arXiv:2303.08587v1 [cs.LG] 14 Mar 2023

DELAY-SDE-NET：一种用于时间序列建模的深度学习方法，兼顾记忆与不确定性估计

Mari Dahl Eggen
奥斯陆大学数学系，挪威奥斯陆
marideg@math.uio.no

Alise Danielle Midtfjord奥斯陆大学数学系，挪威奥斯陆
alisedm@math.uio.no

摘要

准确地对时间序列建模对于众多领域都至关重要。由于现实世界通常过于复杂，无法进行精确建模，因此评估某个动力系统处于特定状态的概率往往很有意义。本文介绍 Delay-SDE-net，这是一种基于随机时滞微分方程（SDDE）的神经网络模型。采用具有多个时滞的 SDDE 作为建模框架，使其适用于存在记忆效应的时间序列，因为它通过系统的先前状态纳入了记忆。Delay-SDE-net 的随机部分为建模中的不确定性估计提供了基础，并拆分为两个神经网络，以分别处理偶然不确定性和认知不确定性。模型能够即时提供不确定性估计，因此适用于时间点稀疏的应用场景。我们推导了 Delay-SDE-net 的理论误差，并通过数值方法分析其收敛率。与类似模型比较时，Delay-SDE-net 始终表现最佳，无论是在时间序列值预测方面，还是在不确定性预测方面。

关键词 物理信息神经网络 · 时间序列 · 不确定性估计 · 随机时滞微分方程 · Barron 空间

1 引言

在经济学、工程学和自然科学等众多领域，准确地对时间序列建模都十分重要。正如 Box 等人（2008）所指出的，某些动力系统可以利用物理定律，几乎精确地在时间上建模。在这种情况下，确定性模型通常是常微分方程（ODE）的解，便足以胜任。然而，现实世界通常过于复杂，无法进行精确建模。为此，我们可以采用随机微分方程（SDE），评估动力系统处于特定状态的概率，从而有可能获得准确的不确定性估计。

尽管许多随机微分方程都已得到了存在唯一性结果（例如 Øksendal（2003）和 Protter（2005）），但大多数方程都无法求得显式解。因此，动力系统通常被假设遵循简单的随机微分方程，例如假设其漂移项和扩散项为常数或线性形式。考虑自相关动力系统的数学模型则更具挑战性。正如 Hu 等人（2001）所解释的，许多物理系统的变化率不仅取决于系统的当前状态，也取决于其过去的状态。对于这种更复杂的情况，可以使用自回归滑动平均（ARMA）模型及其相关模型等统计模型（例如 Brockwell、Davis（1991））。这些统计模型假设动力系统中存在线性的自回归相互作用。已有研究表明，ARMA 模型存在连续时间类比形式，多变量情形下亦是如此（Marquardt、Stelzer，2007）。这种连续时间类比形式称为连续时间 ARMA（CARMA）模型，它是一个具有已知显式解的线性随机微分方程组（例如 Brockwell（2014））。

尽管线性自回归模型有望很好地近似动力系统，但现实世界中的系统往往具有固有的非线性时间依赖关系。人们开发了许多统计方法来对这类非线性

系统进行建模，例如 Corduas（1994年）；Kantz、Schreiber（2003年）以及 Teräsvirta 等人（2010年）。用于表示非线性时间依赖关系的数学模型称为随机时滞微分方程（Mohammed，1998年），其中 CARMA 模型是其一个特例（Basse-O'Connor 等人，2020年）。此外，过去几十年间，人们开发了一系列基于神经网络的非线性时间序列模型，例如时延神经网络（Waibel 等人，1989）、ARIMA 与神经网络混合模型（Zhang，2003年）、循环神经网络（Rumelhart 等人，1986年）、长短期记忆（Hochreiter、Schmidhuber，1997年），以及将时间序列编码并结合卷积神经网络的方法（Wang、Oates，2015年；Gu 等人，2018年；Borovykh 等人，2017年；Zheng 等人，2014年）。最近，Chen 等人（2018）通过引入 ODE-net，为时间序列建模领域作出了贡献。ODE-net 基于常微分方程（ODE）的连续动力学，其导数由神经网络给出。尽管 ODE-net 可用于对动力系统建模，但它并未考虑先前状态。也就是说，ODE-net 没有记忆，因此，与前文提及的网络相比，它并不适合对自回归动力系统进行建模。像 ODE-net 那样，通过显式数学结构约束神经网络，这种方法称为物理信息神经网络（Raissi 等人，2019）。物理信息方法可能为神经网络带来若干优良特性，因为网络获得了物理启发约束和知识等额外信息。神经网络模型有望以更少的训练数据实现更高的可解释性、更强的抗扰动数据鲁棒性，以及在分布外（OOD）数据上更好的稳定性（Karniadakis 等人，2021；Kashinath 等人，2021；Yang、Perdikaris，2019）。

神经网络方法凭借其预测能力受到广泛关注，也因此成为时间序列预测中的有力工具。与某些经典模型相比，神经网络所需的数据预处理较少（例如，需要平稳数据的 ARMA 模型），从而加快了数据准备过程。然而，尽管神经网络能够捕捉复杂模式并提供高准确度的预测，但研究表明它们并不稳定。也就是说，如果输入数据偏离训练数据，例如出现微小扰动、结构变化或对抗性输入，神经网络可能会给出严重错误且过度自信的预测（Antun等人，2020年；Papernot、McDaniel，2018年）。此外，基础神经网络无法提供预测不确定性的估计。在安全关键型应用中，不确定性估计尤为重要，例如医学（Lambrou等人，2010年；Wainberg等人，2018年）、风险分析（Varshney，2016年；Tambon等人，2022年）和自主系统（Pereira、Thomas，2020年）；在这些领域，了解何时可以信任一项预测至关重要。此外，有些应用领域涉及具有较大固有不确定性的动力系统，而量化这种不确定性十分重要，例如气象学（Scher、Messori，2018年；Wang等人，2019a）和金融（Jeon等人，2020年；Zhang等人，2020年）。鉴于这些需求，机器学习研究中出现了一种新趋势：在关注实现高预测准确度的同时，也开始重视提供额外的决策支持，例如可解释性（Zhang等人，2022年；Midtfjord等人，2022年）和不确定性估计（Kabir等人，2018年）。

最常用的不确定性估计方法是贝叶斯神经网络（Denker、LeCun，1990年；MacKay，1992年），该方法通过在模型参数上建立概率分布来量化不确定性。使用这种方法的不便之处包括需要指定合适的先验分布，以及必须执行多次迭代才能对参数的不确定性进行边缘化。其他用于不确定性估计的方法包括集成方法（Kurutach等人，2018年；Rajeswaran等人，2016年）和测试时增强方法（Wang等人，2019b；Kim等人，2020年）。前者通过多个预测器组成的集成来计算预测，并将预测器之间的方差作为不确定性估计；后者则对经过变换的输入数据，使用确定性网络重复进行预测。上述所有方法的一个缺点是，为了获得不确定性估计，需要反复计算预测，这既会增加计算开销，也会耗费时间。关于神经网络预测中不确定性的估计与量化研究综述，参见Abdar等人（2021年）和Gawlikowski等人（2021年）。

正如 Gawlikowski等人（2021年）所述，统称为单一确定性方法 的不确定性估计方法，无需反复计算预测，因此不存在这方面的不便。这类方法会直接估计不确定性，或是在神经网络本身内部进行估计（Mo˙zejko 等人，2018年；Nandy 等人，2020年），或通过并列网络对预测网络进行不确定性推断（Raghu 等人，2019年；Oberdiek 等人，2018年）。采用单一确定性方法，可以即时得出不确定性估计，因此适用于文本翻译、自主系统以及金融领域的预测等现实应用。

另一种较新的方法旨在为确定性预测提供相应的不确定性估计，即采用受随机微分方程启发的神经网络模型，参见 Tzen、Raginsky（2019年）以及 Li 等人（2020a）。此外，Kong 等人（2020年）通过开发 SDE-net，为这一领域作出了贡献。此后，为了实现更灵活的建模，研究者对 SDE-net 进行了进一步创新，例如 Hayashi、Nakagawa（2022年）；Wang 等人（2021年）；Wang、Yao（2021年）以及 Yang 等人（2021年）。SDE-net 与 ODE-net 类似，都是用神经网络对导数进行建模，但 SDE-net 还加入了一个额外的神经网络，用于对随机项进行建模。尽管这些神经随机微分方程可以

视为动力系统的一种状态变换，但它们并未考虑先前状态，因此不太适合对自回归动力系统进行建模。

要准确估计不确定性并非易事，因为不确定性可能源自不同的因素。偶然不确定性 指特定任务中固有的自然随机性，而认知不确定性则源于模型知识不足，与分布外数据检测密切相关（Hüllermeier、Waegeman，2021年）。在许多应用中，区分这两种不确定性十分重要，因为这能为预测结果的可信度提供额外信息，并帮助了解如何改进模型（Der Kiureghian、Ditlevsen，2009）。前文提到的 SDE-net 能够对这两种不确定性进行一定程度的估计，但主要侧重于分布外数据检测。在时间序列建模中，时变方差（异方差性）是一种典型特征。这就需要准确估计偶然不确定性，因为方差较大的时期会增加模型的不确定性。因此，一些 SDE-net 改进方法更侧重于处理偶然不确定性（Wang等人，2019b；Wang、Yao，2021年）。

以上讨论表明，时间序列模型若能描述具有非线性依赖关系的自相关动力系统，并给出精确的不确定性估计，将大有裨益。也就是说，合适的模型可以像 ARMA 模型那样考虑先前状态，像神经网络那样描述非线性，并能够估计和区分偶然不确定性与认知不确定性。本文提出了 Delay-SDE-net，这是 SDE-net（Kong 等人，2020年）的扩展版本，其中纳入了时间序列的滞后值。Delay-SDE-net 基于具有多个时滞的随机时滞微分方程（SDDE）。模型系数由神经网络给出，从而提供了灵活的模型估计框架。此外，Delay-SDE-net 的随机部分被拆分为两个神经网络，分别表示偶然不确定性和认知不确定性。这些网络按照单一确定性方法进行训练，因此能够立即提供预测和不确定性估计。

本文首先在第 2 节介绍预备知识，包括记号、不确定性估计理论、随机时滞微分方程简介，以及双层神经网络的数学理论。第 3 节介绍 Delay-SDE-net，包括其理论框架和训练算法。第 4 节推导双层 Delay-SDE-net 的理论误差界，并在第 5 节通过数值实验进一步检验。本节还使用模拟数据和现实案例研究，将 Delay-SDE-net 与基准模型进行比较。最后，第 6 节总结研究结果并给出结论性评述。

2 预备知识

本节将从偶然不确定性和认知不确定性的角度介绍模型不确定性理论。随后，我们将介绍一类特定的 Itô 随机时滞微分方程及其对应的 Milstein 格式。接下来简要介绍双层神经网络和 Barron 空间。关于各主题的更完整介绍，请参阅相关参考文献。我们首先介绍记号。

2.1 记号

为方便读者，本节介绍并总结本文中使用的记号。首先，假设给定一个完备滤过概率空间 $(\Omega, \mathcal{F}, \{\mathcal{F}_{t\geq 0}\}, P)$，在该空间上定义了一个 d_W 维标准布朗运动过程 $\{W(t) := (W^1(t) \dots, W^{d_W}(t))\}_{t\geq 0}$。

现在，定义 $\mathbb{N} := \mathbb{N}_0 \backslash \{0\}$，并令 $\mathbb{R}^d$ 表示 d 维欧几里得空间，其标量的标准范数记为 $|\cdot|$，向量的标准范数记为 $\|\cdot\|_2$。向量 u 和 v 之间的常规点积记为 uv，相应的 Hadamard积记为 $u \circ v$，而 v^T 表示 v的转置。此外，定义 L_2-范数 $\|\cdot\|_{L_2} := \|\cdot\|_{L_2(\Omega,\mathbb{R}^d)} = (E_{x\sim\mu}[|\cdot|^2])^{1/2}$，其中 $x \in \mathbb{R}^d$，概率测度为 $\mu \in P(\Omega)$；并定义上确界范数 $\|\cdot\|_C := \sup_{-\tau\leq s\leq 0}|\cdot|$，其取遍所有连续函数（多维路径） $C([-\tau, 0]; \mathbb{R}^d)$，其中延迟固定为 $\tau > 0$。连续函数 $C^{i_t,i_x}(T \times \mathbb{R}^{dp}, \mathbb{R}^d)$ 关于时间和空间分别具有至多 i_t 阶和 i_x 阶有界导数。对于某个有界域 D，我们用 $\mathcal{B} := \mathcal{B}_2(D, \mathbb{R})$ 表示本文所关注的 Barron 空间。由 $W(t)$ 驱动的 d 维随机过程 $X(t)$，对应一个连续时间延迟随机微分方程网络模型$X^{(m)}(t)$，其中 $m \in \mathbb{N}$ 表示模型系数中的神经元数量。 $X(t)$ 和 $X^{(m)}(t)$ 的离散时间版本分别记为 $X^\pi(t)$ 和 $X^{\pi,(m)}(t)$，其中 π 表示一个定义好的离散化方案。此类向量过程的分量用额外的上标 $j = 1, \dots, d$表示。

˜在实证分析中，我们使用训练数据集 $\mathcal{D}_0$、验证集和测试数据集 $\mathcal{D}$、分布外数据集 $\mathcal{D}$以及修改后的数据集 $\mathcal{D}^*$。对于每个时间点，我们将 $\boldsymbol{x}_{t_k} = [x_{1,t_k}, \dots, x_{d,t_k}]$记为一个 d维数据点。Delay-SDE-net 的估计模型系数分别记为漂移项的 f_m 和扩散项的 g_m （或者在讨论偶然扩散和认知扩散时，记为 $g_{a,\ m}$ 和 $g_{e,\ m}$

分别）。此外，其他模型的估计模型系数（ˆ）（或一般模型的系数）分别记为漂移项的 f 和扩散项的 $\hat{g}$（或者在讨论偶然扩散和认知扩散时，分别记为 $\hat{g}_a$ 和 $\hat{g}_e$ ）。最后， d维向量过程 $X(t)$ 和向量数据点 $\boldsymbol{x}_{t_k}$ 的各元素均服从正态分布，其均值为 μ、标准差为 σ，记作 $X(t)\ \boldsymbol{x}_{t_k} \sim N(\mu,\ \sigma)^d$。

2.2 不确定性估计

本节介绍与数据驱动模型不确定性量化相关的一些重要因素。我们尤其关注区分两种不同不确定性来源的重要性，即偶然不确定性和认知不确定性（Der Kiureghian、Ditlevsen，2009）。

偶然不确定性是指任务中自然随机性所导致的不确定性，通常认为无法降低。换言之，观测到的数据点 y 来自数据生成函数 f，并受到随机噪声 $y = f(x) + e$ 的干扰，其中由 $e \sim N(0,\ \sigma)$ 导致的不确定性就是偶然不确定性。

ˆ 认知不确定性是指估计模型 $f(x)$ 中知识的缺乏，被认为可以通过获取更多信息来降低。因此，认知不确定性是由估计模型与真实函数 $f(x) - f(x)$ 之间的差异 ˆ 所导致的不确定性，如图 1 所示。文献中通常将认知不确定性的来源分为两类（Sluijterman 等人，2021年）。

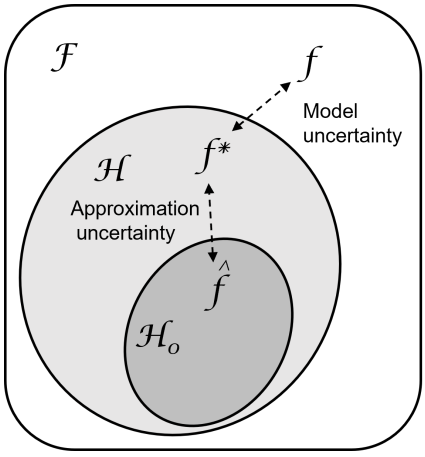


图 1：认知不确定性的两个来源示意图。 $\mathcal{F}$ 是真实函数空间，它将 x 映射到 y。 $\mathcal{H}$ 是假设空间，即算法可以学习的可能模型函数类别。 $\mathcal{H}_0$ 包含于 $\mathcal{H}$中，是算法接受训练所用的假设空间。

ˆ 首先，近似不确定性 源于学习到的模型 f 在假设空间 $\mathcal{H}$中训练得不够充分。这意味着我们尚未在可能模型函数类别中找到最优模型 f^* 。此外，模型接受训练所用的假设空间 $\mathcal{H}_0$可能无法涵盖整个假设空间 $\mathcal{H}$。这会导致模型在 $\mathcal{H}_0$之外不稳定且存在不确定性。增加训练样本数量以及这些样本所覆盖的空间范围，可以降低近似不确定性。

认知不确定性的第二个来源是模型不确定性，它考虑的是假设空间的覆盖范围。也就是说， $\mathcal{H}$ 可能无法涵盖真实函数空间 $\mathcal{F}$。因此，模型可能无法恰当地复现真实函数，例如模型复杂度不足。降低模型不确定性的可行方法包括更改模型结构、更改学习算法，或通过新增或转换解释变量来提供额外信息。由于模型无法判断真实的底层函数是否包含在其假设空间中，模型不确定性常常与随机噪声，也就是偶然不确定性，混为一谈。

为了正确表征模型预测的不确定性，不确定性估计应同时包含偶然不确定性和认知不确定性（Hora, 1996）。然而，传统机器学习方法无法区分两者，并且往往只会给出其中一种不确定性的估计（Hüllermeier, Waegeman, 2021年）。关于偶然不确定性和认知不确定性之间区别的详细讨论，请参阅 Der Kiureghian, Ditlevsen (2009)。

由于神经网络已被证明能够近似几乎任意函数（Hornik 等，1989；Lu 等，2017；Benth 等，2022），即使只有一个隐藏层也不例外，因此理论上，（认知）模型不确定性可以降至零。（认知）近似不确定性也是如此，理论上可以降至零

前提是训练时提供了来自整个假设空间的足够多数据点。尽管如此，正如 Colbrook 等（2022）所述，在实际应用中，大多数情况下几乎不可能训练出最优的神经网络。此外，获取所有可能情形下的数据也很少可行。这意味着，包括神经网络在内的模型，其偶然不确定性和认知不确定性估计仍将持续受到关注。不过，随着模型不断改进、数据量不断增加，偶然不确定性可能会成为影响最大的部分。

2.3 具有多个时滞的随机时滞微分方程与 Milstein 格式

本节介绍 Hu 等（2004）提出的具有多个时滞的 Itô 随机时滞微分方程，以及相应的 Milstein 格式。

为建立随机时滞微分方程框架，需要引入一些概念。首先，定义一个投影 $\Pi: C \to \mathrm{R}^{dp}$，它与给定的固定时间点 $u_1, \ldots, u_p \in [-\tau, 0]$ 相关，定义如下：

$$\Pi(\eta) \coloneqq (\eta(u_1), \ldots, \eta(u_p)) \in \mathbb{R}^{dp}, \tag{1}$$

其中 $\eta: \Omega \to C$ 是一个 $\mathcal{F}_0$-可测的初始过程。此外，对于任意连续的 d 维随机过程 $X(t): [-\tau, T] \times \Omega \to \mathrm{R}^d$，其中 τ, $T \in \mathrm{R}^+$，我们将其分段过程定义为

$$X_t(u) \coloneqq X(t+u), \quad u \in [-\tau, 0], \quad t \in [0, T], \tag{2}$$

其取值属于 C。

给定两个投影 Π_1 和 Π_2 及其对应的时间点 $u_{1,\,1}, \ldots, u_{1,\,p_1} \in [-\tau, 0]$ 和 $u_{2,\,1}, \ldots, u_{2,\,p_2} \in [-\tau, 0]$，分别定义如下这类随机时滞微分方程。设随机（向量）过程 $X(t)$ 是以下方程的（唯一强）解

$$X(t) = \begin{cases} \eta(0) + \int_0^t f(s, \Pi_1(X_s))ds + \int_0^t g(s, \Pi_2(X_s))dW(s), & t \geq 0, \\ \eta(t), & -\tau \leq t < 0, \end{cases} \tag{3}$$

其中，$W(t)$ 是取值于 R^{dw} 的布朗运动过程（例如参见 Øksendal，2003年），系数 $f: [0,T] \times \mathrm{R}^{dp_1} \to \mathrm{R}^d$、$g: [0,T] \times \mathrm{R}^{dp_2} \to \mathrm{R}^{d \times dw}$ 为温和函数（参见 Hu 等，2004年），并满足 Hu 等（2004年）中的 Lipschitz 条件和有界性条件（分别见公式（1.4）和（1.5））。那么，对于每个 $q \geq 1$，都存在一个常数 $C_q = C_q(q, C_L, T) > 0$，其中 C_L 为 Lipschitz常数，使得

$$E\|X_t\|_C^{2q} \leq C_q(1 + E\|\eta\|_C^{2q}),$$

对所有的 $\eta \in C$、$t \in [0,T]$，其中 $\|\cdot\|_C \coloneqq \sup_{-\tau \leq s \leq 0} |\cdot|$（见第2.1节）。

将时间区间 $[-\tau, T]$ 等距离散化，定义如下：

$$\mathcal{T} = \{-\tau = t_{-L} < t_{-L+1} < \cdots < 0 = t_0 < t_1 < \cdots < t_K = T\}, \tag{4}$$

对于整数 $L \geq 0$、$K \geq 1$，且 $\pi \coloneqq \Delta t = t_{k+1} - t_k$ 对于 $k = \{-L, \ldots, K-1\}$。我们假设 $\mathcal{T}$ 定义为包含时滞集合 $\{u_{1,1}, \ldots, u_{1,p_1}, u_{2,1}, \ldots, u_{2,p_2}\} \in \mathcal{T}$。此外，记 $f^j(t, x^{(1)})$ 和 $g^{ji}(t, x^{(2)})$ 分别为 $f \in \mathrm{R}^d$ 和 $g \in \mathrm{R}^{d \times d_W}$ 中的元素 j 和 (j,i)，其中 $x^{(1)} \in \mathrm{R}^{dp_1}$ 且 $x^{(2)} \in \mathrm{R}^{dp_2}$。对于情形 $f \in C^1$，$2(T \times \mathrm{R}^{dp_1}, \mathrm{R}^d)$、$g \in C^1$、$2(T \times \mathrm{R}^{dp_2}, \mathrm{R}^{d \times dW})$ 和 $W(s) = W(0) = 0$、$s \leq 0$，我们将公式（3）中的随机时滞微分方程的 Milstein格式定义为

$$\begin{aligned} X^{j,\pi}(t_{k+1}) =& X^{j,\pi}(t_k) + f^j(t_k, \Pi_1(X^\pi_{t_k}))\Delta t + g^{ji}(t_k, \Pi_2(X^\pi_{t_k}))\Delta W^i(t) \\ &+ F^j\left(g^{j_2 i_2}, \frac{\partial g^{ji}}{\partial x_{j_2 i_2}}\right), \end{aligned} \tag{5}$$

其中，$X^{j,\pi}(t)$ 是 $X(t) \in \mathrm{R}^d$ 的第 j 个分量的离散化版本，上标 π 表示应用了式（4）中的时间离散化。最后，我们有 $\Delta W^i(t) \coloneqq W^i(t_{k+1}) - W^i(t_k) \in \mathrm{R}^{dw}$。$F(\cdot,\cdot)$ 的精确表达式见 Hu 等（2004），并注意到当 $\partial g^{ji}/\partial x_{j_2 i_2} = 0$ 时，$F^j = 0$。

2.4 双层神经网络与 Barron 空间

本文从 Barron 空间的角度讨论由双层神经网络表示的函数的存在性和误差界。正如 E 等人（2021）所述，Barron 空间定义为一类连续函数的集合，这些函数可表示为连续时间版本的双层神经网络，且具有有限的 Barron 范数。也就是说，一个连续函数

若能由双层神经网络近似，且近似误差受 Barron 范数界定，则属于 Barron 空间。下面给出更正式的定义。

如 E 等人（2019）所述，我们定义目标函数 $f^* = E[y \mid x]$，其中 $x \in \mathrm{R}^d, y \in \mathrm{R}$，该函数可通过从底层分布 $p_{x,\ y}$ 中抽取的独立同分布数据样本 $\{\boldsymbol{x}_i, y_i\}_{i=0}^n, n \in \mathrm{N}$ 学得。目标函数 f^* 在图 1 中以函数空间的形式展示。我们将来自 $p_{x,\ y}$ 的数据称为分布内（ID）数据。相反，分布外（OOD）数据是指从不同于分布内数据的概率分布中抽取的数据，记作 $\tilde{p}_{x,\ y}$。以下关于双层神经网络和 Barron 空间的介绍适用于这两种数据类型。关于如何在神经网络训练中使用分布外数据的说明，见第 3.2.3 节。
双层神经网络可表示为

$$f_m(x;\theta) = \sum_{i=1}^{m} a_i \sigma(w_i^T x + b_i), \tag{6}$$

其中 $\theta = \{a_i, w_i, b_i\}_{i=1}^m$，$(a_i,\ w_i,\ b_i) \in \mathrm{R} \times \mathrm{R}^d \times \mathrm{R} \coloneqq \hat{\Omega}$，表示模型参数空间；$\sigma : \mathrm{R} \to \mathrm{R}$ 是激活函数，m 表示隐藏神经元的数量。请注意，在本文的应用中，我们使用 $w_i \in \mathrm{R}^{d+1}$，以将时间作为一个维度纳入其中。式（6）所示隐藏层的无限宽度极限可表示为函数 $f_\rho(x) : D \to \mathrm{R}$；对于有界域 $D \subset \mathrm{R}^d$，该函数为

$$f_\rho(x) = \int_{\hat{\Omega}} a\sigma(w^T x + b)\rho(da \otimes dw \otimes db) = E_{(a,w,b)\sim\rho}\left[a\sigma(w^T x + b)\right]. \tag{7}$$

这里，概率测度 $\rho \in \hat{P}(\hat{\Omega})$ 定义在测度空间 $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ 上，其中 $\hat{\mathcal{F}}$ 是 $\hat{\Omega}$ 上的 Borel σ-代数。

在 E 等人（2021）中，一般激活函数 σ 的 Barron 范数定义为

$$\|f\|_{\mathcal{B}_q} \coloneqq \inf_{\rho} \left(E_\rho\left[|a|^q(\|w\|_1 + |b| + 1)^q\right]\right)^{1/q},$$

其下确界是在满足 $f = f_\rho$ 的 ρ 上取的，其中 $q \in [1, \infty]$。此外，Barron 空间定义为满足 $\|f\|_{\mathcal{B}_q} < \infty$ 的、具有式 (7) 所示形式的连续函数集合。关于 Barron 空间的更全面介绍，可参见例如 E 等人（2019）；E 等人（2021）；E、Wojtowytsch（2021）；E、Wojtowytsch（2022b）以及 E、Wojtowytsch（2022a）。
Delay-SDE-net（在第 3 节中介绍）的理论模型误差界将在第 4 节中推导。该推导结果仅适用于 Barron函数，参见定理 2 及其相应证明。这是因为 Barron 空间中的函数具有一项实用性质：其双层神经网络误差界可表示为神经网络中神经元数量 m 的函数（反比关系）（参见式 (6)）。请注意，本文使用具有 $q = 2$ 的 Barron 空间，并为方便起见记作 $\mathcal{B} \coloneqq \mathcal{B}_2$。在陈述一个给出 Barron函数误差界的定理之前，我们需要以下定义。

定义 1. 给定一个激活函数 σ，定义 $\gamma(\cdot)$函数为

$$\gamma(\sigma) = \gamma_0(\sigma) + \inf_{x\in\mathbb{R}} u(x),$$

其中

$$\gamma_0(\sigma) = \int_{\mathbb{R}} |\sigma''(x)|(|x| + 1)dx \quad \textit{and} \quad u(x) = |\sigma(x)| + (|x| + 2)|\sigma'(x)|.$$

定理 1（Li 等人（2020b），定理 4）。对于任意 $f \in \mathcal{B}$ 和 $m \in \mathbb{N}$，存在一个双层神经网络 $f_m(\cdot;\theta)$，其宽度有限，为 m，且激活函数为 σ，使得

$$\|f_m(x;\theta) - f(x)\|_{L_2} \le \frac{\sqrt{3C_\sigma}\|f\|_{\mathcal{B}}}{\sqrt{m}},$$

其中 $C_\sigma = (\gamma(\sigma) + \min\{|\sigma'(+\infty)|, |\sigma'(-\infty)|\} + |\sigma(0)|)^2$.

请注意，定理 1 省略了 Li 等人（2020b）定理 4 中陈述的一项已证明结果。另请参阅 Li 等人（2020b）附录 C.2 中该定理的证明。

3 Delay-SDE-net

Delay-SDE-net 旨在捕捉神经网络模型输入之间的时间滞后依赖关系，并通过偶然不确定性和认知不确定性的概念来描述建模中可能出现的误差。
原始的 SDE-net（Kong 等人，2020年）使用一个先前的离散化步骤来估计下一步。该方法忽略了数据中可能存在的记忆，这意味着信息可能会丢失。具有多个时滞的随机时滞微分方程（见第 2.3 节）提供了一个数学框架，可将原始 SDE-net 推广为我们所称的 Delay-SDE-net。也就是说，Delay-SDE-net 可基于 p 个先前的离散化步骤进行估计，从而加入有关趋势或快速变化的信息，使模型更能理解当前状况。
Delay-SDE-net 利用滞后时间值信息的方式类似于一类（线性）ARMA 模型。不过，Delay-SDE-net 还具有一项优势：它使用神经网络对漂移和扩散进行建模，能够表示非线性关系，从而扩展了模型的假设空间 $\mathcal{H}$。
此外，正如我们将在第 3.2 节看到的，Delay-SDE-net 使用两个神经网络对扩散系数进行建模，一个用于偶然不确定性，另一个用于认知不确定性，因此能够区分这两种不确定性。这些神经网络的训练方式使 Delay-SDE-net 成为不确定性量化领域单一确定性方法的一员，这意味着无需等待一系列模拟完成，即可立即获得不确定性估计。

3.1 Delay-SDE-net 的理论设定

按照 Kong 等人（2020年）的定义，原始 SDE-net 通过如下形式的伊藤扩散对神经网络模型施加约束

$$dX(t) = f(t, X(t))dt + g(t, X(t))dB(t),$$

其中 $B(t) = \{B(t)\}_{t\geq 0}$ 是一维布朗运动过程（例如参见 Øksendal，2003年）。本节中，我们定义 SDE-net 的一个推广版本。该扩展的建模框架称为 Delay-SDE-net，因为神经网络模型受到随机时滞微分方程的约束（参见第 2.3 节）。Delay-SDE-net 将原始 SDE-net 推广至多维动力系统，以考虑一个或多个滞后项之间的自回归依赖关系。
定义 Delay-SDE-net 时，我们假设给定数据由式（3）所示的随机时滞微分方程生成。为使建模框架保持合理的简洁性，我们对该随机时滞微分方程施加一些约束。也就是说，我们假设给定数据由一个随机过程 $X(t)$ 生成，该过程由随机时滞微分方程的解给出

$$X(t) = \begin{cases} \eta(0) + \int_0^t f(s, \Pi(X_s))ds + \int_0^t g(s, \Pi(\eta(0)))dW(s), & t \geq 0, \\ \eta(t), & -\tau \leq t < 0, \end{cases} \tag{8}$$

其中 $\Pi\colon C \to \mathrm{R}^{dp}$ 如式（1）所定义。注意，扩散项与状态无关，因为它仅取决于初始路径 $\eta\colon \Omega \to C$。作出这一假设是为了简化模型的离散化方案。也就是说，对于式（5）中的所有 $i = 1, \ldots, d$，我们都有 $F^i = 0$，这意味着 Milstein格式等价于 Euler-Maruyama格式（Kloeden、Platen，1992年）。

下面，我们定义一个称为 Delay-SDE-net 的模型。假设给定数据由式（8）所示形式的随机时滞微分方程生成，那么构建一个模型是合理的

$$X^{(m)}(t) = \begin{cases} \eta(0) + \int_0^t f_m(s, \Pi(X_s^{(m)}))ds + \int_0^t g_m(s, \Pi(\eta(0)))dW(s), & t \geq 0, \\ \eta(t), & -\tau \leq t < 0, \end{cases} \tag{9}$$

其初始路径 $\eta\colon \Omega \to C$ 与式（8）中的 η 相同，用于描述数据所代表的动力系统的演化。在式（9）中，f_m 由一个含 m 个神经元的神经网络给出。此外，受第 2.2 节讨论的启发，我们定义

$$g_m \coloneqq g_{a,m} + g_{e,m}, \tag{10}$$

其中 $g_{a,\ m}$ 和 $g_{e,\ m}$ 分别由两个相互独立的 m 神经元神经网络给出，分别表示偶然不确定性和认知不确定性。

请注意，为简化记号，我们假设神经网络 f_m、g_a 和 g_e、m 使用相同数量的神经元 m 进行训练。在应用中，可以毫无问题地省略这一假设。此外，为了分析 Delay-SDE-net 的理论误差界，我们将在第 4 节中将 f_m、g_a 和 g_e、m 视为双层神经网络。在第 5 节的实证分析和应用中，以及一般应用中，神经

网络不受限于仅有两层。请注意，为简单起见，本文第 5 节的所有分析均使用双层神经网络。最后，正如第 2.4 节所述，需要注意的是，方程（9）中的模型是连续时间模型，但是，f_m 和 f_ρ 是不同的函数，因为 f_ρ 是 $m \to \infty$ 的 f_m 极限。

式（9）中 Delay-SDE-net 的连续时间版本采用其离散时间近似进行估计

$$\Delta X^{j,\pi,(m)}(t_k) = f_m^j(t_{k-1}, \Pi(X_{t_{k-1}}^{\pi,(m)}))\Delta t + g_m^j(t_{k-1}, \Pi(\eta^\pi(0)))\Delta W^j(t), \quad t \geq 0, \tag{11}$$

对于 $j = 1, \ldots\ldots, d$，其中 $g_m^j := g_{a,\,m}^j + g_{e,\,m}^j$。请注意，这一离散时间近似基于式（5）中的 Milstein格式，并且扩散项的拆分最初定义于式（10）中。在这种情况下，例如，观测过程 η^π取为 η 的分段线性近似，并使用式（4）中定义的划分。还请注意，为简单起见，我们假设对于 d 维过程 $X(t)$ 的每个元素，都存在一个独立驱动过程，这意味着 $W(t)$ 取值于 R^d。因此，g和 g_m可以分别取为从 $g \in$到 R^d 以及从 $g_m \in$到 R^d 的向量函数（应用向量记号下的 Hadamard积时）。然后，只要每个 g^j（以及 g_m^j）、$j = 1, \ldots\ldots, d$满足存在唯一性所需的条件以及离散化所需的正则性条件，式（8）（以及式（9））中随机时滞微分方程的逐分量随机部分就是 d 个相互独立的一维过程。

本节定义的 Delay-SDE-net 建模框架可采用下一节的方法进行训练，以用于各种应用。

3.2 训练算法

本节介绍一种训练 Delay-SDE-net 的方法。我们将看到，式（9）中的确定性部分和随机部分分别作为独立的神经网络进行训练，并且与相关文献中的方法相比，本文采用了若干新方法。

第 3.2.1、3.2.2 和 3.2.3 节说明如何分别训练神经网络，以实现 Delay-SDE-net 的确定性部分、f_m、随机部分 $g_{a,\,m}$ 和 $g_{e,\,m}$。请注意，如式（ 10 ）所示，随机部分被拆分为两个神经网络，因此 Delay-SDE-net 由三个神经网络组合而成。将确定性部分和随机部分分别训练，与 SDE-net（Kong 等人，2020）及其其他变体（Hayashi、Nakagawa，2022；Wang等人，2021；Wang、Yao，2021；Yang等人，2021）所采用的方法不同。采用独立神经网络的方法，使 f_m 的训练过程更加稳定，因为 $g_{a,\,m}$ 和$g_{e,\,m}$ 的训练不会影响确定性部分（Gawlikowski等人，2021）。此外，$g_{a,\,m}$ 的训练也会更加稳定，因为它是在经过最终优化的 f_m 所产生的固定残差上进行训练的（见第 3.2.2 节）。Raghu 等人（2019）指出了这一点：当使用单个神经网络同时估计不确定性和预测结果时，预测任务会使不确定性估计产生偏差。Hsu等人（2020）、Oberdiek等人（2018）、Lee、AlRegib（2020）以及 Ramalho、Miranda（2020）也曾分别使用两个独立神经网络进行预测和不确定性估计，但仅限于分类任务。

第 3.2.1—3.2.3 节分别介绍的 f_m, $g_{a,\,m}$ 和 $g_{e,\,m}$ 的训练方法，适用于能够进行提前一天预测（即向前预测一个时间步）的 Delay-SDE-net。在第 3.2.4 节中，我们介绍一种训练方法，用于训练能够向前预测 N 个时间步的 Delay-SDE-net。训练算法见算法 1，相应示意图见图 2。该算法的代码可在 https://github.com/alimid/Delay-SDE-net 上找到。

3.2.1 漂移网络

漂移网络 f_m 以时间和先前状态为输入，通过确定性方式进行训练。这相当于一种延迟 ODE-net，是 Chen 等人（2018）提出的 ODE-net 的扩展版本。Yuan 等人（2022）近期发表的研究介绍了一种延迟 ODE-net，其中历史信息作为初始状态加入模型；这一方法不同于本文提出的方法，后者会持续加入新的历史信息。

首先请注意，训练时使用的是公式 (11) 中的离散化延迟随机微分方程网络。为了训练 f_m，我们专注于预测期望值，并忽略随机项。也就是说，通过最小化期望预测误差来求得最优漂移网络

$$\min_{\theta_{f_m}} E_{\mathcal{D}_0}\left[\|\hat{\boldsymbol{x}}_{t_k} - \boldsymbol{x}_{t_k}\|_2\right],$$

其中 $\mathcal{D}_0 :=$ $\{\boldsymbol{x}_{t_k} \in \mathrm{R}^d :\ t_k \in [0, T]\}$ 是训练数据集，θ_{f_m} 表示漂移网络的参数，并且

$$\hat{\boldsymbol{x}}_{t_k} = \boldsymbol{x}_{t_{k-1}} + f_m(t_{k-1}, \boldsymbol{x}_{t_{k-p}}, \ldots, \boldsymbol{x}_{t_{k-1}})\Delta t. \tag{12}$$

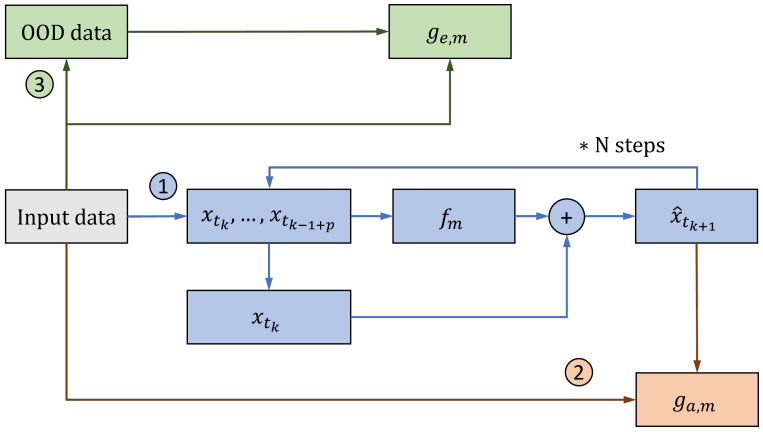


图 2：Delay-SDE-net 训练算法的示意图。首先训练漂移项 f_m。其次，使用漂移项的残差来训练偶然不确定性扩散项$g_{a,\,m}$。第三，认知不确定性扩散项 $g_{e,\,m}$与其他两项分开训练，仅使用输入数据和生成的分布外数据。

受公式（11）的启发，我们使用优化后的漂移网络给出的预测$\hat{\boldsymbol{x}}_{t_k}^{\text{optimal}}$ 来训练偶然不确定性扩散网络$g_{a,\,m}$。也就是说，我们使用这一关系

$$X^{\pi,(m)}(t_k) - g_m(t_{k-1},\cdot)\circ\Delta W(t) \coloneqq \hat{\boldsymbol{x}}_{t_k}^{\text{optimal}} = \boldsymbol{x}_{t_{k-1}} + f_m(t_{k-1}, \boldsymbol{x}_{t_{k-p}}, \ldots, \boldsymbol{x}_{t_{k-1}})\Delta t, \tag{13}$$

其中 $\circ$表示 Hadamard积，并且我们假设 $\Delta W \sim N(0,\ 1)^d$。参见下一节。

3.2.2 偶然扩散网络

漂移项的迭代结束后，便开始训练 Delay-SDE-net 的扩散项。如前所述，我们分别训练偶然不确定性扩散网络 $g_{a,\,m}$和认知不确定性扩散网络 $g_{e,\,m}$。偶然不确定性扩散网络应捕捉数据中的模式，以表示随机噪声的标准差，这对应于漂移网络的偶然不确定性度量。

由于 Delay-SDE-net 由（标准）布朗运动过程驱动（见公式 (13)），与偶然不确定性对应的随机项分布为

$$\int_0^T g_{a,m}(t,\cdot)dW(t) \sim N\left(0, \int_0^T g_{a,m}^2(t,\cdot)dt\right),$$

例如，参见如 Benth（2004）所示，这意味着向前推算一个时间步的近似方差为

$$\int_{t_k}^{t_{k+1}} g_{a,m}^2(t,\cdot)dt \simeq g_{a,m}^2(t_k,\cdot)\Delta t, \tag{14}$$

其中，$g_{a,\,m}(t_k,\ \cdot)$ 是公式（11）中的离散化的偶然不确定性扩散网络。我们基于公式（13）中预测值的残差 $\boldsymbol{e}_{t_k} = \hat{\boldsymbol{x}}_{t_k}^{\text{optimal}} - \boldsymbol{x}_{t_k}$ 训练 $g_{a,\,m}$，并按如下方式进行优化：

$$\min_{\theta_{g_{a,m}}} E_{\mathcal{D}_0}\left[\|g_{a,m}^2(t_k,\cdot)\Delta t - \boldsymbol{e}_{t_k}^2\|_2\right],$$

其中，$\theta_{g_{a,\,m}}$ 表示偶然不确定性扩散网络参数。请注意，Δt 是明确给定的，并且在训练期间保持不变。通过这种方法，经过优化的偶然不确定性扩散网络 $g_{a,\,m}$ 对应于公式（13）中模型预测的标准差。

允许方差随时间变化的模型适用于处理异方差数据集，也就是方差并非常数的数据集。典型例子包括天气建模中的季节性行为（Eggen 等人，2022年），以及股票收益率建模中的日内波动（Andersen 等人，2019年）。请注意，预测残差既包含模型误差，也包含内在随机性，这意味着 $g_{a,\,m}$ 经过优化后会捕捉这两种不确定性。如第 2.2 节所述，$g_{a,\,m}$ 无法区分真实函数中哪些部分属于假设空间 $\mathcal{H}$，哪些部分不属于（见图 1）。因此，$g_{a,\,m}$ 必须视为在给定模型$g_{a,\,m} = E[g_a|f_m]$的条件下对偶然不确定性噪声进行建模。尽管文献中很少提及这一点，但所有基于模型残差优化扩散的模型都包含模型不确定性，例如 Eggen 等人（2022年）中的向量自回归（VAR）模型。

3.2.3 认知不确定性扩散网络

训练 Delay-SDE-net 的最后阶段是训练神经网络 $g_{e,\ m}$，它对应于认知不确定性。在第 2.2 节中，认知不确定性被定义为由两种不确定性构成：近似不确定性和模型不确定性。如第 3.2.2 节所述，由于偶然不确定性扩散网络 $g_{a,\ m}$的训练方法，模型中已经包含了模型不确定性。也就是说，训练后的认知不确定性扩散网络应仅反映近似不确定性。

近似误差源于整个假设空间中训练数据的可用量有限，这意味着真实分布未得到充分覆盖。根据第 2.2 节，理想情况下，认知不确定性扩散网络应能够区分来自其训练所依据的假设空间的数据 $\mathcal{H}_0$ 与此前未见过的数据 $\not\subset \mathcal{H}_0$ （见图 1）。我们采用与训练 SDE-net 类似的方法训练 $g_{e,\ m}$ （Kong 等人，2020年）。也就是说，我们使用一个区分分布内数据和分布外数据的分类器，并对以下目标进行优化

$$\min_{\theta_{g_{e,m}}} E_{\mathcal{D}_0}\left[g_{e,m}^{\text{prob}}(t_k,\cdot)\right] + \max_{\theta_{g_{e,m}}} E_{\tilde{\mathcal{D}}_0}\left[g_{e,m}^{\text{prob}}(t_k,\cdot)\right], \tag{15}$$

其中 $\tilde{\mathcal{D}}_0$ 表示分布外数据集。请注意，式（15）中优化的函数是 $g_{e}^{\text{prob}}{}_{,\ m} \in [0,1]^d$，而不是式（11）中的 $g_{e,\ m} \in \mathrm{R}^d$。这是因为，在分布内数据和分布外数据这两个类别之间执行概率分类任务时，训练得到的网络 $g_{e}^{\text{prob}}{}_{,\ m}$ 表示某个观测属于分布外数据的概率。式（10）中的认知不确定性标准差 $g_{e,\ m}$ 最终近似为 $g_{e,\ m} = \sigma_e g_{e\ m,}^{\text{prob}}$, $\sigma_e \in \mathrm{R}$, 其中，适当的常数 σ_e表示认知不确定性的最大值，可根据所生成的分布外数据的相关知识设定，也可作为调节参数计算。在这两种情况下， σ_e 的幅度都应控制认知不确定性的幅度，并与偶然不确定性的幅度相协调。当 σ_e 为调节参数时，我们最小化

$$\min_{\sigma_e} E_{\mathcal{D}}\left[\|(\sigma_e g_{e,m}^{\text{prob}}(t_k,\cdot) + g_{a,m}(t_k,\cdot)\sqrt{\Delta t})^2 - \boldsymbol{e}_{t_k}^2\|_2\right],$$

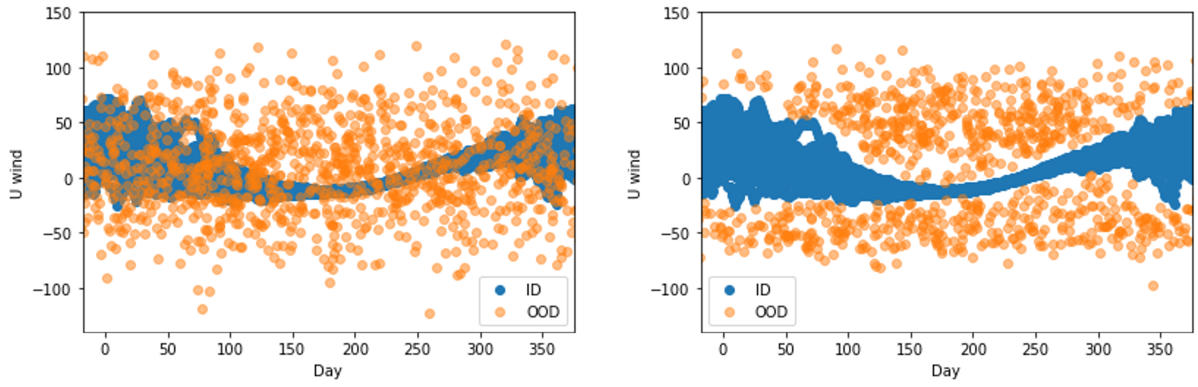
其中 $g_{a,m}(t_k,\cdot)\sqrt{\Delta t}$表示训练得到的偶然不确定性标准差，而 $\mathcal{D}$ 是一个验证数据集 t.

如果可以获取真实的分布外数据，就可以用它来训练 $g_{e\ m,}^{\text{prob}}$。然而，在许多应用中，分布外数据集并不常见，这意味着必须人工生成分布外数据。训练 SDE-net 时，生成分布外数据的方法是从训练数据集中随机抽取数据点，并添加高斯˜噪声。也就是说，SDE-net 使用一个分布外数据集 $\mathcal{D}_0$ 进行训练，其元素为 $\tilde{\boldsymbol{x}}_{t_k} = \boldsymbol{x}_{t_k} + \boldsymbol{\epsilon}_{t_k}$，其中 $\boldsymbol{x}_{t_k} \in \mathcal{D}_0$ 且 $\boldsymbol{\epsilon}_{t_k} \sim N(0,1)^d$。如图3a所示，该方法无法保证生成的都是分布外数据，因为它们可能与训练数据重叠。因此，Delay-SDE-net 改为使用 Moller 等人（2021年）提出的软布朗偏移法生成的分布外数据进行训练。该方法从训练数据 $\boldsymbol{x}_{t_k} \sim \mathcal{D}_0$ 中随机抽取数据点，并迭代地添加高斯噪声， $\boldsymbol{e} = [e_1,\ldots,e_d] \sim N(\mu,\sigma)^d$，直到这些数据点与训练数据集中的所有数据点之间达到预先指定的最小距离 d^- 。简而言之，该方法如下:

当 $d^* < d^-$ **时执行**
$\tilde{\boldsymbol{x}}_{t_k} = \tilde{\boldsymbol{x}}_{t_k} + d^+\boldsymbol{e}$，**直到**
结束

其中， d^* 是 $\mathcal{D}_0$中 $\tilde{\boldsymbol{x}}_{t_k}$ 到最近数据点的距离， d^+ 是为˜添加的噪声预先指定的偏移量长度。软布朗偏移法 $\mathcal{D}_0 = \{\tilde{\boldsymbol{x}}_{t_1},\ldots,\tilde{\boldsymbol{x}}_{t_K}\}$生成的数据保证处于给定训练数据的分布外，如图3b所示。

请注意，用于生成分布外数据的方法应针对具体应用场景进行调整，以确保涵盖特定应用中可能出现的罕见结果。在软布朗偏移法中，例如，可以调整分布内数据与分布外数据之间的最小距离 d^*（由于维度灾难，该距离必须随数据集维度数呈指数增长），也可以通过改变偏移量 d^+ 来调整分布外数据的分散程度。此外，还可以指定分布外数据应偏离分布内数据的哪些属性。在本研究中，我们选择生成涵盖两种特定属性的分布外数据。首先，我们生成反映时间序列局部变异性异常的分布外数据。也就是说，软布朗偏移法对滞后项之间的差异敏感，因为分布外数据集是根据 $[\tilde{\boldsymbol{x}}_{t_k},\ldots,\tilde{\boldsymbol{x}}_{t_{k+1-p}}]^T = [\boldsymbol{x}_{t_k} + d^+\boldsymbol{e}_1,\ldots,\boldsymbol{x}_{t_{k+1-p}} + d^+\boldsymbol{e}_p]^T$ 构建的，其中 $\boldsymbol{e}_i \sim N(\mu,\ \sigma)^d$对于$i = 1,\ldots p$。其次，我们生成反映给定时间数据点幅度异常的分布外数据。也就是说，分布外数据集是根据 $\tilde{\boldsymbol{x}}_{t_k} = \boldsymbol{x}_{t_k} + d^+\boldsymbol{e}$ 构建的。



(a) (b)

图3：使用（a）高斯随机噪声（b）软布朗偏移法为平流层风生成的分布外数据。平流层风数据的描述见第5.2节。蓝色数据点是30年来分散分布的U风数据，黄色数据点则是人工生成的2000个分布外数据点。

3.2.4 TheN步超前 Delay-SDE-net

只需对第 3.2.1 至 3.2.3 节所述 Delay-SDE-net 的训练方法稍作修改，即可实现一个给定初始路径后预测未来 N 步的模型。

N步超前 Delay-SDE-net 的漂移网络通过以下优化进行训练

$$\min_{\theta_{fm}} E_{\mathcal{D}_0}\left[\|\hat{\boldsymbol{x}}_{t_{k+N}} - \boldsymbol{x}_{t_{k+N}}\|_2\right],$$

其中，$\hat{\boldsymbol{x}}_{t_{k+N}}$采用与第 3.2.1 节公式（12）类似的方法，从 $\boldsymbol{x}_{t_k}$ 中预测得到。具体方法如算法 1 所示。同样按照第 3.2.1 节公式 13，我们从训练好的网络 f_m 得到最优预测$\hat{\boldsymbol{x}}^{\text{optimal}}_{t_{k+N}}$，用于训练偶然不确定性扩散网络$g_{a,\ m}$。

与第 3.2.2 节一样，偶然不确定性扩散网络使用残差$\boldsymbol{e}_{t_{k+N}} = \hat{\boldsymbol{x}}^{\text{optimal}}_{t_{k+N}} - \boldsymbol{x}_{t_{k+N}}$进行训练。然而，向前预测 N 步的近似方差与公式（14）不同。也就是说，向前推移 N 步的近似（随时间变化的）方差为

$$\begin{aligned}\int_{t_k}^{t_{k+N}} g^2_{a,m}(t, \Pi(\eta(t_k)))dt &\simeq \sum_{i=0}^{N-1} g^2_{a,m}(t_{k+i}, \boldsymbol{x}_{t_k}, \ldots, \boldsymbol{x}_{t_{k+1-p}})\Delta t \\ &:= V(t_k, \boldsymbol{x}_{t_k}, \ldots, \boldsymbol{x}_{t_{k+1-p}})\Delta t,\end{aligned}$$

其中，V取为从时间 t_k 到 t_{k+N}的点方差之和。请记住，根据第3.1节中 Delay-SDE-net 的定义，模型扩散与状态无关，因此只取决于初始（观测到的）路径。此外，请注意，除观测到的点值 $\boldsymbol{x}_{t_k}$、……、$\boldsymbol{x}_{t_{k+1-p}}$外，$V$仅取决于第一个时间步 t_k。在估计中这一点成立，因为只要 N 和 Δt 在所有训练迭代次数中保持不变，就可以将 t_k 视为给定时间向量 $\boldsymbol{\tau}^T = [t_k, t_{k+1}, \ldots, t_{k+N}]$的充分估计量。这是因为，如果两个观测到的第一个时间步相等 t_a=t_b，那么在这种情况下，两个时间向量也相等 $\boldsymbol{\tau}_a^T = \boldsymbol{\tau}_b^T$，这意味着 t_k 满足 Casella、Berger（2002）定义 6.2.1 中对给定 $\boldsymbol{\tau}^T$的充分估计量所提出的要求。因此，N 步超前方差是在优化过程中训练的

$$\min_{\theta_{g_{a,m}}} E_{\mathcal{D}_0}\left[\|V(t_k, \cdot)\Delta t - (\boldsymbol{e}_{t_{k+N}})^2\|_2\right].$$

最后，N步前向延迟随机微分方程网络的认知不确定性扩散网络分类器 $g^{\text{prob}}_{e\ m,}$ 的训练方式与第 3.2.3 节的式（15）完全相同，仅使用训练中的初始时间 t_k。也就是说，我们假设

$$g^{\text{prob}}_{e,m}(t_{k+N}) = g^{\text{prob}}_{e,m}(t_k),$$

因为分布内和分布外训练数据集的信息仅限于初始时间 t_k，这意味着我们假设 $g^{\text{prob}}_{e\ m,}$ 在从 t_k 到 t_{k+N}的预测步内保持不变。最终的认知不确定性扩散网络 $g_{e,\ m} = \sigma_e^{(N)} g^{\text{prob}}_{e\ m,}$，$\sigma_e^{(N)} \in \mathrm{R}$,

N步前向延迟随机微分方程网络可将 $\sigma_e^{(N)}$作为调节参数进行训练,

$$\min_{\sigma_e^{(N)}} E_{\mathcal{D}}\left[\|(\sigma_e^{(N)} g_{e,m}^{\text{prob}}(t_k,\cdot)+\sqrt{V(t_k,\cdot)\Delta t})^2-\boldsymbol{e}_{t_{k+N}}^2\|_2\right], \tag{16}$$

其中， $\sqrt{V(t_k,\ \cdot)}\Delta t$ 表示训练得到的偶然不确定性标准差。

算法 1 Delay-SDE-net。 p：滞后项数量， N：向前预测的时间步数， π_s：小批量中的采样时间点， Δt：时间步长度

初始化 f_m、 $g_{a,\ m}$ 和$g_{e,\ m}$。
进行# f_m 次训练迭代，**直到**
  采样小批量 $\{[\boldsymbol{x}_{t_k},\ldots,\boldsymbol{x}_{t\ \ \ \ k+1-p}]^T\}_{k=0}^n$, $\boldsymbol{x}_{t\,k}\in\mathrm{R}^d$ 对于 $t_k\in\pi_s$
  对于所有 t **执行**
    $\hat{\boldsymbol{x}}_{t_{k+1}}=\boldsymbol{x}_{t_k}+f_m(\boldsymbol{x}_{t_k},\ldots,\boldsymbol{x}_{t_{k+1-p}}\ t_k)\Delta t$
    $\hat{\boldsymbol{x}}_{t_{k+2}}=\hat{\boldsymbol{x}}_{t_{k+1}}+f_m(\hat{\boldsymbol{x}}_{t\ k+1}\ldots,\boldsymbol{x}_{t\ k+2-p}\ t_{k+1})\Delta t$
    $\vdots$
    $\hat{\boldsymbol{x}}_{t_{k+N}}=\hat{\boldsymbol{x}}_{t_{k+N-1}}+f_m(\hat{\boldsymbol{x}}_{t\ k+N-1},\ldots,\hat{\boldsymbol{x}}_{t_{k+N-p}}\ t_{k+N-1})\Delta t$
  结束循环
  更新 f：
  $\nabla_{f_m}L((\boldsymbol{x}_{t_{k+N}},\ \hat{\boldsymbol{x}}_{t_{k+N}})_{\forall t_k\in\pi_s})$
结束循环

for # $g_{a,m}$ training iterations **do**
  Update $g_{a,m}$ by:
  $\nabla_{g_{a,m}}L((g_{a,m}^2(\boldsymbol{x}_{t_k},\ldots,\boldsymbol{x}_{t_{k+1-p}},t_k)\Delta t,e_{t_{k+N}}^2)_{\forall t_k\in\pi_s}),e_{t_{k+N}}=\boldsymbol{x}_{t_{k+N}}-\hat{\boldsymbol{x}}_{t_{k+N}}$
end for

对于 # $g_{e,\ m}$ 训练迭代次数**执行**
  采样一批分布外数据 $\{[\tilde{\boldsymbol{x}}_{t_k},\ldots,\tilde{\boldsymbol{x}}_{t\ \ \ \ \ \ k+1-p}]^T\}_{k=0}^n$, $\tilde{\boldsymbol{x}}_{t\,k}\in\mathrm{R}^d$ 对于 $t_k\in\tilde{\pi}_s$
  通过以下方式更新$g_{e,\ m}$：
  $\nabla_{g_{e,m}}g_{e,\ m}\ (\tilde{\boldsymbol{x}}_{t_k},\ldots,\tilde{\boldsymbol{x}}_{t\ k+1-p},t_k)\ _{\forall t_k\in\tilde{\pi}_s}-\nabla_{g_{e,m}}g_{e,m}(\boldsymbol{x}_{t_k},\ldots,\boldsymbol{x}_{t\ k+1-p},t_k)\ _{\forall t_k\in\pi_s}$
结束循环

4 双层 Delay-SDE-net 的理论误差

在第3.1节中，我们介绍了 Delay-SDE-net（式（9）），并分别在式（8）和式（11）中给出了它的离散化版本和真实世界对应模型。本节旨在推导双层 Delay-SDE-net 的理论误差和收敛结果。

由于 Delay-SDE-net 的训练采用其离散时间版本，因此必须通过两个近似步骤来推导模型的误差结果。也就是说，我们首先使用连续时间 Delay-SDE-net 近似假设的真实世界随机时滞微分方程模型，然后使用其离散化版本近似连续时间 Delay-SDE-net 模型。下面给出关于真实世界随机时滞微分方程和 Delay-SDE-net 模型系数的必要假设。

假设 1. 给定式 (8) 中的随机时滞微分方程，且其解唯一，我们假设 $f^j,g^j\in\mathcal{B}$, $j=1,\ldots,d$ 是某个有界域 $D\subset T\times\mathbb{R}^{dp}$ 上的有界函数。
假设 2. 给定如式 (9) 所示的连续时间延迟随机微分方程网络，假设 $\eta\in C$ 具有有界变差，且是$(\gamma/2)$-赫尔德连续的，其中 $0<\gamma\leq 1$。此外，假设 $f_m^j,g_m^j\in C^{1,2}(T\times\mathbb{R}^{dp},\mathbb{R})$、 $j=1$、.....、 d 均为有界函数，且其一阶和二阶空间导数有界。

我们采用 Hu 等人（2001）和 Hu 等（2004）推导并提出的公式（3）中动力学方程 Milstein格式的收敛结果，来表示公式（11）中的离散化延迟随机微分方程网络与公式（9）中的连续时间延迟随机微分方程网络之间的收敛率。根据 Hu 等（2004）的定理 5.2 和第 5.3 节中的评注，以及假设 2，我们有：若

$$\sup_{-\tau\leq t\leq 0}\|\eta^\pi(t)-\eta(t)\|_2\leq C_\eta\Delta t^\gamma,$$

其中 C_η 是一个正 常数，则存在一个常数 $C_T > 0$，它取决于 T，且与 Δt无关，使得

$$\sup_{-\tau \le t \le T} \|X^{\pi,(m)}(t) - X^{(m)}(t)\|_{L_2}^2 \le C_T \Delta t^{2\gamma}. \tag{17}$$

正如我们将在本节末尾看到的（推论1），与现实世界模型 X 相比，式（17）中的收敛结果是推导 $X^{\pi,(m)}$ 最终误差界的关键要素。首先，我们需要推导连续时间延迟随机微分方程网络 $X^{(m)}$ 的误差界。下面的定理给出了 $X^{(m)}$ 相对于真实世界随机时滞微分方程 X 的误差界。我们将看到，第2.4节中介绍的 Barron 空间理论，以及第 3.2.3 节（向前预测 N 步时为第 3.2.4 节）训练算法所提供的认知不确定性上界 $g_{e,m}$，将用于得到最终结果。最后请注意，与式（8）中假设的现实世界模型（其中扩散项仅取决于初始路径）相比，定理 2 的结果推广到了状态依赖扩散项。

定理 2. 给定公式（3）中的随机时滞微分方程（*SDDE*）及其对应的（连续时间）双层 *Delay-SDE-net*，在假设 *1* 和 *2* 下，该模型的误差界为

$$\|X^{(m)}(t) - X(t)\|_{L_2} \le C_m,$$

其中

$$C_m \coloneqq \left(4T(C_{g_e} + 2C_B\epsilon_F + \frac{3C_\sigma}{m}(2C_g + TC_f))e^{2pC_L(4+2T)t} \right)^{1/2}.$$

这里，$m \in \mathbb{N}_+$ 是双层神经网络 f_m 和g_a中的神经元数量m （或者说，当两个神经网络中的神经元数量不同时，取较小的神经元数量），C_g、$C_f > 0$分别是 $\|g\|_{\mathcal{B}}^2$和 $\|f\|_{\mathcal{B}}^2$的上界，$C_{g_e} > 0$ 是 $\|g_{e_m}\|_{L_2}^2$ 的上界（见第 *3.2.3–3.2.4* 节），$C_L > 0$ 是f 和 g的 ***Lipschitz***常数，$C_B > 0$ 是函数 $(g_m - g)^2$ 和 $(f_m - f)^2$, 的最大界，ϵ_F 是表示稀有事件的小概率。注意，如果公式（*3*）中的随机时滞微分方程定义在有界域上，我们令 $\epsilon_F = 0$ 。最后，我们有 $C_\sigma = (\gamma(\sigma) + \min\{|\sigma'(+\infty)|, |\sigma'(-\infty)|\} + |\sigma(0)|)^2$,，其中 σ 表示所选的激活函数（ f_m、 g_a、m 以及g_e、m可以各不相同），而 $\gamma(\cdot)$ 见定义 *1*。

证明。 首先，容易看出

$$\sup_{-\tau \le t \le 0} \|X^{(m)}(t) - X(t)\|_{L_2} = \sup_{-\tau \le t \le 0} \|\eta(t) - \eta(t)\|_2 = 0. \tag{18}$$

利用恒等式 $|u+v|^q \le 2^{q-1}(|u|^q + |v|^q)$、Itô等距和柯西-施瓦茨不等式，我们得到

$$\begin{aligned}
\|X^{(m)}(t) - X(t)\|_{L_2}^2 &= E\left[\left(\int_0^t (g_m - g)\, dW(s) + \int_0^t (f_m - f)\, ds \right)^2 \right] \\
&\le 2E\left[\left(\int_0^t (g_m - g)\, dW(s) \right)^2 + \left(\int_0^t (f_m - f)\, ds \right)^2 \right] \\
&\le 2E\left[\int_0^t (g_m - g)^2\, ds \right] + 2E\left[\int_0^t (f_m - f)^2\, ds \int_0^t ds \right] \\
&\le 4\int_0^t \|g_{e,m}\|_{L_2}^2 ds + 4\int_0^t \|g_{a,m} - g\|_{L_2}^2 ds + 2a\int_0^t \|f_m - f\|_{L_2}^2 ds
\end{aligned} \tag{19}$$

其中在最后一步中，我们使用了 $g_m = g_{a,m} + g_{e,m}$ 这一事实、恒等式 $|u+v|^q \le 2^{q-1}(|u|^q + |v|^q)$ 以及单调收敛。注意，为简洁起见，我们使用 $f_m \coloneqq f_m(s, \Pi(X_s^{(m)}))$ 和 $f \coloneqq f(s, \Pi(X_s))$，对于 g_m、 g、$g_{a,m}$ 和 $g_{e,m}$ 也采用类似记法。

现在，令 h_m 和 h 分别表示 $g_{a,m}$、 f_m 和 g、 f。那么我们有

$$\begin{aligned}
&\|h_m(t, \Pi(X_t^{(m)})) - h(t, \Pi(X_t))\|_{L_2} \le \\
&\|h_m(t, \Pi(X_t^{(m)})) - h(t, \Pi(X_t^{(m)}))\|_{L_2} + \|h(t, \Pi(X_t^{(m)})) - h(t, \Pi(X_t))\|_{L_2},
\end{aligned} \tag{20}$$

根据闵可夫斯基不等式。

在　接下来，将分别考察式（20）右侧的两个项　y.

1. 注意，$X_t^{(m)} = [X_t^{1,(m)}, \ldots, X_t^{d,(m)}]^T \in \mathbb{R}^d$，其中 $X^{j,}{}_t{}^{(m)} \in C$ 对于 $j = 1, \ldots, d$。也就是说，对于每个 $t \in [0,T]$，$X_t^{(m)} \in C(\Omega \times J; \mathbb{R}^d)$。此外，$\Pi(X_t^{(m)}) = [X^{(m)}(t+u_1) \ldots, X^{(m)}(t+u_p)]^T : \Omega \to \mathbb{R}^{dp}$，其中 $u_1, \ldots, u_p \in [-\tau, 0], \tau > 0$ 固定，且对于 $i = 1, \ldots, p$，$X^{(m)}(t+u_i) : \Omega \to \mathbb{R}^d$。现在，将 $\Pi(X_t^{(m)})$ 的多元累积分布函数定义为 $F : \mathbb{R}^{dp} \to [0,1]$。此外，将 $\Pi(X_t^{(m)})$ 的值所构成的可测集，即 $\mathbb{R}^{dp}$，划分为一个有界域 $D_r := \{x \in \mathbb{R}^{dp} : |x| < r\}$（其中某个 $r > 0$）及其补集 D_r^c，使得 D_r^c 上的概率很小。于是，对于式（20）右侧的第一项，我们得到以下结果。

$$\begin{aligned}
\|h_m(t, \Pi(X_t^{(m)})) - h(t, \Pi(X_t^{(m)}))\|_{L_2}^2 &= E\left[(h_m(t,x) - h(t,x))^2\right] \\
&= \int_{\mathbb{R}^{dp}} (h_m(t,x) - h(t,x))^2 dF(x) \\
&= \int_{D_r} (h_m(t,x) - h(t,x))^2 dF(x) + \\
&\qquad \int_{D_r^c} (h_m(t,x) - h(t,x))^2 dF(x).
\end{aligned}$$

根据假设 1 和 2，我们知道 h_m 和 h 是有界函数。利用这一点以及 D_r^c 是稀有事件域这一事实，可得

$$\int_{D_r^c} (h_m(t,x) - h(t,x))^2 dF(x) \le C_B F(D_r^c) \le C_B \epsilon_F,$$

对于所有 $t \in [0,T]$，其中 $C_B > 0$ 是一个常数，$\epsilon_F > 0$ 表示 D_r^c 上的小概率。此外，根据假设 1，我们有 $h \in \mathcal{B}$，因此由定理 1 可得

$$\int_{D_r} (h_m(t,x) - h(t,x))^2 dF(x) \le \frac{3C_\sigma \|h\|_{\mathcal{B}}^2}{m}.$$

2. 根据利普希茨连续性以及欧几里得范数的定义，投影 Π（式 (1)）和分段过程 X_t（式 (2)），式 (20) 右侧的第二项为

$$\begin{aligned}
\|h(t, \Pi(X_t^{(m)})) - h(t, \Pi(X_t))\|_{L_2}^2 &= E\left[\left(h(t, \Pi(X_t^{(m)})) - h(t, \Pi(X_t))\right)^2\right] \\
&\le C_L E\left[\|\Pi(X_t^{(m)}) - \Pi(X_t)\|_2^2\right] \\
&= C_L E\left[\sum_{i=1}^p \left(X_t^{(m)}(u_i) - X_t(u_i)\right)^2\right] \\
&= C_L \sum_{i=1}^p \|X_t^{(m)}(u_i) - X_t(u_i)\|_{L_2}^2.
\end{aligned}$$

由于范数随着 $t + u_i$ 增大而增大（见式 (2)），且 $u_i \in [-\tau, 0]$，再利用式 (18)，我们进一步得到

$$\begin{aligned}
&\|h(t, \Pi(X_t^{(m)})) - h(t, \Pi(X_t))\|_{L_2}^2 \\
&\le C_L \sum_{i=1}^p \left(\sup_{0 \le s \le -u_i} \|X_s^{(m)}(u_i) - X_s(u_i)\|_{L_2}^2 + \sup_{-u_i \le s \le t-u_i} \|X_s^{(m)}(u_i) - X_s(u_i)\|_{L_2}^2 \right) \\
&\le C_L \sum_{i=1}^p \|X^{(m)}(t) - X(t)\|_{L_2}^2 = pC_L \|X^{(m)}(t) - X(t)\|_{L_2}^2.
\end{aligned}$$

将恒等式 $|u+v|^q \le 2^{q-1}(|u|^q + |v|^q)$ 应用于式 (20)，并代入上述第 1 点和第 2 点的结果，可得

$$\|h_m(t, \Pi(X_t^{(m)})) - h(t, \Pi(X_t))\|_{L_2}^2 \le 2\left(\xi(h) + pK\|X^{(m)}(t) - X(t)\|_{L_2}^2\right), \tag{21}$$

其中，我们定义 $\xi(h) \coloneqq \frac{3C_\sigma \|h\|_{\mathcal{B}}^2}{m} + C_B \epsilon_F$。

根据第 3.2.3–3.2.4 节，我们知道如何求出认知误差 $\|g_{e,\ m}\|_{L_2}^2 \leq C_{g_e}$ 的上界。将其与公式 (21) 中的结果代入公式 (19)，可得

$$\begin{aligned}\|X^{(m)}(t) - X(t)\|_{L_2}^2 \leq & 4\int_0^T C_{g_e} ds + 8\int_0^T \xi(g) ds + 4T\int_0^T \xi(f) ds \\ & + 8pC_L \int_0^t \|X^{(m)}(s) - X(s)\|_{L_2}^2 ds + 4TpC_L \int_0^t \|X^{(m)}(s) - X(s)\|_{L_2}^2 ds \\ = & 4T(C_{g_e} + 2\xi(g) + T\xi(f)) + 2pC_L(4+2T)\int_0^t \|X^{(m)}(s) - X(s)\|_{L_2}^2 ds.\end{aligned}$$

最后，定义常数 $C_1 \coloneqq 4T(C_{g_e} + 2\xi(g) + T\xi(f))$ 和 $C_2 \coloneqq 2pC_L(4+2T)$，以及函数 $G(t) \coloneqq \|X^{(m)}(t) - X(t)\|_{L_2}^2$，并根据格朗沃尔不等式可得

$$G(t) \leq C_1 + C_2 \int_0^t G(s) ds \quad \leftrightarrow \quad G(t) \leq C_1 \exp(C_2 t)$$

证明至此完成。 □

在估计 Delay-SDE-net 以模拟表示某个动力系统的真实世界随机时滞微分方程时，估计过程是在离散时间域中进行的。也就是说，总误差的上界大于定理 2 中给出的上界。下面的推论给出了所估计 Delay-SDE-net 的总误差理论上界。

推论 1. 给定式(3) 中的随机时滞微分方程（*SDDE*）及其对应的（连续时间）双层 *Delay-SDE-net*，在假设 *1* 和 *2* 下，双层 *Delay-SDE-net* 的估计 π离散化版本（π 由式(4)定义）的误差界为

$$\|X^{\pi,(m)}(t) - X(t)\|_{L_2} \leq C_T^{1/2} \Delta t^{\gamma} + C_m.$$

常数 C_m 由定理 *2* 给出。此外，$C_T > 0$ 是一个依赖于 T 且独立于 Δt的常数，而 γ 是对真实世界随机时滞微分方程初始路径的正则性条件（见假设*2*）。

证明。 根据闵可夫斯基不等式，我们得到

$$\begin{aligned}\|X^{\pi,(m)}(t) - X(t)\|_{L_2} &= \|X^{\pi,(m)}(t) - X^{(m)}(t) + X^{(m)}(t) - X(t)\|_{L_2} \\ &\leq \|X^{\pi,(m)}(t) - X^{(m)}(t)\|_{L_2} + \|X^{(m)}(t) - X(t)\|_{L_2}.\end{aligned} \tag{22}$$

代入式（17）和定理 2 的结果即可完成证明。 □

5 分析与应用

本节介绍双层 Delay-SDE-net 的数值与实证研究。其中包括第5.1.2节的仿真研究，用于探究 Delay-SDE-net 的数值收敛性；以及第 5.1.3 节的比较研究，在该研究中，Delay-SDE-net 与两个基准模型在模拟数据上进行比较。最后，我们在第5.2节中通过一个涉及天气数据的真实世界案例研究，评估 Delay-SDE-net 的性能。

5.1 仿真研究

我们进行仿真研究，以分析双层 Delay-SDE-net 的特性和性能。具体而言，我们定义一个表示现实世界模型的随机时滞微分方程（见式（8）），用它来模拟数据集，以训练和评估 Delay-SDE-net。我们通过在具有不同时间分辨率的数据上测试经训练的 Delay-SDE-net 模型的参考解，研究该模型的数值收敛速度。此外，我们通过评估漂移系数和扩散系数的预测误差，将 Delay-SDE-net 与另外两种时间序列模型进行比较。

5.1.1 模拟设置

考虑式（8）中的真实世界随机时滞微分方程，其中 $d=2$ 和 $p=p_1=p_2=4$ （表示 $\Pi=\Pi_1=\Pi_2$）。也就是说，我们假设一个真实世界动力系统随时间按照 $X(t)\in\mathrm{R}^2$, 演化，并且依赖于 $\Pi(X_t(u))\in\mathrm{R}^8$, 其中 $t\in[0,T]$, $u\in[-\tau,0)$, $T,\tau\geq 0$。请注意，投影所得的点依赖关系 Π 取决于所定义的离散化方案。例如，当 $\Delta t=1$ 时，我们有 $(u_1,u_2,u_3,u_4)=(-4,-3,-2,-1)$ （详见第 5.1.2 节）。我们用参数 $(a^j(i)\,w^j(i)\,b^j)\in\mathrm{R}\times\mathrm{R}^9\times\mathrm{R}$ 定义随机时滞微分方程系数，如下所示。设漂移项$f=[f^1,f^2]^T$为

$$f^j(t,x)\coloneqq f^j(t,\Pi(X_t))=\sum_{i=1}^{2}a_i^j\sigma_1((w_i^j)^T[t,x]^T),\quad j=1,2,$$

其中 $a_1^1=a_1^2=a_2^1=a_2^2=5$，权重 $w_i^j\in\mathrm{R}^9, i=1,2, j=1,2$ 在式 (24) 中给出，且

$$\sigma_1(z)=\frac{e^{\alpha z}-e^{-\alpha z}}{e^{\alpha z}+e^{-\alpha z}},$$

是双曲正切激活函数。此外，扩散 $g=[g^1,g^2]^T$ 定义为

$$g^j(t,x)\coloneqq g^j(t,\Pi(\eta(0)))=a^j\sigma_2((w^j)^T[t,x]^T+b^j),\quad j=1,2,$$

wh其中 $a^1=4, a^2=1/8$，权重 $w^j\in\mathrm{R}^9, j=1,2$ 在式 (24) 中给出， $b^1=0, b^2=1$，且

$$\sigma_2(z)=\frac{1}{1+e^{-\lambda z}},$$

i是 Sigmoid激活函数。注意， $f^1,f^2,g^1,g^2\in\mathcal{B}$ 根据定义成立（见第2.4节）。

使用离散化方案模拟 $t\in[0,T]$ 上的样本路径

$$X^{j,\pi}(t_{k+1})=X^{j,\pi}(t_k)+f^j(t_k,\Pi(X_{t_k}))\Delta t+g^j(t_k,\Pi(\eta(0)))\sqrt{\Delta t}Z^j(t_k),\tag{23}$$

对于每个 $j=1$，2，其中 $Z^j(t_k)\sim N(0,1)$，且 f和 g的权重为

$$\begin{bmatrix}(w_1^1)^T\\(w_2^1)^T\\(w_1^2)^T\\(w_2^2)^T\\(w^1)^T\\(w^2)^T\end{bmatrix}=\begin{bmatrix}0&3&2&2&5&-3&1&-3&-1\\0&1&0&-0.5&0&-1&0&-0.5&0\\0&0&2&0&-3&0&1&0&0\\0&0&1&0&-0.5&0&0&0&-0.5\\-5/365&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&1\end{bmatrix}10^{-2}.\tag{24}$$

每条模拟路径均以 $\eta(t)=[\sin(z_1t),\cos(z_2t)]$, z_1， $z_2\sim N(0,1)$ 为初始值，在 $t\in[-\tau,0]$上。图6展示了一条具有 $\Delta t=1$、 $T=365$ 和 $\tau=4$ 的模拟样本路径示例（包括分布内数据和分布外数据）。

5.1.2 数值收敛性研究

本节通过数值方法研究双层 Delay-SDE-net 的收敛率。所介绍的方法受到 Lord 等人（2014 年）第 8.5 节工作的启发。

根据式（22）和推论 1，我们知道，离散时间双层 Delay-SDE-net 的 L_2-误差 $X^{\pi,(m)}$ 在某个固定时间 $t\in[0,T]$ 处有界，满足

$$\begin{aligned}\|X^{\pi,(m)}(t)-X(t)\|_{L_2}&\leq\|X^{\pi,(m)}(t)-X^{(m)}(t)\|_{L_2}+\|X^{(m)}(t)-X(t)\|_{L_2}\\&\leq C_T^{1/2}\Delta t^{\gamma}+C_m,\end{aligned}\tag{25}$$

其中， X 表示式（8）中的真实世界随机时滞微分方程的解， $X^{(m)}$ 表示式（9）中的连续时间 Delay-SDE-net。 $X^{\pi的(m)}$ L_2误差的数值估计记为

$$\Delta_{X^{\pi,(m)}}\simeq\|X^{\pi,(m)}(t)-X(t)\|_{L_2}.$$

请注意， C_m 表示连续时间 Delay-SDE-net 的 L_2 误差上界， $X^{(m)}$。也就是说，C_m 是该上界

$$\|X^{(m)}(t)-X(t)\|_{L_2}\leq C_m,$$

that 在定理 2 中给出了其明确表达式。 $X^{(m)}$ 的 L_2 误差数值估计记为 a s

$$\Delta_{X^{(m)}} \simeq \|X^{(m)}(t) - X(t)\|_{L_2},$$

如下所示。

根据公式 (25)， $X^{\pi,\,(m)}$ 的收敛率可以通过以下关系进行研究

$$\ln\left(\|X^{\pi,(m)}(t) - X(t)\|_{L_2} - C_m\right) \leq \frac{\ln C_T}{2} + \gamma \ln \Delta t. \tag{26}$$

由于 C_m 是一个上界，并且我们知道真实世界随机时滞微分方程的显式形式，即 $X(t)$，因此在本仿真研究中，我们将使用估计的 L_2-误差 $\Delta_{X^{(m)}}$ 代替 C_m 来评估收敛率 γ。此外，在研究收敛率时，公式 (26) 中的项 $\ln\, C_T/2$并未提供有用信息，因为它只会影响纵轴截距。也就是说，为了研究 $X^{\pi,\,(m)}$的收敛率，我们将考虑以下关系

$$\ln\left(\Delta_{X^{\pi,(m)}} - \Delta_{X^{(m)}}\right) \leq \gamma \ln \Delta t. \tag{27}$$

以下是进行双层 Delay-SDE-net 数值收敛性研究的分步方法:

1. 使用小的 Δt_{ref}，从已知的真实世界随机时滞微分方程 $X(t)$中模拟数据。2. 使用分辨率 Δt_{ref}将双层 Delay-SDE-net 拟合到模拟数据上。这被视为 $X^{(m)}(t)$的精确参考解。3. 计算估计的 L_2-误差 $\Delta_{X^{(m)}}$。4. 对第 1 步生成的模拟数据，针对多个 κ，使用分辨率 $\Delta t = \kappa \Delta t_{\text{ref}}$、N $\ni \kappa > 1$计算估计的 L_2-误差$\Delta_{X^{\pi,\,(m)}}$。5. 按照推论 1 中给出的常规数值收敛性研究方法，评估收敛率 γ。

上述方法用于研究拟合真实世界随机时滞微分方程数据的双层 Delay-SDE-net 的收敛率，该方程定义于第 5.1.1 节。具体而言，我们使用公式 (23) 模拟 $M = 1000$ 个真实世界随机时滞微分方程样本，其中

$$\Delta t_{\text{ref}} \coloneqq t_{k+1} - t_k = 0.01, \quad t_k \in (0, 5], \quad k = 1, \ldots, N_{\text{ref}},$$

其中 $N_{\text{ref}} = 500$，且对于初始路径 η， $t_{-l} \in [-15, 0]$、 $l = 0$、.....、1500。请注意，这意味着 $\tau = 15$ 和$T = 5$。模拟得到的 M 个样本被划分为训练/测试子集（$M_{\text{train}} = 700/M_{\text{test}} = 300$ 划分）。连续时间 Delay-SDE-net 的参考解使用训练数据集中的所有点进行拟合，也就是说，拟合分辨率为 Δt_{ref}。由此得到连续时间延迟随机微分方程网络 $X^{(m)}$ 的参考解 $X^{0.01,\,(m)}$。请注意，使用 $X^{0.01,\,(m)}$预测 X 时，必须使用与生成真实过程 X数据时相同的布朗路径。计算得到的 $\Delta_{X^{(m)}}$ 值见表1。有关数值计算的详细说明，请参见附录 A.1。

在步骤 4 中，我们使用训练好的 Delay-SDE-net（参考解），针对分辨率降低后的测试子集，为模拟测试数据集生成五条新的预测路径

$$\Delta t = \kappa \Delta t_{\text{ref}}, \quad \kappa = 5, 10, 50, 100, 500.$$

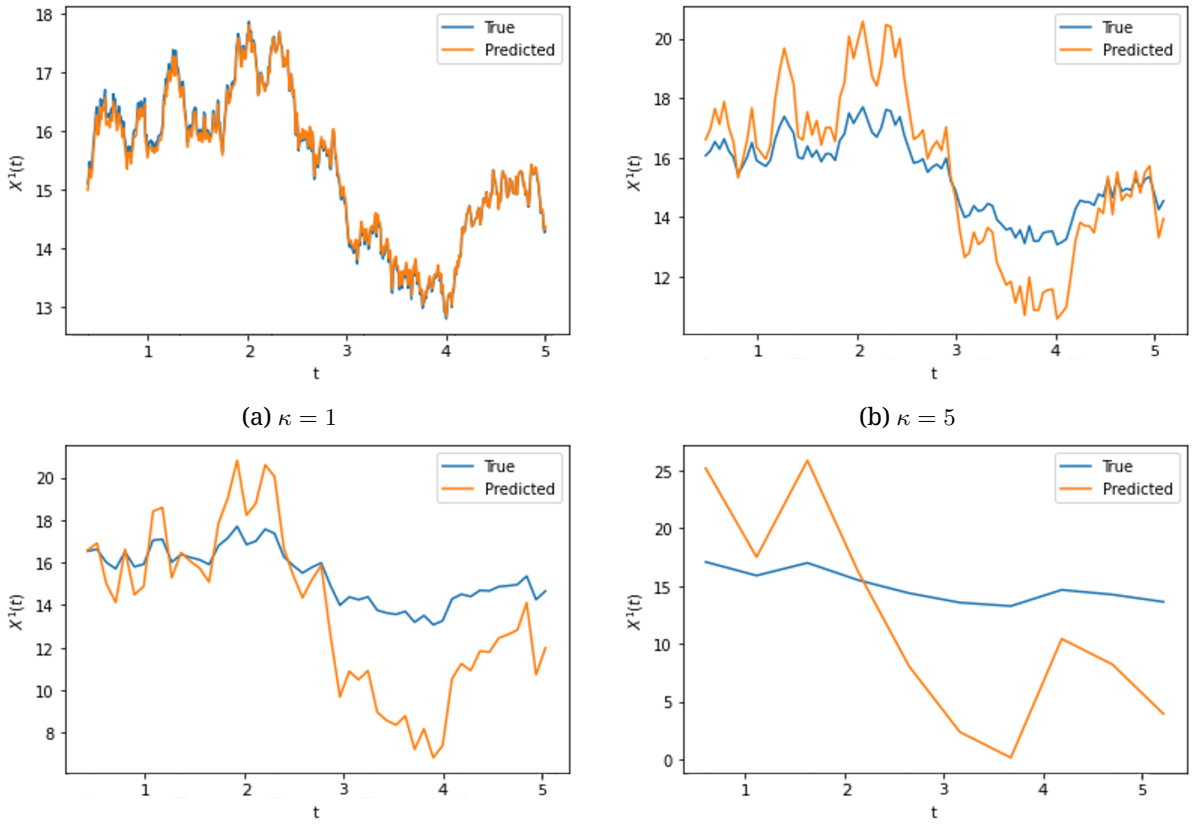
这些分辨率分别生成大小为 $N_\kappa = N_{\text{ref}}/\kappa = 100$，50，10，5，1 的样本路径。请注意，每种分辨率下的预测都代表一个离散化延迟随机微分方程网络模型 $X^{\pi,\,(m)}$。也就是说， $X^{0.05,\,(m)}$、 $X^{0.1,\,(m)}$、 $X^{0.5,\,(m)}$、 $X^{1,\,(m)}$和 $X^{5,\,(m)}$。针对这五个离散化延迟随机微分方程网络，分别计算相应的 L_2-误差估计 $\Delta_{X^{\pi,\,(m)}}$，结果列于表1。关于数值计算的更多详情，请参阅附录A.2。

κ	1	5	10	50	100	500
$\Delta_{X^{(m)}}$	1.956	–	–	–	–	–
$\Delta_{X^{\pi,\,(m)}}$	–	6.030	9.764	26.082	37.658	78.328

表1：连续时间延迟随机微分方程网络 $\Delta_{X^{(m)}}$的误差估计，以及离散时间延迟随机微分方程网络 $\Delta_{X^{\pi,\,(m)}}$ 在五种不同分辨率下的误差估计。

Sett本数值研究中神经网络的训练设置见附录A.3。

正如预期，从表1可以看出，离散化方案的分辨率越高，预测就越准确。显然， $X^{0.01,\,(m)}$（即拟合为连续时间 Delay-SDE-net 的精确参考解）明显优于相应的离散化延迟随机微分方程网络。图4a–4d分别展示了这些模型 $X^{0.01,\,(m)} \simeq X^{(m)}$、 $X^{0.05,\,(m)}$、 $X^{0.1,\,(m)}$ 和 $X^{0.5,\,(m)}$的上述特征，其中绘出了每个模型对一条给定样本路径的预测结果。也就是说，对于分辨率较低的情况，使用 Delay-SDE-net 模型预测的路径往往会偏离真实路径更远。



(a) $\kappa = 1$

(b) $\kappa = 5$

(c) $\kappa = 10$

(d) $\kappa = 50$

图4：Delay-SDE-net 模型 $X^{(m)}(t)$ ($\kappa = 1$)、 $X^{0.05,\,(m)}(t)$ ($\kappa = 5$)、 $X^{0.1,\,(m)}(t)$ ($\kappa = 10$) 和 $X^{0.5,\,(m)}(t)$ ($\kappa = 50$) 对 $X^1(t)$ 的模拟样本路径进行的预测（第 5.1.1 节）。

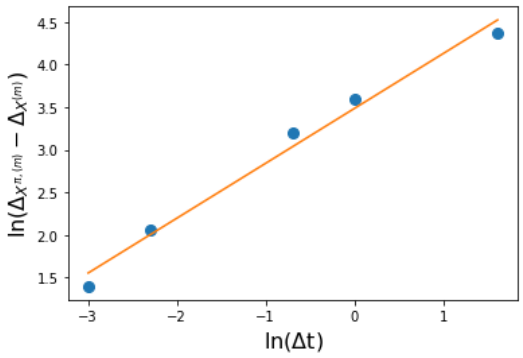


图5：离散化延迟随机微分方程网络的近似 L_2 误差相对于 Δt 的双对数图。图中展示的 Delay-SDE-net 为 $X^{\pi,\,(m)}(t)$，其离散化步长为 0.05、0.1、0.5、1 和 5。同时还展示了相应的线性回归图，表示 $X^{(m)}(t)$的收敛速率。

Delay-SDE-net 的收敛率 γ 可通过将表1中的值代入公式（27）进行评估。结果见图5，并附有相应的线性回归。图5中线性回归的斜率所对应的收敛速率在本研究中为 $\gamma = 0.6359$ 。估计得到的收敛率优于理论上的Euler-Maruyama格式收敛率，即 $\gamma_{\text{Euler}} = 1/2$。请注意，Euler-Maruyama格式是针对具有状态依赖型漂移项和扩散项的 Itô扩散过程推导的。本研究假设 Delay-SDE-net 的扩散项与状态无关（见第3.1节），这意味着相应的随机时滞微分方程模型在这方面较为简单。然而，本研究中的随机时滞微分方程模型比标准的 Itô扩散过程更复杂，因为后者不具备记忆。标准 Itô扩散过程与我们的随机时滞微分方程建模框架各有优劣，这些权衡似乎使我们的模型在收敛率方面略占优势。

请注意，我们将 Delay-SDE-net 的收敛率与 Euler-Maruyama格式进行比较，而不是与 Milstein格式进行比较，尽管推论1中的收敛结果是通过 Milstein格式推导得出的。这样做的原因是，对于采用状态无关扩散的模型，Milstein格式会退化为Euler-Maruyama格式。

5.1.3 与其他模型比较

本节通过将 Delay-SDE-net 与两种用于动力系统的类似模型进行比较，评估其性能。这些基准模型，即 VAR模型和 SDE-net，具有与 Delay-SDE-net 类似的性质。VAR模型能够捕捉动力系统不同滞后版本之间的关系，但只能捕捉线性关系，这一点与 Delay-SDE-net 不同。SDE-net 和 Delay-SDE-net 都基于具有非线性系数的随机微分方程；然而，SDE-net 仅包含系统当前状态的信息，而 Delay-SDE-net 还能够捕捉动力系统滞后状态中的非线性。我们使用第5.1.1节中介绍的随机时滞微分方程模拟生成的数据对这些模型进行测试。

以下几点说明了如何使用第5.1.1节中的模拟设置生成训练数据集和测试数据集，以及如何开展比较研究：

1. 模拟 110 条数据路径，取值为 $\Delta t := t_{k+1} - t_k = 1, \quad t_k \in [1, 365], \quad k = 1, \ldots, 365$。模拟数据对应 110 年的每日值（已剔除闰年）。 2. 将数据划分为训练/验证/测试子集（按 90/10/10 的比例划分）。 ˆ 3. 使用训练数据 $\mathcal{D}_0$训练预测模型 f和 ˆg_a，在验证数据上调整超参数，并使用测试数据 $\mathcal{D}$评估性能。所有这些数据集均由 $X^\pi = [X^{1,\pi}, X^{2,\pi}]^T$生成，见公式（23）。 4. 从由 $\tilde{X}^\pi = [\tilde{X}^{1,\pi}, \tilde{X}^{2,\pi}]^T$生成的数据集 $\tilde{\mathcal{D}}$中抽取分布外数据； $\tilde{X}^\pi = [\tilde{X}^{1,\pi}, \tilde{X}^{2,\pi}]^T$即公式（23）中的随机时滞微分方程，但噪声 $\tilde{g}^1 = 2.5g^1$和 $\tilde{g}^2 = 2.5g^2$有所增加。也就是说，生成比分布内数据取值更极端、但仍符合实际情况的分布外数据。图6展示了一年模拟分布内数据和分布外数据的曲线。˜5. 将测试集 $\mathcal{D}$ 中随机选定的数值区间替换为来自 $\mathcal{D}$的数值区间。附录B.1介绍了所替换区间的详情。这些区间是随机抽取的，以满足以下条件：

$$\max(X^{j,\pi}) < \tilde{X}^{j,\pi} < \min(X^{j,\pi}), \quad j = 1, 2,$$

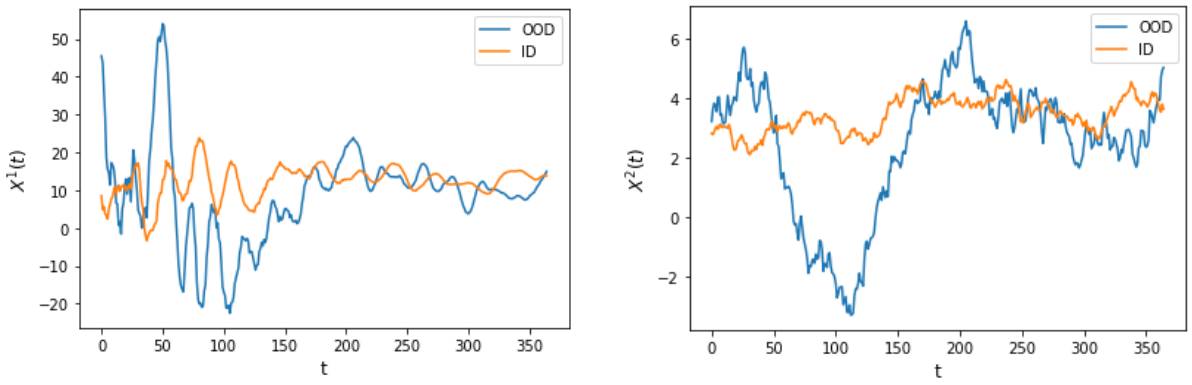
使用 $X^{j,\pi} \in \mathcal{D}_0$，确保这些数据点相对于训练集属于分布外数据。修改后的数据集 $\mathcal{D}^*$包含 3.9% 的分布外数据，模拟了极端值较为罕见的真实情形。

6. 评估 ˆg_e^{prob}在 $\mathcal{D}^*$中识别分布外数据的性能。

SDE-net 根据 Kong 等人（2020）中的回归示例实现，而 VAR 模型的预测则基于 Eggen（2022）的实证研究。SDE-net 和 Delay-SDE-net 的最终调优参数，以及 VAR模型的参数，分别列于附录 B.2 和 B.3。

三种模型在模拟数据上的性能见表 2。漂移项通过真实 X^1 与预测值 ˆX^1 （使用式（12））之间的均方根误差（均方根误差）进行评估，其中假设 X^1 是关注变量， X^2 是解释变量。偶然不确定性扩散项通过真实偶然方差与预测偶然方差 g_a^2 和 ˆg_a^2 之间的均方根误差进行评估。最后，认知不确定性扩散项通过受试者工作特征曲线下面积（ROC曲线下面积）进行评估，用于衡量 ˆg_e^{prob}识别数据是否来自分布外分布 $\tilde{X}$ 的能力（二元分类）。

在预测 f、 g_a 和 g_e^{prob} 时，Delay-SDE-net 的表现始终优于对比模型。显然，在这类时间序列数据上，Delay-SDE-net 相较于 SDE-net 和 VAR模型具有一定优势。VAR



(a) (b)

图6： $X^1(t)$ 的一年模拟分布内和分布外数据，分别对应 (a) 和 (b) $X^2(t)$。

模型在预测 f 和 g_a 时的表现优于 SDE-net。众所周知，正如表2中的结果所反映的那样，线性模型在非线性系统建模中往往表现良好，尤其是对于维度较低的简单系统；本研究中的模拟数据正属于此类。若模拟具有更复杂结构、交互效应且维度更高的数据，我们预计神经网络模型相较于 VAR模型能够取得更好的表现，且优于本次仿真研究中所达到的水平。

与 SDE-net 相比，Delay-SDE-net 对 g_e^{prob} 的估计也更准确，给出的预测几乎完美，接近 1。请注意，g_e^{prob} 与预测步数无关，因为它仅取决于第一步的输入值。最后，与 Delay-SDE-net 和 SDE-net 不同，VAR模型不提供 g_e^{prob} 的预测。

	值（$x|f$） （均方根误差）			偶然不确定性（g_a^2） （均方根误差）			认知不确定性（g_e） （ROC曲线下面积）	
	向量自回归	SDE-net	延迟	向量自回归	SDE-net	延迟	SDE-net	延迟
$t+1$	0.70	1.31	**0.66**	0.58	3.03	**0.26**	0.901	**0.992**
$t+2$	1.09	1.82	**1.02**	1.45	3.37	**0.39**		
$t+3$	1.52	2.32	**1.44**	2.39	3.88	**1.00**		
$t+4$	1.92	2.72	**1.84**	3.35	4.42	**2.06**		
$t+5$	2.25	3.06	**2.17**	4.33	5.04	**2.62**		
$t+6$	2.51	3.28	**2.42**	5.31	5.74	**3.13**		
$t+7$	2.72	3.40	**2.63**	6.30	6.45	**3.63**		

表 2：三个对比模型对 f、g_a^2 和 g_e 的预测性能，其中延迟 代表 Delay-SDE-net。

5.2 现实案例研究

本节将 Delay-SDE-net 应用于一项现实案例研究，并将其性能与 SDE-net 和 VAR模型进行比较（如第 5.1.3 节所示）。该案例研究涉及对现实二维天气系统的短期预测。也就是说，基于风和温度的二维时间序列，我们使用 Delay-SDE-net 对未来一周的风进行预测。下一节将介绍有关该时间序列的更多信息，并阐述为何使用这一特定动力系统来评估 Delay-SDE-net。

5.2.1 数据

本节总结案例研究中使用的数据，并说明为什么值得研究这一二维数据集。

众所周知，天气是一个混沌的非线性动力系统。因此，即使天气变量具有一定程度的记忆，可用于天气预测，可靠的预测也仅限于约10天（Krishnamurthy，2019）。近期研究着重于提升地表天气的长期预测能力。恰当表征平流层天气（平流层是大气层中距地表约15至50公里的区域），有望提升地表天气的长期预测能力，相关示例见 Blanc等人（2018）。尤其是，

在这方面，称为平流层突然增温（SSW）的极端平流层事件尤为重要（Karpechko 等人，2016；Scaife 等人，2022；Hitchcock、Simpson，2014）。平流层突然增温现象是平流层北半球冬季环流突然受阻的结果。具体而言，冬季时，平流层北极周围环绕着高速西风转东风（U风），这种风每十年减弱并最终反向约6次（Pedatella 等人，2018）。更全面的介绍请参见 Eggen 等人（2022年）及其中引用的文献。

在下一节中，我们将评估 Delay-SDE-net 作为平流层纬向U风的短期预测模型。此案例具有重要意义，因为它对 Delay-SDE-net 提出了多方面的挑战，包括季节性趋势、非线性、时间记忆、季节性不确定性（方差）模式，以及发生平流层突然增温（SSW）等极端事件的可能性。根据平流层突然增温的标准定义（Butler 等人，2015年），我们使用10百帕高度处的每日纬向平均平流层温度和U风数据，以及 60°N，对 Delay-SDE-net 进行拟合。数据规格见表3，最终的每日纬向平均时间序列按照 Eggen 等人（2022年）所述进行整理。请注意，基于 Eggen（2022）所识别出的这两种气象变量之间的交叉相关，我们选择研究由U风和温度构成的二维动力系统。

Date	网格	气压层	Time	Area	Unit
1979年1月1日至 2018年12月31日	0.5°	10百帕	00:00、06:00、 12:00、18:00	60°N 和 [−180°E 180°E)	温度：开尔文， U风：米/秒

表3：平流层温度和U风数据的规格。除了单位外，这两个值的规格相似。初始数据集取自ECMWFERA-Interim再分析模型产品。本表摘自Eggen（2022）。

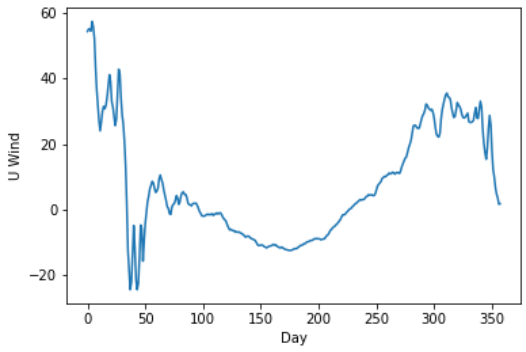


图7：2018年平流层U风的每日纬向平均。2月发生了一次强烈的平流层突然增温，从第45至51天左右U风方向的突然变化可以看出。年底时可以观察到新一轮平流层突然增温的开始。

如上所示，U风的值和方差均呈现季节性模式。冬季，U风通常为幅度较大的正值；到了夏季，则转为幅度较小的负值。此外，冬季的平流层比夏季更加动荡，导致冬季平流层天气的偶然不确定性更高。图7展示了一个示例年份（2018年）的每日纬向平均平流层U风，可以清楚地看到上述特点。另外，请注意2018年发生的大型平流层突然增温事件，其间冬季平流层的风向突然改变。

以下部分的结果基于一个数据集，该数据集分为训练集和测试集（按30/10划分）；模型使用1979—2008年的数据进行训练，并使用2009—2018 年的数据进行评估。

5.2.2 结果

一个成功的预测模型应当能够高性能地预测U风。此外，它还应能够识别平流层突然增温（SSW），即使这类事件可能来自分布外数据，并预测季节性方差；冬季的方差高于夏季。Delay-SDE-net使用经过确定性训练的漂移网络 f_m 来预测期望U风，使用偶然不确定性扩散网络$g_{a,\ m}$ 来预测季节性方差，并使用认知不确定性扩散网络$g_{e,\ m}$ 来预测平流层突然增温等异常事件带来的不确定性。展示结果时，我们将观测值 $\{\boldsymbol{x}_{t_k} = [x_{1,t_k}, x_{2,t_k}] : t_k \in [1, 365]\}$ 称为 x_1 和 x_2，其中U风 x_1是关注变量，而温度 x_2是解释变量。

模型通过向前预测1至7天，与向量自回归模型和SDE-net模型进行比较，结果见表4。漂移通过真实 x_1 与预测U风 $\hat{x}_1$ 之间的均方根误差（使用公式（12））进行评估；扩散则通过预测方差 $\hat{g}^2$ 与真实残差 $e^2 = (x_1 - \hat{x}_1)^2$, 之间的均方根误差进行评估，以衡量模型在数据上预测自身不确定性的能力。对于Delay-SDE-net，$\hat{g}^2 = (g_{a,\ m} + g_{e,\ m})^2$。请注意，对于这个真实世界数据集，我们不知道真实的 g_a 和 g_e，因此必须评估模型估计自身总不确定性的能力，即 $\hat{g}^2 = (\hat{g}_a + \hat{g}_e)^2$, , 其中对于Delay-SDE-net和SDE-net，$\hat{g}_e = \sigma_e^{(N)} \hat{g}_e^{\text{prob}}$。表4中的结果所用的 $\sigma_e^{(N)}$ 是根据公式16估计的；然而，在图8、图9和图10中，为便于说明，$\sigma_e^{(N)}$ 设定为较高值。

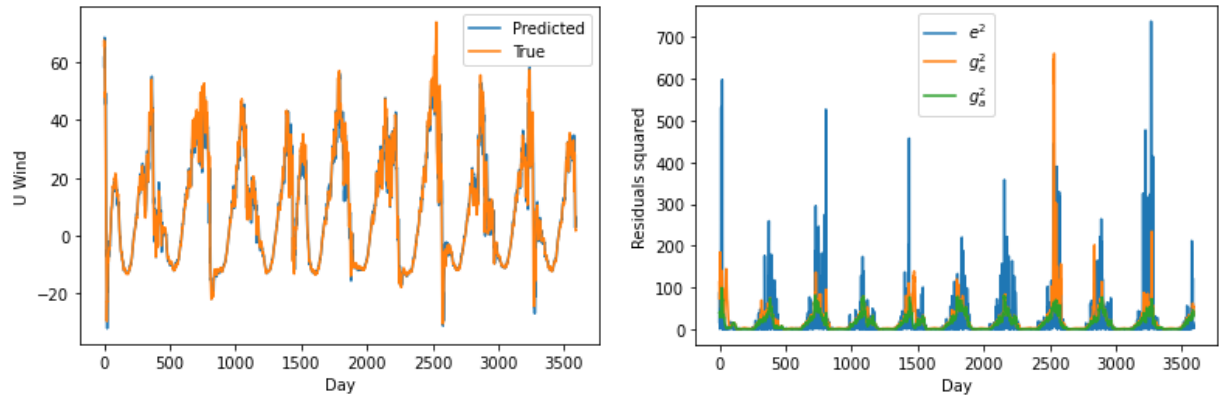
	值 $(x_1 \mid f)$ (均方根误差)			不确定性 (g^2) (均方根误差)		
	VAR	SDE-net	Delay-SDE-net	VAR	SDE-net	Delay-SDE-net
$t+1$	2.31	2.25	**1.48**	15	14	**6**
$t+2$	3.37	3.70	**2.92**	34	36	**24**
$t+3$	4.40	4.64	**4.10**	58	63	**44**
$t+4$	5.20	5.59	**5.02**	81	90	**67**
$t+5$	5.84	6.77	**5.70**	104	130	**90**
$t+6$	6.38	7.36	**6.28**	128	149	**114**
$t+7$	6.82	7.93	**6.73**	148	179	**130**

表4：提前1至7天预测。值 表示 x_1 与 $\hat{x}_1$ 之间的均方根误差，而不确定性 表示 $\hat{g}^2 = (\hat{g}_a + \hat{g}_e)^2$ 与 $e^2 = (x_1 - \hat{x}_1)^2$ 之间的均方根误差。

Delay-SDE-net 对 f 和 g 的预测始终最佳。总体而言，SDE-net 在 f 和 g 方面的表现均最差，也低于 VAR 模型，这表明对于该数据集，纳入时间滞后比使用非线性函数更重要。Delay-SDE-net 通过神经网络捕捉非线性复杂函数，同时利用当前状态和过去状态，发挥类似随机时滞微分方程的作用，因此表现良好。不过，我们发现，在较小的 N 下，Delay-SDE-net 相较于 VAR 模型在提前 N 天预测中的性能优势更大；随着预测时间延长，这一优势会缩小。这与已有认识相符：天气具有有限长度的内在记忆，同时存在大量自然混沌（Krishnamurthy，2019），这意味着随着 N 增大，所有预测都会趋近于随机猜测。

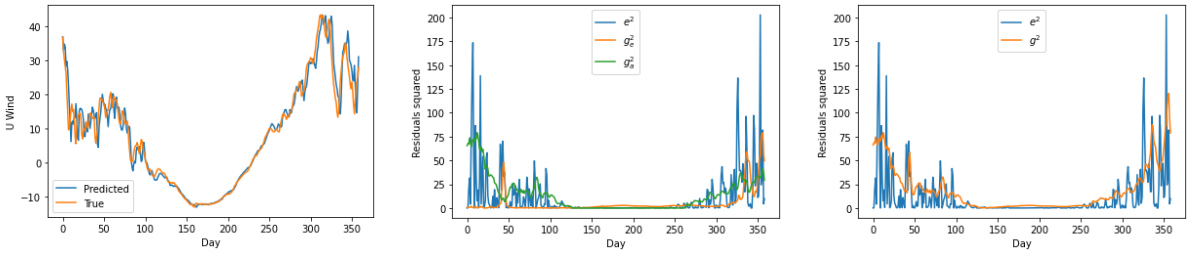
为了展示 Delay-SDE-net 作为预测模型的效果，图 8a 展示了 Delay-SDE-net 对测试数据进行的提前 3 天预测，图 8b 则展示了残差平方及预测不确定性。可以看出，预测的 U风与真实 U风吻合良好，而 $g^2_{a,\ m}$ 捕捉到了季节性所带来的不确定性，以及 $g^2_{e,\ m}$ 一些由平流层突然增温等异常事件引起的残差峰值。不过，由于难以清楚地观察跨度超过 10 年的时间序列上的预测结果，我们以 2012 年为例，在图 9 中放大展示。这是相当动荡的一年，其中发生了两次小型平流层突然增温（分别在年初和年末），并且在年末还开始了一次大型平流层突然增温，该事件持续至 2013 年初。我们观察到，预测的 U风吻合良好，但在方差较高的冬季以及平流层突然增温期间出现了一些误差。$g_{a,\ m}$ 对偶然不确定性给出了良好的预测，既捕捉到了其季节性，也捕捉到了 U风和温度值中的其他影响，并且遵循了图 9b 中残差的变化模式。然而，$g_{a,\ m}$ 未能对第 50 天之前的平流层突然增温期间以及年末的方差给出足够高的预测，这可能是因为训练数据中没有多少类似情况。我们观察到，$g_{e,\ m}$ 在这些情况下预测出较高的认知不确定性，提示可能会发生异常情况。两种不确定性预测结合起来呈现出良好的模式，捕捉到了图 9c 中的大部分残差。

通过结合预期 U风的预测值与预测的不确定性，我们可以绘制带有置信区间的预测图。图 10 展示了 95% 置信区间，分别对应 $g_{a,m}$ 和 $g_{e,m}$，以及图 10c 中的总不确定性。正如预期，$g_{a,m}$ 在多变的冬季给出了较宽的置信区间，而到了夏季则趋近于零，此时 U风稳定得多。与 $g_{e,m}$ 相对应的置信区间总体较窄（由于它是作为概率进行训练的，因此实际上不会为零），但在不熟悉的平流层突然增温附近会变宽。图 10c 中的组合图为预测的 U风提供了合理的置信区间，并显示神经网络的预测 f_m, $g_{a,m}$ 和 $g_{e,m}$ 如何很好地协同工作。



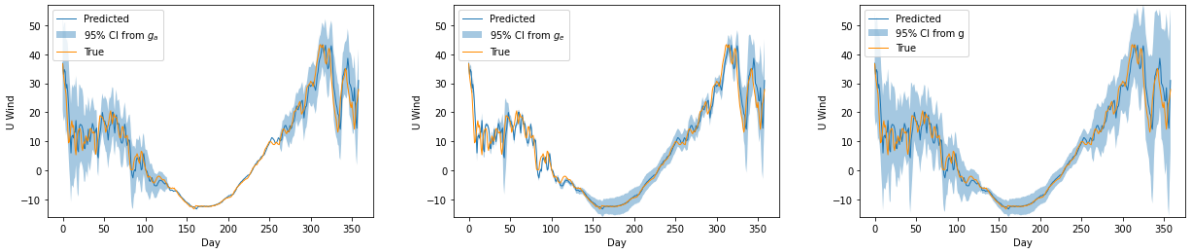
(a) (b)

图 8：提前 3 天的（a）2009—2018 年 U风真实值与预测值，以及（b）预测方差与残差平方的对比图。绿色曲线显示$g^2_{a,\ m}$，橙色曲线则是在$g^2_{a,\ m}$的基础上叠加了$g^2_{e,\ m}$。



(a) (b) (c)

图 9：2012 年提前 3 天的（a）U风真实值与预测值，（b）残差平方以及预测的偶然不确定性和认知不确定性方差，还有（c）残差与预测总方差。



(a) (b) (c)

图 10：2012 年提前 3 天的 U风预测值及其 95% 置信区间，置信区间分别根据（a） g_{a_m}、（b）$g_{e,\ m}$ 和（c） g_m计算得出。

6 结论与后续工作

本文介绍了 Delay-SDE-net，这是一种用于预测具有记忆的时间序列的新型神经网络模型，同时还能分别预测偶然不确定性和认知不确定性。该模型被构造为表现得像随机时滞微分方程（SDDE），也就是说，它基于状态依赖漂移项和随机项，根据时间序列中的滞后数据进行预测。漂移项由一个神经网络建模，并对时间序列的值进行确定性预测。随机项由两个神经网络组成，其中一个对时间序列中的偶然不确定性建模，另一个则对认知不确定性建模。

我们推导了训练后的 Delay-SDE-net 相对于真实世界随机时滞微分方程的理论误差上界。该误差上界由两部分组成：连续时间延迟随机微分方程网络与相应的真实世界随机时滞微分方程之间的误差上界，以及连续时间延迟随机微分方程网络与其

离散化版本之间的误差上界。前者与表示模型系数的神经网络中的神经元数量成反比，后者则与时间离散化的分辨率有关，这与通常情况一致。也就是说，我们证明了，通过在训练网络时增加神经元数量并采用更精细的离散化方案，可以降低模型误差。我们通过数值方法研究了离散时间延迟随机微分方程网络向连续时间延迟随机微分方程网络的收敛率，发现其收敛率为 $\gamma = 0.64$，优于伊藤扩散过程的标准欧拉–丸山收敛速率。

Delay-SDE-net 在根据真实世界随机时滞微分方程模拟的数据上，以及在涉及天气变量二维系统的真实世界案例研究中进行了性能评估。也就是说，我们将 Delay-SDE-net 作出的预测与原始 SDE-net 和 VAR模型的相应预测进行了比较。无论是在预测时间序列的未来值，还是在估计偶然不确定性和认知不确定性方面，Delay-SDE-net 的表现始终最佳。
随着越来越多的领域和应用区域开始利用机器学习等先进统计方法，针对每种具体情况的需求提供合适的模型变得愈发重要。例如，安全关键型应用更加需要提供预测不确定性，尤其是认知不确定性，以便在意外运行条件下也能作出安全决策。金融和天气建模等其他领域通常涉及固有偶然不确定性较高的数据，因此估计这种不确定性十分重要。此外，许多应用领域需要模型预测和不确定性估计能够近乎即时地响应，从而及时作出决策。自动驾驶汽车的运动控制和股票交易就是其中的例子。因此，Delay-SDE-net 采用单一确定性方法生成不确定性估计，这意味着它能够即时提供预测和不确定性估计，不会因反复模拟而延迟。

Delay-SDE-net 的综合特性，例如能够对具有记忆的时间序列建模，并实时预测不确定性，使其目前成为一种独特的模型，适用于众多寻求更先进时间序列数据驱动建模技术的领域。此外，将网络系统限制为遵循随机时滞微分方程（SDDE），有望改善模型的可解释性、鲁棒性以及在分布外数据上的稳定性等特性，同时减少所需的训练数据量，这与其他物理信息神经网络的情况类似（Karniadakis 等人，2021年；Kashinath 等人，2021年；Yang，Perdikaris，2019年）。
由于 Delay-SDE-net 是一种新颖的方法，其训练、建模和数值评估方面仍有若干内容值得进一步探索。作者尤其希望深入研究认知不确定性扩散网络的训练方法。例如，可以采用回归而非分类，以避免使用调节参数 σ_ϵ。此外，也值得对软布朗偏移法在更高维度下的性质进行更全面的研究，从而为该方法应用于时间序列数据时如何恰当选择调优参数制定具体指南。本文通过数值方法评估了离散时间 Delay-SDE-net 的收敛率。另一项有待开展的误差评估是通过数值方法研究连续时间延迟随机微分方程网络的误差界，以验证该误差界是否与网络中的神经元数量成反比。最后，也有必要在其他数据集和其他领域中验证该模型。关于这一点，目前有一项正在进行的工作使用 Delay-SDE-net 对一组经过处理的复杂次声测量数据进行建模，以预测平流层风。初步预测结果已达到该特定领域的前沿水平。这些结果证明了 Delay-SDE-net 对复杂动力系统进行建模的能力。

致谢

特别感谢 Fred Espen Benth 在本研究过程中提供的宝贵建议。作者也感谢 Riccardo De Bin、Quentin Brissaud、Arne Bang Huseby 和 Sven Peter Näsholm 的支持。M.D.E. 获得了 NORSAR 提供的博士研究资助。作者声明不存在任何利益冲突。

A 数值收敛性研究

A.1 连续时间延迟随机微分方程网络参考解

连续时间 Delay-SDE-net 使用 M_{test} 个样本进行测试。也就是说，设 $s = 1$、.....、M_{test}，使得每条样本路径的数据集定义为

$$\mathcal{D}_s := \{\boldsymbol{\eta}_{t_{-l}}^{(s)} \in \mathbb{R}^2 : \quad t_{-l} \in [-15, 0]\} \cup \{\boldsymbol{x}_{t_k}^{(s)} \in \mathbb{R}^2 : \quad t_k \in (0, 5]\},$$

表示真实世界随机时滞微分方程，$X(t)$。对于每个 $\mathcal{D}_s$，使用训练好的连续时间模型 $X^{0.01,\,(m)}(t) \simeq X^{(m)}(t)$，从时间 $t_0 = 0$ 向前预测 500 步，直至时间 $t_{N_{\text{ref}}} = 5$（其中 $X^{0.01,\,(m)}(t) = X^{\pi,\,(m)}(t)$，对于 $\Delta t_{\text{ref}} = 0.01$）。

也就是说，对于每个 s, 使用训练好的双层 Delay-SDE-net 系数 f_m 和 $g_m = (g_{a,\ m} + g_{e,\ m})$ 进行计算

$$
\begin{aligned}
X^{(m)}(0.01) \simeq \hat{\boldsymbol{x}}_{t_1}^{(s)} &= \boldsymbol{\eta}_{t_0}^{(s)} + f_m(t_1, \boldsymbol{\eta}_{t_0}^{(s)}, \boldsymbol{\eta}_{t_{-1}}^{(s)}, \boldsymbol{\eta}_{t_{-2}}^{(s)}, \boldsymbol{\eta}_{t_{-3}}^{(s)})\Delta t_{\text{ref}} + g_m \Delta W_{\text{ref}}^{(1)}, \\
X^{(m)}(0.02) \simeq \hat{\boldsymbol{x}}_{t_2}^{(s)} &= \hat{\boldsymbol{x}}_{t_1}^{(s)} + f_m(t_2, \hat{\boldsymbol{x}}_{t_1}^{(s)}, \boldsymbol{\eta}_{t_0}^{(s)}, \boldsymbol{\eta}_{t_{-1}}^{(s)}, \boldsymbol{\eta}_{t_{-2}}^{(s)})\Delta t_{\text{ref}} + g_m \Delta W_{\text{ref}}^{(2)}, \\
&\vdots \\
X^{(m)}(5) \simeq \hat{\boldsymbol{x}}_{t_{N_{\text{ref}}}}^{(s)} &= \hat{\boldsymbol{x}}_{t_{499}}^{(s)} + f_m(t_{N_{\text{ref}}}, \hat{\boldsymbol{x}}_{t_{499}}^{(s)}, \ldots, \hat{\boldsymbol{x}}_{t_{496}}^{(s)})\Delta t_{\text{ref}} + g_m \Delta W_{\text{ref}}^{(N_{\text{ref}})},
\end{aligned}
$$

其中 $g_m \coloneqq g_m(t_k,\ \boldsymbol{\eta}_{t_0}^{(s)},\ \boldsymbol{\eta}_{t_{-1}}^{(s)},\ \boldsymbol{\eta}_{t_{-2}}^{(s)},\ \boldsymbol{\eta}_{t_{-3}}^{(s)}),\ k = 1,\ \ldots\ldots,\ N_{\text{ref}}$，且 $\Delta W^{(k)}$ 参考$= W(t_{k+1}) - W(t_k)$，$\Delta W_{(k)}$ 参考$\sim N(0,\ \Delta t_{\text{ref}})^d$。现在，相应的 L_2-误差估计为

$$
\Delta_{X^{(m)}} \coloneqq \left(\frac{1}{M_{\text{test}}} \sum_{s=1}^{M_{\text{test}}} \| \hat{\boldsymbol{x}}_{t_{N_{\text{ref}}}}^{(s)} - \boldsymbol{x}_{t_{N_{\text{ref}}}}^{(s)} \|_2^2 \right)^{1/2}. \tag{28}
$$

A.2 离散时间 Delay-SDE-net

对于这五种模型中的每一种，都必须使用与生成随机时滞微分方程以及连续 Delay-SDE-net（参考解）所用的同一布朗路径进行预测，以便所有预测结果都可比较。也就是说，对于每个 κ，我们使用

$$
\Delta W_{\kappa}^{(\tilde{k})} = W(t_{\tilde{k}\kappa+\kappa}) - W(t_{\tilde{k}\kappa}) = \sum_{k=1}^{\kappa} \Delta W_{\text{ref}}^{(k)}, \tag{29}
$$

对于 $\tilde{k} = 1,\ \ldots\ldots,\ N_{\kappa}$，使得 $\Delta W_{(k)_{\kappa}} \sim N(0, \Delta t)^d$。采用这一设置，我们从每个离散化 Delay-SDE-net 模型 $X^{\pi,(m)}(t)$ 中获得 M_{test} 个预测样本，即 $X^{0.05,(m)}(t)$、$X^{0.1,(m)}(t)$、$X^{0.5,(m)}(t)$、$X^{1,(m)}(t)$ 以及 $X^{5,(m)}(t)$。例如，下面的示意图展示了如何根据 $X^{1,(m)}(t)$ 预测 $\mathcal{D}_s$：

$$
\begin{aligned}
X^{1,(m)}(1) = \hat{\boldsymbol{x}}_{t_{100}}^{(s)} &= \boldsymbol{\eta}_{t_0}^{(s)} + f_m(t_{100}, \boldsymbol{\eta}_{t_0}^{(s)}, \boldsymbol{\eta}_{t_{-100}}^{(s)}, \boldsymbol{\eta}_{t_{-200}}^{(s)}, \boldsymbol{\eta}_{t_{-300}}^{(s)})\Delta t + g_m \Delta W_{100}^{(1)}, \\
X^{1,(m)}(2) \simeq \hat{\boldsymbol{x}}_{t_{200}}^{(s)} &= \hat{\boldsymbol{x}}_{t_{100}}^{(s)} + f_m(t_{200}, \hat{\boldsymbol{x}}_{t_{100}}^{(s)}, \boldsymbol{\eta}_{t_0}^{(s)}, \boldsymbol{\eta}_{t_{-100}}^{(s)}, \boldsymbol{\eta}_{t_{-200}}^{(s)})\Delta t + g_m \Delta W_{100}^{(2)}, \\
&\vdots \\
X^{1,(m)}(5) \simeq \hat{\boldsymbol{x}}_{t_{N_{\text{ref}}}}^{(s)} &= \hat{\boldsymbol{x}}_{t_{400}}^{(s)} + f_m(t_{N_{\text{ref}}}, \hat{\boldsymbol{x}}_{t_{400}}^{(s)}, \ldots, \hat{\boldsymbol{x}}_{t_{100}}^{(s)})\Delta t + g_m \Delta W_{100}^{(N_{100})},
\end{aligned}
$$

其中 $g_m \coloneqq g_m(t_k, \boldsymbol{\eta}_{t_0}^{(s)}, \boldsymbol{\eta}_{t_{-100}}^{(s)}, \boldsymbol{\eta}_{t_{-200}}^{(s)}, \boldsymbol{\eta}_{t_{-300}}^{(s)}),\ k = 100, 200, \ldots, N_{\text{ref}}, \Delta t = 1$ 和 $\Delta W_{100}^{(k)}, \tilde{k} = 1, \ldots, 5$ 根据公式（29）定义。注意，$\Delta_{X^{\pi,(m)}}$ 的定义与公式（28）中相同。

A.3 模型参数

最终调优　　用于数值收敛性研究的 Delay-SDE-net 的参数见表　　5.

参数	值
学习率 f_m	0.01
学习率 $g_{a,\ m}$	0.05
迭代次数 f_m	30
迭代次数 $g_{a,\ m}$	22
动量	0.9
权重衰减	0.00005
优化器	SDG

表5：5.1.2 节中模拟数据所用 Delay-SDE-net 的调优参数。

B 比较研究

B.1 分布外数据

其中 th 所在的区间　　测试集 $\mathcal{D}$ 被替换为分布外数据 $\mathcal{D}$，第 5.1.3 节中的结果见表 6。

Year	1	1	3	3	4	5	8	8	9	9
开始 t_k	23	313	79	344	275	67	1	190	48	323
结束 t_k	25	327	91	364	294	71	5	197	52	333

表 6：测试集 $\mathcal{D}$ 被替换为以下值的区间 $\tilde{\mathcal{D}}$，基于模拟数据的比较研究中模拟数据

B.2 模型参数

模拟数据比较研究中 Delay-SDE-net 和 SDE-net 的最终调优参数分别见表 7 和表 8。两个神经网络模型的一些共享参数见表 9。

	学习率 f_m	学习率 $g_{a,m}$	学习率 $g_{e,m}$	$\#f_m$	编号$g_{a,m}$	编号$g_{e,m}$
$t+1$	0.01	0.001	0.005	500	500	20
$t+2$	0.005	0.0001		500	500	
$t+3$	0.001	0.00005		500	500	
$t+4$	0.001	0.000005		500	63	
$t+5$	0.0005	0.000005		500	261	
$t+6$	0.0005	0.000005		500	279	
$t+7$	0.0005	0.000005		500	303	

表 7：调参用于 5.1.3 中模拟数据的 Delay-SDE-net。*lr* 表示学习率，# 表示训练迭代次数。

	学习率 f_m	学习率 g_m	#
$t+1$	0.005	0.005	500
$t+2$	0.001	0.001	500
$t+3$	0.001	0.001	500
$t+4$	0.001	0.001	500
$t+5$	0.0005	0.0005	500
$t+6$	0.0005	0.0005	500
$t+7$	0.0005	0.0005	500

表 8：5.1.3 中模拟数据所用 SDE-net 的调优参数。学习率 是学习率，# 是训练迭代次数。

参数	值
动量	0.9
权重衰减	0.00005
优化器	SGD

表 9：5.1.3节模拟数据和5.2.2节平流层数据所用的 Delay-SDE-net 和 SDE-net 的参数。这些参数未经调优，而是预先确定的。

B.3 向量自回归模型参数

模拟数据比较研究所用的 VAR模型参数见表10。仿真研究中拟合到 VAR模型残差的波动率函数，与 Eggen（2022）中使用的函数不同。也就是说，令 y表示向量自回归模型残差的平方的期望值，令 x 表示一年中的日期。那么，拟合得到的方差函数为

$$y = e^{a+bx},$$

其中 a 和 b 为常数。每个维度的最优参数 a 和 b 通过最小二乘多项式拟合得到。参见表11。

$\hat{\phi}_1$			$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$
	$X^{1,\pi}$	$X^{2,\pi}$			
$X^{1,\pi}$	1.19	−0.0062	$\begin{bmatrix} 0.0015 & 0.0087 \\ 0.028 & -0.077 \end{bmatrix}$	$\begin{bmatrix} -0.13 & -0.0015 \\ 0.31 & 0.025 \end{bmatrix}$	$\begin{bmatrix} -0.17 & -0.0016 \\ 0.39 & -0.010 \end{bmatrix}$
$X^{2,\pi}$	−0.37	1.06			

表10：模拟数据 $X^{\pi} = [X^{1,\pi}, X^{2,\pi}]^{\top}$ 的二维动力系统的 VAR(4)模型参数（参见第 5.1.1 节）。ϕ_1 的标记位置同样适用于其他矩阵参数。

	$X^{1,\pi}$	$X^{2,\pi}$
a	1.0628	−4.3482
b	−0.01694	2.2593×10^{-5}

表11：针对模拟数据的向量自回归模型残差平方的期望值拟合的方差函数参数。

C 现实案例研究

C.1 模型参数

现实案例研究中 Delay-SDE-net 和 SDE-net 的最终调优参数分别列于表12和表13。两个神经网络模型的一些共享参数列于表 9，这些参数与模拟数据比较研究中的参数相同。

	lr f_m	学习率 $g_{a,m}$	学习率 $g_{e,m}$	$\#f_m$	训练迭代次数 $g_{a,m}$	训练迭代次数 $g_{e,m}$
$t+1$	0.05	0.01	0.01	500	486	20
$t+2$	0.01	0.01	0.01	466	114	20
$t+3$	0.01	0.01	0.01	500	229	20
$t+4$	0.01	0.01	0.01	314	63	20
$t+5$	0.005	0.05	0.01	215	189	20
$t+6$	0.005	0.01	0.01	312	500	20
$t+7$	0.005	0.01	0.01	213	500	20

表12：5.2.2节中平流层数据所用 Delay-SDE-net 的调优参数。学习率 指学习率，# 表示训练迭代次数。

	学习率 f_m	学习率 g_m	$\#$
$t+1$	0.0005	0.0005	500
$t+2$	0.001	0.001	27
$t+3$	0.005	0.005	26
$t+4$	0.005	0.005	20
$t+5$	0.001	0.001	46
$t+6$	0.001	0.001	52
$t+7$	0.001	0.001	56

表1：5.2.2节中用于平流层数据的 SDE-net 调优参数。学习率 表示学习率，# 表示训练迭代次数。

C.2 向量自回归模型参数

Eggen（2022）对拟合 U风和温度二维系统的向量自回归模型进行了详细介绍。有关本节所给模型参数的说明，请参阅所引参考文献。

参考文献

Abdar Moloud、*Pourpanah Farhad*、*Hussain Sadiq*、*Rezazadegan Dana*、*Liu Li*、*Ghavamzadeh Mohammad*、*Fieguth Paul*、*Cao Xiaochun*、*Khosravi Abbas*、*Acharya U Rajendra* 等。深度学习中不确定性量化综述：技术、应用与挑战 // Information Fusion。2021年。76。243–297。

c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}	c_{14}	c_{16}	c_{18}	c_{20}
11.47	0.41	22.71	0.20	0.83	0.67	1.048	1.13	0.82	0.60	0.58
c_1	c_3	c_5	c_7	c_9	c_{11}	c_{13}	c_{15}	c_{17}	c_{19}	c_{21}
−0.00018	−0.13	−4.14	−0.64	−0.040	−0.81	0.64	−0.18	−0.0035	0.49	0.38

表14：U风的季节性函数参数。

c_0	c_2	c_4	c_6	c_8	c_{10}	c_{12}	c_{14}	c_{16}	c_{18}	c_{20}
226.28	−0.074	−12.09	0.20	1.80	0.11	0.033	−0.13	−0.12	−0.23	0.093
c_1	c_3	c_5	c_7	c_9	c_{11}	c_{13}	c_{15}	c_{17}	c_{19}	c_{21}
−0.00010	−0.088	1.57	−0.050	2.80	0.14	1.40	0.14	0.19	−0.040	−0.099

表15：温度的季节性函数参数。

$\hat{\phi}_1$			$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$
	U风	温度			
U风	1.72	0.0054	$\begin{bmatrix} -1.03 & -0.024 \\ -0.082 & -0.73 \end{bmatrix}$	$\begin{bmatrix} 0.32 & -0.0002 \\ 0.11 & 0.27 \end{bmatrix}$	$\begin{bmatrix} -0.044 & 0.021 \\ 0.017 & -0.098 \end{bmatrix}$
温度	−0.10	1.53			

表16：U风与温度构成的二维动力系统的四阶向量自回归模型参数。ϕ_1 所用的 $\hat{}$ 位置标记同样适用于其他矩阵参数。

冬季/春季：$w^{(2)}_{0.30,2}$

$d^{(1)}_{0,2}$	$d^{(1)}_{1,2}$	$d^{(1)}_{2,2}$	$d^{(1)}_{3,2}$	$d^{(1)}_{4,2}$	−	−	−	−
−30.63	−339.16	1987.79	373.69	−957.80	−	−	−	−

夏季：$w^{(3)}_{0.05,2}$

$d^{(2)}_{0,2}$	$d^{(2)}_{1,2}$	$d^{(2)}_{2,2}$	$d^{(2)}_{3,2}$	$d^{(2)}_{4,2}$	$d^{(2)}_{5,2}$	$d^{(2)}_{6,2}$	−	−
55.88	−198.28	2460.26	−472.34	2420.60	611.73	−2390.93	−	−

秋季/冬季：$w^{(4)}_{0.05,2}$

$d^{(3)}_{0,2}$	$d^{(3)}_{1,2}$	$d^{(3)}_{2,2}$	$d^{(3)}_{3,2}$	$d^{(3)}_{4,2}$	$d^{(3)}_{5,2}$	$d^{(3)}_{6,2}$	$d^{(3)}_{7,2}$	$d^{(3)}_{8,2}$
128.42	104.79	22.64	37.76	24.94	−65.26	−9.86	−190.33	−96.95

表17：针对U风的向量自回归模型残差拟合的波动率函数参数。

冬季/春季：$w^{(2)}_{0.44,1}$

$d^{(1)}_{0,1}$	$d^{(1)}_{1,1}$	$d^{(1)}_{2,1}$	$d^{(1)}_{3,1}$	$d^{(1)}_{4,1}$
−1.16	−24.02	101.47	25.98	−47.80

夏季：$w^{(2)}_{2.0,1}$

$d^{(2)}_{0,1}$	$d^{(2)}_{1,1}$	$d^{(2)}_{2,1}$	$d^{(2)}_{3,1}$	$d^{(2)}_{4,1}$
0.089	0.10	0.021	0.033	0.014

秋季/冬季：$w^{(2)}_{0.44,1}$

$d^{(3)}_{0,1}$	$d^{(3)}_{1,1}$	$d^{(3)}_{2,1}$	$d^{(3)}_{3,1}$	$d^{(3)}_{4,1}$
16.58	−344.37	113.479	119.90	134.75

表18：拟合温度的VAR模型残差所得波动率函数的参数。

Andersen Torben G.、*Thyrsgaard Martin*、*Todorov Viktor*。日内波动率的时变周期性 // 美国统计学会会刊。2019年。114，528。1695–1707。

Antun Vegard, Renna Francesco, PoonClarice, Adcock Ben, Hansen Anders C。深度学习在图像重建中的不稳定性及人工智能的潜在代价 // 美国国家科学院院刊。2020年。117, 48. 30088–30095。*Basse-O'Connor Andreas, Nielsen Mikkel Slot, PedersenJan, RohdeVictor*。随机时滞微分方程及相关自回归模型 // 随机学。2020年。92, 3. 454–477。*Benth Fred Espen*。《随机分析期权理论：数理金融导论》。柏林，海德堡，2004年。*Benth Fred Espen, Detering Nils, Galimberti Luca*。弗雷歇空间中的神经网络 // 数学与人工智能年刊。2022年。1–29。*BlancE, CerannaL, HauchecorneA,Charlton-Perez A, Marchetti E, Evers L.G, Kvaerna T, Lastovicka J, EliassonL, Crosby N. B, Blanc-Benon P, Pichon A le, Brachet N, Pilger C, Keckhut P, Assink J. D, Smets P.S.M, Lee C.F,Kero J, Sindelarova T,Kämpfer N,Rüfenacht R,FargesT,Millet C, Näsholm S. P, Gibbons S. J, EspyP. J, HibbinsR. E, Heinrich P,Ripepe M, KhaykinS, Mze N, Chum J*。借助ARISE项目改进中层大气动力学的表征 // 地球物理学综述。2018年。39, 2. 171–225。*BorovykhAnastasia, Bohte Sander, Oosterlee Cornelis W*。利用卷积神经网络进行条件时间序列预测 // arXiv预印本 arXiv:1703.04691。2017年。*Box George EP, Jenkins Gwilym M, Reinsel Gregory C*。《时间序列分析：预测与控制》。2008年。*Brockwell P J*。CARMA过程理论与应用的最新成果 // 统计数理研究所年刊。2014年。66, 4. 647–685。*BrockwellP. J.,Davis R. A.* 《时间序列：理论与方法》。1991年。*Butler A H, Seidel D J, Hardiman S C, ButchartN,Birner T,MatchA*。突发平流层增温的定义 // 美国气象学会会刊。2015年。96, 11. 1913–1928。*Casella George, BergerRoger L*。《统计推断》。2002年。*Chen Ricky TQ,RubanovaYulia, Bettencourt Jesse, Duvenaud David K*。神经常微分方程 // 神经信息处理系统进展。2018 年。31。*Colbrook Matthew J, Antun Vegard, Hansen Anders C*。计算稳定且精确的神经网络之难：深度学习的障碍与Smale第18问题 // 美国国家科学院院刊。2022年。119, 12. e2107151119。*Corduas Marcella*。时间序列分析中的非线性检验 // 意大利统计学会期刊。1994年。3, 3. 291–313。*Denker John, LeCun Yann*。将神经网络输出层转换为概率分布 // 神经信息处理系统进展。1990年。3。*Der Kiureghian Armen, Ditlevsen Ove*。偶然不确定性还是认知不确定性？这有关系吗？ // 结构安全。2009。31, 2. 105–112。*E Weinan, Ma Chao, Wu Lei*。双层神经网络总体风险的先验估计 // 数学科学通讯。2019年。17, 5. 1407–1425。*E Weinan, Ma Chao, Wu Lei*。Barron空间与神经网络模型的流诱导函数空间 // 构造性逼近。2021年。55, 1. 369–406。*E Weinan, Wojtowytsch Stephan*。柯尔莫哥洛夫宽度衰减与机器学习中的劣质逼近器：浅层神经网络、随机特征模型与神经切线核 // 数学科学研究。2021年。8, 1。*E Weinan,Wojtowytsch Stephan*。Barron函数的表示公式与逐点性质 // 变分法与偏微分方程。2022a。61, 2。*E Weinan, WojtowytschStephan*。关于具有Barron数据的高维偏微分方程的一些观察 // 第2届数学与科学机器学习会议论文集。145。2022年8月16日至19日。253–269。（机器学习研究论文集）。*Eggen M. D., Dahl K. R., Näsholm S. P., MælandS.* 平流层温度的随机建模 // 数学地球科学。2022年。*Eggen Mari Dahl*。多元ARMA/CARMA变换关系。2022年。*Gawlikowski Jakob, Tassi Cedrique Rovile Njieutcheu, AliMohsin, Lee Jongseok, HumtMatthias, FengJianxiang,Kruspe Anna, Triebel Rudolph, Jung Peter, RoscherRibana, others* 。深度神经网络中不确定性的综述 // arXiv预印本 arXiv:2107

.03342. 2021.

Gu Jiuxiang、*Wang Zhenhua*、*Kuen Jason*、*Ma Lianyang*、*Shahroudy Amir*、*Shuai Bing*、*Liu Ting*、*Wang Xingxing*、*Wang Gang*、*Cai Jianfei* 等。《卷积神经网络的最新进展》//《模式识别》。2018。77。354–377。*Hayashi Kohei*、*Nakagawa Kei*。《分数阶 SDE-Net：生成具有长期记忆的时间序列数据》。2022。*Hitchcock P.*、*Simpson I. R.*《平流层突然增温的向下影响》//《大气科学杂志》。2014。71，10。3856–3876。*Hochreiter Sepp*、*Schmidhuber Jürgen*。《长短期记忆》//《神经计算》。1997。9，8。1735–1780。*Hora StephenC*。《概率引出中的偶然不确定性与认知不确定性：以危险废物管理为例》//《可靠性工程与系统安全》。1996。54，2-3。217–223。*Hornik Kurt*、*Stinchcombe Maxwell*、*White Halbert*。《多层前馈网络是通用逼近器》//《神经网络》。1989。2，5。359–366。*Hsu Yen-Chang*、*Shen Yilin*、*Jin Hongxia*、*Kira Zsolt*。《广义 ODIN：无需使用分布外数据进行训练即可检测分布外图像》//《IEEE/CVF 计算机视觉与模式识别会议论文集》。2020。10951–10960。*Hu Yaozhong*、*Mohammed Salah-Eldin A*、*Yan Feng*。《带记忆随机微分系统的离散时间近似》。2001年。*Hu Yaozhong*、*Mohammed Salah-Eldin A.*、*Yan Feng*。《随机时滞方程的离散时间近似：Milstein格式》//《概率年刊》。2004年。32，1A。265–314。*Hüllermeier Eyke*、*Waegeman Willem*。《机器学习中的偶然不确定性与认知不确定性：概念与方法导论》//机器学习。2021年。110，3。457–506。*Jeon Byounghyun*、*Seo Sung Won*、*Kim Jun Sik*。《期权隐含波动率的不确定性与波动率预测能力》//《期货市场杂志》。2020年。40，7。1109–1126。*Kabir HM Dipu*、*Khosravi Abbas*、*Hosen Mohammad Anwar*、*Nahavandi Saeid*。《基于神经网络的不确定性量化：方法与应用综述》//《IEEE Access》。2018年。6。36218–36234。*Kantz Holger*、*Schreiber Thomas*。《非线性时间序列分析》。2003年。2。*Karniadakis George Em*、*Kevrekidis Ioannis G*、*Lu Lu*、*Perdikaris Paris*、*Wang Sifan*、*Yang Liu*。《物理信息机器学习》//《自然综述：物理》。2021年。3，6。422–440。*Karpechko A.*、*Tummon F.*、世界气象组织秘书处。《平流层的气候可预测性》//《世界气象组织公报》。2016年。65，1。*Kashinath K*、*Mustafa M*、*Albert A*、*Wu JL*、*Jiang C*、*Esmaeilzadeh S*、*Azizzadenesheli K*、*Wang R*、*Chattopadhyay A*、*Singh A* 等。《物理信息机器学习：天气和气候建模案例研究》//《英国皇家学会哲学汇刊A辑》。2021年。379，2194。20200093。*Kim Ildoo*、*Kim Younghoon*、*Kim Sungwoong*。《用于测试时增强的学习损失》//《神经信息处理系统进展》。2020年。33。4163–4174。*Kloeden P. E.*、*Platen E.*《随机微分方程的数值解》。柏林，1992年。*Kong Lingkai*、*Sun Jimeng*、*Zhang Chao*。《SDE-net：为深度神经网络赋予不确定性估计能力》。2020年。*KrishnamurthyV*。《天气和气候的可预测性》//《地球与空间科学》。2019年。6，7。1043–1056。*Kurutach Thanard*、*Clavera Ignasi*、*Duan Yan*、*Tamar Aviv*、*Abbeel Pieter*。《模型集成信赖域策略优化》//arXiv预印本 arXiv:1802.10592。2018年。*Lambrou Antonis*、*Papadopoulos Harris*、*Gammerman Alex*。《利用进化算法为医学诊断提供可靠的置信度度量》//《IEEE生物医学信息技术汇刊》。2010年。15，1。93–99。*Lee Jinsol*、*AlRegib Ghassan*。《将梯度作为神经网络中的不确定性度量》//2020年IEEE国际图像处理会议（ICIP）。2020年。2416–2420。*Li Xuechen*、*Wong Ting-Kam Leonard*、*Chen Ricky T. Q.*、*Duvenaud David*。《随机微分方程的可扩展梯度》。2020a。*Li Zhong*、*Ma Chao*、*Wu Lei*。《使用基于路径的范数度量具有一般激活函数的神经网络复杂度》。2020b。*Lord Gabriel J.*、*Powell Catherine E.*、*Shardlow Tony*。《计算随机偏微分方程导论》。2014年。（剑桥应用数学丛书）。

Lu Zhou, Pu Hongming, Wang Feicheng, Hu Zhiqiang,WangLiwei. 神经网络的表达能力：从宽度的视角看 // 神经信息处理系统进展. 2017. 30.*MacKay David JC*. 反向传播网络的实用贝叶斯框架 // 神经计算. 1992. 4, 3. 448–472.*Marquardt T., Stelzer R.* 多元 CARMA 过程 // 随机过程及其应用. 2007. 117, 1. 96–120.*MidtfjordAliseDanielle, De BinRiccardo,HusebyArne Bang*. 更安全飞机着陆的决策支持系统：使用 XGBoost 和可解释人工智能预测跑道状况 // 寒区科学与技术. 2022. 199. 103556.*Mohammed Salah-Eldin A.* 带记忆的随机微分系统：理论、示例与应用 // 随机分析及相关主题 VI. 马萨诸塞州波士顿：Birkhäuser Boston, 1998. 1–77.*Moller Felix, Botache Diego, Huseljic Denis, Heidecker Florian, BieshaarMaarten, Sick Bernhard*. 使用软布朗偏移采样和自编码器进行分布外检测与生成 // IEEE/CVF 计算机视觉与模式识别会议论文集. 2021. 46–55.*Mo˙zejko Marcin, SusikMateusz, Karczewski Rafał*. 用于神经网络不确定性估计的抑制型 softmax // arXiv预印本 arXiv:1810.01861. 2018.*Nandy Jay, Hsu Wynne, Lee Mong Li*. 力求最大化分布内与分布外样本之间的表征差距 // 神经信息处理系统进展. 2020. 33. 9239–9250.*OberdiekPhilipp, Rottmann Matthias, Gottschalk Hanno*. 基于梯度信息的深度神经网络分类不确定性 // IAPR 模式识别人工神经网络研讨会. 2018. 113–125.*Papernot Nicolas, McDaniel Patrick*. 深度 k 近邻：迈向可信、可解释且鲁棒的深度学习 // arXiv 预印本 arXiv:1803.04765. 2018.*PedatellaN, Chau J, Schmidt H, Goncharenko L, Stolle C, Hocke K, Harvey V, Funke B, Siddiqui T*. 平流层突然增温如何影响整个大气层 // Eos. 2018. 99.*PereiraAna, Thomas Carsten*. 机器学习应用于安全关键型信息物理系统所面临的挑战 // 机器学习与知识提取. 2020. 2, 4. 579–602.*Protter Philip*. 随机积分与微分方程. 柏林，海德堡，2005.*Raghu Maithra, BlumerKaty,SayresRory, Obermeyer Ziad, Kleinberg Bobby, Mullainathan Sendhil, Kleinberg Jon*. 面向医疗第二诊疗意见的直接不确定性预测 // 国际机器学习会议. 2019. 5281–5290.*Raissi Maziar, Perdikaris Paris, Karniadakis George E*. 物理信息神经网络：用于求解涉及非线性偏微分方程的正问题与逆问题的深度学习框架 //《计算物理学杂志》. 2019. 378. 686–707.*Rajeswaran Aravind, Ghotra Sarvjeet,Ravindran Balaraman, Levine Sergey*. Epopt：使用模型集成学习鲁棒神经网络策略 // arXiv预印本 arXiv:1610.01283. 2016.*Ramalho Tiago, Miranda Miguel*. 在表征空间中进行密度估计以预测模型不确定性 // 可信与安全机器学习系统工程国际研讨会. 2020. 84–96.*Rumelhart DavidE, Hinton Geoffrey E, Williams Ronald J*. 通过反向传播误差学习表征 //《自然》. 1986年. 323, 6088. 533–536.*Scaife A. A., Baldwin M. P., Butler A. H., Charlton-Perez A. J., Domeisen D. I. V., Garfinkel C. I., Hardiman S. C., Haynes P., Karpechko A. Y., Lim E.-P., Noguchi S., Perlwitz J., Polvani L., Richter J. H., Scinocca J., Sigmond M., Shepherd T. G., Son S.-W., Thompson D. W. J.* 平流层与长期预测 // 大气化学与物理. 2022年. 22, 4. 2601–2623.*Scher Sebastian, MessoriGabriele*. 使用机器学习预测天气预报不确定性 // 英国皇家气象学会季刊. 2018年. 144, 717. 2830–2841.*Sluijterman Laurens, Cator Eric, Heskes Tom*. 如何评估机器学习回归任务中的不确定性估计？ // arXiv预印本 arXiv:2106.03395. 2021年.*Tambon Florian,Laberge Gabriel, An Le, Nikanjam Amin, Mindom Paulina Stevia Nouwou, Pequignot Yann, KhomhFoutse, Antoniol Giulio, MerloEttore, Laviolette François*. 如何认证基于机器学习的安全关键型系统? 系统性文献综述 // 自动化软件工程. 2022年. 29, 2. 1–74.*Teräsvirta Timo, Tjøstheim Dag, Granger Clive W.J.* 非线性经济时间序列建模. 2010年12月.

Tzen Belinda, Raginsky Maxim. 神经随机微分方程：扩散极限下的深度潜在高斯模型. 2019年.*VarshneyKush R*. 机器学习中的安全工程 // 2016年信息论与应用研讨会 (ITA). 2016年. 1–5.*Waibel Alexander, HanazawaToshiyuki, Hinton Geoffrey, Shikano Kiyohiro, Lang Kevin J*. 使用时延神经网络进行音素识别 // IEEE 声学、语音与信号处理汇刊. 1989. 37, 3. 328–339.*Wainberg Michael, Merico Daniele, DelongAndrew, Frey BrendanJ*. 生物医学中的深度学习 // 《自然·生物技术》. 2018年. 36, 9. 829–838.*Wang Bin, Lu Jie, Yan Zheng, LuoHuaishao, Li Tianrui, Zheng Yu, Zhang Guangquan*. 深度不确定性量化：用于天气预报的机器学习方法 // 第25届 ACM SIGKDD知识发现与数据挖掘国际会议论文集. 2019a. 2087–2095.*WangGuotai, Li Wenqi, Aertsen Michael, Deprest Jan, Ourselin Sébastien, Vercauteren Tom*. 使用测试时增强和卷积神经网络进行医学图像分割的偶然不确定性估计 // 《神经计算》. 2019b. 338. 34–45.*WangYongGuang, Tan HuoBin, Yao ShuZhen*. 曲线随机微分方程网络有助于更好地泛化深度神经网络的不确定性估计 // 人工神经网络与机器学习——ICANN 2021. Cham：施普林格国际出版, 2021. 248–259.*Wang Yongguang, Yao Shuzhen*. 用于深度学习不确定性估计的神经过程家族成员与神经随机微分方程 // 《传感器》. 2021年. 21, 11.*Wang Zhiguang, Oates Tim*. 将时间序列编码为图像，以便使用平铺卷积神经网络进行视觉检查与分类 // 第二十九届 AAAI人工智能会议研讨会论文集. 2015年.*YangLuxuan, GaoTing, LuYubin, Duan Jinqiao, Liu Tao*. 由莱维噪声驱动的集成随机微分方程时间序列预测 // CoRR. 2021年. abs/2111.13164.*Yang Yibo, Perdikaris Paris*. 物理信息神经网络中的对抗性不确定性量化 // 《计算物理学杂志》. 2019年. 394. 136–152.*Yuan Zhaolin, Li Xiaorui, Wu Di, Ban Xiaojuan,Wu Nai-Qi, Dai Hong-Ning, WangHao*. 使用微分常微分方程网络对工业浆料浓密机系统进行连续时间预测 // IEEE/CAA《自动化学报》. 2022年. 9, 4. 686–698.*Zhang Dayong, Hu Min, Ji Qiang*. COVID-19全球大流行下的金融市场 // 《金融研究快报》. 2020年. 36. 101528.*Zhang G Peter*. 使用 ARIMA 与神经网络混合模型进行时间序列预测 // 《神经计算》. 2003年. 50. 159–175.*Zhang Xiaoge, ChanFelix TS, Mahadevan Sankaran*. 图像分类模型中的可解释机器学习：不确定性量化视角 // 《知识型系统》. 2022 年 . 243. 108418.*Zheng Yi, Liu Qi, Chen Enhong, Ge Yong, Zhao JLeon*. 使用多通道深度卷积神经网络进行时间序列分类 // 国际网络时代信息管理会议. 2014年. 298–310.*Øksendal Bernt*. 随机微分方程：应用导论. 柏林，海德堡，2003年.