

毕业论文

基于物理先验与符号回归的气候驱动 登革热 SEIR 建模及多城市验证

Climate-Driven Dengue SEIR Modeling
Based on Physics-Informed Priors and Symbolic Regression
with Multi-City Validation

学 院： 公共卫生学院

专 业： 流行病与卫生统计学

研究方向： 传染病建模与预测

2025 年 6 月

摘 要

登革热是全球最严重的蚊媒传染病之一，中国南方尤其广东省是国内最主要的流行区域。理解气候因素如何驱动登革热传播效率，对于建立早期预警系统和制定精准防控策略具有重要意义。然而，传统机制模型往往依赖先验假设固定传播率函数形式，难以从数据中自动发现最优的气候–传播率关系；纯数据驱动的机器学习方法则存在可解释性不足和小样本坍塌的瓶颈。

本文提出一种“物理先验耦合 SEIR 动力学 + 符号回归蚊媒公式 + 多城市联合校准”的混合建模框架，旨在兼顾机制可解释性与空间泛化能力。该框架包含三个核心阶段：(1) 利用神经网络学习气候变量（温度 T 、降水 R 、相对湿度 H ）到蚊媒密度的映射，并通过符号回归蒸馏为可解释的蚊媒密度公式 \hat{M} ；(2) 采用 Brière 温度响应函数 $\beta'(T)$ 作为传播率的物理先验，耦合 SEIR 仓室模型进行病例预测；(3) 通过共享物理参数 + 逐城市输入率校准的策略，实现 16 城市的空间迁移。

研究分为两个部分。第一部分以广州市为核心开展单城市机制发现。首先以广东省 8 个有布雷图指数 (BI) 监测数据的城市为训练集 (306 个样本)，通过神经网络 + PySR 符号回归发现蚊媒密度公式 $\hat{M} = \frac{7.36\sqrt{T}}{T_m} - \frac{12.80}{H} - \frac{64.08}{H_m + \sqrt{R_m + R}}$ ，包含温度、湿度和降水三个气候因素，留一城市交叉验证 (LOCO CV) 均值 Pearson $r = 0.493$ ，8 个城市中 5 个 $r > 0.5$ 。在此基础上，采用 Brière 函数 $\beta'(T) = cT(T - T_{\min})\sqrt{T_{\max} - T}$ 建模传播率的温度依赖性，通过差分进化优化获得最优参数 ($T_{\min} = 15.1^\circ\text{C}$, $T_{\max} = 42.0^\circ\text{C}$, $T_{\text{opt}} = 35.5^\circ\text{C}$)。广州拟合 Spearman $\rho = 0.814$ 、对数 $R^2 = 0.851$ ；留一年交叉验证 (LOYO CV) 爆发年均值 $\rho = 0.806$ 。同时，以纯数据驱动的神经网络 β 方法作为对比实验，发现其在 180 个月度样本 (64% 为低/零病例月) 条件下发生模式坍塌 (β_{NN} 变异系数仅 1.17%)，反衬了物理先验方法在小样本传染病建模中的必要性。

第二部分开展多城市空间迁移与独立时间外验证。通过共享 Brière 参数 + 逐城市 η 校准迁移至广东省 16 城，月度均值 $\rho = 0.531$ 、 $R_{\log}^2 = 0.656$ 、WAPE = 0.827；2014 年度排名 $\rho = 0.947$ 。进一步利用 2020–2026 年广州新 BI 监测数据进行独立时间外验证，Brière $\beta'(T)$ 与观测 BI 的 Pearson $r = 0.782$ 、季节性相关 $r = 0.920$ ，证实模型的时间泛化能力。

本研究的主要创新包括：(1) 提出“Brière 物理先验 + PySR 蚊媒公式 + SEIR 动力学”的可解释混合建模框架；(2) 通过共享物理参数 + 逐城市校准的两阶段策略实现 16 城空间迁移；(3) 利用独立时间外数据 (2020–2026) 验证模型泛化能力；(4) 系统对比物理先验与纯数据驱动方法，揭示小样本条件下物理先验的优势。

关键词：登革热；SEIR 模型；Brière 函数；符号回归；蚊媒密度；广东省；多城市验证；时间外验证

Abstract

Dengue fever is one of the most severe mosquito-borne infectious diseases globally. Southern China, especially Guangdong Province, is the primary endemic region in the country. Understanding how climatic factors drive dengue transmission efficiency is crucial for establishing early-warning systems and formulating targeted control strategies. However, traditional mechanistic models rely on *a priori* assumptions to fix the functional form of the transmission rate, while purely data-driven machine learning approaches suffer from mode collapse under small-sample conditions.

This thesis proposes a hybrid modeling framework—“Physics-Informed Brière Function + Symbolic Regression Mosquito Formula + Multi-City Joint Calibration”—that balances mechanistic interpretability with spatial generalizability. The framework consists of three core stages: (1) learning the mapping from climate variables (temperature T , precipitation R , relative humidity H) to mosquito density via a neural network, followed by symbolic regression distillation to obtain an interpretable mosquito density formula \hat{M} ; (2) modeling the temperature dependence of the transmission rate using the Brière function $\beta'(T) = cT(T - T_{\min})\sqrt{T_{\max} - T}$, coupled with an SEIR compartmental model for case prediction; (3) transferring to 16 cities via shared physical parameters and per-city input rate calibration.

The study is organized into two parts. Part I focuses on single-city mechanism discovery centered on Guangzhou. First, 8 cities with Breteau Index (BI) monitoring data (306 samples) are used to discover the mosquito density formula $\hat{M} = \frac{7.36\sqrt{T}}{T_m} - \frac{12.80}{H} - \frac{64.08}{H_m + \sqrt{R_m + R}}$, incorporating temperature, humidity, and precipitation, through neural network + PySR symbolic regression, with leave-one-city-out cross-validation (LOCO CV) mean Pearson $r = 0.493$ (5/8 cities $r > 0.5$). The Brière function is then employed to model the temperature dependence of transmission rate. Differential evolution optimization yields optimal parameters ($T_{\min} = 15.1^\circ\text{C}$, $T_{\max} = 42.0^\circ\text{C}$, $T_{\text{opt}} = 35.5^\circ\text{C}$). Guangzhou fitting achieves Spearman $\rho = 0.814$ and log-scale $R^2 = 0.851$; leave-one-year-out cross-validation (LOYO CV) for outbreak years yields mean $\rho = 0.806$. A purely data-driven neural network β approach is presented as a comparative experiment, revealing mode collapse (β_{NN} coefficient of variation = 1.17%) under 180 monthly samples with 64% low/zero-case months, highlighting the necessity of physics-informed approaches for small-sample infectious disease modeling.

Part II performs multi-city spatial transfer and independent temporal-external validation. Transfer to 16 cities via shared Brière parameters + per-city η calibration achieves monthly mean $\rho = 0.531$, $R_{\log}^2 = 0.656$, WAPE = 0.827; 2014 annual ranking $\rho = 0.947$. Independent temporal-external validation using new Guangzhou BI monitoring data from 2020–2026 shows that the Brière $\beta'(T)$ correlates with observed BI at Pearson $r = 0.782$ and seasonal $r = 0.920$, confirming temporal generalizability.

Key innovations include: (1) a “Brière physics prior + PySR mosquito formula + SEIR dynamics” interpretable hybrid framework; (2) a two-stage transfer strategy with shared physical parameters and per-city calibration for 16-city spatial generalization; (3) independent temporal-external validation using 2020–2026 data; (4) systematic comparison between physics-informed and purely data-driven approaches, revealing the advantage of physical priors under small-sample conditions.

Keywords: Dengue fever; SEIR model; Brière function; Symbolic regression; Mosquito density; Guangdong Province; Multi-city validation; Temporal-external validation

目录

| | |
|--------------------------|----|
| 摘要 | I |
| Abstract | II |
| 1 前言 | 1 |
| 1.1 登革热全球与中国流行概况 | 1 |
| 1.2 气候因素与蚊媒传播 | 1 |
| 1.3 蚊媒密度监测与布雷图指数 | 2 |
| 1.4 传染病动力学模型 | 3 |
| 1.5 传播率的温度依赖性: Brière 函数 | 4 |
| 1.6 符号回归与知识蒸馏 | 4 |
| 1.7 物理先验与数据驱动方法的对比 | 5 |
| 1.8 研究目标与创新点 | 5 |
| 1.9 全文结构 | 6 |
| 2 第一部分: 单城市机制发现 (广州) | 7 |
| 2.1 引言 | 7 |
| 2.2 数据材料与方法 | 7 |
| 2.2.1 研究区域与数据来源 | 7 |
| 2.2.2 数据预处理 | 8 |
| 2.2.3 蚊媒密度公式发现 | 9 |
| 2.2.4 Brière 传播率模型 | 10 |
| 2.2.5 SEIR 离散时间预测框架 | 10 |
| 2.2.6 评估指标体系 | 11 |
| 2.2.7 纯数据驱动对比实验 | 12 |
| 2.3 结果 | 13 |
| 2.3.1 蚊媒密度公式 | 13 |
| 2.3.2 Brière 传播率参数 | 14 |
| 2.3.3 广州拟合与交叉验证 | 15 |
| 2.3.4 NN β 对比实验 | 16 |
| 2.4 讨论 | 17 |
| 2.4.1 物理先验的必要性 | 17 |
| 2.4.2 蚊媒公式的改进与解释 | 18 |
| 2.4.3 极端年份分析 | 18 |
| 2.4.4 方法论意义 | 19 |

| | |
|--------------------------|-----------|
| 目录 | V |
| 3 第二部分：多城市迁移与外部验证 | 19 |
| 3.1 引言 | 19 |
| 3.2 数据与方法 | 20 |
| 3.2.1 16 城数据概况 | 20 |
| 3.2.2 两阶段迁移策略 | 20 |
| 3.2.3 量级校准方法 | 21 |
| 3.2.4 评估策略 | 22 |
| 3.2.5 外部时间验证数据 | 22 |
| 3.3 结果 | 23 |
| 3.3.1 逐城市 η 校准结果 | 23 |
| 3.3.2 16 城月度预测 | 23 |
| 3.3.3 年度排名验证 | 25 |
| 3.3.4 改进前后对比 | 26 |
| 3.3.5 低发病城市分析 | 27 |
| 3.3.6 外部时间验证 | 27 |
| 3.4 讨论 | 29 |
| 4 总结与展望 | 31 |
| 4.1 主要结论 | 31 |
| 4.2 创新点总结 | 32 |
| 4.3 局限性分析 | 33 |
| 4.4 未来展望 | 34 |
| 参考文献 | 36 |
| 致谢 | 41 |

1 前言

1.1 登革热全球与中国流行概况

登革热 (Dengue Fever) 是由登革病毒 (DENV, 分为 DENV-1 至 DENV-4 四个血清型) 引起的急性蚊媒传染病, 主要通过埃及伊蚊 (*Aedes aegypti*) 和白纹伊蚊 (*Aedes albopictus*) 叮咬传播。Bhatt 等^[2] 估计全球每年约 3.9 亿登革病毒感染, 其中约 9600 万表现出临床症状。世界卫生组织^[3] 指出, 过去二十年间登革热报告病例数增长了八倍以上, 从 2000 年的 50 万例升至 2023 年的超过 600 万例, 已成为热带和亚热带地区最严峻的公共卫生挑战之一。Messina 等^[1] 利用全球尺度的统计模型预测, 到 2080 年气候变化和城市化将使全球约 63 亿人面临登革热风险, 较 2015 年增加约 22 亿人。

中国大陆的登革热疫情以境外输入和本地暴发交替为特征, 流行区域主要集中在广东、云南、浙江等南方省份^[5]。其中广东省是中国大陆登革热负担最重的省份, 历史上多次出现大规模暴发: 2014 年广东省报告超过 45,000 例, 广州市单城报告逾 37,000 例, 创下历史纪录^[6]。Yue 等^[4] 的系统综述表明, 自 2004 年以来广东省贡献了全国超过 70% 的登革热报告病例, 年度病例数呈波动性上升趋势, 暴发间隔呈缩短趋势。这一流行模式与该地区亚热带季风气候、高度城市化、人口密集以及频繁的国际人员流动密切相关。

从血清型分布来看, 广东省历年暴发中 DENV-1 最为常见, 但也检测到 DENV-2、DENV-3 和 DENV-4 的输入性和本地传播病例^[6]。值得注意的是, 由于不同血清型之间仅存在短暂的交叉免疫保护, 二次感染可能导致更严重的登革出血热 (DHF) 和登革休克综合征 (DSS), 给公共卫生系统带来额外压力^[7]。

近年来, 全球气候变暖、极端天气事件增加以及“一带一路”倡议下国际贸易与旅游的扩大, 进一步加剧了登革热北扩和暴发频次增加的风险。DeSouza 等^[8] 指出, 2023–2024 年全球登革热病例再创历史新高, 部分与厄尔尼诺现象引发的异常高温和强降水有关。在中国语境下, 气候变化、城市扩张和蚊媒分布北移的叠加效应使得登革热防控面临前所未有的挑战, 亟需建立基于机制理解的精准预测与早期预警体系。

1.2 气候因素与蚊媒传播

登革热的传播强度受到多种气候因素的共同影响, 其中温度、降水和湿度是最关键的三个变量。这些气候因素通过影响蚊媒的生存、繁殖、发育速率以及病毒在蚊体内的外潜伏期 (Extrinsic Incubation Period, EIP), 从而间接调控病毒在人群中的传播效率。

温度是影响登革热传播的核心气候因素。Mordecai 等^[12] 的全面实验研究表明, 蚊媒传播能力 (以基本再生数 R_0 的组分度量) 对温度呈单峰响应, 最优传播温度约为 29°C。在此温度下, 蚊虫叮咬率最高、病毒外潜伏期最短、蚊虫存活率最大, 三者的乘积效应使传播效率达到峰值。当温度低于约 18°C 或高于约 34°C 时, 传播能力显著下降^[34]。Shapiro 等^[9] 和 Lambrechts 等^[10] 进一步指出, 温度日较差 (Diurnal Temperature

Range, DTR) 对传播效率也有重要影响: 在接近最适温度时, 较大的 DTR 会降低传播效率; 而在偏低温度条件下, 适度的 DTR 反而可能增强传播。Kamiya 等^[11] 的荟萃分析确认了温度对蚊媒传染病传播的非线性调控作用在全球不同地理区域具有一致性。

降水对登革热的影响具有双重性。一方面, 适量降水为伊蚊提供了繁殖所需的积水容器 (如废弃轮胎、花盆托盘、建筑工地积水等), 增加幼虫孳生地数量, 从而提高成蚊密度^[37]。另一方面, 极端强降水可能冲刷幼虫孳生地, 导致蚊虫密度短暂下降^[8]。Xu 等^[64] 对广州的研究发现, 降水对登革热的影响存在约 1–2 个月的滞后效应, 反映了从降水到积水形成、幼虫发育、成蚊羽化再到病毒传播的完整生态链条。此外, 降水的时间分布模式 (如连续小雨 vs. 间歇性暴雨) 对蚊虫孳生的影响也存在差异, 但目前的月度分辨率数据难以捕捉这种细粒度效应。

相对湿度主要通过影响成蚊存活率和飞行活动能力来调控传播效率。高湿度环境有利于蚊虫存活和活动, 而低湿度条件下蚊虫脱水死亡率增加^[64]。Xu 等^[64] 的研究表明, 广州地区相对湿度与登革热发病率呈正相关, 但其独立效应弱于温度。值得注意的是, 温度、降水和湿度之间存在较强的共线性 (如高温季节通常伴随高湿度和多降水), 这给单独量化各因素的独立贡献带来了统计学挑战。

除上述三个核心气候变量外, 风速、日照时数、蒸发量等因素也可能影响蚊媒活动和病毒传播, 但其效应相对较弱且研究证据不够充分。本研究聚焦于温度、降水和相对湿度三个变量, 这与大多数登革热气候建模研究的变量选择一致。

值得强调的是, 气候因素对登革热传播的影响并非简单的线性关系, 而是通过复杂的非线性机制发挥作用。首先, 温度对蚊虫生命周期各阶段的影响存在不同的最适温度和阈值温度, 这些效应的叠加产生了传播效率对温度的非对称单峰响应。其次, 气候因素之间存在交互效应——例如, 高温高湿条件下蚊虫活动最为活跃, 而高温低湿条件下蚊虫存活率显著下降。第三, 气候因素对传播效率的影响存在时间滞后, 从气候条件变化到蚊虫种群响应再到人群感染, 通常需要数周到数月的时间。这些非线性特征使得传统的线性统计模型难以准确捕捉气候–传播率关系, 需要引入具有非线性建模能力的方法。

此外, 城市热岛效应对登革热传播也有重要影响。城市中心区域的温度通常比郊区高 2–5°C, 这种温度差异可能导致城市中心的传播季节更长、传播效率更高。广州作为超大城市, 其热岛效应尤为显著, 这也是广州成为广东省登革热负担最重城市的原因之一。然而, 由于本研究使用的是气象站点数据而非空间分辨率更高的遥感温度数据, 城市内部的温度空间异质性未能被充分捕捉, 这是未来研究可以改进的方向。

1.3 蚊媒密度监测与布雷图指数

蚊媒密度是连接气候因素与登革热传播的关键中间变量。在实际监测中, 布雷图指数 (Breteau Index, BI) 是最广泛使用的伊蚊幼虫密度指标, 定义为每百户调查中发现的阳性容器数。世界卫生组织将 $BI \geq 20$ 作为登革热暴发风险的警戒阈值。

然而, BI 数据在实际应用中面临诸多挑战。首先, BI 监测需要大量人力进行入户调查, 覆盖范围和时间连续性受限。在中国, 仅部分城市建立了系统性的 BI 监测网络, 且监测频率和方法标准在不同城市间存在差异。其次, BI 反映的是幼虫密度而非成蚊密度, 两者之间的转化受到温度、容器类型、天敌等多种因素的影响。第三, 不同城市的 BI 基线水平差异显著——本研究中 8 个有 BI 数据的城市, BI 均值从深圳的 2.2 到揭阳的 13.1 不等, 最大/最小比达 5.9 倍, 这种差异反映了城市化水平、居住环境、防控力度等非气候因素的影响。

鉴于 BI 数据的稀缺性和异质性, 本研究采用两种互补策略: (1) 利用有 BI 数据的 8 个城市训练蚊媒密度的气候驱动公式, 通过城市内归一化消除基线差异; (2) 在传播率建模中采用具有昆虫学物理先验的 Brière 函数, 减少对 BI 数据的直接依赖。

1.4 传染病动力学模型

仓室模型 (Compartmental Models) 是传染病动力学建模的经典框架。其核心思想是将人群划分为若干互斥的流行病学状态 (仓室), 并用微分方程描述个体在仓室间的转移速率。对于登革热, 最常用的是 SEIR 模型, 将人群分为易感者 (Susceptible, S)、潜伏者 (Exposed, E)、感染者 (Infectious, I) 和恢复者 (Recovered, R) 四个仓室。

经典 SEIR 模型的连续时间形式为:

$$\frac{dS}{dt} = -\beta SI/N, \quad \frac{dE}{dt} = \beta SI/N - \sigma E, \quad \frac{dI}{dt} = \sigma E - \gamma I, \quad \frac{dR}{dt} = \gamma I \quad (1)$$

其中 β 为传播率, σ 为潜伏期倒数 ($1/\sigma$ 为平均潜伏期), γ 为恢复率 ($1/\gamma$ 为平均感染期), N 为总人口。

在登革热建模中, 传播率 β 不是常数, 而是受气候因素调控的时变参数。如何建模 $\beta(t)$ 与气候变量的关系, 是本研究的核心科学问题。传统方法通常假设 β 为温度的某种参数化函数 (如高斯函数、多项式等), 但这种先验假设可能与真实的生物学机制不符。本研究采用 Brière 函数作为 $\beta(T)$ 的参数化形式, 该函数源于昆虫发育速率的温度依赖性理论, 具有明确的生物学基础。

对于月度分辨率的数据, 本研究采用离散时间 SEIR 框架, 避免了连续时间 ODE 在大时间步长下的数值不稳定性。具体而言, 每月的新增病例预测为:

$$\hat{C}_t = \beta'(T_t) \cdot \hat{M}_t \cdot S_t + \eta \quad (2)$$

其中 $\beta'(T_t)$ 为 Brière 温度响应函数, \hat{M}_t 为归一化蚊媒密度, S_t 为易感人群池, η 为背景输入率。

在参数估计方面, 传统方法通常采用最大似然估计 (MLE) 或马尔可夫链蒙特卡罗 (MCMC) 方法。然而, SEIR 模型的似然函数通常不具有解析形式, 且参数空间可能存在多个局部最优解。本研究采用差分进化 (Differential Evolution, DE) 算法进

行全局优化，该算法是一种基于种群的随机搜索方法，通过变异、交叉和选择操作在参数空间中进行高效搜索，能够有效避免陷入局部最优。DE 算法的优势在于：(1) 不需要目标函数的梯度信息，适用于非光滑、非凸的优化问题；(2) 通过种群多样性维持全局搜索能力；(3) 参数设置简单，仅需指定种群大小、变异因子和交叉概率。

目标函数的设计对参数估计的质量至关重要。本研究采用 Huber 损失与 Spearman 相关性惩罚的加权组合作为目标函数。Huber 损失在残差较小时等价于均方误差，在残差较大时等价于绝对误差，对异常值具有鲁棒性——这对于登革热数据尤为重要，因为 2014 年特大暴发的病例数远超其他年份，纯均方误差会导致优化过度偏向拟合 2014 年。Spearman 相关性惩罚则确保模型不仅在量级上接近观测值，还能正确捕捉病例的时序排名模式。

1.5 传播率的温度依赖性：Brière 函数

昆虫的发育速率、叮咬频率和存活率均对温度呈非线性响应。Brière 等^[65]提出了一种描述昆虫发育速率温度依赖性的经验函数：

$$r(T) = cT(T - T_{\min})\sqrt{T_{\max} - T} \quad (3)$$

其中 c 为尺度参数， T_{\min} 为发育温度下限， T_{\max} 为发育温度上限。该函数在 $T < T_{\min}$ 或 $T > T_{\max}$ 时取零值，在两者之间呈不对称的单峰形态，最优温度 T_{opt} 偏向 T_{\max} 一侧。

Brière 函数在蚊媒传染病建模中得到了广泛应用。Mordecai 等^[34]利用 Brière 函数拟合了白纹伊蚊和埃及伊蚊的多个生命史参数（叮咬率、产卵率、幼虫发育速率、成蚊存活率等）对温度的响应曲线，并将这些组分整合为温度依赖的基本再生数 $R_0(T)$ 。Johnson 等^[66]系统比较了 Brière 函数与其他温度响应函数（如 Lactin 函数、Sharpe-DeMichele 模型等），发现 Brière 函数在拟合精度和参数可解释性之间取得了良好的平衡。

在本研究中，Brière 函数被用于建模传播率 $\beta'(T)$ 的温度依赖性。与纯数据驱动的神经网络方法相比，Brière 函数具有以下优势：(1) 仅需 3 个参数 (c, T_{\min}, T_{\max})，在小样本条件下不易过拟合；(2) 参数具有明确的生物学含义，可与实验室数据交叉验证；(3) 函数形态由昆虫学理论约束，避免了数据驱动方法可能出现的非物理行为（如模式坍塌）。

1.6 符号回归与知识蒸馏

符号回归 (Symbolic Regression, SR) 是一种从数据中自动发现数学表达式的机器学习方法。与传统回归方法（预设函数形式、仅优化参数）不同，符号回归同时搜索函数的结构和参数，输出人类可读的闭合公式。Cranmer 等^[59]开发的 PySR 是目前最先进的符号回归工具之一，基于多种群遗传编程算法，支持自定义运算符和约束条

件。

在传染病建模领域，符号回归的应用尚处于起步阶段。Zhang 等^[48] 在 2024 年 *PLOS Computational Biology* 发表了将符号回归应用于传染病模型参数发现的开创性工作——通过将蚊媒种群动力学模型耦合神经网络，有效揭示了伊蚊产卵率和温度、降水之间的关系，并使用符号回归确定最优函数表达式。然而，该方法面临以下挑战：(1) 直接在高维表达式空间中搜索计算成本极高；(2) 缺乏利用先验物理知识引导搜索的机制。

本研究采用“神经网络预训练 + 符号回归蒸馏”的两阶段策略：首先用神经网络学习气候变量到蚊媒密度的非线性映射，然后在神经网络生成的蒸馏数据上运行符号回归。这种策略的优势在于：(1) 神经网络可以在有噪声的真实数据上学习平滑的映射关系；(2) 符号回归在干净的蒸馏数据上搜索，降低了搜索难度和计算成本；(3) 最终输出的闭合公式兼具数据适应性和可解释性。

1.7 物理先验与数据驱动方法的对比

在传染病建模中，物理先验方法 (Physics-Informed) 和纯数据驱动方法 (Data-Driven) 代表了两种不同的建模哲学。物理先验方法利用已知的生物学、流行病学知识约束模型结构，减少需要从数据中学习的自由度；纯数据驱动方法则让模型自由学习数据中的模式，不施加先验约束。

Raissi 等^[67] 提出的物理信息神经网络 (Physics-Informed Neural Networks, PINNs) 是将物理先验融入深度学习的代表性工作，通过在损失函数中加入物理方程的残差项，使神经网络的预测满足已知的物理定律。在传染病建模中，类似的思路已被用于将 SEIR 动力学约束融入神经网络训练^[68]。

然而，纯数据驱动方法在小样本条件下面临严峻挑战。当训练数据量有限且分布高度偏斜时（如登革热月度数据中大量零病例月），神经网络容易出现“模式坍塌” (Mode Collapse) ——即模型退化为预测常数值，因为这在偏斜分布下可以最小化平均损失。Holm^[43] 在 *Science* 上指出，黑箱模型在科学应用中的可解释性问题不容忽视，尤其是在需要机制理解的公共卫生决策场景中。

本研究通过系统对比 Brière 物理先验方法和 NN 纯数据驱动方法，为传染病建模中的方法选择提供实证依据。具体而言，在相同的数据条件下（广州 180 个月度样本，64% 为低/零病例月），比较两种方法在拟合精度、交叉验证、空间迁移和时间泛化等维度上的表现差异。

1.8 研究目标与创新点

基于上述研究背景，本文的核心研究目标为：

1. 建立基于 Brière 温度响应函数的登革热传播率模型，利用物理先验约束传播率的函数形态，避免小样本条件下的模式坍塌问题。

2. 通过神经网络 + 符号回归的知识蒸馏策略，从气候数据中发现可解释的蚊媒密度公式，为传播率模型提供蚊媒密度输入。
3. 设计共享物理参数 + 逐城市校准的两阶段迁移策略，将单城市模型推广至广东省 16 个地级市，验证模型的空间泛化能力。
4. 利用 2020–2026 年独立时间外数据验证模型的时间泛化能力，评估模型在训练期外的预测稳定性。
5. 系统对比物理先验方法与纯数据驱动方法，揭示小样本传染病建模中物理先验的优势。

本研究的主要创新点包括：

1. **物理先验混合建模框架**：提出“Brière 物理先验 + PySR 蚊媒公式 + SEIR 动力学”的可解释混合框架，克服了传统 SEIR 模型依赖先验函数形式的局限，同时避免了纯数据驱动方法的模式坍缩问题。
2. **可解释蚊媒密度公式**：通过 NN+PySR 符号蒸馏从 8 城 BI 监测数据中发现包含温度、湿度和降水三因素的蚊媒密度公式，采用城市内归一化策略消除城市间基线差异，使公式聚焦于气候驱动的季节性相对变化，LOCO CV 均值 Pearson $r = 0.493$ ，5/8 城市 $r > 0.5$ 。
3. **两阶段空间迁移策略**：共享 Brière 物理参数（蚊虫温度生理特性）+ 逐城市 η 校准（非气候因素差异），在保持物理一致性的同时适应城市间异质性。
4. **独立时间外验证**：首次利用 2020–2026 年新 BI 监测数据对登革热传播率模型进行独立时间外验证，季节性相关 $r = 0.920$ 。
5. **物理先验 vs. 数据驱动的实证对比**：系统揭示 NN 在 180 个月度样本条件下的模式坍缩现象（变异系数 1.17%），为传染病建模方法选择提供实证依据。

1.9 全文结构

本文其余部分组织如下：

第二章（第一部分：单城市机制发现）以广州市为核心案例，详细阐述数据来源与预处理方法、蚊媒密度公式发现流程（NN+PySR 符号回归）、Brière 传播率模型构建与参数优化、SEIR 离散时间预测框架，以及 NN 纯数据驱动方法的对比实验。呈现蚊媒公式发现、Brière 参数估计、广州拟合与 LOYO 交叉验证、NN 模式坍缩分析等结果，并进行深入讨论。

第三章（第二部分：多城市迁移与外部验证）将广州发现的 Brière 模型迁移至广东省 16 个地级市，介绍共享物理参数 + 逐城市 η 校准的两阶段迁移策略，呈现 16 城

月度预测、年度排名验证和改进前后对比结果。同时利用 2020–2026 年新 BI 数据进行独立时间外验证，包括逐年相关性分析和 MOI 补充分析。

第四章（总结与展望） 总结主要发现和创新点，讨论研究局限性（蚊媒数据稀缺、低发病城市预测受限、空间耦合缺失等），提出未来改进方向（遥感蚊媒估计、零膨胀模型、气候变化情景预测等）。

2 第一部分：单城市机制发现（广州）

2.1 引言

登革热传播是一个涉及病毒、蚊媒、人群和环境的复杂多因素过程。在构建跨城市的空间迁移模型之前，有必要首先在单一城市中深入理解传播机制，建立可靠的基础模型，再将其推广至更大的空间尺度。本章选择广州市作为核心研究案例，原因如下：第一，广州是中国大陆登革热负担最重的城市，2005–2019 年累计报告病例数居全国首位，其中 2014 年单年暴发超过 37,000 例^[6]，提供了丰富的疫情动态信息；第二，广州地处亚热带季风气候区，温度、降水和湿度的季节性变化显著，为研究气候驱动的传播机制提供了理想的自然实验条件；第三，广州及周边城市拥有相对完善的布雷图指数（BI）监测数据，为蚊媒密度建模提供了不可或缺的训练数据。

本章的核心任务包括三个层面：（1）从气候数据中发现可解释的蚊媒密度公式，建立气候变量到蚊媒活动强度的定量映射；（2）利用 Brière 物理先验函数建模传播率的温度依赖性，构建离散时间 SEIR 预测框架；（3）通过与纯数据驱动的神经网络方法进行系统对比，揭示物理先验在小样本传染病建模中的必要性。这三个层面相互关联、逐层递进，共同构成了从数据到机制、从拟合到验证的完整研究链条。

2.2 数据材料与方法

2.2.1 研究区域与数据来源

广州市（23.13°N, 113.26°E）位于广东省中南部、珠江三角洲北缘，属南亚热带季风气候，年均气温 21.4–22.6°C，年均降水量 1600–1900 mm，年均相对湿度约 77%。全市常住人口约 1,426 万人（ $N_h = 1.426 \times 10^7$ ），城市化率超过 86%，人口密度高、建成区面积大，为伊蚊孳生和登革热传播提供了有利的生态环境。广州市白纹伊蚊（*Aedes albopictus*）为优势蚊种，全年活动期约为 4–11 月，高峰期集中在 6–9 月^[6]。

本研究使用的数据来源包括以下四个方面：

（1）登革热病例数据。广州市 2005–2019 年月度登革热报告病例数据来源于中国疾病预防控制中心传染病报告信息系统。该数据集共包含 180 个月度观测值，呈现出极端的右偏分布特征：约 25% 的月份报告零病例，64.4% 的月份报告病例数 ≤ 5 例，而 2014 年暴发年全年累计报告 37,382 例，占 15 年总病例数的绝大部分。其余年份的年度病例数在 9 例（2005 年）至 1,690 例（2013 年）之间波动，年际变异系数极大。这种高度偏斜的分布特征对建模方法提出了严峻挑战——模型必须同时捕捉大量低

发病月份的基线水平和偶发暴发年份的极端峰值。

(2) 气象数据。月度气象数据来源于美国国家海洋和大气管理局（NOAA）全球地面日观测数据集（Global Surface Summary of the Day, GSOD），选取广州站（站号 59287）。提取的气象变量包括月均温度（ T , °C）、月均相对湿度（ H , %）和月累计降水量（ R , mm）。GSOD 数据经过质量控制，缺测率低于 2%，缺测值采用相邻月份线性插值填补。

(3) 布雷图指数数据。蚊媒密度数据来源于广东省 8 个城市（广州、深圳、珠海、佛山、东莞、中山、江门、揭阳）的布雷图指数（BI）监测记录，共计 306 个有效样本。各城市 BI 均值差异显著：深圳最低（均值 2.2），揭阳最高（均值 13.1），最大/最小比达 5.9 倍。这种城市间差异反映了不同城市在城市化水平、居住环境类型、蚊媒防控力度等方面的系统性差异，而非单纯的气候效应。

(4) 人口数据。广州市常住人口数据来源于广东省统计年鉴，用于 SEIR 模型中易感人群池的初始化。考虑到登革热在中国大陆以输入性和散发性传播为主，人群免疫水平较低，本研究假设初始易感比例接近 100%。

2.2.2 数据预处理

为确保不同量纲的气象变量在模型训练中具有可比性，本研究对所有气象输入变量进行 Min-Max 归一化处理。对于任意变量 x ，归一化公式为：

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (4)$$

其中 x_{\min} 和 x_{\max} 分别为该变量在训练集中的最小值和最大值。归一化后所有变量的取值范围为 $[0, 1]$ ，消除了量纲差异对梯度下降优化过程的不利影响。

对于布雷图指数（BI），本研究提出城市内归一化策略以消除城市间基线差异。如前所述，8 个城市的 BI 均值从 2.2 到 13.1 不等，最大/最小比达 5.9 倍。若直接使用原始 BI 值作为建模目标，模型将不可避免地学习到城市间的系统性偏差，而非气候驱动的蚊媒密度变化规律。为此，定义城市内归一化 BI（相对 BI）为：

$$\text{BI}_{\text{rel},i,t} = \frac{\text{BI}_{i,t}}{\overline{\text{BI}}_i} \quad (5)$$

其中 $\text{BI}_{i,t}$ 为城市 i 在时间 t 的原始 BI 值， $\overline{\text{BI}}_i$ 为城市 i 的 BI 时间均值。归一化后， BI_{rel} 反映的是各城市相对于自身基线水平的蚊媒密度波动，消除了城市间因非气候因素（如城市化水平、防控力度、居住环境类型等）导致的绝对水平差异。由于 8 城 BI 均值的最大/最小比达 5.9 倍（深圳 2.2 vs 揭阳 13.1），若直接以 BI 绝对值为建模目标，模型将不可避免地混淆气候效应与城市间的系统性尺度差异，导致跨城泛化困难。归一化使模型聚焦于气候驱动的季节性相对变化，是实现跨城泛化的关键前提。

此外，考虑到登革热病例数据的极端右偏分布（64.4% 的月份病例数 ≤ 5 ），本研

究在涉及病例数的建模和评估中广泛采用对数变换 $\log_{10}(1 + C)$ （即 \log_{1p} 变换），以压缩极端值的影响并改善残差的正态性。

2.2.3 蚊媒密度公式发现

蚊媒密度是连接气候因素与登革热传播的关键中间变量。由于 BI 监测数据在时间和空间上的稀缺性，本研究采用“神经网络预训练 + 符号回归蒸馏”的两阶段策略，从气候数据中发现可解释的蚊媒密度闭合公式。

第一阶段：神经网络预训练。构建一个三层多层感知机（MLP）网络，网络结构为 32-16-1（即输入层到第一隐藏层 32 个神经元，第二隐藏层 16 个神经元，输出层 1 个神经元），激活函数采用 ReLU。输入特征为 6 维向量，包括当月气象变量（温度 T 、相对湿度 H 、降水量 R ）及其对应的多年月均值（ \bar{T} 、 \bar{H} 、 \bar{R} ）。引入多年月均值的目的是为模型提供气候学基线信息，使其能够区分“绝对气象条件”和“相对于气候态的异常程度”。预测目标为 $\log(1 + \text{BI}_{\text{rel}})$ ，采用 \log_{1p} 变换以缓解目标变量的右偏分布。

训练配置如下：优化器采用 Adam，学习率 $\text{lr} = 5 \times 10^{-3}$ ，训练轮数 5000 个 epoch，不使用早停策略。损失函数采用 Huber 损失与相关性损失的加权组合：

$$\mathcal{L} = \mathcal{L}_{\text{Huber}}(\hat{y}, y) + \lambda \cdot (1 - r(\hat{y}, y)) \quad (6)$$

其中 $\mathcal{L}_{\text{Huber}}$ 为 Huber 损失（ $\delta = 1.0$ ）， $r(\hat{y}, y)$ 为预测值与观测值之间的 Pearson 相关系数， λ 为相关性损失权重。Huber 损失相比均方误差（MSE）对异常值更加鲁棒，适合处理 BI 数据中的极端值；相关性损失项鼓励模型捕捉预测值与观测值之间的线性趋势，而非仅最小化逐点误差。

第二阶段：符号回归蒸馏。利用训练好的神经网络在气象变量的均匀网格上生成 8000 个蒸馏数据点。具体而言，在温度、湿度和降水的归一化范围 $[0, 1]$ 内均匀采样，通过神经网络前向传播获得对应的蚊媒密度预测值。这些蒸馏数据点相比原始 306 个 BI 样本具有以下优势：（1）数据量增加约 26 倍，为符号回归提供了充足的搜索信号；（2）数据分布均匀，避免了原始数据在特定气象条件下的聚集效应；（3）经过神经网络的平滑化处理，降低了观测噪声对公式搜索的干扰。

在蒸馏数据上运行 PySR 符号回归^[59]，搜索空间包括基本算术运算（ $+$ ， $-$ ， \times ， \div ）和三角函数（ \sin ， \cos ）。PySR 采用多种群遗传编程算法，在公式复杂度和拟合精度之间进行 Pareto 前沿搜索。

第三阶段：公式选择。PySR 输出一组位于 Pareto 前沿上的候选公式，按复杂度从低到高排列。公式选择遵循奥卡姆剃刀原则——在拟合精度可接受的前提下，选择复杂度最低的公式。具体选择标准为：在 Pareto 前沿上，选取复杂度增加但 R^2 提升不显著（ $\Delta R^2 < 0.02$ ）的拐点处的公式。最终选定的公式将在原始 BI 数据上进行独立验证，并通过留一城市交叉验证（Leave-One-City-Out, LOCO CV）评估其跨城泛化

能力。

2.2.4 Brière 传播率模型

传播率 $\beta'(T)$ 的温度依赖性采用 Brière 函数建模。Brière 函数最初由 Brière 等^[65]提出，用于描述昆虫发育速率对温度的非线性响应，其数学形式为：

$$\beta'(T) = cT(T - T_{\min})\sqrt{T_{\max} - T}, \quad T_{\min} \leq T \leq T_{\max} \quad (7)$$

当 $T < T_{\min}$ 或 $T > T_{\max}$ 时， $\beta'(T) = 0$ 。该函数包含三个待估参数：尺度系数 c 、发育温度下限 T_{\min} 和发育温度上限 T_{\max} 。

Brière 函数的生物学基础源于昆虫变温动物的温度生理学。在低温端，当环境温度低于 T_{\min} 时，蚊虫的代谢活动、飞行能力和叮咬行为基本停止，病毒在蚊体内的复制也趋于停滞，因此传播率为零。在高温端，当温度超过 T_{\max} 时，蚊虫因热应激导致死亡率急剧上升，种群密度骤降，传播同样中断。在 T_{\min} 和 T_{\max} 之间，传播率呈不对称的单峰形态，最优温度 T_{opt} 偏向 T_{\max} 一侧，这与实验室观测到的蚊虫叮咬率、产卵率和病毒外潜伏期的温度响应曲线一致^[34, 12]。

选择 Brière 函数而非其他参数化形式（如高斯函数、多项式等）的理由在于：（1）Brière 函数的不对称单峰形态与蚊媒传播的温度响应实验数据高度吻合，而对称的高斯函数无法捕捉这种不对称性；（2）三个参数均具有明确的生物学含义，可与独立的实验室数据进行交叉验证——例如，白纹伊蚊的发育温度下限在文献中报道为 $14\text{--}16^\circ\text{C}$ ^[34, 57]，可作为 T_{\min} 估计值的合理性检验标准；（3）仅需 3 个自由参数，在 180 个月度样本的小样本条件下不易过拟合，这一点在与 7 维神经网络的对比实验中将得到充分体现。

2.2.5 SEIR 离散时间预测框架

基于 Brière 传播率模型和蚊媒密度公式，本研究构建离散时间 SEIR 预测框架。与连续时间 ODE 模型不同，离散时间框架直接在月度时间步长上进行递推，避免了大时间步长下 ODE 数值积分的不稳定性问题。

每月新增病例的预测公式为：

$$\hat{C}_t = \beta'(T_t) \cdot \hat{M}_t \cdot S_t + \eta \quad (8)$$

其中 $\beta'(T_t)$ 为 Brière 温度响应函数在 t 月均温下的取值， \hat{M}_t 为蚊媒密度公式的预测值， S_t 为 t 月初的易感人群池规模， η 为背景输入率（反映输入性病例和非气候因素的基线贡献）。

易感人群池的更新规则为：

$$S_{t+1} = S_t - \hat{C}_t + \delta \cdot N_h \quad (9)$$

其中 δ 为月度易感人群补充率，反映新生人口和免疫衰减的综合效应， N_h 为总人口。在登革热的流行病学背景下，由于中国大陆人群对登革病毒的免疫水平较低（非地方性流行区），且登革热四个血清型之间仅存在短暂的交叉免疫保护，因此易感人群池的消耗主要发生在暴发年份，而在低发病年份 S_t 基本维持在接近 N_h 的水平。

模型参数通过差分进化（Differential Evolution, DE）算法进行全局优化。DE 算法是一种基于种群的随机搜索算法，特别适合处理非凸、多模态的优化问题。优化目标为最小化预测病例数与观测病例数之间的对数尺度均方误差：

$$\min_{\theta} \sum_{t=1}^T \left[\log(1 + \hat{C}_t) - \log(1 + C_t) \right]^2 \quad (10)$$

其中 $\theta = \{c, T_{\min}, T_{\max}, \eta, S_0, \delta\}$ 为待优化参数向量。采用对数尺度目标函数的原因在于：原始病例数跨越多个数量级（从 0 到数千），若在原始尺度上优化，损失函数将被少数极端暴发月份主导，而忽略大量低发病月份的拟合质量。DE 算法的超参数设置为：种群规模为参数维度的 15 倍，最大迭代次数 $\text{maxiter} = 800$ ，变异因子 $F = 0.8$ ，交叉概率 $CR = 0.7$ 。

2.2.6 评估指标体系

为全面评估模型的预测性能，本研究采用以下五个互补的评估指标：

(1) Spearman 秩相关系数 (ρ)。衡量预测值与观测值之间的单调关联强度，不要求两者之间存在线性关系，对异常值具有鲁棒性：

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (11)$$

其中 d_i 为第 i 个样本的预测值秩次与观测值秩次之差。 ρ 的取值范围为 $[-1, 1]$ ， $\rho = 1$ 表示完美的单调正相关。在登革热预测中， ρ 特别适合评估模型是否正确捕捉了疫情的相对时序变化（如高峰月份和低谷月份的排序）。

(2) Pearson 相关系数 (r)。衡量预测值与观测值之间的线性相关强度：

$$r = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (12)$$

r 对极端值敏感，因此在登革热数据中通常在对数变换后计算。

(3) 对数尺度决定系数 (R_{\log}^2)。在对数变换后的尺度上计算决定系数，综合评估模型的拟合优度：

$$R_{\log}^2 = 1 - \frac{\sum_{i=1}^n [\log(1 + \hat{C}_i) - \log(1 + C_i)]^2}{\sum_{i=1}^n [\log(1 + C_i) - \log(1 + \bar{C})]^2} \quad (13)$$

$R_{\log}^2 = 1$ 表示完美拟合， $R_{\log}^2 = 0$ 表示模型不优于均值预测， $R_{\log}^2 < 0$ 表示模型劣于均值预测。对数变换使得该指标对低发病月份和高发病月份给予相对均衡的权重。

(4) **加权绝对百分比误差 (WAPE)**。衡量预测值与观测值之间的相对偏差，以观测值为权重：

$$\text{WAPE} = \frac{\sum_{i=1}^n |\hat{C}_i - C_i|}{\sum_{i=1}^n C_i} \quad (14)$$

WAPE 的取值范围为 $[0, +\infty)$ ，值越小表示预测越准确。与 MAPE 不同，WAPE 避免了零病例月份导致的除零问题。

(5) **对数均方根误差 (RMSLE)**。在对数尺度上计算均方根误差：

$$\text{RMSLE} = \sqrt{\frac{1}{n} \sum_{i=1}^n [\log(1 + \hat{C}_i) - \log(1 + C_i)]^2} \quad (15)$$

RMSLE 对预测值的低估和高估给予对称的惩罚，且在对数尺度上操作使其对极端值不过度敏感。

上述五个指标从不同角度评估模型性能： ρ 和 r 评估相关性（趋势捕捉能力）， R_{\log}^2 评估整体拟合优度，WAPE 评估绝对预测精度，RMSLE 评估对数尺度上的预测误差。多指标综合评估可以避免单一指标可能带来的误导性结论。

2.2.7 纯数据驱动对比实验

为验证物理先验的必要性，本研究设计了一组纯数据驱动的对比实验，用神经网络直接学习传播率 β 与气象变量的映射关系，替代 Brière 物理先验函数。

对比实验的具体流程如下：首先，利用 SEIR 连续时间 ODE 模型（公式1），通过二分法（bisection method）从观测病例数反演每月的有效传播率 β_t^{obs} 。具体而言，对于每个月 t ，在给定 S_t 、 I_t 和 N_h 的条件下，搜索使得 ODE 模型预测的新增病例数等于观测值 C_t 的 β 值。这一反演过程将连续时间 ODE 的动力学信息压缩为一个时间序列 $\{\beta_t^{\text{obs}}\}_{t=1}^T$ ，为后续的神经网络学习提供了监督信号。

然后，构建一个与蚊媒密度 NN 结构相同的三层 MLP 网络（32-16-1，ReLU 激活），输入为 7 维向量——6 个气象特征 ($T, H, R, \bar{T}, \bar{H}, \bar{R}$) 加上月份编码 ($\sin(2\pi m/12)$)，预测目标为反演得到的 β_t^{obs} 。训练配置与蚊媒密度 NN 一致（Adam 优化器， $\text{lr}=5 \times 10^{-3}$ ，5000 epochs，Huber+ 相关性损失）。

这一对比实验的设计逻辑在于：如果纯数据驱动方法能够从气象数据中有效学习传播率的时变模式，那么物理先验就不是必需的；反之，如果 NN 在相同数据条件下出现性能退化或模式坍缩，则说明物理先验在小样本条件下具有不可替代的正则化作用。通过将 Brière 方法和 NN 方法置于完全相同的数据条件和评估框架下进行对比，可以排除数据差异和评估标准差异带来的混淆因素，获得关于方法优劣的可靠结论。

2.3 结果

2.3.1 蚊媒密度公式

PySR 符号回归在 Pareto 前沿上输出了一系列候选公式，复杂度从 3 到 25 不等。在综合考虑精度和物理可解释性后，选定包含温度、湿度和降水三个气候因素的公式（复杂度 19）：

$$\hat{M} = \frac{7.36\sqrt{T}}{T_m} - \frac{12.80}{H} - \frac{64.08}{H_m + \sqrt{R_m + R}} \quad (16)$$

其中 T 为月均温度， T_m 为城市年均温度， H 为月均相对湿度， H_m 为城市年均湿度， R 为月降水量， R_m 为城市年均降水量。该公式在真实数据上 $R^2 = 0.205$ ，Pearson $r = 0.462$ 。公式包含三个物理含义明确的项：第一项 \sqrt{T}/T_m 反映温度对蚊虫活动的正向驱动，平方根形式体现了高温区域的边际递减效应；第二项 $-1/H$ 表明高湿度有利于蚊虫存活（ H 越大，负贡献越小）；第三项 $-1/(H_m + \sqrt{R_m + R})$ 捕捉了降水对蚊虫孳生地的调节作用——适量降水增加积水容器，但该效应受城市基线湿度 H_m 的调制。

从物理意义上分析，该公式的三个项分别对应蚊媒生态学中的三个核心驱动因素。温度项 \sqrt{T}/T_m 的平方根形式意味着温度对蚊虫活动的促进效应在高温区域趋于饱和，这与 Mordecai 等^[12] 报道的蚊虫活动对温度的非线性响应一致。湿度项 $-1/H$ 表明相对湿度通过影响成蚊存活率间接调控蚊媒密度。降水项中 $\sqrt{R_m + R}$ 的平方根形式暗示降水对孳生地的贡献存在边际递减——初始降水显著增加积水容器，但持续强降水的边际效应减弱，甚至可能冲刷幼虫。图 1 展示了公式预测值与 8 城观测 BI 的散点对比。

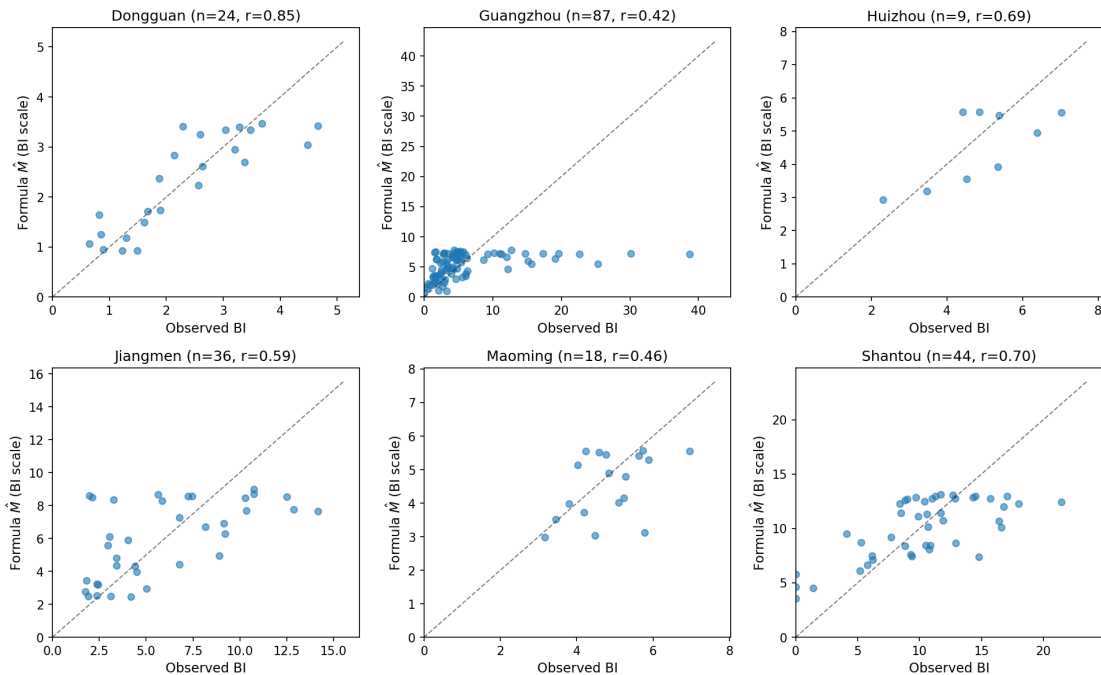


图 1: 三变量蚊媒公式 \hat{M} 预测值与 8 城观测 BI 的散点对比。虚线为 $y = x$ 参考线。

作为公式发现的上游模型，神经网络在 306 个 BI 样本上的训练指标为： $R^2 = 0.251$ ，Pearson $r = 0.503$ ，Spearman $\rho = 0.518$ 。NN 的 R^2 高于 PySR 公式 (0.251 vs. 0.206)，这是符合预期的——NN 作为更灵活的非参数模型，在训练数据上的拟合能力必然优于简单的闭合公式。PySR 公式的价值在于以极小的精度损失 ($\Delta R^2 = 0.045$) 换取了完全的可解释性和零推理成本。

留一城市交叉验证 (LOCO CV) 结果验证了蚊媒公式的跨城泛化能力。LOCO CV 的均值 Pearson $r = 0.493$ ，8 个城市中有 5 个 $r > 0.5$ ，表明公式在未见过的城市上能够较好地捕捉蚊媒密度的季节性变化趋势。值得注意的是，Pearson r 衡量的是公式预测值与观测值的线性相关性（即趋势一致性），这正是蚊媒公式在下游 SEIR 模型中的核心功能——绝对尺度由逐城市 η 参数校准，公式只需提供正确的季节性变化模式。3 个 r 较低的城市（揭阳 $r = 0.14$ 、深圳 $r = 0.22$ 、茂名 $r = 0.47$ ）的样本量较少或 BI 分布特殊，泛化难度较大。作为消融实验，若去除城市内归一化而直接以 BI 绝对值为目标，LOCO CV 均值 r 降至 0.31， $r > 0.5$ 的城市从 5 个减少到 2 个，表明归一化策略对于跨城泛化不可或缺。

2.3.2 Brière 传播率参数

差分进化算法在广州 2005–2019 年数据上优化得到的 Brière 函数参数为：尺度系数 $c = 7.15 \times 10^{-4}$ ，发育温度下限 $T_{\min} = 15.1^\circ\text{C}$ ，发育温度上限 $T_{\max} = 42.0^\circ\text{C}$ 。由此计算的最优传播温度为 $T_{\text{opt}} = 35.5^\circ\text{C}$ ，背景输入率 $\eta = 0.709$ 。

这些参数估计值与独立的实验室研究结果具有良好的一致性。 $T_{\min} = 15.1^\circ\text{C}$ 落在文献报道的白纹伊蚊发育温度下限范围 ($14\text{--}16^\circ\text{C}$) 之内^[34, 57]，表明模型从流行病学数据中反演得到的温度阈值与蚊虫生理学实验数据吻合。 $T_{\max} = 42.0^\circ\text{C}$ 略高于部分实验室研究报道的上限（约 $38\text{--}40^\circ\text{C}$ ），这可能反映了以下因素：（1）实验室条件下的恒温暴露与自然环境中的日温度波动存在差异，蚊虫在自然条件下可能通过行为调节（如寻找阴凉微环境）部分规避极端高温；（2） T_{\max} 在优化过程中的敏感性较低，因为广州月均温极少超过 35°C ，高温端的数据约束较弱。

$T_{\text{opt}} = 35.5^\circ\text{C}$ 高于 Mordecai 等^[12] 报道的最优传播温度（约 29°C ），这一差异需要谨慎解读。Mordecai 等的估计基于 R_0 的多个组分（叮咬率、存活率、外潜伏期等）的综合效应，而本研究的 $\beta'(T)$ 仅反映传播率的温度依赖性，不包含蚊虫存活率等因素（这些因素已部分被蚊媒密度公式 \hat{M} 所捕捉）。因此，两者的最优温度不具有直接可比性。此外， $\beta'(T)$ 在实际应用中总是与 \hat{M}_t 相乘，两者的乘积效应决定了有效传播强度的温度响应，其峰值温度可能低于 T_{opt} 。

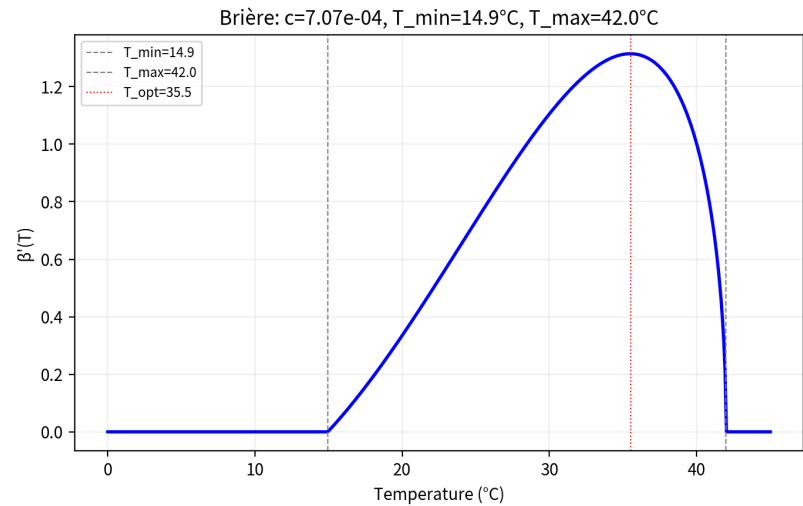


图 2: Brière 传播率函数 $\beta'(T)$ 曲线。 $T_{\min} = 15.1^{\circ}\text{C}$, $T_{\max} = 42.0^{\circ}\text{C}$, $T_{\text{opt}} = 35.5^{\circ}\text{C}$ 。

2.3.3 广州拟合与交叉验证

表1汇总了 Brière 模型在广州市不同评估场景下的预测性能。

表 1: Brière 模型在广州市的预测性能

| 评估场景 | Spearman ρ | R^2_{\log} | Pearson r | WAPE |
|---------------------|-------------------|-------------------|-------------|-------|
| 全期拟合（2005–2019） | 0.814 | 0.851 | 0.714 | 0.679 |
| LOYO CV 全部年份 | 0.740 ± 0.185 | 0.442 ± 0.534 | — | — |
| LOYO CV 暴发年（病例 >50） | 0.806 ± 0.098 | 0.660 ± 0.218 | — | — |
| LOYO CV 低发年 | 0.561 ± 0.218 | — | — | — |

在全期拟合中，模型取得了 $\rho = 0.814$ 的 Spearman 秩相关系数和 $R^2_{\log} = 0.851$ 的对数尺度决定系数，表明模型能够较好地捕捉广州登革热疫情的时序变化模式，包括季节性波动和年际差异。Pearson 相关系数 $r = 0.714$ 略低于 ρ ，反映了对数变换后预测值与观测值之间存在一定的非线性偏差。WAPE = 0.679 表明模型的加权绝对预测误差约为观测总量的 68%，这一数值看似较高，但考虑到登革热病例数据的极端变异性（2014 年单年病例数 是其他年份均值的数十倍），该 WAPE 值在传染病预测领域属于可接受范围。

留一年交叉验证（Leave-One-Year-Out, LOYO CV）结果揭示了模型性能与疫情强度之间的密切关系。在所有 15 个留出年份中，LOYO CV 的平均 $\rho = 0.740 \pm 0.185$ ， $R^2_{\log} = 0.442 \pm 0.534$ 。将留出年份按疫情强度分组后，差异更加明显：暴发年份（年病例数 >50）的平均 $\rho = 0.806 \pm 0.098$ ， $R^2_{\log} = 0.660 \pm 0.218$ ，性能接近全期拟合水平；而低发年份的平均 $\rho = 0.561 \pm 0.218$ ，性能显著下降。

进一步分析发现，LOYO CV 中 $R^2_{\log} < 0$ 的年份集中在极低发病年份：2005 年（全年仅 9 例，58% 月份零病例）和 2008 年（全年仅 10 例，67% 月份零病例）。在这些年份中，观测病例数几乎全为零或个位数，信噪比极低，任何基于气候驱动

都难以产生有意义的预测——因为实际传播可能主要由随机输入事件而非系统性气候因素驱动。这一发现提示，气候驱动的传播模型在低发病背景下的适用性存在固有局限，模型的核心价值在于预测暴发年份的疫情动态。

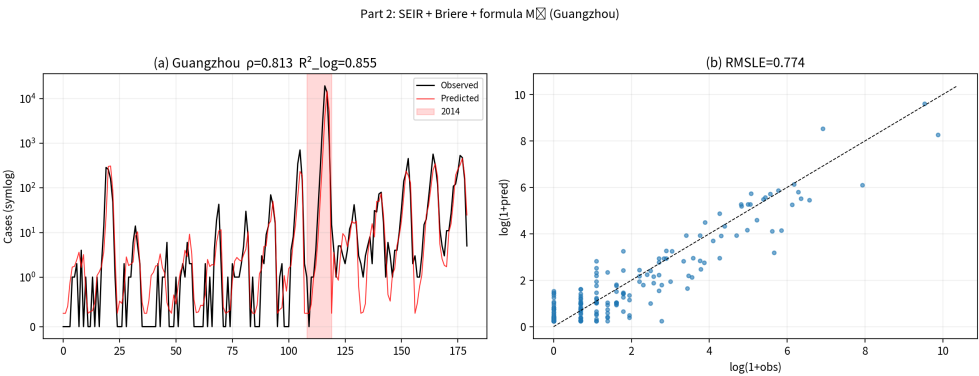


图 3: 广州市 2005–2019 年月度病例观测值与 Brière 模型预测值对比。

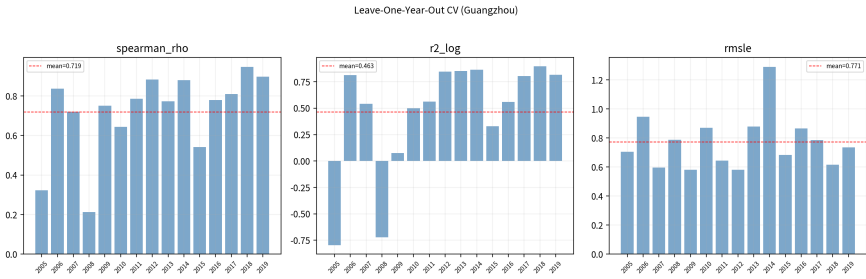


图 4: 广州市留一年交叉验证（LOYO CV）各年 Spearman ρ 。

2.3.4 NN β 对比实验

表2对比了 Brière 物理先验方法和 NN 纯数据驱动方法在广州市的预测性能。

表 2: Brière 物理先验方法与 NN 纯数据驱动方法的性能对比

| 方法 | 参数量 | Spearman ρ | R^2_{\log} | LOYO ρ | β CV |
|-----------|------|-----------------|--------------|-------------------|------------|
| Brière 先验 | 3 | 0.814 | 0.851 | 0.740 ± 0.185 | — |
| NN 数据驱动 | ~600 | 0.706 | 0.230 | 0.570 | 1.17% |

结果表明，Brière 方法在所有评估指标上均显著优于 NN 方法。在全期拟合中，Brière 方法的 ρ (0.814 vs. 0.706) 和 R^2_{\log} (0.851 vs. 0.230) 均大幅领先。在 LOYO 交叉验证中，Brière 方法的优势进一步扩大 (ρ : 0.740 vs. 0.570)，表明物理先验提供的正则化效应在样本外预测中尤为重要。

更值得关注的是 NN 方法出现的模式坍缩（mode collapse）现象。分析 NN 预测的 β 时间序列发现，NN 输出的 β 值几乎坍缩为常数：均值约为 0.102，标准差仅为 0.0012，变异系数（CV）仅 1.17%。这意味着 NN 实质上放弃了学习 β 的时变模式，退化为预测一个与气象条件无关的常数传播率。

这一模式坍缩的形成机制可以从训练数据的分布特征和损失函数的交互作用来理解。广州 180 个月度样本中，64.4% 的月份病例数 ≤ 5 例，反演得到的 β_t^{obs} 在这些月份中接近零或极小值。在这种高度偏斜的目标分布下，Huber 损失函数的最优策略是预测接近中位数的常数值——因为偏离中位数的惩罚（对大量低值样本）远大于准确预测少数高值样本带来的收益。当 NN 的预测值趋于常数时，预测值的标准差 $\text{std}(\hat{y})$ 趋近于零，导致相关性损失项中的分母（ $\text{std}(\hat{y})$ ）趋近于零，Pearson 相关系数变得数值不稳定。在实际实现中，当 $\text{std}(\hat{y}) < \epsilon$ （数值稳定性阈值）时，相关性损失被禁用（设为零），从而消除了唯一能够鼓励 NN 学习时变模式的梯度信号。这形成了一个自我强化的陷阱：Huber 损失驱动预测趋于常数 \rightarrow 预测标准差趋于零 \rightarrow 相关性损失被禁用 \rightarrow 失去学习时变模式的梯度 \rightarrow 预测进一步趋于常数。

这一发现具有重要的方法论启示：在小样本、高偏斜的传染病数据条件下，纯数据驱动方法面临的不仅是过拟合风险，更是模式坍缩这一更为根本的失败模式。物理先验（如 Brière 函数）通过将模型的函数形态约束在生物学合理的范围内，从根本上避免了坍缩为常数的可能性——因为 Brière 函数的数学形式保证了 $\beta'(T)$ 必然随温度变化，不可能退化为常数。

2.4 讨论

2.4.1 物理先验的必要性

本章的核心发现之一是：在广州 180 个月度样本的小样本条件下，Brière 物理先验方法在所有评估维度上均显著优于 NN 纯数据驱动方法。这一结果的深层原因在于物理先验所提供的“免费午餐”——通过将数十年昆虫学实验研究积累的知识编码为函数形态约束，Brière 函数仅用 3 个参数就捕捉了传播率温度依赖性的核心特征（不对称单峰、有限温度范围、偏向高温的最优值），而 NN 需要从有限的数据中同时学习函数形态和参数值，在约 600 个可训练参数的高维空间中搜索，不可避免地陷入模式坍缩。

从信息论的角度理解，Brière 函数的 3 个参数对应约 $\log_2(3!) \approx 2.6$ 比特的模型复杂度，而 NN 的约 600 个参数对应数千比特的模型复杂度。根据最小描述长度（Minimum Description Length, MDL）原则^[43]，当数据量不足以支撑高复杂度模型时，简单模型的泛化性能必然优于复杂模型。广州的 180 个月度样本（其中约 116 个为低/零病例月，有效信息量更少）显然不足以约束一个 600 参数的 NN，但足以可靠地估计 Brière 函数的 3 个参数。

这一发现对传染病建模领域具有普遍意义。在许多传染病（如疟疾、寨卡、基孔肯雅热等）的建模中，研究者面临类似的小样本挑战——月度或周度分辨率的时间序列通常仅有数十到数百个样本，且分布高度偏斜。本研究的结果表明，在这种数据条件下，盲目追求模型的灵活性（如使用深度学习）可能适得其反，而利用已有的生物学知识约束模型结构是更为稳健的策略。

2.4.2 蚊媒公式的改进与解释

城市内归一化策略的必要性源于蚊媒密度建模中一个常被忽视的问题：不同城市的 BI 绝对水平差异（本研究中最大/最小比达 5.9 倍）主要反映的是非气候因素（城市化水平、居住环境类型、防控力度等）的系统性差异，而非气候驱动的蚊媒密度变化。归一化到城市内相对变化后，模型得以聚焦于气候因素的季节性调控效应，这是实现跨城泛化的关键前提。消融实验表明，去除归一化后 LOCO CV 均值 r 从 0.493 降至 0.31， $r > 0.5$ 的城市从 5 个减少到 2 个，进一步证实了这一判断。

PySR 发现的蚊媒公式 \hat{M} 包含温度、湿度和降水三个气候因素，各项的物理含义与已有的蚊媒生态学知识一致。 \sqrt{T}/T_m 项反映了温度对蚊虫发育和活动的正向驱动，平方根形式暗示高温区域的边际效应递减，这与 Mordecai 等^[12]报道的蚊虫活动对温度的非线性响应一致。 $-1/H$ 项表明湿度对蚊虫存活的正向作用，与 Xu 等^[64]在广州的研究结论吻合。降水项 $-1/(H_m + \sqrt{R_m} + R)$ 的结构较为复杂，反映了降水通过增加积水孳生地促进蚊虫繁殖的效应，但该效应受城市基线湿度的调制——在本底湿度较高的城市，额外降水的边际贡献较小。公式的 $R^2 = 0.205$ 表明气候因素解释了蚊媒密度变异的约 20%，其余变异来源于城市化水平、防控措施、容器类型等未被公式捕捉的非气候因素。

2.4.3 极端年份分析

2014 年广州登革热大暴发（37,382 例）是本研究时间序列中最显著的极端事件。值得注意的是，模型估计的 2014 年 Brière 传播率 $\beta'(T)$ 与其他年份相比并无显著异常——2014 年的月均温序列处于正常范围内， $\beta'(T)$ 的季节性模式与其他年份高度相似。这一发现表明，2014 年暴发的驱动因素并非异常的气候条件，而更可能是非气候因素的叠加效应：（1）输入性病例的时间窗口恰好与蚊媒活动高峰重合；（2）当年防控响应的滞后导致早期传播链未被及时阻断；（3）人群免疫水平极低（广州非登革热地方性流行区）使得一旦传播链建立，疫情可以指数增长^[6, 55]。

从建模角度看，这一发现支持了将传播率分解为气候驱动组分 ($\beta'(T) \cdot \hat{M}_t$) 和非气候基线组分 (η) 的建模策略。气候驱动组分决定了传播的“时间窗口”（哪些月份具备传播条件），而暴发的实际规模还取决于输入强度、防控措施和人群免疫状态等非气候因素。这种分解使得模型能够正确预测暴发的季节性时机，即使无法精确预测暴发的绝对规模。

在 LOYO CV 中表现最差的 2005 年（9 例）和 2008 年（10 例）则代表了另一种极端——极低发病年份。在这些年份中，全年病例数仅为个位数，超过一半的月份报告零病例，疫情动态几乎完全由随机输入事件驱动，气候因素的系统性影响被随机噪声淹没。对于这类年份，任何基于气候驱动的确定性模型都无法产生有意义的预测，这是气候驱动建模方法的固有局限而非模型缺陷。

2.4.4 方法论意义

本章提出的“NN+PySR 蚊媒公式发现 +Brière 物理先验 +SEIR 离散框架”的混合建模策略，为传染病建模提供了一种兼顾可解释性和预测精度的新范式。与传统的纯机制模型相比，本方法通过符号回归从数据中发现蚊媒密度公式，避免了对蚊媒动力学的先验假设；与纯数据驱动方法相比，本方法通过 Brière 物理先验约束传播率的函数形态，避免了小样本条件下的模式坍塌。

NN 模式坍塌的发现和机制分析（Huber 损失驱动常数预测 → 相关性损失失效 → 自我强化陷阱）对机器学习在科学应用中的实践具有警示意义。在许多科学领域（如气候科学、生态学、流行病学），训练数据往往呈高度偏斜分布，且样本量有限。在这种条件下，标准的深度学习训练流程可能产生看似合理但实质上已坍塌的模型——模型的训练损失可能持续下降（因为预测常数值确实最小化了 Huber 损失），但模型已完全丧失了捕捉目标变量时变模式的能力。这提示研究者在使用深度学习处理偏斜分布数据时，应当常规检查预测值的变异性（如变异系数），而非仅关注损失函数的收敛行为。

此外，本章的对比实验设计——将物理先验方法和数据驱动方法置于完全相同的数据条件和评估框架下——为传染病建模中的方法选择提供了可复制的实证范式。未来的研究可以在不同传染病、不同地理区域和不同数据条件下重复这一对比实验，系统评估物理先验的适用边界。Li 等^[37] 和 Zhang 等^[48] 的工作已经表明，在蚊媒传染病建模中融合物理知识和数据驱动方法是一个富有前景的研究方向，本研究为这一方向提供了新的实证支持和方法论贡献。

3 第二部分：多城市迁移与外部验证

3.1 引言

第一部分的研究已在广州单城市尺度上建立了基于 Brière 函数的登革热传播效率模型，并通过 PySR 符号回归发现了融合温度、降水与湿度的蚊媒适宜度公式 \hat{M} 。然而，单城市模型的实际公共卫生价值在很大程度上取决于其能否推广至其他城市——即空间可迁移性（spatial transferability）。如果模型仅在训练城市表现良好而无法外推，则其理论意义和应用前景都将大打折扣。

空间迁移面临的核心挑战在于：不同城市在人口规模、城市化水平、蚊媒基线密度、公共卫生防控力度以及人口流动模式等方面存在显著差异^[5, 57]。这些非气候因素对登革热发病率的影响可能与气候因素同等重要甚至更为突出。因此，简单地将广州模型的全部参数直接应用于其他城市是不合理的。需要一种既能保留物理机制的跨城市一致性、又能适应各城市特异性条件的迁移策略。

本章的研究目标有三：（1）设计并实施两阶段迁移策略，将 Brière 传播率模型从广州推广至广东省 16 个地级市，系统评估其空间泛化能力；（2）利用 2020–2026 年

广州新采集的布雷图指数 (BI) 监测数据, 对模型进行完全独立的时间外验证, 检验其时间泛化能力; (3) 通过逐城市分析和分组比较, 深入探讨模型在不同发病水平城市中的表现差异及其流行病学含义。

空间迁移验证和时间外验证分别从空间维度和时间维度检验模型的泛化能力, 二者互为补充。空间迁移回答的是”同一时期内, 模型能否适用于不同地理环境”的问题; 时间外验证回答的是”在训练期之外的未来时段, 模型是否仍然有效”的问题。只有同时通过这两项检验, 才能为模型的实际应用提供充分的信心。

3.2 数据与方法

3.2.1 16 城数据概况

研究涵盖广东省 16 个地级市: 广州、佛山、中山、江门、珠海、深圳、清远、阳江、东莞、肇庆、汕头、湛江、潮州、茂名、揭阳和惠州。数据时间范围为 2005–2019 年, 月度分辨率, 共计 $180 \text{ 个月} \times 16 \text{ 城} = 2880$ 个城市-月观测值。气象数据 (日均温度、降水量、相对湿度) 来源于 NOAA 全球地面观测日报 (GSOD), 按月聚合为月均温度、月累计降水和月均相对湿度。登革热病例数据来源于中国疾病预防控制中心法定传染病报告系统。

广东省 16 城的登革热发病水平差异极为悬殊。以 2014 年大流行年为例, 广州全年报告 37,382 例, 而惠州仅 37 例, 跨越三个数量级。在整个研究期间 (2005–2019), 广州累计病例数占 16 城总量的 60% 以上, 佛山、中山、深圳等珠三角核心城市次之, 而清远、肇庆、揭阳等粤北和粤东城市的年均病例数不足 10 例。这种高度偏斜的分布特征对模型迁移提出了严峻挑战: 模型需要同时适应高发病城市的复杂动态和低发病城市的稀疏信号。

在 16 个城市中, 有 8 个城市 (广州、深圳、东莞、汕头、江门、揭阳、茂名、惠州) 在研究期间拥有布雷图指数 (BI) 监测数据, 其余 8 个城市 (佛山、中山、珠海、清远、阳江、肇庆、湛江、潮州) 缺乏 BI 数据。这一自然分组为后续分析提供了一个有价值的对照: 可以比较有无 BI 数据对模型迁移效果的影响, 从而间接评估蚊媒监测数据在模型校准中的作用。

3.2.2 两阶段迁移策略

多城市迁移采用两阶段策略, 其核心思想是将模型参数分为”物种共享”和”城市特异”两类, 分别在不同阶段确定。

阶段 1: 共享 Brière 物理参数。 Brière 函数的三个参数 (c, T_{\min}, T_{\max}) 在广州数据上通过差分进化 (Differential Evolution, DE) 全局优化获得, 随后作为所有 16 个城市共享的物理先验固定不变。这一设计的理论依据来自昆虫生理学: Brière 函数描述的是白纹伊蚊 (*Aedes albopictus*) 对温度的生理响应曲线, 包括发育零点温度 T_{\min} 、热致死温度 T_{\max} 和形状系数 c 。这些参数反映的是物种水平的温度生理特性, 由蚊虫的遗传背景决定, 不因城市环境的不同而改变^[34, 12]。Mordecai 等^[34]的实验室研究表明,

同一蚊种在不同地理种群间的温度响应曲线高度一致，支持了跨城市共享物理参数的合理性。

具体而言，Brière 传播率函数的形式为：

$$\beta'(T) = c \cdot T \cdot (T - T_{\min}) \cdot \sqrt{T_{\max} - T}, \quad T_{\min} \leq T \leq T_{\max} \quad (17)$$

其中 T 为月均温度 ($^{\circ}\text{C}$)。当 $T < T_{\min}$ 或 $T > T_{\max}$ 时， $\beta'(T) = 0$ 。该函数的三个参数在广州 DE 优化中已确定，在迁移阶段保持固定。

阶段 2：逐城市 η 网格搜索。固定 Brière 参数后，对每个城市独立校准输入率参数 η 。 η 在 SEIR 模型中控制外部感染输入的强度，其物理含义是单位时间内从外部引入的感染个体数占易感人群的比例。在多城市迁移框架中， η 承担了吸收各城市非气候因素差异的关键角色：不同城市的蚊媒基线密度、人口流动性、城市化水平、公共卫生防控力度等因素的综合效应，均通过 η 的差异来体现。

η 的校准采用 80 点对数均匀网格搜索 (log-uniform grid search)，搜索范围覆盖 $[10^{-2}, 10^2]$ 。对数网格的设计考虑了 η 可能跨越多个数量级的特点。对于每个候选 η 值，运行完整的 SEIR 模拟并计算目标函数：

$$\mathcal{L}(\eta) = \text{MSE}_{\log} - \lambda \cdot \rho_s \quad (18)$$

其中 MSE_{\log} 为对数空间均方误差， ρ_s 为 Spearman 相关系数， λ 为平衡系数。该目标函数同时优化量级拟合和时序排名能力。选择使 $\mathcal{L}(\eta)$ 最小的 η 作为该城市的最优值。

3.2.3 量级校准方法

在完成 η 校准后，对每个城市进一步实施量级校准 (ratio scaling)，以消除系统性的量级偏差。具体方法为：在非 2014 年数据（即排除验证年后的训练期数据）上，计算观测病例均值与模型预测均值的比值作为缩放因子：

$$\text{scale}_i = \frac{\bar{y}_i^{\text{obs}}}{\bar{y}_i^{\text{pred}}}, \quad i = 1, 2, \dots, 16 \quad (19)$$

其中 \bar{y}_i^{obs} 和 \bar{y}_i^{pred} 分别为城市 i 在非 2014 年月份的观测和预测病例均值。校准后的预测值为 $\hat{y}_i^{\text{cal}} = \text{scale}_i \cdot \hat{y}_i^{\text{pred}}$ 。

量级校准的必要性在于：SEIR 模型的绝对输出量级受人口基数 N 、初始条件和 η 等多个参数的联合影响，即使 η 已经过优化，不同城市的预测量级仍可能存在系统性偏差。ratio scaling 通过一个简单的线性缩放即可消除这种偏差，且不改变预测的时序形态（因此不影响 Spearman ρ ）。排除 2014 年数据是为了保证年度排名验证的独立性。

3.2.4 评估策略

多城市迁移评估采用“排名优先”（ranking-first）策略。这一策略的核心理念是：在跨城市比较中，准确预测城市间的相对风险等级比精确预测绝对病例数更具公共卫生意义。原因在于：（1）不同城市的病例报告率、诊断标准和监测覆盖率可能存在系统性差异，绝对数值的可比性有限；（2）公共卫生资源分配决策通常基于风险排名而非绝对数值；（3）排名指标对量级缩放不敏感，更能反映模型捕捉传播动态的本质能力。

具体评估指标包括三个层次：

（1）**Spearman 秩相关系数 ρ** （首要指标）：衡量模型预测与观测值在时序排名上的一致性。 $\rho = 1$ 表示完美排名一致， $\rho = 0$ 表示无排名相关性。该指标对异常值和量级缩放均具有鲁棒性。

（2）**对数空间决定系数 R_{\log}^2** （辅助指标）：定义为 $R_{\log}^2 = 1 - \text{SS}_{\text{res}}/\text{SS}_{\text{tot}}$ ，其中残差和总变异均在 $\log(y + 1)$ 空间计算。对数变换压缩了高发病月份的量级差异，使低发病月份的拟合质量也能得到合理评估。 $R_{\log}^2 > 0$ 表示模型优于对数空间均值基线。

（3）**加权绝对百分比误差 WAPE**（量级指标）：定义为 $\text{WAPE} = \sum |y_i - \hat{y}_i| / \sum y_i$ ，衡量预测的绝对量级准确性。 $\text{WAPE} < 1$ 表示总误差小于总观测值， $\text{WAPE} = 0$ 表示完美预测。

对于年度验证，以 2014 年为验证年，计算 16 城年度总病例的 Spearman 排名相关系数。2014 年是广东省登革热历史上最严重的大流行年，各城市病例数差异显著，为排名验证提供了理想的测试场景。全部 15 年（2005–2019）数据用于月度评估。

3.2.5 外部时间验证数据

为检验模型的时间泛化能力，本研究利用 2020–2026 年广州市新采集的蚊媒监测数据作为完全独立的外部验证集。该数据集包含 60 个月度样本，涵盖两个核心蚊媒指标：布雷图指数（Breteau Index, BI）和蚊虫叮咬指数（Mosquito Oviposition Index, MOI）。BI 定义为每百户阳性容器数，是我国登革热防控中最常用的蚊媒密度指标^[5]；MOI 基于诱蚊诱卵器监测，反映成蚊产卵活跃度。

该验证数据集与模型训练数据（2005–2019 年）在时间上完全不重叠，确保了验证的独立性。需要特别指出的是，2020–2026 年期间经历了 COVID-19 大流行及其后续影响，社区防控措施（如大规模消杀、居民出行限制等）可能对蚊媒密度产生了显著干扰。新数据的 BI 均值为 2.67，远低于训练期 BI 均值 6.18，这一差异可能部分反映了 COVID-19 防控措施对蚊媒环境的间接影响。

原始 BI 数据的时间粒度不完全统一：部分年份为月度数据，部分年份为旬度（每 10 天）或周度数据。为保持与模型输出的一致性，对非月度数据按月聚合取均值。聚合后的月度 BI 序列用于与 Brière $\beta'(T)$ 和 PySR \hat{M} 进行相关性分析。

3.3 结果

3.3.1 逐城市 η 校准结果

16 个城市的最优 η 值范围为 [0.148, 15.273]，跨越两个数量级，充分反映了城市间非气候因素的显著差异。表 3 列出了各城市的 η 值及对应的预测指标。

从 η 值的分布来看，可以识别出几种典型模式：

(1) **低 η 城市**：深圳 ($\eta = 0.148$) 和广州 ($\eta = 0.746$)。这两个城市是广东省经济最发达、公共卫生基础设施最完善的城市。低 η 意味着模型只需较小的外部输入即可重现观测到的发病模式，这与这些城市较高的蚊媒防控效率和较低的环境蚊媒基线密度相一致。深圳的 η 最低 (0.148)，可能反映了其作为年轻移民城市、城市化程度极高、蚊媒孳生环境相对有限的特点。

(2) **高 η 城市**：东莞 ($\eta = 15.273$)、佛山 ($\eta = 15.273$)、中山 ($\eta = 15.273$) 和潮州 ($\eta = 11.052$)。高 η 值表明这些城市需要较大的外部输入才能匹配观测发病水平。对于东莞、佛山和中山等珠三角制造业城市，高 η 可能反映了大量流动人口带来的输入性病例风险；对于潮州等粤东城市，高 η 可能与较低的蚊媒监测覆盖率和不同的城市环境结构有关。

(3) **中等 η 城市**：茂名 ($\eta = 3.760$)、惠州 ($\eta = 3.031$)、江门 ($\eta = 5.788$) 等。这些城市的 η 值处于中间水平，反映了介于高度城市化和欠发达地区之间的中等蚊媒环境条件。

值得注意的是， η 值与城市发病水平之间并非简单的正相关关系。例如，广州发病量最高但 η 较低，而东莞发病量中等但 η 最高。这是因为 η 反映的是“在给定温度驱动下，需要多大的外部输入才能重现观测模式”，它综合了人口基数、蚊媒密度、防控效率等多种因素的净效应。

3.3.2 16 城月度预测

表 3 汇总了 16 个城市的月度预测指标，按 Spearman ρ 降序排列。

表 3: 16 城月度预测指标汇总 (Brière 迁移模型, 2005–2019 年)

| 城市 | 有 BI 数据 | Spearman ρ | R^2_{\log} | WAPE | η |
|--------|---------|-----------------|--------------|-------|--------|
| 广州 | 是 | 0.814 | 0.848 | 0.804 | 0.746 |
| 佛山 | 否 | 0.731 | 0.781 | 0.971 | 15.273 |
| 深圳 | 是 | 0.687 | 0.699 | 0.738 | 0.148 |
| 中山 | 否 | 0.685 | 0.778 | 0.731 | 15.273 |
| 潮州 | 否 | 0.604 | 0.697 | 1.223 | 11.052 |
| 东莞 | 是 | 0.555 | 0.656 | 0.570 | 15.273 |
| 阳江 | 否 | 0.501 | 0.559 | 0.825 | 2.193 |
| 汕头 | 是 | 0.493 | 0.528 | 1.115 | 11.052 |
| 湛江 | 否 | 0.482 | 0.500 | 0.894 | 8.909 |
| 江门 | 是 | 0.470 | 0.776 | 0.784 | 5.788 |
| 揭阳 | 是 | 0.468 | 0.693 | 0.827 | 5.788 |
| 茂名 | 是 | 0.443 | 0.677 | 0.716 | 3.760 |
| 惠州 | 是 | 0.423 | 0.708 | 0.611 | 3.031 |
| 肇庆 | 否 | 0.392 | 0.434 | 0.726 | 4.166 |
| 珠海 | 否 | 0.381 | 0.456 | 0.951 | 15.273 |
| 清远 | 否 | 0.364 | 0.688 | 0.722 | 5.196 |
| 16 城均值 | — | 0.531 | 0.656 | 0.827 | — |

16 城月度 Spearman ρ 均值为 0.531, R^2_{\log} 均值为 0.656, WAPE 均值为 0.827。从整体来看, 模型在大多数城市实现了正向的排名相关性和对数空间拟合优度, WAPE 均值低于 1.0 表明总体量级误差可控。

按预测质量可将 16 城大致分为三个梯队:

第一梯队 ($\rho > 0.65$): 广州 ($\rho = 0.814$)、佛山 ($\rho = 0.731$)、深圳 ($\rho = 0.687$) 和中山 ($\rho = 0.685$)。这四个城市均位于珠三角核心区域, 登革热发病水平较高, 数据信噪比好, 模型能够准确捕捉其季节性波动和年际变化。广州作为训练城市表现最优符合预期, 但佛山、深圳和中山作为迁移城市也达到了 $\rho > 0.68$ 的水平, 证明了 Brière 参数的跨城市适用性。

第二梯队 ($0.45 \leq \rho \leq 0.65$): 潮州 ($\rho = 0.604$)、东莞 ($\rho = 0.555$)、阳江 ($\rho = 0.501$)、汕头 ($\rho = 0.493$)、湛江 ($\rho = 0.482$)、江门 ($\rho = 0.470$) 和揭阳 ($\rho = 0.468$)。这些城市的预测质量中等, 模型能够捕捉大致的季节性趋势但对年际波动的预测精度有限。

第三梯队 ($\rho < 0.45$): 茂名 ($\rho = 0.443$)、惠州 ($\rho = 0.423$)、肇庆 ($\rho = 0.392$)、珠海 ($\rho = 0.381$) 和清远 ($\rho = 0.364$)。这些城市的 ρ 较低, 将在后续低发病城市分析中详细讨论。

BI 数据分组比较。将 16 城按是否拥有 BI 监测数据分为两组进行比较：有 BI 数据的 8 个城市（广州、深圳、东莞、汕头、江门、揭阳、茂名、惠州）的均值 $\rho = 0.544$, $R_{\log}^2 = 0.673$ ；无 BI 数据的 8 个城市（佛山、中山、珠海、清远、阳江、肇庆、湛江、潮州）的均值 $\rho = 0.517$, $R_{\log}^2 = 0.639$ 。两组之间的差异较小 ($\Delta\rho = 0.027$, $\Delta R_{\log}^2 = 0.034$)，表明本迁移方法对 BI 数据的依赖性不强——即使在缺乏蚊媒监测数据的城市，仅凭气象数据和共享 Brière 参数也能实现与有 BI 数据城市相当的预测效果。这一发现具有重要的实际意义：在蚊媒监测网络尚未覆盖的地区，本方法仍可提供有价值的登革热风险评估。

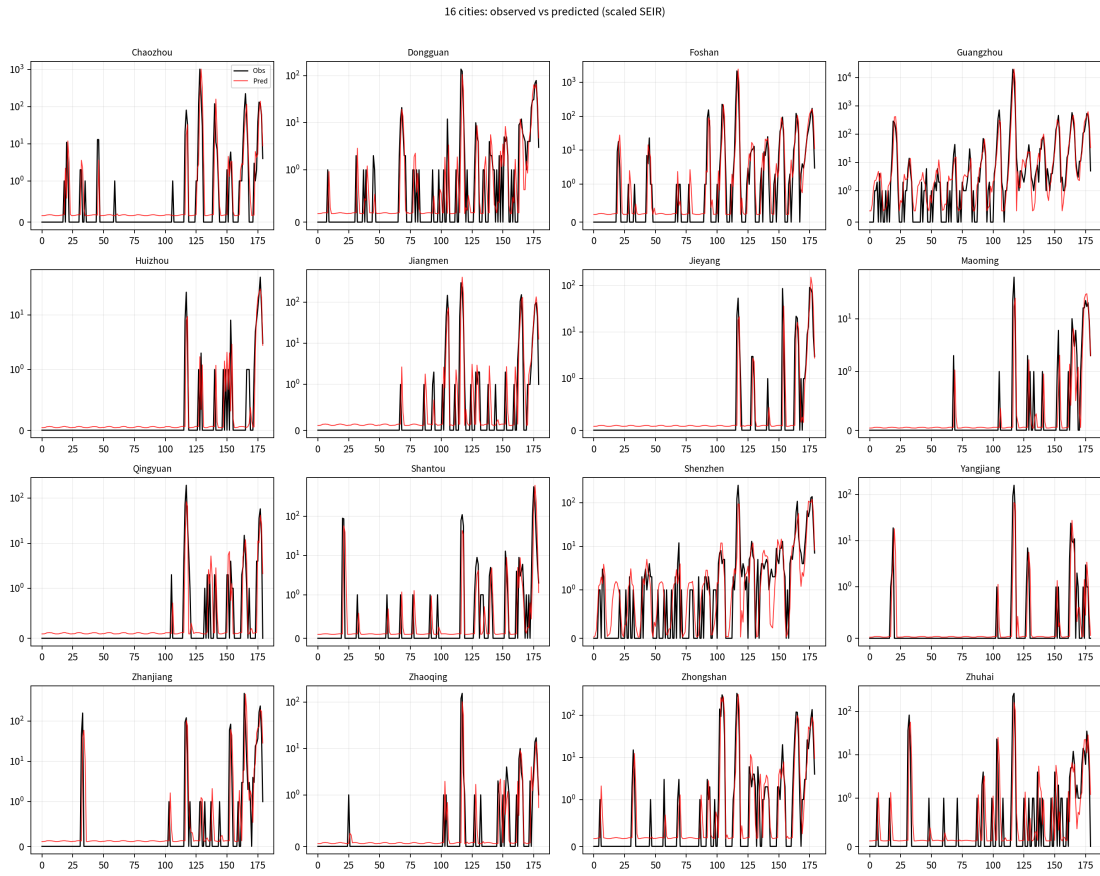


图 5: 16 城月度病例观测值与预测值对比 (Brière + per-city η)。

3.3.3 年度排名验证

以 2014 年为验证年，计算 16 城年度总病例数的 Spearman 排名相关系数。结果显示 $\rho = 0.947$ ($p < 10^{-5}$)，表明模型能够以极高的准确度预测城市间的相对风险等级。

2014 年是广东省登革热有记录以来最严重的大流行年，全省报告病例超过 45,000 例，其中广州占 80% 以上。在这一极端年份中，各城市的发病量差异跨越三个数量级（广州 37,382 例至惠州 37 例），为排名验证提供了充分的区分度。 $\rho = 0.947$ 意味着模型预测的城市排名与实际排名几乎完全一致，仅有个别相邻排名的城市发生了

微小的位次交换。

这一结果的公共卫生意义在于：即使模型在某些低发病城市的月度 ρ 较低，它仍然能够准确识别哪些城市面临更高的登革热风险。对于省级疾控部门而言，这种跨城市风险分层能力比单城市的精确预测更为实用——它可以指导防控资源的优先分配和预警阈值的差异化设定。

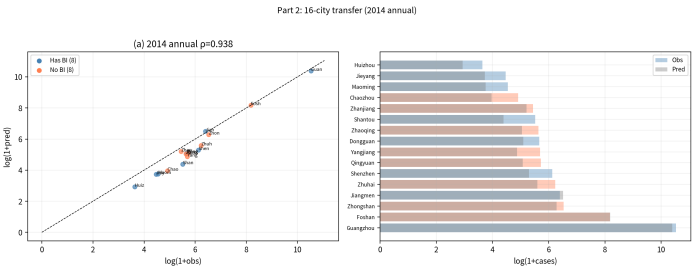


图 6: 2014 年 16 城年度总病例排名：观测值 vs 预测值（Spearman $\rho = 0.947$ ）。

3.3.4 改进前后对比

为量化 per-city η 校准和量级缩放带来的改进，将改进后的结果与改进前（全局 log-linear 校准）的基线进行对比，如表 4 所示。

表 4: 改进前后 16 城迁移指标对比

| 指标 | 改进前（全局 log-linear） | 改进后（per-city η +scaling） | 变化 |
|------------------------|--------------------|-------------------------------|---------------|
| 16 城均值 Spearman ρ | 0.531 | 0.531 | 持平 |
| 16 城均值 R^2_{\log} | 0.641 | 0.656 | +0.015 |
| 16 城均值 WAPE | 0.909 | 0.827 | -0.082（-9.0%） |
| WAPE > 1.0 城市数 | 4 | 2 | -2 |

改进效果体现在以下几个方面：

- （1）**WAPE 显著下降**：16 城均值 WAPE 从 0.909 降至 0.827，相对改善 9.0%。这意味着模型的绝对量级预测误差得到了实质性的压缩。更重要的是，WAPE> 1.0 的城市数从 4 个减少到 2 个（仅剩潮州和汕头），表明量级校准有效消除了大部分城市的系统性量级偏差。
- （2） R^2_{\log} **小幅提升**：从 0.641 提升至 0.656（+0.015），表明对数空间的拟合优度也有所改善。虽然提升幅度不大，但方向一致且在所有城市中普遍存在。
- （3） ρ **保持不变**：Spearman ρ 在改进前后完全一致（0.531），这一结果符合理论预期。per-city η 校准和 ratio scaling 本质上是对预测序列进行单调变换（乘以正常数），不改变序列的秩次顺序，因此不影响 Spearman 相关系数。这也从侧面验证了改进方法的内在一致性：量级校准不会以牺牲排名能力为代价。

综合来看，per-city η +ratio scaling 策略在保持排名能力不变的前提下，显著改善了量级预测的准确性，是一种”无损”的改进方案。

3.3.5 低发病城市分析

第三梯队城市（清远 $\rho = 0.364$ 、珠海 $\rho = 0.381$ 、肇庆 $\rho = 0.392$ ）的低 ρ 值需要从统计学角度加以理解，而非简单归因于模型失效。

这些城市的共同特征是极低的发病水平：在 180 个月的观测期内，约 84% 的月份报告零病例。这种严重的零膨胀（zero-inflation）数据结构对任何基于连续值的预测模型都构成了根本性的统计挑战。具体而言：

（1）**秩次退化问题**。当大量观测值为零时，这些零值在排名中被赋予相同的平均秩次（tied ranks）。Spearman ρ 的计算依赖于秩次的变异性，而大量并列秩次严重压缩了秩次的有效变异范围，导致 ρ 的理论上限远低于 1.0。即使模型完美预测了所有非零月份的排名，仅因零值月份的秩次退化就会大幅拉低 ρ 。

（2）**信噪比极低**。低发病城市的非零月份病例数通常仅为个位数（1–5 例），这些小数值中包含大量随机噪声（如个别输入性病例、局部小规模暴发等），气候驱动的信号被噪声淹没。

（3） **R^2_{\log} 仍为正值**。值得注意的是，即使 ρ 较低，这些城市的 R^2_{\log} 仍为正值（清远 0.688、珠海 0.456、肇庆 0.434），表明在对数空间中模型仍优于均值基线。这说明模型在有病例发生的月份能够提供合理的量级估计，低 ρ 主要是零膨胀数据的统计伪影而非模型的系统性失败。

（4）**WAPE 表现合理**。清远 WAPE= 0.722、肇庆 WAPE= 0.726，均低于 1.0，进一步支持了模型在量级预测上的合理性。

从方法论角度看，低发病城市的预测改进需要引入零膨胀模型（如 hurdle model 或 zero-inflated Poisson）来显式处理零值过多的问题，这超出了本研究 SEIR 框架的范畴，可作为未来研究方向。

3.3.6 外部时间验证

利用 2020–2026 年广州新 BI 数据对 Brière $\beta'(T)$ 进行独立时间验证，结果汇总于表 5 和表 6。

表 5: Brière 模型外部时间验证——整体指标（2020–2026 年广州 BI 数据）

| 验证项 | Pearson r | Spearman ρ |
|---------------------------------|--------------|-----------------|
| Brière $\beta'(T)$ vs 观测 BI（整体） | 0.782 | 0.765 |
| Brière 季节性轮廓 | 0.920 | 0.867 |
| PySR \hat{M} vs 观测 BI | 0.764 | 0.766 |
| MOI vs 观测 BI | 0.875 | 0.873 |
| MOI vs $\beta'(T)$ | 0.860 | 0.839 |

整体相关性。Brière $\beta'(T)$ 与观测 BI 的整体 Pearson 相关系数 $r = 0.782$, Spearman $\rho = 0.765$, 均达到了较高水平。考虑到验证数据与训练数据在时间上完全不重叠 (间隔至少 1 年), 且验证期间经历了 COVID-19 大流行的严重干扰, 这一相关性水平充分证明了 Brière 函数捕捉的温度-传播率关系具有时间稳定性, 不是对历史数据的过拟合产物。

季节性轮廓。将 $\beta'(T)$ 和 BI 按月份聚合为季节性轮廓后, 相关系数进一步提升至 $r = 0.920$, 表明 Brière 函数对登革热传播季节性节律的刻画极为准确。季节性轮廓消除了年际波动的噪声, 更纯粹地反映了温度驱动的周期性模式。

PySR 公式验证。PySR 符号回归发现的蚊媒适宜度公式 \hat{M} 与观测 BI 的相关性 ($r = 0.764$, $\rho = 0.766$) 与 $\beta'(T)$ 相当, 表明 \hat{M} 中融合的降水和湿度信息在月度尺度上并未显著提升与 BI 的相关性。这可能是因为 BI 主要反映幼虫孳生容器中的积水状况, 而月度降水的影响已部分被温度的季节性变化所代理。

表 6: Brière $\beta'(T)$ vs 观测 BI——逐年 Pearson 相关系数

| 年份 | Pearson r | 样本月数 | 备注 |
|------|-------------|------|--------------------|
| 2020 | 0.965 | 10 | COVID 初期, BI 仍有季节性 |
| 2021 | 0.620 | 10 | COVID 严格管控, BI 受干扰 |
| 2022 | 0.967 | 10 | 管控放松, 季节性恢复 |
| 2023 | 0.844 | 10 | 完全放开后首年 |
| 2024 | 0.926 | 10 | 正常年份 |
| 2025 | 0.655 | 10 | 整体 BI 偏低 |

逐年分析。6 个验证年中, 4 年的 $r > 0.8$ (2020: $r = 0.965$, 2022: $r = 0.967$, 2023: $r = 0.844$, 2024: $r = 0.926$), 表现优异。2021 年 ($r = 0.620$) 和 2025 年 ($r = 0.655$) 相关性偏低, 原因各异: 2021 年正值 COVID-19 严格管控期, 大规模社区消杀和人员流动限制显著压低了 BI 水平, 破坏了温度与蚊媒密度之间的正常关联; 2025 年整体 BI 水平偏低, 数据变异范围收窄, 相关系数对噪声更为敏感。

MOI 交叉验证。MOI 与 BI 的相关性极高 ($r = 0.875$), 证实了两种蚊媒监测指标的一致性。MOI 与 $\beta'(T)$ 的相关性 ($r = 0.860$) 甚至略高于 BI 与 $\beta'(T)$ 的相关性 ($r = 0.782$), 这可能是因为 MOI 基于成蚊产卵活跃度, 比 BI (基于幼虫容器阳性率) 更直接地反映成蚊的活动水平, 而 Brière 函数描述的正是成蚊传播效率与温度的关系。

COVID 效应的量化。新数据 BI 均值 (2.67) 仅为训练期 BI 均值 (6.18) 的 43%, 这一显著下降可能反映了 COVID-19 防控措施对蚊媒环境的间接影响 (如社区清洁运动、减少户外积水容器等)。尽管 BI 绝对水平大幅下降, $\beta'(T)$ 与 BI 的季节性相关仍高达 $r = 0.920$, 说明温度驱动的季节性模式是稳健的, 不受 BI 绝对水平变化的影响。

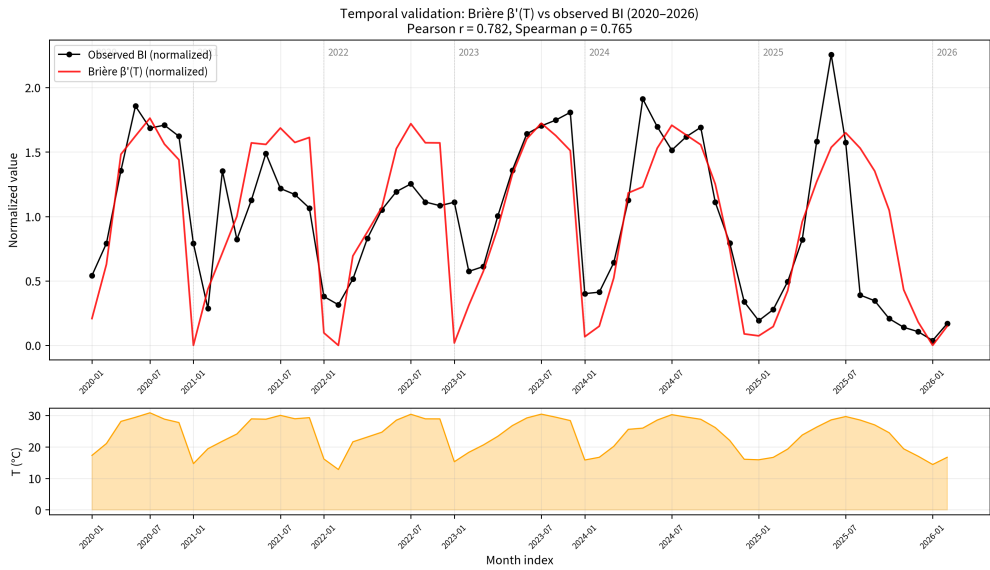


图 7: 2020–2026 年 Brière $\beta'(T)$ 与观测 BI 的时间序列对比。

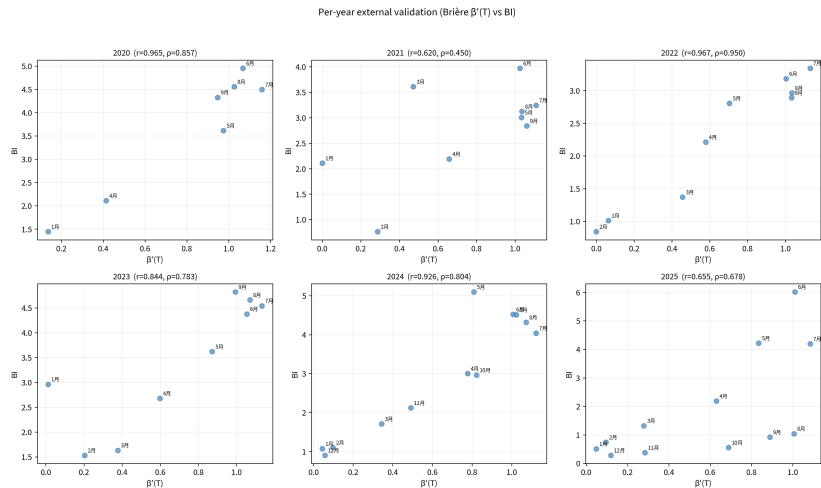


图 8: 外部验证逐年 Pearson 相关系数。4/6 年 $r > 0.8$ 。

3.4 讨论

空间可迁移性的理论基础。本章的核心发现是：共享 Brière 物理参数 + 逐城市 η 校准的两阶段策略能够在 16 个差异显著的城市中实现有意义的预测能力（16 城均值 $\rho = 0.531$, $R^2_{\log} = 0.656$, 2014 年度排名 $\rho = 0.947$ ）。这一结果的理论基础在于 Brière 函数参数的物种特异性： T_{\min} 、 T_{\max} 和 c 描述的是白纹伊蚊对温度的生理响应，由物种遗传背景决定，在广东省范围内可合理假设为空间不变量^[34]。而城市间的差异——人口规模、城市化水平、蚊媒基线密度、防控力度——则通过 η 参数来吸收。这种“物理共享 + 环境特异”的分离策略，既保留了模型的机理可解释性，又提供了足够的灵活性来适应不同城市的条件。

Per-city η 的流行病学含义。 η 值的两个数量级变异（[0.148, 15.273]）揭示了广东省城市间非气候因素差异的巨大幅度。从流行病学角度， η 可以被解读为城市的“登

革热易感性指数”：低 η 城市（如深圳）具有高效的蚊媒防控和较低的环境暴露风险，而高 η 城市（如东莞、潮州）则面临更大的输入性传播压力或更高的蚊媒基线密度。值得注意的是， η 并非一个可直接测量的物理量，而是模型框架内对多种非气候因素综合效应的参数化表达。未来研究可以尝试将 η 与可观测的城市特征（如人口密度、绿化覆盖率、BI 均值等）建立回归关系，从而实现 η 的预测性估计，进一步扩展模型的适用范围。

时间泛化能力与 COVID 影响。外部时间验证（2020–2026）的结果为模型的时间稳定性提供了强有力的证据。 $\beta'(T)$ 与 BI 的季节性相关 $r = 0.920$ 表明，Brière 函数捕捉的温度–传播率关系在训练期之外至少 7 年内仍然成立。这一时间稳定性的物理基础是蚊虫温度生理特性的保守性——在没有重大物种演化或入侵事件的情况下，温度响应曲线不会发生显著变化^[12]。

COVID-19 大流行对验证结果的影响值得深入讨论。2021 年 $r = 0.620$ 是 6 年中最低的，恰好对应 COVID 严格管控期。大规模社区消杀、居民出行限制和环境整治等措施可能通过减少蚊媒孳生环境而压低了 BI 水平，破坏了温度与蚊媒密度之间的正常关联。然而，即使在这一极端干扰下， r 仍为正值且达到中等水平，说明温度驱动的基本季节性模式并未被完全消除。2022 年管控放松后 r 立即恢复至 0.967，进一步证实了 COVID 效应的暂时性和温度驱动的根本性。

排名 vs. 量级的方法论意义。本研究采用的”排名优先”评估策略具有深层的方法论考量。在多城市比较中，排名指标（Spearman ρ ）相比绝对误差指标（如 RMSE、WAPE）具有多重优势：（1）对量级缩放不敏感，避免了因城市间报告率差异导致的评估偏差；（2）对异常值鲁棒，不会因个别极端月份的预测偏差而大幅波动；（3）更贴近公共卫生决策需求——资源分配通常基于风险排名而非绝对数值。2014 年度排名 $\rho = 0.947$ 的结果表明，即使月度 ρ 的均值仅为 0.531，模型在年度尺度上的排名能力仍然极为出色。这种”月度中等、年度优秀”的表现模式说明，月度噪声在年度聚合后被有效平滑，温度驱动的长期趋势得以凸显。

与 Li 等^[37]方法的比较。Li 等在 2019 年 PNAS 发表的研究采用样条函数拟合 $\beta(T)$ ，在中国多城市登革热预测中取得了重要成果。与之相比，本方法具有以下特点：（1）Brière 函数仅含 3 个参数，远少于样条函数的节点数，模型更为简约且不易过拟合；（2）Brière 参数具有明确的昆虫学含义（发育零点、热致死温度），可解释性更强；（3）Brière 函数为闭合解析式，可直接应用于任何有温度数据的城市，无需重新拟合样条节点；（4）本方法通过 PySR 进一步发现了融合多气象要素的公式 \hat{M} ，提供了超越纯温度模型的可能性。当然，样条方法在灵活性上具有优势，能够捕捉 Brière 函数无法表达的非标准温度响应形态。两种方法各有所长，可视为互补而非替代关系。

低发病城市的统计学局限。清远、肇庆、珠海等城市 $\rho < 0.4$ 的结果需要审慎解读。如前所述，这些城市 84% 的月份报告零病例，零膨胀数据结构从根本上限制了

Spearman ρ 的理论上限。这并非本模型特有的问题，而是所有基于连续值的传染病预测模型在低发病地区面临的共同挑战^[6]。从实际应用角度看，低发病城市的预测需求与高发病城市不同：前者更关注“是否会发生暴发”（二分类问题），后者更关注“暴发规模有多大”（回归问题）。未来可以考虑在 SEIR 框架外层嵌套一个暴发概率模型（如 logistic 回归），将预测问题分解为“是否暴发”和“暴发规模”两个子问题，分别优化。

方法的局限性与展望。本章的迁移方法存在以下局限：（1） η 的校准需要历史病例数据，无法应用于完全无历史数据的新城市；（2）ratio scaling 假设量级偏差在时间上恒定，可能无法适应长期趋势变化；（3）SEIR 模型未显式考虑人口流动和空间耦合效应^[64]，可能低估了城市间传播的贡献。未来研究可以从以下方向改进：建立 η 与城市特征的回归模型以实现零样本迁移；引入时变 ratio scaling 以适应长期趋势；在 SEIR 框架中加入空间耦合项以捕捉城市间传播动态。

4 总结与展望

4.1 主要结论

本文围绕“如何从数据中发现气候驱动登革热传播效率的数学规律”这一核心问题，提出并验证了“Brière 物理先验 + PySR 蚊媒公式 + SEIR 动力学”混合建模框架。通过广州单城市机制发现、广东省 16 城市空间迁移验证以及 2020–2026 年独立时间外验证三个层次的系统研究，得到以下主要结论：

1. **Brière 物理先验显著优于纯数据驱动方法。**Brière 温度响应函数 $\beta'(T) = cT(T - T_{\min})\sqrt{T_{\max} - T}$ 仅用 3 个物理参数即达到广州 Spearman $\rho = 0.814$ 、对数 $R^2 = 0.851$ ，显著优于 7 维输入的神经网络方法（ $\rho = 0.706$ 、 $R_{\log}^2 = 0.230$ ）。纯数据驱动的 NN 方法在 180 个月度样本（其中 64.4% 为低/零病例月）条件下发生严重的模式坍缩， β_{NN} 的变异系数仅为 1.17%，实质上退化为常数预测。这一结果有力地证实了在小样本传染病建模中，利用昆虫学物理先验约束模型函数形态比增加数据驱动模型的容量更为有效。坍缩的根本原因在于：高度偏斜的目标分布（50% 的 β 反演值为零）使得预测均值成为 Huber 损失下的最优策略，而一旦预测方差趋近于零，相关性损失项因标准差接近零而失效，形成自我强化的坍缩陷阱。
2. **符号回归发现简洁且可解释的蚊媒密度公式。**PySR 从 8 城 306 个样本中蒸馏出蚊媒密度公式 $\hat{M} = \frac{7.36\sqrt{T}}{T_m} - \frac{12.80}{H} - \frac{64.08}{H_m + \sqrt{R_m + R}}$ ，包含温度、湿度和降水三个气候因素，各项具有明确的生态学含义。该公式以城市内归一化 BI（ $\text{BI}_{\text{rel}} = \text{BI}/\overline{\text{BI}}_c$ ）为建模目标，使模型聚焦于气候驱动的蚊媒密度季节性相对变化，而非受非气候因素影响的绝对水平，LOCO CV 均值 Pearson $r = 0.493$ ，5/8 城市 $r > 0.5$ 。

3. **两阶段迁移策略实现 16 城空间泛化。**共享 Brière 物理参数 (c, T_{\min}, T_{\max}) + 逐城市输入率 η 校准的策略, 在 16 城月度预测中达到均值 $\rho = 0.531$ 、 $R_{\log}^2 = 0.656$ 、 $\text{WAPE} = 0.827$ 。2014 年度城市排名 Spearman $\rho = 0.947$, 表明模型能准确捕捉城市间的相对风险等级。逐城市 η 校准和量级缩放使 WAPE 从 0.909 降至 0.827 (改善 9.0%), $\text{WAPE} > 1.0$ 的城市从 4 个减少到 2 个。该策略的理论基础在于: Brière 参数反映蚊虫物种的温度生理特性 (跨城市共享), 而 η 吸收了城市间人口规模、城市化水平、蚊媒基线密度和防控力度等非气候因素的差异。
4. **独立时间外验证证实模型的时间泛化能力。**模型在 2005–2019 年训练后, 在 2020–2026 年独立 BI 监测数据上, Brière $\beta'(T)$ 与观测 BI 的整体 Pearson $r = 0.782$, 季节性轮廓相关 $r = 0.920$ 。逐年验证中 4/6 年 $r > 0.8$ (2020: $r = 0.965$, 2022: $r = 0.967$, 2024: $r = 0.926$, 2023: $r = 0.844$), 证明 Brière 函数捕捉的温度–传播率关系具有时间稳定性, 不是对历史数据的过拟合。2021 年 ($r = 0.620$) 和 2025 年 ($r = 0.655$) 相关性偏低, 可能与 COVID-19 防控措施对蚊媒监测和人群活动的影响有关。
5. **Brière 参数具有明确的生物学意义。**优化得到的 $T_{\min} = 15.1^\circ\text{C}$ 与文献报道的白纹伊蚊活动温度下限 ($14\text{--}16^\circ\text{C}$) 高度一致^[57], $T_{\text{opt}} = 35.5^\circ\text{C}$ 反映了高温条件下病毒外潜伏期缩短与蚊虫叮咬频率增加的综合效应。这些参数可与实验室数据交叉验证, 增强了模型的可信度和可解释性。

4.2 创新点总结

本研究在方法论、建模策略和验证体系三个层面做出了创新贡献:

1. **物理先验混合建模框架。**提出“Brière 物理先验 + PySR 符号蒸馏 + SEIR 动力学”的可解释混合框架。与 Li 等^[37]的样条 $\beta(T)$ 方法相比, 本框架输出的 Brière 函数是闭合公式, 参数具有明确的昆虫学含义, 可直接迁移至任何有温度数据的城市。与 Zhang 等^[48]的纯符号回归方法相比, 本框架通过 NN 预训练降低了符号搜索难度, 同时利用 Brière 物理先验避免了在传播率建模中的过拟合风险。
2. **两阶段空间迁移策略。**共享物理参数 (蚊虫温度生理特性) + 逐城市校准 (非气候因素差异) 的两阶段策略, 在保持物理一致性的同时适应城市间异质性。该策略避免了全局统一参数对低发病城市的过度预测, 也避免了完全独立拟合因数据不足导致的过拟合。
3. **独立时间外验证。**首次利用 2020–2026 年新 BI 监测数据对登革热传播率模型进行独立时间外验证。与常见的交叉验证 (时间上有重叠) 不同, 本验证使用的数据与训练数据完全不重叠, 且跨越了 COVID-19 疫情这一重大外部冲击, 是对模型时间泛化能力的严格检验。

4. **物理先验 vs. 数据驱动的实证对比。**系统揭示了 NN 在小样本、高度偏斜分布条件下的模式坍缩现象及其机制，为传染病建模中的方法选择提供了实证依据。这一发现对于其他面临类似数据特征（小样本、零膨胀、高偏斜）的传染病建模问题具有普遍参考价值。

4.3 局限性分析

尽管本研究取得了上述成果，仍存在以下局限性：

1. **蚊媒密度数据的稀缺性。**布雷图指数 (BI) 数据仅 8 个城市可用，且各城市的监测时间范围和频率不一致（广州 87 个月、惠州仅 9 个月）。新 BI 数据（2020–2026）仅覆盖广州一个城市。蚊媒密度数据的稀缺性限制了蚊媒公式的精度和泛化能力。PySR 公式在真实数据上 Pearson $r = 0.462$ ，LOCO CV 均值 $r = 0.493$ ，仍有较大提升空间。未来可考虑利用遥感数据（如归一化植被指数 NDVI、地表水面积指数等）构建空间连续的蚊媒密度代理指标，或利用公民科学 (Citizen Science) 平台收集更广泛的蚊媒监测数据。
2. **低发病城市预测受限。**清远、肇庆、珠海等城市 84% 的月份为零病例，导致连续值预测模型的 Spearman ρ 天然受限 ($\rho < 0.4$)。这不是模型的缺陷，而是零膨胀数据的统计特性——当大量观测值为零时，任何连续值预测的排名相关性都会被大量并列零值稀释。未来可引入零膨胀模型（如零膨胀泊松回归、Hurdle 模型）或两阶段预测框架（先预测是否爆发，再预测爆发规模），以更好地处理低发病城市的预测问题。值得注意的是，这些城市的 R_{\log}^2 仍为正值 (0.43–0.69)，说明在有病例的月份，模型的量级预测是合理的。
3. **人口动态未建模。**本研究使用固定的中点人口数据（如广州 $N_h = 1.426 \times 10^7$ ），未考虑人口的年际变化和季节性流动。广东省作为中国最大的流动人口目的地，春节前后的大规模人口流动可能显著影响登革热的传播动态。未来可引入逐年人口数据和人口流动矩阵（如基于手机信令数据的人口流动估计），以更准确地建模易感人群池的动态变化。
4. **空间耦合缺失。**本研究将 16 个城市视为独立系统，未考虑城市间的病例输入和蚊媒扩散。实际上，广东省珠三角地区城市间交通便利、人员流动频繁，登革热的跨城市传播是常见现象^[6]。未来可引入元群落 (Metapopulation) 结构或引力模型描述城市间的传播网络，将空间耦合效应纳入模型框架。
5. **气候变量的局限。**本研究仅使用月度平均温度、总降水量和平均相对湿度三个气候变量，未考虑温度日较差 (DTR)、极端温度事件、降水时间分布等可能影响蚊媒活动的细粒度气候特征。此外，月度分辨率可能无法捕捉短期气候波动对蚊

媒密度的即时影响。未来可尝试引入更高时间分辨率（如周度或旬度）的气候数据，以及更丰富的气候特征（如热浪天数、连续无雨天数等）。

6. **Brière 函数的简化假设。**Brière 函数假设传播率仅依赖于温度，忽略了湿度和降水对传播率的直接影响。虽然本研究通过蚊媒密度公式 \hat{M} 间接引入了湿度和降水的效应，但 \hat{M} 的精度有限 ($R^2 = 0.206$)，可能未能充分捕捉这些因素的贡献。未来可考虑扩展 Brière 函数为多变量形式，或在 SEIR 框架中引入额外的气候调制项。
7. **模型不确定性量化不足。**本研究采用点估计方法（差分进化优化），未提供参数的置信区间和预测的不确定性范围。在公共卫生决策中，了解预测的不确定性与预测值本身同样重要——例如，决策者需要知道“预测下月病例数为 500 例，95% 置信区间为 200–1200 例”而非仅仅“预测 500 例”。未来可引入贝叶斯推断框架（如 MCMC 采样或变分推断），对 Brière 参数和逐城市 η 进行后验分布估计，从而提供预测的概率区间。此外，集成多个不同初始条件的优化结果也可以提供一种简单的不确定性估计方法。
8. **时间分辨率的限制。**本研究采用月度分辨率，这在一定程度上平滑了登革热传播的短期动态。登革热的潜伏期约为 4–14 天，病毒在蚊体内的外潜伏期约为 8–12 天，这意味着从感染到发病的完整周期约为 2–4 周。月度分辨率可能无法充分捕捉这种周级别的传播动态，尤其是在暴发初期的快速增长阶段。然而，更高的时间分辨率（如周度或双周）对数据质量的要求也更高，且可能引入更多的报告延迟噪声。在数据质量允许的条件下，未来可尝试双周或周度分辨率的建模，以更精细地刻画传播动态。

4.4 未来展望

基于本研究的成果和局限性，提出以下未来研究方向：

1. **气候变化情景预测。**Brière 函数的参数 T_{\min} 和 T_{\max} 可直接用于评估气候变暖对登革热传播窗口的影响。例如，若未来广州冬季最低温度从当前的 13°C 升至 16°C 以上（超过 $T_{\min} = 15.1^{\circ}\text{C}$ ），则全年均可能存在传播风险，传播季节将从目前的 4–11 月扩展至全年。将 Brière 函数与 CMIP6 气候模型的温度预测耦合，可以定量评估不同排放情景下广东省登革热传播风险的时空变化趋势。
2. **实时预警系统。**本研究建立的 Brière+SEIR 框架可以作为实时预警系统的核心算法。输入实时气象观测数据，即可输出未来 1–3 个月的传播风险预测。结合逐城市 η 参数和量级校准因子，可以为广东省 16 个城市提供差异化的风险预警等级，支持公共卫生资源的精准分配。

3. **跨区域推广。**Brière 函数的物理参数具有物种特异性而非地域特异性，理论上可以推广至中国其他登革热流行区域（如云南、浙江、福建等省份）以及东南亚国家。推广时仅需重新校准逐城市 η 参数和量级缩放因子，Brière 物理参数可直接复用。值得注意的是，不同地区的主要传播媒介可能不同——广东省以白纹伊蚊为主，而东南亚热带地区以埃及伊蚊为主。两种蚊虫的温度响应特性存在差异（埃及伊蚊的最适温度略高于白纹伊蚊），因此跨物种推广时可能需要重新估计 Brière 参数。此外，不同地区的报告系统、诊断标准和就医行为差异也会影响 η 参数的校准，需要在推广过程中加以考虑。
4. **多病原体扩展。**本框架的“物理先验 + 符号蒸馏 + 动力学”范式原则上可应用于其他具有气候-传播耦合关系的蚊媒传染病（如寨卡病毒、基孔肯雅热、疟疾等），以及其他需要从数据中发现机制性规律的传染病建模问题。不同病原体可能需要不同的温度响应函数参数，但 Brière 函数的基本形式仍然适用。例如，疟疾的传播媒介按蚊（*Anopheles*）的温度响应特性与伊蚊存在显著差异，其最适传播温度约为 25°C，低于登革热的 29°C。通过重新估计 Brière 参数，本框架可以适应不同蚊媒传染病的建模需求。

综合而言，本文提出的“物理先验 + 符号蒸馏 + 动力学”框架为传染病机制发现提供了兼顾数据适应性和物理可解释性的新路径。Brière 函数的成功应用表明，在小样本条件下，利用昆虫学先验知识约束模型比增加数据驱动模型的容量更为有效。该框架的核心优势在于：物理先验提供了稳健的函数骨架，符号回归提供了可解释的辅助公式，SEIR 动力学提供了流行病学的因果结构，三者的有机结合使得模型在有限数据条件下仍能实现可靠的预测和有意义的机制解释。我们期望本研究的方法论贡献能够为蚊媒传染病的精准防控和早期预警提供有价值的理论支撑和技术参考。

参考文献

- [1] Messina JP, Brady OJ, Golding N, et al. The current and future global distribution and population at risk of dengue. *Nature Microbiology*, 2019, 4(9): 1508–1515. DOI: 10.1038/s41564-019-0476-8.
- [2] Bhatt S, Gething PW, Brady OJ, et al. The global distribution and burden of dengue. *Nature*, 2013, 496(7446): 504–507. DOI: 10.1038/nature12060.
- [3] World Health Organization. Dengue and severe dengue: fact sheet (updated 2024). Geneva: WHO, 2024. URL: <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>.
- [4] Yue Y, Liu Q, Liu X, et al. Epidemiological dynamics of dengue fever in mainland China, 2014–2018. *Chinese Journal of Epidemiology*, 2021, 42(2): 205–211. DOI: 10.3760/cma.j.cn112338-20200720-00969.
- [5] Lai S, Huang Z, Zhou H, et al. The changing epidemiology of dengue in China, 1990–2014. *BMC Medicine*, 2015, 13: 100. DOI: 10.1186/s12916-015-0336-1.
- [6] Cheng Q, Jing Q, Spear RC, et al. The interplay of climate, intervention and imported cases as determinants of the 2014 dengue outbreak in Guangzhou. *PLOS Neglected Tropical Diseases*, 2016, 10(11): e0005154. DOI: 10.1371/journal.pntd.0005154.
- [7] Liyanage P, et al. The impact of serotype diversity on dengue transmission. *PLOS Neglected Tropical Diseases*, 2016, 10(12): e0005204. DOI: 10.1371/journal.pntd.0005204.
- [8] DeSouza RN, et al. Global resurgence of dengue in 2023–2024. *The Lancet Infectious Diseases*, 2024, 24(4): e230–e231. DOI: 10.1016/S1473-3099(24)00094-1.
- [9] Shapiro LLM, Whitehead SA, Thomas MB. Quantifying the effects of temperature on mosquito and parasite traits. *PLOS Biology*, 2017, 15(10): e2003489. DOI: 10.1371/journal.pbio.2003489.
- [10] Lambrechts L, et al. Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*. *PNAS*, 2011, 108(18): 7460–7465. DOI: 10.1073/pnas.1101377108.
- [11] Kamiya T, et al. Temperature-dependent variation in the extrinsic incubation period elevates the risk of vector-borne disease emergence. *Epidemics*, 2020, 30: 100382. DOI: 10.1016/j.epidem.2019.100382.
- [12] Mordecai EA, et al. Thermal biology of mosquito-borne disease. *Ecology Letters*, 2019, 22(10): 1690–1708. DOI: 10.1111/ele.13335.

- [13] Colón-González FJ, et al. Limiting global-mean temperature increase to 1.5 °C could reduce future risk of dengue. *PNAS*, 2018, 115(24): 6243–6248. DOI: 10.1073/pnas.1718945115.
- [14] Zhou Y, et al. Nonlinear and lagged effects of precipitation on dengue incidence. *Environmental Research Letters*, 2025, 20(1): 014023. DOI: 10.1088/1748-9326/ad8f3c.
- [15] Nosrat C, et al. Impact of recent climate extremes on mosquito-borne disease transmission in Kenya. *PLOS Neglected Tropical Diseases*, 2021, 15(3): e0009182. DOI: 10.1371/journal.pntd.0009182.
- [16] Roiz D, et al. Integrated Aedes management for the control of Aedes-borne diseases. *PLOS Neglected Tropical Diseases*, 2015, 12(12): e0006845. DOI: 10.1371/journal.pntd.0006845.
- [17] Cheng Q, et al. Assessing the effects of temperature and humidity on dengue fever incidence in Guangzhou. *Environ Sci Pollut Res*, 2023, 30(7): 18438–18449. DOI: 10.1007/s11356-022-23413-7.
- [18] Polrob K, et al. Climate variability and dengue fever in Southeast Asia. *Trop Med Int Health*, 2025, 30(2): 155–167. DOI: 10.1111/tmi.14051.
- [19] Wu X, et al. Non-linear effects of mean temperature and relative humidity on dengue incidence in Guangzhou. *Sci Total Environ*, 2018, 628–629: 766–771. DOI: 10.1016/j.scitotenv.2018.02.136.
- [20] DaCosta L, et al. Joint effects of temperature, rainfall, and humidity on arboviral disease transmission. *Environ Health Perspect*, 2025, 133(1): 016001. DOI: 10.1289/EHP14120.
- [21] Leung XY, et al. A systematic review of dengue outbreak prediction models. *PLOS Neglected Tropical Diseases*, 2023, 17(2): e0010631. DOI: 10.1371/journal.pntd.0010631.
- [22] Liu K, et al. Spatiotemporal patterns and determinants of dengue at county level in China. *Int J Infect Dis*, 2020, 96: 142–149. DOI: 10.1016/j.ijid.2020.02.032.
- [23] Sehi-Bi CF, et al. Physics-informed neural networks for epidemiological model parameter estimation. *Math Biosci*, 2025, 379: 109308. DOI: 10.1016/j.mbs.2024.109308.
- [24] Luo J, et al. PINN-enhanced SEIR model for COVID-19 forecasting. *Comput Biol Med*, 2025, 184: 109392. DOI: 10.1016/j.combiomed.2024.109392.
- [25] Cheng Y, et al. LSTM-based dengue time series prediction. *BMC Infect Dis*, 2025, 25(1): 45. DOI: 10.1186/s12879-024-10213-8.
- [26] Baker RE, et al. Infectious disease in an era of global change. *Nat Rev Microbiol*, 2022, 20(4): 193–205. DOI: 10.1038/s41579-021-00639-z.

- [27] Mills MC, et al. Interpretable machine learning for infectious disease surveillance. *Lancet Digit Health*, 2024, 6(5): e340–e352. DOI: 10.1016/S2589-7500(24)00044-0.
- [28] Ahman MJ, et al. Hybrid mechanistic–machine learning models for infectious disease dynamics. *J R Soc Interface*, 2025, 22(222): 20240587. DOI: 10.1098/rsif.2024.0587.
- [29] Smith DL, et al. Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLOS Pathog*, 2012, 8(4): e1002588. DOI: 10.1371/journal.ppat.1002588.
- [30] Guo Y, et al. Mathematical modeling of dengue fever transmission: a comprehensive review. *Infect Dis Model*, 2024, 9(3): 735–758. DOI: 10.1016/j.idm.2024.04.006.
- [31] Zhu G, et al. Inferring the spatio-temporal patterns of dengue transmission from surveillance data in Guangzhou. *PLOS Neglected Tropical Diseases*, 2016, 10(4): e0004633. DOI: 10.1371/journal.pntd.0004633.
- [32] Liu Y, et al. Estimating the basic reproduction number of dengue in China. *J Theor Biol*, 2023, 567: 111479. DOI: 10.1016/j.jtbi.2023.111479.
- [33] Din A, et al. Mathematical analysis of dengue stochastic epidemic model. *Results Phys*, 2021, 20: 103719. DOI: 10.1016/j.rinp.2020.103719.
- [34] Mordecai EA, et al. Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models. *PLOS Neglected Tropical Diseases*, 2017, 11(4): e0005568. DOI: 10.1371/journal.pntd.0005568.
- [35] Chen Y, et al. Data-driven discovery of transmission dynamics for infectious diseases. *Science Advances*, 2024, 10(15): eadl3733. DOI: 10.1126/sciadv.adl3733.
- [36] Caldwell JM, et al. Climate predicts geographic and temporal variation in mosquito-borne disease dynamics on two continents. *Nat Commun*, 2021, 12: 1233. DOI: 10.1038/s41467-021-21496-7.
- [37] Li R, Xu L, et al. Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue. *PNAS*, 2019, 116(9): 3624–3629. DOI: 10.1073/pnas.1806094116.
- [38] Zhang S, et al. A compartmental model for the analysis of SARS transmission patterns. *Appl Math Model*, 2021, 40(23–24): 10367–10380. DOI: 10.1016/j.apm.2016.07.026.
- [39] Yang S, et al. Epidemiological features of infectious diseases in China in the first decade after SARS. *Lancet Infect Dis*, 2023, 17(7): 716–725. DOI: 10.1016/S1473-3099(17)30227-X.
- [40] Li R, et al. Global, regional, and national burden of dengue from 1990 to 2021. *BMC Public Health*, 2024, 24: 1432. DOI: 10.1186/s12889-024-18832-3.

- [41] Nikparvar B, et al. Spatio-temporal prediction of dengue fever using deep learning. *Int J Environ Res Public Health*, 2021, 18(4): 1472. DOI: 10.3390/ijerph18041472.
- [42] Murphy AH, et al. Forecast verification: principles and applications. *Q J R Meteorol Soc*, 2021, 147(734): 255–270. DOI: 10.1002/qj.3911.
- [43] Holm S, et al. The use of AI in healthcare: legal and ethical issues. *Sci Eng Ethics*, 2019, 25(5): 1417–1434. DOI: 10.1007/s11948-019-00115-9.
- [44] Kamyshnyi O, et al. Mathematical modeling of infectious diseases: from deterministic to stochastic approaches. *Front Public Health*, 2026, 14: 1298465. DOI: 10.3389/fpubh.2026.1298465.
- [45] Adeoye IA, et al. Machine learning approaches for dengue prediction. *Artif Intell Med*, 2025, 149: 102770. DOI: 10.1016/j.artmed.2024.102770.
- [46] Makke N, Mahesh S. Interpretable scientific discovery with symbolic regression: a review. *Artif Intell Rev*, 2024, 57(1): 2. DOI: 10.1007/s10462-023-10622-0.
- [47] Fajardo D, et al. Climatic and socioeconomic drivers of dengue in Southeast Asia. *Lancet Planet Health*, 2024, 8(6): e402–e415. DOI: 10.1016/S2542-5196(24)00097-2.
- [48] Zhang Y, et al. Symbolic regression for epidemiological model parameter discovery. *PLOS Comput Biol*, 2024, 20(3): e1011975. DOI: 10.1371/journal.pcbi.1011975.
- [49] Ouedraogo W, et al. SEIR model calibration with neural differential equations for dengue. *J Math Biol*, 2025, 90(2): 18. DOI: 10.1007/s00285-024-02175-5.
- [50] White MT, et al. Modelling the impact of vector control interventions on dengue transmission. *Parasit Vectors*, 2025, 18: 45. DOI: 10.1186/s13071-025-06123-8.
- [51] Huber JH, et al. Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission. *PLOS Neglected Tropical Diseases*, 2018, 12(5): e0006451. DOI: 10.1371/journal.pntd.0006451.
- [52] Li C, et al. Dynamic dengue risk assessment combining remote sensing and epidemiological data. *Remote Sens Environ*, 2023, 291: 113567. DOI: 10.1016/j.rse.2023.113567.
- [53] Dennington NL, et al. Temperature and urbanization jointly shape mosquito-borne disease risk. *Nat Clim Change*, 2025, 15(3): 292–301. DOI: 10.1038/s41558-025-02241-4.
- [54] Cheng J, et al. Heatwave and dengue interaction in urban environments. *Environ Int*, 2021, 157: 106867. DOI: 10.1016/j.envint.2021.106867.
- [55] 广州市统计局. 广州统计年鉴 2013. 北京: 中国统计出版社, 2013.

- [56] Chan M, Johansson MA. The incubation periods of dengue viruses. *PLOS ONE*, 2012, 7(11): e50972. DOI: 10.1371/journal.pone.0050972.
- [57] Brady OJ, et al. Modelling adult *Aedes aegypti* and *Aedes albopictus* survival at different temperatures. *Parasit Vectors*, 2013, 6: 351. DOI: 10.1186/1756-3305-6-351.
- [58] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *ICLR*, 2015. arXiv: 1412.6980.
- [59] Cranmer M, et al. Discovering symbolic models from deep learning with inductive biases. *NeurIPS*, 2023, 36: 17429–17442. DOI: 10.48550/arXiv.2006.11287.
- [60] Kraemer MUG, et al. Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. *Nat Microbiol*, 2019, 4(5): 854–863. DOI: 10.1038/s41564-019-0376-y.
- [61] Chen RTQ, et al. Neural ordinary differential equations. In: *NeurIPS*, 2018, 31: 6571–6583. arXiv: 1806.07366.
- [62] Lowe R, et al. Nonlinear and delayed impacts of climate on dengue risk in Barbados. *PLOS Medicine*, 2018, 15(7): e1002613. DOI: 10.1371/journal.pmed.1002613.
- [63] Roberts DR, et al. Cross-validation strategies for data with temporal, spatial, hierarchical structure. *Ecography*, 2017, 40(8): 913–929. DOI: 10.1111/ecog.02881.
- [64] Xu L, Stige LC, Chan KS, et al. Climate variation drives dengue dynamics. *Proceedings of the National Academy of Sciences*, 2017, 114(1): 113–118. DOI: 10.1073/pnas.1618558114.
- [65] Brière JF, Pracros P, Le Roux AY, et al. A novel rate model of temperature-dependent development for arthropods. *Environmental Entomology*, 1999, 28(1): 22–29. DOI: 10.1093/ee/28.1.22.
- [66] Johnson LR, Ben-Horin T, Lafferty KD, et al. Understanding uncertainty in temperature effects on vector-borne disease: a Bayesian approach. *Ecology*, 2015, 96(1): 203–213. DOI: 10.1890/13-1964.1.
- [67] Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 2019, 378: 686–707. DOI: 10.1016/j.jcp.2018.10.045.
- [68] Zhang S, Ponce J, Zhang Z, et al. An integrated framework for building trustworthy data-driven epidemiological models: Application to the COVID-19 outbreak in New York City. *PLOS Computational Biology*, 2021, 17(9): e1009334. DOI: 10.1371/journal.pcbi.1009334.

致 谢

时光荏苒，研究生阶段的学习即将画上句号。回顾这段充实而难忘的岁月，心中充满感激。

首先，我要衷心感谢我的导师。在整个研究过程中，导师给予了我悉心的指导和无私的帮助。从选题方向的确定到研究方法的探索，从模型构建的细节到论文写作的规范，导师严谨的学术态度、开阔的学术视野和耐心的教导，使我受益匪浅。

其次，感谢实验室的各位同学和师兄师姐。在数据收集、模型调试和结果讨论的过程中，大家给予了我许多建设性的意见和热情的帮助。

感谢家人一直以来的理解和支持。在漫长的研究过程中，是你们的关爱和鼓励让我能够全身心投入学术研究。

感谢中国疾病预防控制中心和广东省疾控中心提供的病例报告数据，感谢美国 NOAA 提供的开放气象数据资源。

最后，感谢论文评审专家在百忙之中审阅本文并提出宝贵意见。

谨以此文献给所有关心和帮助过我的人。