Fast and Accurate Influenza Forecasting in the United States with Inferno

Dave Osthus

Los Alamos National Laboratory, Statistical Sciences Group

Abstract

Infectious disease forecasting is an emerging field and has the potential to improve public health through anticipatory resource allocation, situational awareness, and mitigation planning. By way of exploring and operationalizing disease forecasting, the U.S. Centers for Disease Control and Prevention (CDC) has hosted FluSight since the 2013/14 flu season, an annual flu forecasting challenge. Since FluSight's onset, forecasters have developed and improved forecasting models in an effort to provide more timely, reliable, and accurate information about the likely progression of the outbreak. While improving the predictive performance of these forecasting models is often the primary objective, it is also important for a forecasting model to run quickly, facilitating further model development, improvement, and scalability. In this vein I introduce Inferno, a fast and accurate flu forecasting model inspired by Dante, the top performing model in the 2018/19 FluSight challenge. When compared to all models that participated in FluSight 2018/19, Inferno would have placed 2nd in both the national and state challenges, behind only Dante. Inferno, however, runs in minutes and is trivially parallelizable, while Dante takes hours to run, representing a significant operational improvement with minimal impact to performance. A future consideration for forecasting competitions like FluSight will be how to encourage improvements to secondarily important properties of forecasting models, such as runtime, generalizability, and interpretability.

1 Introduction

Infectious disease outbreaks can be disruptive, deadly, and complex. By the end of November 2020, COVID-19 had killed almost 1.5 million people globally and over 250 thousand people in the United States (U.S.) [6]. Each year in the U.S., seasonal influenza kills tens of thousands of people and hospitalizes hundreds of thousands [25]. Life saving resources, such as respirators, antivirals, vaccines, and medical professionals must be allocated to ensure locations are

prepared and ready for the impending outbreak.

This is where infectious disease forecasting comes in. If forecasts can reliably anticipate the progression of an outbreak, we may be better prepared to confront it when it arrives. Infectious disease forecasting is still relatively young, but can no longer claim novelty. There has been a flurry of infectious disease forecasting challenges/collaborations in the last ten years, including the Defense Advanced Research Projects Agency's 2014/15 Chikungunya challenge [5], a collection of vector-borne disease challenges hosted by the U.S. Centers for Disease Control and Prevention (CDC) for dengue (2015) [9], Aedes (2019) [23], and West Nile virus (2020) [24], the U.S. CDC COVID-19 forecasting collaboration (2020) [27], and the U.S. CDC's flagship influenza forecasting challenge, FluSight, held annually since the 2013/14 flu season. The FluSight challenge alone has resulted in a wave of infectious disease forecasting model development [13, 12, 11, 14, 20, 19, 4, 3, 10, 15, 1, 28, 26].

The organizing body of a forecasting challenge (in the case of FluSight, the U.S. CDC) provides immense operational and research value by determining forecasting targets of public health relevance through interactions with their state and local public health partners, identifying relevant data sources and making them publicly available to forecasters, and defining the forecast evaluation criteria — a more challenging task than it may first appear (see [2] and [21]).

For instance, the FluSight challenge asks forecasters to predict seven targets on a weekly basis throughout the flu season: 1 through 4-week-ahead forecasts of influenza-like illness (ILI), the week of flu season onset, the week the flu season will peak, and the peak value of ILI for the flu season. Forecasts are made for states, Health and Human Services (HHS) regions, and the United States. ILI data collected by the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) are used for forecasting; targets are defined as summaries of ILI data. FluSight uses the log scoring rule to evaluate forecasts. The log scoring rule evaluates probabilistic forecasts, requiring forecasters to not only provide a prediction of what they think will happen in the future but also quantify how sure they are of that. The choice made by the U.S. CDC to use a log scoring rule makes clear their position that uncertainty quantification is of value to public health. Given a set of forecasting targets and an evaluation metric, forecasters develop models capable of forecasting the targets with the goal of maximizing their forecast evaluation score.

While a forecasting model's predictive performance is and should be of primary importance, it is not exclusively important. This seems obvious upon even cursory consideration. For instance, all else equal, an interpretable forecasting model is better than a black box forecasting model. All else equal, a generalizable forecasting model applicable to many disease forecasting contexts is better than a highly-tailored model to a specific disease context. All else equal, a forecasting model that runs quickly and is scalable is better than one that is slow and computationally expensive. While all of these seem obvious, none of these secondary factors are incorporated into FluSight's forecast evaluation criteria; it *only* measures the pre-

dictive performance of the model. As a result, much of the forecasting research of the past decade has focused more on developing models that improve forecasting scores and less on developing models that are generalizable, interpretable, scalable, and fast.

In this paper, I focus on improving the runtime of flu forecasting models while maintaining high prediction standards with the presentation of Inferno, a fast and accurate flu forecasting model. Inferno is a parallelizable, empirical Bayesian forecasting model inspired by Dante, the top performing model in FluSight 2018/19 [14]. The achieved goal of Inferno is to maintain the high predictive performance of Dante but substantially decrease the runtime. As will be discussed later, Inferno would have placed 2nd only to Dante in the 2018/19 FluSight challenge, but runs in minutes rather than hours, constituting a significant speed-up in operational performance.

In the remainder of this paper, I describe the details to Inferno (Section 2) and present Inferno's forecasting performance as compared to all participating models in FluSight 2018/19 (Section 3). I conclude the paper by raising important questions the infectious disease forecasting community must grapple with in order to improve the utility for forecasting challenges for public health.

2 Methods

2.1 Dante Background

Dante is a multiscale, probabilistic, influenza forecasting model. Dante has two sub-models: a state forecasting model and an aggregation model which combines state forecasts to produce HHS regional and United States forecasts. The state forecasting model is a statistical model where the expectation of ILI on a given week, state, and season is modeled as a function of four components: an overall trend component, a state-specific deviation component, a season-specific deviation component, and a state/season-specific deviation component. These four components are each modeled as random or reverse-random walks — flexible time series models that capture temporal correlation. By modeling all states and past flu seasons jointly, Dante achieves self-consistency and is able to borrow information across seasons and space. By modeling the HHS regional and United States forecasts as U.S. Census weighted averages of state forecasts, Dante ensures self-consistency across geographic scales. For more details on Dante, see [14].

Dante is a fully Bayesian model, capturing uncertainty in all model parameters, latent states, and forecasts through its posterior (predictive) distribution. The fully Bayesian formulation and self-consistency of Dante comes at a computational price, however. Dante represents a large model that will grow each year as more historical data are added and is not well-positioned to scale with possible future changes/expansions to FluSight (e.g., county-level forecasting). Nothing is precomputed and due to its interconnected model structure, it

is not obvious how to break up Dante to exploit parallelization.

Inferno was developed to addresses these computational shortcomings. Inferno is an empirical Bayesian analogue to the fully Bayesian Dante, where instead of modeling historical flu seasons directly, Inferno uses historical flu seasons to precompute model parameters. Inferno trades in self-consistency for parallelization, allowing all states, HHS regions, and the United States to be fit independently. In Section 2.2, I describe the Inferno forecasting model.

2.2 Inferno

Let $y_t \in (0, 1)$ for $t = 1, 2, \ldots, T$ be ILI/100 for states or state-weighted ILI/100 for HHS regions or the United States (collectively referred to as (w)ILI) for week of season t, where $t = 1$ corresponds to Morbidity and Mortality Weekly Report (MMWR) week 40, roughly the beginning of October, and $T = 35$ roughly corresponds to the end of May. Inferno's generative model is defined as follows:

$$y_t|\theta_t, \alpha \sim \text{Beta}(\alpha\theta_t, \alpha(1-\theta_t)) \tag{1}$$

$$\theta_t = \text{logit}^{-1}(\gamma_t + \delta_t) \tag{2}$$

$$\boldsymbol{\delta}|\mu, \boldsymbol{\Sigma} \sim \text{GP}(\mu\mathbf{1}, \boldsymbol{\Sigma}) \tag{3}$$

$$\mu|\sigma_\mu^2 \sim \text{N}(0, \sigma_\mu^2) \tag{4}$$

$$\boldsymbol{\Sigma}_{i,i} = \sigma_\Sigma^2 \tag{5}$$

$$\boldsymbol{\Sigma}_{i,j\neq i} = \phi\sigma_\Sigma^2 \text{exp}(-\lambda(i-j)^2), \tag{6}$$

where y_t is the noisy but observable measurement of ILI/100 on week t, θ_t is the true but unobservable value of ILI/100 on week t, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_T)'$ is a $T \times 1$ vector, $\mathbf{1}$ is a $T \times 1$ vector of 1s, $\boldsymbol{\Sigma}$ is a $T \times T$ positive semi-definite matrix, $\text{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a Gaussian process (GP) with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the scalar parameters α, σ_μ^2, σ_Σ^2, λ are all greater than 0, and $\phi \in [0, 1]$. In this paper, bold quantities represent vectors or matrices, while non-bold quantities represent scalars. The Beta distribution of Equation 1 requires $y_t \in (0, 1)$. Thus, all y_t below a low threshold l are set equal to l and all y_t above $1 - l$ are set to $1 - l$. For this work, $l = 0.0005$.

Inferno takes an empirical Bayesian approach, where unknown parameters are estimated from historical training data. The following outlines a six step procedure to estimate the unknown parameters α, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_T)'$, σ_μ^2, σ_Σ^2, λ, and ϕ and to use Markov chain Monte Carlo (MCMC) to sample and forecast from Inferno's posterior predictive distribution.

2.2.1 Step 1: Estimate $\theta_{s,t}$

For a given geographical unit (e.g., state, region country), let $y_{s,t}$ by (w)ILI for training season $s \in 1, 2, \ldots, S$ and week of season t. Fit $\hat{\beta}_{s,t}$ as a 3-week moving average:

$$\hat{\beta}_{s,1} = \frac{1}{2}(y_{s,1} + y_{s,2}) \tag{7}$$

$$\hat{\beta}_{s,t\neq 1,T} = \frac{1}{3}(y_{s,t-1} + y_{s,t} + y_{s,t+1}) \tag{8}$$

$$\hat{\beta}_{s,T} = \frac{1}{2}(y_{s,T-1} + y_{s,T}). \tag{9}$$

Figure 1 shows the moving average fit to ILI in Illinois. By construction, the moving average captures the shape of the ILI curve. The moving average, however, can miss sharp changes in ILI caused by differences in reporting practices over holidays. For instance, we see that the moving average most often underestimates ILI the week of Christmas ($t = 13$, or MMWR week 52).

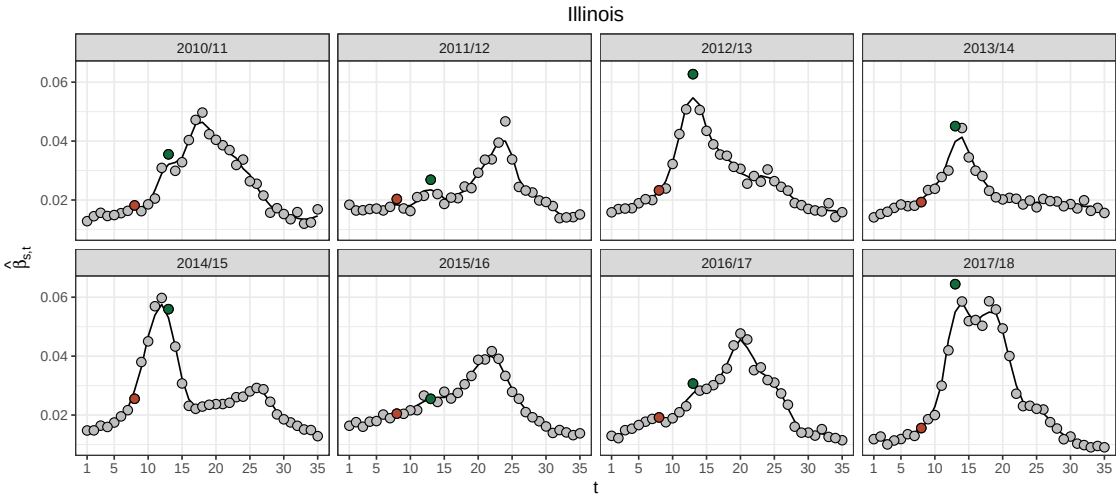


Figure 1: ILI (grey points) and $\hat{\beta}_{s,t}$ (black line) for the historical seasons for Illinois. ILI for the week of Thanksgiving ($t = 8$) and Christmas ($t = 13$) are highlighted in brown and green, respectively. $\hat{\beta}_{s,t}$ typically underestimates the sharp uptick in ILI observed on Christmas and to a lesser extent Thanksgiving, which is likely a result of changes in reporting and care-seeking behavior over the holidays.

To capture the systematic sharp changes in ILI that are common across training seasons, Inferno computes the quantity τ_t:

$$\hat{\tau}_t = \frac{1}{S}\sum_{s=1}^{S}(y_{s,t} - \hat{\beta}_{s,t}). \tag{10}$$

Figure 2 plots τ_t for all states. τ_t captures the holiday effects in ILI, with a small but consistent positive τ_t on the week of Thanksgiving ($t = 8$, or MMWR week 47) and a larger positive effect the week of Christmas.

Finally, the quantity $\theta_{s,t}$ captures both the ILI profile ($\hat{\beta}_{s,t}$) and the holiday effects ($\hat{\tau}_t$):

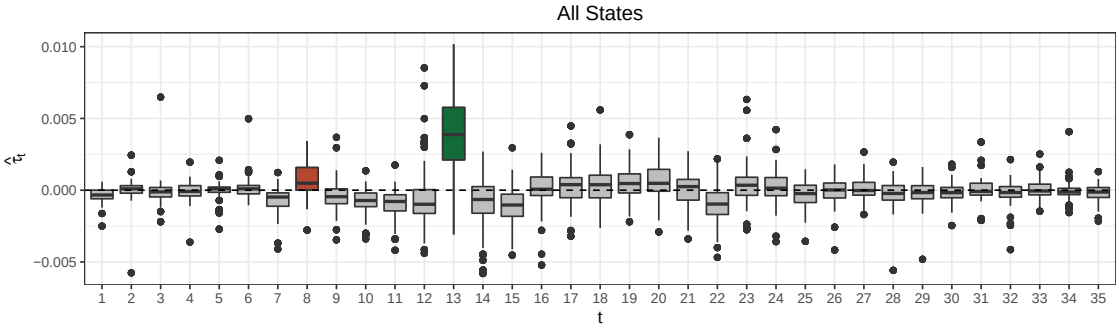


Figure 2: The quantity $\hat{\tau}_t$ for all states. $\hat{\tau}_t$ the week of Thanksgiving (brown) and Christmas (green) are systematically positive, likely as a result of systematic changes to reporting and care-seeking behavior over the holidays.

$$\hat{\theta}_{s,t} = \begin{cases} l & \text{if } \hat{\beta}_{s,t} + \hat{\tau}_t < l \\ 1 - l & \text{if } \hat{\beta}_{s,t} + \hat{\tau}_t > 1 - l \\ \hat{\beta}_{s,t} + \hat{\tau}_t & \text{otherwise.} \end{cases} \tag{11}$$

Figure 3 shows how $\hat{\theta}_{s,t}$ tracks the profile of the ILI season, like $\hat{\beta}_{s,t}$, but better tracks ILI on the holidays, especially Christmas.

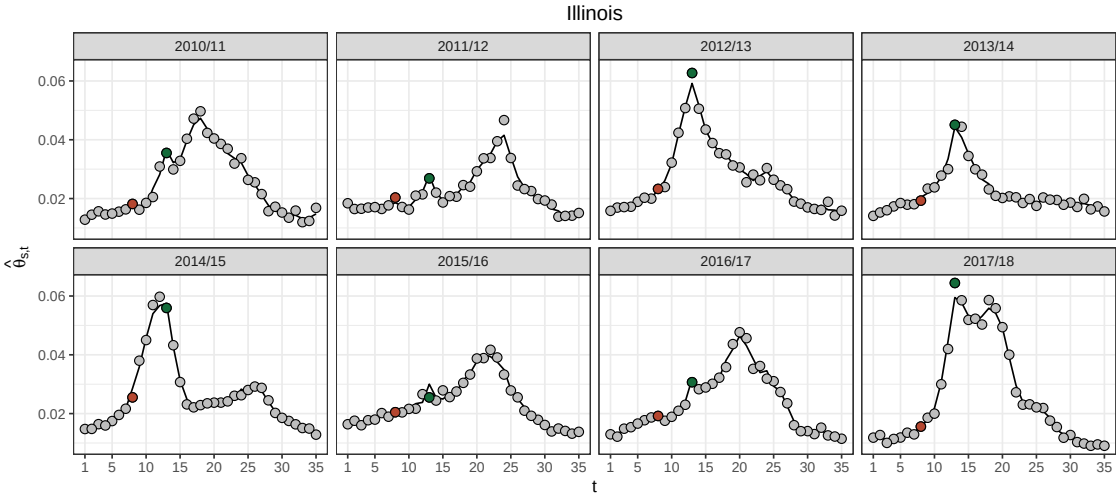


Figure 3: ILI (grey points) and $\hat{\theta}_{s,t}$ (black line) for the historical seasons for Illinois. ILI for the week of Thanksgiving ($t = 8$) and Christmas ($t = 13$) are highlighted in brown and green, respectively. $\hat{\theta}_{s,t}$ better matches ILI data on the holidays than $\hat{\beta}_{s,t}$ (Figure 1) by accounting for the systematic reporting and care-seeking changes over the holidays, as accounted for by $\hat{\tau}_t$.

2.2.2 Step 2: Estimate α

Inferno estimates $\hat{\theta}_{s,t}$ in order to facilitate the estimation of the other unknown quantities of Inferno's generative model. The expectation and the variance of Inferno's data model (Equation 1) are,

$$\mathrm{E}(y_t|\theta_t, \alpha) = \theta_t \tag{12}$$

$$\mathrm{Var}(y_t|\theta_t, \alpha) = \frac{\theta_t(1-\theta_t)}{1+\alpha}. \tag{13}$$

The parameter α controls the variance of the data model, capturing the week-to-week variability in the ILI data. The larger α is, the smaller the variance reflecting less week-to-week noise in the ILI data. The smaller α is, the larger the variance reflecting more week-to-week noise in the ILI data. We estimate $\alpha > 0$ as the maximum likelihood estimate (MLE) of Inferno's data model by minimizing the negative log likelihood:

$$\hat{\alpha} = \underset{\alpha}{\mathrm{argmin}} \sum_{s=1}^{S}\sum_{t=1}^{T} -\log(\mathrm{Beta}(y_{s,t}|\hat{\theta}_{s,t}, \alpha)), \tag{14}$$

where $\log(x)$ is the natural log of x,

$$\mathrm{Beta}(y_{s,t}|\hat{\theta}_{s,t}, \alpha) = \frac{y_{s,t}^{a-1}(1-y_{s,t})^{b-1}}{\mathrm{B}(a,b)} \tag{15}$$

$$\mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{16}$$

$$a = \alpha\hat{\theta}_{s,t} \tag{17}$$

$$b = \alpha(1-\hat{\theta}_{s,t}), \tag{18}$$

and $\Gamma()$ is the gamma function.

The top of Figure 4 shows $\hat{\alpha}$ for all states, territories, and cities (collectively referred to as states). States like the U.S. Virgin Islands, North Dakota, and Puerto Rico have the smallest $\hat{\alpha}$s, reflecting they have the largest week-to-week noise in their ILI data, while states like California, Illinois, and New York City have the largest $\hat{\alpha}$s, reflecting they have the smallest week-to-week noise in their ILI data. The bottom of Figure 4 shows the 95% prediction interval for $\mathrm{Beta}(\hat{\alpha}\hat{\theta}_{s,t}, \hat{\alpha}(1-\hat{\theta}_{s,t}))$ for North Dakota, Nevada, and Illinois, illustrating the different levels of week-to-week noise in ILI data across states.

2.2.3 Step 3: Estimate γ_t

Seasonal flu has a typical shape to it in the United States. ILI starts at low levels early in the season, rises to a peak between December and March, and reverts to low levels by the end of May. The role of $\boldsymbol{\gamma}$ is to capture this typical seasonal flu profile. Inferno computes γ_t as follows:

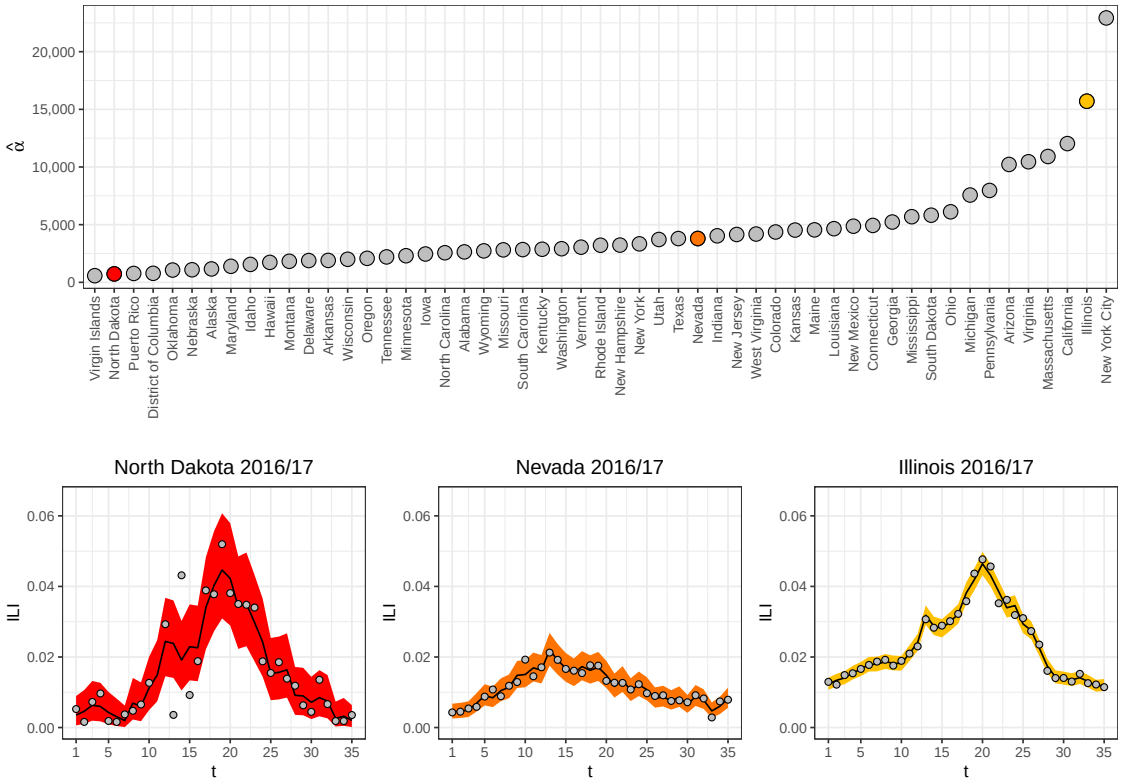


Figure 4: (Top) $\hat{\alpha}$ for all states based on training data from 2010/2011 through 2017/18. (Bottom) ILI (grey points), $\hat{\theta}_{s,t}$ (black line) and 95% prediction interval for Beta$(\hat{\alpha}\hat{\theta}_{s,t}, \hat{\alpha}(1-\hat{\theta}_{s,t}))$ (ribbon) for North Dakota, Nevada, and Illinois in 2016/17. $\hat{\alpha}$ captures the week-to-week noise in ILI data that systematically varies from state-to-state, where North Dakota has more week-to-week noise than Illinois.

$$\hat{\gamma}_t = \frac{1}{S}\sum_{s=1}^{S} \text{logit}(\hat{\theta}_{s,t}), \tag{19}$$

where $\text{logit}(p) = \log(p/(1-p))$.

Figure 5 shows $\hat{\boldsymbol{\gamma}}$ for North Dakota, Nevada, and Illinois. We see for all states, $\hat{\boldsymbol{\gamma}}$ captures the typical profile of seasonal flu on the logit scale, with low levels at the beginning of the flu season, ramping up to a peak in the middle, then reverting back to low levels by the end.

2.2.4 Step 4: Estimate σ_μ^2

Equation 2 is the mean of Inferno's data model. While $\boldsymbol{\gamma}$ captures the typical profile of seasonal flu, $\boldsymbol{\delta}$ captures season-specific deviations from $\boldsymbol{\gamma}$. Inferno models $\boldsymbol{\delta}$ with a Gaussian process (GP), a stochastic process where any finite collection of random variables has a multivariate normal distribution. That is,

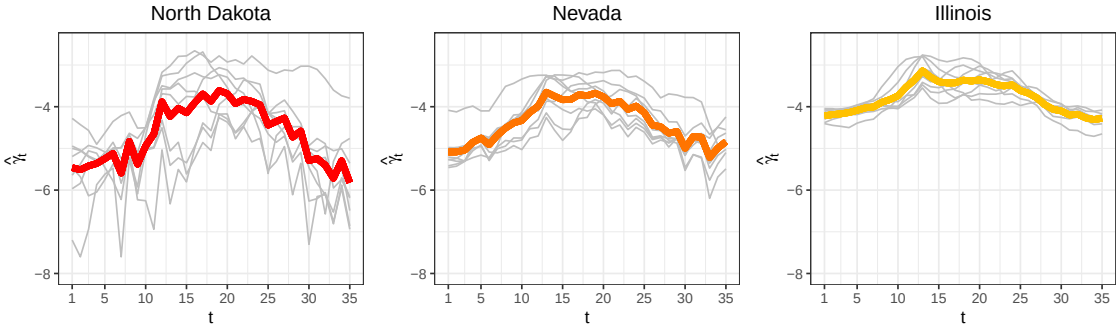


Figure 5: $\hat{\boldsymbol{\gamma}}$ (colored line) and $\text{logit}(\hat{\boldsymbol{\theta}}_s)$ (grey lines) for North Dakota, Nevada, and Illinois. $\hat{\boldsymbol{\gamma}}$ captures the typical profile of seasonal flu specific to each state on the logit scale.

$$\text{GP}(\boldsymbol{\delta}|\mu\mathbf{1}, \boldsymbol{\Sigma}) = (2\pi)^{-T/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{\delta}-\mu\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}-\mu\mathbf{1})\right), \tag{20}$$

where $\mathbf{1}$ is a $T \times 1$ vector of ones, $\boldsymbol{\Sigma}$ is a $T \times T$ positive semi-definite matrix, $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}^{-1}$ is the inverse of $\boldsymbol{\Sigma}$. The model for the mean of the GP μ is

$$\mu \sim \text{N}(0, \sigma_\mu^2). \tag{21}$$

Step 4 describes how to estimate σ_μ^2.

First compute the following quantities:

$$\hat{\delta}_{s,t} = \text{logit}(\hat{\theta}_{s,t}) - \hat{\gamma}_t \tag{22}$$

$$\hat{\mu}_s = \frac{1}{T}\sum_{t=1}^{T}\hat{\delta}_{s,t}. \tag{23}$$

The top of Figure 6 shows $\hat{\boldsymbol{\delta}}_s$ and $\hat{\mu}_s$ for North Dakota, Nevada and Illinois. The quantity $\hat{\mu}_s$ captures how far, on average, $\hat{\boldsymbol{\delta}}_s$ deviates from $\mathbf{0}$.

The quantity $\hat{\sigma}_\mu^2$ is computed as the unbiased sample variance of $\hat{\boldsymbol{\mu}}$:

$$\hat{\sigma}_\mu^2 = \frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\mu}_s - \frac{1}{S}\sum_{s=1}^{S}\hat{\mu}_s\right)^2. \tag{24}$$

The bottom of Figure 6 shows $\hat{\sigma}_\mu^2$ for all states. Some states, like North Dakota, have appreciable average season-to-season variation while other states, like Illinois, have smaller average season-to-season deviations from their typical seasonal flu profiles.

2.2.5 Step 5: Estimate σ_Σ^2, λ, ϕ

Step 5 estimates the covariance parameters in $\boldsymbol{\Sigma}$. The covariance matrix captures different characteristics of $\boldsymbol{\delta}$. Recall Equations 5 and 6:

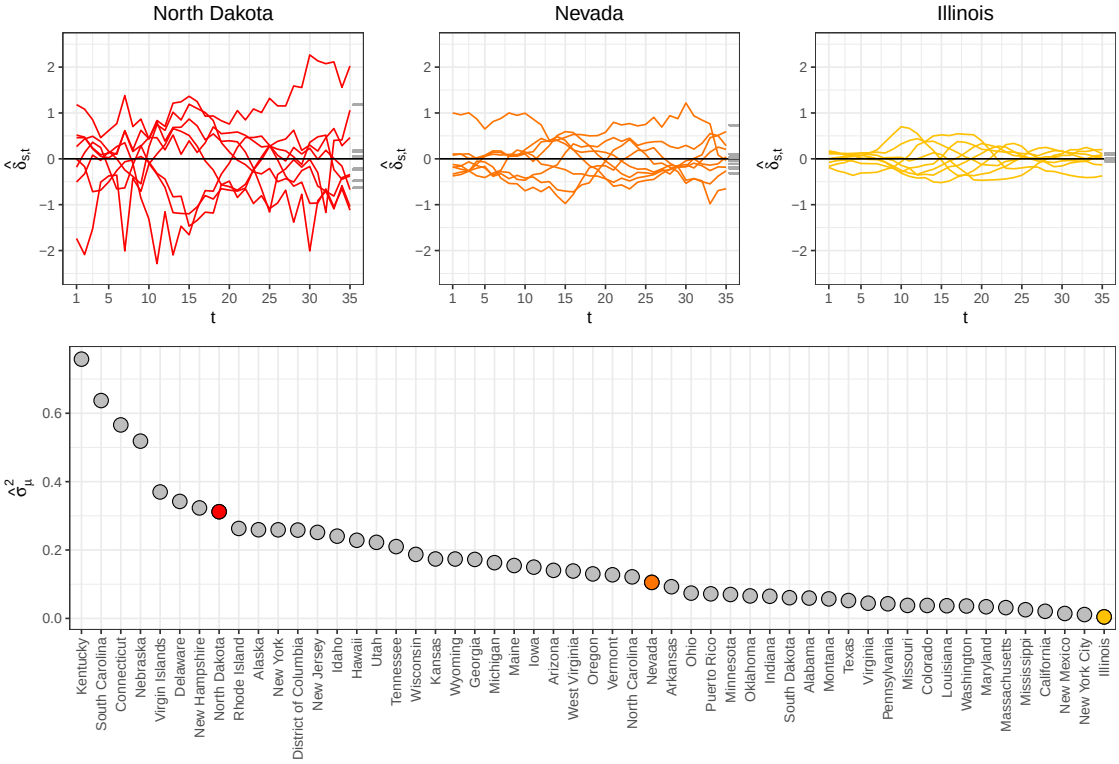


Figure 6: (Top) $\hat{\boldsymbol{\delta}}_s$ (colored lines) and $\hat{\mu}_s$ (grey tick marks) for North Dakota, Nevada, and Illinois. North Dakota exhibits more season-to-season variability in $\hat{\mu}_s$ than Illinois, as can be seen in the spread of $\hat{\mu}_s$. (Bottom) $\hat{\sigma}^2_\mu$ for all states. Considerable variation in $\hat{\sigma}^2_\mu$ across states is observed.

$$\boldsymbol{\Sigma}_{i,i} = \sigma^2_\Sigma$$
$$\boldsymbol{\Sigma}_{i,j\neq i} = \phi\sigma^2_\Sigma \exp(-\lambda(i-j)^2).$$

The parameter σ^2_Σ is the marginal variance for the GP. It captures how far $\boldsymbol{\delta} - \mu\mathbf{1}$ typically deviates from 0. The top of Figure 7 plots $\hat{\boldsymbol{\delta}}_s - \hat{\mu}_s\mathbf{1}$ for North Dakota, Nevada, and Illinois. North Dakota exhibits more variability than Illinois as can be seen with its wider range of values. Inferno estimates σ^2_Σ as

$$\hat{\sigma}^2_\Sigma = \frac{1}{ST-1}\sum_{s=1}^{S}\sum_{t=1}^{T}(\hat{\delta}_{s,t} - \hat{\mu}_s)^2. \tag{25}$$

The remaining parameters of $\boldsymbol{\Sigma}$ are ϕ and λ. They collectively capture two different characteristics of $\boldsymbol{\delta}$. The parameter ϕ captures the smoothness of $\boldsymbol{\delta}$. For instance, $\hat{\boldsymbol{\delta}}_s$ for Illinois in Figure 6 are much smoother than $\hat{\boldsymbol{\delta}}_s$ for North Dakota. The parameter ϕ captures this feature, with ϕ close to 1 resulting in smoother $\boldsymbol{\delta}$s. The second characteristic of $\boldsymbol{\delta}$ captured by ϕ and λ is the correlation between entries of $\boldsymbol{\delta}$. For instance, the correlation between δ_i

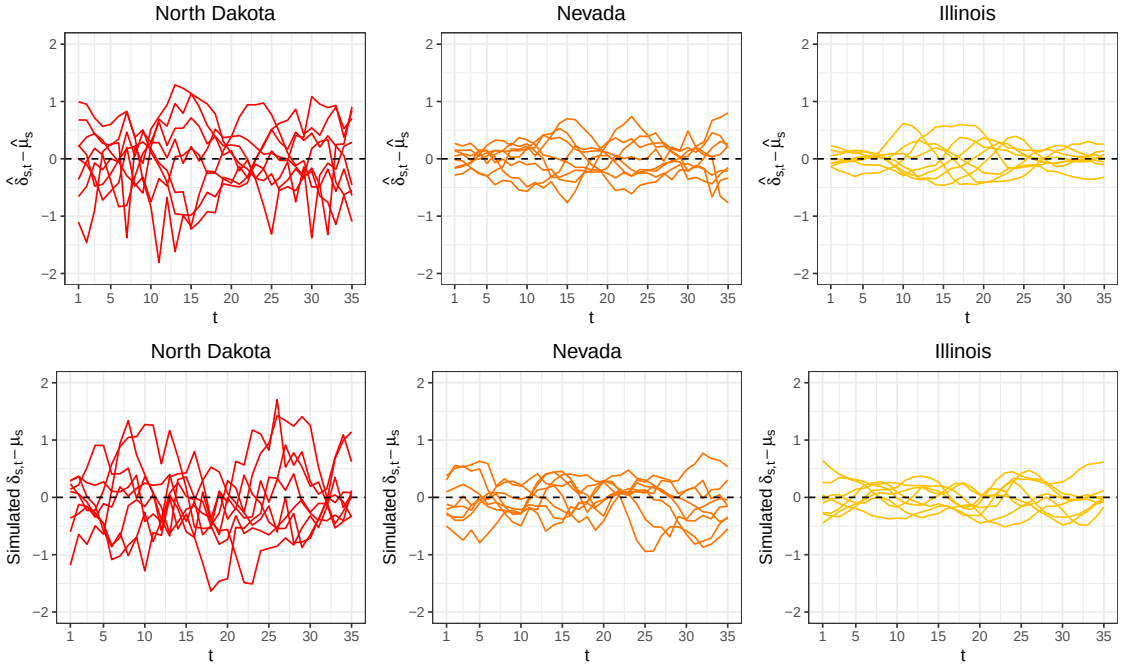


Figure 7: (Top row) The estimated times series from training data for the quantities $\hat{\boldsymbol{\delta}}_s - \hat{\mu}_s \mathbf{1}$. (Bottom row) Realizations drawn from $\text{GP}(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$. Good visual agreement is seen between the simulated $\boldsymbol{\delta}_s - \mu_s \mathbf{1}$ and the $\hat{\boldsymbol{\delta}}_s - \hat{\mu}_s \mathbf{1}$ calculated from training data, suggesting the Gaussian process is able to capture heterogenous discrepancy characteristics across states.

and δ_{i+1} is

$$\text{Cor}(\delta_i, \delta_{i+1}) = \frac{\text{Cov}(\delta_i, \delta_{i+1})}{\sqrt{\text{Var}(\delta_i)}\sqrt{\text{Var}(\delta_{i+1})}} = \frac{\sigma_\Sigma^2 \phi \exp(-\lambda(i-(i+1))^2)}{\sigma_\Sigma^2} = \phi \exp(-\lambda). \tag{26}$$

Inferno estimates ϕ and λ by minimizing the negative log likelihood:

$$\hat{\lambda}, \hat{\phi} = \underset{\lambda,\phi}{\text{argmin}} \sum_{s=1}^{S} -\log\Big(\text{GP}(\hat{\boldsymbol{\delta}}_s | \hat{\mu}_s, \hat{\sigma}_\Sigma^2, \lambda, \phi)\Big). \tag{27}$$

Figure 8 plots the covariance parameter estimates for all states. North Dakota has larger marginal variance (larger $\hat{\sigma}_\Sigma^2$), less smoothness (smaller $\hat{\phi}$), and lower correlation (smaller $\hat{\phi}\exp(-\hat{\lambda})$) than Illinois.

The bottom of Figure 7 shows realizations drawn from $\text{GP}(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$. The fitted GP does a good job capturing the different characteristics of the empirical quantities $\hat{\boldsymbol{\delta}}_s - \hat{\mu}_s \mathbf{1}$, suggesting the GP is a defensible generative model for $\boldsymbol{\delta}$.

2.2.6 Step 6: Sample Forecasts from Inferno

The sixth and final step of Inferno is to replace parameters with their estimates and sample from the posterior predictive distribution. Specifically, the generative model with parameters

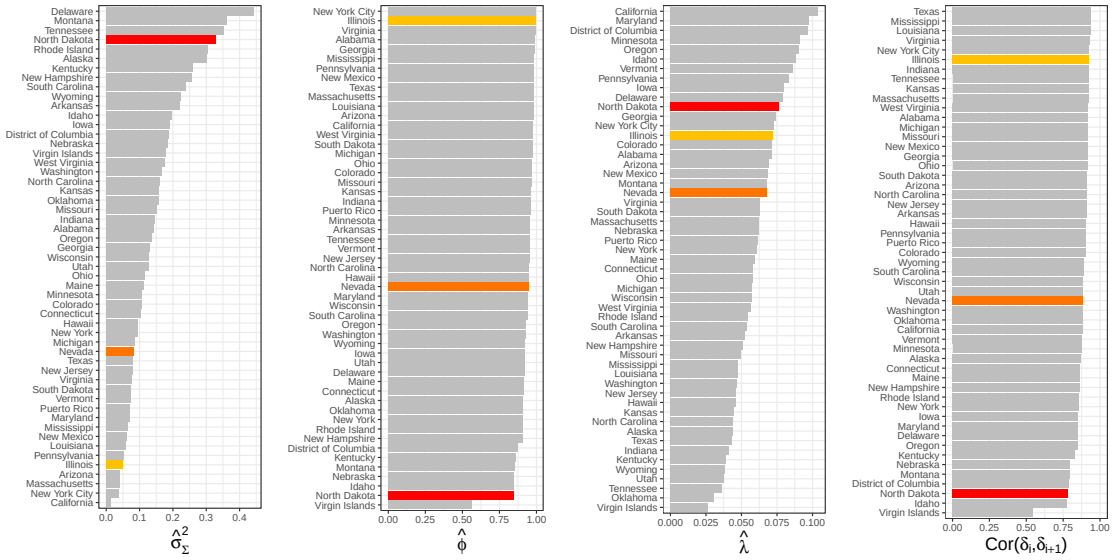


Figure 8: Estimated GP covariance parameter estimates for all states. Parameter estimates for North Dakota, Nevada, and Illinois are highlighted in red, orange, and yellow, respectively. North Dakota has larger marginal variance (larger $\hat{\sigma}^2_\Sigma$), less smoothness (smaller $\hat{\phi}$), and lower correlation (smaller $\text{Cor}(\delta_i, \delta_{i+1})$) than Illinois.

replaced by their estimates is

$$y_t|\theta_t \sim \text{Beta}(\hat{\alpha}\theta_t, \hat{\alpha}(1-\theta_t)) \tag{28}$$

$$\theta_t = \text{logit}^{-1}(\hat{\gamma}_t + \delta_t) \tag{29}$$

$$\boldsymbol{\delta}|\mu \sim \text{GP}(\mu\mathbf{1}, \hat{\boldsymbol{\Sigma}}) \tag{30}$$

$$\mu \sim \text{N}(0, \hat{\sigma}^2_\mu) \tag{31}$$

$$\hat{\boldsymbol{\Sigma}}_{i,i} = \hat{\sigma}^2_\Sigma \tag{32}$$

$$\hat{\boldsymbol{\Sigma}}_{i,j\neq i} = \hat{\phi}\hat{\sigma}^2_\Sigma \text{exp}(-\hat{\lambda}(i-j)^2). \tag{33}$$

Inferno forecasts the entire flu season by sampling from the posterior predictive distribution given the first t weeks of ILI observations:

$$[\tilde{\boldsymbol{y}}|\boldsymbol{y}_{1:t}] = \int [\tilde{\boldsymbol{y}}, \boldsymbol{\theta}|\boldsymbol{y}_{1:t}] d\boldsymbol{\theta} = \int [\tilde{\boldsymbol{y}}|\boldsymbol{\theta}][\boldsymbol{\theta}|\boldsymbol{y}_{1:t}] d\boldsymbol{\theta}, \tag{34}$$

where $[X|Y]$ is the conditional distribution of X given Y and $\tilde{\boldsymbol{y}}$ is assumed to be independent of $\boldsymbol{y}$, given $\boldsymbol{\theta}$ where $\boldsymbol{\theta}$ generically represents all parameters and latent states of Inferno. The posterior predictive distribution of Equation 34 is not known in closed form. Markov chain Monte Carlo (MCMC) sampling is used to draw from the posterior predictive distribution. The probabilistic programming language JAGS (Just Another Gibbs Sampler) [16] is used to execute the MCMC sampling. JAGS is called with functions from the `rjags` package [17] in

the programming language R [18]. The results are J draws from the posterior predictive distribution of Equation 34. For this paper, forecasts are based on $J = 25,000$ draws, discarding the first 12,500 draws as burn-in and thinning the remaining 12,500 draws by two, resulting in forecasts based on 6,250 MCMC draws. The JAGS code that implements Inferno can be found in Appendix A.

Figure 9 shows the forecasts for North Dakota, Nevada, and Illinois throughout the 2018/19 flu season. The posterior predictive mean and the 95% posterior prediction interval are shown as summaries of the forecasts.

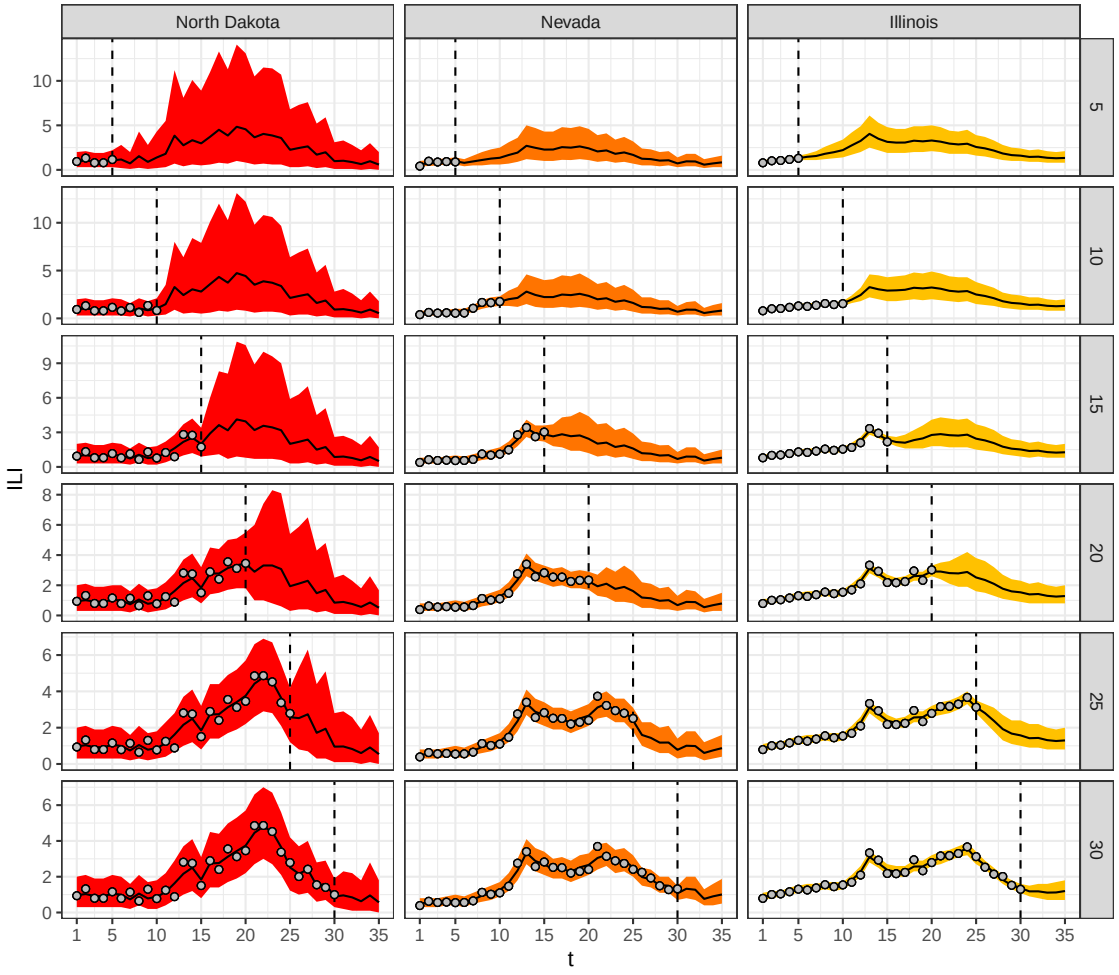


Figure 9: Inferno forecasts for the 2018/19 flu season for North Dakota, Nevada, and Illinois (columns) made $t = 5, 10, 15, 20, 25, 30$ weeks into the flu season based on summaries of draws from the posterior predictive distribution $[\tilde{\boldsymbol{y}}|\boldsymbol{y}_{1:t}]$ of Equation 34 (rows). Posterior mean (black line) and 95% prediction intervals (ribbons) are displayed, along with $\boldsymbol{y}_{1:t}$ (grey points).

3 Results

Inferno is compared to all models that participated in the U.S. CDC's 2018/19 National and Regional FluSight challenge as well as the State challenge. Forecasting follows the guidelines outlined by the CDC FluSight challenge; see [22] for details. The forecasts and the evaluation procedure is briefly described below.

Forecasts are made for four short-term targets (1, 2, 3, and 4-week-ahead) and three seasonal targets (the peak week, the peak percentage, and the onset week — onset is not forecasted for the State challenge). All forecast targets are binned and a probability is assigned to each bin such that the sum of all probabilities over all bins for a target equals 1. Say bin b is the bin containing the correct target and $p_b \in [0, 1]$ is the probability assigned to the correct bin. The log score is then $\max(-10, \log(p_b))$. When b is the bin of the correct target, the evaluation criteria is called the *single-bin log score*; single-bin log score is the scoring criteria used starting with the 2019/20 FluSight challenge and is a proper score. When b is the set of bins containing the correct target plus/minus a set of predefined neighboring bins, the evaluation criteria is called the *multi-bin log score*; multi-bin log score is the scoring criteria used in the 2018/19 FluSight challenge. The multi-bin log score is an improper scoring rule [21]. Multi-bin skill and single-bin skill are derived by exponentiating the multi-bin and single-bin log scores, respectively.

ILI data is subject to weekly revisions. As a result, it is important to use the ILI estimates that were available at the time to make faithful comparisons to models that participated in the real-time FluSight challenges. Data available at historical dates are made available by the Carnegie Mellon University Delphi group's API [7] and were used to produce the results in this section.

Figure 10 and Table 1 show the multi- and single-bin skills for Inferno and all models that participated in the 2018/19 FluSight challenges. Inferno would have placed 2nd only to Dante in the 2018/19 FluSight National and Regional as well as State challenges. FluSight 2018/19 used multi-bin skill as the forecast evaluation. Starting with FluSight 2019/20, single-bin skill will be used. While single-bin and multi-bin skills are correlated, as can be seen in Figure 10, the relationship is not perfect. Models can rise or fall in the relative ranking depending on which evaluation metric is used for scoring, highlighting that the evaluation metric the forecasting challenge organizing body selects is of consequence. Inferno and Dante both perform better under the multi-bin skill evaluation than single-bin skill, but are both top 4 models by either evaluation metric. Most importantly, the drop in predictive performance from Dante to Inferno is small.

The small drop in predictive performance from Dante to Inferno is offset by Inferno's significant improvement in runtime and preparation for future scalability to more granular forecasting geographies. Figure 11 shows the runtime comparison between Dante and Inferno at different stages of the flu season. Dante takes between 90 and 105 minutes to produce 25,000 MCMC samples throughout the 2018/19 flu season. Inferno takes between 30 seconds

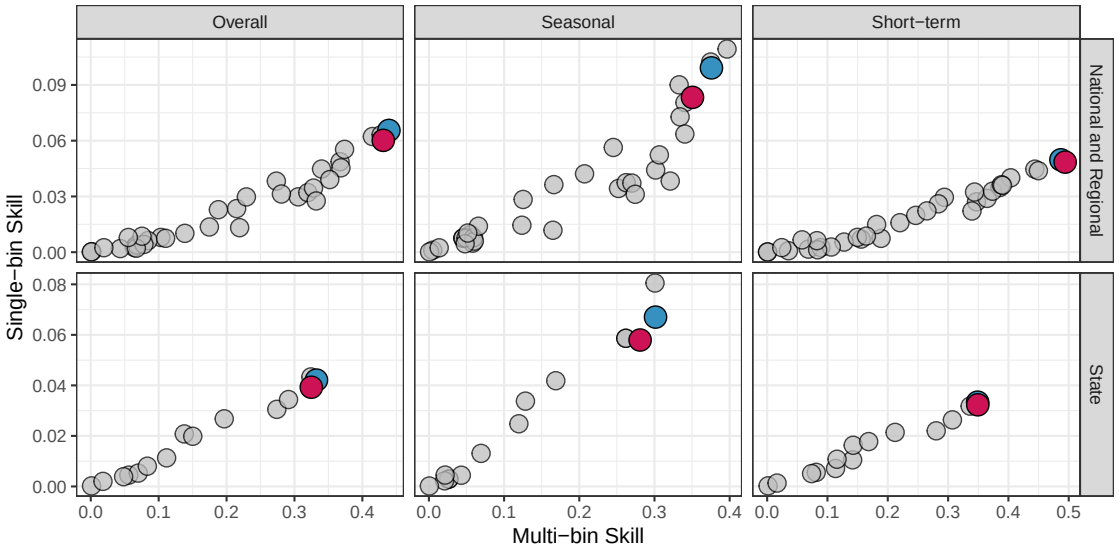


Figure 10: Results for the 2018/19 FluSight National and Regional challenge (top row) and State challenge (bottom row) for Inferno (red point), Dante (blue point) and all other models that participated in the 2018/19 FluSight challenges (grey points). The 2018/19 FluSight challenge evaluated models using multi-bin skill (x-axis), but starting with the FluSight 2019/20 challenge, will be using single-bin skill (y-axis). Skill scores are presented overall (left column), but also by seasonal targets (middle column) and short-term targets (right column). Inferno is a leading forecasting model overall, excelling in short-term forecasting, with good but not leading seasonal forecasting performance.

and 2 minutes to produce the same number of MCMC samples. The runtimes in Figure 11 are not directly comparable, as Inferno's runtimes are for only one of the 64 geographies (53 states, 10 HHS regions, and the United States), while Dante's runtimes are for all 64 geographies. If Inferno ran sequentially over all geographies, it would take roughly 30 minutes at the beginning of the season and 130 minutes at the end of the season, resulting in no computational gains over Dante by the end of the flu season. However, Inferno is trivially parallelizable. With cluster computing, all 64 geographies of Inferno could be computed simultaneously in two minutes or less. The significant advantage Inferno has over Dante is its scalability via parallelization. Thus, while Inferno's predictive performance is comparable to but slightly worse than Dante's, its significantly improved runtime and scalability make it a more attractive alternative for both the present and the future.

Table 1: The rank by challenge and target for Inferno and Dante as measured by single-bin and multi-bin skill. Inferno would have placed 2nd in both the National and Regional and the State challenges as measured by multi-bin skill, only finishing behind Dante. Inferno would have placed 4th (National and Regional) and 3rd (State) were the forecasts evaluated with single-bin skill. For both challenges and both evaluation metrics, Inferno achieved better short-term than seasonal performance.

2018/19 FluSight Challenge	**Target**	**Multi-bin Rank**		**Single-bin Rank**	
		Inferno	**Dante**	**Inferno**	**Dante**
National and Regional	**Overall**	2	1	4	1
	1 wk ahead	1	2	1	2
	2 wk ahead	1	2	2	1
	3 wk ahead	1	2	2	1
	4 wk ahead	2	1	2	1
	Season peak percentage	5	1	5	3
	Season peak week	11	8	11	8
	Season onset	5	1	7	1
State	**Overall**	2	1	3	2
	1 wk ahead	3	1	3	1
	2 wk ahead	2	1	2	1
	3 wk ahead	1	2	2	1
	4 wk ahead	1	2	1	2
	Season peak percentage	3	2	5	2
	Season peak week	3	1	3	1

4 Discussion

In this paper, I argued that while predictive performance is the most important measure of a forecasting model, it is not singularly important. Other factors like interpretability, generalizability, scalability, and runtime are also important. Developing a model with leading predictive performance but drastically improved runtime was the motivation behind Inferno. I laid out a six step procedure to estimate the parameters of Inferno from historical ILI data, greatly reducing the MCMC computations as executed by the probabilistic programming language JAGS. Furthermore, by forecasting each geography separately, Inferno can take advantage of parallelization, both improving forecast runtimes in the present while being scalable and well-positioned for the more spatially granular future of flu forecasting (e.g., county-level forecasting).

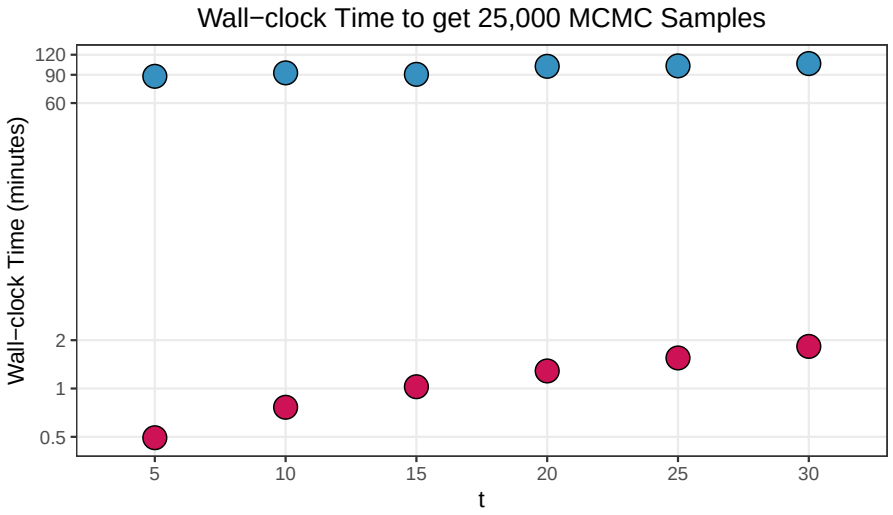


Figure 11: The average wall-clock runtime for a geography (e.g., a state) of Inferno (red) and actual wall-clock runtime for Dante (blue) to get 25,000 MCMC samples. Runtime increases as the size of the conditioning data increases for both Dante and Inferno. Inferno runs in 30 seconds early in the season (t=5) and takes approximately 2 minutes by the end of the flu season (t=30). In contrast, Dante takes approximately 90 minutes to run at the beginning of the flu season and 105 minutes by the end as a result of it conditioning on a much larger set of data.

Inferno's predictive performance was comparable to but worse than Dante's. This may be for a couple different reasons, both of which are addressable. Firstly, Dante explicitly models backfill; previous work has shown that accounting for and modeling backfill can result in improved predictive performance [4, 11]. Similar modeling can be incorporated into Inferno at little additional computational cost. Secondly, Dante achieves self-consistency in its forecasts by modeling and forecasting all nested geographical units jointly. This self-consistency comes at a computational cost. The price Inferno pays to achieve significant computational speed-ups is the loss of self-consistency. There has been some recent work that takes independently generated probabilistic forecasts and, using principles of coherence, produces self-consistent forecasts that have improved predictive performance [8]. The combination of backfill modeling and coherence exploitation may result in equal or even better predictive performance at minimal computational cost.

Inferno's development was motivated by the desire to build a forecasting model that maintains world-leading predictive performance while improving forecasting model properties not directly evaluated by FluSight. It is relatively straightforward to list characteristics we desire in an infectious disease forecasting model that extend well beyond predictive performance. For instance, we want forecasting models to

- reliably and accurately forecast public health relevant targets with actionable lead times
- quantify their forecast uncertainties

- incorporate public health interventions, facilitating "what-if" scenario assessments
- be transferrable between disease contexts and geographies
- run at all spatial scales
- be nimble and adaptable to ever-changing forecast settings
- run quickly, facilitating fast development and testing.

No forecasting model in existence today does all of these things well, but developing such a model should be the goal.

Forecasting challenges have proven to be highly effective organizing tools to help focus forecasters around a common goal, providing real value to public health responses. The singular goal of forecasters to maximize their predictive score is both a blessing and a curse. The predictive score helps generate competition which drives innovation and improvement. It also puts up blinders to all other characteristics we want forecasting models to have. A big question the forecasting community must address going forward is how forecasting challenges can be modified to expand their definition of what a "good" forecasting model is. That is, should forecasting challenges explicitly incorporate aspects of forecasting model generalizability, interpretability, utility, scalability, and speed and if so, how? Or are forecasting challenges not meant to assess forecasting models holistically but rather only assess one specific aspect of those models — their predictive performance?

Acknowledgements

The author is thankful to C.C. Essix for her support and encouragement. This work was developed with the support of LANL's LDRD research Project 20190546ECR. The author is thankful to the U.S. CDC FluSight team for making historical forecast submissions publicly available. Approved for unlimited release under LA-UR-20-30384.

Appendix A

The following inputs are supplied to the JAGS model that implements Inferno:

- `T` is 35, the number of weeks of the flu season.
- `y` is a 35×1 vector where `y[t]` is the observed (w)ILI value for week `t` if it has been observed or `NA` if it has not. If `y[t]` is less than 0.0005 or greater than 0.9995, `y[t]` is set equal to 0.0005 or 0.9995, respectively.
- `alpha` is $\hat{\alpha}$, computed from Equation 14.
- `gamma` is a 35×1 vector where `gamma[t]` is $\hat{\gamma}_t$, computed from Equation 19.
- `invCholUpper` is the inverse of the upper triangular Cholesky decomposition of $\hat{\mathbf{\Sigma}}^{-1}$, where $\hat{\mathbf{\Sigma}}$ is computed from Equation 32 and 33.
- `sigma_mu` is the square root of $\hat{\sigma}^2_\mu$, computed from Equation 24.

The JAGS code implementing Inferno is as follows:

```
model{
  for(t in 1:T){
    ## draw from posterior predictive distribution
    ypred[t] ~ dbeta(alpha*theta[t], alpha*(1-theta[t]))
    ## data model
    y[t] ~ dbeta(alpha*theta[t], alpha*(1-theta[t]))
    ## compute theta
    theta[t] <- ilogit(gamma[t] + delta[t])
  }
  ## discrepancy GP model
  delta[1:T] <- mu + invCholUpper %*% Z[1:T]
  ## sample standard normals
  for(t in 1:T){
    Z[t] ~ dnorm(0,1)
  }
  ## discrepancy mean model
  mu ~ dnorm(0,pow(sigma_mu,-2))
}
```

References

[1] M. Ben-Nun, P. Riley, J. Turtle, D. P. Bacon, and S. Riley. Forecasting National and Regional Influenza-like Illness for the USA. *PLoS Computational Biology*, 15(5):e1007013, 2019.

[2] J. Bracher. On the multibin logarithmic score used in the FluSight competitions. *Proceedings of the National Academy of Sciences*, 116(42):20809–20810, 2019.

[3] L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, and R. Rosenfeld. Flexible Modeling of Epidemics with an Empirical Bayes Framework. *PLoS Computational Biology*, 11(8):e1004382, 2015.

[4] L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, and R. Rosenfeld. Nonmechanistic Forecasts of Seasonal Influenza with Iterative One-week-ahead Distributions. *PLoS Computational Biology*, 14(6):e1006134, 2018.

[5] S. Y. Del Valle, B. H. McMahon, J. Asher, R. Hatchett, J. C. Lega, H. E. Brown, M. E. Leany, Y. Pantazis, D. J. Roberts, S. Moore, et al. Summary Results of the 2014-2015 DARPA Chikungunya Challenge. *BMC Infectious Diseases*, 18(1):245, 2018.

[6] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020.

[7] D. C. Farrow, L. C. Brooks, A. Rumack, R. J. Tibshirani, and R. Rosenfeld. Delphi Epidata API, 2015. Available at `https://github.com/cmu-delphi/delphi-epidata`.

[8] G. Gibson, K. Moran, N. Reich, and D. Osthus. Improving Probabilistic Infectious Disease Forecasting Through Coherence. *bioRxiv*, 2019.

[9] M. A. Johansson, K. M. Apfeldorf, S. Dobson, J. Devita, A. L. Buczak, B. Baugher, L. J. Moniz, T. Bagley, S. M. Babin, E. Guven, T. K. Yamana, J. Shaman, T. Moschou, N. Lothian, A. Lane, G. Osborne, G. Jiang, L. C. Brooks, D. C. Farrow, S. Hyun, R. J. Tibshirani, R. Rosenfeld, J. Lessler, N. G. Reich, D. A. T. Cummings, S. A. Lauer, S. M. Moore, H. E. Clapham, R. Lowe, T. C. Bailey, M. García-Díez, M. S. Carvalho, X. Rodó, T. Sardar, R. Paul, E. L. Ray, K. Sakrejda, A. C. Brown, X. Meng, O. Osoba, R. Vardavas, D. Manheim, M. Moore, D. M. Rao, T. C. Porco, S. Ackley, F. Liu, L. Worden, M. Convertino, Y. Liu, A. Reddy, E. Ortiz, J. Rivero, H. Brito, A. Juarrero, L. R. Johnson, R. B. Gramacy, J. M. Cohen, E. A. Mordecai, C. C. Murdock, J. R. Rohr, S. J. Ryan, A. M. Stewart-Ibarra, D. P. Weikel, A. Jutla, R. Khan, M. Poultney, R. R. Colwell, B. Rivera-García, C. M. Barker, J. E. Bell, M. Biggerstaff, D. Swerdlow, L. Mier-y Teran-Romero, B. M. Forshey, J. Trtanj, J. Asher, M. Clay, H. S. Margolis, A. M. Hebbeler, D. George, and J.-P. Chretien. An Open Challenge to Advance Probabilistic Forecasting for Dengue Epidemics. *Proceedings of the National Academy of Sciences*, 116(48):24268–24274, 2019.

[10] S. Kandula and J. Shaman. Near-term Forecasts of Influenza-like Illness: An Evaluation of Autoregressive Time Series Approaches. *Epidemics*, 27:41–51, 2019.

[11] D. Osthus, A. R. Daughton, and R. Priedhorsky. Even a Good Influenza Forecasting Model can Benefit from Internet-based Nowcasts, but those Benefits are Limited. *PLoS Computational Biology*, 15(2):e1006599, 2019.

[12] D. Osthus, J. Gattiker, R. Priedhorsky, and S. Y. Del Valle. Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy (with Discussion). *Bayesian Analysis*, 14(1):261–312, 2019.

[13] D. Osthus, K. S. Hickmann, P. C. Caragea, D. Higdon, and S. Y. Del Valle. Forecasting Seasonal Influenza with a State-space SIR Model. *The Annals of Applied Statistics*, 11(1):202, 2017.

[14] D. Osthus and K. R. Moran. Multiscale Influenza Forecasting. *arXiv preprint arXiv:1909.13766*, 2019.

[15] S. Pei, S. Kandula, W. Yang, and J. Shaman. Forecasting the Spatial Transmission of Influenza in the United States. *Proceedings of the National Academy of Sciences*, 115(11):2752–2757, 2018.

[16] M. Plummer. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 10. Vienna, Austria., 2003.

[17] M. Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2018. R package version 4-8.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[19] E. L. Ray and N. G. Reich. Prediction of Infectious Disease Epidemics via Weighted Density Ensembles. *PLoS Computational Biology*, 14(2):e1005910, 2018.

[20] N. G. Reich, L. C. Brooks, S. J. Fox, S. Kandula, C. J. McGowan, E. Moore, D. Osthus, E. L. Ray, A. Tushar, T. K. Yamana, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, and J. Shaman. A Collaborative Multiyear, Multimodel Assessment of Seasonal Influenza Forecasting in the United States. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019.

[21] N. G. Reich, D. Osthus, E. L. Ray, T. K. Yamana, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, and J. Shaman. Reply to Bracher: Scoring Probabilistic Forecasts to Maximize Public Health Interpretability. *Proceedings of the National Academy of Sciences*, 116(42), 9 2019.

[22] The Epidemic Prediction Initiative. FluSight 2018–2019 Guidance Documents, September 2018. Available at `https://predict.cdc.gov/post/5ba1504e5619f003acb7e18f`.

[23] The Epidemic Prediction Initiative. Aedes Forecasting 2019, February 2019. Available at `https://predict.cdc.gov/post/5c4f6d687620e103b6dcd015`.

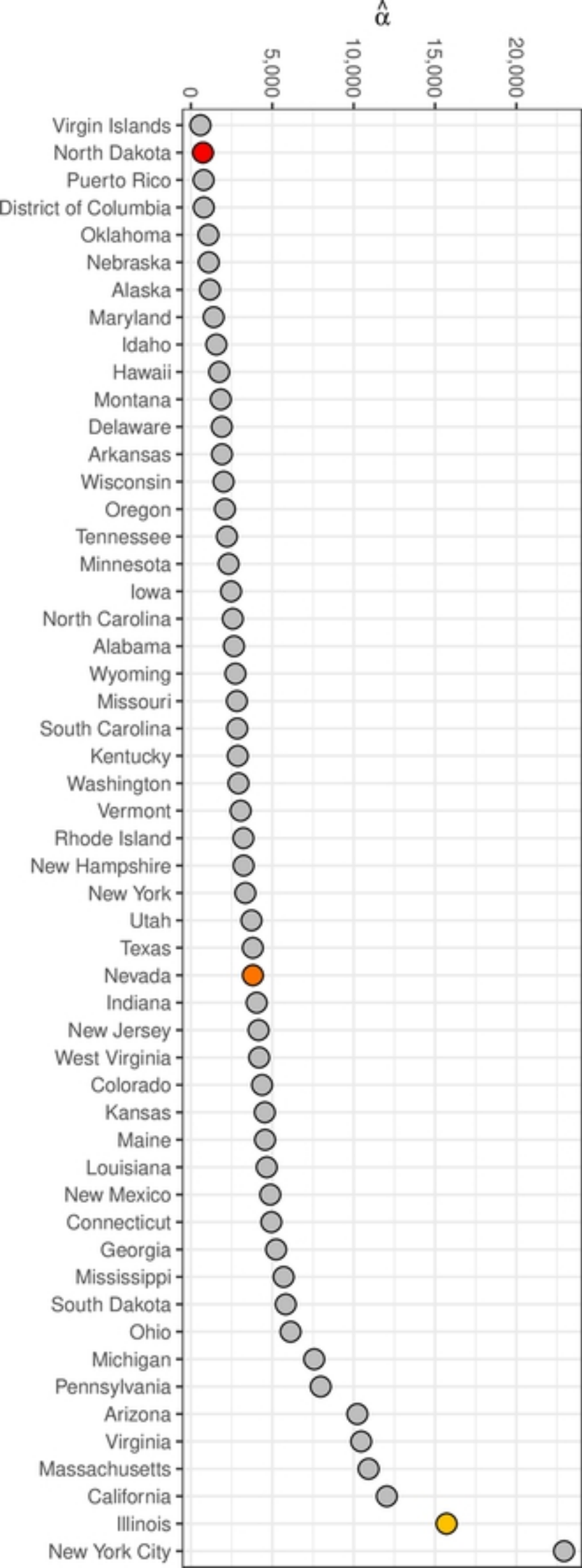
[24] The Epidemic Prediction Initiative. West Nile Virus Forecasting 2020, January 2020. Available at `https://predict.cdc.gov/post/5e18a08677851c0489cf10b8`.

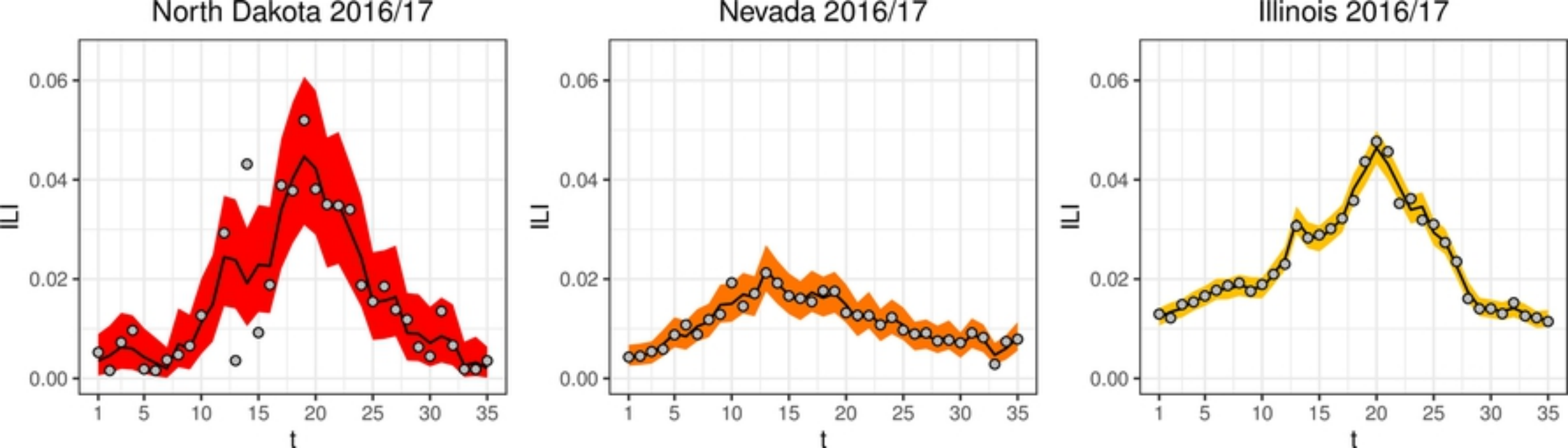
[25] The United States Centers for Disease Control and Prevention. Disease Burden of Influenza, December 2020. Available at `https://www.cdc.gov/flu/about/burden/index.html`.

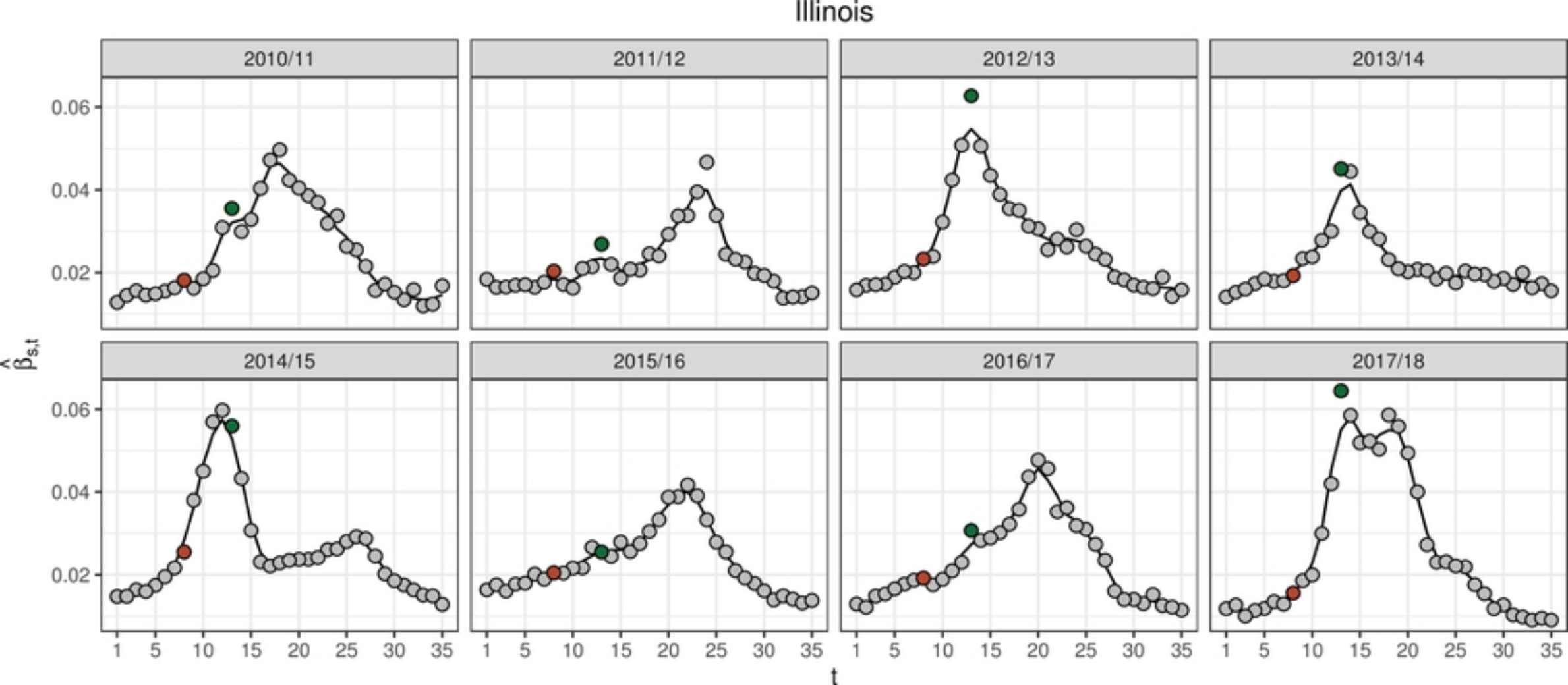
[26] S. Venkatramanan, B. Lewis, J. Chen, D. Higdon, A. Vullikanti, and M. Marathe. Using Data-driven Agent-based Models for Forecasting Emerging Infectious Diseases. *Epidemics*, 22:43–49, 2018.

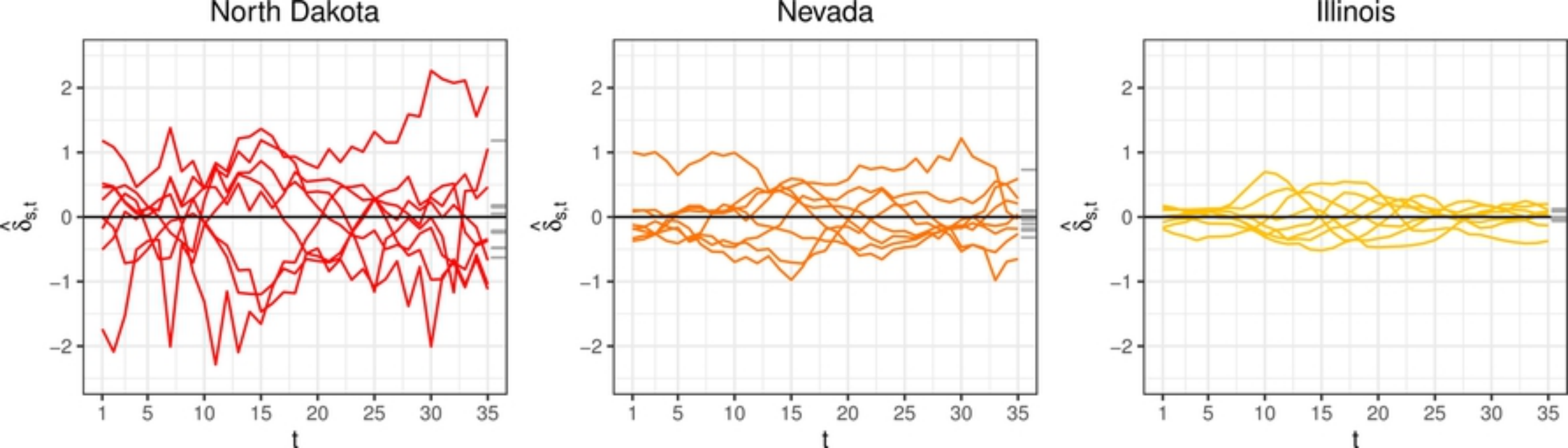
[27] N. Wattanachit, S. Wang, N. Reich, E. Ray, J. Niemi, K. Le, A. Khandelwal, A. H. Kanji, K. House, E. Cramer, M. Cornell, A. Brennen, and J. Bracher. COVID-19 Forecast Hub. Nov. 2020. Available at `https://github.com/reichlab/covid19-forecast-hub`.

[28] S. Yang, M. Santillana, and S. C. Kou. Accurate Estimation of Influenza Epidemics using Google Search Data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.

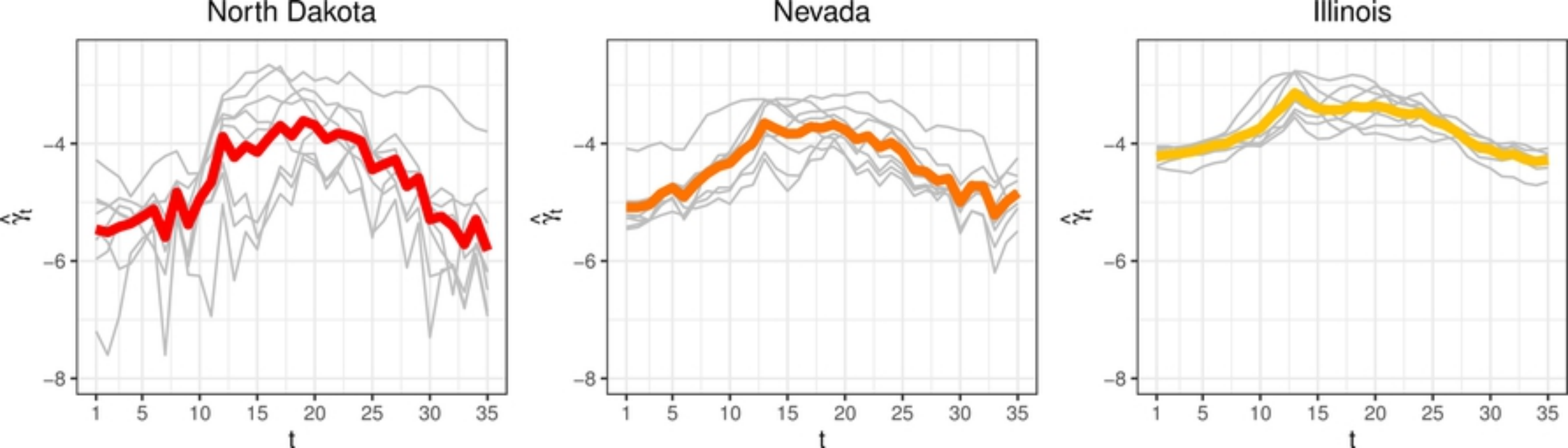

$\hat{\alpha}$
0
5,000
10,000
15,000
20,000
Virgin Islands
North Dakota
Puerto Rico
District of Columbia
Oklahoma
Nebraska
Alaska
Maryland
Idaho
Hawaii
Montana
Delaware
Arkansas
Wisconsin
Oregon
Tennessee
Minnesota
Iowa
North Carolina
Alabama
Wyoming
Missouri
South Carolina
Kentucky
Washington
Vermont
Rhode Island
New Hampshire
New York
Utah
Texas
Nevada
Indiana
New Jersey
West Virginia
Colorado
Kansas
Maine
Louisiana
New Mexico
Connecticut
Georgia
Mississippi
South Dakota
Ohio
Michigan
Pennsylvania
Arizona
Virginia
Massachusetts
California
Illinois
New York City

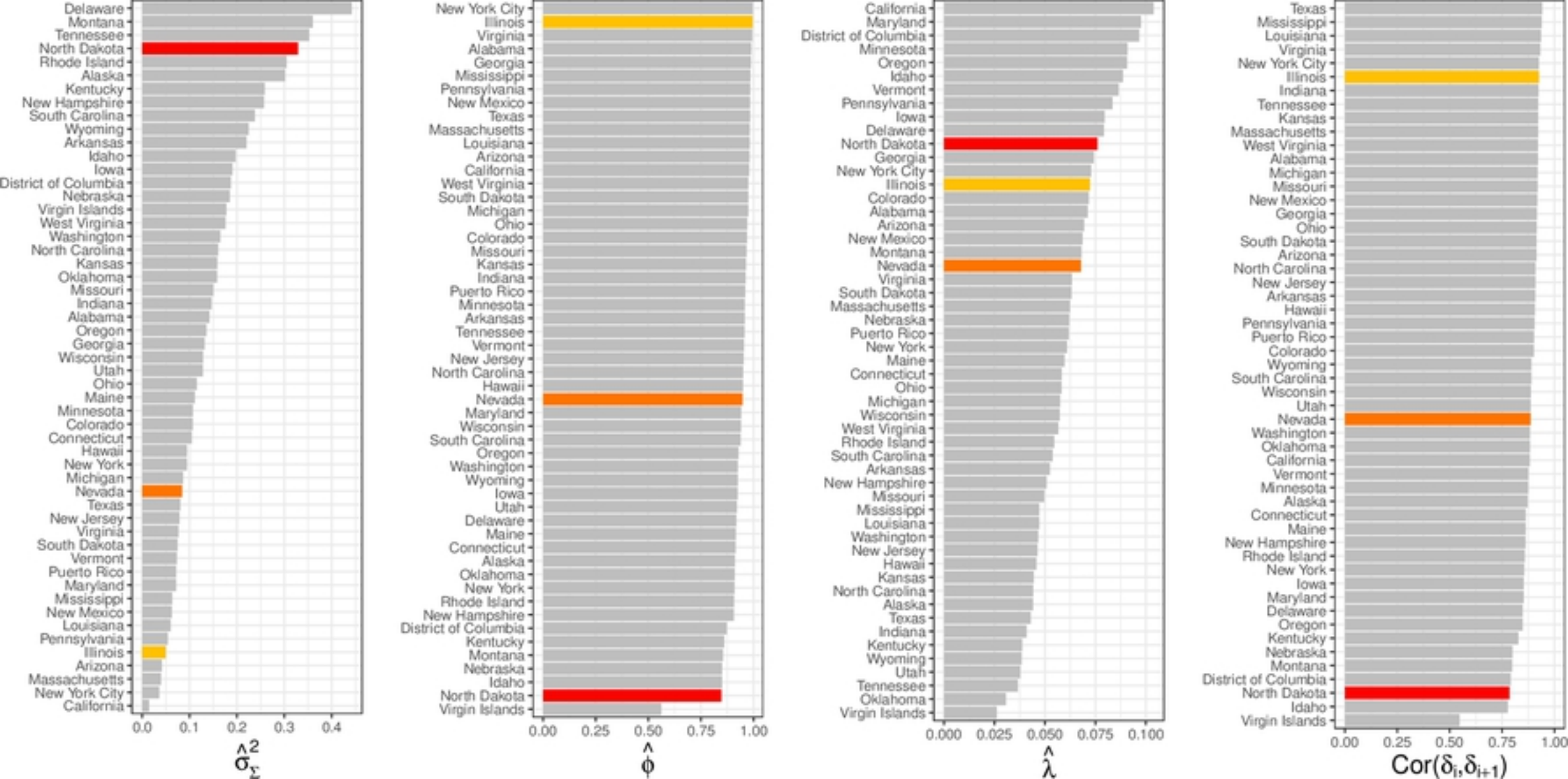

North Dakota 2016/17
Nevada 2016/17
Illinois 2016/17
ILI
t
0.00
0.02
0.04
0.06
1
5
10
15
20
25
30
35

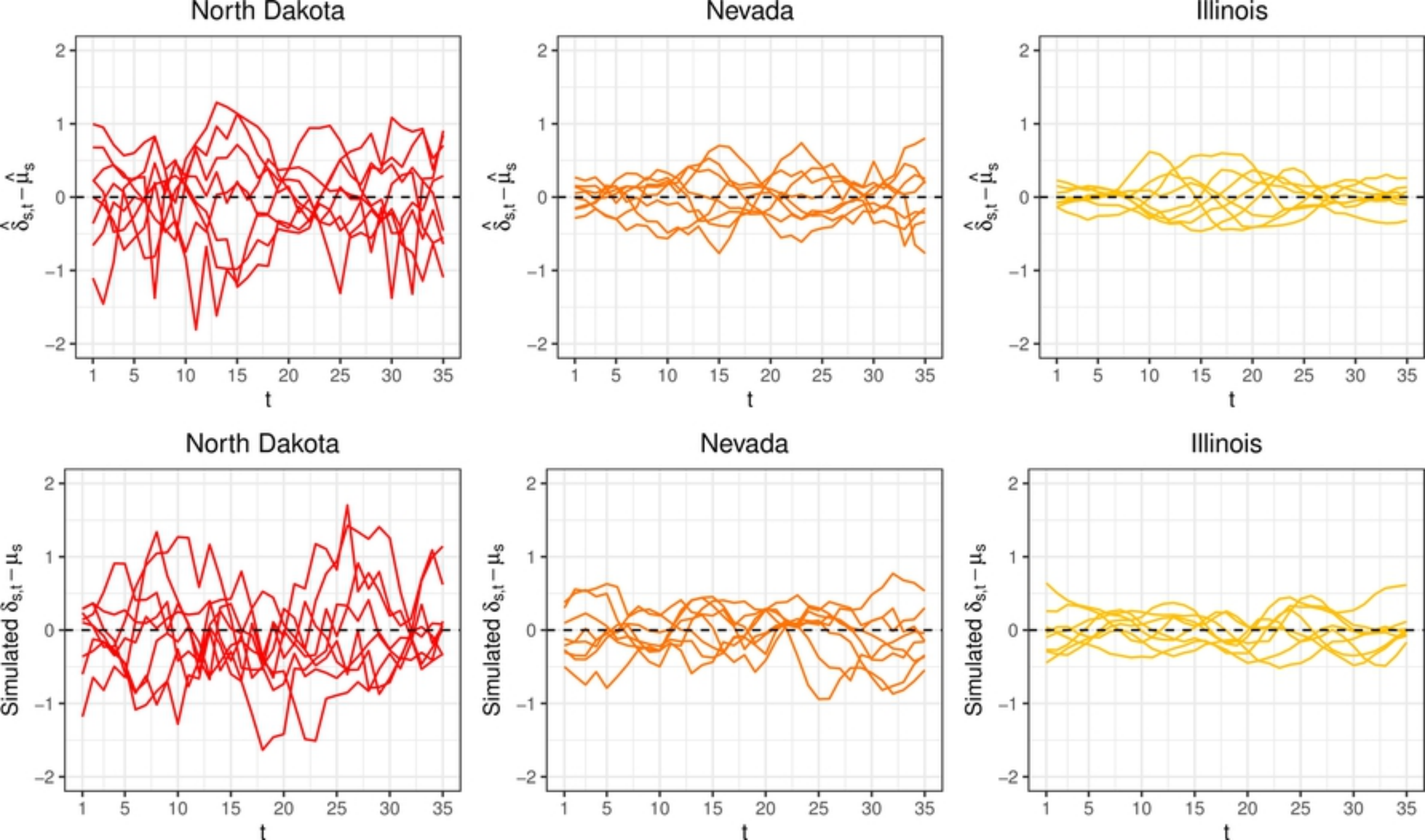

Illinois
2010/11
2011/12
2012/13
2013/14
2014/15
2015/16
2016/17
2017/18
0.06
0.04
0.02
$\hat{\beta}_{s,t}$
1 5 10 15 20 25 30 35
t

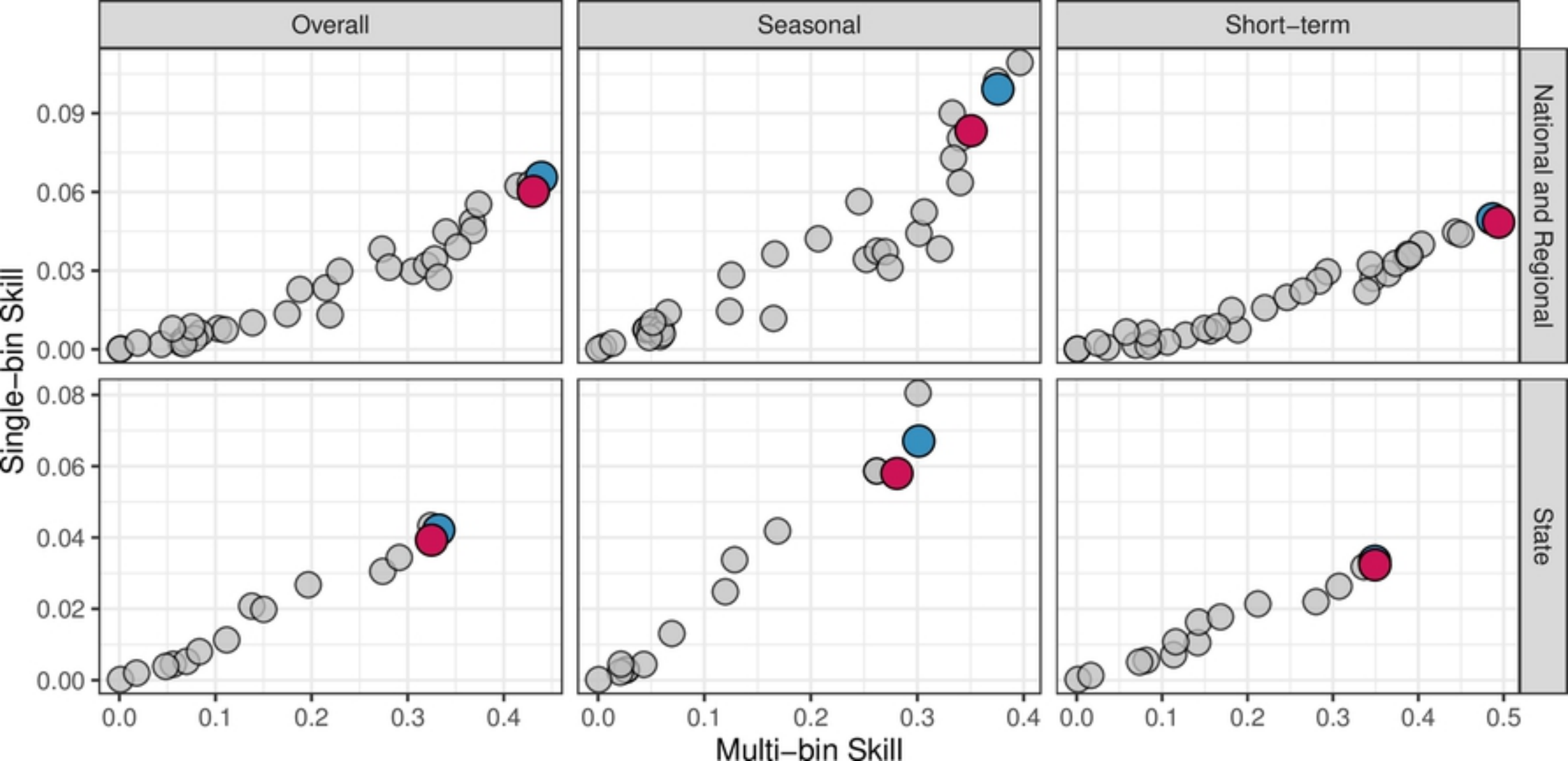

North Dakota
Nevada
Illinois
$\hat{\delta}_{s,t}$
t
2
1
0
−1
−2
1
5
10
15
20
25
30
35

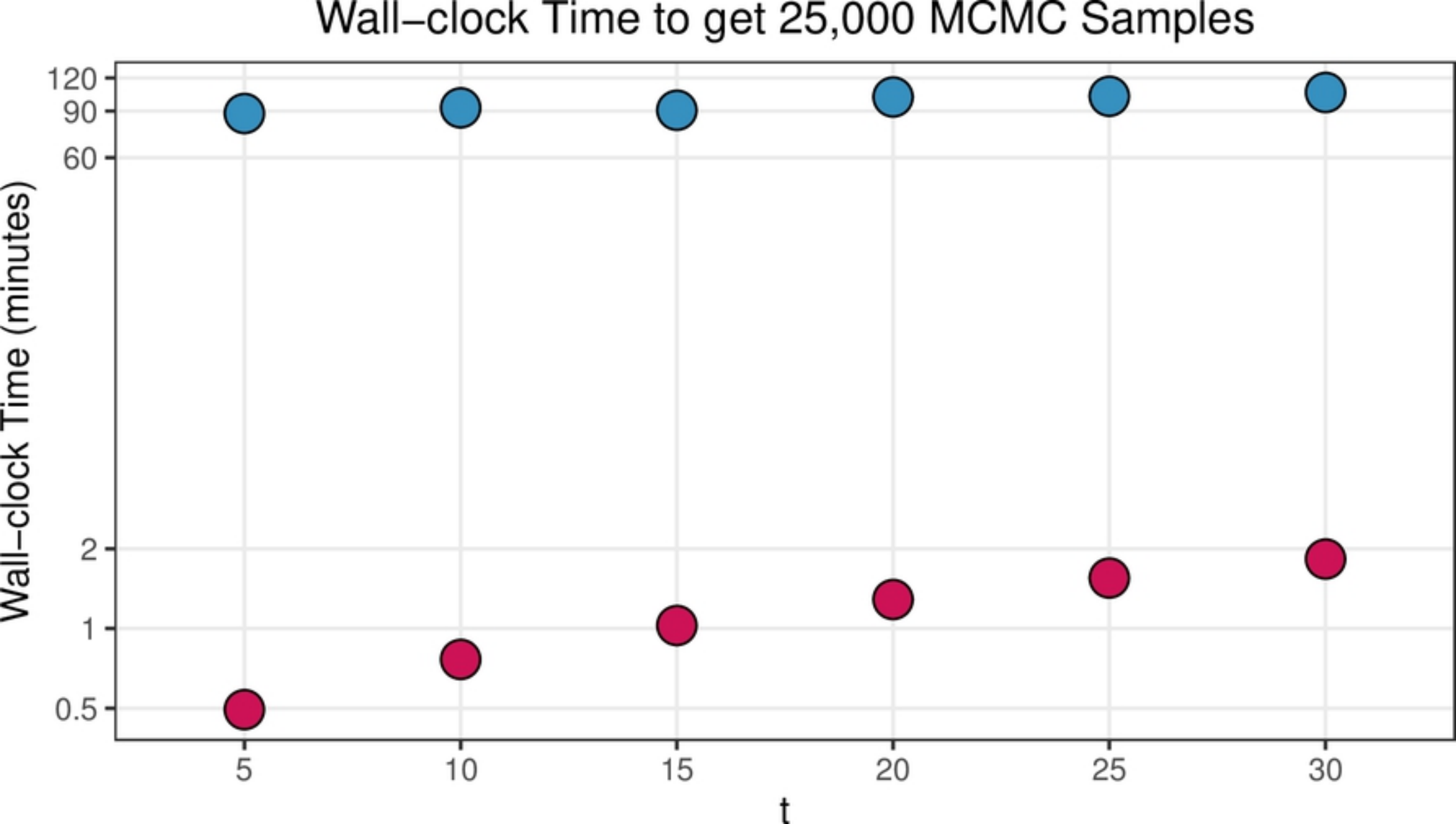
North Dakota
Nevada
Illinois
5
10
15
20
25
30
ILI
t

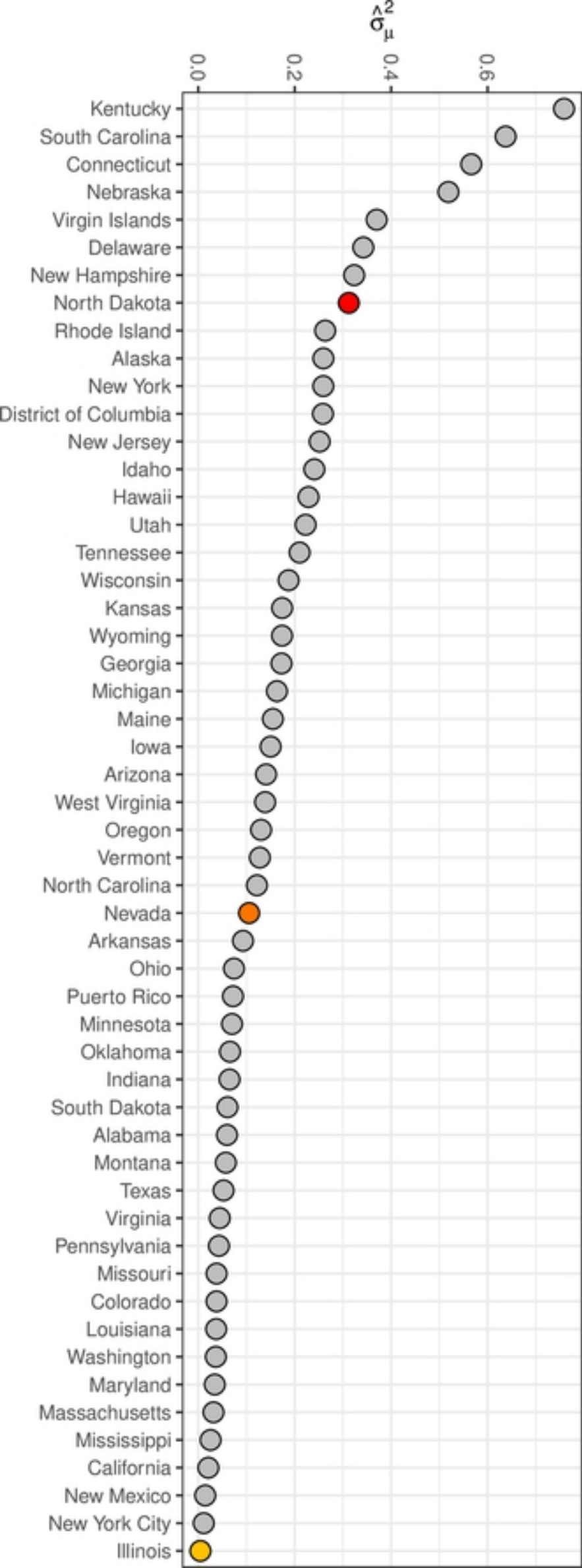

North Dakota
Nevada
Illinois
$\hat{\gamma}_t$
−4
−6
−8
1
5
10
15
20
25
30
35
t

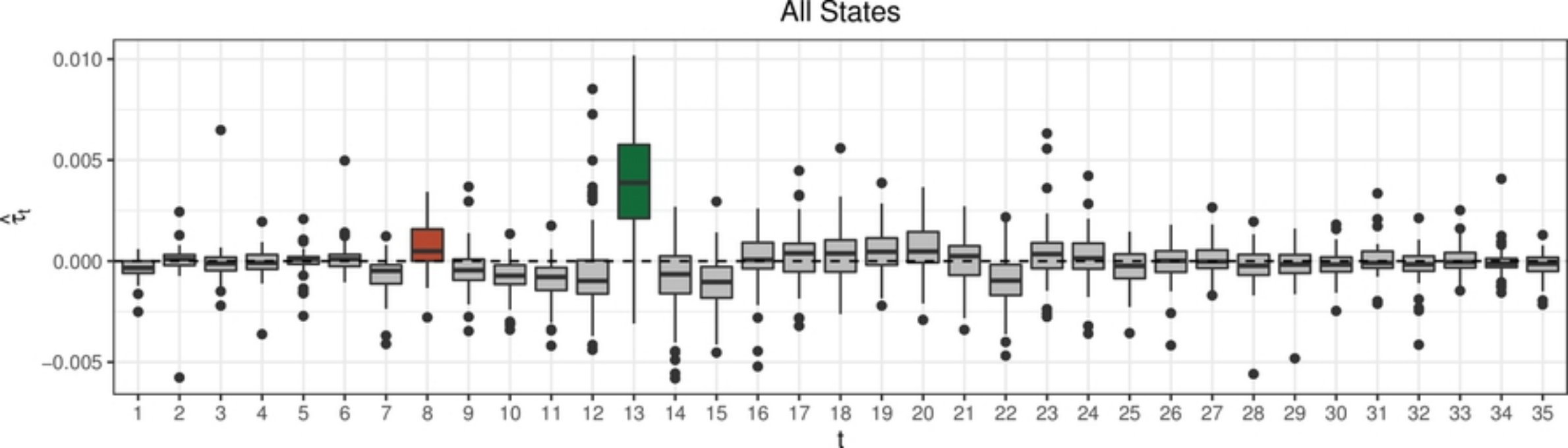

Delaware
Montana
Tennessee
North Dakota
Rhode Island
Alaska
Kentucky
New Hampshire
South Carolina
Wyoming
Arkansas
Idaho
Iowa
District of Columbia
Nebraska
Virgin Islands
West Virginia
Washington
North Carolina
Kansas
Oklahoma
Missouri
Indiana
Alabama
Oregon
Georgia
Wisconsin
Utah
Ohio
Maine
Minnesota
Colorado
Connecticut
Hawaii
New York
Michigan
Nevada
Texas
New Jersey
Virginia
South Dakota
Vermont
Puerto Rico
Maryland
Mississippi
New Mexico
Louisiana
Pennsylvania
Illinois
Arizona
Massachusetts
New York City
California
0.0
0.1
0.2
0.3
0.4
$\hat{\sigma}^2_\Sigma$
New York City
Illinois
Virginia
Alabama
Georgia
Mississippi
Pennsylvania
New Mexico
Texas
Massachusetts
Louisiana
Arizona
California
West Virginia
South Dakota
Michigan
Ohio
Colorado
Missouri
Kansas
Indiana
Puerto Rico
Minnesota
Arkansas
Tennessee
Vermont
New Jersey
North Carolina
Hawaii
Nevada
Maryland
Wisconsin
South Carolina
Oregon
Washington
Wyoming
Iowa
Utah
Delaware
Maine
Connecticut
Alaska
Oklahoma
New York
Rhode Island
New Hampshire
District of Columbia
Kentucky
Montana
Nebraska
Idaho
North Dakota
Virgin Islands
0.00
0.25
0.50
0.75
1.00
$\hat{\phi}$
California
Maryland
District of Columbia
Minnesota
Oregon
Idaho
Vermont
Pennsylvania
Iowa
Delaware
North Dakota
Georgia
New York City
Illinois
Colorado
Alabama
Arizona
New Mexico
Montana
Nevada
Virginia
South Dakota
Massachusetts
Nebraska
Puerto Rico
New York
Maine
Connecticut
Ohio
Michigan
Wisconsin
West Virginia
Rhode Island
South Carolina
Arkansas
New Hampshire
Missouri
Mississippi
Louisiana
Washington
New Jersey
Hawaii
Kansas
North Carolina
Alaska
Texas
Indiana
Kentucky
Wyoming
Utah
Tennessee
Oklahoma
Virgin Islands
0.000
0.025
0.050
0.075
0.100
$\hat{\lambda}$
Texas
Mississippi
Louisiana
Virginia
New York City
Illinois
Indiana
Tennessee
Kansas
Massachusetts
West Virginia
Alabama
Michigan
Missouri
New Mexico
Georgia
Ohio
South Dakota
Arizona
North Carolina
New Jersey
Arkansas
Hawaii
Pennsylvania
Puerto Rico
Colorado
Wyoming
South Carolina
Wisconsin
Utah
Nevada
Washington
Oklahoma
California
Vermont
Minnesota
Alaska
Connecticut
Maine
New Hampshire
Rhode Island
New York
Iowa
Maryland
Delaware
Oregon
Kentucky
Nebraska
Montana
District of Columbia
North Dakota
Idaho
Virgin Islands
0.00
0.25
0.50
0.75
1.00
$Cor(\delta_i,\delta_{i+1})$

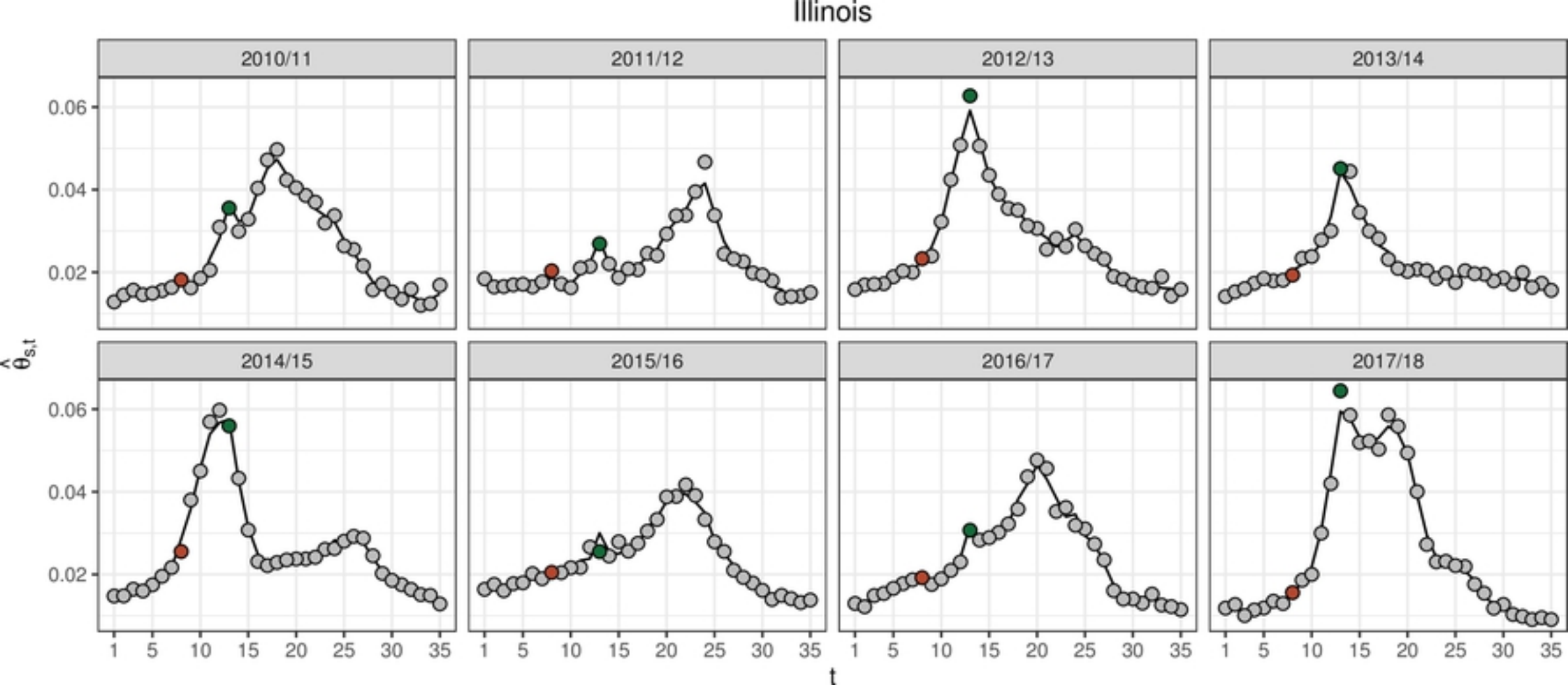

North Dakota
Nevada
Illinois
$\hat{\delta}_{s,t} - \hat{\mu}_s$
t
North Dakota
Nevada
Illinois
Simulated $\delta_{s,t} - \mu_s$
t


Overall
Seasonal
Short−term
National and Regional
State
Single−bin Skill
Multi−bin Skill
0.00
0.03
0.06
0.09
0.02
0.04
0.08
0.0
0.1
0.2
0.3
0.4
0.5


Wall−clock Time to get 25,000 MCMC Samples
Wall−clock Time (minutes)
120
90
60
2
1
0.5
5
10
15
20
25
30
t


$\hat{\sigma}^2_\mu$
0.0
0.2
0.4
0.6
Kentucky
South Carolina
Connecticut
Nebraska
Virgin Islands
Delaware
New Hampshire
North Dakota
Rhode Island
Alaska
New York
District of Columbia
New Jersey
Idaho
Hawaii
Utah
Tennessee
Wisconsin
Kansas
Wyoming
Georgia
Michigan
Maine
Iowa
Arizona
West Virginia
Oregon
Vermont
North Carolina
Nevada
Arkansas
Ohio
Puerto Rico
Minnesota
Oklahoma
Indiana
South Dakota
Alabama
Montana
Texas
Virginia
Pennsylvania
Missouri
Colorado
Louisiana
Washington
Maryland
Massachusetts
Mississippi
California
New Mexico
New York City
Illinois


All States
0.010
0.005
0.000
−0.005
$\hat{\tau}_t$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
t


Illinois
2010/11
2011/12
2012/13
2013/14
2014/15
2015/16
2016/17
2017/18
0.06
0.04
0.02
$\hat{\theta}_{s,t}$
1 5 10 15 20 25 30 35
t