的循环性。

示踪剂在双周期域上演化，其原始笛卡尔坐标在跨越边界时会出现不连续。这使得标准的欧几里得表示不再适用，因为跨越 2π 边界的小位移会表现为一次大幅跳变。为避免此类伪影，所有示踪粒子位置均采用角度嵌入进行编码

$$(x, y) \mapsto (\cos x,\ \sin x,\ \cos y,\ \sin y),$$

该表示保持了环面上的周期性和连续性。随后，单位向量表示被映射回物理坐标，并进一步与傅里叶特征结合，然后输入网络 $\mathbf{f}_\theta$ 和 $\mathbf{G}_\theta$。这种表示确保神经网络接收到几何上一致的输入，并防止训练数据中出现无意义的不连续性。

A.2. 用于流场的循环卷积神经网络自编码器。

64×64 双周期网格上的欧拉流场通过卷积自编码器进行编码和解码，其所有层在两个空间方向上均使用环绕（循环）填充。因此，每个卷积实际上都作用于一个环面上，不会产生人为边缘。编码器 $\mathcal{E}$ 通过堆叠循环 3×3 卷积和池化，将多通道流场压缩为潜在状态 $\mathbf{z}^n$；解码器 $\mathcal{D}$ 则使用循环转置卷积来反转这一映射。该设计生成了紧凑的平移等变潜在表示，并通过结构设计确保重构流场具有周期性。

参考文献

Anderson, J. L., 2001：用于数据同化的集合调整卡尔曼滤波器。《月刊天气评论》，**129(12)**，2884–2903，https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2。

Anderson, J. L., and S. L. Anderson, 1999：通过蒙特卡洛实现非线性滤波问题，以生成集合资料同化和预报。《月刊天气评论》，https://doi.org/https://doi.org/10. 1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2。

Apte, A., C. K. R. T. Jones, and A. M. Stuart, 2008：一种用于拉格朗日数据同化的贝叶斯方法。*Tellus A*：动力气象学与海洋学，**60 (2)**。

Asefi, N., L. Lupin-Jimenez, T. Wu, R. He 和 A. Chattopadhyay，2025：极端稀疏性条件下海洋动力学的生成式拉格朗日数据同化。arXiv，URL http://arxiv.org/abs/2507.06479，arXiv:2507.06479[physics]，https://doi.org/10.48550/arXiv.2507.06479。

Bach, E.、R. Baptista、E. Calvello、B. Chen 和 A. Stuart，2025：学习增强型集合滤波器。arXiv，URLhttp://arxiv.org/abs/2504.17836，arXiv:2504。17836 [stat],https://doi.org/10.48550/arXiv.2504.17836。

Baptista, R., A. M. Stuart, and S. Tran, 2025: 对比学习的数学视角. arXiv, URLhttp://arxiv.org/abs/2505.24134, arXiv:2505.24134 [stat],https://doi.org/10.48550/arXiv.2505.24134.

Bocquet, M., A. Farchi, T. S. Finn, C. Durand, S. Cheng, Y. Chen, I. Pasmans, and A. Carrassi, 2024: 通过识别不稳定性（无需集合）实现混沌动力学的精准深度学习滤波。《混沌：非线性科学跨学科期刊》, 34 (9), 091 104, https://doi.org/10.1063/5.0230837. Boudier, P., A. Fillion, S. Gratton, S. Gürol, and S. Zhang, 2023: 数据同化网络。《地球系统建模进展》, 15 (4), e2022MS003 353, https://doi.org/10.1029/2022MS003353. Businger, S., S. R. Chiswell, W. C. Ulmer, and R. Johnson, 1996: 气球作为大气研究的拉格朗日测量平台。《地球物理研究杂志：大气》, 101 (D2), 4363–4376, https://doi.org/ 10.1029/95JD00559. Centurioni, L., A. Horányi, C. Cardinali, E. Charpentier, and R. Lumpkin, 2017: 用于测量海平面大气压力的全球海洋观测系统：对数值天气预报的作用与影响。《美国气象学会通报》, https://doi.org/10.1175/BAMS-D-15-00080.1. Chen, C., N. Chen, and J.-L. Wu, 2024: CGNSDE：用于复杂系统建模和数据同化的条件高斯神经随机微分方程。《计算机物理通讯》, 304, 109 302, https://doi.org/10.1016/j.cpc.2024.109302. Chen, C., N. Chen, Y. Zhang, and J.-L. Wu, 2025a: CGKN：用于复杂动力系统建模和高效数据同化的深度学习框架。《计算物理学杂志》, 532, 113 950, https://doi.org/ 10.1016/j.jcp.2025.113950. Chen, C., Z. Wang, N. Chen, and J.-L. Wu, 2025b: 通过离散时间条件高斯 Koopman 网络对部分观测非线性动力系统建模并实现高效数据同化。《计算力学与工程方法》, 445, 118 189, https://doi.org/10.1016/j.cma.2025.118189. Chen, C., and J.-L. Wu, 2025: 神经动力算子：结合基于梯度和无导数优化方法的连续时空模型。《计算物理学杂志》, 520, 113 480, https://doi.org/10. 1016/j.jcp.2024.113480. Chen, N., S. Fu, and G. E. Manucharyan, 2022: 一种高效且统计精确的拉格朗日数据同化算法及其在离散元海冰模型中的应用。《计算物理学杂志》, 455, 111 000, https://doi.org/10.1016/j.jcp.2022.111000. Chen, N., and A. J. Majda, 2018: 多尺度非线性随机系统的条件高斯系统：预测、状态估计与不确定性量化。《熵》, 20 (7), 509, https://doi.org/ 10.3390/e20070509. Chen, R. T. Q., Y. Rubanova, J. Bettencourt, and D. Duvenaud, 2019: 神经常微分方程. arXiv, URL http://arxiv.org/abs/1806.07366, arXiv:1806.07366 [cs, stat], https://doi.org/10.48550/arXiv.1806. 07366. Cheng, S., and Coauthors, 2023: 动力系统数据同化与不确定性量化的机器学习：综述. arXiv, URL http://arxiv.org/abs/2303.10462, arXiv:2303.10462 [cs]. Dong, X., C. Chen, and J.-L. Wu, 2025: 通过条件扩散模型和神经算子进行数据驱动的随机闭合建模。《计算物理学杂志》, 534, 114 005, https://doi.org/10.1016/j.jcp.2025.114005. Du, P., M. H. Parikh, X. Fan, X.-Y. Liu, and J.-X. Wang, 2024: 用于生成时空湍流的条件神经场潜扩散模型。《自然·通讯》, 15 (1), https://doi.org/10.1038/ s41467-024-54712-1. Evensen, G., 1994: 使用蒙特卡洛方法预报误差统计的非线性准地转模型序贯数据同化。《地球物理研究杂志》, 99 (C5), 10 143, https://doi.org/10.1029/ 94JC00572.

Fablet, R., B. Ch围裙, L. Drumetz, E. Mémin, O. Pannekouck[e, 一个]d F. R 2021 Learning Variational Dat
a 同化模型与求解器。《地球系统建模进展》，13（10），e2021MS002 ousseau：572，
https://doi.org/10.1029/2021MS002572。Farchi, A., M. Chrust, M. Bocquet, P. Laloyaux, and M. Bonavita, 2023：在增量式四维变分框架中利用神经网络进行在线模型误差校正。《地球系统建模进展》，15（9），e2022MS003 474，https://doi.org/10.1029/2022MS003474。Gaspari, G., and S. Cohn, 1999：二维和三维相关函数的构造。《英国皇家气象学会季刊》，125（554），723–757，https://doi.org/10.1256/smsqj.55416。Gauthier, D. J., E. Bollt, A. Griffith, and W. A. S. Barbosa, 2021：新一代储层计算。《自然·通讯》，12（1），5564，https://doi.org/10.1038/s41467-021-25801-2。Gould, J., and Coauthors, 2004：Argo剖面浮标开启海洋现场观测新时代。《Eos：美国地球物理联合会汇刊》，85（19），185–191，https://doi.org/10.1029/2004EO190002。Hochreiter, S., and J. Schmidhuber, 1997：长短期记忆。《神经计算》，9（8），1735–1780，https://doi.org/10.1162/neco.1997.9.8.1735。Ide, K., L. Kuznetsov, and C. K. R. T. Jones, 2002：点涡系统的拉格朗日数据同化*。《湍流期刊》，3（1），053，https://doi.org/10.1088/1468-5248/3/1/053。Jazwinski, A., 1970：5 滤波理论导论。《随机过程与滤波理论》，第64卷，爱思唯尔，142–161，https://doi.org/10.1016/S0076-5392(09)60374-X，网址：https://linkinghub.elsevier.com/retrieve/pii/S007653920960374X。Kalman, R., and R. Bucy, 1961：线性滤波与预测理论的新成果。《流体工程期刊》，美国机械工程师学会汇刊，83（1），95–108，https://doi.org/10.1115/1.3658902。Kuznetsov, L., K. Ide, and C. K. R. T. Jones, 2003：一种拉格朗日数据同化方法。《月刊天气评论》，https://doi.org/https://doi.org/10.1175/1520-0493(2003)131<2247:AMFAOL>2.0.CO;2。Legler, D., and Coauthors, 2015：面向业务海洋学的全球海洋实时现场观测系统现状。《业务海洋学期刊》，8（sup2），s189–s200，https://doi.org/10.1080/1755876X.2015.1049883。Li, S., T. Zheng, A. Farchi, M. Bocquet, and P. Gentine, 2025：利用生成式机器学习进行概率数据同化，以投影集合分布：Lorenz ’96 概念验证。《地球物理研究快报》，52（12），e2024GL112 523，https://doi.org/10.1029/2024GL112523。Li, Z., N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, 2021：用于参数化偏微分方程的傅里叶神经算子。《arXiv》，网址：http://arxiv.org/abs/2010.08895，arXiv:2010.08895 [cs]，https://doi.org/10.48550/arXiv.2010.08895。Lorenc, A. C., 1986：数值天气预报的分析方法。《英国皇家气象学会季刊》，112（474），1177–1194，https://doi.org/10.1002/qj.49711247414。Lorenc, A. C., 2003：利用四维变分数据同化对误差协方差进行建模。《英国皇家气象学会季刊》，129（595），3167–3182，https://doi.org/10.1256/qj.02.131。Lu, L., P. Jin, G. Pang, Z. Zhang, a nd G. E. Karniadakis, 2021：基于算子普适逼近定理，通过DeepONet学习非线性算子。《自然·机器智能》，3（3），218–229，https://doi.org/10.1038/s42256-021-00302-5。Lumpkin, R., T. Özgökmen, and L. Centurioni, 2017：表层漂流浮标应用的进展。《海洋科学年度评论》，9（1），59–81，https://doi.org/10.1146/annurev-marine-010816-060641。

Lusch, B., J. N. Kutz, and S. L. Brunton, 2018：用于非线性动力学通用线性嵌入的深度学习。《自然·通讯》，9（1），4950，https://doi.org/10.1038/s41467-018-07210-0。Mariano, A. J., A. Griffa, T. M. Özgökmen, and E. Zambianchi, 2002：《拉格朗日分析与海岸及海洋动力学预测 2000》。《大气与海洋技术杂志》。Martin, S. A., G. E. Manucharyan, and P. Klein, 2025：利用多模态卫星观测进行海洋表层状态估计的生成式数据同化。《地球系统建模进展》，17（8），e2025MS005 063，https://doi.org/10.1029/2025MS005063。McPhaden, M. J., K. J. Connell, G. R. Foltz, R. C. Perez, and K. Grissom, 2023：天气与气候的热带海洋观测：全球热带系泊浮标阵列十年回顾 ı《海洋学》。《海洋学》，https://doi.org/https://doi.org/10.5670/oceanog.2023.211。Mildenhall, B., P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, 2020：NeRF：将场景表示为神经辐射场以进行视图合成。《arXiv》，网址：http://arxiv.org/abs/2003.08934，arXiv:2003.08934 [cs]，https://doi.org/10.48550/arXiv.2003.08934。Molcard, A., A. Griffa, and T. M. Özgökmen, 2005：多层原始方程海洋模型中的拉格朗日数据同化。《大气与海洋技术杂志》，https://doi.org/10.1175/JTECH-1686.1。Molcard, A., L. I. Piterbarg, A. Griffa, T. M. Özgökmen, and A. J. Mariano, 2003：同化漂流浮标观测以重建欧拉环流场。《地球物理研究杂志：海洋》，108（C3），https://doi.org/10.1029/2001JC001240。Penny, S. G., T. A. Smith, T.-C. Chen, J. A. Platt, H.-Y. Lin, M. Goodliff, and H. D. I. Abarbanel, 2022：将循环神经网络与数据同化相结合，实现可扩展的数据驱动状态估计。《地球系统建模进展》，14（3），e2021MS002 843，https://doi.org/10.1029/2021MS002843。Perez, E., F. Strub, H. d. Vries, V. Dumoulin, and A. Courville, 2017：FiLM：利用通用条件层进行视觉推理。《arXiv》，网址：http://arxiv.org/abs/1709.07871，arXiv:1709.07871 [cs]，https://doi.org/10.48550/arXiv.1709.07871。Qi, D., and A. J. Majda, 2016：统计响应与不确定性量化的低维降阶模型：双层斜压湍流。《大气科学杂志》，https://doi.org/10.1175/JAS-D-16-0192.1。Qu, Y., J. Nathaniel, S. Li, and P. Gentine, 2024：多模态场景下的深度生成式数据同化。《arXiv》，网址：http://arxiv.org/abs/2404.06665，arXiv:2404.06665 [cs]。Raissi, M., P. Perdikaris, and G. E. Karniadakis, 2019：物理信息神经网络：一种用于求解涉及非线性偏微分方程的正问题与逆问题的深度学习框架。《计算物理学杂志》，378，686–707，https://doi.org/10.1016/j.jcp.2018.10.045。Schuster, M., and K. Paliwal, 1997：双向循环神经网络。《IEEE信号处理汇刊》，45（11），2673–2681，https://doi.org/10.1109/78.650093。Slivinski, L., E. Spiller, A. Apte, and B. Sandstede, 2015：用于拉格朗日数据同化的混合粒子—集合卡尔曼滤波器。《月刊天气评论》，https://doi.org/10.1175/MWR-D-14-00051.1。Spantini, A., R. Baptista, and Y. Marzouk, 2022：非线性集合滤波的耦合技术。《arXiv》，网址：http://arxiv.org/abs/1907.00389，arXiv:1907.00389 [stat], https://doi.org/10.48550/arXiv.1907.00389。Stamatelopoulos, S., and T. P. Sapsis, 2025：扩散模型能否捕捉极端事件统计？《计算力学与工程方法》，435，117 589，https://doi.org/10.1016/j.cma.2024.117589。

Sun, L., 和 S. G. Penny，2019年：使用局地集合变换卡尔曼滤波器对双环流海洋模型中的表层漂流浮标进行拉格朗日数据同化。《月刊天气评论》，https://doi.org/10.1175/ MWR-D-18-0406.1。Tancik, M., 及合作者，2020年：傅里叶特征使网络能够在低维域中学习高频函数。arXiv，网址http://arxiv.org/abs/2006.10739，arXiv:2006。10739 [cs], https://doi.org/ 10.48550/arXiv.2006.10739。Vallis, G. K.，2017年：《大气与海洋流体动力学：基础与大尺度环流》。剑桥大学出版社，网址
https://www.cambridge.org/core/books/atmospheric-and-oceanic-fluid-dynamics/41379BDDC4257CBE11143C466F6428A4，iSBN：9781107065505 9781107588417，https://doi.org/10.1017/9781107588417。van Leeuwen, P. J., H. R. Künsch, L. Nerger, R. Potthast, 和 S. Reich，2019年：高维地球科学应用中的粒子滤波器：综述。《英国皇家气象学会季刊》，**145 (723)**，2335– 2365，https://doi.org/10.1002/qj.3551。Wang, Z., N. Chen, 和 D. Qi，2025：一种用于多层流场的闭式近似非线性数据同化算法。《月刊天气评论》，https://doi.org/10.1175/MWR-D-24-0277.1。Yang, S., 及合作者，2025：面向天气与气候的生成式同化与预测。arXiv，网址http://arxiv.org/abs/2503.03038，arXiv:2503。03038 [cs],https://doi.org/10.48550/arXiv.2503.03038。