使用机制模型和统计时间序列模型对 COVID-19 发病率进行短期概率预测的比较

Nicolas Banholzer[1,2], Thomas Mellan[3],H Juliette T Unwin[3], Stefan Feuerriegel[4,5], Swapnil Mishra[6],和 Samir Bhatt[3,7,*]

[1]苏黎世联邦理工学院管理、技术与经济系，瑞士苏黎世[2]伯尔尼大学社会与预防医学研究所，瑞士伯尔尼[3]伦敦帝国理工学院公共卫生学院，英国伦敦[4]慕尼黑大学管理学院，德国慕尼黑[5]慕尼黑机器学习中心 (MCML)，德国慕尼黑[6]新加坡国立大学及新加坡国立大学健康系统公共卫生学院与数据科学研究所，新加坡[7]哥本哈根大学公共卫生系，丹麦哥本哈根[*]通讯作者：
samir.bhatt@sund.ku.dk

arXiv:2305.00933v1 [stat.AP] 1 May 2023

摘要

传染病传播的短期预测是风险评估和公共卫生决策中的关键组成部分。尽管人们已经开发了多种用于短期预测的模型，但关于这些模型相对性能的诸多问题仍有待解答。在此，我们将基于更新方程的常用机制模型所生成的短期概率预测，与统计时间序列模型的预测值进行比较。我们的实证比较基于美国六个大州在疫情第一年期间的 COVID-19 每日发病数数据。我们发现，平均而言，统计时间序列模型的概率预测总体上至少与机制模型的预测值同样准确。此外，统计时间序列模型能够更好地捕捉波动性。我们的研究结果表明，通过对疾病动态进行假设而融入机制模型的领域知识，并不能提高疾病发病率的短期预测精度。然而，我们指出，预测往往只是众多研究目标之一，因此机制模型仍然具有重要意义，例如，可用于模拟疫苗的影响或新变异株的出现。

关键词 ： 概率预测、机制模型、时间序列模型、COVID-19

1 引言

流行病预测旨在预测疾病发病率等观测指标的发展情况，为卫生政策提供参考 [1–6], 例如，影响资源分配或公共卫生干预措施的实施时机等决策，以控制疫情暴发。预测可以针对长短不同的时间范围，服务于不同目的 [7]。长期预测（数月至数年）通常分析假设情景，例如疫苗接种率提高或出现一种传播力更强的新病毒株时，流行病将如何发展。相比之下，短期预测（数日至数周）估计近期将会发生什么。因此，短期预测的表现通常通过与实际观测结果进行比较来评估。

流行病预测可以采用不同类型的模型，但大体上可分为机制模型和统计时间序列模型。机制模型 [8, 9] 会对数据生成过程作出假设，并基于对疾病动态的明确数学表达式。它们建立了观测到的公共卫生结果（如报告病例数或死亡人数）与未观测到的、与传播直接相关的结果（如感染人数或瞬时再生数）之间的联系。将这些联系形式化，需要掌握特定疾病的知识，例如代际间隔或潜伏期等流行病学参数。统计时间序列模型 [10–12] 无需对数据生成过程作出假设即可进行估计。相反，这些模型会识别观测时间序列中的统计模式，例如利用数据中的自相关。由于无需领域知识即可识别这些模式，这类模型的统计特性得到了充分刻画，因此可用于各种预测应用。

本文比较了机制模型与统计时间序列模型给出的短期概率预测。机制模型通过考虑流行病学过程的不确定性来提供概率预测。例如，机制模型会使用对代际间隔或序列间隔的先验估计，而这些估计存在不确定性，尤其是在流行病早期。统计时间序列模型则通过贝叶斯推断提供概率预测，其中模型参数被视为需要积分处理的随机量。比较机制模型与统计时间序列模型给出的概率预测，有助于了解领域知识带来的额外价值。如果机制模型的预测值更准确，这

将表明，关于疾病动态和流行病学参数的知识能够为短期疫情预测带来额外价值。

我们比较了三种统计时间序列模型（季节性自回归积分移动平均模型（SARIMA） [10], Prophet [11, 13],高斯过程（GP） [12]）的短期概率预测性能，并与两种基于更新方程的常用机制模型（EpiEstim [8, 14] 和 EpiNow2 [9, 15]）进行比较。我们指出，基于更新方程的机制模型与易感者-感染者-康复者模型 [16]之间存在等价关系。我们的实证比较基于美国六个大州在疫情首年（2020年3月15日（美国大流行开始之日）至2021年3月15日）COVID-19发病率数据。我们使用最近两个月的数据，针对每一天反复估计模型；与相关研究 [7, 17, 18], 类似，我们对接下来两周进行概率预测。我们通过回顾性计算基于模型的预测值与观测发病率之间的差异来评估预测性能。

2 材料与方法

2.1 数据

我们使用2020年1月19日至2021年3月15日期间美国六个大州（亚利桑那州、加利福尼亚州、伊利诺伊州、马里兰州、新泽西州和纽约州）报告的 COVID-19 每日新增病例数数据。数据通过 R软件包 `covidcast` 0.4版2 [19], 获取，该软件包汇总了约翰斯·霍普金斯冠状病毒资源中心的 COVID-19 每日确诊病例数数据 [20] ，并使用世界银行的人口数据计算发病率（即每10万人新增病例数）。数据截至2022年1月1日，因此包含对实时数据的修订。这些修订在少数情况下导致新增病例数为负值，例如，删除大量重复记录时。在数据预处理中，我们将负值（这类数值不合理）替换为前一天的数值。

2.2 预测任务

预测任务是对美国各州 $s = 1, \ldots, 6$ 在第 t 天的 COVID-19 发病率进行每日两周预测。令 $N_{s,t}$ 表示 COVID-19 新增病例的观测数量，令 $N_{s,t+1}$ 至 $\hat{}N_{s,t+14}$ 表示新增病例数的每日提前 2 周预测值。基于模型的预测

在2020年3月15日至2021年3月15日之间的51周（即1年）内生成。每周日对从下周一开始的14天进行预测。每个预测值都基于使用截至第 t 天的历史时间序列估计的模型。具体而言，使用距 t 两个月前的观测数据（即 $N_{s,t-56},\ldots,N_{s,t}$）作为模型 m 的训练数据，以生成概率预测 $\hat{N}^{m,s}_{d,t+k}$，其中 $\hat{N}$ 是预测的新增病例数，$d=1,\ldots,D$ 表示从模型 $F^m_{s,t}$ 的后验分布中抽取的样本（抽样总数为 D），而模型 $m=1,\ldots,M$ 的总数为 M。

2.3 预测模型概述

表 1 概述了本研究所考虑的预测模型。各模型的性质及其实现将在接下来的两个章节中详细介绍。请注意，我们通过调整重要的模型参数或设定，为每个模型考察了不同的建模选择。在结果部分，我们首先报告每个模型中总体预测评分最高的建模选择。随后，我们选择这些建模选择所对应的模型进行相互比较。

2.4 机制模型

我们使用以下基于更新方程的机制模型：EpiEstim [8, 14] 和 EpiNow2 [9, 15]。

(i) EpiEstim

EpiEstim 建立在 Cori 等人 [8] 提出的框架之上，用于估算时变再生数；该方法已广泛用于监测传染病传播 [5, 21–23]。该框架假设新增感染数可通过泊松更新过程建模，其中新增感染者人数与现患病例数乘以其传染性成正比 [24]。形式上，第 t 天在 s 州的预期发病率估算为

$$E[I_{s,t}] = R_{s,t}\sum_{j=0}^{n} w_j I_{t-j,s}, \tag{1}$$

表1。 概率预测模型和建模选择摘要。

机制模型	
(i) EpiEstim	• 估计固定窗口 τ 内的时变再生数。 • 再生数的后验分布通过解析方法推导得出。 • 假设病例按感染日期记录，并采用恒定的时间偏移。 • 不确定的代际间隔分布。 ⇒ 建模选择： τ 的不同取值。
(ii) EpiNow2	• EpiEstim 的扩展。 • 时变再生数建模的不同方法。 • 从感染到报告的延迟分布存在不确定性。 • 针对星期效应进行调整的观测模型。 ⇒ 建模选择：采用不同方法对时变再生数建模。
统计时间序列模型	
(i) SARIMA	• 季节性自回归移动平均模型。 • 通过每周差分处理星期效应。 ⇒ 建模选择：模型非季节性部分的不同构成。
(ii) Prophet	• 加性时间序列模型。 • 具有线性增长模型的趋势分量。 • 增长率在由参数 τ 确定的估计变点处发生变化。 • 星期效应的季节性分量。 ⇒ 建模选择： τ的不同取值。
（iii）高斯过程	• 组合核函数，以建模时间序列的不同方面。 • 用于趋势的长程平方指数核。 • 用于短期变化的短程平方指数核。 • 用于星期效应的周期核。 ⇒ 建模选择：短程核的不同长度。

其中$R_{s,\ t}$ 为时变再生数，w 为代际时间（通常假设其服从代际间隔分布 [25]），而 $I_{t-j,\ s}$ 分别表示在 t 之前$j = 0,\ldots,n$ 天的新增感染数。Cori 等人使用贝叶斯方法估计 $R_{s,\ t}$，并通过假设 $R_{s,\ t}$服从共轭伽马先验分布，解析地推导出$R_{s,\ t}$ 的后验分布。Cori 等人的框架已在 R 软件包 `EpiEstim`（版本 2.2-4） [14]中实现。

我们应用 EpiEstim，并假设潜伏期恒定；在这种情况下，确诊病例的发病率与感染的发病率完全相同，只是时间上有所偏移。我们采用不确定的代际间隔分布，其依据是对 SARS-CoV-2 [26]代际时间均值（4.2，95% 区间为 3.3 至 5.3）和标准差（4.9，95% 区间为 3.0 至 8.3）的先前估计。此外，我们假设基本再生数 R_0 的伽马先验分布，其均值为 1.10，标准差为 0.04，与相关研究 [27]一致。EpiEstim 在时间窗口 $[t-\tau+1;t]$ 上估计再生数 $R_{s,\ t}$。参数 τ 可控制 $R_{s,\ t}$估计值的变异程度。较大的时间窗口（即 τ 的取值较高）会降低 R_t 估计值的变异，并使发病率的变化更加平滑。实证分析中，我们考察了不同的建模选择，将时间窗口从一天调整到两周。

EpiEstim 返回每个时间窗口中 R_t 的估计均值和标准差。我们使用最后一个时间窗口（即进行预测之前的那个时间窗口）的均值和标准差，计算R_s的后验伽马分布的形状参数和尺度参数$_{,\ t}$。基于此，对于代际间隔分布的每个样本，我们抽取后验抽样值 $R^1_{s,\ t}$，.. ，$R^D_{s,\ t}$，并将其与观测发病率 $N_{s,\ 0}$，.. ，$N_{s,\ t}$结合，用于预测未来发病率 $\hat{}N^d_{s,\ t+1}$至 $\hat{}N^d_{s,\ t+21}$。发病率采用泊松分布建模。预测结果使用 R软件包 `projections` 版本 0.5.4 [28]生成。

（*ii*）*EpiNow2*

EpiNow2 基于 Abbott 等人提出的框架 [9]来估计时变再生数，并已用于近期研究，以预测 COVID-19 [7, 9, 29] 的发病率。EpiNow2 采用与 EpiEstim 类似的方法，但进行了多项扩展。首先，可以使用随机游走或高斯过程等方法对再生数估计值的不确定性进行建模。其次，可以调整潜在新增感染数，以考虑仍易感于病毒的人口比例。第三，EpiNow2 纳入了一种观测

模型，通过不确定的延迟分布将预期新增感染数映射为观测病例数，并对从感染到病例报告的时间进行调整。第四，可以根据星期效应调整观测病例数。Abbott 等人提出的框架由广受欢迎的 R软件包 `EpiNow2` (版本 1.3.2) [15]实现。

我们通过纳入观测模型来应用 EpiNow2。延迟分布的不确定性通过代际时间的均值和标准差先验估计值 [26] （与 EpiEstim 相同）、潜伏期 [30], 和报告延迟 [31]来设定。此外，我们假设基本再生数 R_0 服从对数正态先验，其参数设定使其均值和标准差与 EpiEstim 中伽马分布的相同。在实证分析中，我们采用不同模型估计时变再生数，包括非参数反向计算、周随机游走，或按照相关研究 [32]设定参数的高斯过程。与 EpiEstim 不同的是，观测到的病例采用负二项分布而非泊松分布建模，并对星期效应进行调整。

2.5 统计时间序列模型

我们使用以下统计时间序列模型：季节性自回归积分移动平均模型（SARIMA） [10], Prophet [11], 和高斯过程（GP） [12]。与机制模型不同，我们直接对发病率对数$Y_{s,\ t}$ （即每日每10万人确诊新增病例数的对数）进行建模。统计时间序列模型的预测值会经过转换，以得到新增病例数$N_{s,\ t}$ ，从而与机制模型的预测值进行比较。

(i) SARIMA

SARIMA模型 [10] 已被多项研究用于预测COVID-19 [33–37]的发病率。它们是两个组成部分的线性组合：自回归（AR）部分和移动平均（MA）部分。在AR部分中，结果$Y_{s,\ t}$ 仅取决于滞后的结果$Y_{s,\ t-j}$，其中 j表示时间滞后。在MA部分中，$Y_{s,\ t}$ 仅取决于滞后的误差$\epsilon_{s,\ t-j} = Y_{s,\ t-j} - Y_{s,\ t-j-1}$, 即观测结果与估计结果之间的差值。

一个SARIMA($p = 1, d = 0, q = 1$)$(0,1,0)_7$ 其中包含一个自回归成分和一个移动平均成分，可写为

$$\Delta Y_{s,t} = \alpha + \beta \Delta Y_{s,t-1} + \phi \Delta \epsilon_{s,t}, \tag{2}$$

其中 $\Delta Y_{s,t} = Y_{s,t} - Y_{s,t-7}$ 和 $\Delta \epsilon_{s,t} = \epsilon_{s,t} - \epsilon_{s,t-7}$分别表示相应参数。请注意，我们在模型的季节性部分使用发病率对数的周差分，以考虑星期效应。在实证分析中，我们通过改变自回归（p）、移动平均（q）和差分（d）的阶数，考察了不同的非季节性部分。我们对模型截距（α）、自回归参数（β）和移动平均参数（ϕ）使用弱信息先验

$$\alpha \sim \text{Normal}(0, 2.5, 6), \quad \beta \sim \text{Normal}(0, 0.5), \quad \phi \sim \text{Normal}(0, 0.5). \tag{3}$$

该模型使用 R软件包 `bayesforecast` 版本 1.0.1 [38进行估计]

(ii) Prophet

Prophet [11] 是一款广泛应用于各类预测任务的热门开源软件，并已被多项研究用于对 COVID-19 [35,39,40]发病率建模。它是一种加法模型，由三个部分组成，分别用于建模时间序列的趋势、季节性和节假日效应（此处未考虑）。其中关键部分是趋势。趋势可以采用线性增长或逻辑斯蒂增长建模，并允许增长率在时间点 s_j, $j = 1$、.....、J发生变化。增长率的变化 δ_j采用层级先验 $\delta_j \sim$ 拉普拉斯(0分布 τ)建模。可以通过调整先验尺度（τ）来控制增长率可发生变化的次数。 τ 值越大，趋势变化就越频繁。

我们假设趋势部分呈线性增长，并在季节性部分考虑星期效应，以此使用 Prophet。根据经验，我们考察了变点数量的先验尺度在 $\tau \in [0.10、 0.45]$范围内取不同值时的情况。该实现基于 R软件包 `prophet` 版本 1.0 [13]。

（*iii*）高斯过程（高斯过程）

高斯过程 [12]已用于流行病预测 [41–43]。高斯过程是一种随机过程，或一组随机变量，其中每个有限子集的联合分布均为多元正态（高斯）分布。高斯过程完全由均值函数和协方差函数（通常称为核函数）定义。均值函数可以任意指定，但在数据转换后（例如进行中心化）通常设为零。因此，关键组成部分是核函数，它刻画了输入数据对之间的相关性。先验信息反映在所选的核函数中，该函数用于描述随机过程的行为。一种常用的核

是平方指数（SE）核 k_{SE}，定义如下：

$$k_{\mathrm{SE}}(t_i,t_j)=\alpha^2\exp\left(-\frac{(t_i-t_j)^2}{2\rho^2}\right), \tag{4}$$

其中 α^2 是决定输出尺度的信号噪声参数，而 ρ 是长度尺度（或带宽）参数，用于决定输入对之间的相关性，即两个时间点(t_i，t_j)。较大的 α 允许偏离均值的幅度更大，而较大的 ρ 会使函数路径更加粗糙。由于可辨识性较弱，通常会固定这两个参数之一 ρ 和 α 中的一个 [12,44]。对于时间序列分析，可以使用周期核（PE）来建模季节性模式

$$k_{\mathrm{PE}}(t_i,t_j)=\alpha^2\exp\left(-\frac{2\sin\left(\pi\left|t_i-t_j\right|/p\right)^2}{2\rho^2}\right), \tag{5}$$

其中 p 是一个额外参数，用于表示周期性。

我们使用由三个核函数组合而成的零均值高斯过程对 $Y_{s,\ t}$ 进行建模

$$Y_{s,t}\sim\mathrm{GP}\left(0,k_{\mathrm{SE}}^{\mathrm{long}}+k_{\mathrm{SE}}^{\mathrm{short}}+k_{\mathrm{PE}}^{\mathrm{week}}+\sigma^2\delta_{ij}\right)\ , \tag{6}$$

其中 $k^{\mathrm{long}}{}_{\mathrm{SE}}$ 是具有较长长度尺度的 SE 核，$k_{\mathrm{SE}}^{\mathrm{short}}$是具有较短长度尺度的 SE 核，$k_{\mathrm{PE}}^{\mathrm{week}}$是用于表示星期效应的 PE 核，而 $\sigma^2\delta_{ij}$ 是一个带有克罗内克δ函数的正则化项，用于确保仅当 $i=j$时才添加 σ^2 高斯噪声。SE 核通过长程和短程相关结构 [44]分别对长期趋势和短期变化进行建模。

各核的先验设定如下。对于 $k^{\mathrm{long}}{}_{\mathrm{SE}}$，我们将长度尺度设为 $\rho=56$天（训练期的长度），并对 $\alpha\sim$学生t($\nu=$3,$\mu=0,\sigma=5$)使用弱信息先验。对于 $k_{\mathrm{PE}}^{\mathrm{week}}$，我们将长度尺度设为 $\rho=1$，周期性设为 $p=7$ 天，并使用$\alpha\sim$学生t($\nu=5,\mu=0,\sigma=2$)。在短程核中，我们根据经验考虑了不同的长度尺度 $\rho\in(1,7,14)$天，使高斯过程能够或快或慢地响应短期变化。对 $\alpha\sim$ 学生t($\nu=7,\mu=0,$ $\sigma=2$)给出了弱信息先验。我们使用概率编程语言 Stan 2.21.0 [45]版实现了高斯过程。

2.6 评估指标

我们在提前2周预测的时间窗口内评估概率预测。我们按周汇总每日预测值，以消除潜在的星期效应，而 EpiEstim 模型未对此进行考虑。针对提前1周和2周的预测值，分别计算以下评估指标。

预测性能主要依据连续排序概率分数（CRPS） [46]进行评估。它衡量预测值与观测值之间的差异，是一种严格适当的评分规则。其定义如下：

$$CRPS(F_f, x_t) = \int \left[F(y_t) - \mathbf{1}(y_t \geqslant x_t)\right]^2 \mathrm{d}y_t, \tag{7}$$

其中， F 是观测值 x 在时间 t的预测值 y 的累积分布函数。请注意，CRPS 是预测值与观测值之间绝对误差的推广。因此，CRPS 越低，表示预测越准确。

我们按照建议，对预测发病率 [47]的对数计算 CRPS。这相当于评估乘法预测误差，而非绝对预测误差。值得注意的是，取对数可以减轻模型在流行病呈指数增长期间所做预测中常见的极端异常值的影响。

我们为每个预测值计算 CRPS，并按模型、州和流行病阶段汇总得分（人工标注，见 S1 附录图??）。此外，我们进行模型两两比较 [48]，以衡量各模型相对于其他模型的预测性能。具体做法是先计算得分比，再通过对所有得分比取几何平均数来计算相对技巧。在两两比较中，得分比（以及相对技巧）越低，表示预测性能越好。

次要性能指标包括尖锐度、校准度和偏差。尖锐度以预测样本中各值与其中位数的绝对偏差的归一化中位数来衡量 [49]。校准度根据概率积分变换（PIT）值的密度 [46]以及 50% 和 95% 预测区间（PI）的经验覆盖率 [7]进行评估。偏差则通过后验抽样值高于观测值的平均比例来评估。

主要和次要性能指标使用 R软件包 `scoringRules`版本 1.0.2 [50] 和 `scoringutils` [48] 版本 1.0.1 计算。

流行病预测的一项重要特性是能够预判指数增长。连续排序概率得分无法专门衡量这一特性。因此，我们采用 McDonald 等人提出的另一项指标， [51] 用于检测所谓的“热点”。如果某个州在时间 t s 的发病率每周增幅超过 25%，则将其定义为热点，即

$$H_{s,t} = \begin{cases} 1, & \text{if } (I_{s,t} - I_{s,t-7})/(I_{s,t-7}) \geqslant 0.25, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

模型 m 预测某地区为热点的概率，等于预测中每周变化幅度超过 25% 的后验抽样值所占的比例。我们使用受试者工作特征曲线下面积（曲线下面积）评估性能。与 McDonald 等人 [51],类似，我们排除每周发病率低于每10万人 70 例新增病例的数据。这剔除了 35% 的观测值。在其余观测值中，25% 被归类为热点。

2.7 软件与模型估计

所有分析均使用 R 软件（版本 4.2.0） [52]完成。概率模型采用马尔可夫链蒙特卡罗（MCMC）抽样进行估计（EpiEstim 除外，其后验分布通过解析方法推导得出）。每个模型均使用 4 条马尔可夫链进行估计，每条链迭代 1,000 次，其中前 500 次迭代作为预热过程予以舍弃。最终得到 2,000 个后验样本。

2.8 数据和代码的可用性

数据来自公开可用的数据源（流行病学数据来自约翰斯·霍普金斯冠状病毒资源中心；人口数据来自世界银行）。预处理后的数据文件及可复现代码可从 https://github.com/nbanho/covid_predict 获取。

3 结果

具体的建模选择对各模型整体预测性能的影响微乎其微（S1 附录图 ??）。表 2展示了每个模型中整体连续排序概率得分最高的建模选择。后续对机制模型与统计时间序列模型进行模型间比较时，将采用这些选择对应的模型。

表2. 总体连续排序概率得分最高的建模选择。所有选择的结果见 S1 附录图。**??**。

模型	总体连续排序概率得分最高的建模选择
机制模型	
EpiEstim	在7天时间窗口内估计的时变再生数。
EpiNow2	使用高斯过程建模时变再生数。
统计时间序列模型	
SARIMA	SARIMA(1,0,1)(0,1,0)$_7$ 模型，包含一个自回归成分和一个移动平均成分。
Prophet	先验尺度 $\tau = 0.45$ 允许增长率存在较多变点。
GP	长度尺度为 $\rho = 7$天的短程核。

图 1展示了亚利桑那州各模型的提前1周和2周预测值，作为示例。所有模型的提前1周预测值校准良好，大多数时候都涵盖了观测到的每周发病率，但 SARIMA 和 Prophet 在 2021年1月 的峰值除外。提前2周预测值的覆盖率较差，尤其是大多数模型高估了 2021年1月 的峰值，也未能预见此后的下降。EpiNow2 的预测区间比 EpiEstim 更宽。同样，高斯过程的预测区间比 SARIMA 和 Prophet 更宽。

图2显示了各州次要性能指标（校准度、尖锐度和偏差）的结果。EpiEstim、SARIMA 和 Prophet 的概率积分变换分布呈 U 形（图 2a），表明预测分布过窄；而 EpiNow2 和 GP 的分布呈驼峰形，表明预测分布过宽。GP 和 EpiNow2 对 50%预测区间和 95%预测区间实现了良好的覆盖率，而 EpiEstim、SARIMA 和 Prophet 对两个预测区间的覆盖率均低于理论预期（图2b）。EpiNow2 的预测值尖锐度最低，尤其是提前2周预测，其离散度明显最高（图2c）。EpiEstim 往往高估真实发病率，而 GP 往往低估真实发病率（图 2d）。在偏差方面，其他模型未表现出类似行为。

CRPS评分见图3。所有模型的评分分布均呈右偏，表明存在相当数量的较大预测误差（图3a）。对于提前1周预测，各模型的评分相近。GP的均值和中位数 CRPS最低（表现最佳），Prophet最高。后者在95%分位数上的CRPS也最高。提前2周预测的CRPS评分分散程度更高，但各模型之间的差异仍然相似，只是更加明显。GP的评分最佳，且分布更集中于较低值

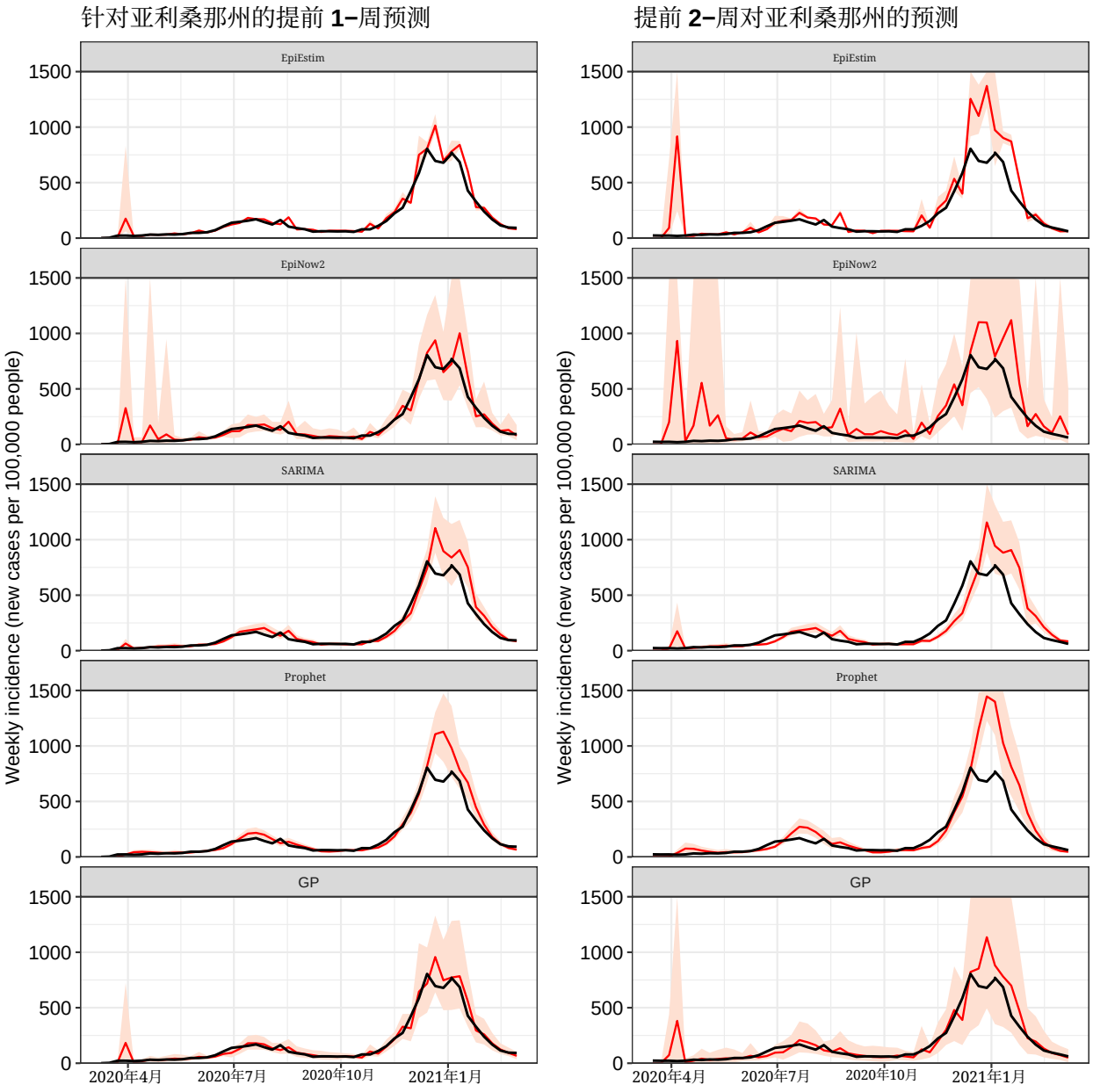


图1. 按模型划分的亚利桑那州概率预测。各模型对亚利桑那州的概率预测（后验均值以红线表示，95%预测区间以阴影区域表示）。（**a**）提前1周预测。（**b**）提前2周预测。观测发病率以黑线表示。各模型对美国各州的概率预测见 S1 附录**？？-？？**。

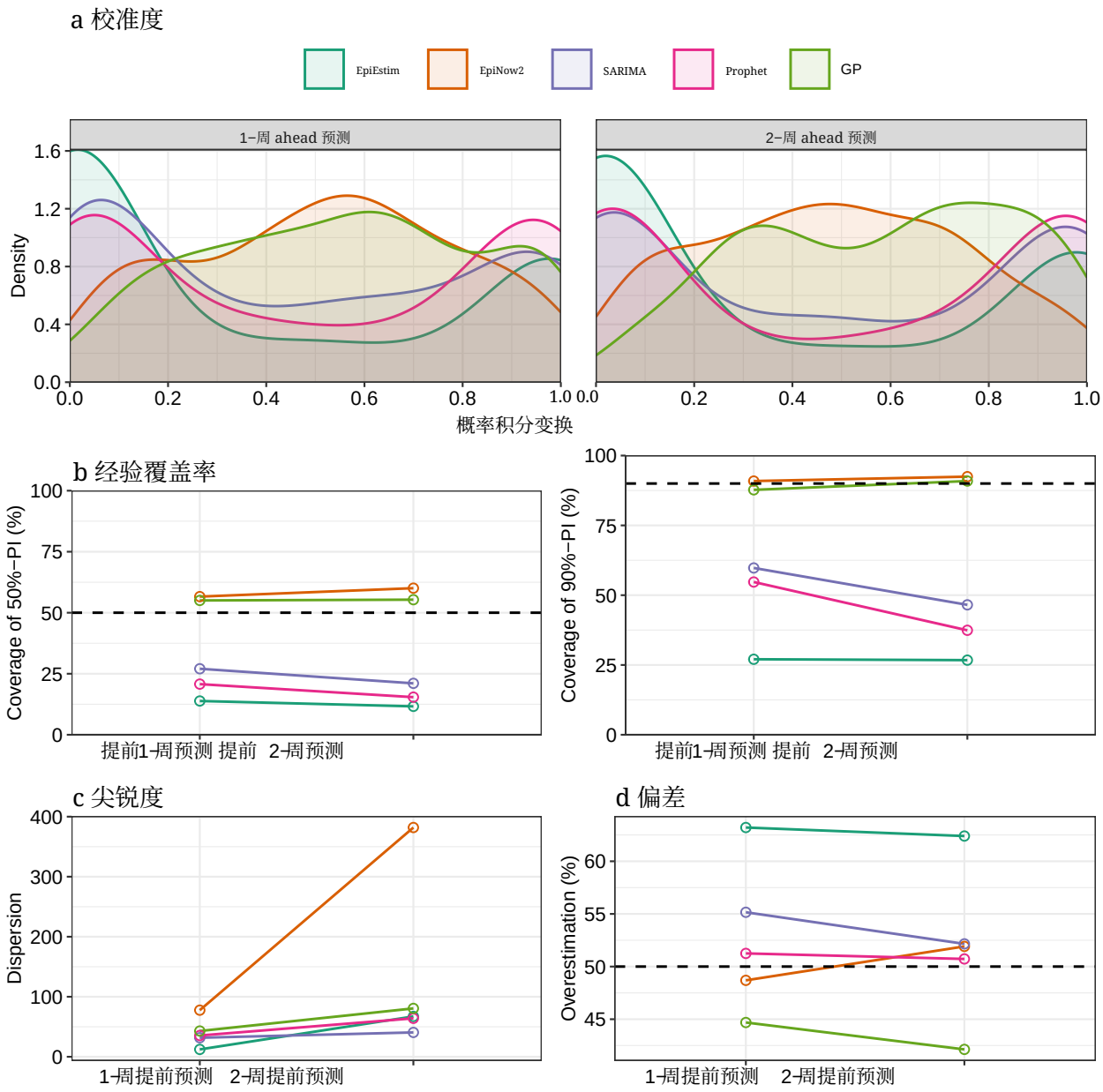


图 2. 次要性能指标的总体评估。(a) 通过概率积分变换的密度分布评估概率校准。(b-c) 50 %（左图）和 95 %（右图）预测区间 [PI]的覆盖率。(d) 通过离散度评估尖锐度，离散度以预测样本相对于其中位数的绝对偏差的中位数衡量。(e) 通过后验抽样值高于观测发病率的平均比例评估偏差。

CRPS。相比之下，Prophet的分布更分散，其90%和95%分位数对应的CRPS更高。SARIMA、EpiNow2和EpiEstim的CRPS评分相近，介于GP和Prophet之间。模型两两比较也证实了预测性能的差异。在提前1周和2周的预测中，GP的相对技巧最低（最佳）（图 3b）。与其他模型相比，GP的分数比最高（最佳）；不过，在提前2周预测中，它与SARIMA和EpiEstim之间的差异在统计学上并不显著（图 3c）。此外，在美国各州中的全部6个州，GP的CRPS评分均最低（S1 附录图 ??），并且在（亚）指数下降阶段表现尤为出色（S1 附录图 ??），因为它能更早预测发病率峰值及其随后下降。

CRPS用于评估整体预测性能。预测模型的另一项重要能力是预测热点，即每周发病率大幅上升的情况。对热点预测AUC的比较显示，总体而言，统计时间序列模型比机制模型更准确地预测热点（图 4）。在提前1周和2周的预测中，Prophet的AUC最高，其次是GP和SARIMA。

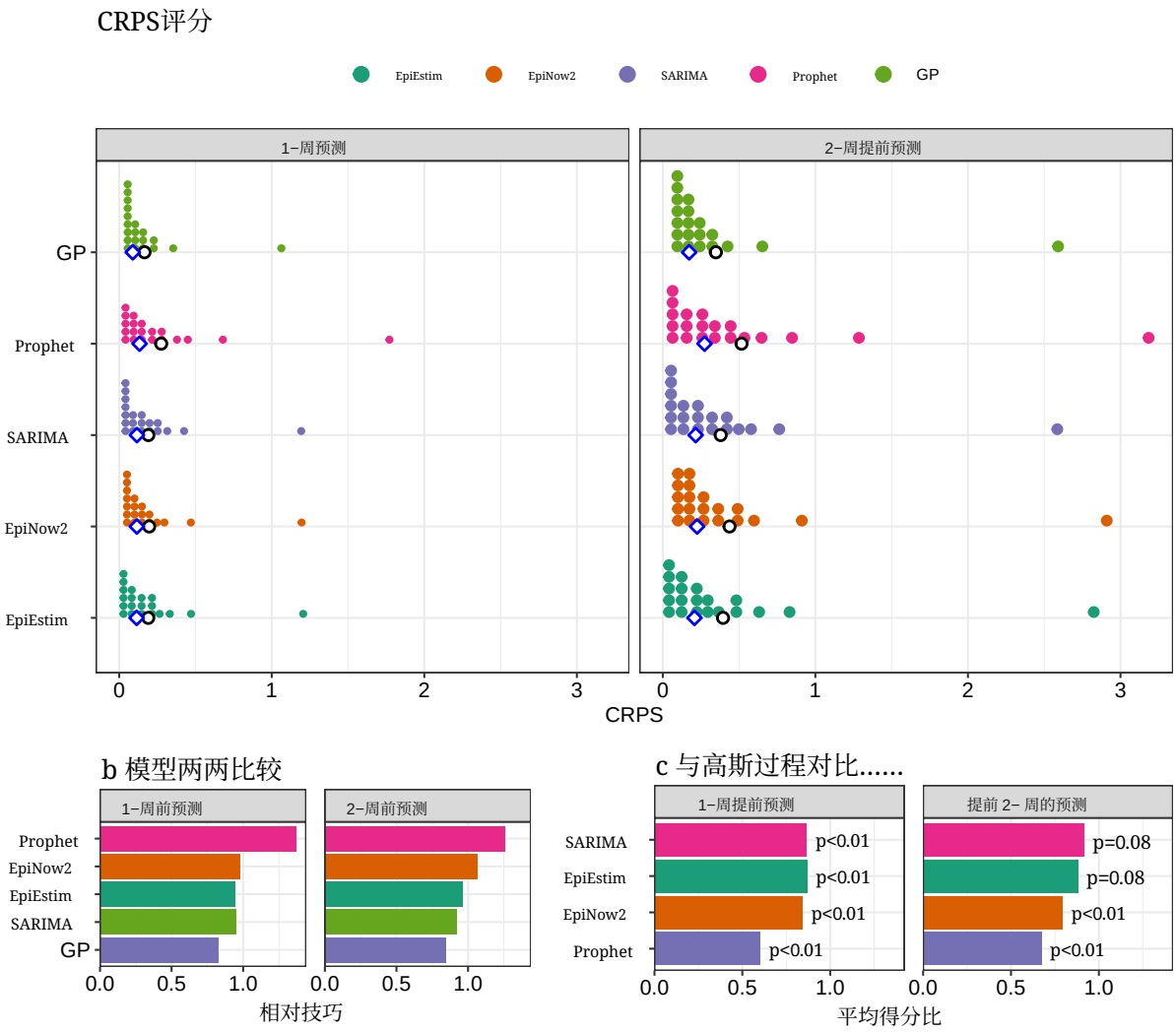
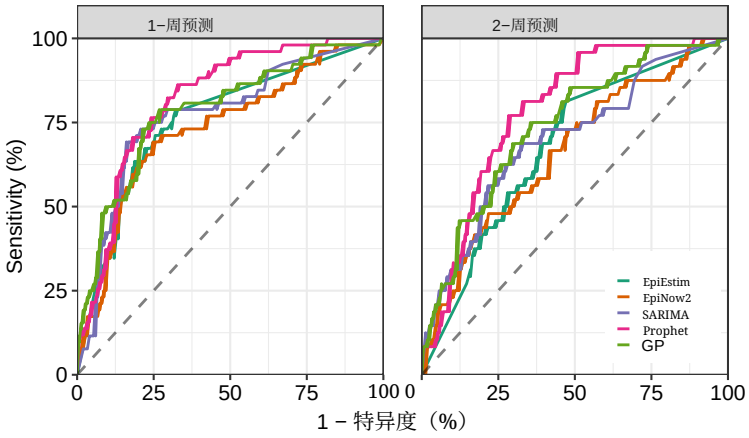


图 3. 预测性能评估。(a) 按模型划分的连续排序概率分数（CRPS）经验分位数分布。每个点代表经验分布的 20 个分位数之一。中位数 CRPS 以蓝色方块表示，均值 CRPS 以黑色圆点表示。CRPS 越低，预测性能越好。(b) 根据 CRPS 评分进行模型两两比较所得的相对技巧。数值越低，表示相对预测性能越好。(c) 高斯过程（GP）与其他模型相比的平均得分比。得分比越接近 1，表示预测性能与 GP 越相似。



模型	1-周曲线下面积	2-周曲线下面积
EpiEstim	76	68
EpiNow2	73	66
SARIMA	77	70
Prophet	82	78
GP	79	74

图4. 热点预测评估。用于预测热点的受试者工作特征曲线下面积（AUC）。热点定义为每周发病率增幅超过25%，且每10万人新增病例数超过10例。AUC越高，表明检测热点的能力越强。

4 讨论

4.1 总结

我们比较了大流行第一年期间美国六个大州中用于预测 COVID-19 发病率的机制模型和统计时间序列模型的短期预测值。统计时间序列模型的预测准确度至少不低于机制模型。高斯过程的整体预测性能最佳，但与 SARIMA 和 EpiEstim 等其他模型的差异不大，尤其是在提前2周预测方面。高斯过程和 EpiNow2 的预测值不如其他模型（EpiEstim、SARIMA 和 Prophet）的预测值尖锐，但其经验覆盖率更好。最后，统计时间序列模型对发病率大幅上升的预测比机制模型更准确。

4.2 研究发现的背景分析

我们的研究表明，统计时间序列模型的短期预测性能与机制模型相当，甚至更好。这与先前的研究结果一致：统计时间序列模型有多种已知应用，也能够对疾病传播提供相当准确的预测值 [35, 36, 41]。尽管如此，这两类模型都难以进行流行病预测。原因在于，预测值更适合被理解为完全基于结果历史时间序列的预测结果。因此，这些预测结果假设一切都没有变化

，且历史趋势会延续到未来。然而，影响疾病传播的因素可能并未被历史时间序列捕捉到，例如自愿行为改变或非药物干预措施 [25, 53, 54]。

一项相关研究比较了对 COVID-19 病例和死亡人数未来四周内的人工预测与基于模型的预测 [32]。研究通过网络应用程序征集人工预测，参与者可选择预测分布（如对数正态分布），然后查看并交互式调整数值预测。研究采用逐分位数均值汇总个体预测，并将群体预测与机制模型的概率预测进行比较。研究表明，群体预测比机制模型的预测更准确。具体而言，机制模型往往会高估疫情高峰期的发病率。类似地，我们在研究中发现，机制模型和统计时间序列模型也经常高估发病率峰值。与人工预测相比，基于模型的预测之所以可能高估峰值，或许是因为只有人工预测能够考虑到在必须做出预测之时恰好实施的非药物干预措施可能产生的影响。

通过使用集成或纳入辅助指标，可以进一步改进基于模型的疾病发病率预测。以往研究表明，多个模型的组合预测可能比各个单独模型的预测更准确 [7, 55, 56],，而额外的数据流（如在线调查或流动性数据）可以提升预测性能 [51, 57–59]。集成预测和辅助指标并非本研究的重点，但我们认可它们的潜力。

4.3 局限性

我们的研究存在局限性。首先，我们通过比较发病率对数来评估预测值的乘法误差。我们选择乘法误差，是因为它对异常值不太敏感，并能确保不同美国州和流行病阶段之间具有可比性 [32]。乘法误差的一个缺点是，无论发病率的绝对水平如何，其数值都相同。这可能会使在发病率较低阶段出现较大误差的模型处于不利地位，而这些阶段并不那么令人担忧。其次，我们使用连续排序概率得分（CRPS）作为主要性能指标来评估预测值。连续排序概率得分是一种适当评分规则，用于在多个领域评估数值预测 [46]。然而，政策制定者可能

会使用不同的方法评估预测值，例如采用一种效用函数，对低估疾病发病率的惩罚大于对高估的惩罚。第三，尽管我们比较了通常用于预测发病率的模型，但这些模型仍只代表所有可能模型中的一部分。因此，我们承认，对本文所考虑模型进行扩展，或采用本研究未考虑的其他类型模型，可能会取得更优的性能，例如分隔室模型 [60], 梯度提升树 [36], 或深度学习 [59]。第四，预测值是在回顾性情况下生成并分析的，这可能会影响对其进行的评估。相比之下，实时生成的预测值通常基于不完整的发病率数据 [51] ，或者预测者可能会在提交前进行人工调整 [7]。第五，我们根据报告的 COVID-19 发病率评估了预测值。因此，我们承认，COVID-19 的实际病例数可能高于报告病例数 [61], ，也就是说，我们的分析未处理识别偏倚。第六，我们在一年时间内，根据经过仔细标注的流行病阶段，对美国六个大州的 COVID-19 发病率进行了预测评估。未来的研究可以扩展我们的分析，例如考虑不同的地理区域、其他流行病学结局（例如死亡人数 [7, 55] 或住院人数 [37]），甚至考虑其他病原体引起的流行病（例如流感 [1, 3]）。

4.4 总结与展望

机制模型纳入了有关疾病动态的知识。例如，在模拟疫苗接种率的影响或新病毒变异株的出现时，这类领域知识十分重要。然而，它对疾病发病率短期预测的附加价值尚不明确。我们对机制模型和统计时间序列模型所做的疾病短期概率预测进行了实证比较。结果发现，统计时间序列模型的总体表现与机制模型相近，这表明无需对疾病动态进行建模，也能对疾病发病率作出短期预测。未来研究可评估能否通过其他方式进一步改进预测值，例如采用集成模型或纳入辅助指标。

参考文献

1. Biggerstaff M, Johansson M, Alper D, Brooks LC, Chakraborty P, Farrow DC, 等。美国流感季节预测协作项目第二年的成果。Epidemics. 2018;24:26–33. doi:10.1016/j.epidem.2018.02.003.

2. Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, 等。估算埃博拉疫情中病例的未来数量——利比里亚和塞拉利昂，2014–2015年。《发病率和死亡率周报》。2014;63(3)。获取地址： https://stacks.cdc.gov/view/cdc/24901。

3. Bogoch II, Brady OJ, Kraemer MUG, German M, Creatore MI, Kulkarni MA, 等。预测寨卡病毒从巴西向国际传播。《柳叶刀》。2016;387(10016):335–336。doi:10.1016/S0140-6736(16)00080-5。

4. Nsoesie E, Mararthe M, Brownstein J。季节性流感疫情峰值预测。《PLoS 当前》。2013年;5。doi:10.1371/currents.outbreaks.bb1e879a23137022ea79a8c508b030bc。

5. Bhatia S, Parag KV, Wardle J, Imai N, Van Elsland SL, Lassmann B 等。全球 COVID-19 传播趋势短期至中期预测：回顾性评估。medRxiv[预印本]； 2021 [引自 2022年3月 3]。可从以下地址获取： https://www.medrxiv.org/content/10.1101/2021.07.19.21260746.abstract。doi:10.1101/2021.07.19.21260746。

6. Massonnaud C, Roux J, Crépey P. COVID-19：法国短期住院需求预测。medRxiv [预印本]； 2020 [引自 2022年3月 3]。可从以下地址获取： https://www.medrxiv.org/content/10.1101/2020.03.16.20036939v1。doi:10.1101/2020.03.16.20036939。

7. Bracher J、Wolffram D、Deuschel J、Görgen K、Ketterer JL、Ullrich A 等。第二波疫情期间德国和波兰 COVID-19 的预注册短期预测研究。《自然·通讯》。2021年；12：5173。doi:10.1038/s41467-021-25207-0。

8. Cori A, Ferguson NM, Fraser C, Cauchemez S. 一种用于估计流行病期间时变再生数的新框架和软件。美国流行病学杂志。2013年;178(9):1505–1512。doi:10.1093/aje/kwt133。

9. Abbott S、Hellewell J、Thompson RN、Sherratt K、Gibbs HP、Bosse NI 等。使用国家和次国家级病例数估算 SARS-CoV-2 的时变再生数。Welcome Open Research。2020年；(5)：112。doi:10.12688/wellcomeopenres.16006.2。

10. Hyndman RJ, Athanasopoulos G. 第8章：ARIMA模型。载于：《预测：原理与实践》。OTexts；2018年。第221–274页。

11. Taylor SJ, Letham B. 大规模预测。Am Stat. 2018年；72(1)：37–45。doi:10.1080/00031305.2017.1380080。

12. Rasmussen CE.《机器学习中的高斯过程》。麻省理工学院出版社；2006年。

13. Taylor S, Letham B. Prophet：自动预测程序；2021年。R软件包版本1.0。可从以下网址获取： https://cran.r-project.org/web/packages/prophet。

14. Cori A, Kamvar Z, Stockwin J, Jombart T, Dahlqwist E, FitzJohn R, 等。EpiEstim：一种用于估算流行病期间随时间变化的瞬时再生数的工具；2021年。R软件包版本2.2-4。可从以下网址获取： https://cran.r-project.org/web/packages/EpiEstim。

15. Abbott S、Hellewell J、Sherrat K、Gostic KM、Hickson J、Badr HS 等。EpiNow2：实时估算病例数和时变流行病学参数；2021年。R软件包版本1.3.2。可查阅：https://cran.r-project.org/web/packages/EpiNow2。

16. Champredon D、Dushoff J、Earn DJD。服从厄朗分布的SEIR流行病模型与更新方程的等价性。《应用数学学会应用数学杂志》。2018;78(6):3258–3278。

17. Jewell NP、Lewnard JA、Jewell BL。COVID-19大流行的预测性数学模型：预测结果的基本原理与价值。《美国医学会杂志》。2020年;323(19):1893–1894。doi:10.1001/jama.2020.6585。

18. Krymova E、Béjar B、Thanou D、Sun T、Manetti E、Lee G 等。全球COVID-19病例和死亡人数的趋势估计与短期预测。《美国国家科学院院刊》。2022;119(32):e2112656119。doi:10.1073/pnas.2112656119。

19. Arnold T、Bien J、Brooks L、Colquhoun S、Farrow D、Grabman J 等。Covidcast：Delphi 的 COVIDcast流行病学数据API 客户端；2021年。R软件包版本 0.4.2。可从以下网址获取：https://cran.r-project.org/web/packages/covidcast。

20. Dong E、Du H、Gardner L。用于实时追踪 COVID-19 的交互式网络仪表板。《柳叶刀》传染病。2020年；20(5)：533–534。doi:10.1016/S1473-3099(20)30120-1。

21. Riccardo F, Ajelli M, Andrianou XD, Bella A, Manso MD, Fabiani M，等。意大利COVID-19流行病暴发1个月后的病例流行病学特征及再生数估计，2020年1月28日至3月31日。Euro Surveill. 2020;25(49):2000790. doi:10.2807/1560-7917.ES.2020.25.49.2000790.

22. Al Wahaibi A, Al Manji A, Al Maani A, Al Rawahi B, Al Harthy K, Alyaquobi F，等。非药物干预措施后的COVID-19流行病监测：在移民人口众多的国家使用时变再生数。Int J Infect Dis. 2020;99:466– 472. doi:10.1016/j.ijid.2020.08.039.

23. Huisman JS、Scire J、Angst DC、Li J、Neher RA、Maathuis MH 等。SARS-CoV-2 有效再生数的估算与全球监测。eLife。2022;11:e71345。doi:10.7554/eLife.71345。

24. Fraser C。新发流行病中个体和家庭再生数的估算。《公共科学图书馆·综合》。2007;2(8):e758。doi:10.1371/journal.pone.0000758。

25. Flaxman S、Mishra S、Gandy A、Unwin HJT、Mellan TA、Coupland H 等。估算非药物干预措施对欧洲 COVID-19 疫情的影响。《自然》。2020年;584(7820):257– 261。doi:10.1038/s41586-020-2405-7。

26. Ganyani T、Kremer C、Chen D、Torneri A、Faes C、Wallinga J 等。根据症状发作数据估算冠状病毒病（COVID-19）的代间隔，2020年3月。《欧洲疾病监测》。2020;25(17):2000257。doi:10.2807/1560-7917.ES.2020.25.17.2000257。

27. Bosse NI、Abbott S、Bracher J、Hain H、Quilty BJ、Jit M 等。比较德国和波兰COVID-19 的人工预测值与模型预测值。medRxiv [预印本]； 2021 [引文日期：2022年3月3]。可查阅： https://www.medrxiv.org/content/10.1101/2021.12.01.21266598v1。doi:10.1101/2021.12.01.21266598。

28. Jombart T、Nouvellet P、Bhatia S、Kamvar ZN、Taylor T、Ghozzi S。Projections：预测未来病例发病率；2021年。R软件包版本 0.5.4。获取地址： https://cloud.r-project.org/web/packages/projections。

29. Davies NG、Abbott S、Barnard RC、Jarvis CI、Kucharski AJ、Munday JD 等。对英格兰 SARS-CoV-2 谱系 B.1.1.7 传播能力及影响的估计。《科学》。2021年；372(6538)：eabg3055。doi:10.1126/science.abg3055。

30. Lauer SA、Grantz KH、Bi Q、Jones FK、Zheng Q、Meredith HR，等。基于公开报告的确诊病例估算2019冠状病毒病（COVID-19）的潜伏期及其应用。Ann Intern Med。2020年；172(9)：577–582。doi:10.7326/M20-0504。

31. Cereda D、Tirani M、Rovida F、Demicheli V、Ajelli M、Poletti P，等。意大利伦巴第大区COVID-19疫情暴发的早期阶段。arXiv [预印本]； 2020 [引用于2022年3月 3]。可从以下地址获取： https://arxiv.org/abs/2003.09320。doi:10.48550/ARXIV.2003.09320。

32. Bosse NI、Abbott S、Bracher J、Hain H、Quilty BJ、Jit M，等。比较德国和波兰COVID-19的人类预测值与模型预测值。PLoS Comput Biol。2022；18(9)：e1010405。doi:10.1371/journal.pcbi.1010405。

33. Alabdulrazzaq H、Alenezi MN、Rawajfih Y、Alghannam BA、Al-Hassan AA、Al-Anzi FS。基于自回归积分滑动平均模型预测 COVID-19 传播的准确性。Results Phys. 2021年；27：104509。doi:10.1016/j.rinp.2021.104509。

34. Kufel T。基于自回归积分滑动平均模型预测若干欧洲国家确诊 COVID-19 病例动态。Equilibrium. 2020年；15(2)：181–204。doi:10.24136/eq.2020.009。

35. Wang Y、Yan Z、Wang D、Yang M、Li Z、Gong X 等。COVID-19 每日新增病例和累计病例的预测与分析：时间序列预测和机器学习模型。BMC Infect Dis. 2022；22(1)：495。doi:10.1186/s12879-022-07472-6。

36. Rahman MS、Chowdhury AH、Amrin M。用于预测孟加拉国 COVID-19 发病率的自回归积分滑动平均模型与 XGBoost 预测模型的准确性比较。《PLoS 全球公共卫生》。2022;2(5):e0000495。doi:10.1371/journal.pgph.0000495。

37. Perone G。自回归积分滑动平均模型、指数平滑法、神经网络自回归模型、TBATS模型和混合模型在预测意大利 COVID-19 第二波疫情住院人数方面的比较。《欧洲卫生经济学杂志》。2022;23(6):917–940。doi:10.1007/s10198-021-01347-4。

38. Matamoros AA、Torres CC、Dala A、Hyndman R、O'Hara-Wild M。Bayesforecast：使用 Stan 进行贝叶斯时间序列建模；2021年。R软件包版本 1.0.1。可从以下网址获取： https:

//cran.r-project.org/web/packages/bayesforecast。

39. Battineni G, Chintalapudi N, Amenta F. 使用 Fb-Prophet 机器学习模型预测四个疫情严重国家（美国、巴西、印度和俄罗斯）的 COVID-19 流行病规模。Appl Comput Inform. 2020年; doi:10.1108/ACI-09-2020-0059.

40. Aditya Satrio CB, Darmawan W, Nadia BU, Hanafiah N. 使用 ARIMA 模型和 PROPHET 对印度尼西亚冠状病毒病进行时间序列分析与预测。Procedia Comput Sci. 2021年;179:524–532. doi:10.1016/j.procs.2021.01.036.

41. Johnson LR, Gramacy RB, Cohen J, Mordecai E, Murdock C, Rohr J, et al. 使用异方差高斯过程对疾病发病率进行现象学预测：一项登革热案例研究。Ann Appl Stat. 2018;12(1):27–66. doi:10.1214/17-AOAS1090.

42. Arias Velásquez RM，Mejía Lara JV。使用降维高斯过程回归预测并评估美国 COVID-19 的传播。Chaos Solitons Fractals。2020年；136：109924。doi:10.1016/j.chaos.2020.109924。

43. Qian Z，Alaa AM，van der Schaar M。载于：《神经信息处理系统进展》，第33卷；2020年。第10729–10740页。可从以下网址获取：https://proceedings.neurips.cc/paper/2020/hash/79a3308b13cd31f096d8a4a34f96b66b-Abstract.html。

44. Hawryluk I，Hoeltgebaum H，Mishra S，Miscouridou X，Schnekenberg RP，Whittaker C，等。高斯过程即时预测：应用于 COVID-19 死亡率报告。《第三十七届人工智能不确定性会议论文集》。2021年；161：1258–1268。可从以下网址获取：https://proceedings.mlr.press/v161/hawryluk21a.html。

45. Carpenter B、Gelman A、Hoffman MD、Lee D、Goodrich B、Betancourt M 等。Stan：一种概率编程语言。《J Stat Softwb》。2017；76：1–32。doi:10.18637/jss.v076.i01。

46. Gneiting T、Balabdaoui F、Raftery AE。概率预测、校准度与尖锐度。《J R Stat Soc Series B》。2007；69(2)：243–268。doi:10.1111/j.1467-9868.2007.00587.x。

47. Bosse NI、Abbott S、Cori A、van Leeuwen E、Bracher J、Funk S。在流行病学背景下转换预测结果以评估预测性能。medRxiv [预印本]；2023 [引用于 2022 年 3 月 3]。可从以下网址获取：https://doi.org/10.1101/2023.01.23.23284722。doi:10.1101/2023.01.23.23284722。

48. Bosse N、Abbott S、Gruson H、Bracher J、Funk S。Scoringutils：评分与评估预测的实用工具；2022。R软件包版本 1.0.1。可从以下网址获取： https://cran.r-project.org/web/packages/scoringutils。

49. Funk S、Camacho A、Kucharski AJ、Lowe R、Eggo RM、Edmunds WJ。实时流行病预测的表现评估：塞拉利昂西部地区埃博拉疫情的案例研究，2014—2015年。《PLoS Computational Biology》。2019；15(2)：e1006785。doi:10.1371/journal.pcbi.1006785。

50. Jordan A、Krueger F、Lerch S。ScoringRules：参数化与模拟分布预测的评分规则；2022。R软件包版本1.0.2。可通过以下网址获取： https://cran.r-project.org/web/packages/scoringRules。

51. McDonald DJ、Bien J、Green A、Hu AJ、DeFries N、Hyun S等。辅助指标能否改进COVID-19预测和热点预测？《美国国家科学院院刊》。2021年；118(51)：e2111453118。doi:10.1073/pnas.2111453118。

52. R 核心团队。R：一种用于统计计算的语言和环境。R 统计计算基金会；2022。版本 4.2.1。可从以下网址获取： https://www.R-project.org/。

53. Brauner JM、Mindermann S、Sharma M、Johnston D、Salvatier J、Gavenčiak T 等。推断政府干预措施对抗 COVID-19 的有效性。《科学》。2021年；371(6531)：eabd9338。doi:10.1126/science.abd9338。

54. Banholzer N、Weenen Ev、Lison A、Cenedese A、Seeliger A、Kratzwald B 等。估算非药物干预措施对第一波疫情期间 COVID-19 新增感染人数的影响。《公共科学图书馆·综合》。2021年；16(6)：e0252827。doi:10.1371/journal.pone.0252827。

55. Cramer EY、Ray EL、Lopez VK、Bracher J、Brennen A、Castro Rivadeneira AJ 等。美国 COVID-19 死亡率个体和集成概率预测的评估。《美国国家科学院院刊》。2022；119(15)：e2113561119。doi:10.1073/pnas.2113561119。

56. Sherratt K、Gruson H、Grah R、Johnson H、Niehus R、Prasse B，等。欧洲国家COVID-19多模型集合预测的预测性能。medRxiv [预印本]；2022 [引用于2022年3月 3]。可从以下网址获取： https://www.medrxiv.org/content/10.1101/2022.06.16.22276024v1。doi:10.1101/2022.06.16.22276024。

57. Reinhart A、Brooks L、Jahja M、Rumack A、Tang J、Agrawal S，等。COVID-19实时指标开放存储库。《美国国家科学院院刊》。2021;118(51):e2111452118。doi:10.1073/pnas.2111452118。

58. Schwabe A、Persson J、Feuerriegel S。基于大规模流动性数据预测 COVID-19 的传播。载于：《第27届 ACM SIGKDD 知识发现与数据挖掘大会论文集》；2021年。第3531–3539页。可从以下网址获取： https://doi.org/10.1145/3447548.3467157。doi:10.1145/3447548.3467157。

59. Rodríguez A、Tabassum A、Cui J、Xie J、Ho J、Agarwal P 等。DeepCOVID：用于可解释实时 COVID-19 预测的运营型深度学习驱动框架。载于：《AAAI人工智能大会论文集》。第35卷；2021年。第15393–15400页。可从以下网址获取：https://ojs.aaai.org/index.php/AAAI/article/view/17808。doi:10.1609/aaai.v35i17.17808。

60. Reiner RC, Barber RM, Collins JK, Zheng P, Adolph C, Albright J 等. 美国 COVID-19 情景建模。Nat Med. 2021年;27(1):94–105. doi:10.1038/s41591-020-1132-9.

61. Noh J, Danuser G. 美国各州和世界各国 COVID-19感染者比例的估计。《公共科学图书馆·综合》。2021年；16(2)：e0246772。doi:10.1371/journal.pone.0246772。

致谢

感谢 Anne Cori 在 EpiEstim 实施方面给予的支持。感谢 Nikos Bosse 就如何评估概率预测提出的建议。

伦理审批

本研究无需获得伦理审批。

利益冲突

SF 报告称，除本研究提交的工作外，其还获得了瑞士国家科学基金会的资助。所有其他作者均声明不存在利益冲突。

资金支持

NB 和 SF 感谢瑞士国家科学基金会（SNSF）通过 Eccellenza 项目资助该研究，资助项目编号为 186932，项目主题为“数据驱动的健康管理”。TM、SM、HJTU 和 SB 感谢 MRC 全球传染病分析中心的资助（参考编号 MR/R015600/1）。该中心由英国医学研究理事会（MRC）和英国外交、联邦及发展事务部（FCDO）根据 MRC/FCDO 协议共同资助，也是由欧盟支持的 EDCTP2 计划的一部分；此外，中心还感谢 Community Jameel 的资助。资助机构未参与本研究的设计、实施、数据处理、分析、审查、报告或研究结果解读。SB 获得国家健康与护理研究所（NIHR）健康保护研究小组（建模与卫生经济学）的资助。该研究小组由英国健康安全局、伦敦帝国理工学院和 LSHTM 合作设立（资助代码 NIHR200908）。免责声明：“本文所述观点仅代表作者本人，并不一定反映 NIHR、英国健康安全局或卫生与社会保健部的观点。”SB 感谢 Novo Nordisk 基金会通过 Novo Nordisk 青年研究者奖（NNF20OC0059309）提供的资金支持，该奖项也支持 SM。SB 感谢丹麦国家研究基金会通过讲席资助提供的资金支持。SB 感谢 Eric and Wendy Schmidt 基金会战略创新基金通过 Schmidt 多领域学者奖（G-22-63345）提供的资金支持。

作者贡献

NB、TM、SM 和 SB 共同构思了本研究。NB 收集并分析了数据。NB、TM、SM 和 SB 共同参与了研究方法的制定。所有作者均参与了数据和结果的解读与可视化。NB 撰写了初稿。所有作者均参与了撰写、审阅与编辑。

支持信息

S1 附录。**补充材料**。图 S1。疫情轨迹的标记。图 S2。不同建模选择下的预测性能。图 S3。按模型划分的亚利桑那州概率预测。图 S4。按模型划分的伊利诺伊州概率预测。图 S5。按模型划分的马里兰州概率预测。图 S6。按模型划分的新泽西州概率预测。图 S7。按模型划分的纽约州概率预测。图 S8。按美国州评估预测性能。图 S9。按流行病阶段评估预测性能。

S1 附录

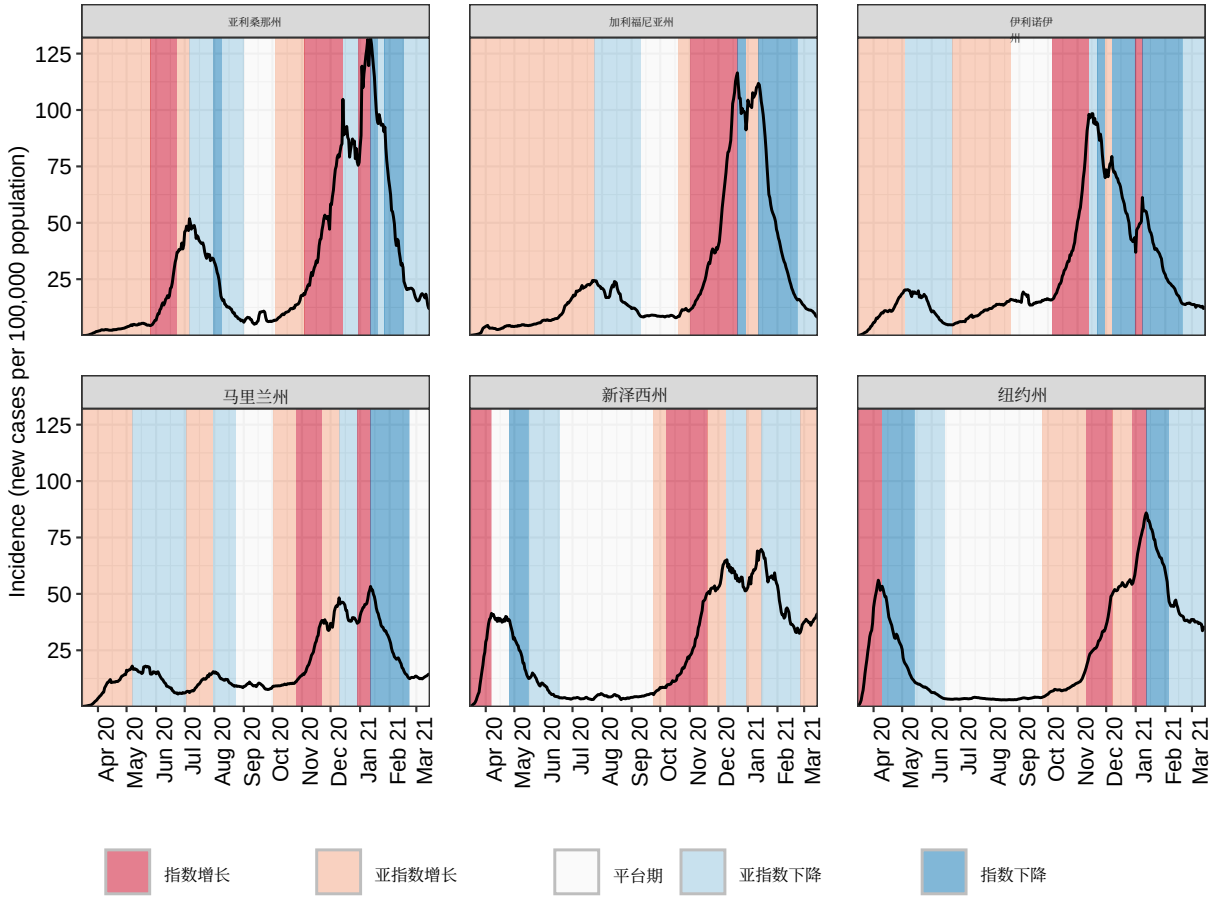


图 S1. 疫情轨迹的标注。美国各州的疫情轨迹根据五个阶段进行了人工标注：指数增长、次指数增长、平台期、亚指数下降、指数下降。

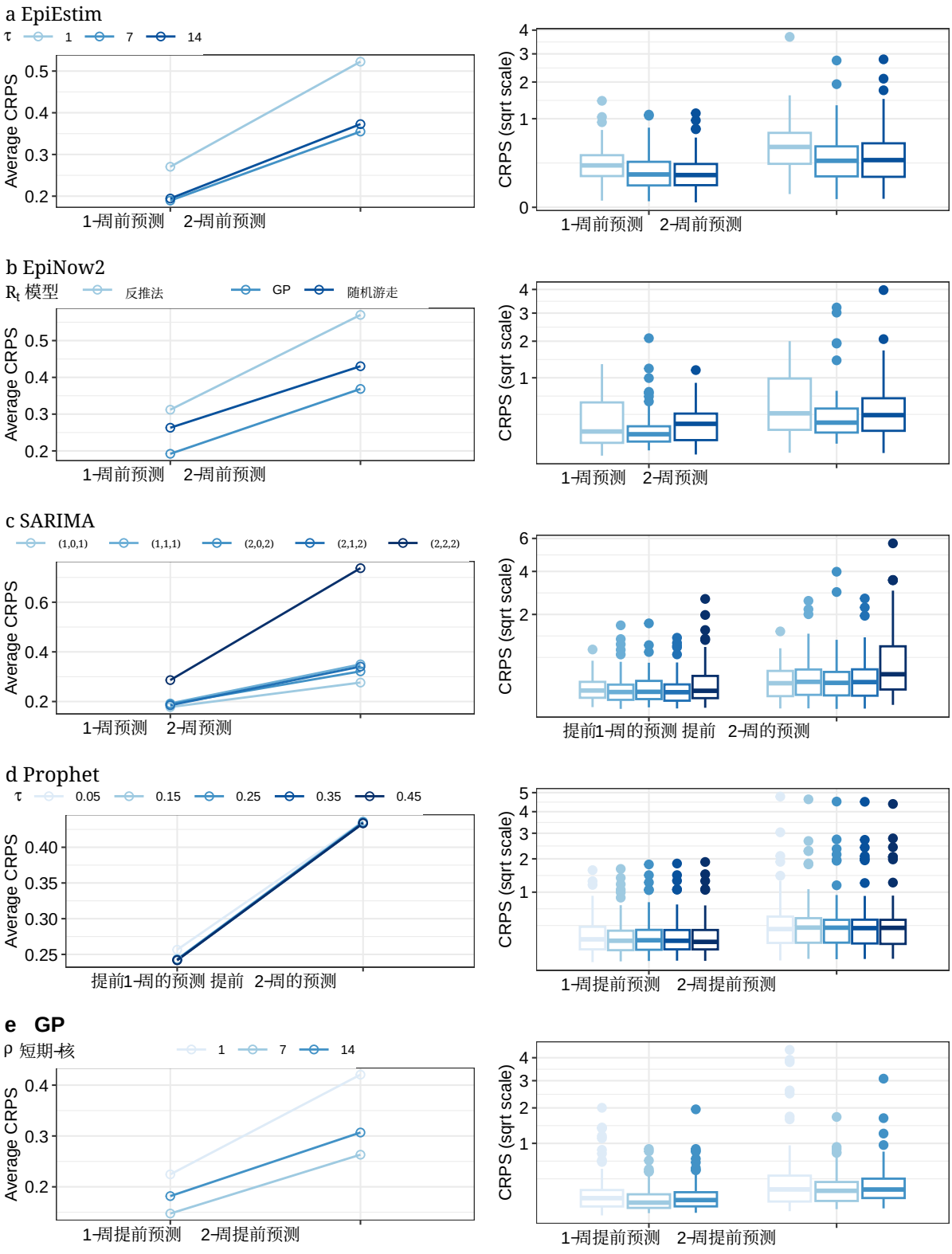


图 S2. 不同建模选择的预测性能。按模型设定划分的提前1周和2周预测的连续排序概率得分（左图为平均得分，右图为得分箱线图）。(a) EpiEstim。(b) EpiNow2。(c) SARIMA。(d) Prophet。(e)高斯过程。

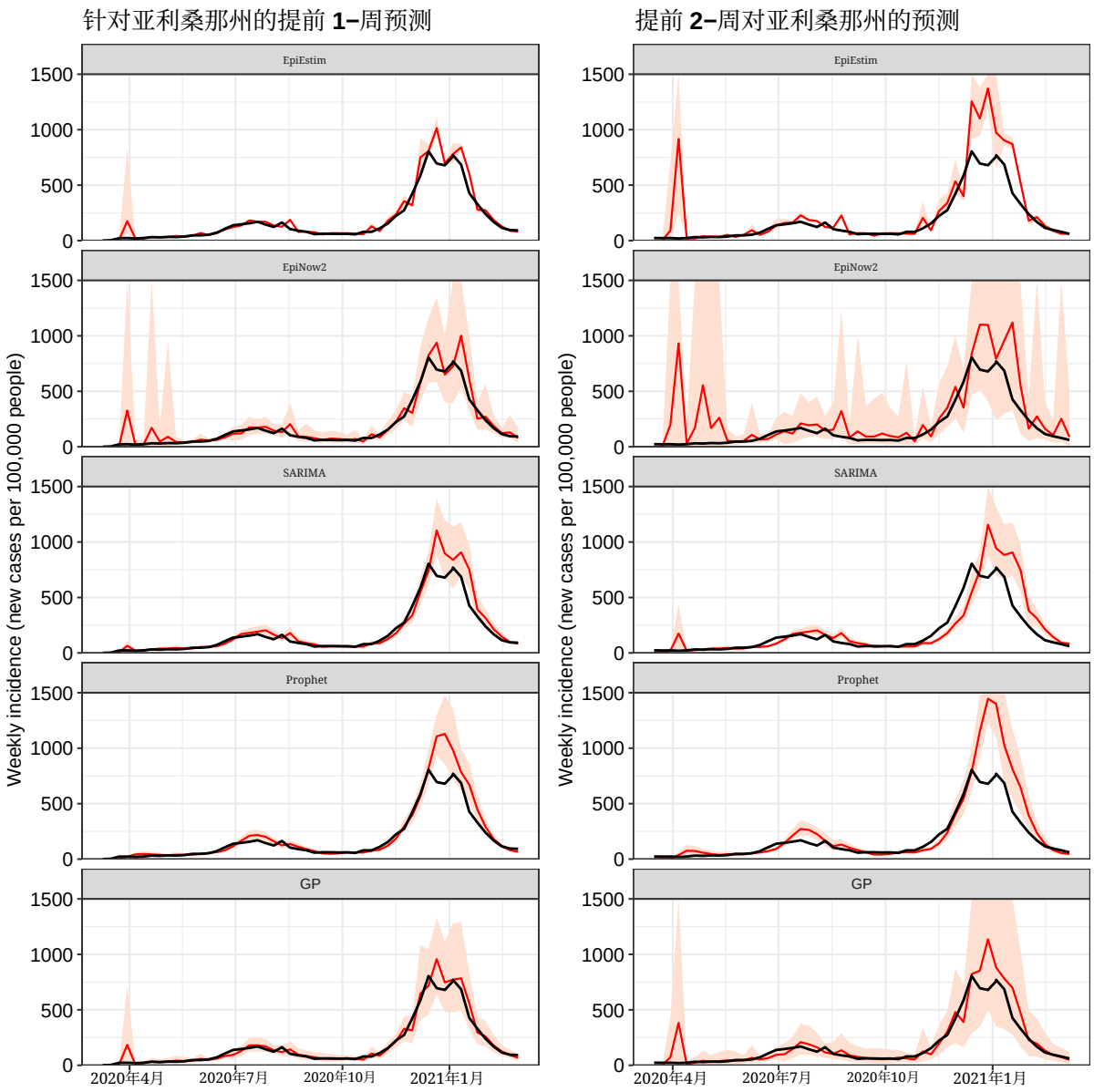


图 S3. 按模型划分的亚利桑那州概率预测。各模型对亚利桑那州的概率预测（后验均值以红线表示，95%预测区间以阴影区域表示）。(a) 提前1周预测。(b) 提前2周预测。观测发病率以黑线表示。

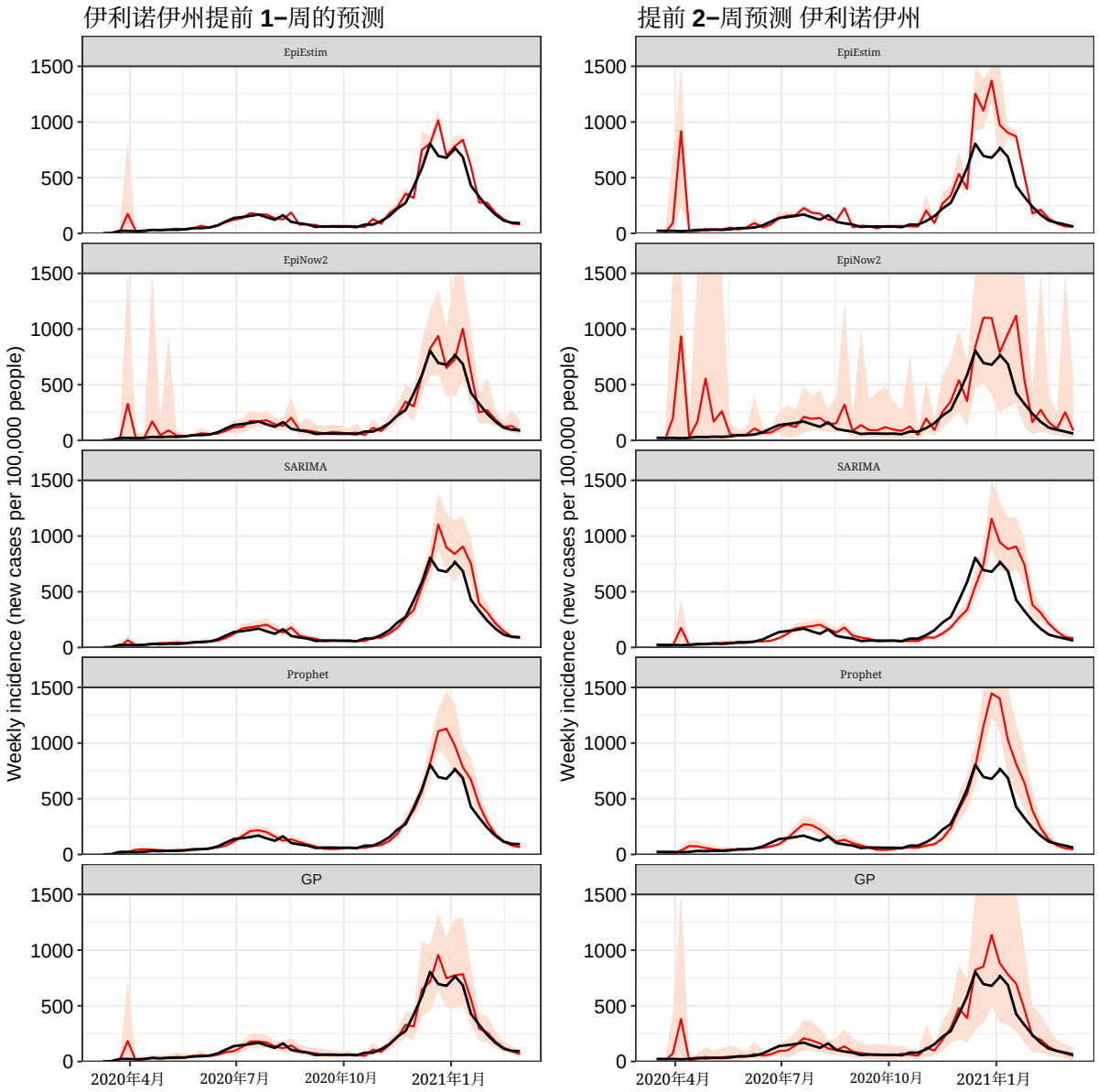


图 S4. 按模型划分的伊利诺伊州概率预测。各模型对伊利诺伊州的概率预测（后验均值以红线表示，95%预测区间以阴影区域表示）。(a) 提前1周预测。(b) 提前2周预测。观测发病率以黑线表示。

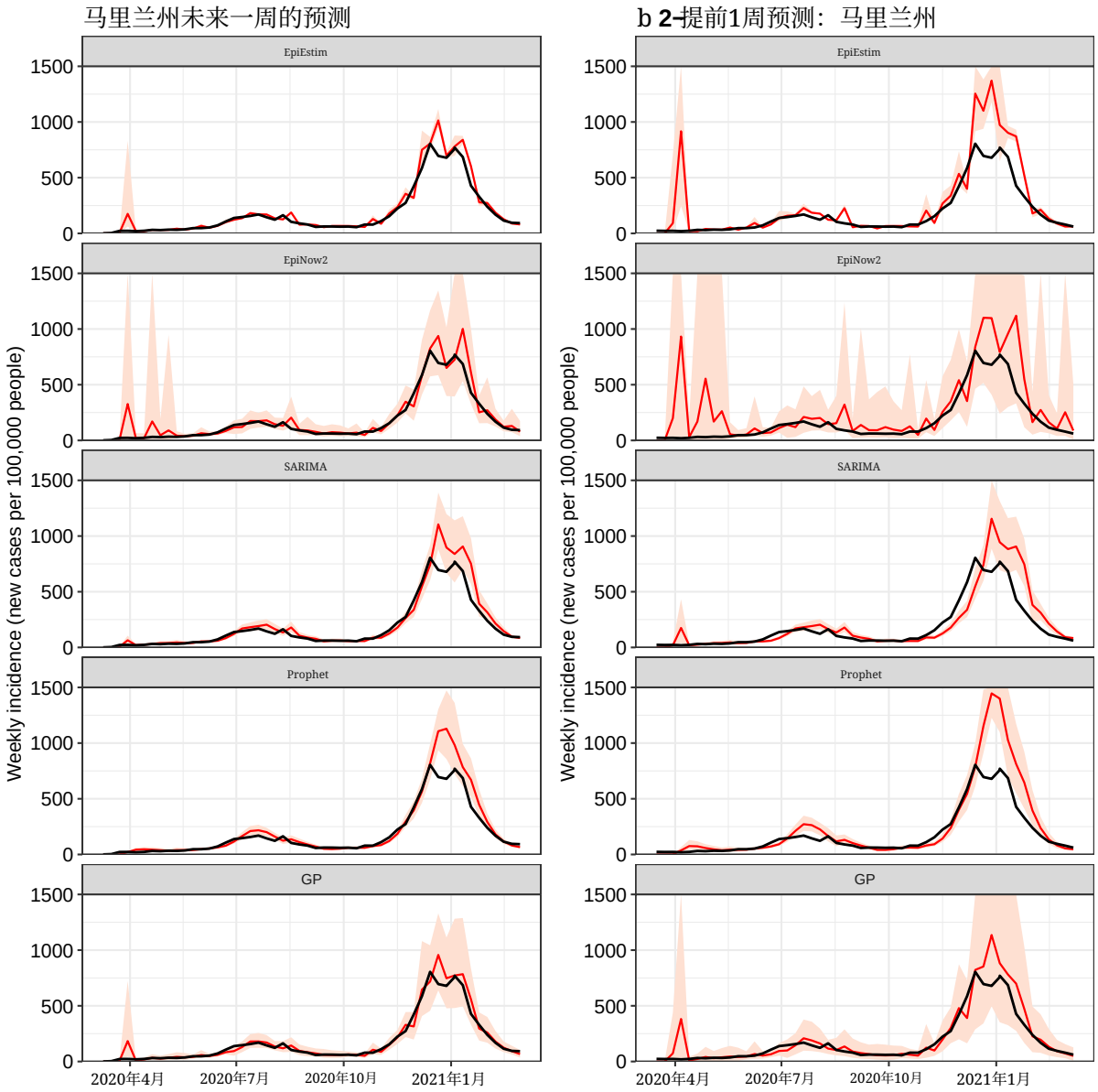


图 S5. 马里兰州各模型的概率预测。马里兰州各模型的概率预测（后验均值以红线表示，95%预测区间以阴影区域表示）。(a) 提前1周预测。(b) 提前2周预测。黑线表示观测发病率。

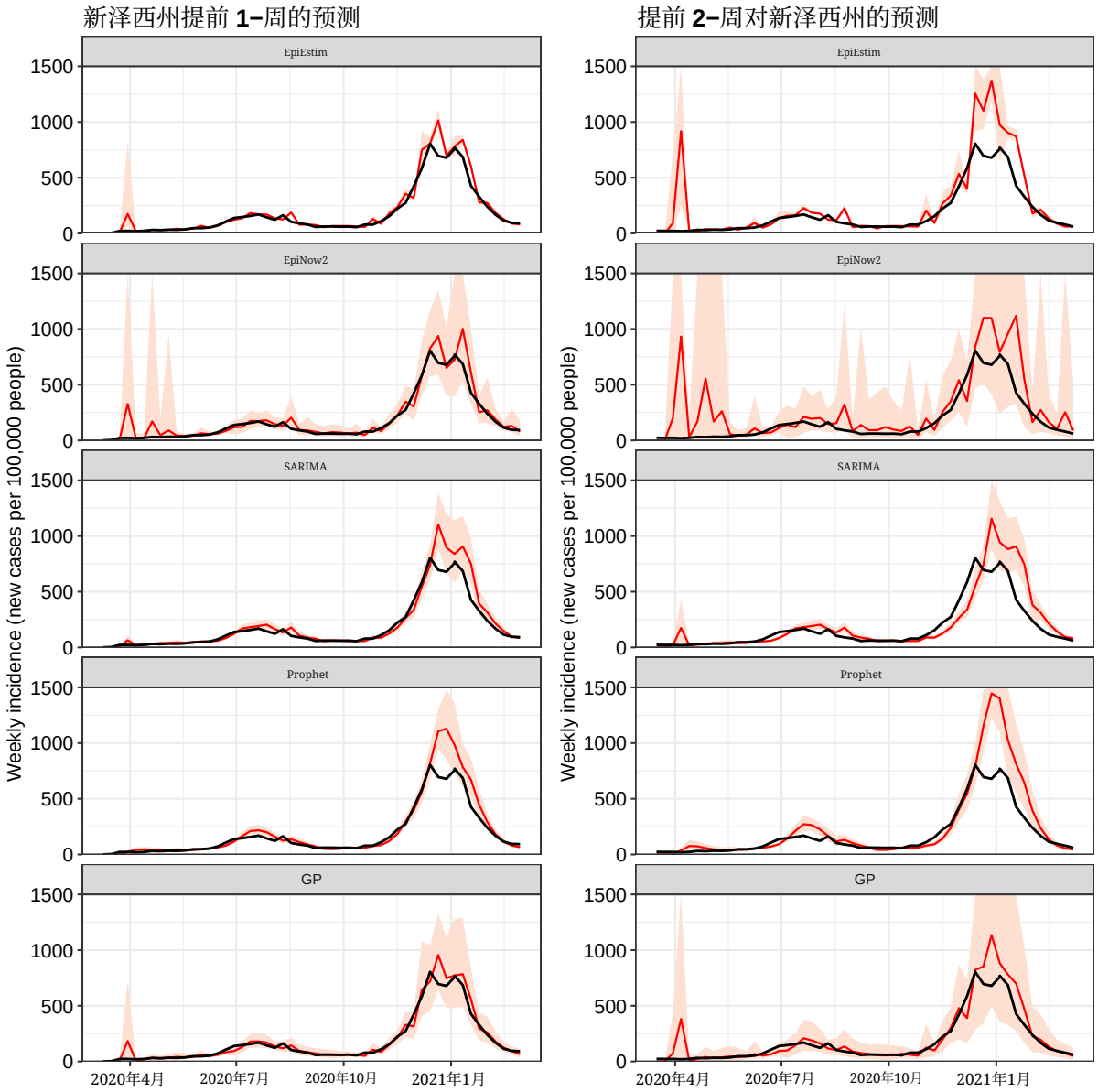


图 S6. 各模型对新泽西州的概率预测。 各模型对新泽西州的概率预测（后验均值以红线表示，95%预测区间以阴影区域表示）。(a) 提前1周预测。(b) 提前2周预测。黑线表示观测发病率。

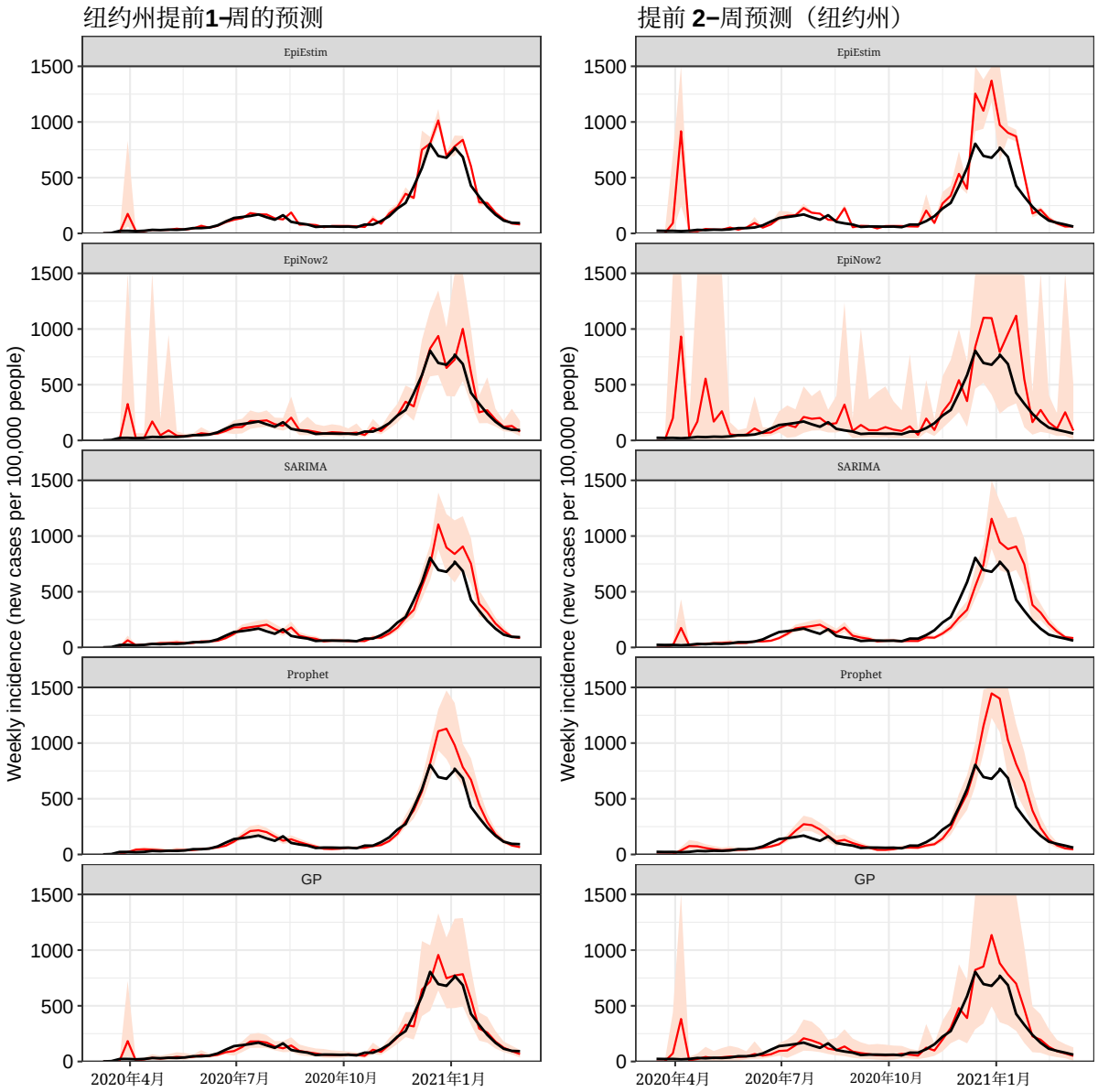


图 S7. 各模型对纽约州的概率预测。各模型对纽约州的概率预测（后验均值以红线表示，95%预测区间以阴影区域表示）。(a) 提前1周预测。(b) 提前2周预测。黑线表示观测发病率。

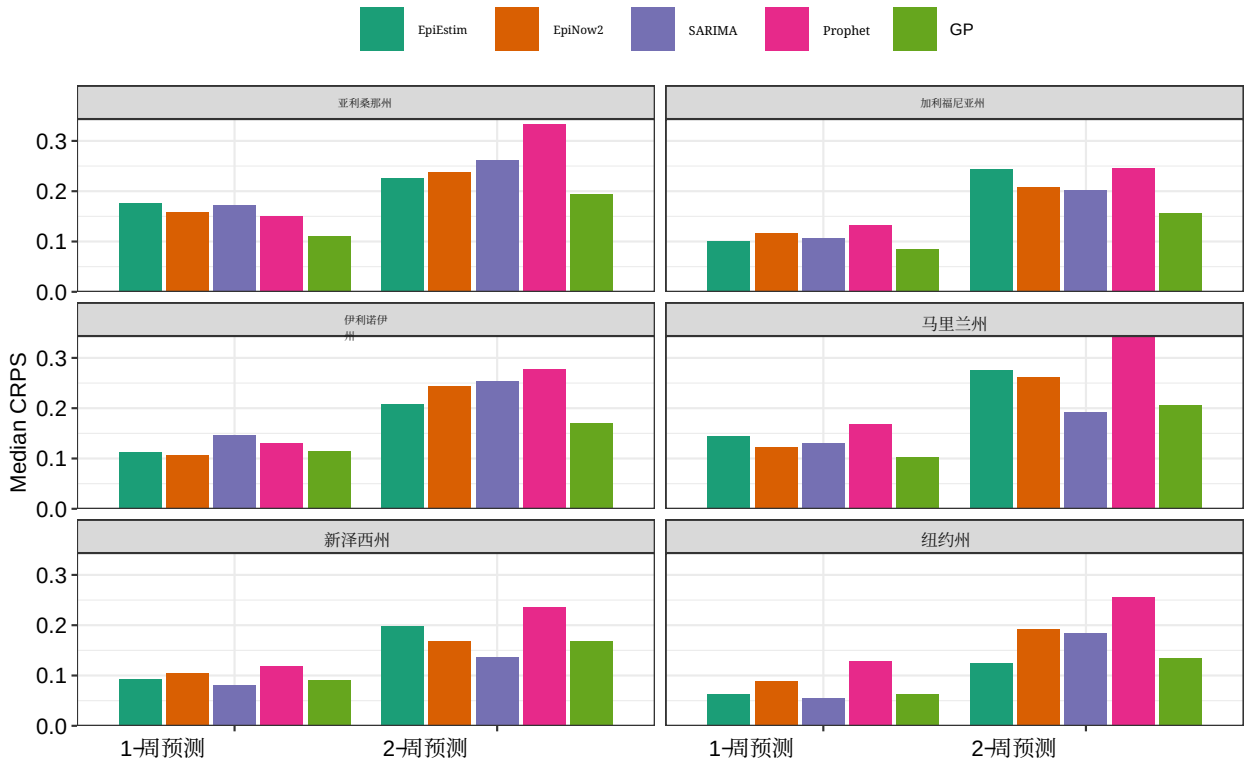


图 S8. 按美国州评估预测性能。各模型提前1周和2周预测的美国州中位数 CRPS。

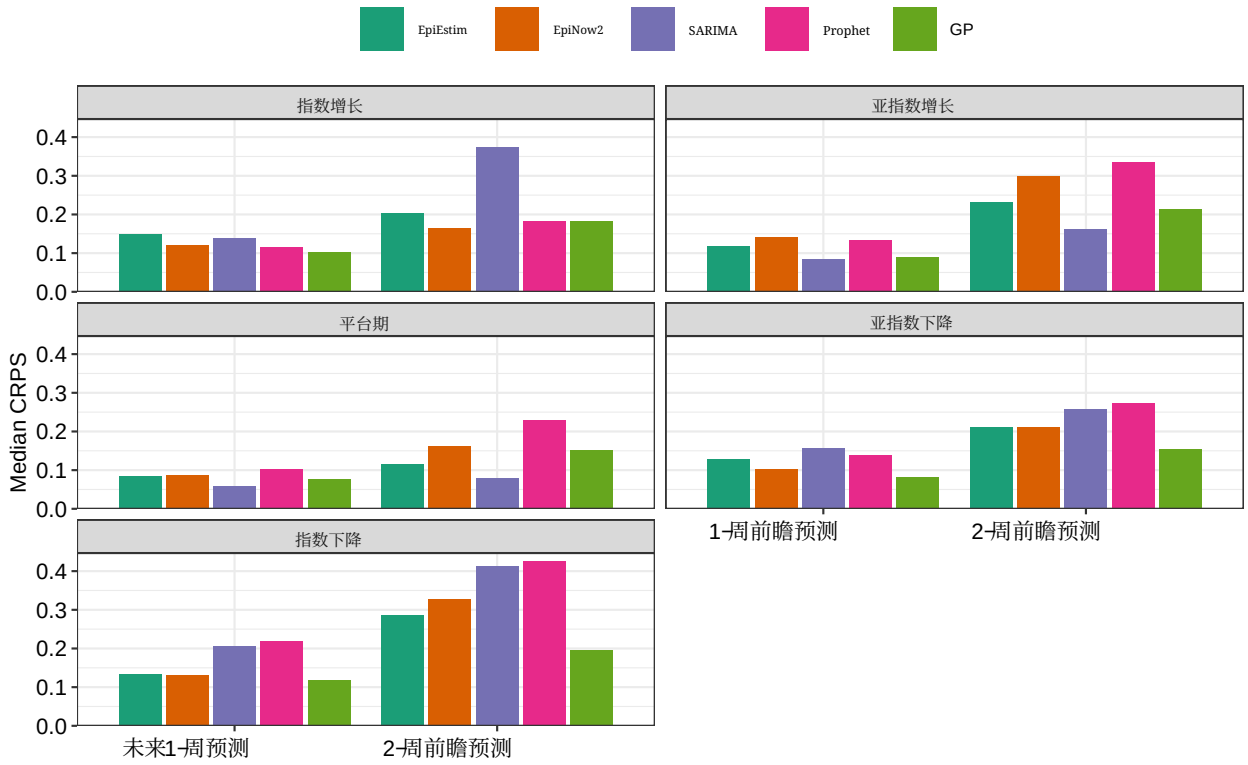


图 S9. 按流行病阶段评估预测性能。不同流行病阶段的中位数 CRPS（标签见图 S1）下模型提前1周和2周的预测。