此处将在排版后的文章中显示适用于文章类型的横幅

arXiv:2304.03362v1 [physics.flu-dyn] 6 Apr 2023

将蝴蝶变换为图：混沌与湍流系统的统计量

Andre N.Souza[1]†

[1]麻省理工学院，剑桥，马萨诸塞州，美国 4

（收稿日期：xx；修订日期：xx；接受日期：xx） 5

[{ "id": 0, "output": "我们提出了一种数据驱动的方法，用于构建动力系统底层连续性方程/福克-普朗克方程的有限体积离散化。该方法具有灵活划分状态空间的特点，可推广至函数空间，适用于任意长度的时间序列数据，对噪声具有鲁棒性，并能够量化有限样本效应所带来的不确定性。应用该方法后，可以得到马尔可夫状态（单元中心）以及生成元的随机矩阵近似。二者结合使用时，能够模拟底层系统的统计特性。我们将该方法应用于洛伦兹方程（三维常微分方程）以及一种经过修改的Held-Suarez大气模拟（在薄球壳上包含重力和旋转的可压缩欧拉方程的通量差分间断伽辽金离散）。我们证明，粗离散化能够捕捉系统的许多基本统计性质，例如稳态矩、时间自相关以及状态空间子集的驻留时间。" }]

1. 引言

复杂系统建模的目标通常不是确定动力学方程，而是构建在分布上收敛于相关统计量的模型。在湍流建模中，这可以视为大涡模拟（LES）的目标之一：将部分统计量（通常是动能谱）与直接数值模拟（DNS）的结果进行比较。类似地，在地球系统建模中，天气模式在长时间尺度上的不可预测性，要求开发可用于查询相关统计量的非线性模型。因此，这些模型的目标并非收敛于动力学轨迹，而是在分布上收敛于目标可观测量。

当前工作源于构建复杂物理现象（如湍流）的简化统计模型的需求。我们采用了最初由 Hopf（1948年）提出、由 Lorenz（1963年）补充，并在 Cvitanović等人（2016年）那里形成现代形式的湍流动力系统视角。因此，这种方法旨在对与混沌或湍流动力学相关的统计量进行直接离散化；我们假设这些动力学具有遍历性、混合性，并且与分形流形相关。

有许多种直接针对统计量的离散化方法；这里的统计量指对底层连续性方程（确定性动力学）、福克-普朗克方程（随机动力学）、佩龙–弗罗贝尼乌斯／转移算子（离散时间动力学）或 Koopman 算子（佩龙–弗罗贝尼乌斯／转移算子的伴随算子）进行离散化。

† 通信邮箱：andrenogueirasouza@gmail.com

摘要不得延续到第2页

离散化方法包括 Ulam（1964）的方法；Dellnitz 等人（2001、2005）的方法；以及针对随机洛伦兹方程的 Allawala 和 Marston（2016）的方法。现代方法采用算子理论与数据驱动方法相结合的思路，进而构造保持测度的 Koopman 算子，Colbrook（2022）；这类方法建立在 Schmid（2010）等人对线性化动力算子的早期研究之上。Stefan Klus、Péter Koltai 和 Christof Schütte（2016），以及 Fernex 等人（2021）综述了佩龙–弗罗贝尼乌斯算子的数据驱动构造。在各种假设下的收敛保证可见于例如 Froyland（1997）、Colbrook 和 Townsend（2021）、Schütte等人（2022）。

这里的方法最接近 Froyland 等人（2013年）与 Fernex 等人 （2021年）方法的结合，其目标是构造生成元（连续性／Fokker-Planck 算子）的离散化。此外，为了推进对偏微分方程所产生统计量的直接离散化，我们采用场论视角，例如 Hopf（1952年）所提出的视角。

本文其余部分安排如下：第2节讨论基础理论与近似方法。本文的方法深受 Hopf（1952年）和 Cvitanović等人 （2016年）的启发。核心思路是利用描述动力学的方程（“拉格朗日”量）来离散化描述统计量的方程（“欧拉”量）。随后再进行进一步近似，以计算感兴趣的可观测量。

第3节介绍一种数据驱动方法，该方法能够量化不确定性，用于计算生成元的近似值。此方法适用于任意长的时间序列数据、具有较大状态空间的动力系统（1,000，000+ 个自由度），并能进一步提供离散化生成元各项的不确定性估计。

第 4 节介绍两个例子：第一个例子以 Lorenz（1963）为例，介绍该方法的概念与局限性。第二个例子采用 Held 和 Suarez（1994）提出的大气基准案例，研究薄壳球面上的可压缩欧拉方程，并考虑重力和旋转。该系统呈现湍流，可作为地球气候学的替代模型。可以看到，即使对统计量进行粗离散化，也能捕捉原始系统的许多特征。

如果只想了解这些计算能实现什么，可以放心跳过第 2-3节，直接阅读第 4节。此外，附录 A、B、C分别通过讨论 Koopman 模态、简单对称性的纳入方式，以及可压缩欧拉方程的数值离散化，对正文内容作了扩展。

2. 理论

2.1. 有限维动力系统

We 从一个 d 维的一般连续时间动力系统开始，其形式为 y

$$\dot{\boldsymbol{s}} = \boldsymbol{U}(\boldsymbol{s}) \tag{2.1}$$

其中 $\boldsymbol{s}(t)$： $\mathbb{R} \to \mathbb{R}^d$ 是系统的状态，而 $\boldsymbol{U}$： $\mathbb{R}^d \to \mathbb{R}^d$ 是演化方程。方程 2.1 为确定动力系统的演化提供了简洁规则；然而，在存在混沌的情况下，不确定的初始条件会使未来预测迅速失效，见 Lorenz（1963）。因此，研究概率密度的统计演化更为自然，如 Hopf（1952）所示。因此，我们关注的不是方程 d–给出的维常微分方程2.1 ，而是支配状态空间中概率密度演化的 d–维偏微分方程。

为此，我们用 $\boldsymbol{\vartheta} \in \mathbb{R}^d$ 表示状态空间中的固定向量，用 $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, \ldots\ldots, \vartheta_d)$ 表示状态 $\boldsymbol{\vartheta}$ 的各个分量，并用 $\boldsymbol{U} = (U_1, U_2, \ldots\ldots U_d)$ 表示演化规则的各个分量。为阐明概念上的差异，我们不犯通常的记号错误，即令

$s = \mathcal{s}$。方程 2.1 中统计量的演化方程以概率分布函数为特征，

$$\mathcal{P} = \mathcal{P}(\mathcal{s}_1, \mathcal{s}_2, ..., \mathcal{s}_d, t) = \mathcal{P}(\boldsymbol{\mathcal{s}}, t), \tag{2.2}$$

由连续性方程给出

$$\partial_t \mathcal{P} + \sum_{i=1}^{d} \frac{\partial}{\partial \mathcal{s}_i} (U_i(\boldsymbol{\mathcal{s}})\mathcal{P}) = 0. \tag{2.3}$$

上述方程表示概率守恒。它与可压缩纳维–斯托克斯方程中的质量连续性方程完全类似。不过，这里的“质量”被解释为概率密度。分布，$\mathcal{P}$，受流动动力学引导 $\boldsymbol{U}$ 流向状态空间中概率较高的区域。因此，我们关注的不是动力学方程 2.1 所给出的“拉格朗日”视角，而是由 2.3 所给出的“欧拉”视角。

2.2. 无限维动力系统

当基础动力系统是偏微分方程时，我们假设存在定义适当的离散化，将其约化为形式上的 $\mathbb{R}^d$ 维动力系统。我们满足于研究 $\mathbb{R}^d$ 近似的统计量。人们希望不同的离散化能够导出关于基础偏微分方程的相似统计表述；因此，值得像 Hopf（1952）所做的那样，为偏微分方程的类比连续性方程引入记号。

我们谨慎处理记号。前面的 d 维向量现在变成函数空间中的一个向量，其分量由连续指标 $\boldsymbol{x}$（区域 Ω 中的位置）和离散指标 j（感兴趣的场的指标）标记。因此，对于固定指标 i，分量选择 $s_i(t)$ 类似于$)(t)$，其中位置 $\boldsymbol{x}$ 和场指标 $j \in \{1, ..., d_s\}$ 固定，例如不可压缩纳维–斯托克斯方程中的三个速度分量。在离散情形下，单个指标 i 遍历所有速度分量和所有模拟网格点。我们选择不采用通常的记号约定，即使用 $\tilde{\boldsymbol{s}}(\boldsymbol{x}, t)$，因为 $\boldsymbol{s}: \mathbb{R} \to \mathrm{X}$，其中 X 是一个函数空间。另一种记号则暗示一种形式为 $\tilde{\boldsymbol{s}}: \Omega \times \mathbb{R} \to \mathbb{R}^{d_s}$ 的映射。

具体来说，我们考虑一个定义在区域Ω上的状态 $\boldsymbol{s}$ 所满足的偏微分方程，并配以适当的边界条件，

$$\partial_t \boldsymbol{s} = \boldsymbol{\mathcal{U}}[\boldsymbol{s}] \tag{2.4}$$

其中，算子 $\mathrm{UUU} : \mathrm{X} \to \mathrm{X}$ 描述系统的演化。UUU 在位置 $\boldsymbol{x}$和场指标 j处的分量记为 $\mathrm{U}_{(\boldsymbol{x},j)}$。

概率密度泛函的类似演化，

$$\mathcal{P} = \mathcal{P}[\mathcal{s}_{(\boldsymbol{x},1)}, \mathcal{s}_{(\boldsymbol{x},2)}, ..., \mathcal{s}_{(\boldsymbol{x},d_s)}, t] = \mathcal{P}[\boldsymbol{\mathcal{s}}, t] \tag{2.5}$$

记为

$$\partial_t \mathcal{P} + \sum_{j=1}^{d_s} \int_\Omega d\boldsymbol{x} \frac{\delta}{\delta \mathcal{s}_{(\boldsymbol{x},j)}} \left(\mathcal{U}_{(\boldsymbol{x},j)}[\boldsymbol{\mathcal{s}}]\mathcal{P}\right) = 0. \tag{2.6}$$

式2.3 中的求和被2.6中的位置指标积分和场指标求和所取代。此外，偏导数被变分导数所取代。这里的变分导数采用物理学家的定义，也就是说，

$$\frac{\delta \mathcal{s}_{(\boldsymbol{x},i)}}{\delta \mathcal{s}_{(\boldsymbol{y},j)}} = \delta(\boldsymbol{x} - \boldsymbol{y})\delta_{ij} \Leftrightarrow \frac{\partial \mathcal{s}_{i'}}{\partial \mathcal{s}_{j'}} = \delta_{i'j'} \tag{2.7}$$

与离散恒等式类似。在典型的物理学记法中，通常会省略对位置 $\boldsymbol{x}$ 的依赖，并明确写出场变量的各个分量（而不是像我们这里这样采用索引记法），例如，

$$\frac{\delta}{\delta \mathcal{J}_{(\mathbf{y},1)}} \sum_{i=1}^{d_s} \int_\Omega d\boldsymbol{x} (\mathcal{J}_{(\boldsymbol{x},i)})^2 = 2\mathcal{J}_{(\mathbf{y},1)} \Rightarrow \frac{\delta}{\delta u} \int_\Omega (u^2 + v^2 + w^2) = 2u \tag{2.8}$$

对于不可压缩纳维–斯托克斯方程的预报变量 u， v， w

为了推导方程 2.6，我们假设方程 2.1 是方程 2.4的离散化。从方程 2.3出发，首先在索引 i 处引入一个控制体，如下所示 $\Delta \boldsymbol{x}_i$，将方程改写为

$$\partial_t \mathcal{P} + \sum_{i=1}^{d} \Delta \boldsymbol{x}_i \frac{1}{\Delta \boldsymbol{x}_i} \frac{\partial}{\partial \mathcal{J}_i} (U_i(\mathcal{J})\mathcal{P}) = 0. \tag{2.9}$$

在“极限”中，我们有

$$\sum_{i=1}^{d} \Delta \boldsymbol{x}_i \frac{1}{\Delta \boldsymbol{x}_i} \frac{\partial}{\partial \mathcal{J}_i} \to \sum_{j=1}^{d_s} \int_\Omega d\boldsymbol{x} \frac{\delta}{\delta \mathcal{J}_{(\boldsymbol{x},j)}} \text{ and } U_i \to \mathcal{U}_{(\boldsymbol{x},j)} \tag{2.10}$$

本文重点研究如何对方程 2.3 和方程 2.6 进行离散化，这些方程定义在通常被认为具有混沌或湍流特性的状态空间子集上，而我们仅能获得方程 2.1和方程2.4分别给出的轨迹信息。我们采用现代数据驱动方法，从轨迹数据中累积统计量，方法与 Stefan Klus、Péter Koltai 和 Christof Schütte（2016年）；Fernex 等人 （2021年）的方法类似，但采用了不同的划分方法，使我们能够推广到方程 2.6给出的函数情形，并处理任意数量的时间序列数据。

2.3. 有限体积离散化

为聚焦讨论，我们采用有限维情形。不过，这些论证同样适用于无限维情形†。我们首先假设，底层动力系统位于与状态空间紧子集 $\mathcal{M} \subset \mathbb{R}^d$相关联的混沌吸引子上。我们引入 N对 $\mathcal{M}$的分区，记作 $\mathcal{M}_n$，其中 $n = 1,\ldots, N$。

粗粒化离散化变量 P_n 为

$$\int_{\mathcal{M}_n} d\mathcal{J}\mathcal{P} = \int_{\mathcal{M}_n} \mathcal{P} = P_n \tag{2.11}$$

这与有限体积法中的常见做法一致。若上下文明确，我们将省略无穷小状态空间体积$d\mathcal{J}$。此处 $P_n(t)$ 表示在时刻 $\mathcal{M}_n$ 在时间 t内处于状态空间子集中的概率。对分区积分，可得方程 2.3：

$$\frac{d}{dt} P_n = \int_{\mathcal{M}_n} \left[\sum_{i=1}^{d} \frac{\partial}{\partial \mathcal{J}_i} (U_i(\mathcal{J})\mathcal{P}) \right] = \int_{\partial \mathcal{M}_n} \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \mathcal{P} \tag{2.12}$$

其中 $\partial \mathcal{M}_n$ 是分区的边界，而$\hat{n}$ 是法向量。有限体积法的精妙之处在于，通过适当选择数值通量，将方程右端 2.12表示为粗粒化变量 P_n 的形式。

我们以一种迂回的方式计算数值通量。我们列出数值离散化的一些期望条件。

† 当然，这是假设极限在某种意义上是合理的。此外，在离散情形下，状态空间中的无穷小体积 $d\mathcal{J}$ 在函数空间情形下则表示为 $\mathrm{D}[\mathcal{J}]$。

《流体专题》文章不得超过此页长度

（i）离散方程用瞬时粗粒化变量 P_n 表示。

(ii) 离散方程是线性的，这与无限维方程类似。(iii) 方程必须守恒概率。(iv) 概率必须始终为正。

前两项要求指出，

$$\int_{\partial \mathcal{M}_n} \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \mathcal{P} \approx \sum_m Q_{nm} P_m \tag{2.13}$$

对于某个矩阵 Q。因此，我们希望得到如下形式的方程

$$\frac{d}{dt} \hat{P}_n = \sum_m Q_{nm} \hat{P}_m \tag{2.14}$$

ˆ 我们引入了一个“帽子”来区分数值近似 P_n 与精确解P_n。第三项要求指出，

$$\sum_n \hat{P}_n = 1 \Rightarrow \frac{d}{dt} \sum_n \hat{P}_n = 0 = \sum_{mn} Q_{nm} \hat{P}_m \tag{2.15}$$

ˆ 对每个 P_m，因此

$$\sum_n Q_{nm} = 0 \tag{2.16}$$

对于每个 m，也就是说，矩阵的列之和必须为零。此外，最后一个要求指出，Q_{nm} 的非对角项必须全部为正。为了说明这一点，我们采用反证法。假设存在一个负的非对角元素，不失一般性地设为分量 Q_{21}。那么，如果在零时刻，我们的概率向量从 ˆ $P_1(0) = 1$ 开始，且对于 $n > 1$，ˆ $P_n(0) = 0$，经过一个无穷小时间步 dt 后，我们有

$$\hat{P}_2(dt) = dt \sum_m Q_{2m} \hat{P}_m(0) = Q_{21} < 0, \tag{2.17}$$

这会产生矛盾，因为概率必须始终保持为正。在所有要求中，最容易舍弃的正是第四项，因为不可能构造出既保正又线性的高阶离散化（Zhang & Shu（2011））。

这四项要求结合起来，足以将矩阵 Q 识别为有限状态空间上连续时间马尔可夫过程的生成元。这一观察构成了数据驱动方法的基础，用于对方程 2.3和2.6进行离散化。矩阵的对角元素与在某个分区中停留的平均时间有关，而列中的非对角元素则与离开某个分区后进入给定分区的概率成正比。这意味着，我们通过对运动方程进行蒙特卡罗积分来构造边界上的数值通量。

直观地说，当一个动力系统通过状态空间划分进入时，它便与处于该分区内联系起来。轨迹在一个分区内停留的时间称为停留时间，随后它最终将从该分区退出，进入状态空间中的某个其他分区。对运动方程进行足够长时间的积分，可以构建停留时间分布，以及前往状态空间中不同分区的退出概率。此外，为了进行计算，我们将状态空间的每个区域与一个“单元中心”相关联，在本文中称其为马尔可夫状态。马尔可夫状态将作为该状态空间区域内分布的狄拉克δ函数近似的中心。利用转移矩阵 Q

以及与一个分区相关联的马尔可夫状态，我们可以计算任何感兴趣变量的矩、稳态分布和自相关。

2.4. 时间计算与系综计算

至此，我们已经讨论了动力系统的统计量方程、无限维情形的记法，以及如何将连续时间马尔可夫过程与有限体积离散化的连续性方程联系起来。现在，我们将讨论如何基于离散化进行统计计算，以及如何确认离散化捕捉到了底层连续性方程的统计量。简而言之，我们将比较时间平均与系综平均，以及自相关的类似计算。

我们必须引入额外的记号。如前所述，我们假设存在一个遍历混沌吸引子，因此存在唯一的不变测度，我们将其记为 $\mathscr{P}(\boldsymbol{s})$。相对于分区 M_n 的条件不变测度记为 $\mathscr{P}(\boldsymbol{s}|\mathrm{M}_n)$，而在分区 M_n 中找到某个状态的概率为 $\mathbb{P}(\mathrm{M}_n)$，因此不变测度可分解为

$$\mathscr{P}(\boldsymbol{s}) = \sum_n \mathscr{P}(\boldsymbol{s}|\mathcal{M}_n)\mathbb{P}(\mathcal{M}_n). \tag{2.18}$$

此外，我们引入传递算子 $\mathscr{T}^\tau$的记号，该算子通过以下关系定义

$$\mathcal{P}(\boldsymbol{s}, t+\tau) = \mathscr{T}^\tau \mathcal{P}(\boldsymbol{s}, t) \tag{2.19}$$

其中 $\mathcal{P}$ 是方程 2.3 的一个解。因此，传递算子规定了如何通过方程 $\mathcal{P}$，依据方程 2.3 将密度演化到未来的某个时刻 τ。此外，

$$\lim_{\tau\to\infty} \mathscr{T}^\tau \mathcal{P}(\boldsymbol{s}, t) = \mathscr{P}(\boldsymbol{s}) \tag{2.20}$$

对于任意密度 $\mathcal{P}(\boldsymbol{s},\ t)$，包括 $\delta(\boldsymbol{s}-\boldsymbol{s}')$ 对于初始状态 $\boldsymbol{s}' \in \mathcal{M}$，由遍历性假设可得†。

对于可观测量 $g : \mathrm{M} \to \mathbb{R}$，我们计算长时间平均

$$\langle g\rangle_T = \lim_{T\to\infty} \frac{1}{T}\int_0^T g(\boldsymbol{s}(t))dt \tag{2.21}$$

并与系综平均进行比较

$$\langle g\rangle_E = \int_{\mathcal{M}} g(\boldsymbol{s})\mathscr{P}(\boldsymbol{s}) \tag{2.22}$$

此外，我们比较时间相关观测量。时间序列的计算为

$$R_T(g,\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T g(\boldsymbol{s}(t+\tau))g(\boldsymbol{s}(t))dt, \tag{2.23}$$

˜由此得到自协方差 C_T 和自相关 C_T，

$$C_T(g,\tau) \equiv R_T(g,\tau) - \langle g\rangle_T^2 \text{ and } \tilde{C}_T(g,\tau) \equiv C_T(g,\tau)/C_T(g,0). \tag{2.24}$$

系综平均版本需要更多说明。我们将变量 $g(s(t))$与 $g(s(t+\tau))$ 进行关联，这涉及两个变量的联合分布。为了陈述系综平均版本，我们首先回顾一个关于随机变量 X，Y及其联合密度 $\rho(x,\ y)$ 的事实，

†这里对极限的处理有些不严谨，但应理解为，趋于狄拉克δ函数密度的极限是最后取的极限。

条件密度 $\rho(x|y)$和边际密度 $\rho_y(y)$。期望值、两个可观测量的相关性，计算为

$$\langle g(X)g(Y)\rangle = \iint dxdyg(x)g(y)\rho(x,y) = \iint dxdyg(x)g(y)\rho(x|y)\rho_y(y) \tag{2.25}$$

$$= \int dyg(y)\rho_y(y)\left[\int dxg(x)\rho(x|y)\right]. \tag{2.26}$$

要将上述计算转换到当前情形，我们将ρ_y 视为不变测度，$\mathscr{P}$。条件分布 $\rho(x|y)$被视为未来 τ 时刻的概率密度，已知其在 $\tau = 0$时刻最初处于状态 $\boldsymbol{s}$ 。

因此，在当前情形下， $\rho(x|y)$变为 $\mathscr{T}^\tau\delta(\boldsymbol{s}-\boldsymbol{s}')$，其中 δ 函数密度精确表示我们对 $\tau = 0$时刻状态的了解。总之，系综时间自相关计算如下：

$$R_E(g,\tau) = \int_{\mathcal{M}} d\boldsymbol{s}'g(\boldsymbol{s}')\mathscr{P}(\boldsymbol{s}')\left[\int_{\mathcal{M}} d\boldsymbol{s}g(\boldsymbol{s})\mathscr{T}^\tau\delta(\boldsymbol{s}-\boldsymbol{s}')\right]. \tag{2.27}$$

˜自协方差 C_E和自相关 C_E分别为

$$C_E(g,\tau) \equiv R_E(g,\tau) - \langle g\rangle_E^2 \text{ and } \tilde{C}_E(g,\tau) \equiv C_E(g,\tau)/C_E(g,0). \tag{2.28}$$

这些计算概述了我们希望进行比较的精确关系。不过，首先，我们将通过长时间的有限轨迹来近似时间平均，并通过第 2.3 节中的有限体积离散化来近似系综平均。

2.5. 时间计算与系综计算的近似

前一节给出了我们希望据以进行计算的数学理想情形；然而，由于我们采用数据驱动的构造，因此必须在有限维空间和有限时间范围内进行计算。

给定一个状态在均匀间隔时刻 times t_n 至 N_t（时间间隔为 Δt）的时间序列，我们将可观测量 g 的均值和长时间平均近似为

$$\langle g\rangle_T \approx \frac{1}{N_t}\sum_{n=1}^{N_t} g(\boldsymbol{s}(t_n)) \tag{2.29}$$

$$R_T(g,\tau) \approx \frac{1}{N_t'}\sum_{n=1}^{N_t'} g(\boldsymbol{s}(t_n + \text{round}(\tau/\Delta t)\Delta t))g(\boldsymbol{s}(t_n)) \tag{2.30}$$

其中 round 函数计算最接近的整数，且 $N_t' = N_t - \text{round}(\tau/\Delta t)$。

我们使用第 2.3 节中的构造来计算系综平均。回想一下，最后我们用一个矩阵 Q 近似了该过程的生成元，该矩阵描述了与状态空间分区相关的概率演化。此外，我们选取一个与分区 $\mathcal{M}n$ 相关的状态，作为“单元中心”进行计算。我们用上标表示不同的状态，因为下标保留用于计算状态的各个分量。此外，我们不要求马尔可夫状态 $\boldsymbol{\sigma}^n$ 属于分区 $\mathcal{M}_n$。例如，我们可以选择 $\boldsymbol{\sigma}^n$ 作为动力系统的不动点，或者选择混沌吸引子内一条周期轨道上的若干点$\mathcal{M}$。

可观测量的系综平均通过利用不变测度的分解来计算2.18，但随后对其进行近似

$$\mathscr{P}(\boldsymbol{s}|\mathcal{M}_n) \approx \delta(\boldsymbol{s}-\boldsymbol{\sigma}^n) \tag{2.31}$$

也就是 一种简单但粗略的近似。因此，系综平均的计算方式是将一个 s

$$\langle g\rangle_E = \int_{\mathcal{M}} g(\boldsymbol{\sigma}) \left[\sum_n \mathcal{P}\left(\boldsymbol{\sigma}|\mathcal{M}_n\right)\mathbb{P}(\mathcal{M}_n)\right] = \sum_n \left[\int_{\mathcal{M}} g(\boldsymbol{\sigma})\mathcal{P}\left(\boldsymbol{\sigma}|\mathcal{M}_n\right)\right]\mathbb{P}(\mathcal{M}_n) \tag{2.32}$$

$$\approx \sum_n \left[\int_{\mathcal{M}} g(\boldsymbol{\sigma})\delta\left(\boldsymbol{\sigma}-\boldsymbol{\sigma}^n\right)\right]\mathbb{P}(\mathcal{M}_n) = \sum_n g(\boldsymbol{\sigma}^n)\mathbb{P}(\mathcal{M}_n) \tag{2.33}$$

对于时间自相关的系综平均版本，除了近似不变测度之外，我们还必须近似作用于状态的 δ 函数密度的转移算子， $\mathcal{T}^\tau\delta(\boldsymbol{\sigma}-\boldsymbol{\sigma}')$。我们计算

$$R_E(g,\tau) = \int_{\mathcal{M}}\int_{\mathcal{M}} d\boldsymbol{\sigma}' d\boldsymbol{\sigma} g(\boldsymbol{\sigma}')\mathcal{P}(\boldsymbol{\sigma}')g(\boldsymbol{\sigma})\mathcal{T}^\tau\delta(\boldsymbol{\sigma}-\boldsymbol{\sigma}') \tag{2.34}$$

$$\approx \sum_n \int_{\mathcal{M}}\int_{\mathcal{M}} d\boldsymbol{\sigma}' d\boldsymbol{\sigma} g(\boldsymbol{\sigma}')\delta(\boldsymbol{\sigma}'-\boldsymbol{\sigma}^n)\mathbb{P}(\mathcal{M}_n)g(\boldsymbol{\sigma})\mathcal{T}^\tau\delta(\boldsymbol{\sigma}-\boldsymbol{\sigma}') \tag{2.35}$$

$$= \sum_n g(\boldsymbol{\sigma}^n)\mathbb{P}(\mathcal{M}_n)\int_{\mathcal{M}} d\boldsymbol{\sigma} g(\boldsymbol{\sigma})\mathcal{T}^\tau\delta(\boldsymbol{\sigma}-\boldsymbol{\sigma}^n) \tag{2.36}$$

然后进一步近似

$$\mathcal{T}^\tau\delta(\boldsymbol{\sigma}-\boldsymbol{\sigma}^n) \approx \sum_{m=1}^N \delta(\boldsymbol{\sigma}^m-\boldsymbol{\sigma})[\exp(Q\tau)]_{mn}. \tag{2.37}$$

矩阵指数 $\exp(Q\tau)$ 是离散系统中类似的佩龙–弗罗贝尼乌斯／转移算子。该矩阵是一个（列）随机矩阵，其元素之和为一。此近似的直观解释是，将以状态 $\boldsymbol{\sigma}^n$ 为中心的狄拉克分布在传递算子下的向前演化视为以状态 $\boldsymbol{\sigma}^m$为中心的狄拉克函数的加权和。将各部分组合起来，得到

$$R_E(g,\tau) \approx \sum_{n=1}^N g(\boldsymbol{\sigma}^n)\mathbb{P}(\mathcal{M}_n)\left[\sum_{m=1}^N g(\boldsymbol{\sigma}^m)[\exp(Q\tau)]_{mn}\right] \tag{2.38}$$

此外，所有协方差和相关性均利用上述近似进行计算。至此，我们完成了关于如何从生成元的有限体积离散化中近似求得系综平均和协方差的讨论。然而，如何根据数据构造矩阵和马尔可夫状态仍有待说明。生成元的构造将在下一节讨论。

3. 方法论

本节将概述如何根据轨迹数据构造近似生成元的一般方法。离散化中最关键的环节是定义嵌入函数 $\mathrm{E}:\ \mathbb{R}^d \to \{1,2,...,N\}$，它将任意状态 $\boldsymbol{\sigma}\in\mathbb{R}^d$ 映射为一个整数 $n\in\{1,2,...,N\}$。该函数通过以下关系隐式定义了一个分区

$$\mathcal{M}_j = \{\boldsymbol{\sigma}:\mathcal{E}(\boldsymbol{\sigma}) = j \text{ for each } \boldsymbol{\sigma}\in\mathbb{R}^d\}\cap\mathcal{M}. \tag{3.1}$$

Th与流形 M 的交集对该方法的成功至关重要。

此外，马尔可夫状态（单元中心）的选择满足 $\mathcal{E}(\boldsymbol{\sigma}^n) = n$ 对每个$n\in\{1,...,N\}$。定义嵌入函数时有极大的自由度，我们将在第 4节介绍三个例子。我们还将讨论

第 5 节将讨论实际考量与推广。可以使用 K均值算法同时求解嵌入函数和马尔可夫状态，参见 Lloyd (1982)，但我们不希望在此将自己限制于这一选择。嵌入函数是一个分类器（机器学习意义上的），用于区分不同的流状态，并以整数作为类别标签。暂且假设已有这样一个函数，并将重点放在构造生成元 Q。

马尔可夫嵌入函数E将动力学轨迹转换为整数序列，我们将其解释为有限状态空间上的马尔可夫过程的实现。在此阶段，可以采用传统方法从数据构造转移/佩龙-弗罗贝尼乌斯算子，参见 Stefan Klus、Péter Koltai 和 Christof Schütte（2016年）；Fernex等（2021年）。鉴于我们关注的是构造连续时间马尔可夫过程，该算法将按照 Froyland 等（2013年）的方法进行修改。要构造 Q，必须为每个分区计算两个量

(i) 停留时间：动力学轨迹在分区 M_n 中退出前所停留的时间。

(ii) 退出概率：退出分区 M_j 时，从分区 M_j 移动到 M_i 的概率。

设 T_j 为与分区 j 相关联的停留时间分布，E_{ij} 表示退出概率矩阵的元素。按照我们的约定，对于所有 i，$\sum_j E_{ij} = 1$ 和 $E_{ii} = 0$。矩阵 Q_{ij} 的元素构造如下

$$Q_{ij} = E_{ij}/\langle T_j\rangle \text{ for } i \neq j \text{ and } Q_{jj} = -1/\langle T_j\rangle \text{ for } j \in \{1, ..., N\} \tag{3.2}$$

wh其中 $\langle T_j\rangle$ 表示分区停留时间分布的期望值 j.

在接下来的小节中，我们将概述如何根据有限数据经验构造矩阵，以及如何通过贝叶斯方法纳入有限抽样效应导致的不确定性。采用后一种方法时，我们不将 Q_{ij} 矩阵的各项视为确定的数值，而是将其视为分布。由此得到的生成元随机矩阵表示纳入了不确定性。

3. 经验构造

我们从生成元的经验构造开始。只需关注与矩阵 Q_{ij}第 j 列相关联的分区 j。为了计算经验停留时间分布和经验均值，我们统计状态 j 在转移到状态 $i \neq j$之前出现的次数。例如，假设有三个状态 $j = 1$，并考虑以下由马尔可夫嵌入应用于时间间隔为 Δt 的时间序列后得到的整数序列，

$$1, 1, 1, 2, 2, 1, 1, 3, 1, 2, 1, 1. \tag{3.3}$$

我们将该序列分组如下

$$(1, 1, 1), 2, 2, (1, 1), 3, (1), 2, (1, 1) \tag{3.4}$$

以确定停留时间。因此，状态 1 的停留时间为

$$3\Delta t, 2\Delta t, \Delta t, 2\Delta t \tag{3.5}$$

其经验期望值为 $2\Delta t$，这意味着转移率为 $1/(2\Delta t)$。

要计算分区 j 的退出概率，我们统计转变到分区 i 的次数，并除以转变总数。在此示例中，要计算分区1转入分区 2 或 3 的退出概率，我们将它们按如下方式组合在一起

$$1, 1, (1, 2,)2, 1, (1, 3), (1, 2), 1, 1 \tag{3.6}$$

因此我们看到了三个出口，其中两个通向状态 2，一个通向状态 3；因此退出概率为 $E_{21} = 2/3$ 和 $E_{31} = 1/3$。

其余状态也以类似方式构造，从而得到矩阵

$$Q = \frac{1}{\Delta t}\begin{bmatrix} -1/2 & 2/3 & 1 \\ 1/3 & -2/3 & 0 \\ 1/6 & 0 & -1 \end{bmatrix} \tag{3.7}$$

按照目前的实现方式，生成元的精度仅为 Δt 阶，因为我们不会通过时间插值来求出“精确的”停留时间。我们并未着力改进这一点，因为我们认为，误差的主要来源是有限抽样效应。在下一节中，我们将基于有限抽样和贝叶斯框架，为经验构造补充不确定性估计。

3. 贝叶斯构造

要通过贝叶斯构造得到生成元，我们需要四个要素。

(i) 停留时间的似然分布 (ii) 与停留时间相关的转移率的先验分布 (iii) 退出概率的似然分布 (iv) 与退出概率相关的概率值的先验分布

我们作出一些假设，使得可以针对每一列独立地从连续时间马尔可夫过程中抽样

(i) 停留时间的似然分布呈指数分布，其速率参数为 λ_i

(ii) 退出概率的似然分布为参数为 $\vec{p} \in [0,1]^{N-1}$、满足关系 $\sum_{i=1}^{N-1} p_i = 1$ 的多项分布

(iii) 指数分布的速率参数的先验分布服从参数为 $(\alpha,\ \beta)$ 的伽马分布，记作 $\Gamma(\alpha,\ \beta)$

(iv) 多项分布中概率的先验分布来自参数向量为 $\vec{\alpha}$、长度为 $N-1$ 的狄利克雷分布，记作 Dirichlet($\vec{\alpha}$)

这些分布是共轭先验，因此后验分布与先验分布属于同一族；参见 Gelman 等人（2013年）。例如，在这种构造下，3×3 矩阵的形式始终为

$$Q = \begin{bmatrix} -1 & [\vec{D}_2]_1 & [\vec{D}_3]_1 \\ [\vec{D}_1]_1 & -1 & [\vec{D}_3]_2 \\ [\vec{D}_1]_2 & [\vec{D}_2]_2 & -1 \end{bmatrix} \begin{bmatrix} G_1 & 0 & 0 \\ 0 & G_2 & 0 \\ 0 & 0 & G_3 \end{bmatrix} \tag{3.8}$$

其中，$G_i \sim \Gamma(\alpha_i,\ \beta_i)$、$\vec{D}_i \sim$狄利克雷$(\vec{\alpha}_i)$，而 $[\vec{D}_i]_j$ 表示随机向量 D_i的第 j个分量。

数据采集时，每列的参数 $(\alpha_i,\ \beta_i)$ 和 $\vec{\alpha}_i$ 都会根据贝叶斯法则进行更新。例如，假设我们观察到与分区 i相关的以下经验计数

(i) M 离开分区 $\vec{i}$ (ii) $[M]_j$ 从分区 i 离开，进入分区 j。

Rapids 文章不得超过此页长度

(iii) T_1, T_2, ... T_M 经验观测的停留时间

并且我们以（α^0, β^0）和 $\vec{\alpha}^0$ 作为先验分布的参数。关系$\sum_j [\vec{M}]_j = M$ 成立†。后验分布的参数 (α^1, β^1) 和 $\vec{\alpha}^1$ 为

$$\alpha^1 = \alpha^0 + M \, , \beta^1 = \beta^0 + \sum_i^M \hat{T}_i, \text{ and } \vec{\alpha}^1 = \vec{\alpha}^0 + \vec{M}. \tag{3.9}$$

当 α^0, β^0 和 $|\vec{\alpha}^0|$ 趋近于零时，前一节介绍的经验方法与贝叶斯方法得到的期望值一致。

当前方法是构造具有量化不确定性的矩阵的众多方法之一，但并非唯一方法‡。当前构造在许多方面并不完善（例如，停留时间不服从指数分布，或系统在无穷小步长下不是马尔可夫的），但我们坚持认为，某种不确定性的量化总比没有好。我们使用不确定性量化来排除虚假结果，而不是增强对推断正确性的信心。佩龙–弗罗贝尼乌斯／转移算子也可以采用此处的方法，即使用多项分布作为似然函数，并使用狄利克雷分布作为先验。

下面给出运用该理论和方法构造生成元的数据驱动近似并量化不确定性的示例。

4. 应用

我们将上一节的方法应用于两个不同的系统。第一个系统的动力学由洛伦兹方程描述：

$$\dot{x} = -\sigma(x - y) \tag{4.1}$$
$$\dot{y} = -y + (r - z)x \tag{4.2}$$
$$\dot{z} = -bz + xy \tag{4.3}$$

其中，我们令 $x = s_1$、 $y = s_2$、 $z = s_3$。相应的连续性方程为

$$\partial_t \mathcal{P} + \partial_x \left([-\sigma(x - y)] \, \mathcal{P} \right) + \partial_y \left([-y + (r - z)x \,] \mathcal{P} \right) + \partial_z \left([-bz + xy] \mathcal{P} \right) = 0 \tag{4.4}$$

其中，我们使用记号 x 表示 ϑ_1， y 表示 ϑ_2， z 表示 ϑ_3。

我们考虑的第二个系统是球面上包含旋转和重力的可压缩欧拉方程的通量差分间断 Galerkin 离散。选取的预报变量为密度 ρ、动量 $\rho\boldsymbol{u}$和总能量 ρe。其动力学由以下方程给出：

$$\partial_t \rho = -\nabla \cdot (\rho \boldsymbol{u}) \tag{4.5}$$
$$\partial_t (\rho \boldsymbol{u}) = -\nabla \cdot (\boldsymbol{u} \otimes \rho \boldsymbol{u} + p\mathbb{I}) - \rho \nabla \Phi + \mathbf{S}_{\rho \boldsymbol{u}} (\rho, \rho \boldsymbol{u}, \rho e) \tag{4.6}$$
$$\partial_t (\rho e) = -\nabla \cdot (\boldsymbol{u} \, (p + \rho e)) + S_{\rho e} \, (\rho, \rho \boldsymbol{u}, \rho e) \, . \tag{4.7}$$

其中源项的具体内容，$\mathbf{S}_{\rho \boldsymbol{u}}$ 和 $S_{\rho e}$，详见附录C， Φ 为位势，而 p 为气压。我们与先验记法作如下对应$s_{(\boldsymbol{x},1)} = \rho$， $s_{(\boldsymbol{x},2)} = \rho u$， $s_{(\boldsymbol{x},3)} = \rho v$， $s_{(\boldsymbol{x},4)} = \rho w$， $s_{(\boldsymbol{x},5)} = \rho e$。相应的连续性

†从技术上说，这里可能存在一个“差一”错误，为便于呈现我们忽略了它。

‡ 例如，可以考虑列之间的相关性，或者将每个出口概率的似然单独视为伯努利分布，并采用与之共轭的贝塔分布先验。

方程是

$$\partial_t \mathcal{P} + \int_\Omega \frac{\delta}{\delta \rho} \left[-\mathrm{div}\left(\varrho \boldsymbol{u}\right) \mathcal{P} \right] \tag{4.8}$$

$$+ \int_\Omega \frac{\delta}{\delta \varrho u} \left[\left(-\mathrm{div}\left(\boldsymbol{u} \varrho u + p\hat{x}\right) - \varrho \hat{x} \cdot \mathrm{grad}(\Phi) + S_{\rho u} \right) \mathcal{P} \right] \tag{4.9}$$

$$+ \int_\Omega \frac{\delta}{\delta \varrho v} \left[\left(-\mathrm{div}\left(\boldsymbol{u} \varrho v + p\hat{y}\right) - \varrho \hat{y} \cdot \mathrm{grad}(\Phi) + S_{\rho v} \right) \mathcal{P} \right] \tag{4.10}$$

$$+ \int_\Omega \frac{\delta}{\delta \varrho w} \left[\left(-\mathrm{div}\left(\boldsymbol{u} \varrho w + p\hat{z}\right) - \varrho \hat{z} \cdot \mathrm{grad}(\Phi) + S_{\rho w} \right) \mathcal{P} \right] \tag{4.11}$$

$$+ \int_\Omega \frac{\delta}{\delta \varrho e} \left[\left(-\mathrm{div}\left(\boldsymbol{u} \left(p + \varrho e\right)\right) + S_{\rho e} \right) \mathcal{P} \right] = 0, \tag{4.12}$$

其中我们建立对应关系 $\mathfrak{z}_{(\boldsymbol{x},1)} = \varrho$, $\mathfrak{z}_{(\boldsymbol{x},2)} = \varrho u$, $\mathfrak{z}_{(\boldsymbol{x},2)} = \varrho v$, $\mathfrak{z}_{(\boldsymbol{x},2)} = \varrho w$,$\mathfrak{z}_{(\boldsymbol{x},5)} = \varrho \mathrm{e}$, 以及 $\boldsymbol{u} = (u, v, w)$。请注意，源项$\mathbf{S}_{\rho \boldsymbol{u}}$已被拆分为三项： $S_{\rho u}$、 $S_{\rho v}$和 $S_{\rho w}$。此外，我们使用“grad”和“div”记号，以强调映射函数与函数之间的联系。

可压缩欧拉方程的数值离散化在附录C中概述，但对于当前目的无关紧要。相反，我们将该系统视为一个有限但高维的空间，其中 $d = 1,\ 481,\ 760$ 在我们的具体情况下。

我们选择 Held-Suarez 测试案例进行分析，因为它具有湍流，已得到大气科学界的广泛研究，而且是一种与地球物理相关的配置，产生的风场和温度模式与地球上观测到的相似。此外，其统计量在多种离散化策略、耗散机制和方程表述下都很稳健。它不会呈现亚稳态，因此可作为检验该方法的严格测试。

为离散化第一个连续性方程，我们使用三个分区；对于后者，我们采用两种划分策略：先使用400个分区，之后使用100个分区。嵌入函数 E 和马尔可夫状态 $\boldsymbol{\sigma}^n$ 的选择将在相关章节中说明。

4.1. *Lorenz*：不动点分区

我们为洛伦兹系统选择经典参数值 $r = 28$, $\sigma = 10$, 以及 $b = 8/3$；该系统已知会呈现混沌解。选定马尔可夫状态$\boldsymbol{\sigma}^n$和嵌入函数 $\mathcal{E}$后，我们依据第 3 节的方法自动构造生成元。我们用以下虚构内容来指导我们的选择：

据说，流动的相干结构组织并引导着混沌的动力学。当一条轨迹在状态空间中游走时，它会在相干结构附近停留不成比例的时间，并继承它们的性质。随后，相干结构将其行为印刻在混沌轨迹上，并体现在系综平均中。因此，混沌平均可以通过更简单结构之间的转变来理解。这个富有画面感的故事促使人们使用不动点作为马尔可夫状态，

$$\boldsymbol{\sigma}^1 = [-\sqrt{72}, -\sqrt{72}, 27], \tag{4.13}$$

$$\boldsymbol{\sigma}^2 = [0, 0, 0], \tag{4.14}$$

$$\boldsymbol{\sigma}^3 = [\sqrt{72}, \sqrt{72}, 27], \tag{4.15}$$

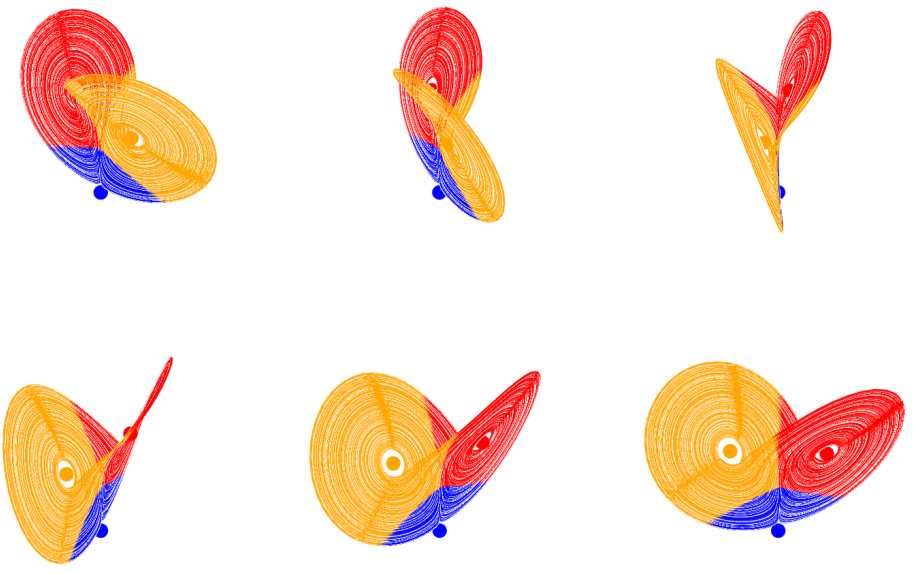

图 1：洛伦兹不动点划分。这里我们从多个角度展示生成的划分。颜色对应于与给定不动点“最接近”的轨迹所关联的不同分区。

并根据最近的不动点划分状态空间，

$$\mathcal{E}(\boldsymbol{\imath}) = \begin{cases} 1 & \text{if } \|\boldsymbol{\imath} - \boldsymbol{\sigma}^1\| < \|\boldsymbol{\imath} - \boldsymbol{\sigma}^2\| \text{ and } \|\boldsymbol{\imath} - \boldsymbol{\sigma}^3\| \\ 2 & \text{if } \|\boldsymbol{\imath} - \boldsymbol{\sigma}^2\| < \|\boldsymbol{\imath} - \boldsymbol{\sigma}^1\| \text{ and } \|\boldsymbol{\imath} - \boldsymbol{\sigma}^3\| \\ 3 & \text{if } \|\boldsymbol{\imath} - \boldsymbol{\sigma}^3\| < \|\boldsymbol{\imath} - \boldsymbol{\sigma}^1\| \text{ and } \|\boldsymbol{\imath} - \boldsymbol{\sigma}^2\| \end{cases} \tag{4.16}$$

其中 $\|\cdot\|$ 表示标准欧几里得范数。嵌入函数通过将轨迹与最近的不动点关联起来，确定该划分。换言之，这一划分策略是混沌吸引子 $\mathcal{M}$ 与整个状态空间上的 Voronoi 镶嵌 $\mathbb{R}^3$ 的交集。我们从多个角度展示了这一选择所诱导的划分，见图1 。各区域按照最近的不动点进行颜色编码。

我们使用四阶龙格－库塔法和时间步长 $\Delta t = 5 \times 10^{-3}$，根据洛伦兹方程构造时间序列。我们将初始条件设为$(x(0),\ y(0),\ z(0)) = (14,\ 20,\ 27)$，并积分至时间 $T = 10^5$，从而得到 2×10^7 个时间快照。在每个时刻，我们应用嵌入函数，生成表示划分动力学的整数序列†。图2 将这一过程可视化。

根据整数序列，我们应用第3.2 节中的方法，构造带有量化不确定性的生成元的数据驱动近似。对于先验分布，我们使用无信息先验，使随机矩阵的均值与3.1节中的经验构造一致。生成元每个元素的均值（保留两位小数）

† 可以将其视为定义一个符号序列。

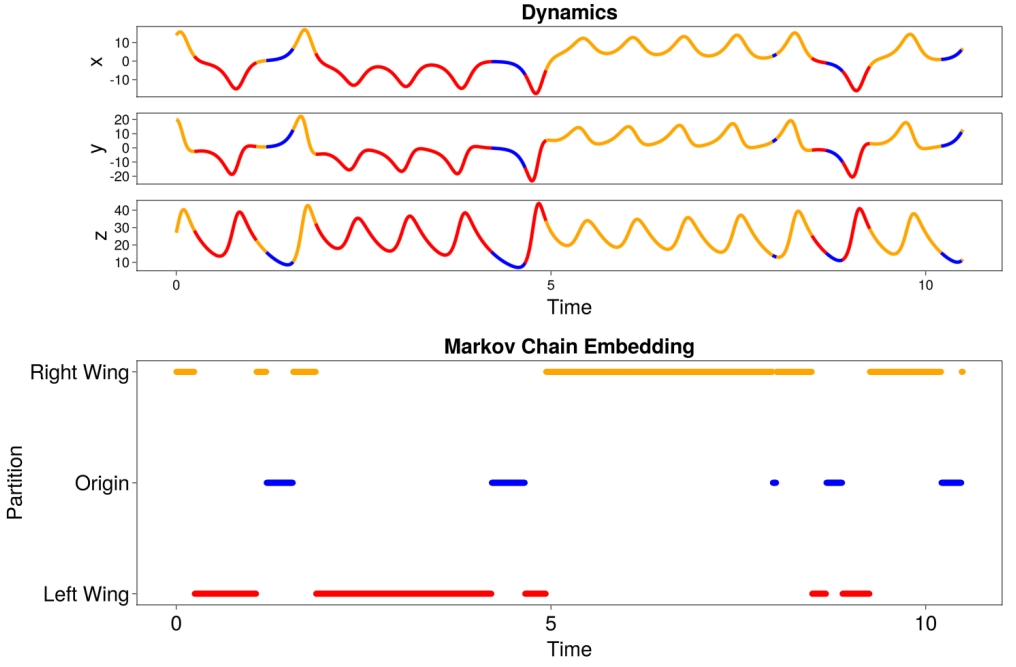

图2：Lorenz固定点划分马尔可夫链嵌入。x, y, z 变量的动力学显示在顶部的3个面板中，相关嵌入显示在底部面板中。随着动力学轨迹在状态空间中移动，它会根据其与最近不动点的接近程度进行标记。

位置）是

$$\langle Q \rangle = \begin{bmatrix} -1.17 & 1.93 & 0.65 \\ 0.52 & -3.86 & 0.52 \\ 0.65 & 1.93 & -1.17 \end{bmatrix} \tag{4.17}$$

矩阵中表面上的对称性是由截断至小数点后两位以及大量数据造成的。在附录B中，我们展示了如何纳入 Lorenz 方程的对称性，并报告系综平均统计量。

使用随机矩阵表示不确定性的用途总结于图3中。图中展示了不同时段子集所对应的各个矩阵元素的分布。使用较少的数据（以较短的采集时间表示， T）会导致矩阵元素存在显著的不确定性。此外，使用互不连续的时段子集表明矩阵元素呈现出明显的收敛。

现在我们可以计算统计量了。为简便起见，我们仅报告由生成元的均值 $\langle Q \rangle$计算得到的一阶、二阶和三阶矩。与特征值 $\lambda = 0$ 对应的 $\langle Q \rangle$ 的稳态分布报告为小数点后两位：

$$[\mathbb{P}(\mathcal{M}_1), \mathbb{P}(\mathcal{M}_2), \mathbb{P}(\mathcal{M}_3)] \approx [0.44, 0.12, 0.44] \tag{4.18}$$

由此，我们利用第2.4节中的近似和马尔可夫状态 $\boldsymbol{\sigma}^n$ 计算 $n = 1$、2、3的稳态统计量。具体而言，系综平均

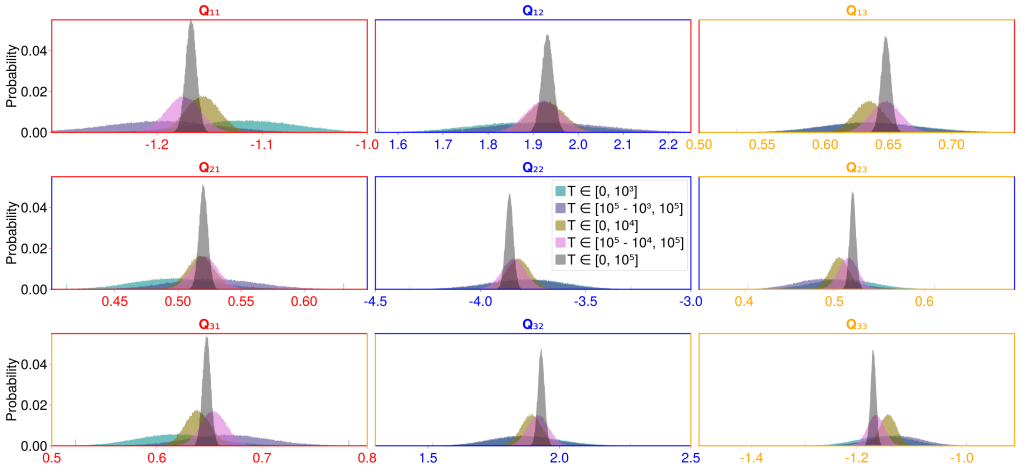

图 3：生成元的洛伦兹不动点划分分布。上图展示了 3×3 生成元各项的不确定性估计。图中各面板的分布与矩阵各项之间存在一一对应关系。面板内不同颜色的分布表示基于可用数据量对各项所作的不同估计，此处以洛伦兹系统的模拟时间来表示可用数据量。我们看到，随着模拟时间间隔增加、数据量增多，我们对矩阵各项的估计也更加确信。此外，各分布的范围彼此重叠。

可观测量的，

$$g^1(\boldsymbol{\jmath}) = \jmath_3 = x,\ g^2(\boldsymbol{\jmath}) = (\jmath_3)^2 = x^2,\ \text{or } g^3(\boldsymbol{\jmath}) = (\jmath_1)^2\jmath_3 = x^2x \tag{4.19}$$

通过方程 2.33进行近似，现为方便起见再次列出，

$$\langle g^j\rangle_E = g^j(\boldsymbol{\sigma}^1)\mathbb{P}(\mathcal{M}_1) + g^j(\boldsymbol{\sigma}^2)\mathbb{P}(\mathcal{M}_2) + g^j(\boldsymbol{\sigma}^3)\mathbb{P}(\mathcal{M}_3) \text{ for each } j \tag{4.20}$$

得出

$$\langle x\rangle_E \approx 27\times0.44 + 0\times0.12 + 27\times0.44 \approx 24 \tag{4.21}$$

$$\langle x^2\rangle_E \approx 27^2\times0.44 + 0^2\times0.12 + 27^2\times0.44 \approx 642 \tag{4.22}$$

$$\langle x^2x\rangle_E \approx \left(-\sqrt{72}\right)^2\times27\times0.44 + 0^3\times0.12 + \left(\sqrt{72}\right)^2\times27\times0.44 \approx 1711. \tag{4.23}$$

表1 展示了时间平均和系综平均的结果。所有平均值之间都存在对应关系，最显著的差异出现在涉及y^2 项的平均值中，其相对误差在 25%以内。动力系统的不动点具有独特性，因为它们满足与统计稳态相同的所有动力学平衡关系。尽管我们重点讨论了矩，但也可以比较任意可观测量的统计量，例如，

$$\langle x\log(x)\rangle_E \approx 78.4 \text{ and } \langle z\log(z)\rangle_T \approx 76.0, \tag{4.24}$$

其中我们使用了 $z\log(z) \to 0 \text{ as } z\to0$。根据对称性，可以预期:

$$\langle x\rangle = \langle y\rangle = \langle xz\rangle = \langle yz\rangle = \langle yyy\rangle = \langle xxy\rangle = \langle xyy\rangle = \langle xzz\rangle = \langle yzz\rangle = 0 \tag{4.25}$$

但有限抽样效应会阻止这种情况发生。正如附录B中所做的那样，纳入对称性后，系综计算可以达到机器精度。

除了包含有关稳态分布的信息外，生成元 Q提供时间信息：自相关以及在给定状态内的平均停留时间

	$\langle x \rangle$	$\langle y \rangle$	$\langle z \rangle$	$\langle xx \rangle$	$\langle xy \rangle$	$\langle xz \rangle$	$\langle yy \rangle$	$\langle yz \rangle$	$\langle zz \rangle$
系综	-0.0	-0.0	23.8	63.5	63.5	-0.1	63.5	-0.1	642.4
时间	-0.0	-0.0	23.5	62.8	62.8	-0.2	81.2	-0.2	628.9
	$\langle xxy \rangle$	$\langle xxz \rangle$	$\langle xyy \rangle$	$\langle xyz \rangle$	$\langle xzz \rangle$	$\langle yyy \rangle$	$\langle yyz \rangle$	$\langle yzz \rangle$	$\langle zzz \rangle$
系综	-0.3	1713.2	-0.3	1713.2	-3.4	-0.3	1713.2	-3.4	17346.1
时间	-0.4	1879.7	-0.4	1677.2	-4.1	-0.4	1997.2	-4.2	18446.3

表 1：洛伦兹吸引子的经验矩。系综平均与时间平均的比较。

分区。我们展示了六个可观测量的自相关，

$$g^1(\mathfrak{z}) = x,\ g^2(\mathfrak{z}) = y,\ g^3(\mathfrak{z}) = z,\ g^4(\mathfrak{z}) = \begin{cases} 1 & \text{if } \mathcal{E}(\mathfrak{z}) = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{4.26}$$

$$g^5(\mathfrak{z}) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases}, \text{ and } g^6(\mathfrak{z}) = \begin{cases} 1 & \text{if } \mathcal{E}(\mathfrak{z}) = 2 \\ 0 & \text{otherwise} \end{cases} \tag{4.27}$$

见图 4，这些结果是通过公式2.30 和 2.38计算得到的，并做了适当修改以考虑均值，并将高度归一化为 1。这里我们看到该方法在捕捉自相关方面既有成效，也有局限。一般而言，如果某个可观测量在给定分区内近似为常数，那么马尔可夫模型就能捕捉其去相关过程，例如可观测量 g^4 和 g^5。不过，有时即使某个量在一个区域内并非常数，也可能得到“不错”的结果，例如 g^1 或 g^2。顺便提一下，g^5 是由该分区诱导的 Koopman 模态数值近似（生成元的左特征向量 Q）。更多详情请参见附录 A.1 。

无法捕捉 g^6 的自相关，而其在 $\mathcal{M}_2$，内为常数，部分原因在于停留时间分布远非呈指数分布。为展示这种失效模式，我们绘制了图 5 中各分区的停留时间分布。我们展示了该分布的几种分箱策略，以说明指数分布捕捉经验停留时间分布的分位数的能力。

根据所关注的时间尺度，M_1 和 M_3 分区近似呈指数分布，尽管就停留时间分布而言，它们会呈现出类似分形的特征。相比之下，在细化分箱后，分区 M_2 的停留时间分布远非呈指数分布。此外，生成元的构造本身隐含了一个假设，即转移概率与在状态空间某一特定子集内停留的时间长短无关。更好的统计模型应纳入以在某个分区内停留的时间为条件的退出概率。

图 6 总结了由此得到的统计动力学，其中生成元和转移概率定义了一个图结构。该图结构包含状态空间不同区域之间的拓扑连通性信息，以及由转移概率所刻画的动力学在不同时间尺度下的连通性“强度”。生成元本身是离散系统的广义“图拉普拉斯算子”。

接下来 t 节中，我们转向一个偏微分方程示例，其中相同 e

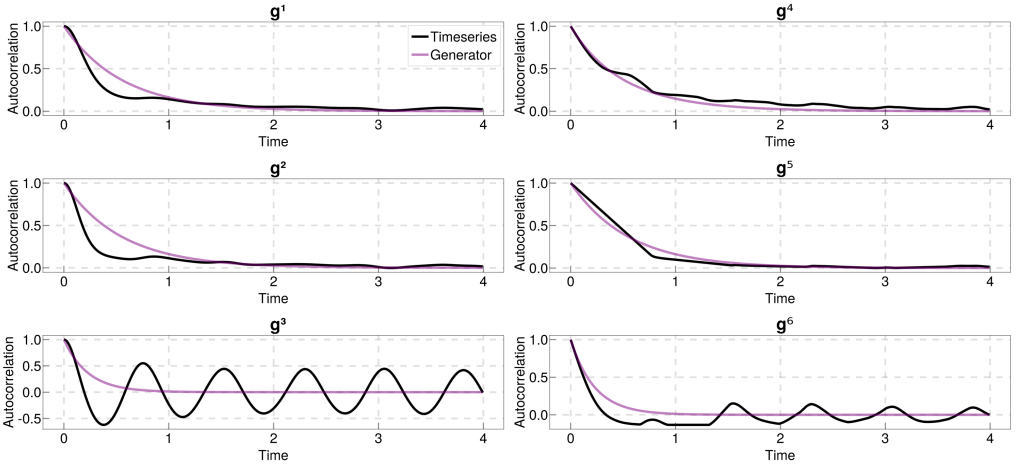

图 4：Lorenz 自相关：生成元与时间序列的比较。图中展示了六个可观测量的自相关。半透明的紫色线条根据生成元计算得出，黑色线条则根据时间序列计算得出。即使是粗略的分区，也能捕捉可观测量 g^4 和 g^5，但难以刻画振荡相关性。

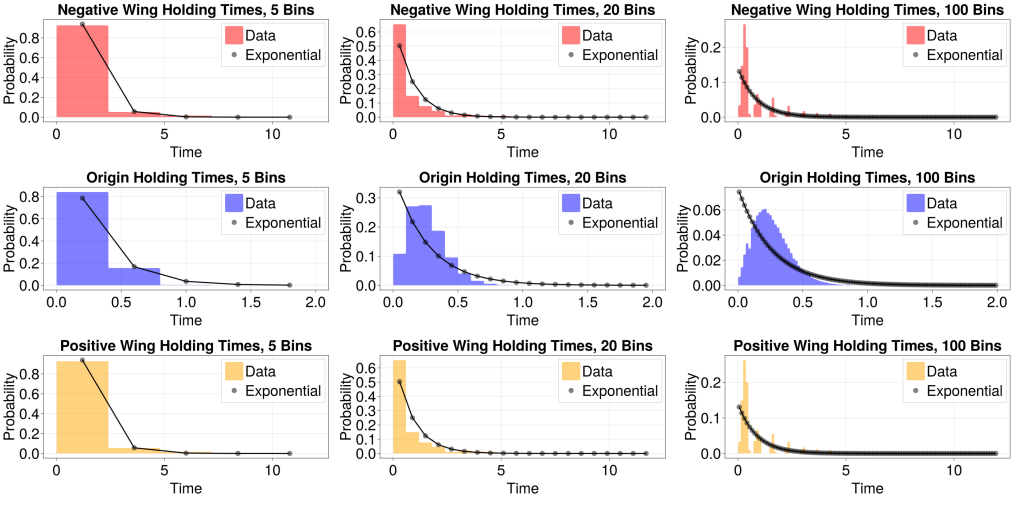

图 5：Lorenz 不动点划分的停留时间。对给定划分策略使用生成元时，其基本假设之一是状态中的停留时间呈指数分布。这里，我们考察了不同分箱数量所对应分区的停留时间分布分位数。黑点表示等效指数分布的分位数，生成元给出了其速率参数。

该方法适用，但就使用该方法能够提取出哪些信息而言，要微妙得多。

4.2. *Held-Suarez*：高维离散化的微妙之处

我们已经看到，该方法适用于 Lorenz 统计量的粗离散化。现在，我们将相同的方法应用于 Held-Suarez 大气测试案例。在图7中，我们展示了 Held-Saurez 模拟中预报变量的典型快照。

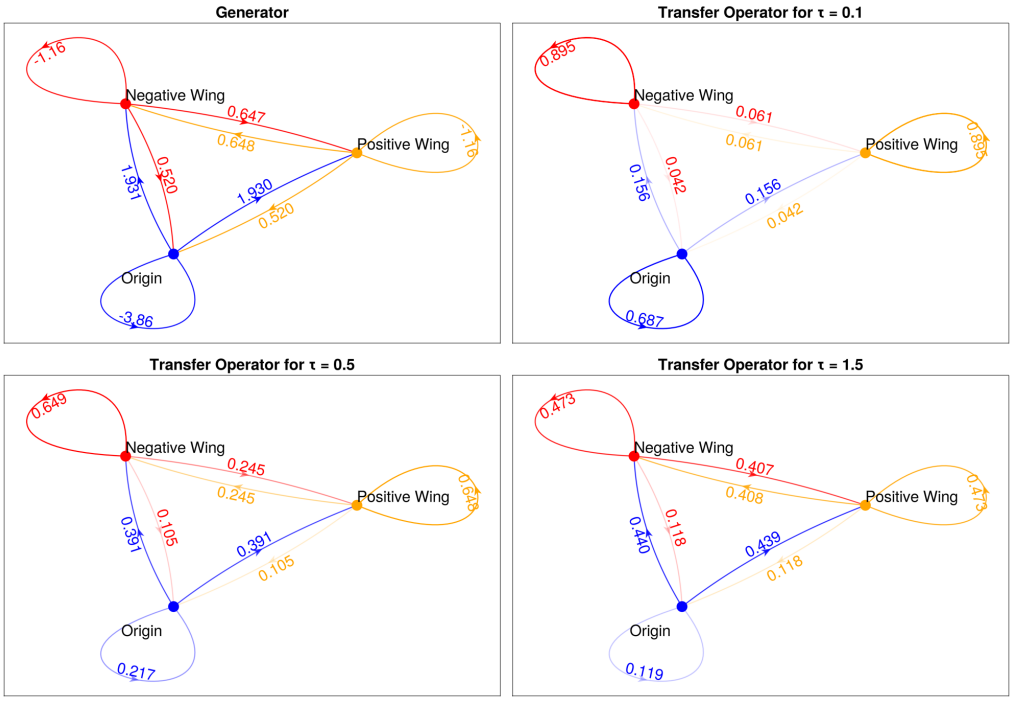

图 6：洛伦兹不动点图。生成元（左上）和多个时间尺度下的转移概率以图形形式呈现。转移概率会随时间尺度变化。

纬向速度是沿东西方向流动的风速，而经向速度则沿南北方向流动。

由于系统具有高维性，有一些细节需要考虑。例如，在上一节中，我们看到，如果某个可观测量与给定分区不相关，那么自相关就不太可能被生成元准确捕捉。此外，根据方程 2.33 和 2.38，我们预计只有在一个分区内大致为常数的可观测量，才能忠实地表示其系综平均统计量。

然而，这两种直觉都未必成立。在高维情形下，我们必须区分两类可观测量：一类与特定划分策略高度相关，另一类则不相关。对于与划分策略不相关的可观测量，我们依靠蒙特卡洛采样来计算系综统计量。实际上，相对于各个分区，可观测量是一个随机向量。

假设近似生成元针对稳态分布产生均匀分布。在这种情况下，预期只要各个马尔可夫状态彼此“独立”，表现至少不会逊于蒙特卡洛采样。如果某种划分策略与感兴趣的可观测量具有较强相关性，那么我们预计其表现会优于对马尔可夫状态进行“随机”采样。

对于自相关，也会出现类似现象。与各分区不相关的可观测量，相对于这些分区而言就是一个随机向量。只要许多可观测量具有相似的自相关，这种策略就能很好地捕捉这些可观测量。

当然，我们不希望依赖运气，也不希望任由算法摆布来针对感兴趣的目标可观测量。因此，在高维情形下，所选的划分策略，

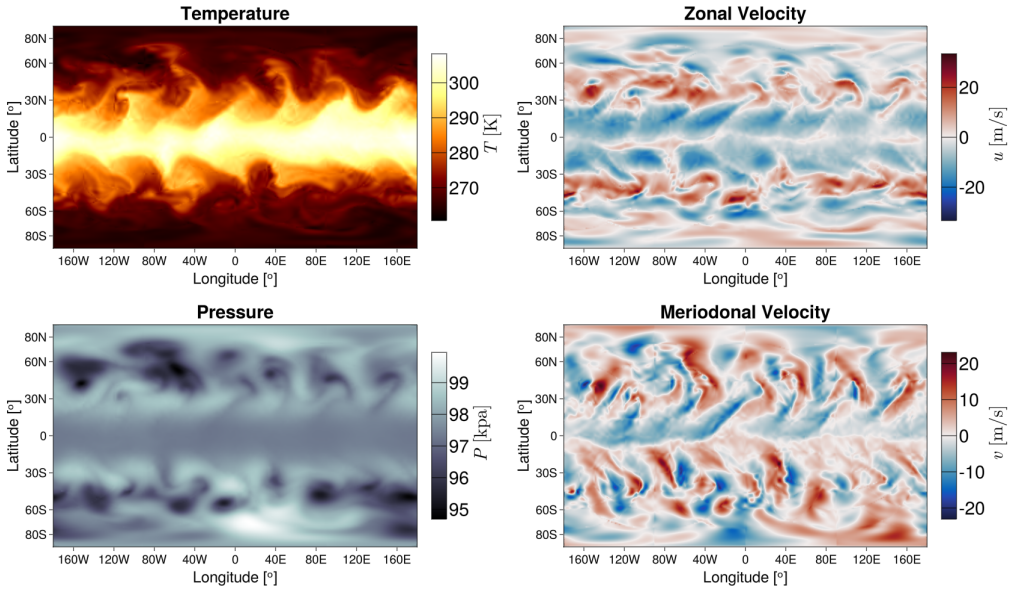

图 7：Held-Suarez 大气测试案例的表面场。图中展示了表面温度（左上）、纬向速度（右上）、经向速度（右下）和气压（左下）。

由 E 函数所封装的内容对于针对感兴趣的可观测量至关重要。我们将介绍两个选择分区的例子。第一个与上一节类似：基于马尔可夫状态定义适当的距离函数。第二个旨在针对感兴趣的可观测量，即球壳内半径上的温度“极值”（旨在表示固定位置处的热浪）。后者更类似于统计力学中的做法，在统计力学中，人们定义一个“宏观态”，但除此之外，我们还将挑选出与该宏观态对应的几个“微观态”。一般来说，可以通过按照一个（或多个）可观测量进行划分来创建整个状态空间的分区，这在执行降维时十分常见；然而，我们认为，为了计算与整个状态空间相关的系综平均统计量和相关性，我们始终需要一个与某个分区相关联的代表性状态。

第一种划分策略旨在揭示湍流吸引子的拓扑结构。第二种策略则表明，针对性划分策略能够为感兴趣的可观测量实现数据驱动统计建模。

4.3. *Held-Suarez*：距离划分

第一种划分策略首先进行一次初始模拟，以达到湍流状态，详情见附录 C。以一个“模拟日”作为时间单位，对应于根据行星角速度矢量完成一次旋转 $\boldsymbol{\Omega}$。据称，大气中天气的去相关时间约为两周。因此，每隔 15 个模拟日采集一次马尔可夫状态，直至累积 400 个状态。这一选择对应于湍流吸引子的随机样本。

嵌入函数 E 与之前一样，对应于“最近”马尔可夫

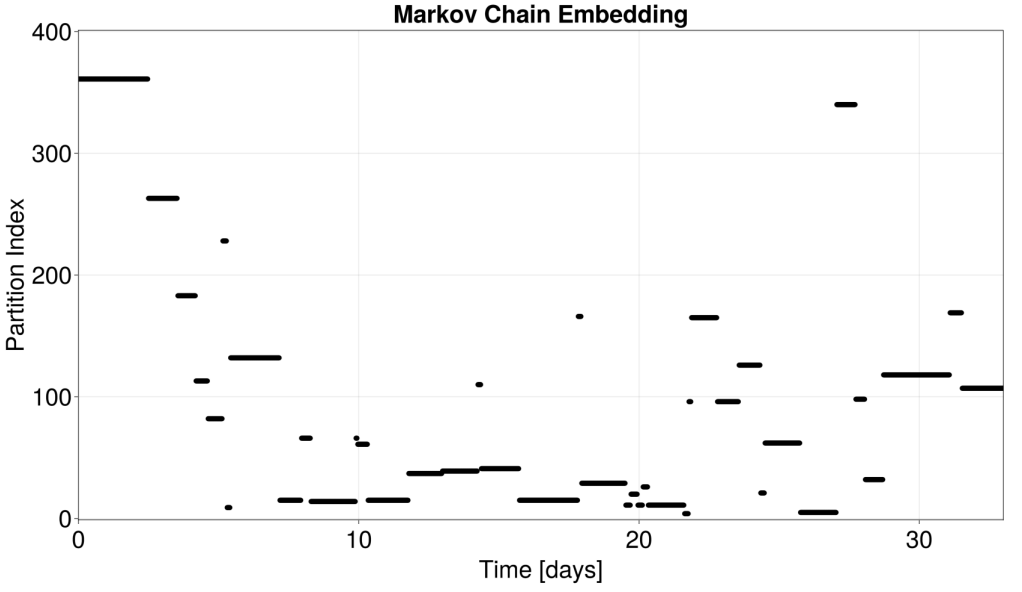

图8：Held Suarez马尔可夫链嵌入。动力学被简化为一个整数序列。我们按照处于某个分区的稳态概率对索引排序，使索引1对应概率最高的分区，索引400对应概率最低的分区。

状态。我们对“接近”的定义基于距离函数，

$$d(\boldsymbol{\imath}^1, \boldsymbol{\imath}^2) = \sqrt{\int_\Omega d\boldsymbol{x} \sum_i (\alpha_i)^{-2} \left(\boldsymbol{\imath}^1_{(\boldsymbol{x},i)} - \boldsymbol{\imath}^2_{(\boldsymbol{x},i)}\right)^2} \tag{4.28}$$

它是系统不同场之间的加权 L^2 范数（这样我们就能以无量纲的方式将各场相加）。 α_i 为

$$\alpha_1 = 1.3[kg/m^3], \alpha_2 = \alpha_3 = \alpha_4 = 60[m/s], \alpha_5 = 2.3 \times 10^6 [kg/(ms^2)]. \tag{4.29}$$

嵌入函数总体上为

$$\mathcal{E}(\boldsymbol{\imath}) = n \text{ if } d(\boldsymbol{\imath}, \boldsymbol{\sigma}^n) < d(\boldsymbol{\imath}, \boldsymbol{\sigma}^m) \text{for each } m \neq n \tag{4.30}$$

我们继续演化系统额外 的 $\Delta t = 0$。03 模拟年，并将嵌入函数应用于瞬时状态。这总计为2,000，000+个时间快照，但没有保存任何快照，因为这将占用超过20 TB的数据。嵌入函数是“即时”应用的，只记录整数序列。此过程最初30个模拟天的情况如图8所示。我们事后对索引进行了排序，使概率最高的分区对应索引1，概率最低的分区对应索引400。

对于生成元的先验分布，我们取 $\alpha = 1$，并取 $\beta = \Delta t$，其中 Δt 是时间序列的采样时间间隔。对于矩阵的每一列，我们使用 $\vec{\alpha} = 10^{-4}\mathbf{1}$，其中111是全1向量。该先验分布的解释如下：如果某个分区未被观测到，则假定其停留时间低于 Δt 天这一采样阈值。我们采取这一预防措施，是因为事先无法确定在有限采样周期内是否会重新访问每个分区。尽管如此，200个模拟年足以

重新访问每个分区。我们假定未观测状态的连接较为稀疏，这体现在选择 $\vec{\alpha} = 10^{-4}\mathbf{1}$ 上。

显示一个 400×400 随机矩阵的均值和方差并没有特别的启发意义。因此，我们在图9中总结了均值生成元的四个性质：逆特征值的实部、与某个分区相关联的稳态概率值、给定分区与其他每个分区之间的连通性，以及一个分区的平均停留时间。逆特征值的实部与系统中衰减最慢的自相关相关联，这些自相关由划分策略捕捉。我们看到，特征值聚集在一个模拟日附近。此外，我们看到前几个特征值与主体之间存在明显的谱隙†。稳态概率向量并不均匀（左上），然而在每个状态上花费的时间（右下）对每个状态来说大致相同。通过考察给定分区的连通性，可以解释概率不均匀的原因（左下）。连通性定义为从给定分区离开或进入给定分区的经验数量。我们看到，与其他分区相比，概率更高的分区与其余状态空间的连通性更强。一个分区的连通性可以被视为与该分区相关联的有效动力学可预测性。例如，周期解经过充分采样的分区只与另一个分区相连，因为未来可以从过去精确预测。

此外，最大连通性约为 200，这一数字小于使用基于距离、位于 1,000，000+ 维空间中的 400个分区所预期的数值。正是这种降低的连通性，使当前的划分策略摆脱了 Ulam方法中存在的“维数灾难”，参见Ulam（1964）。控制“体积”适应湍流吸引子的形状，而动力学引导概率流出，仅流向与盒状控制体积相关的无数个面中的一个子集。Ulam 方法是第 3 节中方法的一个特例，因为可以将 Ulam 方法中的盒子中心作为马尔可夫状态，并使用 L^∞ 范数来计算状态之间的距离。因此，使用 Ulam 方法中的相同盒子，将得到相同的生成元近似。

事先，在马尔可夫状态按时间均匀采样的情况下，没有理由认为某个分区会不同于另一个分区；然而，图 9 却表明情况并非如此。状态空间中概率最高的区域充当中心枢纽，将状态空间的各个区域连接在一起。这些区域或许与相干结构有关，例如不动点或不稳定方向很少的周期轨道。

我们已经讨论了生成元的拓扑特征。关于有限抽样效应和停留时间分布的更多细节，请参见附录C.3。总之，最可能状态的停留时间近似呈指数分布；生成元最大去相关时间尺度存在显著的不确定性，而其他特征值的不确定性则较小。现在我们转而计算统计量。

如前所述，我们区分与分区相关的可观测量和不相关的可观测量。当前的划分策略并不针对某个可观测量，但确实会产生与慢、快去相关时间尺度相关联的涌现可观测量，即全局 Koopman 模态。尽管这不是本文的重点，我们仍在附录A中对其进行了讨论。本文此处的重点是计算与分区**不**相关的可观测量的统计量。随后，我们将在第4.4节中，为一个目标可观测量构造分区。

† 目前尚不清楚，在不断细化粗粒化状态空间时，是否存在唯一的极限。这可能意味着，在分区不断细化的极限下，连续谱和离散谱都存在。

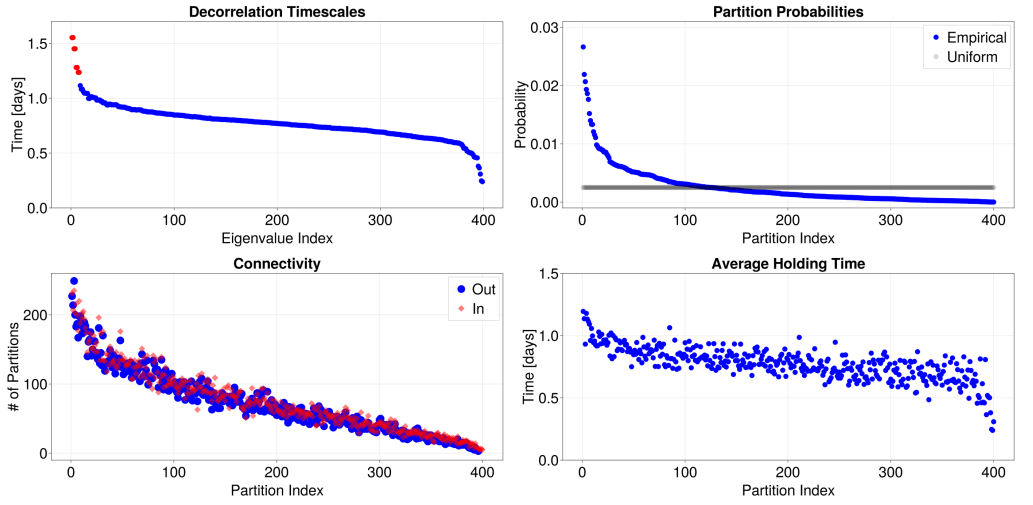

图 9：生成器性质。左上图展示了生成元特征值实部倒数，对应于与划分相关的去相关时间尺度。右上图为某个分区对应的稳态概率。右下图为给定分区的停留时间。左下图根据经验观测到的转变，总结了一个分区与其他分区之间的连通性。

我们考察该可观测量的直方图

$$g(\boldsymbol{\imath}) = \boldsymbol{u}_{\boldsymbol{x}} \cdot \hat{\varphi} \tag{4.31}$$

其中 $\hat{\varphi}$ 是沿经向的单位向量，$\boldsymbol{x}$ 是内壳层（表面）上纬度为 $\theta = 35S$、经度为 $\varphi = 135E$ 的一个点，见图 10。我们展示了两个相互重叠的直方图。一个直方图根据生成元计算得出，另一个则通过汇集时间序列计算得出。可观测量的时间序列是在一个为期 30 年的时段内累积得到的，该时段与用于构建生成元的数据时段不重叠。紫色区域表示两个直方图重叠的部分，红色区域表示马尔可夫模型高估概率的部分，蓝色区域表示马尔可夫模型低估概率的部分。与前文一样，我们展示了多个区间，以体现“分位数收敛”的概念。当区间数与马尔可夫状态数相同（本例中为 400 个区间）时，狄拉克δ函数近似开始显现。δ函数的高度与生成元的稳态概率分布相关。该分布完全是由于马尔可夫状态的选取而得以捕捉，这些状态是湍流吸引子的随机样本。假设马尔可夫状态的概率服从均匀分布，也能达到同等保真度；无需构建生成元即可得到这一结果。在当前情境下，这是一大优势，因为这些分布是“免费”获得的。

我们通过计算连续统可观测量的均值进一步强调这一点。我们使用每个纬度和每个高度处纵向速度场的纵向平均，

$$g^{(\theta,r)}(\boldsymbol{\imath}) = \frac{1}{2\pi}\int_0^{2\pi} (\boldsymbol{u}_{(\theta,\varphi,r)} \cdot \hat{\varphi}) d\varphi. \tag{4.32}$$

固定的纬度 θ 和高度 r 构成一个可观测量，我们将位置$\boldsymbol{x} = \theta\hat{\theta} + \varphi\hat{\varphi} + r\hat{r}$用球基中的分量展开。我们计算图 11中每个可观测量的系综均值和时间平均。时间平均是对三个

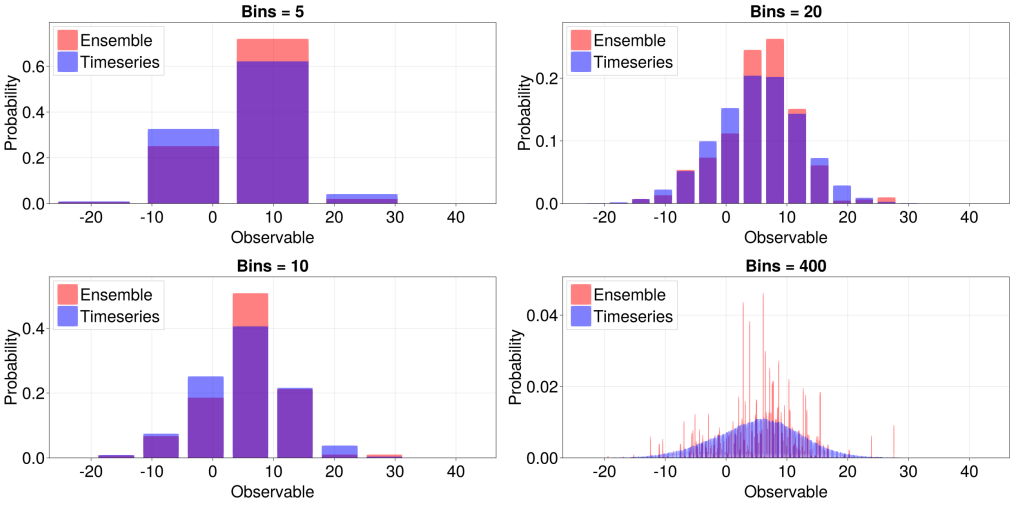

图10：可观测量的稳态分布。此处，我们使用狄拉克δ函数近似来处理分区内的概率密度，并查看基于分布的不同粗粒化方式得到的推断分布。重叠区域以紫色显示；红色条形表示概率“高估”，蓝色条形表示概率“低估”。时间平均值和系综平均值分别为 5.3 和 5.7。时间标准差和系综标准差分别为.1 和 6.5。

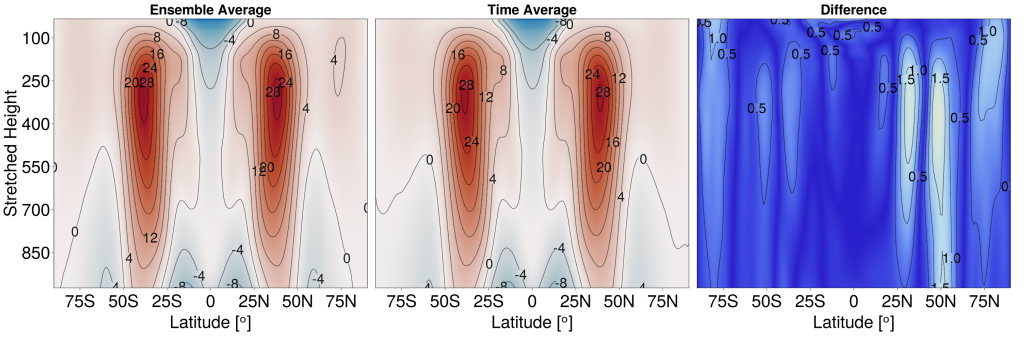

图11：连续统可观测量的平均值。系综平均（左）和时间平均（中）纵向均值风展示了连续统可观测量的均值。最右侧的图显示了两个均值之间的逐点绝对差。

模拟年得出的。为了与该场通常的可视化方式相联系，我们根据赤道处气压的纵向平均重新缩放高度轴。这种缩放模拟了大气科学文献中使用“气压坐标”的效果。我们看到，纵向风“蝴蝶”翅膀右半部分的系综均值与时间平均相差不到每秒两米，除此之外，两者吻合良好。

四个可观测量的自相关，

$$g^1(\boldsymbol{\imath}) = \varrho_{\boldsymbol{x}}\,,\; g^2(\boldsymbol{\imath}) = \boldsymbol{u}_{\boldsymbol{x}} \cdot \hat{\varphi}\,,\; g^3(\boldsymbol{\imath}) = \boldsymbol{u}_{\boldsymbol{x}} \cdot \hat{\theta}\,,\; \text{and } g^4(\boldsymbol{\imath}) = \mathcal{T}_{\boldsymbol{x}}\,, \tag{4.33}$$

如图12 所示，位置与 $\boldsymbol{x}$ 之前相同。变量 $\mathcal{T}_{\boldsymbol{x}}$ 是温度

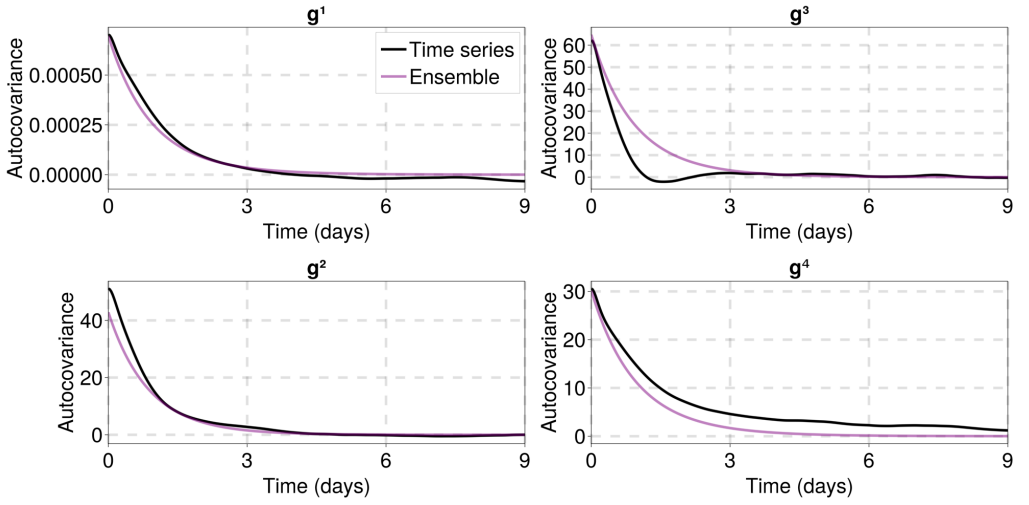

图12：Held-Suarez 大气测试中多个可观测量的自协方差。图中展示了基于时间序列（黑色）和生成元（紫色）的多个可观测量的自协方差。这些可观测量与分区不相关，因此，马尔可夫模型预测的自相关在四种情况下都相似。

在同一点处，并通过以下关系定义，

$$\mathcal{T}_{\boldsymbol{x}} \equiv \frac{\gamma - 1}{R_d \varrho_{\boldsymbol{x}}} (\varrho e_{\boldsymbol{x}} - \frac{1}{2} \varrho_{\boldsymbol{x}} \|\boldsymbol{u}_{\boldsymbol{x}}\|^2 - \varrho_{\boldsymbol{x}} \Phi_{\boldsymbol{x}}) \tag{4.34}$$

其中， $R_d = 287$ 为理想气体常数， $\gamma = 1.4$ 为空气的比热比， Φ 为位势。我们以蓝色展示从时间序列中经验获得的自相关，以紫色展示生成元。由于生成元的大多数特征值聚集在一天左右，随机向量很可能产生一天的去相关时间。对于可观测量 g^3，这种分区的近似效果较差，但对其他变量则不然。原则上，有无穷多个可观测量的去相关时间可以由随机向量很好地近似，也有无穷多个可观测量的近似效果较差。

因此，下一节将展示针对某个特定可观测量时会发生什么。

4.4. *Held-Suarez*：温度极值统计量划分

现在，我们以不同方式对湍流吸引子进行分区，以针对某个特定可观测量的统计量：区域中特定位置的温度极值。具体而言，

$$g(\boldsymbol{\imath}) = \begin{cases} 1 & \text{if } \mathcal{T}_{\boldsymbol{x}} > 290[K] \\ 0 & \text{otherwise} \end{cases} \tag{4.35}$$

这里， $\boldsymbol{x}$ 是球面内壳层上纬度-经度坐标为（$\theta = 35$S， $\varphi = 135$E）的一个点。我们使用脚本 $\mathcal{T}$ 来表示温度，与前文的记法类似。选择 290[K] 是因为在一次较短的模拟运行中，该点温度的 95% 分位数为此值。

我们首先将任意状态划分为两类，以收集马尔可夫状态：$g(\boldsymbol{\imath}) = 1$ 和 $g(\boldsymbol{\imath}) = 0$。前者代表“极端状态”，后者代表“良性状态”。随后，我们收集十个有代表性的“极端”状态和九十个有代表性的良性状态。具体而言，我们运行一次模拟，并每两周检查一次状态。我们使用可观测量（即分类器） g 来判断该状态是否为极端状态。这个过程会持续进行，直到收集到至少十个极端状态和九十个良性状态。

我们总共只保留一百个状态。因此，任何多余的状态都会被舍弃。极端状态的索引设为 1–10，良性状态的索引设为 11–100。

有了这些马尔可夫状态后，嵌入函数定义如下

$$\mathrm{E}(\boldsymbol{s}) = \begin{cases} n & \text{若 } g(\boldsymbol{s}) = 1 \text{ 且 } d(\boldsymbol{s},\boldsymbol{\sigma}^n) < d(\boldsymbol{s},\boldsymbol{\sigma}^m) \text{对每个 } m \neq n, \text{ 其中 } m,n \in \{1,\ldots,10\} \text{ 若 } g(\boldsymbol{s}) \\ n & = 0 \text{ 且 } d(\boldsymbol{s},\boldsymbol{\sigma}^n) < d(\boldsymbol{s},\boldsymbol{\sigma}^m) \text{ 对每个 } m \neq n, \text{ 其中 } m,n \in \{11,\ldots,100\} \end{cases} \tag{4.36}$$

也就是说，我们首先根据可观测量 g 对状态进行分类，然后在相应类别中计算距离最近的马尔可夫状态。最后，我们运行模型 100 个模拟年，并构建马尔可夫链嵌入。

我们首先关注处于与极端事件相关的分区中的停留时间分布。该分布被用作热浪持续时间的代理。鉴于我们有十个对应于极端事件的可能状态，我们还考虑极端事件持续时间内状态之间的转变。启发式地说，极端事件期间不同全局状态之间的转变很少见，因为持续时间相较于处于某一状态的停留时间而言很短。

然而，它们确实会在此次模拟中发生。图 13 总结了与极端状态相关的分区之间的转移路径。在此图中，状态11-100 被归并为单个状态，并展示了转移路径的图结构。红线的透明度对应于不同极端状态分区之间转移的概率，蓝线对应于离开某一极端状态的转移概率。图结构表明，当地球上某个特定位置发生极端事件时，全局状态会发生变化。

此外，这表明一个极端状态对应许多“微观状态”（由 $g(\boldsymbol{s}) = 1$ 所诱导的分区定义的宏观状态），并且在给定的宏观状态配置期间，这些“微观状态”之间存在非零转变。这一细微之处部分解释了极端事件预测的复杂性。

处于极端状态的停留时间分布（由马尔可夫嵌入计算）涵盖了不同状态之间的转变。此外，我们从时间上互不重叠的集合中收集温度可观测量的统计量，并在图中展示其停留时间分布 14。马尔可夫状态表示表明，该停留时间分布得到了很好的刻画。

极端状态的存在可以从随机过程的角度视为一个退出时间问题。极端事件对应于状态空间中的一个特定子集，此处由十个分区表征。处于极端状态的平均持续时间必须将极端事件持续期间的状态转变纳入考量。

我们还比较了 $\boldsymbol{x}$ 处由 400 状态系统、100 状态系统和时间序列计算得到的温度分布，见图 15。特别是，我们看到 100 状态系统能更好地捕捉 95% 尾部分布。因此，选择一种针对感兴趣的可观测量的划分策略是可行的，与朴素的通用划分相比，其保真度更高。此过程等同于数值方法中的局部网格细化。

作为本节的最后一点说明，分区的选择不必是二元的。在运行模拟时，可以使用相同的马尔可夫状态，同时计算多种不同的划分策略。如果划分策略利用独立可观测量，那么构造张量积生成元便是顺理成章的做法。本文未进一步探讨这些策略。

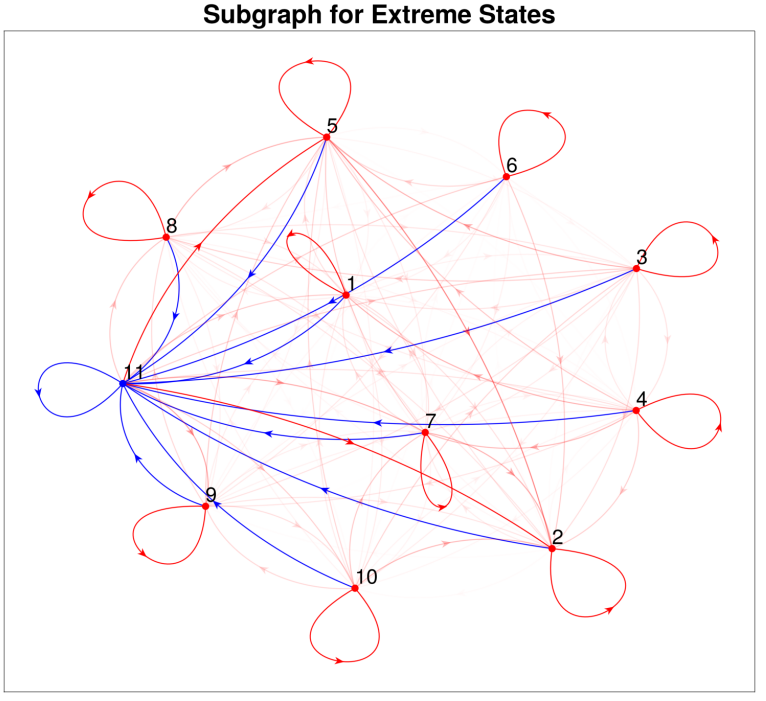

图13：极端转变的网络结构。图中显示十一个状态，其中1–10对应极端状态，第11个对应其余90个状态；为便于可视化，将这90个状态合并为一个节点。蓝线表示转变到良性状态，红线表示转变到极端状态。线条的不透明度与状态之间转变的概率成正比。极端状态之间也存在转变。

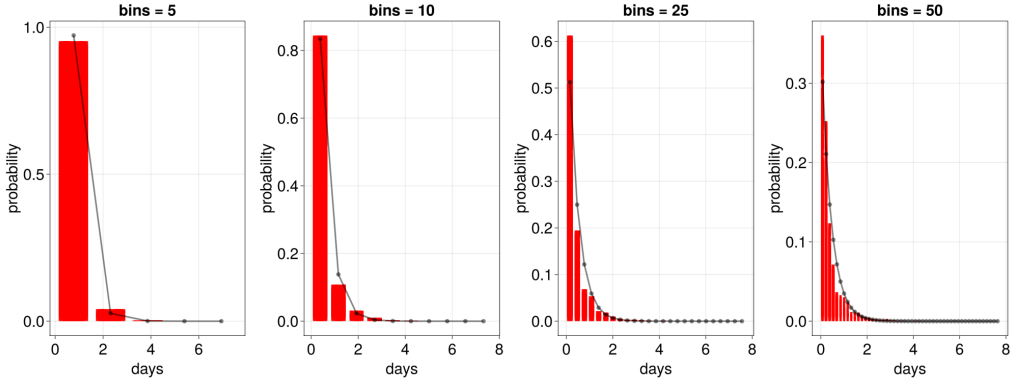

图14：Held Suarez 极端状态停留时间。图中以红色显示了根据时间序列计算得出的极端状态持续时间的若干分位数。为简化起见，我们用黑点表示停留时间的指数分布；通过考察单个极端状态分区的停留时间，计算得出去相关时间约为半天。

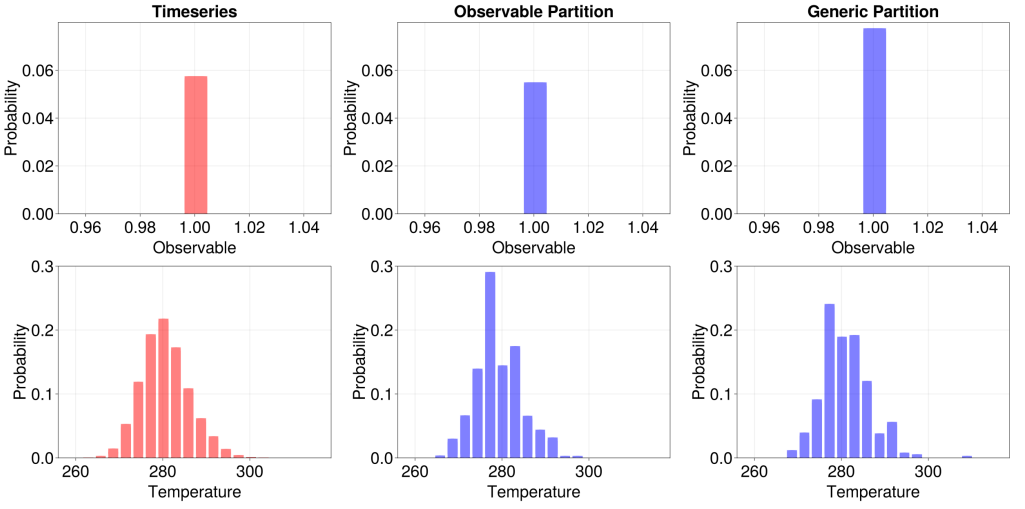

图 15：Held Suarez 可观测量比较。我们展示了网格细化对一个感兴趣的可观测量的影响。与通用的 400 状态划分相比，该细化策略更好地捕捉了尾部概率分位数（上方三个面板）。下方三个面板展示了温度分布，供比较参考。

5. 结论

总而言之，我们完成了三项工作

(i) 第 2 节：回顾并扩展了将动力系统转化为具有有限状态空间的连续时间马尔可夫过程的理论框架。

(ii) 第 3 节：开发了一种贝叶斯流式数据驱动算法，用于构建具有有限状态空间的连续时间马尔可夫过程的生成元。

(iii) 第 4 节：将该方法应用于两个系统：洛伦兹系统的连续性方程和可压缩欧拉方程。

我们已经看到，即使采用粗离散化，也能捕捉到许多统计量。在 Lorenz 系统中，我们将动力系统的不动点既用作马尔可夫状态，也用来确定划分策略的锚点。这三个状态足以捕捉均值和二阶矩。此外，即使采用粗离散化，某些自相关和驻留时间也能得到较好的捕捉，具体取决于所关注的时间尺度。

在高维情形下，计算颇为微妙。与划分策略不相关的可观测量，仍能通过蒙特卡洛抽样得出合理的稳态统计量。此外，对于不相关的随机变量，也能够捕捉其自协方差，但并非所有可能的可观测量都可以。当划分策略选为与感兴趣的可观测量相一致时，驻留时间和统计量都得到了较好的捕捉。

总的来说，我们看到，系统统计表示中最关键的组成部分是划分策略的选择。未来的方向需要开发新颖的划分策略。例如，根据动态模态分解给出的模态振幅进行划分，或使用机器学习方法（如自编码器）来降低状态空间的维度。Koopman 算子与 Perron-Frobenius算子之间的一致性很可能会带来最大的改进。在马尔可夫状态划分中纳入部分时间相干性似乎也很有前景†。

在实际应用中，划分策略应当在计算上可行。使用

† 例如，在 Held-Suarez 情况下，可以连续一个月每隔一天选取马尔可夫状态，然后间隔几个月，再反复收集马尔可夫状态。

采用树结构对状态空间进行分区，可以大幅减轻计算负担，但需要预先占用更多内存。例如，使用树结构和二元分类将生成 2^N 个分区，需要存储 $2^{N+1} - 2$ 个场，并且需要应用 $2N$ 次分类函数求值。此时，对马尔可夫状态进行压缩也变得很重要，因此应利用对称性并采用较低精度的算术。

既然该方法已被表述为一种数值离散化，就可以考虑一些直接的推广。例如，除了使用有限体积法对空间进行离散化之外，还可以使用间断伽辽金法对时间进行离散化。这样，时间轨迹便可表示为分段多项式，而非分段常数。此外，流向状态空间不同区域的概率通量（在离散时间下）将取决于历史。

与本文提出的方法相比，一种更为根本的变革是采用生成模型（类似于 Ho 等人（2020）），来表示分区内部的分布。状态空间的分区可能比整个流形更易于表示。此外，使用非线性模型作为生成元，以纳入所有需要赋予“注意力”的不同物理特征，或许能够带来更好的整体表示，例如 Vaswani等人（2017）。

撰写本文并采用这一视角，主要是为了借此初步理解如何运用算子理论方法研究气候变化，这与 Froyland 等人（2021）的做法类似。气候变化通常被描述为“统计量随时间变化”，因此需要给出精确定义。我们着重研究了一个相对于时间保持不变的高维测度。这一特性并不适用于气候系统，因为其统计量是非平稳的。“稳定”气候的主要信号并非稳定不变，而是受昼夜周期和季节周期影响，呈现时间周期性。因此，首先要做的简化是考虑一个生成元，其各项为时间的周期函数，且其马尔可夫状态也随时间呈周期性变化；关于分子动力学中的类似考量，参见 Wang 和 Schütte（2015）。气候变化由此可表述为偏离这一时间周期性（高维）流动的现象。

补充数据。补充材料可在以下地址获取https://github.com/sandreza/ MarkovChainHammer.jl。

致谢。 Tobias Bischoff、Keaton Burns、Glenn Flierl、Raf Ferrari、Ludovico、Peter、MG、Fabri、Matthew、Simone、Pedram、Predrag。作者谨向 2022 年地球物理流体动力学项目致谢，本研究的大部分工作是在该项目期间完成的。

资助。本研究得到了 Eric 和 Wendy Schmidt 的慷慨支持，该支持由 Schmidt Futures 项目推荐。地球物理流体动力学项目由美国国家科学基金会（美国）和美国海军研究办公室（美国）提供支持。

利益声明。作者声明不存在利益冲突。

附录 A. 全局 Koopman 模态

我们不问“我们能否预测感兴趣的可观测量？”，而是问“我们能预测什么？”。后一个问题是系统的一种涌现属性，由底层系统的 Koopman 模态所刻画。那些去相关时间尺度较长的 Koopman 模态，构成了系统在较长时间尺度上最可预测的特征。

Koopman 模态既是可观测量，也是转移概率算子 T^τ 的左特征向量。例如，如果 g_λ 是 T^τ 的左特征向量，且其特征值为 $e^{\lambda\tau}$，那么我们有

如下

$$R_E(g_\lambda,\tau) = \int_{\mathcal{M}} d\boldsymbol{\mathfrak{s}}' g_\lambda(\boldsymbol{\mathfrak{s}}')\mathscr{P}(\boldsymbol{\mathfrak{s}}') \left[\int_{\mathcal{M}} d\boldsymbol{\mathfrak{s}} g_\lambda(\boldsymbol{\mathfrak{s}})\mathscr{T}^\tau \delta(\boldsymbol{\mathfrak{s}}-\boldsymbol{\mathfrak{s}}')\right] \tag{A 1}$$

$$= \int_{\mathcal{M}} d\boldsymbol{\mathfrak{s}}' g_\lambda(\boldsymbol{\mathfrak{s}}')\mathscr{P}(\boldsymbol{\mathfrak{s}}') \left[\int_{\mathcal{M}} d\boldsymbol{\mathfrak{s}} g_\lambda(\boldsymbol{\mathfrak{s}}) e^{\lambda\tau} \delta(\boldsymbol{\mathfrak{s}}-\boldsymbol{\mathfrak{s}}')\right] \tag{A 2}$$

$$= e^{\lambda\tau} \int_{\mathcal{M}} d\boldsymbol{\mathfrak{s}}' g_\lambda(\boldsymbol{\mathfrak{s}}')^2 \mathscr{P}(\boldsymbol{\mathfrak{s}}') \tag{A 3}$$

$$= e^{\lambda\tau} \langle g_\lambda^2 \rangle_E \tag{A 4}$$

因此，从可预测性的角度来看，最有用的 Koopman 模态是那些随时间缓慢去相关的模态，即 $\text{real}(\lambda) \approx 0$，但还具有振荡分量，使得 $\text{real}(\lambda)/\text{imaginary}(\lambda) \approx 0$ 成立。

如果 $\text{real}(\lambda) = 0$ 位于混沌吸引子上，那么我们预期它是“平凡的”可观测量†$g_\lambda(\boldsymbol{\mathfrak{s}}) = c$，对应于常数 c。否则，我们预期所有对应于传递算子的特征值都满足 $\text{real}(\lambda) < 0$，也就是说，我们预期所有非平凡可观测量最终都会去相关。这意味着 $\langle g_\lambda \rangle_E = 0$ 因为

$$\langle g_\lambda \rangle_E = \int_{\mathcal{M}} d\boldsymbol{\mathfrak{s}} g_\lambda(\boldsymbol{\mathfrak{s}})\mathscr{P}(\boldsymbol{\mathfrak{s}}) = \lim_{\tau\to\infty} \int_{\mathcal{M}} d\boldsymbol{\mathfrak{s}} g_\lambda(\boldsymbol{\mathfrak{s}})\mathscr{T}^\tau \delta(\boldsymbol{\mathfrak{s}}-\boldsymbol{\mathfrak{s}}') = \lim_{\tau\to\infty} e^{\lambda\tau} g_\lambda(\boldsymbol{\mathfrak{s}}') = 0 \tag{A 5}$$

其中 $\boldsymbol{\mathfrak{s}}'$ 是吸引子 M 上的任意状态。

The 关于一个 Koopman 模态的以下四个陈述不能同时成立

- Koopman 模态满足关系 $g_\lambda(\boldsymbol{s}(t+\tau)) = e^{\lambda\tau} g_\lambda(\boldsymbol{s}(t))$
- Koopman 模态 g_λ 是状态空间上的连续函数
- 动力学轨迹上存在任意多个近复现
- 与 Koopman 模态相关联的特征值满足 $\text{real}(\lambda) < 0$。

证明如下。假设这四个条件均成立。令 $\boldsymbol{\mathfrak{s}}'$ 为未来 $\boldsymbol{\mathfrak{s}}$ 某个时刻 τ 的一个近复现，使得 $\|\boldsymbol{\mathfrak{s}}-\boldsymbol{\mathfrak{s}}'\| < \epsilon$，其中范数为某个范数。根据范数的连续性，$g_\lambda$ 可得

$$|g_\lambda(\boldsymbol{\mathfrak{s}}) - g_\lambda(\boldsymbol{\mathfrak{s}}')| < \delta \tag{A 6}$$

但是根据假设，$g_\lambda(\boldsymbol{\mathfrak{s}}') = e^{\lambda\tau} g_\lambda(\boldsymbol{\mathfrak{s}})$，因此

$$|g_\lambda(\boldsymbol{\mathfrak{s}})||1 - e^{\lambda\tau}| < \delta \tag{A 7}$$

这是一个矛盾，因为 τ 可以任意大，而 δ 可以任意小。数值证据证实，满足 $g_\lambda(\boldsymbol{s}(t+\tau)) = e^{\lambda\tau} g_\lambda(\boldsymbol{s}(t))$ 的 Koopman 模态并非在整个状态空间上都存在 Parker & Page (2020)。只要湍流吸引子具有混合性质，我们就不应期待 Koopman 模态存在有限维线性子空间（常数可观测量除外）。关于 Lorenz 吸引子的类似论述，参见 Arbabi & Mezić (2017)。我们认为，上述证明为在 Koopman 模态的表示中使用分段不连续基提供了一个合理的论据。

对于随机动力系统，人们预期 Koopman 模态（在此情境下，Koopman 算子定义为 Fokker–Planck 算子的伴随算子）是状态的连续泛函，但不再满足关系 $g_\lambda(\boldsymbol{s}(t+\tau)) = e^{\lambda\tau} g_\lambda(\boldsymbol{s}(t))$。从启发式角度看，这是因为动力学中的噪声在概率空间中起到扩散作用，从而平滑非光滑场。Maćešić & Črnjarić-Žic 的方程 6.55

† 纯虚特征值的存在意味着，在混沌吸引子上存在能够预测未来任意时刻的可观测量，这与 SRB 测度的定义不相容。

(2020）颇具启发性。由此可知，对于随机微分方程

$$d\boldsymbol{s} = \boldsymbol{U}(\boldsymbol{s})dt + \epsilon d\boldsymbol{W} \tag{A 8}$$

其中， ϵ 是噪声方差， $d\boldsymbol{W}$ 是一个 d–维维纳过程，Koopman 模态按照以下方式演化。

$$dg_\lambda(\boldsymbol{s}(t)) = \lambda g_\lambda(\boldsymbol{s}(t))dt + \epsilon \nabla g_\lambda(\boldsymbol{s}(t))d\boldsymbol{W}, \tag{A 9}$$

与状态变量的组合是必要的，因为 g_λ： $\mathbb{R}^d \to \mathbb{R}$，但需要特别指出的是，不能脱离状态空间中的评估位置来单独考虑 g_λ 的演化。当噪声趋近于零时， $\epsilon \to 0$，梯度项 ∇g_λ 可能在状态空间中的某些特定点趋于无穷大，这正如在双阱随机势中所预期的那样。因此，在线性系统中不太可能出现病态现象。

另外需要指出的是，尽管在偏微分方程（PDE）的语境中它们常被称为 Koopman 模态，但不应将它们与空间模态混为一谈。Koopman 模态是特征算子（类似于低维情境中的特征向量和特征函数），也就是说，是作用于状态的泛函。另一方面，传递算子的右特征向量确实起到将 Koopman 模态投影出来的算子作用。下一节将通过一个具体示例进行说明，但此处我们暂且保持抽象。

由 $\boldsymbol{x}$ 索引的连续统可观测量定义统计模态为

$$G_\lambda(\boldsymbol{x}) \equiv \int_{\mathcal{M}} d\boldsymbol{\jmath} g^{\boldsymbol{x}}(\boldsymbol{\jmath})\mathcal{P}_\lambda(\boldsymbol{\jmath}) \tag{A 10}$$

其中，P_λ 是传递算子的右特征向量，即

$$\mathcal{T}^\tau \mathcal{P}_\lambda = e^{\lambda\tau}\mathcal{P}_\lambda. \tag{A 11}$$

方程 A10 将可观测量 $g^{\boldsymbol{x}}$ 中相应的部分投影到对应的 Koopman 模态上†。我们将 $G_\lambda(\boldsymbol{x})$ 视为一种模态，不过称其为相应 Koopman 模态的模态振幅或许更为恰当。Koopman 模态是否构成一个完备基，使得 $g^{\boldsymbol{x}} = \sum_\lambda G_\lambda(\boldsymbol{x})g_\lambda$，目前尚不明确。这意味着，如果任意可观测量 $g^{\boldsymbol{x}}$ 无法表示为 Koopman 模态之和，那么它从根本上来说可能是不可预测的。

I在下一节中，我们将讨论 Koopman 模态的数值近似。

A.1. 数值近似

数值 Koopman 模态是矩阵 Q的左特征向量，记作 $\boldsymbol{g}_\lambda$；而将其视为作用于状态的泛函时，其近似形式为

$$g_\lambda(\boldsymbol{\jmath}) \approx [\boldsymbol{g}_\lambda]_{\mathcal{E}(\boldsymbol{\jmath})}, \tag{A 12}$$

其中， $[\boldsymbol{g}_\lambda]_n$ 是特征向量 $\boldsymbol{g}_\lambda$的第 $n'th$ 个分量。因此，我们先将嵌入函数应用于状态，然后使用整数标签选取特征向量$\boldsymbol{g}$的相应分量。

在每个时刻，我们绘制近似的 Koopman 模态

$$g_\lambda(\boldsymbol{s}(t)) \approx [\boldsymbol{g}_\lambda]_{\mathcal{E}(\boldsymbol{s}(t))}, \tag{A 13}$$

其中，向量 $\boldsymbol{g}$ 的分量由 $\mathcal{E}(\boldsymbol{s}(t))$给出。图16展示了 Held-Suarez 设置最初 30 个模拟天的这些动力学。此外，我们通过两种方式计算自相关，以检验数值 Koopman 模态的保真度。该计算见图 16。从右上图可以看出，这两种方法

† Due to bi-orthogonality of left and right eigenvectors.

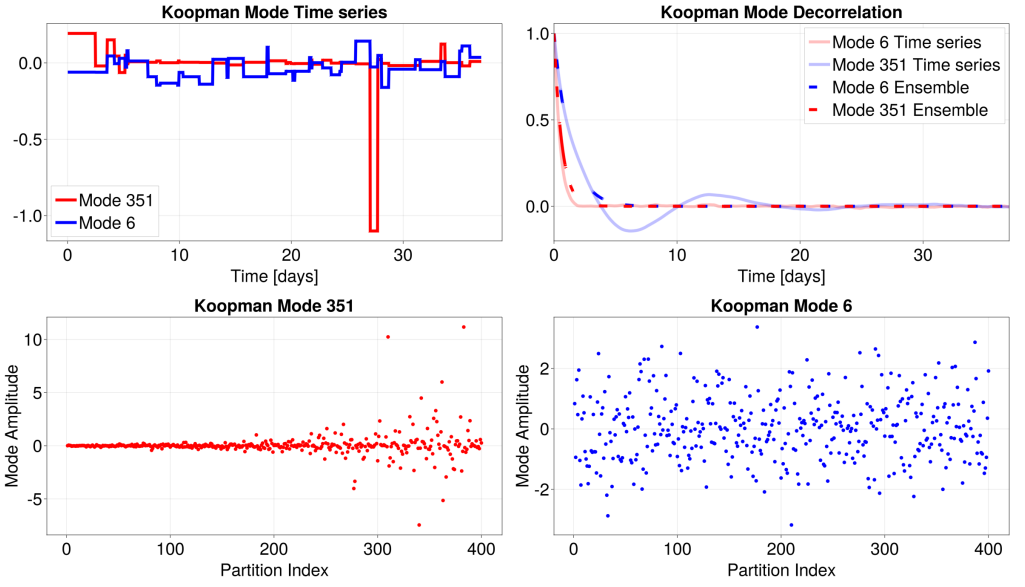

图 16：Held-Suarez Koopman 模态。我们以红色和蓝色展示两个数值 Koopman 模态。左上图为数值 Koopman 模态随时间的变化。右上图为通过生成元（虚线）和时间序列（实线）计算的数值 Koopman 模态的去相关时间尺度。下方各图为 Koopman 模态实部随状态索引的变化，因此也隐含地表示其作为状态函数的变化。

对于模态 351（红色），两种计算结果在所有时间上都一致；但对于模态 6（蓝色），这种一致性仅在最初约 5 天内成立。鉴于利用时间序列得到的衰减呈近指数结构，这表明可能存在对现有生成元的扰动，使时间序列计算与系综计算相一致。或许，约 14 天处的峰值与通常假定的大气去相关时间相同。底部两个面板显示了 Koopman 模态的实部随分区索引的变化。我们看到，模态 351 捕捉到了状态空间中不太可能出现的分区，而模态 6 则均匀分布于所有状态之间。二元分类算法会根据该模态的正值和负值对分区进行划分。

最后，我们展示了图 17中与表面温度场相关的 Koopman 模态振幅。因此，可观测量为

$$g^{\boldsymbol{x}_s}(\boldsymbol{s}) = \mathcal{T}_{\boldsymbol{x}_s} \tag{A 14}$$

其中 $\boldsymbol{x}_s$ 是表面温度，$\mathcal{T}_{\boldsymbol{x}_s}$ 是由方程4.34定义的温度可观测量。投影采用A10的离散类比计算，

$$G_\lambda(\boldsymbol{x}_s) \approx \sum_n g^{\boldsymbol{x}_s}(\boldsymbol{\sigma}^n)[\vec{\mathrm{P}}_\lambda]_n \tag{A 15}$$

其中，$\vec{\mathrm{P}}_\lambda$ 是生成元 Q 对应于特征值 λ的右特征向量。此外，$[\mathrm{P}_\lambda]_n$ 表示该特征向量的 $n'th$ 分量。我们可以看到，与波状模式相关联的振荡模态。

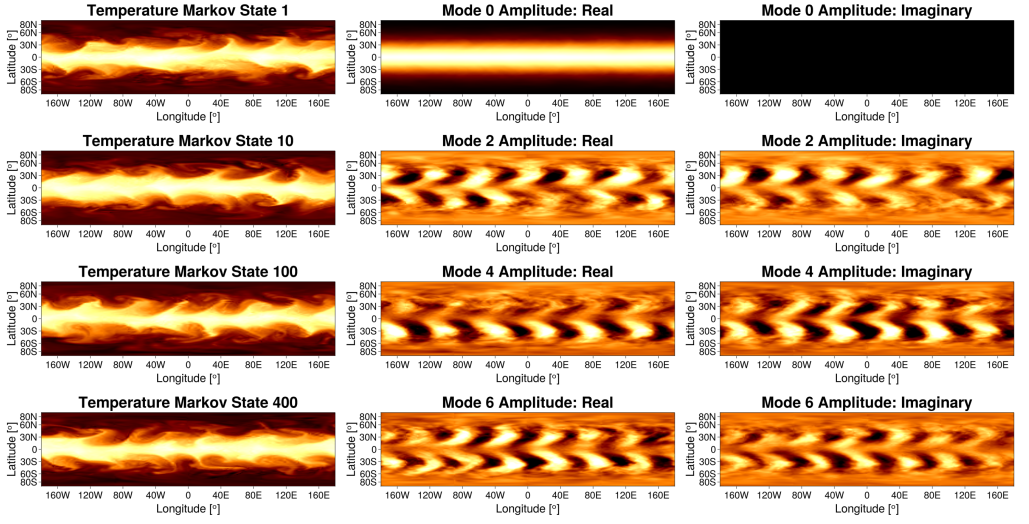

图 17：Held-SuarezKoopman 模态振幅。我们展示了四个具有代表性的表面温度场，这些温度场由马尔可夫状态构造而成，并通过将这些场投影到相应 Koopman 模态的实部和虚部，得到其 Koopman 模态振幅。统计稳态与模态零相关联。

A.2. 去相关与振荡的矩阵分解

为了进一步理解生成元 Q所关联的时间尺度，我们将矩阵分解为一个负半定分量和一个特征值纯为虚数的分量。首先，我们假设生成元 Q 是遍历的，因此它有一个零特征值，而所有其他特征值的实部都严格为负。

令 $\vec{\mathbb{P}} = [\mathbb{P}(\Omega_1), ..., \mathbb{P}(\Omega_n)]$ 为与特征值$\lambda = 0$对应的归一化特征向量，其中归一化方式取为

$$\mathbf{1}^T \vec{\mathbb{P}} = 1 \tag{A 16}$$

其中 $\mathbf{111}$ 是由所有 $1's$ 构成的向量。在遍历假设下，所有向量 $\vec{p}$的元素均为正。我们将矩阵 Q 分解为负半定部分和纯虚部分，如下所示

$$Q = \frac{1}{2}\underbrace{Q + PQ^T P^{-1}}_{\text{negative semi-definite}} + \frac{1}{2}\underbrace{Q - PQ^T P^{-1}}_{\text{imaginary eigenvalues}} \tag{A 17}$$

其中 $P = \text{Diagonal}(\vec{\mathbb{P}})$ 是一个对角矩阵，其对角线元素为稳态分布 $\vec{\mathbb{P}}$。关系式 $P^{-1}\vec{\mathbb{P}} = \mathbf{1}$ 成立。证明矩阵 $Q + PQ^T P^{-1}$ 为负半定矩阵如下。我们首先通过柯西-施瓦茨不等式观察到 $R_E(g, dt) \leqslant R_E(g, 0)$。更直观地说，这正是“可观测量最终会去相关”这一说法。接下来，

$$R_T(g, dt) = \boldsymbol{g}^T \exp(Qdt) P\boldsymbol{g} \approx \boldsymbol{g}^T P\boldsymbol{g} + dt\boldsymbol{g}^T QP\boldsymbol{g} \leqslant \vec{g}^T P\boldsymbol{g} \Rightarrow \boldsymbol{g}^T QP\boldsymbol{g} \leqslant 0 \tag{A 18}$$

由于 QP 是负半定的，且我们可以将 $\boldsymbol{g}$ 重缩放为 $\boldsymbol{hhh} = P^{-1/2}\boldsymbol{g}$，因此矩阵 $\tilde{Q} = P^{1/2} Q P^{-1/2}$ 是负半定的。因此，对称部分 $\tilde{Q}$ 是负半定的。

注意到相似变换

$$Q + PQ^T P^{-1} = P^{1/2}\left[P^{-1/2}QP^{1/2} + (P^{-1/2}QP^{1/2})^T\right]P^{-1/2} \tag{A 19}$$

$$Q - PQ^T P^{-1} = P^{1/2}\left[P^{-1/2}QP^{1/2} - (P^{-1/2}QP^{1/2})^T\right]P^{-1/2} \tag{A 20}$$

由于相似矩阵具有相同的特征值，证明得证。启发式地，分解的$Q+PQ^T P^{-1}$部分会导致时间上的去相关，而$Q+PQ^T P^{-1}$部分则会导致时间上的振荡。

对于 Held-Suarez 生成元，$Q + PQ^T P^{-1}$ 特征值对应的时间尺度范围为 $\approx$ 0。25 天至 $\approx$ 1。6 天，而 $Q - PQ^T P^{-1}$ 特征值对应的时间尺度范围为 25 天至 44 年。由于时间演化过程中两个矩阵分量之间的相互作用，纯振荡时间尺度从未被观测到。我们指出，当两个矩阵分量对易时，这种分解尤其有用。此外，$Q + PQ^T P^{-1}$ 矩阵包含与数值格式相关的数值耗散。

Froyland（2005）在从不可逆马尔可夫链定义时间可逆马尔可夫链时，曾对该矩阵的对称部分进行过评论。此外，关于漂移项和扩散项的拆分，另请参见 Klus 等人（2020）。

附录 B. 对称性

在第4.1节中，洛伦兹方程的对称性并未直接纳入生成元。我们在此弥补这一不足，并概述一种纳入对称性的方法。洛伦兹方程在变换 $(x, y, z) \mapsto (-x, -y, z)$下保持不变。只要存在一个混沌吸引子，就预期这种对称性也适用于混沌轨迹。为纳入这种对称性，我们分两步进行。

第一步是验证马尔可夫状态也满足这一对称性。由于$\boldsymbol{\sigma}^1 \mapsto \boldsymbol{\sigma}^3$、$\boldsymbol{\sigma}^3 \mapsto \boldsymbol{\sigma}^1$和$\boldsymbol{\sigma}^2 \mapsto \boldsymbol{\sigma}^2$ 在对称操作下，马尔可夫状态由洛伦兹方程的不动点定义，因此包含了该对称性。一般而言，必须将对称算子应用于每个马尔可夫状态，并在必要时纳入“对称状态”。

第二步是将对称性纳入所得的马尔可夫嵌入。例如，对于洛伦兹方程，如果我们观察序列

$$\text{first sequence} = 1, 1, 1, 2, 2, 3, 3, 1, 1 \tag{B 1}$$

然后，对上述序列应用对称操作可得

$$\text{second sequence} = 3, 3, 3, 2, 2, 1, 1, 3, 3 \tag{B 2}$$

然后，我们对第一个序列应用贝叶斯矩阵构造并计算后验分布。接着，我们将这些后验分布用作第二个序列进行贝叶斯矩阵构造时的新先验。这样得到的矩阵通过数据增强纳入了对称性。

我们在表 2 中展示了洛伦兹不动点生成元在这种对称性增强下的期望值。我们可以看到，本应为零的量的期望值现在确实为零。

类似的考虑也适用于其他类型的对称性。例如，连续对称性可近似为离散对称性，随后便可采用本文的方法。

	$\langle x\rangle$	$\langle y\rangle$	$\langle z\rangle$	$\langle xx\rangle$	$\langle xy\rangle$	$\langle xz\rangle$	$\langle yy\rangle$	$\langle yz\rangle$	$\langle zz\rangle$
系综	-0.0	-0.0	23.8	63.5	63.5	-0.0	63.5	-0.0	642.4
	$\langle xxy\rangle$	$\langle xxz\rangle$	$\langle xyy\rangle$	$\langle xyz\rangle$	$\langle xzz\rangle$	$\langle yyy\rangle$	$\langle yyz\rangle$	$\langle yzz\rangle$	$\langle zzz\rangle$
系综	0.0	1713.2	0.0	1713.2	0.0	0.0	1713.2	0.0	17346.1

表 2：洛伦兹吸引子的经验矩。系综平均与时间平均的比较。

附录 C. Held-Suarez

Isaac Held 和 Max Suarez 在 Held & Suarez (1994) 中引入了一项简化的大气模型测试。该测试案例有意不指定耗散机制，并旨在灵活适配计算中采用的预报变量或坐标系。其主要目的是作为一项稳健的“物理测试”，以便在不同的数值格式和运动方程之间进行比较。在第 C.1 节中，我们说明方程；在第C.2节中，我们说明所采用的数值离散化。最后，我们在第C.3节中总结第4节所讨论的若干要点，包括停留时间、矩阵元素的收敛以及特征值灵敏度。

C.1. 偏微分方程设置

我们选择使用一组保留完全可压缩动力学的方程，并以密度、总能量和笛卡尔动量作为预报变量，其方程如下

$$\partial_t \rho + \nabla \cdot (\rho \boldsymbol{u}) = 0 \tag{C 1}$$

$$\partial_t (\rho \boldsymbol{u}) + \nabla \cdot (\boldsymbol{u} \otimes \rho \boldsymbol{u} + p\mathbb{I}) = -\rho \nabla \Phi - 2\,(\boldsymbol{\Omega} \cdot \hat{r})\,\hat{r} \times \rho \boldsymbol{u} - k_v\,(\mathbb{I} - \hat{r} \otimes \hat{r})\,\rho \boldsymbol{u} \tag{C 2}$$

$$\partial_t (\rho e) + \nabla \cdot (\boldsymbol{u}\,(p + \rho e)) = -k_T \rho c_v \left(T - T_{\text{equilibrium}}\right) \tag{C 3}$$

其中 $\Phi = 2GM_P r_{\text{planet}}^{-1} - GM_P r^{-1}$ 为位势，$\boldsymbol{\Omega} = \Omega \hat{\mathbf{z}}$ 为行星角速度，$\hat{\mathbf{z}}$ 为行星自转轴方向，且 r 为球坐标系中的径向方向。科里奥利力投影到径向分量上，使得小行星类比对象（我们在第4节中用来进行模拟）具有与地球相似的气候态。此外，变量 $T_{\text{equilibrium}}$ 为依赖于纬度(φ)和气压 $\sigma = p/p_0$的辐射平衡温度，

$$T_{\text{equilibrium}}(\varphi, \sigma) = \max\left(T_{\min}, [T_{\text{equator}} - \Delta T_y \sin^2(\varphi) - \Delta\theta_z \ln(\sigma)\cos(\varphi)]\sigma^{R_d/c_p}\right), \tag{C 4}$$

而参数 k_v, k_T分别为动量阻尼和温度松弛的倒数时间尺度，其中

$$k_v = k_f \Delta\sigma \quad \text{and} \quad k_T = k_a + (k_s - k_a)\Delta\sigma \cos^4(\varphi), \tag{C 5}$$

其中 $\Delta\sigma = \max\{0, (\sigma - \sigma_b)/(1 - \sigma_b)\}$。温度和气压为，

$$T = \frac{1}{c_v \rho}\left(\rho e - \frac{1}{2}\rho \|\boldsymbol{u}\|^2 - \rho\Phi\right) \quad \text{and} \quad p = \rho R_d T. \tag{C 6}$$

参数	值	unit	描述
$\mathcal{X}$	80	-	缩放参数
z_{top}	3×10^4	m	大气高度
$r_{行星}$	6。371×10^6/X	m	行星半径
R_d	287	$m^2\ s^{-2}\ K^{-1}$	干空气气体常数
Ω	2π/86400 ×X	s^{-1}	科里奥利参数
p_0	1×10^5	$kg\ m^{-1}\ s^{-2}$	参考海平面气压
T_{min}	200	K	最低平衡温度
$T_{equator}$	315	K	赤道平衡温度
σ_b	0.7	-	无量纲阻尼高度
c_v	717.5	$J\ kg^{-1}\ K^{-1}$	干空气定容比热容
c_p	1004.5	$J\ kg^{-1}\ K^{-1}$	干空气定压比热容
k_f	X/86400	s^{-1}	动量阻尼时间尺度
k_a	X/(40 × 86400)	s^{-1}	极区松弛时间尺度
k_s	X/(4 × 86400)	s^{-1}	赤道松弛时间尺度
ΔT_y	60	K	纬向温差
$\Delta\theta_z$	10	K	垂直温差
G	6。67408×10^{-11}	千克$^{-1}$ 米^{3}秒$^{-2}$	万有引力常数
M_P	$5.9722/X^2 \times 10^{24}$	kg	行星质量

表 3：Held-Suarez 测试案例的参数值。数值 X = 1 对应标准测试案例，而 X = 80 是我们在此使用的版本。

模拟设置的参数值见表3。

我们对密度和总能量采用无通量边界条件，对水平动量采用自由滑移边界条件，对垂直动量采用不可穿透边界条件。初始条件是流体从静止状态开始，$\rho \boldsymbol{u} = 0$，处于等温大气中，

$$p(r) = p_0 \exp\left(-\frac{\Phi(r) - \Phi(r_{\text{planet}})}{R_d T_I}\right) \text{ and } \rho(r) = \frac{1}{R_d T_I} p(r) \tag{C 7}$$

其中我们使用 $T_I = 285K$。

C.2. 数值方法

为近似上一节的方程，我们采用 Souza等 （2022）概述、并由 Waruszewski 等（2022）精确定义的通量差分间断 Galerkin 方法。我们选择能够保持动能+势能的数值通量，以帮助确保流动的非线性稳定性，并采用 Roe 通量进行耗散。此外，Niegemann等（2012）提出的低存储四阶 14 阶段 Runge-Kutta 方法用于时间推进，并会产生一种数值耗散。所有模拟均在 NVidia Titan V 图形处理器上运行。

该区域是对半径为r_{planet}、高度为 z_{top}的薄球壳的分段多项式近似。该薄球形区域被划分为曲线单元，并采用区域的等参表示，以及 Ronchi等（1996）提出的立方球映射。从本质上说，这种选择将区域表示为分段多项式函数，其中多项式的阶数对应于离散化的阶数（Winters等）

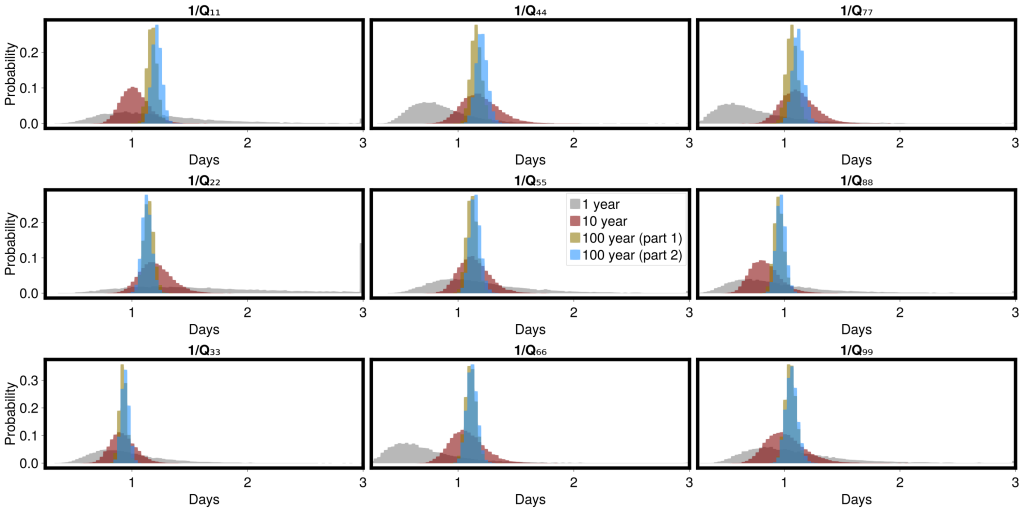

图 18：Held-Suarez生成器速率条目。图中展示了不同时间间隔下逆速率的不确定性。灰色、红色、蓝色和金色分别表示 1 年、10 年、100 年以及另一场（独立的）100 年模拟所对应的不确定性。可以看出，两组 100 年估计结果有显著重叠。

（2021年）。度量项按照 Kopriva（2006）的处理方式处理，并满足如下离散性质：常向量场的散度为零，即度量项具有自由流保持性质。

我们在垂直方向使用 4 个单元，在球面上使用 6×6^2 个单元（每个立方球面板使用 6^2 个单元），并在每个单元内使用 6 阶多项式。鉴于我们有 5 个预报状态（密度、笛卡尔动量的三个分量和总能量），自由度总数为 $5{\times}4{\times}(6{\times}6^2){\times}7^3 = 1,481,760$——这决定了水平声学 CFL 限制的时间步长。

C.3. 划分性质与不确定性量化

我们展示两幅图来考察收敛。在图 18 中，我们展示了生成元最可能的前 16 个状态对应的逆停留时间（对角项）。可以看出，在不同的时间间隔下，矩阵元素似乎趋于收敛。

在图 19中，我们展示了生成器矩阵随机样本所对应的分布中的逆特征值实部。该变量对应于分区给出的去相关时间尺度。贝叶斯方法表明，我们无法信赖数值解所得的最慢去相关时间尺度，因为它在 1.5 天到 20 天之间变化。另一方面，其他特征值也不能被视为毫无意义，因为对于在不相交时间子集上收集的数据，其概率分布彼此重叠。作为一项技术说明，可以利用均值生成器的特征值分解作为迭代过程的初始猜测，从而加速特征值的不确定性传播。

如第4.1 节所述，我们展示了特定状态的停留时间分布。由于共有四百个状态，我们在图20中展示了概率最高的前三个状态的停留时间。分位数近似呈指数分布，但仔细观察后会发现并非完全如此，这与预期相符。

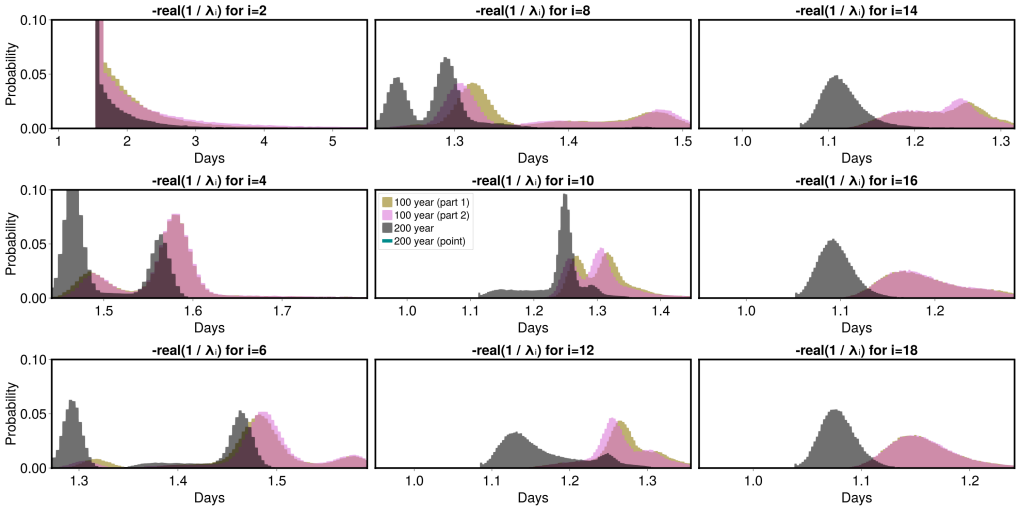

图19：Held-Suarez 生成元的去相关时间尺度。我们传播随机生成元的不确定性，以考察推断出的特征值分布。我们针对三种情况传播不确定性：完整的200年、前100年和后100年。此外，我们还展示了根据随机矩阵均值计算得出的200年生成元点估计。最慢的去相关时间尺度对扰动最为敏感，其他特征值则不那么敏感。

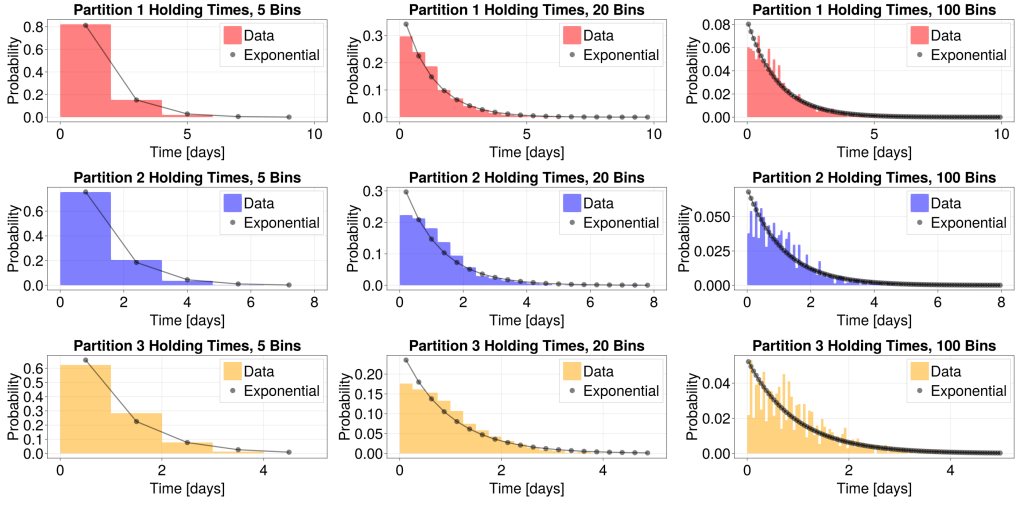

图20：Held Suarez 停留时间：连续时间马尔可夫模型与时间序列经验分布。状态中停留的时间大致呈指数分布。从可能的时间范围可以看出，状态3的停留时间最长。另一方面，状态1出现的概率最高，且其停留时间不太符合指数分布。

参考文献

Allawala、Altan 和 Marston，J. B. 2016年 随机受迫 Lorenz 吸引子的统计量：福克-普朗克方程与累积量展开。*Phys. Rev. E* **94**，052218。Arbabi、Hassan 和 Mezić，Igor 2017 遍历理论、动态模态分解，以及计算

Koopman 算子的谱性质。《*SIAM* 应用动力系统期刊》**16** (4), 2096–2126，arXiv：https://doi.org/10.1137/17M1125236。Colbrook, Matthew & Townsend, Alex 2021 动力系统 Koopman 算子谱性质的严格数据驱动计算。*arXiv:2111.14889* 。Colbrook, Matthew J. 2022 用于数据驱动计算保测动力系统的 mpedmd 算法。Cvitanović, P., Artuso, R., Mainieri, R., Tanner, G. & Vattay, G. 2016 《混沌：经典与量子》。哥本哈根：尼尔斯·玻尔研究所。Dellnitz, Michael, Froyland, Gary & Junge, Oliver 2001 gaio 背后的算法——动力系统的集合定向数值方法。载于《遍历理论、分析与动力系统的高效模拟》（编者 Bernold Fiedler），第145–174页。柏林、海德堡：Springer Berlin Heidelberg。Dellnitz, Michael, Junge, Oliver, Koon, Wang Sang, Lekien, Francois, Lo, Martin W., Marsden, Jerrold E., Padberg, Kathrin, Preis, Robert, Ross, Shane D. & Thiere, Bianca 2005 动力天文学与多体问题中的输运。《国际分岔与混沌期刊》**15** (03), 699–727，arXiv：https://doi.org/10.1142/S0218127405012545。Fernex, Daniel, Noack, Bernd & Semaan, Richard 2021 基于簇的网络建模——从快照到复杂动力系统。《科学进展》**7**, eabf5006。Froyland, Gary 1997 相关性衰减速率的计算机辅助界限。《数学物理通讯》**189** (1), 237–257。Froyland, Gary 2005 统计最优的近似不变集。《物理学*D*辑：非线性现象》**200** (3), 205–219。Froyland, Gary, Giannakis, Dimitrios, Lintner, Benjamin R., Pike, Maxwell & Slawinska, Joanna 2021 运用算子理论方法分析气候动力学的谱。《自然·通讯》**12**(1), 6570。Froyland, Gary, Junge, Oliver & Koltai, Péter 2013 无需轨迹积分估计流的长期行为：无穷小生成元方法。《*SIAM* 数值分析期刊》**51** (1), 223–247，arXiv：https://doi.org/10.1137/110819986。Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki & Rubin, Donald B. 2013 《贝叶斯数据分析》，第3版。Chapman and Hall/CRC。Held, I. M. & Suarez, M. J. 1994 关于大气环流模式动力核心互比的一项提议。《美国气象学会公报》**75** (10), 1825–1830，版权：Copyright 2018 Elsevier B.V.，保留所有权利。Ho, Jonathan, Jain, Ajay & Abbeel, Pieter 2020 去噪扩散概率模型。*arXiv* 预印本 *arxiv:2006.11239* 。Hopf, Eberhard 1948 展现湍流特征的一个数学示例。《纯粹与应用数学通讯》**1** (4), 303–322，arXiv：https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.3160010401。Hopf, Eberhard 1952 统计流体力学与泛函演算。《印第安纳大学数学期刊》**1**, 87–123。Klus, Stefan, Nüske, Feliks, Peitz, Sebastian, Niemann, Jan-Hendrik, Clementi, Cecilia & Schütte, Christof 2020 Koopman 生成元的数据驱动近似：模型降阶、系统辨识与控制。《物理学*D*辑：非线性现象》**406**, 132416。Kopriva, David A. 2006 曲线网格上的度量恒等式与间断谱元方法。《科学计算期刊》**26** (3), 301。Lloyd, S. 1982 PCM 中的最小二乘量化。《*IEEE* 信息论汇刊》**28** (2), 129–137。Lorenz, E. N. 1963 确定性非周期流。《大气科学期刊》**20**, 130–141。Maćešić, Senka & Črnjarić-Žic, Nelida 2020 《非自治与随机系统的 *Koopman* 算子理论》，第131–160页。Cham：Springer International Publishing。Niegemann, Jens, Diehl, Richard & Busch, Kurt 2012 稳定区域优化的高效低存储 Runge-Kutta 数值格式。《计算物理学期刊》**231**, 364–372。Parker, Jeremy P. & Page, Jacob 2020 孤立锋面与孤子的 Koopman 分析。《*SIAM* 应用动力系统期刊》**19** (4), 2803–2828，arXiv：https://doi.org/10.1137/19M1305033。Ronchi, C., Iacono, R. & Paolucci, P.S. 1996 “立方球面”：球面几何中偏微分方程求解的新方法。《计算物理学期刊》**124** (1), 93–114。

Schmid, Peter J. 2010 数值和实验数据的动态模态分解。流体力学杂志 **656**, 5–28。Schütte, Christof, Klus, Stefan & Hartmann, Carsten 2022 克服分子动力学中的时间尺度障碍：传递算子、变分原理与机器学习。技术报告 22-25。ZIB, Takustr. 7, 14195 Berlin。Souza, Andre, He, Jia, Bischoff, Tobias, Waruszewski, Maciej, Novak, Lenka, Barra, Valeria, Gibson, Thomas, Sridhar, Akshay, Kandala, Sriharsha, Byrne, Simon, Wilcox, Lucas, Kozdon, Jeremy, Giraldo, Frank, Knoth, Oswald, Ferrari, Raffaele, Marshall, John & Schneider, Tapio 2022 通量差分间断伽辽金法在理想化全可压缩非静力干燥大气中的应用。期刊=*ESSOAR* 预印本 。Stefan Klus, Péter Koltai, and Christof Schütte 2016 Perron-Frobenius 算子与 Koopman 算子的数值近似。计算动力学杂志 **3** (1), 51–79。Ulam, S. M. 1964 现代数学中的问题。载于现代数学中的问题。Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Ł ukasz & Polosukhin, Illia 2017 注意力就是你所需要的一切。载于神经信息处理系统进展 (编者 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett), , vol. 30。Curran Associates, Inc. Wang, Han & Schütte, Christof 2015 为周期驱动的非平衡系统构建马尔可夫状态模型。化学理论与计算杂志 **11** (4), 1819–1831。Waruszewski, Maciej, Kozdon, Jeremy, Wilcox, Lucas, Gibson, Thomas & Giraldo, Francis X. 2022 非守恒形式平衡律的熵稳定间断伽辽金法：在含重力欧拉方程中的应用。计算物理学杂志 **(即将出版)**。Winters, A.R., Kopriva, D.A., Gassner, G.J. & Hindenlang, F. 2021 可压缩纳维–斯托克斯方程的现代稳健节点间断伽辽金谱元法的构建。*CISM* 国际机械科学中心，课程与讲座 **602**, 117–196。Zhang, Xiangxiong & Shu, Chi-Wang 2011 满足最大值原理且保持正性的守恒律高阶数值格式：综述与最新进展。英国皇家学会会刊 *A*：数学、物理与工程科学**467** (2134), 2752–2776, arXiv: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2011.0153。