arXiv:2304.02966v1 [physics.flu-dyn] 6 Apr 2023

湍流对流中大尺度状态之间的集体变量

Priyanka Maity, [1] Andreas Bittracher, [2] P´eter Koltai, [3,2] 以及 J¨org Schumacher[1,4]

[1]*Department of Mechanical Engineering, Technische Universität Ilmenau, Postfach 100565, D-98684 Ilmenau, Germany* [2]*Department of Mathematics, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin, Germany*
[3]*Department of Mathematics, University of Bayreuth, Universitätsstraße 30, D-95440 Bayreuth, Germany*
[4]*Tandon School of Engineering, New York University, New York, NY 11201, USA*（日期：2023年4月7日）

受限湍流对流流动中的动力学由多个长寿命的宏观环流状态主导，系统随后以马尔可夫型跳跃过程访问这些状态。在本研究中，我们通过一种数据驱动的学习算法，分析这些相继宏观系统状态之间的短过渡路径。该算法提取出低维转移流形以及相关的新坐标；我们将这些坐标称为复杂湍流流动状态空间中的集体变量。因此，我们将随机微观系统（例如大分子动力学）中构象跃迁的概念迁移并拓展到确定性的宏观流动中。我们的分析基于湍流对流的长期直接数值模拟轨迹，研究对象为封闭立方体单元中的流动，其普朗特数为 $Pr = 0.7$，瑞利数为$Ra = 10^6$ 和 10^7，时间滞后为 10^5 个对流自由落体时间单位。模拟解析了所有具有物理相关性的尺度上的涡旋和羽流，从而得到一个由超过350万个自由度张成的状态空间。通过仅使用两个集体变量，转移流形分析即可捕捉大尺度环流状态之间的跃迁动力学，这意味着数据维度降低了超过一百万倍。我们的方法表明，在当前设置下，大尺度流动的停滞及随后发生的反转并不常见，从而为开发宏观复杂非线性动力系统的高效降阶模型铺平了道路。

关键词：非线性动力学、流体动力学、湍流对流

I. 引言

复杂的非线性系统通常涵盖数量级跨度极大的相关动力学尺度。例子始于微观的随机层面：蛋白质大分子在某种构象中保持数毫秒后，会在数纳秒内切换到另一种构型，从而导致其化学功能发生显著变化 [1, 2]。而在宏观的确定性层面，受限几何结构中的湍流，或延展层中的湍流，可以呈现不同有序程度的大尺度空间图样；在长期演化过程中，这些图样会经历更长时间的瞬态阶段 [3–6]。从一种构型到另一种构型的（快速）转换，是由次级流动结构、较小涡旋、剪切层或羽流的涨落触发的，而这些结构能够影响热量或动量的湍流输运 [7–13]。宏观流动的状态空间或相空间是无限维的——或者至少具有极高维数——因此需要进行大幅度降维，才能有效地对观测到的大尺度动力学进行建模 [14]。

本研究将一种无监督数据约简策略——转移流形框架——从随机分子动力学 [15–17]迁移到确定性湍流中。具体而言，我们将分析一种湍流热对流流动，也称为瑞利–贝´纳德对流（RBC）流动。该流动由浮力驱动，受限于一个封闭的立方体单元 [18–20]。尽管其动力学和物理空间几何形态都很简单，这种构型仍是自然界和技术领域诸多应用的典型范例。当前的湍流流动与前述微观示例有若干相似之处。它会呈现不同的构型，此处表现为不同的大尺度环流（LSC）状态 [21–26] ，这些状态在相空间中占据不同区域，并维持许多个对流时间单位，随后迅速在相空间中彼此转换。一方面，可以认为这些LSC状态对应于相空间中速度场和温度场高度相似的状态，因此可将它们视为形成了集中的簇。另一方面，LSC状态之间的暂态动力学可能任意复杂，因而构成了该系统降阶建模的一个主要障碍——甚至可以说是首要障碍。我们在此关注的正是这种暂态动力学，更确切地说，是不同LSC状态之间的演变过程。具体而言，我们将提供基于数据的证据，表明这一过程具有低维性。

封闭立方体单元中的 LSC 呈现为四种沿对角线 [27–31] 排列、充满整个单元的离散环流滚动状态。近期研究表明，这种从一种长寿命 LSC（LL-LSC）状态跃迁至另一种状态的过程，会经过短寿命大尺度环流（SL-LSC）状态；基于对 LL-LSC 寿命和跃迁概率的分析，该过程可以近似描述为马尔可夫过程 [32]。

对长期轨迹中快速跃迁事件的分析，将得到一组最小的新坐标，用以张成低维替代状态空间，足以表征

该系统的统计动力学行为。随后，这些新坐标所给出的描述将与物理空间中湍流RBC流的动力过程联系起来；具体而言，局部角涡在LSC旁间歇性地相互作用，并短暂增强，进而将对角环流卷推入下一个宏观态。因此，在这一特定应用中，我们的方法为构建LSC动力学的高效降阶动力学模型打开了大门，即将其表示为替代状态空间上的马尔可夫过程。

这些新坐标在原始化学语境中通常称为反应坐标 [16], ，我们将在此应用中使用集体变量这一概念。湍流数据源自一段长期模拟轨迹，包含 10^5 个对流时间单位 t_f ，对应两个不同的瑞利数 $Ra = 10^6$, 10_7 以及一个普朗特数 $Pr = 0.7$。关于热对流的瑞利–贝纳德流动模型，以及 Ra和 Pr的精确定义，请参见下一节。考虑到模拟中所涉及的 4 个湍流场的谱分辨率，我们统计得到描述湍流对流流动的自由度超过 350 万个。

近年来，利用（深度）神经网络寻找替代动力学模型，尤其是寻找用于表达这些模型的低维变量的方法大量涌现 [33–38]。这些方法的成功很大程度上依赖于此类网络表示来自广泛通用类别的坐标的能力。然而，在这些方法中，使其表现良好所必需的动力学条件仍是隐含的。本文提出的方法依赖于非常明确的动力学假设，并在数据驱动的计算过程中对这些假设进行验证。

II. 湍流瑞利–B ´ENARD流

A. 模型方程与模拟方法

我们模拟表示瑞利–贝纳德对流动力学的无量纲三维不可压缩 Boussinesq 运动方程。其形式如下:

$$\partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\nabla p + \sqrt{\frac{Pr}{Ra}}\nabla^2\mathbf{u} + T\hat{z}\,, \tag{1}$$

$$\partial_t T + (\mathbf{u}\cdot\nabla)T = \frac{1}{\sqrt{RaPr}}\nabla^2 T\,, \tag{2}$$

$$\nabla\cdot\mathbf{u} = 0\,. \tag{3}$$

这里，$\mathbf{u}(\mathbf{x}\ t)$ 是流体的速度场， $T(\mathbf{x}\ t)$ 是温度场，而 $p(\mathbf{x}\ t)$是压力场。流动的无量纲控制参数是瑞利数 Ra，它衡量浮力驱动的强度；以及普朗特数 Pr，它是动量扩散率与热扩散率之比。两个数分别为

$$Ra = \frac{\alpha g\,\delta T d^3}{\nu\kappa} \quad \text{and} \quad Pr = \frac{\nu}{\kappa}\,, \tag{4}$$

其中， α 表示等压热膨胀系数， g 表示重力加速度， $\delta T = T_{\text{bottom}} - T_{\text{top}}$表示沿厚度为 d的流体层维持的温度差， ν 和 κ 分别表示流体的运动黏度和热扩散率。通过以单元高度 d （此处等于两个水平边长）重新缩放长度、以自由落体速度$U_f = \sqrt{\alpha g \delta T d}$重新缩放速度，以及以外部温差 δT重新缩放温度，对方程进行了无量纲化。由此得到的无量纲时间单位为自由落体时间尺度 $t_f = \sqrt{d/}$（$\alpha g \delta T$），这是本研究中的大尺度对流时间单位。

我们假设立方体单元 $V = d^3$ 的六个面均满足无滑移速度边界条件$\mathbf{u} = 0$。系统从底部均匀加热，条件为 $z = 0$ with $T = T_{\text{bottom}}$ ，并从顶部冷却，条件为 $z = 1$ with $T = T_{\text{top}}$。我们还假设侧面绝热，$\mathbf{n}\cdot\nabla T = 0$，因此底部输入的所有热量都必须经由流体传递至顶部。

模拟采用基于谱元方法的开源代码 Nek5000（版本17）执行 [39, 40]。模拟假定沿每个空间方向使用16个谱元，并在每个空间方向及每个谱元上采用5阶（对于 $Ra = 10^6$）和7阶（对于 $Ra = 10^7$）的拉格朗日插值多项式，分别得到884,736个和2,097,152个配置点。通过分析平均动能耗散率的垂直剖面，验证该谱分辨率对于分析大尺度环流而言是否足够充分。模拟针对流体的两个瑞利数值 $Ra = 10^6$ 和 10^7 进行，其中普朗特数为 $Pr = 0.7$，分别对应有效雷诺数$Re = \sqrt{Ra/Pr}\,u_{\text{rms}}$为467和1350；其中， u_{rms}表示均方根速度，见表I。

$(Pr,\ Ra)$	$(0.7, 10_6)$	$(0.7, 10_7)$
Re	467	1350
$u_{\rm rms} = \langle u_i^2 \rangle_{V,t}^{1/2}$	0.39	0.36
$\omega_{\rm rms} = \langle \omega_i^2 \rangle_{V,t}^{1/2}$	2.86	3.83
$\theta_{\rm rms} = \langle \theta^2 \rangle_{V,t}^{1/2}$	0.29	0.29
$t_{\rm pers}$ (SL-LSC)	$4.2t_f$	$8.6t_f$
$t_{\rm pers}$ (LL-LSC)	$12.1t_f$	$26t_f$

表 I：基本参数和湍流量汇总。比较在瑞利数 Ra and普朗特数 Pr的各种组合下获得的速度（$u_{\rm rms}$）、涡量（$\omega_{\rm rms}$）和温度（$\theta_{\rm rms}$）的均方根值。此处，$\langle \cdot \rangle_{V,\ t}$是对立方体积和时间的综合平均值。长寿命大尺度环流（LL-LSC）和短寿命大尺度环流（SL-LSC）的平均持续时间（$t_{\rm pers}$）在最后两行中以对流自由落体时间单位 t_f列出。它们由将指数定律拟合到持续时间分布所得。

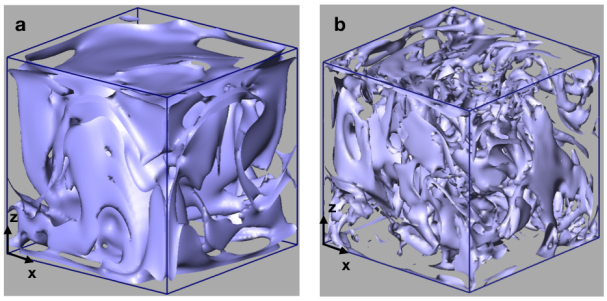

图1：流体 Pr $= 0.7$ 在瑞利数为 (a) $Ra = 10^6$ 和 (b) $Ra = 10^7$ 时绘制的涡量幅值等值面。图中展示了对流流动在两个时刻的状态。等值面绘制在数值 $|\omega| = 1.5$ 处，该数值大致对应于两种流动中最可能出现的涡量幅值。

从随机初始条件开始，我们等待了 5000 个自由落体时间，让系统达到稳态。随后让轨迹再演化 10^5 个自由落体时间，并每隔 t_f 个自由落体时间输出一次湍流场，从而为这里考察的每种情况收集 $N_s = 10^5$ 个完整流场快照。

B. 速度和涡量统计信息

表 I 列出了两次运行中系统的速度、涡量和温度涨落。这些涨落是相对于流场体积 V 和时间 t 计算的均方根。我们的数据分析发现，在比较两次长期运行（分别在 $Ra = 10^6$ 和 10^7 以及 $Pr = 0.7$）时，尽管对流流动中的涡量涨落水平更高，但观测到长期存在的 LSC 状态之间的切换频率较低。如图所示，速度、 $u_{\rm rms}$ 和温度场 $\theta_{\rm rms}$ 的均方根值几乎保持不变。

较高瑞利数下涡量涨落的幅值更大 $\omega_{\rm rms}$ ，这意味着在较高的 Ra下，涡旋结构更加碎片化且更薄；而在较低的瑞利数下，我们观察到的涡旋结构则更为连贯。图 1 中的涡量等值面图证实了这一点。在 $Ra = 10^7$ 的情况下，缺少空间尺度较大的连贯涡旋结构，这或许可以解释为何 LSC 结构在某一特定宏观态中的寿命更长。我们的观察结果与此前针对纵横比相同的圆柱形容器开展的研究一致 [41, 42]。根据文献 [42], , 在 $Pr = 0.7$ 下，湍流对流中的平均动能耗散率随瑞利数的变化遵循标度律

$$\langle \epsilon \rangle_{V,t} \sim Ra^{-0.2} \tag{5}$$

动能耗散率场

$$\epsilon(\mathbf{x},t)=\sqrt{\frac{Pr}{4Ra}}\left[\boldsymbol{\nabla}u+(\boldsymbol{\nabla}u)^T\right]^2 . \tag{6}$$

平均动能耗散率反过来与涡量平方的均值近似成正比，这一点已在 [41, 43]中发现。这意味着

$$\langle\epsilon\rangle_{V,t}\approx\sqrt{\frac{Pr}{Ra}}\omega_{\rm rms}^2\sim Ra^{-0.2}\,, \tag{7}$$

因此，按照 $\omega_{\rm rms}\sim Ra^{0.15}$，其随 Ra 增长。请注意，在均匀各向同性湍流中，这是一个精确等式；比例常数为运动黏度。

III. 大尺度环流

A. 大尺度环流状态的检测

下面我们讨论如何从数值模拟数据中获得大尺度流动状态。在此分析中，将标量场和矢量场通过谱插值映射到均匀笛卡尔网格上。LSC取向角随时间的演化可视化与计算沿用了 [32]中的方法。随后，我们将垂直速度分量的均匀网格数据插值到半径固定为 $r=0.45$ 的圆上，角度 θ 以 5°为步长变化。角度 θ 以顺时针方向相对于 y轴测量。然后，我们对插值到该圆上的垂直速度分量数据进行方位角傅里叶变换。LSC取向角 θ 的识别取决于两个量：（i）最大傅里叶模态所携带的能量比，其表示为

$$\delta(t)=\frac{\max_{k_\theta}|\hat{u}_z(k_\theta,t)|^2}{\sum_{k_\theta}|\hat{u}_z(k_\theta,t)|^2}\,, \tag{8}$$

以及（ii）最大傅里叶模态的相位。在具有显著LSC结构的时间区间内，最大能量将由该特定傅里叶模态拥有，并且我们将得到较大的 $\delta(t)$值。LSC结构的方向对应于最大傅里叶模态的相位。在不存在任何LSC结构时，动能将分布于所有模态之间，并且我们将得到较低的 $\delta(t)$值。非相干态极为罕见且不稳定。为了系统地识别大尺度环流，我们首先计算了 δ的概率密度函数（PDF）的标准差 σ。所有低于 $\mu-2\sigma$ 的流动状态（其中 μ为PDF的均值），因而具有极低 δ 值的状态，都被识别为非相干态。所有其余值大于 $\mu+2\sigma$的状态，则根据对应于最大傅里叶模态相位的方向角 θ，被归类为四种不同LSC结构中的一种。

我们根据傅里叶模态所含的能量，启发式地识别出对流流动中的六种不同离散大尺度环流状态。请注意，这种分类很大程度上依赖于我们对物理空间几何形状以及已观测到的大尺度环流构型的先验认知。我们发现，系统主要偏好沿对角线排列的长寿命大尺度环流（LL-LSC）状态，分别记为 $S^{\pi/4}$、$S^{3\pi/4}$、$S^{5\pi/4}$ 和 $S^{7\pi/4}$。LL-LSC 状态之间的跃迁经由短寿命大尺度环流（SL-LSC）状态 $S^{n\pi/2}$（沿侧面排列）以及非相干状态 S_0 发生。我们每隔一个对流自由落体时间监测一次大尺度状态，并从两次模拟运行中各收集了 10^5 个完整的三维对流流场快照。图2（a、b）展示了两次运行中大尺度环流方向 θ 的长期轨迹演化。值得注意的是，$Ra=10^7$ 情况下的跃迁频率低于 $Ra=10^6$。准二维流动中也观察到了这一现象 [44]。较低的角度涨落水平使得底部面板看起来明显包含更少的数据点；但两种情况下的数据点数均为 $N_s=10^5$ 。

图2（c–j）详细展示了两个长寿命 LSC 状态之间的一次转换，作为典型示例：系统在相对较短的18个对流自由落体时间内，从长寿命状态$S^{5\pi/4}$ 经由短寿命状态$S^{n\pi/2}$ 和非相干零状态S_0 ，切换到长寿命状态$S^{3\pi/4}$ 。这大约相当于一个流体微团沿大尺度环流涡卷循环一周所需的时间；因此，与总时间滞后相比，这一过程可视为快速过程。图中间一行展示了从侧面观察的三维流线视图，并叠加了高涡量幅值的等值面（灰色）。我们清楚地看到，这些强烈的涡量结构始终存在于大尺度环流平均流的两侧，见（c）–（f）图。因此，涡量幅值的不平衡应是导致大尺度环流涡卷失稳的原因。高涡量区域对大尺度

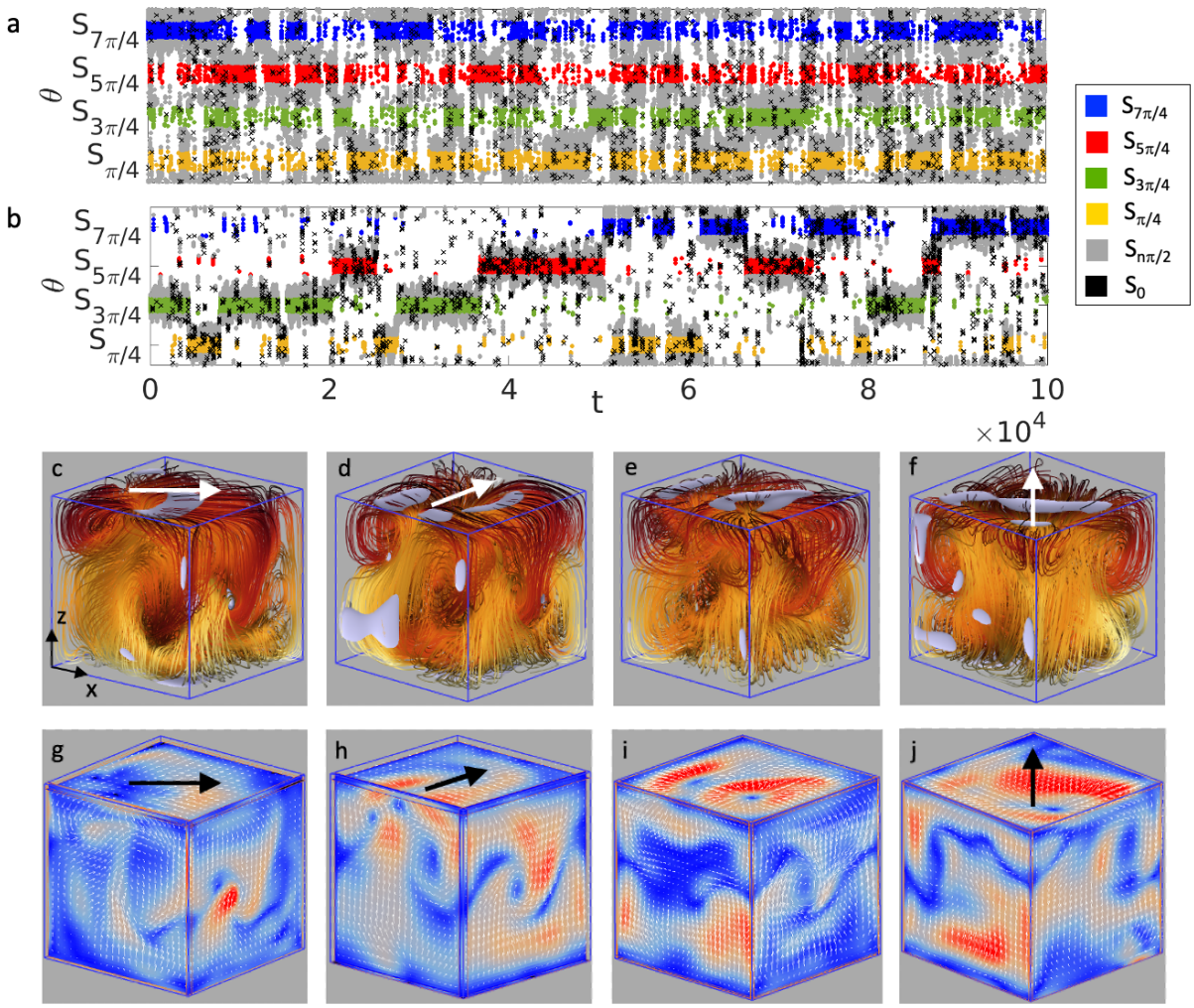

图2：立方体容器中对流流动的大尺度流动行为。普朗特数为 $Pr = 0.7$ 、瑞利数为两种不同取值时，大尺度环流（LSC）的取向角 θ 随时间的演变。（a）Ra $= 10^6$。（b）Ra $= 10^7$。两个面板中的图例均标示了六种不同的大尺度环流状态。（c–j）不同大尺度环流状态之间快速转变的一个示例的体积可视化。面板（c）–（f）为速度流线图，展示转变过程中各大尺度环流状态的演变：从面板（c，g）中的长寿命大尺度环流状态$S^{5\pi/4}$，转变为（d，h）中的中间短寿命大尺度环流状态$S^{n\pi/2}$，再经由（e，i）中的非相干零状态S_0，最终转变为（f，j）中的新长寿命大尺度环流状态$S^{3\pi/4}$。速度流线按该时刻的温度着色，其中黄色表示最高温度，黑色表示最低温度（$0 \leq T \leq 1$）。除零状态外，白色箭头表示大尺度环流的流动方向。图中还加入了涡量幅值为 $|\boldsymbol{\omega}| \geq 7$ 的等值面。面板（g）–（j）绘出了对应的速度矢量，其中蓝色和红色等值线分别表示最小涡量幅值 $|\boldsymbol{\omega}| = 0$和最大涡量幅值 $|\boldsymbol{\omega}| = 8$。黑色箭头再次表示大尺度环流的对齐方向。

流动停止的重要性已在准二维情形中得到认识 [44]；然而，在三维情形下，角涡的构成有所不同。图底部一行（g）–（j）图中涡量幅值等值线的可视化结果表明，高涡量区域是上升羽流卷曲臂的一部分。大尺度环流涡卷的对齐方向由系统中三维上升和下降羽流之间的相互作用决定，例如，图2（d，h）中左侧前表面中心的上升羽流和右侧前表面的下降羽流。我们还在第二节B中表明，涡量涨落的幅值随瑞利数增加而增大，涡旋结构也变得更加丝状，见图1和表I。该示例还表明，复杂的全三维模式融入了切换过程，例如S_0。要构建描述大尺度流动行为的降阶模型，必须用一组最小坐标来描述这种动力学。

B. 大尺度环流状态之间的跃迁

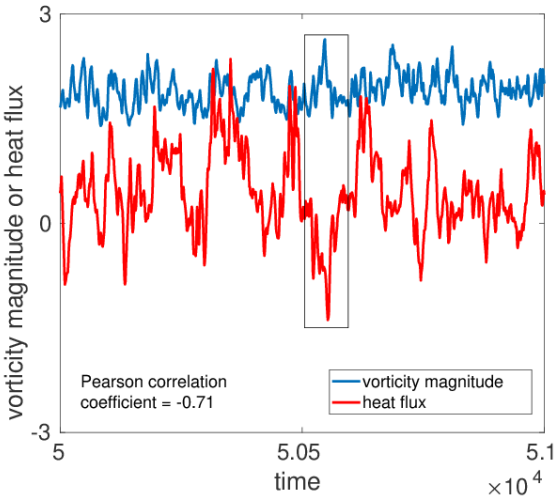

图 3：通过比较沿 LSC 对齐方向计算的子体积平均对流热通量 $\langle u_z T(t)\rangle_s$ 与在强烈涡旋结构所在的对角子体积中计算的子体积平均涡量幅值 $\langle|\boldsymbol{\omega}|(t)\rangle_s$，监测 LSC 状态 $S_{5\pi/4}$ 到 $S_{7\pi/4}$ 经由短暂状态发生的跃迁，其中 $Ra = 10^7$。为便于与涡量幅值进行视觉比较，对流热通量还乘以 50 的乘法因子进行放大。长寿命 LSC 的失稳发生在涡量达到峰值且热通量降至最低时。图中框选的 100 个自由落体时间窗口内发生了这样一次失稳事件。此窗口内两个时间信号的皮尔逊相关系数为 -0.71，凸显了跃迁附近的强负相关。

我们的观察表明，在转变为另一种长寿命 LSC 状态之前，长寿命 LSC 状态总会先失稳为 $S_{n\pi/2}$（短寿命状态）或 S_0（非相干状态）。为了证实我们关于环流方向垂直平面四角处的相干涡旋使 LSC 失稳的机制这一假设，我们比较了沿流动对齐方向的热通量与触发失稳的涡量幅值。为此，我们提取了以立方体模拟区域八个角为中心的子体积 $V_s = (d/4)^3$，。随后，我们计算了八个 V_s 各自内部的涡量中心，定义如下：

$$\mathbf{r}_s(t) = \frac{\int_{V_s} |\boldsymbol{\omega}(\mathbf{x},t)|\mathbf{x}\, dV}{\int_{V_s} |\boldsymbol{\omega}(\mathbf{x},t)|\, dV}\,. \tag{9}$$

在这八个涡量中心$\mathbf{r}_s(t)$处（它们是 V_s中高涡量区域的“质心”），我们以其为中心，选取另一个更小的立方体子体积 V_c，其中包含 8^3 个网格点，用以计算以下时间序列

$$\langle|\boldsymbol{\omega}|(t)\rangle_s = \frac{1}{V_c}\int_{V_c} |\boldsymbol{\omega}(\mathbf{x}+\mathbf{r}_s(t),t)|\, dV\,. \tag{10}$$

对流热通量也采用类似的处理方法，

$$\langle u_z T(t)\rangle_s = \frac{1}{V_c}\int_{V_c} u_z(\mathbf{x}+\mathbf{r}_s(t),t) T(\mathbf{x}+\mathbf{r}_s(t),t)\, dV\,. \tag{11}$$

最后，我们对穿过立方体中两个垂直对角平面的四个 $\langle|\boldsymbol{\omega}|(t)\rangle_s$ 和 $\langle u_z T(t)\rangle_s$ 取算术平均 $(\cdot)$。因此，现在对于这两个物理量中的每一个，我们都有两条时间序列。当长寿命的 LSC 沿对角线排列时（表示 $S_{7\pi/4}$ 或 $S_{3\pi/4}$ 沿一个对角平面排列，或 $S_{5\pi/4}$ 或 $S_{\pi/4}$ 沿另一个对角平面对齐），沿该对角平面的热通量达到最大值。在 LSC 的失稳阶段，位于垂直于 LSC 平面的平面内的涡旋变得强烈，相应的涡量幅值时间序列达到峰值。同时，沿 LSC 方向的热通量下降。这在图 3 中较短的 $1000t_f$时间窗口内有所展示。系统从 $S_{5\pi/4}$ 状态过渡到 $S_{7\pi/4}$ 状态。需要特别注意的是，该过渡并不平滑。在过渡到 $S_{7\pi/4}$ 状态之前，系统会在 $S_{5\pi/4}$ 状态和短寿命的（$S_{n\pi/2}$ 或非相干状态）之间反复切换。因此，图 3 捕捉到了许多这样的失稳 ന๐ɐ

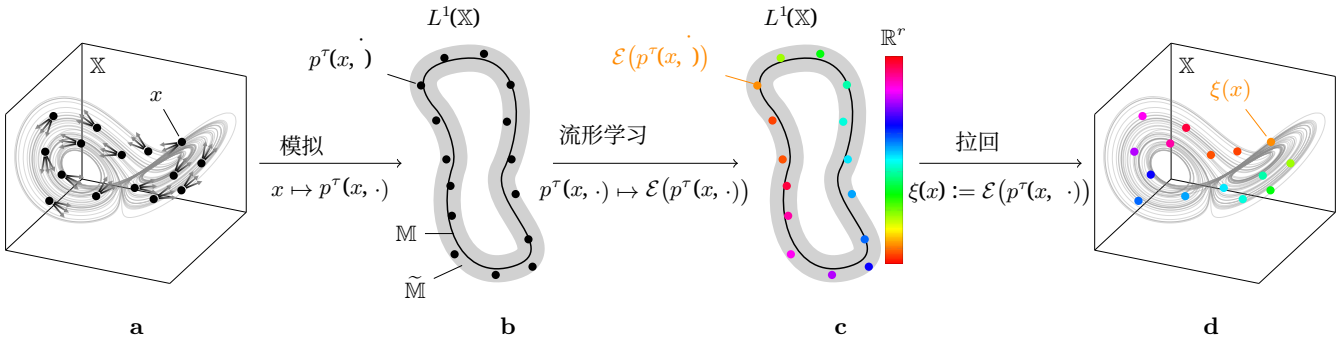

图4：噪声系统的转移流形框架示意图。(a) 包含初始状态 x的状态空间。图中从各点发出的多条箭头强调了系统的概率性质。(b) 初始状态通过其转移密度映射到 $\widetilde{\mathbb{M}} \subset L^1$，它们可能聚集在低维流形 $\mathbb{M}$附近。(c)找到了 $\widetilde{\mathbb{M}}$ 的低维参数化，如颜色所示。(d) 该参数化被平凡地拉回到初始状态集合，以定义集体变量 ξ。

事件。在图3中，矩形盒内观测到了一个此类失稳事件，其持续时间为 $100t_f$ 自由落体时间；在该事件中，系统从长寿命 $S^{5\pi/4}$ 至短寿命$S^{n\pi/2}$ 状态发生跃迁。当涡量幅值在垂直于LSC对齐方向的方向上达到峰值时，可以清楚地观察到沿LSC对齐方向的热通量下降。在此时间窗口内，所得皮尔逊相关系数为 -0.71。我们计算了许多此类失稳事件中两个信号之间的皮尔逊相关系数，并且始终得到负相关系数（未显示）。在跃迁事件之外，即长寿命大尺度环流稳定时，皮尔逊相关系数或为正，或略微为负。这证明了我们的命题：位于LSC环流方向垂直方向上的强烈涡旋结构触发了失稳机制，使长寿命大尺度环流发生失稳，并经由短寿命或非相干状态跃迁至另一个长寿命大尺度环流状态。

IV. 通过转移流形框架进行数据降维

转移流形框架（TMF）能够识别低维但可能是多维的集体变量，即完整状态空间中的可观测量，用于刻画关键的动力学现象。该算法是无监督且数据驱动的，也就是说，它不需要任何关于湍流流动的物理知识。尽管该框架最初是为随机过程设计的，我们将展示如何将其应用于混沌系统（此处采用通常意义，而非严格的数学意义），但确定性系统，例如宏观尺度的RBC流。我们已经看到，在最大尺度上，该流动表现得像一个“无记忆”的随机过程，参见持续时间的指数分布。这表明，在两者之间存在一个粗粒化层级，在该层级上，系统可以被良好地描述为随机过程，从而适用于本框架。

下文中，变量 x 表示对流流动的一个状态，即高维状态空间 $\mathbb{X}$，而不是像前几节那样表示位置向量的 $\mathbf{x} \in \mathbb{R}^3$。

A. 随机系统的原始框架

现在我们简要列出一些基本的数学关系，以使本文内容自成一体。考虑一个定义在有界状态空间 $\mathbb{X} \subset \mathbb{R}^N$（初始条件可变，为 X_0）上、时间和空间连续且齐次的马尔可夫过程 $\{X_t\}_{t\geq 0}$（或简称 $\{X_t\}$），以及底层概率空间 $(\Omega\ \mathcal{F},\ \mathrm{Prob})$。假设 $\{X_t\}$ 是遍历的，并具有唯一的不变密度 $\pi : \mathbb{X} \to \mathbb{R}^+$。我们还将 π 表示为由 π 诱导的概率测度，并用 Leb 表示勒贝格测度。在对 $\{X_t\}$ 的正则性作出较弱假设的条件下，该过程的统计性质由一族转移密度函数 $\{p^t\}_{t\geq 0} \subset L^1_{\pi\times \mathrm{Leb}}(\mathbb{X}\times\mathbb{X})$刻画，其含义是

$$\mathrm{Prob}[\,X_t \in A \mid X_0 = x\,] = \int_A p^t(x,y)\,dy. \tag{12}$$

换言之，$p^t(x,\ \cdot)$ 是以 $X_0 \sim \delta_x$ 为起始分布的 X_t 分布。写成条件分布的形式，即 $(X_t \mid X_0 = x) \sim p^t(x,\ \cdot)$。

现在考虑某个候选集体变量 $\xi:\ \mathbb{X}\to\mathbb{R}^r$, $r\ll N$，并令 $\mathbb{Y}:=\xi(\mathbb{X})$。由于我们关注 RBC 流中的系统演化最缓慢的 LSC 现象和子过程，因此可以非正式地称 ξ 为一个“良好”的集体变量，条件是投影到 ξ 后能保留系统的长期统计特性。也就是说，当 t 大于某个阈值时间尺度 τ 时，$\{\xi(X_t)\}$ 的统计特性将在某种意义上与 $\{X_t\}$ 相似。在上文引入的形式化框架中，这意味着

$$\|p^t(x,\cdot)-\bar{p}^t(\xi(x),\cdot)\|_{L^1}\le\varepsilon\quad\forall x\in\mathbb{X},t\ge\tau, \tag{13}$$

对于某个阈值滞后时间 $\tau>0$、某族关于其第一个坐标光滑的函数 $\{\bar{p}^t(\cdot,\cdot)\}_{t\ge0}\subset L^1_{\bar{\pi}\times\mathrm{Leb}}(\mathbb{Y}\times\mathbb{X})$，以及某个较小的 $\varepsilon>0$。这里，$\bar{\pi}$ 表示测度 π 经 ξ 推前后的测度。我们称满足式 (13) 的集体变量为 ε-一致。

注意，式（13）表明转移密度的集合AV视频

$$\widetilde{\mathbb{M}}:=\{p^\tau(x,\cdot)\mid x\in\mathbb{X}\}\subset L^1(\mathbb{X}) \tag{14}$$

在 ε 附近紧密地积聚于 r 维流形 $\mathbb{M}\subset L^1(\mathbb{X})$ 上，

$$\mathbb{M}:=\{\bar{p}^\tau(z,\cdot)\mid z\in\mathbb{Y}\}, \tag{15}$$

转移流形。我们强调，转移流形并不位于状态空间中，因此不应将其视为连接状态空间某些部分的低维流形；异宿轨道或流形才具有这种作用。

反过来，可以从转移流形的参数化构造出与 ε 一致的集体变量。令 $\sim$ 对于此 $\mathcal{E}:\ \mathbb{M}\to\mathbb{R}^r$ 是 $\mathbb{M}$ 的任意参数化，即 $\mathbb{M}$ 与其像之间的一一映射，并且 $\mathcal{Q}:\ \mathbb{M}\to\mathbb{M}$ 是具有该性质的任意映射。

$$\left\|p^\tau(x,\cdot)-\mathcal{Q}\big(p^\tau(x,\cdot)\big)\right\|_{L^1}\le\varepsilon. \tag{16}$$

在.....之下 根据 (13) 的假设，这样的映射存在，而且可以证明该集体变量】【。 e

$$\xi(x):=\mathcal{E}\left(\mathcal{Q}\left(p^\tau(x,\cdot)\right)\right) \tag{17}$$

确实满足 (13)。我们将 ξ 称为 跃迁流形集体变量。与转移流形的情况一样，ξ 不一定参数化状态空间中每条轨迹的时间演化，而是以集体方式参数化暂态动力学期间轨迹的进程。

如 [16], 所示，如果原过程 $\{X_t\}_{t\ge0}$是（随机）可逆的，则过程 $\{\xi(X_t)\}_{t\ge0}$ 具有与原过程 $\{X_t\}_{t\ge0}$的主导时间尺度（相关性衰减率） $\mathcal{O}(\varepsilon)$-接近的主导时间尺度。这量化了集体变量 ξ 的“优良”程度，并为一种构造性方法铺平了道路，该方法可以找出转移流形的此类集体变量；参见图4。关于集体变量计算的算法实现的详细讨论可见附录B。

B. 确定性系统中集体变量的基于集合的近似

对于映射 Φ^τ 所定义的确定性流，转移“密度” $p^\tau(x,\ \cdot)$ 是狄拉克分布 $\delta_{\Phi^\tau(x)}(\cdot)$，因此不属于 $L^1(\mathbb{X})$。然而，复杂的混沌系统在许多方面表现得如同真正的随机系统，其分析具有概率或统计性质。这促使我们将 TMF 应用于确定性混沌系统，例如本文研究的系统。

上述问题的一种解决方案是引入小随机扰动 [45]的概念。直观地说，转移密度（或一般情况下的测度）族 $(p_\varepsilon)_{\varepsilon>0}$ 是映射 Φ 的小随机扰动，当测度空间中的 $p_\varepsilon(x,\ \cdot)\xrightarrow{w}\delta_{\Phi(x)}$ 在 $\varepsilon\to0$ 时于弱意义下收敛。随着扰动 [45, 46] 趋于零，这些小随机扰动会保留原系统的某些动力学性质。这是一种结构稳定性，我们现在假定它也适用于集体变量。我们的思路是利用具有 $L^1(\mathbb{X})$ 中转移密度的小随机扰动，将 TMF 推广到确定性系统。

我们提出在确定性流 Φ^τ 的集体变量计算流程中使用以下离散化。可以证明，该离散化会对原系统产生微小的随机扰动，其转移密度为 [47]。更确切地说，我们不再在单个点处计算跃迁流形集体变量 ξ，而是计算状态空间 $\mathbb{X}$ 的划分元素 $\{A_1,\dots,A_L\}$ 上 ξ 的平均值：

$$\xi(x)\rightsquigarrow\Xi(A):=\mathbb{E}[\,\xi(X)\mid X\sim\pi|_A\,]. \tag{18}$$

算法 4.1跃迁流形集体变量的集合式计算

输入： 具有 $y_i = \Phi^\tau x_i$ 的数据集对 $\mathcal{X} = \{x_1, \ldots, x_{N_s}\}, \mathcal{Y} = \{y_i, \ldots, y_{N_s}\}$。1: 选择 $\mathbb{X}$ 的一个划分 $\{A_1, \ldots, A_L\}$，使亚稳态区域和过渡区域得到均匀覆盖。2: 将数据集 $\mathcal{X}$ 排序到划分元素中，即对于每个 j，计算索引集

$$\mathcal{I}_j := \{i \mid x_i \in A_j\}.$$

集合

$$\mathcal{Y}_j = \{y_i \mid i \in \mathcal{I}_j\}$$

因此是 $p^\tau(A_j, \cdot)$ 的经验近似。
3: 近似距离矩阵 $D \in \mathbb{R}^{L\times L}$，

$$D_{ij} = d\left(p^\tau(A_i, \cdot), p^\tau(A_j, \cdot)\right)$$

由采样得到的 $p^\tau(A, \cdot)$。
4: 对 D 应用无监督距离型流形学习算法。
输出： 方法　　对转移流形集体变量 ξ 的近似，在集合 $\{A_1, \ldots, A_L\}$ 上取平均，即

$$\{\Xi(A_1), \ldots, \Xi(A_L)\}.$$

这将使我们在整个状态空间上，而不只是孤立点上，获得（近似的） ξ，同时还能够从长期连续轨迹数据中计算它，而不是进行模拟“突发段”；参见算法 4.1。

在我们的设定中，数据集对 $\mathcal{X}, \mathcal{Y}$将由一条很长的轨迹构成，该轨迹由

$$\mathcal{T} := \{x_0, \Phi^\tau x_0, \Phi^{2\tau} x_0, \ldots, \Phi^{N_s\tau} x_0\} \tag{19}$$

其中 Φ^τ 表示动力系统的流映射， N_s 是快照的数量，此处为 $N_s = 10^5$。我们使用

$$\mathcal{X} = \{x_0, \ldots, \Phi^{N_s-1} x_0\}, \tag{20}$$

$$\mathcal{Y} = \{\Phi^\tau x_0, \ldots, \Phi^{N_s\tau} x_0\}, \tag{21}$$

其中我们使用滞后时间 $\tau = 1t_f$。以此采样率输出的轨迹足以解析LSC环流涡卷中流体微团的翻转过程。

为了获得 $\mathbb{X}$ 的划分，我们采用基于距离的方法。由于这种方法的结果可能严重依赖于所用的距离，尤其是在维度非常高的赋范空间中，因此我们将通过辅助观测函数 ψ，对映射到低维空间中的轨迹$\mathcal{T} \in \mathbb{R}^d$ 进行划分，即映射到 $z_i = \psi(x_i)$。这种做法的可行性以及 ψ 的选择将在下一节讨论。随后，根据为 $\psi(\mathcal{T})$ 找到的划分，诱导出 $\mathcal{T}$ 的划分。为了划分集合 $z_1, \ldots, z_{N_s} \subset \mathbb{R}^d$，我们使用由 $\{z^{(1)}, \ldots, z^{(L)}\}$ 个中心点构成的 *Voronoi* 剖分，这些中心点依据某种合理规则选取。对 $\mathbb{X}$ 的Voronoi剖分由中心点 $\{x^{(1)}, \ldots, x^{(L)}\}$ 构成，是一组集合 $A_1, \ldots, A_L$，其中 $\mathbb{X} = \cup_{i=1}^L A_i$，且对于 $j \neq i$，有 $x \in A_i \Leftrightarrow \|x - x^{(i)}\| \leq \|x - x^{(j)}\|$。换言之，集合 A_i 由那些距 $x^{(i)}$ 比距任何其他中心点 $x^{(j)}$ 都更近的点组成。我们在此将 k-均值聚类算法应用于 $\psi(\mathcal{T})$，其中 $k = L$，并将得到的聚类质心用作 Voronoi 中心，正如 [48] 所提出的。已有研究 [17] 表明，这种方法能够利用中心点对数据点进行均匀覆盖。更多细节见附录B。

V. 瑞利–贝纳德流的集体变量

A. 通过时间滞后独立成分分析（时间滞后独立成分分析）进行预处理

RBC的流状态 $x \in \mathbb{X}$ 最初是无限维的，即使经过用于直接数值模拟的空间离散化，仍然具有数百万个维度。从原始计算网格进行后续降采样后，在当前情况下仍得到一个12288维的数据空间，另请参见下一小节。在如此高维的空间中表示一般分布是一项具有挑战性的任务，而事实证明，进一步使用某种粗粒度降维技术对数据进行预处理是有益的。请注意，TMF˜所依据的假设是，(14)中的集合 $\mathbb{M}$ 几乎是一个低维光滑流形。这意味着，几乎任何映射到一个足够但低维的空间的线性投影，都能保留其核心的流形结构 $\mathbb{M}$[70]

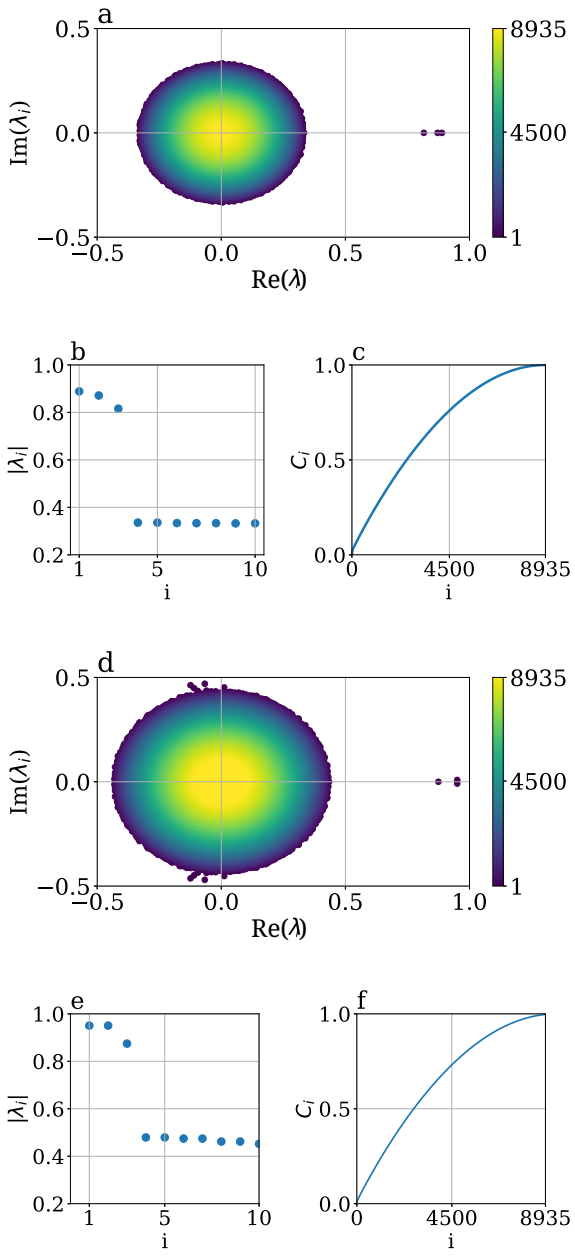

图 5：（a）运行 $Ra = 10^6$（a–c）和 $Ra = 10^7$（d–f）时TICA特征值的实部和虚部。（a、d）颜色对应于特征值索引。（b、e）前10 个TICA特征值的绝对值。可以清楚地看到谱隙。（c、f）累积动力学方差 c_i，表示通过投影到前 i TICA坐标上所保留的动力学方差部分。

我们将考虑分布 $p^\tau(x,\cdot)$、 $x \in \mathbb{X}$经过线性函数 ψ 的推前： $\mathbb{X} \to \mathbb{R}^d$具有适中的值域维数 d。更准确地说，当 $(X_\tau \mid X_0 = x) \sim p^\tau(x,\cdot)$ 时，我们将考虑 $(\psi(X_\tau) \mid X_0 = x)$ 在 $\mathbb{R}^d$ 上的分布。除此之外，这样做的优势在于，与 $p^\tau(x,\cdot)$相比，在相同样本数量下，对后一个分布进行采样更加准确，因为 $d \ll N$。

我们选择由 ψ 时间滞后独立成分分析 （TICA）构成的函数。简而言之，TICA 会确定状态的线性坐标，并按照所谓的“动力学方差”对其排序；关于这一数据分析算法的更多详情，请参见附录 A。通过这种方式，预期主导的TICA坐标能够捕捉大量动力学变异性，因此适合作为解析转移密度的可观测量。需要注意的是，不能指望它们能够参数化暂态动力学——对于一个高度非线性（湍流）系统而言，这确实是

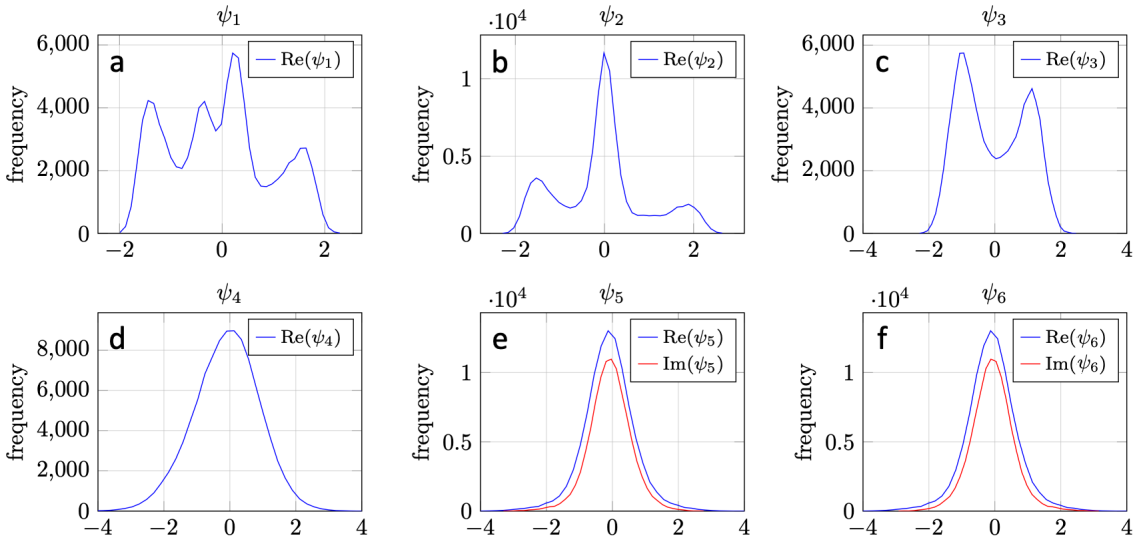

图 6：轨迹上前六个 TICA 坐标取值的直方图。前四个坐标完全为实数，因此面板 (a–d) 中仅绘制实部；对于面板 (e、f) 中第 5 和第 6 个主导坐标，则绘制其实部和虚部。我们观察到 ψ_1 至 ψ_3 中存在多峰分布，而较高坐标则近似服从正态分布（实数或复数）。还存在更高阶的纯实 TICA 坐标，其同样近似服从正态分布（未显示）。数据对应于 $Ra = 10^6$。

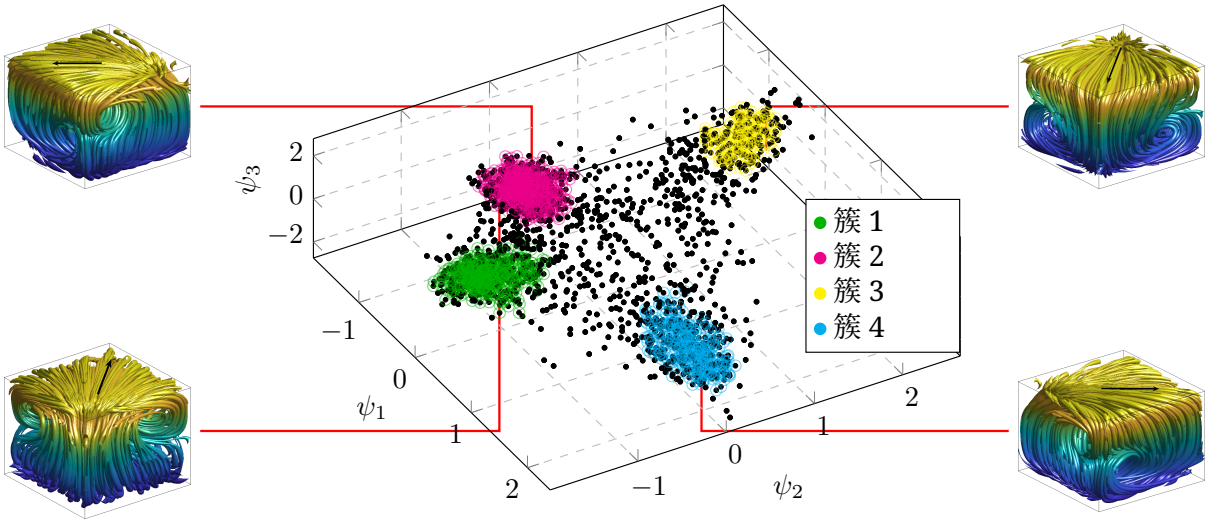

图 7：由时间滞后独立成分分析（TICA）确定的前三个坐标 ψ_k 所张成的超平面中的大尺度环流状态。图中心显示了轨迹的一个子样本（2000 个点）在 (ψ_1, ψ_2, ψ_3)-空间中的投影。以颜色标记的点表示基于密度的噪声应用空间聚类（DBSCAN）算法在该投影中识别出的四个簇的归属。图两侧显示了与这四个簇相对应的时间平均速度场，以流线图的形式呈现。数据对应于 $Ra = 10^6$。

对线性可观测量要求过高了。

B. TICA坐标：Rayleigh–B´enard 情形

由于我们关注的是对流流动较大尺度上的宏观和介观动力学效应，因此我们减少数据量，并对直接数值模拟（DNS）的原始网格分辨率进行降采样。三个分量的

因此，速度向量场在具有 $16 \times 16 \times 16$ 个点的更粗均匀网格上进行谱插值。因此，RBC流轨迹嵌入在维数为 $N = 16^3 \times 3 = 12288$的相空间中。

随后，对这条长度为 $N_s = 10^5$的 N维轨迹 $(x_i)_{i=1}^{N_s}$ 进行时间滞后独立成分分析，得到流动状态 x_i的*TICA*坐标$\psi_1(x_i)$ $\psi_2(x_i)$ $\ldots, \psi_N(x_i)$ 。图5（a、d）中的时间滞后独立成分谱显示，第三个特征值之后存在明显间隙；该特征值见同图的（b、e）图。这个间隙表明，前三个TICA坐标中的每一个所包含的动力学方差，都是其他任一坐标的两倍，因此建议采用上一节引入的观测函数 $\psi = (\psi_1, \psi_2, \psi_3)$（注意 $d = 3$），见图5（c、f）。如附录B 6所述，对于转移流形分析，增加TICA坐标的数量并不会带来信息增益，因此对于当前讨论的两种情况，前三个主导坐标已足够。

在继续分析系统的暂态动力学之前，我们先从静态角度分析数据$\{\psi(x_i)\}_{i=1}^{N_s}$，即不考虑数据点的先后顺序信息。图6展示了前六个TICA坐标的直方图，显示出 ψ_1, ψ_2, ψ_3 at$Ra = 10^6$具有多模态结构。这表明，瑞利-贝纳德对流中的大尺度结构可能对应于映射到 ψ空间中的数据簇。事实上，在 ψ空间中绘制轨迹的一个子集 [71]，可以看到四个密集分布的区域，如图7所示。采用标准的基于密度的噪声应用空间聚类（DBSCAN）算法 [49]，参数为 $\varepsilon = 0.12$ 、minPts为 $= 15$，我们识别出四个簇，也就是四组轨迹索引集；在DBSCAN算法的参数适度变化时，这些簇仍然稳健。将这四组索引映射回物理空间，并对相应的流场求平均后，可以看到它们分别对应四种大尺度环流状态；侧面的图7展示了流线图。 $Ra = 10^7$ 的结果类似（未展示）。

到目前为止，基于时间滞后独立成分分析的瑞利–贝纳德流分析本质上是线性的。借助以下转移流形框架，我们将推进到非线性约化。如上所述，尽管 TMF 可应用于任意系统，但它仅对可逆系统具有定量性能保证。RBC 是一个确定性系统，因此不可能具有可逆性。补救措施在于， [32] 中的结果以及图 5(a) 中主导时间滞后独立成分谱的实数性，都表明大尺度过程具有可逆性。

C. 过渡流形分析

虽然投影到主导的三个 TICA 坐标上显然可以识别大尺度环流状态，但无法深入了解它们之间的过渡路径。因此，我们现在应用基于集合的过渡流形算法，见第 IV B 节。我们再次将该框架的技术细节移至附录B。该算法已在算法 4.1 中概述。现在将其应用于投影到前三个 TICA 坐标上的轨迹。

第一步是构建 $\mathbb{X}$ 的沃罗诺伊剖分，尤其要解析过渡区域；或者等价地说，选取点 $z^{(1)}, \ldots, z^{(L)}$，将其作为沃罗诺伊单元的中心点。基于附录 B 1 和 B 2 中详述的原因，我们选择 k-均值聚类算法，并将簇数量设为 $k = L = 2000$ ，以选取中心点。这样，每个沃罗诺伊单元平均对应 50 个轨迹快照。在以往涉及 TMF 的研究中，已有证明数量相当的样本能够捕捉所关注的动力学性质 [17, 48]。与等距子采样相比，所选中心点 $z^{(1)}, \ldots, z^{(L)}$对簇区域的采样略显不足，而对过渡区域的覆盖则更加密集。需要注意的是，由于 k-均值算法的性质，中心点通常并不是数据集（即轨迹）中的实际状态，而是数据集中某些子集的质心。

接下来，我们应用算法 4.1，使用与中心点 $z^{(1)}, \ldots, z^{(L)}$相关联的沃罗诺伊单元 $A_1, \ldots, A_L$。对于转移密度之间的距离度量，我们使用最大均值差异（见附录B3），并使用扩散映射从所得距离中学习集体变量 ξ，见附录B4。所得分析（见附录B 5和附录B6）表明，相关联的转移流形至多是二维的，因此我们得到一个二维集体变量： $\xi(x_i) \in \mathbb{R}^2$。沃罗诺伊中心的集体变量 ξ 在图8(a)中以黑点表示。现在，数据聚集在该集体变量空间的四个角落，对应于大尺度环流。

然而，与图7 中的 TICA嵌入相比，簇要集中得多。这表明，尽管来自同一簇的两个点 x_i, x_j 来自同一簇可能在物理空间或时间滞后独立成分分析-ψ-空间中存在显著差异，但它们的动力学距离，即密度 $p^\tau(x_i, \cdot)$ 和 $p^\tau(x_j, \cdot)$之间的最大均值差异，要小得多。这使我们能够根据它们的未来统计更好地对状态进行分组，即识别具有相同长期演化的状态。

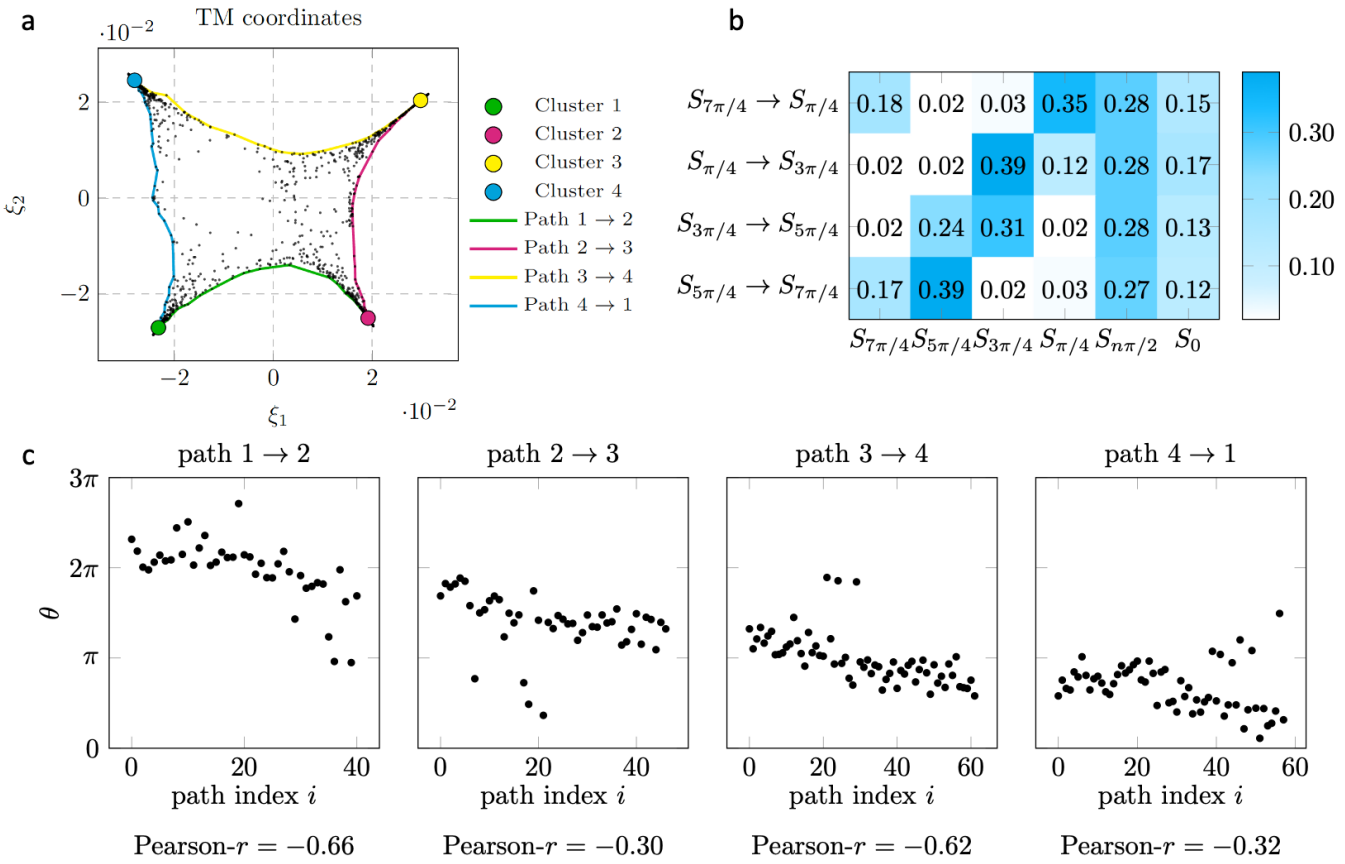

图 8：瑞利–贝纳德流动的转移流形分析。(a) 在由二维集体变量 ξ 张成的转移流形空间中确定的簇中心之间的最短路径。(b) 转移路径处于六种大尺度环流状态中的每一种状态内或状态之间的时间比例。(c) 沿各条转移路径的进程（以路径上各点的索引表示）与中平面角 θ 的比较。可以直观看到二者存在很强的线性相关性，这也得到了较高皮尔逊相关系数的证实。注意，由于 θ 具有周期性，绘制其原始值无法正确反映它沿路径的变化趋势。因此，为得到这些结果，我们进行了预处理步骤，根据哪个值能使其与 θ_{i-1} 的距离最小，将每个 θ_i 可能平移 $\pm 2\pi$。数据对应 $Ra = 10^6$。

D. 集体变量中的暂态动力学

由于转移流形上彼此邻近的点具有相似的演化过程，因此集体变量空间中的最短路径很可能具有独特的动力学意义。为验证这一点，我们考察图8(a) 中各角点之间的最短路径；这些路径即为彩色曲线。更确切地说，对于两个簇中心 c_1，$c_2 \in \mathbb{R}^2$，我们考察连接它们的直线

$$\gamma(r) := rc_1 + (1-r)c_2, \qquad r \in [0,1],$$

并将其定义为转移流形采样点之间从 c_1 到 c_2 的离散路径，即

$$\Gamma(r) := \mathrm{argmin}_{i=1,\dots,L} \left\| \gamma(r) - \xi(z^{(i)}) \right\|.$$

请注意，这些路径并非系统的轨迹。沿各条路径的进展与中平面角 θ （回顾图2）表现出很强的相关性，如图8(c)所示。θ 与 $\Gamma(r)$ 之间相关性不完美，这是意料之中的，因为通过中平面角对环流状态进行分类是基于（i）几何—物理直觉，以及（ii）仅基于时间瞬时的流场信息，这与TMF中的动力学信息形成对比。然而，这两种分类方法的结果之间存在明显的相关性。这些过渡路径与第三节A中的启发式分类吻合良好。图8(b)所示的路径在各自大尺度环流中所经历的部分进一步印证了这一点。我们观察到，一条连接两个

与两个长寿命大尺度环流相关的角落大部分时间都处于这些状态，而在经过短寿命大尺度环流和非相干态时发生转变，并且明显几乎不在其他长寿命大尺度环流中停留。

图9展示了集体变量空间中不同大尺度环流状态的出现频率，即它们在对应沃罗诺伊单元中的出现频率。轨迹快照中的大尺度环流仍通过图2所用的分类方法进行计算。我们观察到，长寿命大尺度环流的快照分别集中在转移流形的一个角上。此外，短寿命大尺度环流倾向于均匀分布在该结构的边缘和角上，而非相干快照则更倾向于落在转移流形的中心。这些结果表明，转移流形的几何形状能够编码并凸显系统的长期稳定性和转变动力学。

最后，我们注意到，转移流形分析还表明不存在流动反转。为此，我们在附录B6的图12(a)中考察了转移流形的一种替代参数化方式——因此也得到略有失真的集体变量——以突出小尺度几何特征。该参数化方式显示，转移流形对角相对的角之间不存在相连的路径，而这些角对应于环流方向相反的长寿命大尺度环流。

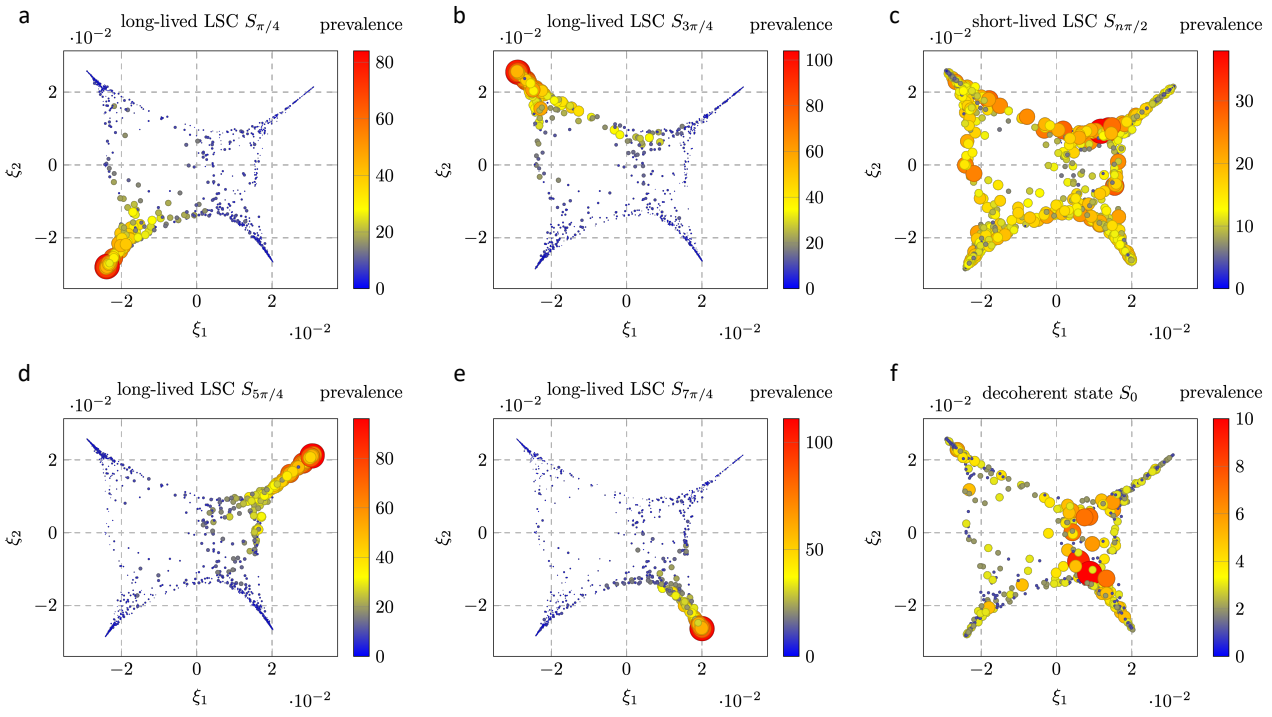

图9：对于 $Ra = 10^6$情形，在所有轨迹快照中，四种长寿命状态（a、b、d、e）、短寿命状态（c）和非相干状态（f）的出现频率。点的颜色和大小均表示落入与该点对应的Voronoi单元中的轨迹快照数量，这些快照属于六种状态之一；这些状态由图2所采用的方法识别，并在第三节A中进行了详细描述。数据对应 $Ra = 10^6$。

我们对 $Ra = 10^7$ 的数据重复进行了 TMF分析（未显示）。数据处理流程相同，首先得到 3 个 TICA 坐标，如图5(d–f)所示。TMF框架再次给出了 4 个不同的簇。图10（附录B 3）所示的 MMD直方图表明，数据的聚簇程度更高，这一点也可以从图2(b)中看出。然而，与 $Ra = 10^6$ 的情况相比存在差异。具体而言，该结构不再完全是二维的，而是三维的，甚至可能更多。不过，第三个维度并未赋予转移流形任何额外的可解释结构；它似乎只是使流形“变厚”，也就是说，仿佛仅仅是噪声。可以预期，随着瑞利数的增加，转移流形的维度也会增加；这一点只能留待未来研究。

VI. 总结与讨论

本研究提出了对湍流瑞利–B´énard对流流动中从一个大尺度状态过渡到另一个大尺度状态的交叉动力学的数据驱动描述——这是确定性的宏观复杂非线性动力系统的典范。众所周知，封闭立方体单元中的RBC流会以沿对角线方向的大尺度环流滚动形式停留在宏观状态中数百个对流时间单位，随后快速切换到另一个对角线流状态。这些跃迁的具体机制是我们研究的重点。正如我们在第二和第三节中详细讨论的那样，它们的特征在于构建大尺度环流的热羽流与LSC两侧振荡的高振幅涡旋之间复杂的相互作用。另请参见展示该跃迁的图1和图2。这一动力学过程完全是三维的，并且包含了我们所研究的动力系统中的大部分自由度。

我们将转移流形框架应用于当前系统。该框架最初是为随机微观系统开发的，例如大分子的构象跃迁。在我们的宏观系统中，成功将系统显著约化为 2 个坐标，并将其称为集体变量。对于 4 个湍流场（u_x，u_y，u_z，T），初始计算网格在两次运行中均先下采样至 (u_x，u_y，u_z)，其自由度总数至少为 $N_{\mathrm{full}} \gtrsim 3.5{\cdot}10^6$。这构成了我们对跃迁动力学进行约化描述的起点。

随后，对跨越非线性转移流形的 2 个坐标进行的数据驱动计算，预先通过时间滞后独立成分分析（TICA）进行了处理，该分析将 12288 维的数据空间降维。TICA 确定了线性坐标，使我们能够识别大尺度环流状态，如图7所示。随后，我们在经过TICA-预处理的时间序列上运行转移流形分析，发现实际上有两个坐标描述了其转移动力学。将动力学投影到这些集体变量上，显著改善了簇之间的区分效果，这一点在比较图 7 和图9时可以看出。大尺度环流状态与转移流形的尖点清晰对应。此外，图 9清楚地表明，短寿命大尺度环流沿着流形边界排列，连接各个尖点，而非相干状态位于内部；也就是说，所有六个宏观态都可以在由两个集体变量张成的平面中得到适当区分。新的集体变量还表明，对于所选参数组，没有观察到流动反转，因此我们可以得出结论：在当前的三维 RBC 对流情况下，流动反转极为罕见。

未来的工作可以利用这些降维坐标，识别系统中大尺度过程的代理动力学。通过研究纵横比稍大的封闭胞腔中的对流流动构型，可以进一步拓展这一框架。此时，多个大尺度环流涡卷将充满模拟区域。第一步应如参考文献 [32] 对当前情形所做的那样，验证大尺度流动动力学的马尔可夫性质。届时可能会出现更多寿命较长的宏观态构型，因此，在 TMF框架中描述这些跃迁所需的集体变量也可能超过两个。

A. 数据处理管道摘要

附录中详细概述了所用方法的若干具体内容，包括附加图表和算法表格。其中包含有关转移流形框架以及通过时间滞后独立成分分析进行中间降维的技术细节。在此，我们再次简要总结学习集体变量的数据处理流程，分为以下 5 个步骤：这些步骤是：

1. 将网格从分别对应于 $Ra = 10^6$ 和 10^7,的 $N_{\mathrm{full}} \approx 3.54 \times 10^6$ 和 3.84×10^6 个自由度降采样至 $N = 12888$ 个自由度。这是合理的，因为我们关注的是出现在介观尺度及更大尺度上的过程。 2. 计算时间滞后独立成分分析坐标 $\psi_1, \ldots, \psi_N$，并将 $\psi := (\psi_1, \psi_2, \ldots, \psi_d)$以及 $d = 3$ 设为高维原系统的可观测量。这一降维遵循以下假设：转移密度是低维的，并且对于 $d \ll N$，密度估计更加准确。 3. 随后对 $\{\psi(x_i) \mid i = 1, \ldots, N_s\}$ 进行沃罗诺伊剖分，以找到沃罗诺伊单元$^{\intercal}$ $A_i \subset \mathbb{R}^d$, $i = 1, \ldots, L = 2000$。这是合理的，因为复杂的湍流表现得像随机系统。作为主流程的分支，我们对单元中心点应用 DBSCAN，并将这些簇与对流流动的大尺度环流相关联。 4. 通过样本 $\{\psi(x_{j+\tau}) \mid \psi(x_j) \in A_i\}$，近似转移密度 $p^\tau(x_i, \cdot)$ 的 ψ-推前，并使用最大均值差异计算它们之间的距离矩阵 $D \in \mathbb{R}^{L \times L}$。其可行性

这一步的效果已在 [17], 中得到证明，其中最大均值差异最终允许进行“准嵌入”。

5. 对 D 应用扩散映射，计算转移流形，从而估计其维数为二，并得到相应的集体变量 $\xi = (\xi_1, \xi_2)^\mathsf{T}$。

致谢

P.M. 的工作获得德国研究联合会（DFG）资助，资助项目编号为 SCHU 1410/29-1 和 SCHU 1410/30-1。A.B. 获得 DFG 通过 CRC 1114“复杂系统中的尺度级联”（项目编号 235221301，项目 B03“多尺度问题的多层次粗粒化”）提供的资助。P.K. 部分获得 DFG 通过 CRC 1114“复杂系统中的尺度级联”（项目编号 235221301，项目 A08“准定常大气状态的表征与预测”）提供的资助，并获得德国卓越战略——柏林数学研究中心 MATH+（EXC-2046/1，项目 ID：390685689 ）的资助。全面的长期模拟在伊尔梅瑙工业大学（德国）大学计算中心的计算集群上进行。

附录 A：时间滞后独立成分分析（时间滞后独立成分分析）

时间滞后独立成分分析（TICA）用于对模拟数据进行预处理。TICA将数据降维至状态空间中的一个低维超平面，该超平面将作为随后进行转移流形分析以确定集体变量的起点。

1. 时间滞后独立成分分析算法

大尺度环流在高维构型空间（或相空间）中演化，由缓慢演化的或亚稳态的 LL-LSC状态，以及快速演化的过渡状态组成，其中包括SL-LSC 和非相干态。由于数据具有高维性，提取系统的基本构型状态和过渡状态颇具挑战。这表明应采用降维技术，以有效处理所生成的数据。

或许最广泛使用的线性降维技术是主成分分析（PCA）。PCA 寻找输入自由度的高方差线性组合，而时间滞后独立成分分析 [50] 是一种基于变分方法的线性变换方法，它将一组高维输入坐标转换为一组具有最大自相关性的低维输出坐标。因此，它可以被视为 PCA 的动态版本。时间滞后独立成分分析已在信号处理中得到发展 [50] ，并且最近被应用于分子动力学 [15, 51, 52]。其使用有两个目标，即：（i）降低构型空间的维度，以便更快地处理数据；（ii）提取慢变量参数，从而帮助我们了解缓慢演化的状态。

流体动力学中常用的另一种方法是动态模态分解（DMD） [53, 54], ，它通过最小二乘拟合，对观测到的时间序列建立线性模型。值得注意的是，正如 [55, 第 3 节2]所示，从算法层面来看，时间滞后独立成分分析和动态模态分解本质上是等价的。然而，动态模态分解旨在利用线性模型对流体流动的演变进行可视化和预测，而时间滞后独立成分分析则用于生成一组（降维后的线性）坐标，以解释系统中大部分的动力学变化。因此，以下内容将采用时间滞后独立成分分析的术语。

我们从一条 N 维动力系统轨迹 $\mathbf{x}(t) = (x_1(t) \ldots, x_N(t)) \in \mathbb{R}^N$ 开始。这里，笛卡尔坐标取为速度向量场的三个分量，这些分量经谱插值后定义在均匀$16 \times 16 \times 16$ 网格上（比计算网格更粗）。因此，RBC流轨迹嵌入在维数为 $N = 16^3 \times 3 = 12288$ 的相空间中。请注意，我们不将温度场纳入分析。输入数据应满足零均值，因此我们将使用

$$\mathbf{x}'(t) = \mathbf{x}(t) - \langle \mathbf{x}(t) \rangle_t \,, \tag{A1}$$

其中，$\langle \cdot \rangle_t$ 表示时间平均。随后，我们计算各种滞后时间 τ 下的自协方差矩阵 $C(\tau)$，其定义为

$$C_{ij}(\tau) = \langle x_i'(t) x_j'(t+\tau) \rangle_t \in \mathbb{R}^{N \times N} \,. \tag{A2}$$

C 的矩阵元通过以下方式进行计算

$$C_{ij}(\tau) = \frac{1}{N_s - \tau - 1} \sum_{t=1}^{N_s-\tau} x_i'(t) x_j'(t+\tau)\,, \tag{A3}$$

其中， t 和 τ 是输出时间间隔 $1t_f$的整数倍。因此，在本例中，时间对应于从 1 到 $N_s = 10^5$ 的整数。TICA特征向量 $\mathbf{\Psi}_k \in \mathbb{R}^N$ 可通过求解广义特征值问题得到

$$C(\tau)\mathbf{\Psi}_k = \lambda_k C(0)\mathbf{\Psi}_k\,. \tag{A4}$$

其中第 k个 TICA 特征值为 λ_k。*TICA*坐标 （或 TICA 模态） $\psi_1(x'(i))$ $\psi_2(x'(i))\ldots$ 是以 TICA 特征向量 $\mathbf{\Psi}_k$构成的基来表示 $x'(i)$时的系数。我们注意到，对于数据矩阵

$$\begin{aligned} X &:= \begin{bmatrix} | & & | \\ x'(1) & \cdots & x'(N_s-\tau) \\ | & & | \end{bmatrix} \quad \text{and} \\ Y &:= \begin{bmatrix} | & & | \\ x'(1+\tau) & \cdots & x'(N_s) \\ | & & | \end{bmatrix}, \end{aligned} \tag{A5}$$

人们通常写作

$$\begin{aligned} C(0) &= \frac{1}{N_s - \tau - 1} X X^\intercal \quad \text{and} \\ C(\tau) &= \frac{1}{N_s - \tau - 1} X Y^\intercal\,. \end{aligned} \tag{A6}$$

2. 坐标降维

我们使用 PyEMMA 软件包 [56]计算滞后时间为 $\tau = 50$ 时的 TICA 坐标。图 5(a) 显示了自协方差矩阵的特征值 $\lambda_i,\ \ i \geq 1$ ，下文将其称为*TICA*特征值 。

这本身也值得单独探讨，不过我们注意到，TICA（以及 DMD）提供了对与动力学相关的复合算子的近似，该算子以 Koopman 命名，参见 [55] 及其中的参考文献。因此，TICA谱不可避免地只是对系统谱的一种粗略近似，进而也是对动力学中相关性衰减的一种粗略近似。

我们观察到，谱中的大部分都局限在半径小于 1 的复圆盘内。这是拟紧转移算子（以及 Koopman 算子）的典型情形；这类算子出现在具有一定混沌性或混合程度的复杂动力学中， [57, 尤其参见第 1.3 和 2.3 节]。此外，占主导地位（按绝对值计）的 TICA特征值均为纯实数，这表明在与这些特征值相关的长时间尺度上，系统的表现具有统计可逆性。事实上，从一个 LSC 跃迁到其他 LSC 似乎并没有周期性倾向。TICA特征值的绝对值也称为其对应特征向量的动力学方差 ，可用于指示投影到该坐标后保留了多少“动力学信息” [58]。图 5(b) 展示了前 20 个坐标的动力学方差。 λ_3之后的谱隙显而易见，这暗示前三个特征向量在动力学上的重要性远超其余特征向量。

累积动力学方差现在定义为截至某个TICA特征对的动力学方差之和，相对于所有动力学方差之和：

$$c_i := \frac{\sum_{k=1}^{i} |\lambda_k|}{\sum_{k=1}^{N} |\lambda_k|} \tag{A7}$$

c_i 在 i 中的变化过程如图 5 (c) 所示。我们观察到曲线逐渐变平，这表明较低阶 TICA 坐标的影响更大（这一点很明显，因为特征值按绝对值递减排序），但在 c_3之后没有明显的跳跃或拐点。此外，与例如许多分子动力学数据集的分析中常见的情况相比，这条曲线的斜率要平缓得多 [51]。这说明动力学方差的很大一部分由中小尺度过程产生，并非过度集中于全局大尺度过程；后者可由前三个 TICA 坐标很好地表示（见正文图 7）。

因此，尽管 λ_i, $i = 1$，2，3可能包含不成比例的大量信息，但要通过线性模型或多或少完整地解析系统动力学，则需要数量多得多的 TICA 坐标。另外，请注意，在 PyEMMA 使用的单精度浮点运算中， c_{8935} 等于 1，因此数据集可以减少 $N - 8935 = 3353$ 个维度，而不会造成可测量的动力学方差损失，其中 $N = 12288$ 是原系统的维数。

图 6 通过直方图展示了沿我们的轨迹取值的前六个 TICA 坐标。与前三个最大的纯实特征值相对应，前三个坐标也都是纯实数。此外，属于谱中分离良好的主导部分的坐标呈现多峰分布，而属于上述复圆盘中特征值的坐标则近似服从正态分布。

ψ_1 的多模态结构,2,3 让人们希望能够据此识别瑞利-贝纳德对流中的大尺度结构。事实上，在（ψ_1, ψ_2, ψ_3）空间中绘制轨迹的一个子集 [72] ，可以发现四个高密度区域（见正文中的图7）。应用基于密度的噪声应用空间聚类（DBSCAN） [49] ，使用参数 $\varepsilon = 0.12$ 和 minPts $= 15$，我们识别出四个簇，即四个轨迹索引集的集合；这些集合对于DBSCAN算法中参数的适度变化具有鲁棒性。将这四个索引集映射回物理空间，并对相应的流场进行平均后，我们看到它们分别对应于四种大尺度环流状态（再次参见图7两侧）。

附录B: 随机系统转移流形框架的算法实现 s

在实际中，转移流形 $\mathbb{M}$ 和映射 $\mathcal{E}$， $\mathcal{Q}$ 均未知，而 $p^\tau(x, \cdot)$ 只能通过经验获知（例如，在 x 中使用不同的随机种子启动大量 DNS 模拟，并运行至时间 τ）。因此，算法策略是在 ~ 集合 $\mathbb{M}$ 的有限子样本中，同时学习 参数化 $\mathcal{E}$ 和投影 $\mathcal{Q}$ 。我们首先概述该算法，然后更详细地说明各个步骤:

算法 2.2跃迁流形集体变量的逐点计算

1: 从 $\mathbb{X}$ 中采样 M 个点 $\{x_1, \ldots, x_L\}$，使亚稳态区域和过渡区域得到覆盖。2: 对于每个 x_i，通过并行直接数值模拟采样 $p^\tau(x_i, \cdot)$。3: 近似距离矩阵 $D \in \mathbb{R}^{L \times L}$,

$$D_{ij} = d\left(p^\tau(x_i, \cdot), p^\tau(x_j, \cdot)\right).$$

从采样的 $p^\tau(x, \cdot)$

4: 将无监督距离型流形学习算法应用于 D。**输出:** 跃迁流形集体变量 ξ^* 的近似值，在点 $\{x_1, \ldots, x_L\}$ 处进行评估，即

$$\{\xi^*(x_1), \ldots, \xi^*(x_L)\}.$$

1. 状态空间采样

评估点 $\{x_1, \ldots, x_L\}$ 应均匀覆盖 $\mathbb{X}$中具有动力学相关性的区域，通常是亚稳态集合及其之间的过渡区域，以便 ξ^* 能够在全局范围内捕捉相关动力学。存在多种策略，各自具有不同的优点和缺点。虽然均匀采样 $\mathbb{X}$ 可能是最容易实现的，但它会在动力学无关的区域大量过度采样，即 π-测度较小的区域。尤其是在高维情况下，这一点更为明显，因为 π 往往集中在 $\mathbb{X}$中一个（勒贝格）测度较小的部分。因此，需要大量样本才能以足够的细致程度覆盖相关区域。

另一方面，从不变密度中采样 π 可以解决这一问题，但会导致对亚稳态区域过度采样的偏差（ π 在这些区域中高度集中），而过渡区域则被忽略。为减轻这种偏差，可对 π 进行第二步子采样，以调整分布，使样本分布得更加均匀。文献 [48], 提出了两种子采样策略，即泊松圆盘和*k*均值子采样。这两种策略分别使 ξ^*的特定近似误差最小化。有关这些算法的详细信息也可参见 [48]。在后续的数据分析中，我们将使用 k均值，这主要是因为其数值实现更易获得且更稳健。

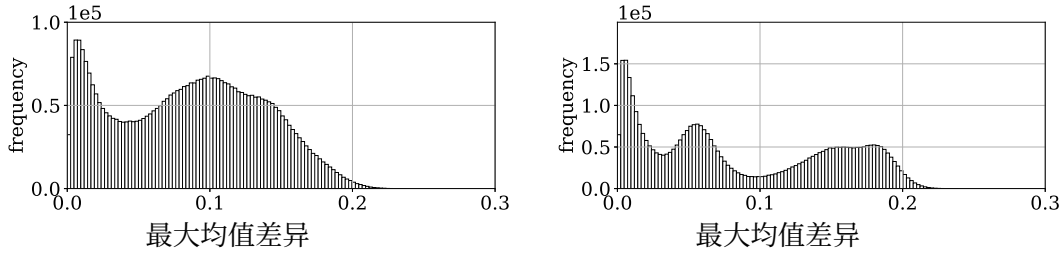

图 10：嵌入式转移密度之间最大平均差异的分布。我们观察到距离聚集成两簇，一簇接近零，另一簇约为 0.1，分别对应簇内距离和簇间距离。数据分别对应 $Ra = 10^6$ （上）和 10^7 （下）。

2. 跃迁密度采样

这一步旨在对转移密度 $p^\tau(x_i, \cdot) \quad i = 1, \ldots, L$进行经验近似，因为除最简单的动力系统外，这些转移密度都无法以闭式形式表示。为此，令 $\Phi_\omega \colon \mathbb{R}^+ \times \mathbb{X} \to \mathbb{X}$ 表示过程 $\{X_t\}$的样本函数，其中 $\omega \in \Omega$（样本空间以 $\{X_t\}$为基础）。对于服从 Prob 分布的$\omega_1, \ldots, \omega_M$，由此得到 $p^\tau(x, \cdot)$的经验 M-样本代表：

$$\{\Phi_{\omega_1}(\tau, x), \ldots, \Phi_{\omega_M}(\tau, x)\} \sim p^\tau(x, \cdot).$$

如果 $\{X_t\}$ 由右端项可微的随机微分方程（SDE）刻画， Φ_ω 可以通过数值积分器实现，例如 Euler–Maruyama 方案或更高阶的方案。

在确定性系统 $\Phi(\tau, \cdot)$ 的基于 Voronoi 单元的离散化中，随机性来自从状态 x 所属的 Voronoi 单元中抽取的初始条件。样本 ω_i 随后指的是这些随机抽取的初始条件。

3. 密度距离矩阵

距离矩阵 D 的组装仅用于为流形学习算法的应用准备数据（见下一小节），但包含密度度量 d 的关键选择：$L^1(\mathbb{X}) \times L^1(\mathbb{X}) \to \mathbb{R}^+$，该选择决定了算法的数值可行性及其输出质量。虽然

$$d(u, v) = \|u - v\|_{L^1}$$

可能看似是显而易见的选择，而转移流形框架正是针对这一度量构建的，但在数值上实现它却异常困难（涉及对 N 维空间 $\mathbb{X}$的积分），并且无法直接根据 u, v 的样本进行估计（必须推导出 u 或 v 的近似解析表达式）。由于这些困难，在 [16, 59], 中提出了几种替代性的统计距离，包括最大均值差异 （最大均值差异） [60], ，该距离将在后续的数据分析中使用。

最大均值差异取决于正半定核函数 k 的选择： $\mathbb{X} \times \mathbb{X} \to R^+$，其解析定义为

$$d_k(u, v) := \|\mu(u) - \mu(v)\|_{\mathscr{H}(k)},$$

其中$\|\cdot\|_{\mathscr{H}(k)}$ 是由内积诱导的范数。这里， μ 表示映射到 $\mathscr{H}(k)$ 中的核均值嵌入 [61] 。由于在 $\mathscr{H}(k)$ 中有如下内积表示式，可直接据此得到计算诱导范数的公式，因此其具体形式与 $\mathscr{H}(k)$ 的具体形式无关：

$$\langle \mu(u), \mu(v) \rangle_{\mathscr{H}(k)} = \int_{\mathbb{X}} \int_{\mathbb{X}} u(x) v(y) k(x, y) \, d\pi(x) \, d\pi(y).$$

由内积$\langle \cdot, \cdot \rangle_{\mathscr{H}(k)}$ 所诱导的希尔伯特空间 $\mathscr{H}(k)$ 称为与 k 相关的再生核希尔伯特空间（RKHS）。关于 RKHS 及其泛函分析性质，已有大量文献 [62,63]。此外，部分作者在 [59] 中证明，映射 $\mathbb{M} \mapsto \mu(\mathbb{M})$ 是一种嵌入，即保持中心的。

$\mathbb{M}$的拓扑性质。重要的是，从样本中估计$\langle\cdot,\cdot\rangle_{\mathscr{H}(k)}$，进而估计 $d_k(\cdot,\cdot)$是直接的：对于

$$\langle u, v\rangle^M_{\mathscr{H}(k)} := \frac{1}{M^2}\sum_{k,\ell=1}^{M} k(y_i^k, y_j^\ell),$$

其中 $y_i^k \sim u$, $y_j^\ell \sim v$ 成立

$$\langle\cdot,\cdot\rangle^M_{\mathscr{H}(k)} \to \langle\cdot,\cdot\rangle_{\mathscr{H}_k} \quad (M\to\infty).$$

如正文所述，我们将通过观测函数 $\psi = (\psi_1, \psi_2, \psi_3)$ 对其进行前推，比较转移密度 $p^\tau(x,\cdot)$，其中 ψ_i 表示第 i 个 TICA 坐标。换言之，对于不同的 $x, y \in \mathbb{X}$ 对，我们需要计算密度 $\psi(X)$ 和 $\psi(Y)$ 的最大均值差异，其中 $X \sim p^\tau(x,\cdot)$ 和 $Y \sim p^\tau(y,\cdot)$。为此，我们选择核函数 $k(\xi,\eta) = \exp(-\|\xi-\eta\|^2/\sigma)$，其中 $\sigma = 7$ 以及 $\xi,\eta \in \mathbb{R}^3$（因为 ψ映射到 $\mathbb{R}^3$）。两个数据集的最大均值差异分布如图 10 所示。$Ra = 10^6$ 的数据呈现出 MMD直方图的双峰结构，这也将在关于流形学习方法带宽选择的讨论中出现。$Ra = 10^7$ 的数据呈现出三峰结构。在这里，我们发现，第一个最小尺度源于强烈的聚类，并使数据集断开（参见 B 6 节的讨论）；直方图中的第二个峰值甚至表明，相应的长度尺度无法捕捉数据的全局性质。正是 MMD直方图的第三个局部最大值提供了正确的嵌入。

4. 流形学习

现在，我们将流形学习方法应用于距离矩阵 D。从概念上讲，这相当于学习主文中式 (6) 里的组合参数化与投影映射 $\mathcal{E}\circ\mathcal{Q}$ 的近似。将此映射拉回到起始点 x_i，即可得到最终的集体变量 $\xi^\star$。

我们选择的流形学习方法是扩散映射算法 [64], ，因为过去的实践表明，它在实际场景中表现良好 [17, 48], ，不过也可以使用多维尺度分析 [65], 、局部线性嵌入 [66], 或许多其他方法。

这里我们不详细推导扩散映射的解析形式，而只介绍算法。首先，根据距离矩阵 $D \in \mathbb{R}^{L\times L}$构造马尔可夫转移矩阵 $M \in \mathbb{R}^{L\times L}$：

$$M_{ij} = \frac{K_{ij}}{s_i}, \tag{B1}$$

其中，$K_{ij} = \exp\left(-D_{ij}^2/\varepsilon\right)$是一个相似度矩阵，带有某个带宽参数 $\varepsilon > 0$（不要将其与式 (6) 中给出的动能耗散率场混淆），且 $s_i = \sum_j K_{ij}$。作为一个马尔可夫矩阵，M的主特征值为 1，对应的特征向量为常数向量 $\phi_1 = (1,\ldots,1)$。因此，扩散映射（即在采样点 $\{x_1,\ldots,x_L\}$处取值的集体变量 ξ^*）由以下次主特征向量 $\phi_2,\ldots\phi_{r+1}$给出。这里，r通常可以根据 M的谱间隙确定，也可以将主特征向量 ϕ_i 彼此作图，并舍弃那些不包含额外几何信息的特征向量；这类特征向量有时称为高阶谐波。

我们注意到，扩散映射如 [64] 中所述，具有一个参数 α，该参数用于消除数据样本密度不均匀所导致的偏差。当前的简化描述对应于 $\alpha = 0$的情况。然而，我们的所有实现都使用 $\alpha = 1$，这是在极限$L\to\infty$下使结果不依赖于采样密度的取值。此时，坐标 ϕ_i 仅取决于底层数据流形的几何特征。

5. 带宽选择扩散映射

让我们简要回顾一种在扩散映射方法中自动确定“良好”核带宽参数值的方法 εin。该流程源于 [67], ，后来在 [68] 中得到进一步完善。 [69, 附录 A2]中也可以找到总结。

回顾一下，扩散映射方法首先将数据集 $\{x_i\}_{i=1}^L$ 中的成对距离转换为相似度矩阵 $K \in \mathbb{R}^{L\times L}$，其元素为 $K_{ij}(\varepsilon) = \exp\left(-\varepsilon^{-1}D_{ij}^2\right)$。注意，本文始终使用高斯核。通过对相似度矩阵的元素取平均，我们定义

$$S(\varepsilon) := \frac{1}{L^2}\sum_{i,j} K_{ij}(\varepsilon)\,. \tag{B2}$$

我们注意到以下两种极限行为：

- 由于 $\varepsilon \to 0$，我们有 $K_{ij}(\varepsilon) \to \delta_{ij}$，即克罗内克 delta。因此，$S(\varepsilon) \to \frac{1}{L}$。
- 由于 $\varepsilon \to \infty$，我们有 $K_{ij}(\varepsilon) \to 1$，因此 $S(\varepsilon) \to 1$。

根据理论考量，在这两个极端之间，$\log(S(\varepsilon))$ 应在 $\log(\varepsilon)$ 中存在一段仿射线性增长区间。建议通过最大化以下表达式来确定该区间

$$\frac{d\log(S(\varepsilon))}{d\log(\varepsilon)}, \tag{B3}$$

关于 ε。其思路是，这种 ε 既不能太小（相对于数据点密度而言），也不能太大（相对于数据点云的直径而言）。如果数据确实具有流形结构，这样的带宽参数便能解析数据的流形结构，并在扩散映射方法中充分地表示它。相反，带宽过小往往会使数据集实际上彼此断开，而带宽过大则会通过“模糊”局部几何特征而将其忽略。

请注意，该过程还可估算数据流形的维数。对于处于适当范围内的带宽 ε，其结果为式 (B3) 中数值的两倍。若数据密度变化很大，该过程可能会针对不同长度尺度给出多个“最优”带宽 ε，因此需要在分辨率与连通性之间进行折中（或者将数据拆分为连通分量，再逐一处理）。对于更加不规则的数据，可能需要使用可变带宽核函数，如 [68] 中所述。

6. 瑞利–贝纳德对流数据的带宽选择

为了评估更多的 TICA 坐标是否能够捕捉到有关 RBC 系统转变动力学几何特征的更多信息，我们针对分别由前 3 个、5 个和 10 个 TICA 坐标计算得到的 MMD 距离矩阵，运行了附录B5中的带宽调优分析。图 11 展示了 $Ra = 10^6$ 对应的 $S(\varepsilon)$ 值和量 (B3)。

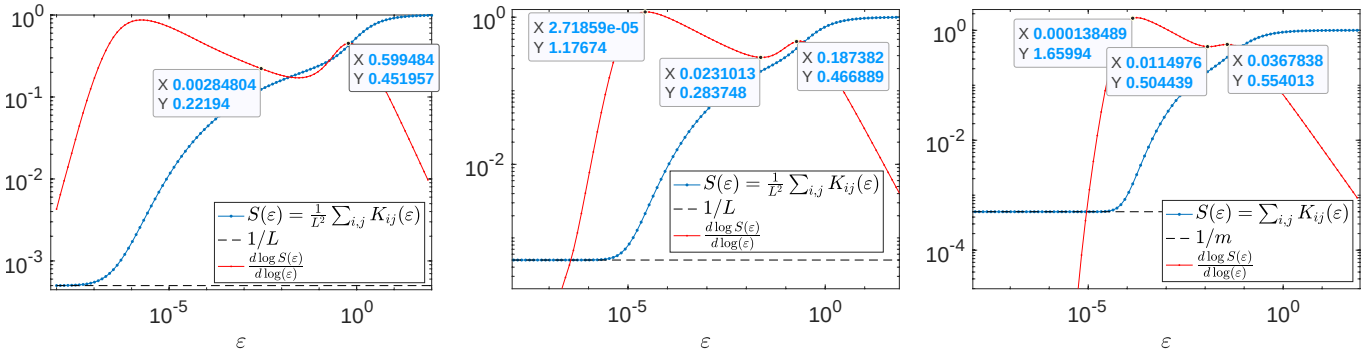

图 11：由前三个（左）、五个（中）和十个（右）TICA坐标获得的最大均值差异距离矩阵的量 $S(\varepsilon)$ 及其双对数导数。虚线表示 $1/L$，即 $S(\varepsilon)$ 的下界。图中的刻度标示了一些数值，以供参考。

如果距离矩阵表明数据密度至多缓慢变化，那么对数导数（红色曲线）将呈现出单一平台，而与该平台相关联的每个带宽 ε 都是表示数据的同样好选择。然而，这些曲线略显双峰结构，表明数据集存在一定的不规则性，并且存在多个邻近尺度，突出了数据集的不同特征：

- 左侧最大值周围的尺度较小，可以解析局部变异，但可能导致数据断开。
- 围绕右侧最大值的更大尺度，将数据集视为一个连通流形。

我们对针对三个主导的TICA坐标所获得的数据集进行了详细分析。原因在于：(i) 本质上，三个案例都估计数据集的维数至多为二维（参见小节B 5），并且(ii)在适当的尺度 ε 下，三个数据集的参数化本质上是等价的（未展示）。

这表明，前三个 TICA坐标已经捕获了系统跃迁统计的足够信息，因此增加更多 TICA 坐标无法对此带来进一步改善（至少对于目前可获得的RBC数据而言）。此外，对于Voronoi中心选择而言， k-均值等辅助算法众所周知在高维中会表现恶化。

图11左侧对应三个 TICA 坐标（ψ_1、 ψ_2、 ψ_3）之间的距离，建议考虑带宽 $10^{-6} < \varepsilon < 0.5$。我们观察到，对于 $10^{-6} < \varepsilon < 10^{-4}$ ，数据集实际上是不连通的，即参数化受少数离群点主导，无法提取涉及多个数据点的结构。对于$\varepsilon = 10^{-4}$，离群点减少到寥寥几个，扩散映射特征向量 ϕ_8、 ϕ_9 已呈现出一些结构，另见图12(a)。

增加带宽后，对于 $\varepsilon = 3 \cdot 10^{-3}$ 可实现“数据流形”的连通性，见图 12(b)，因此这是我们在正文进一步分析中使用的带宽参数。图 12(c) 展示了 $\varepsilon = 0.02$ 的扩散映射，以说明带宽进一步增加如何开始模糊局部特征，并突出数据集的全局几何特征。如果增加到 $\varepsilon = 0.5$ ，这一点也不会改变（未显示）。超过此阈值，即对于非常大的带宽，数据集开始显得越来越像点。

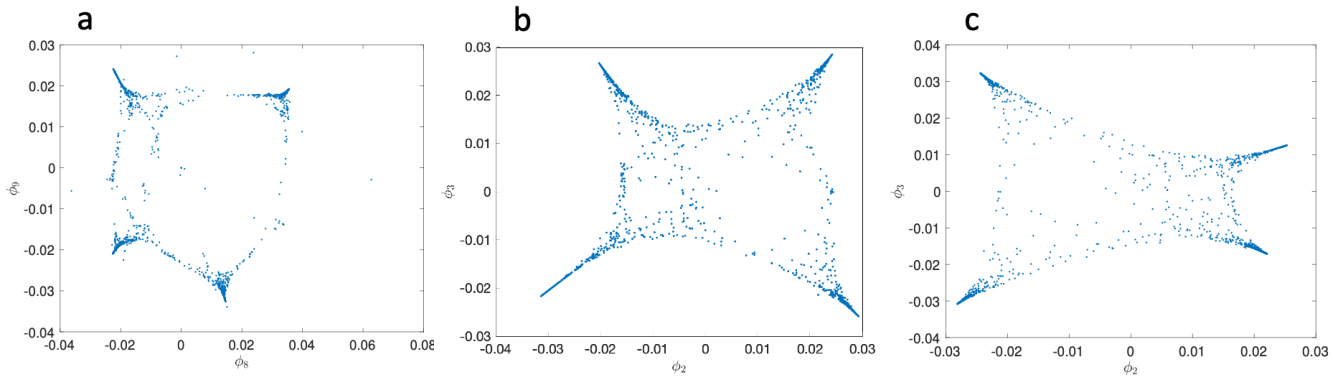

图12：不同带宽下转移流形的扩散映射参数化： $\varepsilon = 10^{-4}$ 对应(a)， 3×10^{-3} 对应(b)， 2×10^{-2} 对应(c)。底层的最大均值差异距离矩阵由三个主导 TICA 坐标计算得出。请注意，由于扩散映射坐标是作为特征向量计算的，其符号并不唯一，在不同计算中可能有所不同。因此，要在不同子图之间进行插值，可能需要将一个或两个坐标轴绕零点翻转。

[1] S. E. Jackson。小型单结构域蛋白质如何折叠？ *Folding and Design*，3:R81–R91，1998。[2] G. D. Rose、P. J. Fleming、J. R. Banavar 和 A. Maritan。基于主链的蛋白质折叠理论。

Proc. Natl. Acad. , : –。

Sci. USA 103 16623 16633, 2006

[3] E. A. Spiegel。恒星中的对流 I：基本 Boussinesq对流。 *Annu. Rev. Astron. Astrophys.*，9(1):323–352，1971。[4] R. Markson。有组织对流的大气电探测。 *Science*，188(4194):1171–1177，1975。[5] J. Marshall 和 F. Schott。开阔海洋对流：观测、理论与模型。 *Rev. Geophys.*，37(1):1–64，1999。[6] J. Schumacher 和 K. R. Sreenivasan。专题讨论：太阳中不寻常的对流动力学。 *Rev. Mod. Phys.*，92:041001，2020。[7] H.-D. Xi 和 K.-Q. Xia。湍流热对流中的流动模式跃迁。 *Phys. Fluids*，20(5):055104，2008。[8] E. P. van der Poel、R. J. A. M. Stevens 和 D. Lohse。连接湍流瑞利–B´纳德对流中的流动结构与热通量。 *Phys. Rev. E*，84:045303，2011。[9] S. Weiss 和 G. Ahlers。在纵横比为 $\tau = 0.50$ 、普朗特数 Pr 为 $= 4.38$ 的圆柱容器中的湍流瑞利–B´纳德对流。 *J. Fluid. Mech.*，676:5–40，2011。[10] S. G. Huisman、R. C. A. van der Veen、C. Sun 和 D. Lohse。高度湍流 Taylor–Couette流中的多重状态。*Nat. Commun.*，5:3820，2014。[11] Y. Bao、J. Chen、B.-F. Liu、Z.-S. She、J. Zhang 和 Q. Zhou。分隔式热对流中的增强热输运。*J. Fluid Mech.*，784:R5，2015。[12] L. Zwirner、A. Tilgner 和 O. Shishkina。受限湍流瑞利–B´纳德对流中大尺度环流的椭圆不稳定性与多辊流动模式。 *Phys. Rev. Lett.*，125:054502，2020。[13] F. Schindler、S. Eckert、T. Z¨urner、J. Schumacher 和 T. Vogt。强湍流液态金属对流中相干大尺度流动的崩溃。 *Phys. Rev. Lett.*，128:164501，2022。[14] D. Ruelle。 *Turbulence, strange attractors and chaos*。1995。

[15] S. Schultze和 H.Grubm¨uller。《随机游走与蛋白质动力学的时间滞后独立成分分析》。*J. Chem. Theory Comput.*，17(9):5766–5776，2021。[16] A. Bittracher、P. Koltai、S.Klus、R.Banisch、M.Dellnitz 和 C.Sch¨utte。《复杂亚稳态系统的转移流形：有效动力学的理论与数据驱动计算》。 *J. Nonlinear Sci.*，28(2):471–512，2018年。[17] A. Bittracher、J. Moschner、B. Koksch、R. Netz 和 C. Sch¨utte。《通过转移流形分析探索NFGAILS淀粉样蛋白原纤维化的锁定阶段》。 *Eur. Phys. J. B*，94:471–512，2021。[18] S.Chandrasekhar。《 *Hydrodynamic and hydromagnetic stability*》。1961。[19] G. Ahlers、S. Grossmann 和D.Lohse。《湍流瑞利–B´énard 对流中的热传递与大尺度动力学》。*Rev. Mod. Phys.*，81:503–537，2009。[20] F.Chill´a 和 J. Schumacher。《湍流瑞利–B´énard对流的新视角》。 *Eur. Phys. J. E*，35:58，2012。[21] J. J. Niemela、L.Skrbek、K. R.Sreenivasan和 R.J. Donelly。《受限热对流中的风》。 *J. Fluid. Mech.*，449 169 178，2001：–。[22] K. R. Sreenivasan、A.Bershadskii和 J.J.Niemela。《热对流中的平均风及其反转》。 *Phys. Rev. E*，65 056306，2002：。[23] A. Parodi、J. vonHardenberg、G. Passoni、A.Provenzale 和E.A Spiegel。《湍流对流中羽流的聚类》。*Phys. Rev. Lett.*，92:194503，2004。[24] B.A.Puthenveettil和 J.H. Arakeri。《高瑞利数对流中的羽流结构》。 *J. Fluid. Mech.*，542 249，2005 217：–。[25] E. Brown和 G. Ahlers。《湍流瑞利–B´énard 对流中的大尺度环流旋转与停止》。*J. Fluid. Mech.*，568:351–386，2006。[26] H.-D.Xi 和K.-Q.Xia。《湍流热对流中大尺度环流的停止与反转》。

Phys. Rev. , ：。 *E* 75 066307, 2007

[27] N. Foroozani、J. J. Niemela、V.Armenio和K. R. Sreenivasan。《容器形状对对流中湍流涨落标度的影响》。 *Phys. Rev. E*，90:063003，2014。[28] K.Bai、D.Ji和 E. Brown。《低维模型预测湍流中大尺度相干结构几何依赖动力学的能力》。 *Phys. Rev. E*，93:023117，2016。[29] N. Foroozani、J. J. Niemela、V.Armenio 和K. R. Sreenivasan。《立方体中湍流对流大尺度流动的重新定向》。 *Phys. Rev. E*，95:033107，2017。[30] D. Giannakis、A. Kolchinskaya、D.Krasnov 和 J. Schumacher。《湍流对流单元中长期演化的 Koopman 分析》。 *J. Fluid. Mech.*，847:735–767，2018年。[31] A. Vasiliev、P. Frick、A. Kumar、R.Stepanov、A.Sukhanovskii 和M. K. Verma。《立方单元中大尺度对流的瞬态流动与重新定向》。 *Int. Commun. Heat Mass Transf.*，108:104319，2019。[32] P. Maity、P.Koltai 和 J. Schumacher。《立方瑞利–B´énard单元中的大尺度流动：长期湍流统计与宏观态跃迁的马尔可夫性》。 *Phil. Trans. R. Soc. A*，380(2225):20210042，2022。[33] S. Brandt、F.Sittel、M. Ernst和G.Stock。《生物分子反应坐标的机器学习》。 *J. Phys. Chem. Lett.*，9 9 2144 2150，2018( )：–。[34] C.Wehmeyer 和 F. No

´ *J.* e。 *Chem. Phys.* 148 24 241703, 2018

《时间滞后自编码器：用于分子动力学的慢集体变量深度学习》。，()：。[35] B. Lusch、J. N.Kutz和S. L.Brunton。《用于非线性动力学通用线性嵌入的深度学习》。

Nat. , ：。 *Commun.* 9 4950, 2018

[36] A.Mardt、L. Pasquali、H. Wu 和 F. No

´ 9 1 1 11, e。 2018

《用于分子动力学深度学习的 VAMPnets》。 *Nat. Commun.*，()：–。[37] K. Champion、B. Lusch、J.N. Kutz 和S. L. Brunton。《坐标与控制方程的数据驱动发现》。*PNAS*，116(45):22445–22451，2019。[38] S.E.Otto和 C.W. Rowley。《用于学习动力学的线性递归自编码器网络》。 *SIAM J. Appl. Dyn. Syst.*，181 558 593，2019( )：–。[39] P. F.Fischer。《用于不可压缩 Navier–Stokes方程谱元求解的重叠Schwarz 方法》。*J. Comput. Phys.*，133(1):84–101，1997。[40] J. D. Scheel、M.S.Emran 和 J. Schumacher。《解析湍流瑞利–B´énard对流中的精细尺度结构》。*New J. Phys.*，15(11):113063，2013。[41] J.D.Scheel和J. Schumacher。《低普朗特数下湍流对流中的整体与局部统计》。

J. Fluid. , ：– *Mech.* 802 147 173, 2016

。[42] J.D.Scheel 和 J.Schumacher。《预测低普朗特数对流流动中过渡到完全湍流黏性边界层的范围》。 *Phys. Rev. Fluids*，2:123501，2017。[43] A.Pandey、D. Krasnov、R. Samtaney、J. Schumacher和K. R.Sreenivasan。《受限域与延展域中对流湍流特征之间的相似性》。 *Physica D*，442:133537，2022。[44] K. Sugiyama、R. Ni、R.J. A. M. Stevens、T. S.Chan、S.-Q. Zhou、H.-D. Xi、C. Sun、S. Grossmann、K.-Q. Xia 和 D. Lohse。《热驱动湍流中的流动反转》。 *Phys. Rev. Lett.*，105:034503，2010。[45] Y. Kifer。《双曲变换与扩张变换的一般随机扰动》。 *J. Anal. Math.*，47:111–150，1986。[46] R.Z.Khas'minskii。《抛物型与椭圆型微分方程以及小扩散马尔可夫过程的平均原理》。 *Theory Probab. Appl.*，8:1–21，1963。[47] G. Froyland。《二维Anosov 系统中 Sinai–Bowen–Ruelle 测度的有限近似》。 *Random Comput. Dyn.* 3 251 264, 1995

, ：–

。[48] A.Bittracher 、 R.Banisch 和 C. Sch¨utte。《分子反应坐标的数据驱动计算》。 *J. Chem. Phys.*，

149(15):154103, 2018年。[49] M. Ester、H.-P.Kriegel、J. Sander和 X.Xu。用于在含噪声的大型空间数据库中发现簇的基于密度的算法。载于 *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*，第 226–231页，1996 年。[50] L.Molgedey和H.G.Schuster。使用时间延迟相关性分离独立信号的混合。

Phys.，：–。[51] G.P´

Rev. Lett. 72 3634 3637, 1994

erez-Hern´andez、F. Paul、T. Giorgino、G.De Fabritiis和F.No´e。用于构建马尔可夫模型的慢分子序参量识别。 *J. Chem. Phys.*，139(1):015102，2013 年。[52] F. N

¨ ´ ´ ´ *J.* uske、B. G.

Chem. Theory Comput. 10 4 1739 1752, 2014

Keller、G. Perez-Hernandez、A. S. J. S.Mey 和F. Noe。分子动力学的变分方法。，( ): –。[53] P. J.Schmid和 J.Sesterhenn。实验数据的动态模态分解。载于*Particle Image Velocimetry Melbourne Victoria Australia* 2009 *8th International Symposium on*

，、、、、。

[54] P. J.Schmid。数值与实验数据的动态模态分解。 *J. Fluid Mech.*，656:5–28，2010 年。[55] S. Klus、F. N¨uske、P. Koltai、H. Wu、I. Kevrekidis、C.Sch¨utte和 F.No´e。数据驱动的模型降阶与转移算子近似。 *J. Nonlinear Sci.*，28(3):985–1010，2018年。[56] M. K.Scherer、B. Trendelkamp-Schroer、F. Paul、G. P´erez-Hern´andez、M. Hoffmann、N. Plattner、C.Wehmeyer、J.H.Prinz和 F. No

´ 2 *J. Chem.* e。PyEMMA：用

Theory Comput. 11 1 5525 5542, 2015

于马尔可夫模型估计、验证和分析的软件包。，(): –。[57] V. Baladi。 *Positive transfer operators and decay of correlations*，第 16卷。World Scientific，2000年。[58] F. No´e和C.Clementi。分子动力学模拟中的动力学距离与动力学映射。 *J. Chem. Theory Comput.*，1110 5002 5011，2015 年 (): –。[59] A. Bittracher、S. Klus、B. Hamzi、P.Koltai 和C. Sch¨utte。通过转移流形的核函数嵌入对复杂亚稳态系统进行降维。 *J. Nonlinear Sci.*，31(1):3，2020年。[60] A.Gretton、K. Borgwardt、M. Rasch、B.Sch¨olkopf 和 A.Smola。核函数双样本检验。 *J. Mach. Learn. Res.*，13723 773，2012 年：–。[61] A. Smola、A. Gretton、L.Song和B.Sch

¨ *International Conference* olkopf。分布的希

on Algorithmic Learning Theory 13 31 2007

尔伯特空间嵌入。载于，第– 页。Springer，。[62] T. Hofmann、B. Sch¨olkopf 和A. J Smola。机器学习中的核函数方法。 *Ann. Stat.*，36(3):1171–1220，2008 年。[63] A.Berlinet 和 C. Thomas-Agnan。 *Reproducing kernel Hilbert spaces in probability and statistics*。SpringerScience &Business Media，2011年。[64] R.R. Coifman 和 S.Lafon。扩散映射。 *Appl. Comput. Harmon. Anal.*，21(1):5–30，2006年。[65] F. W. Young。 *Multidimensional Scaling: History*、 *Theory*、 *and Applications*。Psychology Press，2013 年。[66] S. T.Roweis 和L. K.Saul。通过局部线性嵌入进行非线性降维。

Science 290 5500 2323 2326, ，(): –。[67] R.

2000

R. Coifman、Y. Shkolnisky、F.J.Sigworth和 A.Singer。来自未知随机投影的图拉普拉斯层析成像。 *IEEE Trans. Image Process.*，17(10):1891–1899，2008年。[68] T. Berry和 J.Harlim。可变带宽扩散核函数。 *Appl. Comput. Harmon. Anal.*，40(1):68–96，2016年。[69] P. Koltai和 S. Weiss。湍流瑞利–B´enard 对流中大尺度流动结构的扩散映射嵌入与转移矩阵分析。 *Nonlinearity*，33(4):1723–1756，2020 年。[70] 我们注意到，这个“投影定理”是一个有用的定量陈述，也用于证明惠特尼的（弱）嵌入定理。[71] 以步长50 等时间间隔采样，因此得到2000个点。[72] 以步长 50 等时间间隔采样，因此得到 2000 个点。