The Annals of Applied Statistics
2022, Vol. 16, No. 1, 300–325
https://doi.org/10.1214/21-AOAS1505

SEQUENTIAL MODELING, MONITORING, AND FORECASTING OF STREAMING WEB TRAFFIC DATA

BY KAORU IRIE[1,a], CHRIS GLYNN[2,b] AND TEVFIK AKTEKIN[3,c]

[1]*Faculty of Economics, University of Tokyo,* [a]*irie@e.u-tokyo.ac.jp*
[2]*Economic Research, Zillow Group,* [b]*christophergl@zillowgroup.com*
[3]*Paul College of Business and Economics, University of New Hampshire,* [c]*tevfik.aktekin@unh.edu*

In this paper we introduce strategies for modeling, monitoring, and forecasting sequential web traffic data using flows from the Fox News website. In our analysis we consider a family of Poisson-gamma state space (PGSS) models that can accurately quantify the uncertainty exhibited by web traffic data, can provide fast sequential monitoring and prediction mechanisms for high frequency time intervals, and are computationally feasible when structural breaks are present. As such, we extend the family of PGSS models to include the state augmented (sa-)PGSS model whose state evolution structure is flexible and responsive to sudden changes. Such adaptability is achieved by augmenting the state vector of the PGSS model with an additional state variable for a time-varying discount factor. We develop an efficient particle-based estimation procedure that is suitable for sequential analysis, allowing us to estimate dynamic state variables and static parameters via closed-form conditional sufficient statistics. We compare the performance of the PGSS family of models against viable alternatives from the literature and argue that, especially in the presence of structural breaks, our proposed approach yields superior sequential model fit and predictive performance while preserving computational feasibility. We provide additional insights by designing a simulation study that mimics potential web traffic data patterns.

1. Introduction. Web traffic data is a crucial component of many modern applications, such as web analytics, consumer analytics, and network analysis, among others. Web traffic data, generated by the visitors to a website, are often used as a performance metric and an input for decision making by online retailers and various e-commerce businesses. Worldwide e-commerce sales is a significant portion of the global economy and is predicted to reach nearly $3.46 trillion in 2019, up from $2.93 trillion in 2018.[1] Web traffic data is typically defined by the number of visitors (or clicks) a site receives or the number of pages a visitor browses within a given website. Many e-commerce businesses rely on accurate and fast prediction of streaming web traffic data in high frequency time intervals to continually allocate resources. Google, Yahoo, and various online news sites are often interested in modeling and predicting traffic flows among webpages that are in turn used as inputs for selling and pricing of online advertising campaigns. Online retailers, such as Amazon, Walmart, and Target, use web traffic data as part of recommender systems and search engine optimization schemes to better understand consumer behavior and purchasing trends.

The analysis of streaming web traffic data poses a number of statistical modeling challenges. In our view the most prominent challenges are threefold: (i) Adequately representing the uncertainty characteristics typically observed in web traffic data, (ii) Developing fast and efficient learning and prediction mechanisms suitable for sequential analysis in high frequency time intervals, and (iii) Designing models that can rapidly adapt to sudden structural

Received March 2019; revised April 2021.
Key words and phrases. Web traffic, count data, high frequency, Poisson-gamma, sequential Monte Carlo.
[1]https://www.digitalcommerce360.com/article/global-ecommerce-sales/

changes in web traffic while preserving computational feasibility. In many e-commerce settings, resource allocation decisions are made on the order of seconds to minutes in a sequential manner where balancing model complexity and computational speed is imperative. When business operations of these web platforms depend on accurate short-term predictions of consumer demand (measured in counts of visitors or clicks), the ability to quickly identify structural breaks and adjust forecasts of customer counts is critical. Markov switching models have traditionally been used to model regime changes in time series data; however, on e-commerce platforms, consumer demand changes rapidly, and resources are reallocated frequently. In high-frequency applications the computational cost of fitting Markov switching models is prohibitive, as they are not suitable for designing fast learning and forecasting mechanisms in a sequential manner. The computational cost is amplified significantly when switching points are unknown, as is the case in the analysis of web traffic data. In this paper we analyze various streams of web traffic from the Fox News website in high-frequency time intervals and develop computationally competitive Bayesian state space models that can adequately represent characteristics of such high-frequency, bursty web traffic data. Specifically, we develop the state-augmented Poisson-gamma state space (sa-PGSS) model, an integer-valued state-space model whose structure flexibly adapts to newly observed counts and admits a sequential Monte Carlo algorithm for online updates of posterior and one-step-ahead predictive distributions. It advances the literature on Poisson-gamma state-space (PGSS) models by introducing a mechanism to sequentially adapt model structure, as called for by data. To achieve this, we augment the state-variable in the PGSS model with a dynamic discount factor that enables rapid model adaption to structural changes in observed counts. The methodological novelty of our approach stems from this state variable augmentation which increases the flexibility of the PGSS model while allowing us to develop a fast estimation algorithm suitable for sequential parameter learning, system monitoring, and demand forecasting.

The remainder of our paper is structured as follows. In Section 2 we introduce the details of the data. In Section 3 we summarize the PGSS model and its properties. We illustrate the inability of the base PGSS model to rapidly adapt to structural breaks. In Section 4 we introduce the sa-PGSS model and in Section 5 develop its particle-based algorithm. Section 6 discusses the numerical analysis of two web traffic streams from Fox News website and a simulated study. Section 7 concludes with a summary and a discussion on future directions.

2. Web traffic data: Fox news streams. To investigate the characteristics of web traffic data in a setting where fast sequential online learning, monitoring, and prediction are essential, we consider observations from the Fox News website. The data itself was obtained from the raw access log of the Fox News website, which is a collection of individual URL access logs (date and time), and is the flow (number of accesses) from one category of news articles to another. The counts are observed at 30 second intervals sequentially which precludes the use of Markov chain Monte Carlo methods necessary for implementing many Markov switching models. For our investigation we consider two flows within the Fox News website, a high-count and a low-count flow, each aimed at addressing different characteristics of web traffic data. The first flow we considered was from the top (main) page of the website to the category titled "World" between 9:05 and 9:55 a.m. on February 23, 2015. The first observation at 9:05 is omitted from the series. The total length of the time series is $T = 99$. In this illustration the web traffic counts range between approximately 70 and 270, thus is a suitable example of a stream with relatively high counts. The second example involves flows between pages that are visited less frequently, which we refer to as the low count stream. These include flows between the pages titled "Politics" and "Leisure" recorded between 13:05 and 13:33 p.m. on March 2, 2015. In this example, counts generally vary between zero and 10. Our eventual goal is to show how we can build sequential models and estimation methods

for either one of these flows 30 seconds in advance, allowing advertising impressions to be optimally allocated across sections. Robust forecasting and real-time monitoring systems of web traffic are of great interest to many e-commerce firms, as optimal online ad placement and efficient web-server maintenance are top priorities.

3. Poisson-gamma state space (PGSS) model. Forecasting count data in high-frequency settings that potentially exhibit bursts poses a number of statistical and computational challenges. One such challenge is to flexibly model temporal dependence in a way that facilitates rapid and online estimation of model parameters. There are two main approaches for modeling count time series: The first assumes that time-varying counts are generated by a stationary stochastic process (Freeland and McCabe (2004)); the second approach models temporal dependence via state space models and allows for the possibility that counts are nonstationary (Aktekin, Polson and Soyer (2018), Berry and West (2020), Chen, Banks and West (2019), Chen et al. (2018), Frühwirth-Schnatter and Wagner (2006), Gamerman, Rezende dos Santos and Franco (2013), Harvey and Fernandes (1989), Glynn et al. (2019)). The state-space approach exploits the conditional independence of counts given that state parameters themselves follow a stochastic process, inducing temporal dependence in counts marginally; see Prado and West (2010) and Davis et al. (2015) for recent reviews of state space models and time series of counts.

The PGSS model (Aktekin and Soyer (2011), Aktekin, Soyer and Xu (2013), Chen et al. (2018)) is a popular choice for modeling time-varying count data, since the Poisson-gamma conjugacy admits online, closed-form calculation of posterior and forecast distributions. The PGSS model is one in a broader class of gamma-beta random walk models for Poisson rates. The gamma-beta state transition was first introduced by Smith and Miller (1986) for state space models with exponential likelihoods and was later utilized to model stochastic volatility in financial markets by Uhlig (1994), Uhlig (1997). Recently, the same state transition structure has been used to model a general class of non-Gaussian state space models (Gamerman, Rezende dos Santos and Franco (2013)). One attractive feature common to gamma-beta random walk models is that the beta-distributed innovations in the state equation yield a state variable that is marginally gamma distributed (assuming that the initial state prior is also gamma distributed), leading to closed-form updates of posterior and forecast distributions in the PGSS model. While online, analytically available posterior and predictive distributions are attractive features, the single-process PGSS model is unable to capture sudden bursts or regime switches in counts. The lack of flexibility in the PGSS model stems from the static discount parameter used in defining state transitions.

In this section we introduce necessary notation and the conjugacy preliminaries for the standard PGSS model which yields tractable filtering as well as one-step-ahead predictive densities. Let N_t for $t = 1, \ldots, T$ represent a univariate time series of counts and $\mathcal{D}_t = \{N_1, \ldots, N_t\}$ a collection of these counts until time t. The likelihood (observational equation) is defined by the Poisson distribution,

$$(N_t|\theta_t) \sim \text{Po}(\theta_t), \tag{1}$$

where, given θ_t, N_t is assumed to be conditionally independent of N_{t-1}. Temporal dependence of N_t on N_{t-1} is governed by the stochastic evolution of θ_{t-1} to θ_t. The state transition (evolution) equation follows a multiplicative gamma-beta random walk. Conditional on θ_{t-1} and $\mathcal{D}_{t-1}$,

$$\theta_t = \theta_{t-1}\eta_t/\gamma, \qquad \eta_t \sim \text{Beta}\big(\gamma\alpha_{t-1}, (1-\gamma)\alpha_{t-1}\big), \tag{2}$$

which implies a state transition equation given by

$$(\theta_t|\theta_{t-1}, \gamma, \mathcal{D}_{t-1}) \sim \text{ScaledBeta}\big(\gamma\alpha_{t-1}, (1-\gamma)\alpha_{t-1}\big), \tag{3}$$

for $\theta_t \in (0, \theta_{t-1}/\gamma)$, $\alpha_{t-1} > 0$, and $0 < \gamma < 1$. The shape parameter, α_{t-1}, is a function of the past observations $\mathcal{D}_{t-1}$ in general, and its specific functional form is given later. We note here that the state transition density (3) is a function of the past observations, $\mathcal{D}_{t-1}$, unlike traditional linear state space models. Here, γ is referred to as the discount factor and controls the persistence of the state variables. For instance, when $\gamma \uparrow 1$, θ_t and θ_{t-1} will be similar (strong dependence and persistence). Whereas, when $\gamma \downarrow 0$, θ_t and θ_{t-1} will likely be less similar, implying more volatile state dynamics (weak dependence and persistence).

Various versions of the PGSS model have been considered in the literature. Gamerman, Rezende dos Santos and Franco (2013) consider a general class of non-Gaussian state space models where the Poisson sampling model appears as a special case. Aktekin, Soyer and Xu (2013) consider it for modeling mortgage default counts, Chen et al. (2018) utilize it to model web traffic in network flow data, and Aktekin, Polson and Soyer (2018) extend it to account for multivariate time series of counts. Further details of the PGSS model can be found in these papers and the references therein. In what follows, we provide a summary of some of the relevant results of the PGSS model. Given the initial state prior of $\theta_0 \sim \text{Ga}(\alpha_0, \beta_0)$, we can show four key model properties:

i The time $t-1$ posterior distribution $(\theta_{t-1}|\gamma, \mathcal{D}_{t-1})$ is gamma-distributed

$$(\theta_{t-1}|\gamma, \mathcal{D}_{t-1}) \sim \text{Ga}(\alpha_{t-1}, \beta_{t-1}). \tag{4}$$

ii The prior distribution for θ_t is a discounted version of equation (4), inflating the prior variance of θ_t relative to the posterior at $t-1$,

$$(\theta_t|\gamma, \mathcal{D}_{t-1}) \sim \text{Ga}(\gamma\alpha_{t-1}, \gamma\beta_{t-1}). \tag{5}$$

iii The time t posterior is also gamma distributed

$$(\theta_t|\gamma, \mathcal{D}_t) \sim \text{Ga}(\alpha_t, \beta_t), \tag{6}$$

where $\alpha_t = \gamma\alpha_{t-1} + N_t = \sum_{s=0}^{t-1} \gamma^s N_{t-s} + \gamma^t \alpha_0$ and $\beta_t = \gamma\beta_{t-1} + 1 = \frac{1-\gamma^t}{1-\gamma} + \gamma^t \beta_0$. Observe that α_t combines a γ-discounted shape parameter from the posterior of θ_{t-1} and the most recently observed N_t, while β_t increments the γ-discounted rate parameter from the posterior of θ_{t-1} by one to reflect an additional data point.

iv The one-step-ahead predictive distribution is Negative Binomial,

$$(N_t|\gamma, \mathcal{D}_{t-1}) \sim \text{NegBin}\left(\gamma\alpha_{t-1}, \frac{\gamma\beta_{t-1}}{\gamma\beta_{t-1}+1}\right). \tag{7}$$

Conditional on γ, the filtering density $p(\theta_t|\gamma, \mathcal{D}_t)$ and the one-step-ahead predictive density $p(N_t|\gamma, \mathcal{D}_{t-1})$ are available in closed form which makes the PGSS model attractive for practical applications of web traffic count data. Another noteworthy property of the PGSS model is the closed form availability of the marginal likelihood that can be used to estimate static model parameters like γ. Typically these marginal likelihoods cannot be obtained analytically outside of linear and Gaussian models, such as the well-known dynamic linear model (West and Harrison (1986), West and Harrison (1997)). With the negative binomial one-step ahead densities in (7), we can construct the marginal likelihood from the product

$$\begin{aligned} p(\mathcal{D}_T|\gamma) &= \prod_{t=1}^{T} p(N_t|\gamma, \mathcal{D}_{t-1}) \\ &= \prod_{t=1}^{T} \frac{\Gamma(\gamma\alpha_{t-1} + N_t)}{N_t!\Gamma(\gamma\alpha_{t-1})} \left(\frac{\gamma\beta_{t-1}}{\gamma\beta_{t-1}+1}\right)^{\gamma\alpha_{t-1}} \left(\frac{1}{\gamma\beta_{t-1}+1}\right)^{N_t}. \end{aligned} \tag{8}$$

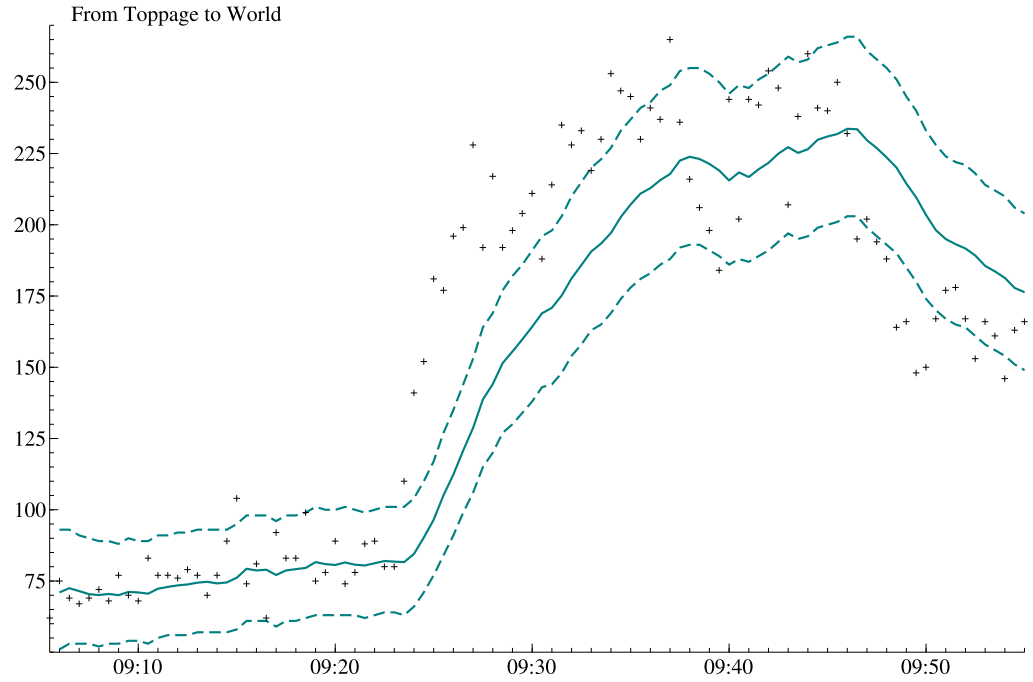


FIG. 1. *The one-step-ahead predictive distribution of N_t (web traffic flow data with* 30 *second time intervals) with static γ. The optimal value of γ is computed by the empirical Bayes method that makes use of the closed form availability of the marginal likelihood. The solid line shows the median of the one-step-ahead predictions and the dashed lines represent the* 95% *predictive intervals.*

If we do not fix γ but treat it as a parameter to be estimated, the sequential analysis of posterior and predictive distributions becomes more complicated. For any given continuous prior choice of γ, it is not possible to obtain an analytically tractable posterior analysis. However, given (8), we can obtain a discrete posterior distribution for γ if we assume a discrete prior defined over the region $(0, 1)$. Alternatively, we can compute a point estimate of γ by maximizing (8). In both cases the computations are straightforward and fast.

When γ is static, regardless of whether it is treated as a tuning parameter or a parameter to be estimated, the PGSS model is slow to adapt to structural changes in counts. Such a structural change is illustrated in Figure 1, where a sudden surge in web traffic on the Fox News website occurs at approximately 9:25 a.m (a similar graphic is presented in Figure 14 of Chen et al. (2018) which contains a full PGSS analysis of the Fox News data). Observe that the median of the one-step-ahead predictive distribution (solid green line) fails to rapidly adapt to the surge. In fact, the predictive distribution from the PGSS model effectively smooths the web traffic data, due to the discount structure in (7). When volume surges at 9:25, the PGSS model underpredicts traffic, and when the number of visitors drops after 9:45 a.m., the PGSS model overpredicts traffic. In both directions the predictions are sluggish in responding to rapid changes in observed data. This is largely due to the static treatment of γ. During the stable period from 9:00 to 9:25, the data provides evidence for a reasonably high value of γ, and past counts significantly contribute to forecasts, $(N_{t+1}|\gamma, \mathcal{D}_t)$. This feature—a strength from 9:00 to 9:25—becomes a weakness when a surge in traffic occurs. The high value of γ gives significant weight to past counts in one-step-ahead forecasts, but the forecasts fail to adequately adapt to the structural change in the time series. At 9:25, a small γ is needed so that less information is inherited from past counts and forecasts rapidly adapt to recently observed data. We view this static γ as a major shortcoming of the PGSS model. In Section 4 we augment the state variable with a dynamic discount factor γ_t that adaptively weights previous

information, based on predictive errors, providing increased model flexibility when structural changes occur.

4. The state-augmented PGSS model. In this section we extend the PGSS model to account for dynamic changes in the discount factor, γ. In doing so, we preserve the properties of the base PGSS model, conditional on the dynamic discount factor. The motivation for modeling γ as dynamic stems from the lack of adaptability of the PGSS model to sudden shifts in the web traffic data flow presented in Figure 1. This adaptability can be achieved by allowing γ_t to be relatively large in stable regions and small in regions where sudden shifts occur, eliminating the need for prospective intervention, as in Chen et al. (2018).

Assuming the same Poisson observation equation (1), we define the state evolution conditional on $\gamma_{1:t} = \{\gamma_1, \ldots, \gamma_t\}$, as $p(\theta_t|\theta_{t-1}, \mathcal{D}_{t-1}, \gamma_{1:t})$, which will be

$$(\theta_t|\theta_{t-1}, \mathcal{D}_{t-1}, \gamma_{1:t}) \sim \text{ScaledBeta}\big(\gamma_t\alpha_{t-1}, (1-\gamma_t)\alpha_{t-1}\big), \tag{9}$$

where $\theta_t \in (0, \theta_{t-1}/\gamma_t)$.

We note that the state equation depends on all the past observations $\mathcal{D}_{t-1}$ and discount factors $\gamma_{1:(t-1)}$ whose contributions are embedded in α_{t-1}. Assuming the same initial state prior as before, $\theta_0 \sim \text{Ga}(\alpha_0, \beta_0)$, the online state update—conditional on $\gamma_{1:t}$—will be $(\theta_t|\gamma_{1:t}, \mathcal{D}_t) \sim \text{Ga}(\alpha_t, \beta_t)$, where

$$\begin{aligned} \alpha_t &= \gamma_t\alpha_{t-1} + N_t, \\ \beta_t &= \gamma_t\beta_{t-1} + 1. \end{aligned} \tag{10}$$

Similarly, the one-step-ahead predictive density can be shown to follow

$$(N_t|\gamma_{1:t}, \mathcal{D}_{t-1}) \sim \text{NegBin}\left(\gamma_t\alpha_{t-1}, \frac{\gamma_t\beta_{t-1}}{\gamma_t\beta_{t-1}+1}\right). \tag{11}$$

The dynamic nature of the discount factors can be described by any Markovian process, such as

$$(\gamma_t|\gamma_{1:(t-1)}) \sim p(\gamma_t|\gamma_{t-1}),$$

which needs to be selected carefully so that it facilitates sequential estimation but is also flexible enough to increase the adaptability of the PGSS model. With this in mind and the fact that γ_t's are defined between 0 and 1, we consider a logistic transformation of the following form:

$$g_t = \text{logit}(\gamma_t) = \log\frac{\gamma_t}{1-\gamma_t},$$

where the transformed series, $g_1, g_2, \ldots, g_{t-1}, g_t, \ldots$, follows a first order autoregressive model, as in

$$g_t = (1-\phi)\mu + \phi g_{t-1} + N\big(0, \sigma^2\big). \tag{12}$$

We take a fully Bayesian point of view and assume priors on the above AR(1) triplet, (μ, ϕ, σ^2), and allow them to be updated sequentially in the face of new count data, substantially increasing temporal adaptability of the PGSS model. To be more specific, for another parametrization $\phi_0 = (1-\phi)\mu$, $\phi_1 = \phi$ and $w = \sigma^{-2}$, we assume the following normal-inverse gamma distribution as the hyperprior:

$$p(\phi_0, \phi_1, w|\mathcal{D}_0) = N(\phi_0, \phi_1|m_0, C_0/w)\,\text{Ga}(w|a_0/2, b_0/2). \tag{13}$$

As we see in the next section, this prior is conditionally conjugate in our model.

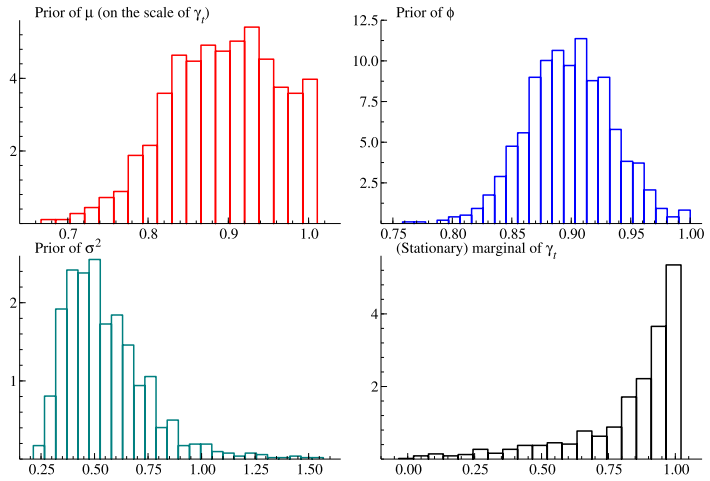


FIG. 2. *The histograms of samples generated from the prior of* $\text{logit}^{-1}(\mu)$ (*top left*), ϕ (*top right*), σ^2 (*bottom left*), *and the stationary marginal of* γ_t (*bottom right*) *implied by the prior in* (13) *with the choice of hyperparameters we consider in analyzing web traffic data. These priors are highly concentrated around the values of* $\text{logit}^{-1}(\mu) = 0.9$, $\phi = 0.9$, *and* $\sigma^2 = 0.5$ *to ensure the persistency of discount factors favoring higher values.*

Hyperparameter selection. If the distribution of g_0 is

$$g_0 \sim N\big(\mu, \sigma^2/\sqrt{1-\phi^2}\big),$$

then the g_t process would be stationary with marginal distribution

$$g_t \sim N\big(\mu, \sigma^2/\sqrt{1-\phi^2}\big)$$

which indirectly implies the stationary distribution of γ_t by inverse logistic transformation. The selection of hyperparameters in the prior of the AR(1) triplet controls the implied stationary distribution for γ_t. While the discount factor is now allowed to be dynamically changing, it is preferable in many cases that the value of γ_t is high and stable over time. This can be realized by high μ, high ϕ, and small σ^2. Assuming a relatively strong prior on (μ, ϕ, σ^2) has practical advantages where the discount factor would be fairly constant to avoid overfitting and high levels of flexibility. This allows the discount factor to be more adaptive only when steep changes are observed in the flow of web traffic data. For all of our numerical examples, the hyperparameters of normal-inverse gamma prior in equation (13) are set at $m_0 = [(1-0.9)\text{logit}(0.9), 0.9]'$, $C_0 = (0.05)^2 I_2$, $a_0 = 10$, and $b_0 = 5$. This prior reflects our preference on the specific values of $(\mu, \phi, \sigma^2) = (\text{logit}(0.9), 0.9, 0.5)$ with small variance. The implied priors of the AR(1) parameters and the stationary distribution of γ_t are shown in Figure 2. Figure 2 shows that the prior distribution favors γ_t values near 1—implying persistent counts—but still allowing for the possibility of γ_t values close to 0—implying a lower level of dependence in N_t on previously observed counts $\mathcal{D}_{t-1}$. We recommend those values of hyperparameters as the default choice. In addition, our experiments using real web traffic data (not shown here) revealed that using a diffuse prior may not be preferred in scenarios that require higher and more stable discount factor processes.

5. Sequential estimation of the sa-PGSS model. The sa-PGSS model requires fast and efficient computational strategies for online updates of posterior and predictive distributions. As pointed out by Storvik (2002) and Carvalho et al. (2010a), traditional Markov chain Monte Carlo (MCMC) methods, especially the forward filtering backward sampling (FFBS) algorithm of Carter and Kohn (1994) and Frühwirth-Schnatter (1994), are computationally expensive, as state variables must be reestimated each time new data is observed. With this in mind,

we develop a particle-based algorithm that allows us to update static as well as the dynamic (state) variables in a fast sequential manner. The initial idea of particle filtering (PF) dates back to the work of Gordon, Salmond and Smith (1993). Since then there have been several successful applications of the PF algorithm in various settings, such as those discussed in Carvalho et al. (2010b) for general mixtures, Gramacy and Polson (2011) for Gaussian process models in sequential optimization, Lopes and Polson (2016) for heavy-tailed distributions, and Prado and Lopes (2013) for estimating parameters in autoregressive time series models. One of challenges common to all PF applications is the particle degeneracy issue that arises in learning static parameters. We overcome this issue using a particle learning algorithm by obtaining the conditional sufficient statistics for the static parameters in a similar vein to the methods proposed by Storvik (2002), Carvalho et al. (2010a) and Prado and Lopes (2013). For recent surveys of particle-based methods, we refer readers to the works of Lopes and Tsay (2011) and Singpurwalla, Polson and Soyer (2018).

In our proposed extension of the PGSS model, the new state vector consists of the (θ_t, γ_t) pair and the static parameters vector is defined by $\vartheta = (\mu, \phi, \sigma)$. The full joint density of all model parameters can be summarized via $p(\theta_{1:t}, \gamma_{1:t}, \vartheta|\mathcal{D}_t)$. However, as our main target is to sequentially update the relevant parameters and to obtain one-step-ahead forecasts, our goal reduces to generating samples from $p(\theta_t, \gamma_t, \vartheta|\mathcal{D}_t)$ which is not available in analytical form. Markov chain Monte Carlo (MCMC) and particle filtering (PF) methods are the two options for generating samples from this density. As pointed out by Storvik (2002), MCMC requires restarting each simulation as new data is observed, increasing the computational burden significantly as the dimension t increases in state space models. As our goal is fast sequential online updating and prediction, we consider PF algorithms that are based on rebalancing of a finite number of particles of the state posterior distributions proportional to the likelihood. As pointed out by Carvalho et al. (2010a), estimating static parameters in sequential models is surprisingly difficult, due to potential particle degeneracy. A potential remedy is to use conditional sufficient statistics of the static parameters when they are analytically available, as considered by Carvalho et al. (2010a), Fearnhead (2002), Storvik (2002). As these conditional sufficient statistics for ϑ can be obtained analytically in our model, we can devise a fast PF algorithm to generate samples from $p(\theta_t, \gamma_t, \vartheta|\mathcal{D}_t)$. This approach is typically referred to as particle learning (PL) algorithm due to the use of conditional sufficient statistics in estimating static parameters. In developing the algorithm, we exploit three major features of our model: 1) Closed-form availability of the state filtering density conditional on the dynamic discount parameters, 2) Closed-form availability of the marginal likelihoods, and 3) Analytical tractability of the conditional sufficient statistics for the static parameters.

Our goal is to eventually obtain samples from $p(\theta_t, \gamma_t, \vartheta|\mathcal{D}_t)$, which can be achieved by augmenting the density by adding α_t, β_t as

$$\begin{aligned} &p(\theta_t, \gamma_t, \alpha_t, \beta_t, \vartheta|\mathcal{D}_t) \\ &\quad = p(\theta_t|\gamma_t, \alpha_t, \beta_t, \vartheta, \mathcal{D}_t)\, p(\alpha_t, \beta_t|\gamma_t, \vartheta, \mathcal{D}_t)\, p(\gamma_t, \vartheta|\mathcal{D}_t) \\ &\quad = p(\theta_t|\alpha_t, \beta_t, \mathcal{D}_t)\, p(\alpha_t, \beta_t|\gamma_t, \mathcal{D}_t)\, p(\gamma_t, \vartheta|\mathcal{D}_t), \end{aligned} \tag{14}$$

where $p(\theta_t|\alpha_t, \beta_t, \mathcal{D}_t)$ is a Gamma distribution with parameters α_t, β_t and $p(\alpha_t, \beta_t|\gamma_t, \mathcal{D}_t)$ is not a known density but can be computed via

$$p(\alpha_t, \beta_t|\gamma_t, \mathcal{D}_t) = \int p(\alpha_t, \beta_t|\gamma_t, \alpha_{t-1}, \beta_{t-1}, \mathcal{D}_t)\, p(\alpha_{t-1}, \beta_{t-1}|\mathcal{D}_t)\, d\alpha_{t-1}\, d\beta_{t-1},$$

where $p(\alpha_t, \beta_t|\gamma_t, \alpha_{t-1}, \beta_{t-1}, \mathcal{D}_t)$ is a degenerate density with deterministic parameter updating given by (10). We note here that to sequentially compute α_t and β_t, we would need

samples from $p(\alpha_{t-1}, \beta_{t-1}|\mathcal{D}_t)$, which we discuss in the sequel. (See the paragraph after the algorithm on page 14.)

The next step is to sample from $p(\gamma_t, \vartheta|\mathcal{D}_t)$, which can be decomposed using a similar augmentation approach via

$$\begin{aligned} p(\gamma_t, \vartheta|\mathcal{D}_t) &= \int p(\gamma_t, \gamma_{t-1}, \alpha_{t-1}, \beta_{t-1}, \vartheta|\mathcal{D}_t)\, d\gamma_{t-1}\, d\alpha_{t-1}\, d\beta_{t-1} \\ &\propto \int p(N_t|\gamma_t, \alpha_{t-1}, \beta_{t-1}, \mathcal{D}_{t-1})\, p(\gamma_t|\gamma_{t-1}, \vartheta, \mathcal{D}_{t-1})\, p(\vartheta|\gamma_{t-1}, \mathcal{D}_{t-1}) \times \cdots \\ &\quad \times p(\gamma_{t-1}, \alpha_{t-1}, \beta_{t-1}|\mathcal{D}_{t-1})\, d\gamma_{t-1}\, d\alpha_{t-1}\, d\beta_{t-1}, \end{aligned}$$

where $p(N_t|\gamma_t, \alpha_{t-1}, \beta_{t-1}, \mathcal{D}_{t-1})$ is a negative binomial density given by (11) and $p(\gamma_t|\gamma_{t-1}, \vartheta, \mathcal{D}_{t-1})$ is the state transition for γ_t given by (12). In addition, we can approximate the online posterior $p(\gamma_{t-1}, \alpha_{t-1}, \beta_{t-1}|\mathcal{D}_{t-1})$ at $t-1$ by S particles as

$$p(\gamma_{t-1}, \alpha_{t-1}, \beta_{t-1}|\mathcal{D}_{t-1}) \approx \sum_{i=1}^{S} w_{t-1}^i \delta_{\{\gamma_{t-1}^i, \alpha_{t-1}^i, \beta_{t-1}^i\}}(\gamma_{t-1}, \alpha_{t-1}, \beta_{t-1}),$$

where $\delta_{\{x\}}(\cdot)$ is the point-mass distribution at x and $\{w_{t-1}^i\}_{i=1:S}$ are nonnegative mixture weights whose sum over i must be equal to one. The final step is to generate from the density $p(\vartheta|\gamma_{t-1}, \mathcal{D}_{t-1})$ to fully implement the above sequential scheme. To do so, we utilize the conditional sufficient statistics updating of ϑ which is available analytically in our model. After the reparametrization of the hyperparameters as $\phi_0 = (1-\phi)\mu$, $\phi_1 = \phi$ and $w = \sigma^{-2}$, we can use a bivariate normal-gamma prior as

$$p(\phi_0, \phi_1, w|\mathcal{D}_0) = N(\phi_0, \phi_1|m_0, C_0/w)\,\mathrm{Ga}(w|a_0/2, b_0/2)$$

for a given collection of prior parameters $\mathcal{S}_0 = \{m_0, C_0, a_0, b_0\}$. The likelihood function for the triplet ϕ_0, ϕ_1, w is obtained via the AR(1) model

$$g_t = \phi_0 + \phi_1 g_{t-1} + N(0, \sigma^2), \tag{15}$$

and for all t, the conditional posterior would be

$$\begin{aligned} p(\phi_0, \phi_1, w|\gamma_{1:t}, \mathcal{D}_t) &= p(\phi_0, \phi_1, w|\mathcal{S}_t) \\ &= N(\phi_0, \phi_1|m_t, C_t/w)\,\mathrm{Ga}(w|a_t/2, b_t/2), \end{aligned} \tag{16}$$

where the set of conditional sufficient statistics is given by $\mathcal{S}_t = \{m_t, C_t, a_t, b_t\}$ updated as a function of $\mathcal{S}_{t-1}$, g_t, and g_{t-1} via

$$\begin{aligned} m_t &= m_{t-1} + A_t e_t \qquad C_t = C_{t-1} - q_t A_t A_t', \\ a_t &= a_{t-1} + 1 \qquad b_t = b_{t-1} + e_t^2/q_t, \end{aligned} \tag{17}$$

and

$$\begin{aligned} G_t &= [1, g_{t-1}] \qquad e_t = g_t - G_t' m_{t-1}, \\ q_t &= 1 + G_t' C_{t-1} G_t \qquad A_t = C_{t-1} G_t / q_t. \end{aligned} \tag{18}$$

The above approach can be implemented with a minor modification under the constraint on ϕ for stationarity as in (Prado and Lopes (2013)). Namely, the prior and posterior distributions are truncated such that the generated particles of ϕ that do not fall in the region $(-1, 1)$ (or $(0, 1)$) are rejected in sampling. Consequently, we can obtain samples from $p(\phi_0, \phi_1, w|\mathcal{S}_{t-1})$ and, in turn, from $p(\vartheta|\gamma_{t-1}, \mathcal{D}_{t-1})$ that is required for updating (14).

In what follows, we present our algorithm that is based on the sequential decomposition of model parameters of interest summarized by (14). Our approach can be viewed as a combination of the auxiliary particle filter (APF) of Pitt and Shephard (1999) with conditional sufficient statistics updating of static parameters. Our algorithm can be summarized via the following steps:

Given a particle set $(\theta_{t-1}^i, \gamma_{t-1}^i, \alpha_{t-1}^i, \beta_{t-1}^i, \vartheta^i | \mathcal{D}_{t-1})$ with weights w_{t-1}^i, repeat the following step 1-6 for each $j \in 1{:}S$.

1. Resample an auxiliary index $i(j)$ with probability $w_{t-1|t}^{i(j)} \propto p(N_t | \hat{\gamma}_t^i, \alpha_{t-1}^i, \beta_{t-1}^i, \mathcal{D}_{t-1}) w_{t-1}^i$ for each i.
2. Propagate g_t^j from the state transition density, $N((1-\phi^{i(j)})\mu^{i(j)} + \phi^{i(j)} g_{t-1}^{i(j)}, (\sigma^2)^{i(j)})$ and set $\gamma_t^j = \text{logit}^{-1}(g_t^j)$.
3. Resample using normalized weights $w_t^j \propto p(N_t | \gamma_t^j, \alpha_{t-1}^{i(j)}, \beta_{t-1}^{i(j)}, \mathcal{D}_{t-1}) / p(N_t | \hat{\gamma}_t^{i(j)}, \alpha_{t-1}^{i(j)}, \beta_{t-1}^{i(j)}, \mathcal{D}_{t-1})$.
4. Compute $\alpha_t^j = \gamma_t^j \alpha_{t-1}^j + N_t$ and $\beta_t^j = \gamma_t^j \beta_{t-1}^j + 1$ and sample θ_t^j from $\text{Ga}(\alpha_t^j, \beta_t^j)$.
5. Update $\mathcal{S}_t^j = f(\mathcal{S}_{t-1}, \gamma_t^j, \gamma_{t-1}^{i(j)})$ via (17) and (18).
6. Sample ϑ^j from $p(\vartheta | \mathcal{S}_t^j)$ given by (16).

Use the particle set $(\theta_t^j, \gamma_t^j, \alpha_t^j, \beta_t^j, \vartheta^j | \mathcal{D}_t)$ for the next time period $t+1$.

We note here that in step 1, $\hat{\gamma}_t^i$ is set equal to γ_{t-1}^i as an estimator for γ_t (similar to the APF approach). At the end of step 3 and as a consequence of resampling, we obtain samples from $p(\gamma_t, \alpha_{t-1}, \beta_{t-1} | \mathcal{D}_t)$ that are used in updating α_t and β_t in step 4. We do not need to propagate θ_t from θ_{t-1}, as the conditional filtering density is available analytically. An important feature of our sa-PGSS model is that the vector of past discount terms, $\gamma_{1:t}$, can be summarized by a lower dimensional vector $(\gamma_t, \alpha_{t-1}, \beta_{t-1})$, thus reducing the dimension of the state vector for γ_t's to 3 from t. This avoids the need to generate from the t dimensional state vector (can be achieved using a forward filtering and backward sampling (FFBS) step) and reduces the computational burden significantly.

Particle dimension and effective sample size. Our experiments with the sa-PGSS model typically suggest that a particle size of $N = 5000$ was more than sufficient in all the numerical examples. We also investigated the implications of using smaller particle sizes (1000, 2000, and 3000) on the estimation paths of both the state and static parameters of our model and found no clear differences. We omit the details of these experiments to preserve space in the narrative and use $N = 5000$ as a very conservative particle size in all our subsequent numerical examples.

To assess the existence of potential particle degeneracy in the estimates obtained using our PF algorithm, we also keep track of the so-called effective sample size (ESS) via

$$\text{ESS}_t = \frac{1}{\sum_{i=1}^N (w_t^i)^2},$$

where w_t^i represents the weight of particle i at t before the resampling step (if any). We note there that $1 \leq \text{ESS}_t \leq N$ where lower values indicate evidence in favor of degeneration and vice versa. The ESS estimates can be used as a monitoring tool for assessing the need to resample at each point in time and to detect anomalies (such as structural breaks or sudden bursts in data). We investigate the implications of monitoring the ESS over time and how it can be used as a practical tool in our numerical examples.

6. Analysis of web traffic data. In this section we analyze three examples of web traffic data to investigate the implications of using the family of PGSS models, including sa-PGSS extension, and to show the implementation of our particle based algorithms. We start our discussion with a simulated study of a web traffic stream that exhibits clear and sudden shifts. The second and third examples consider web traffic data from the Fox News website, as described in Section 2. In all three scenarios we provide comparisons of online learning and forecasting results for various models. A brief description of each model included in our comparison is as follows:

1. *sa-PGSS*: The state-augmented Poisson-gamma state space model where the dynamic discount factor, γ_t, evolves over time via the transition equation defined in equation (12).

2. *PGSS-random*: The Poisson-gamma state space model where γ is assumed to be static but random. We assume that the prior distribution of γ is a uniform discrete distribution defined over $\{0.01, 0.02, \ldots, 0.99\}$, an approach considered in Aktekin, Soyer and Xu (2013). The posterior distribution of γ is then obtained via

$$p(\gamma|\mathcal{D}_t) \propto p(\gamma) \prod_{s=1}^{t} p(N_s|\mathcal{D}_{s-1}, \gamma),$$

where $p(N_s|\mathcal{D}_{s-1}, \gamma)$ is the negative binomial marginal likelihood from (8).

3. *PGSS-deterministic*: The Poisson-gamma state space model where the discount factor γ_t evolves dynamically but in a deterministic manner as considered by Chen et al. (2018). More specifically, γ_t is assumed to exhibit the following functional form

$$\gamma_t = d + (1 - d)\exp(-k\alpha_{t-1}),$$

where d represents the baseline, k is a tuning parameter controlling the speed of the information decay, and α_{t-1} is the shape parameter of the time $t-1$ posterior distribution from (4). The motivation of using the above specification stems from scenarios with zero counts and to mitigate the numerical issues caused by extremely small α_{t-1}'s. When α_{t-1} is large, the exponential term approaches 0 and $\gamma_t \approx d$, leading to an approximately constant discount factor. In Chen et al. (2018) and our study, the decay parameter is set to $k = 1$. Formally, the optimal value of d can be estimated using an empirical Bayes approach as in

$$d^* = \arg\max\{p(d|\mathcal{D}_T)\} = \arg\max\left\{p(d) \prod_{t=1}^{T} p(N_t|\mathcal{D}_{t-1}, d)\right\}$$

with some constraint on the support of d, such as $d \in (0.9, 1)$. In the Fox News dataset, $d = 0.9$ is obtained by following this procedure with the training dataset. We choose $d = 0.9$ for the simulation example since a training dataset was not available.

4. *Dynamic linear model (DLM)*: DLMs are commonly used state space models for nonstationary time series with Gaussian observations. Similar to the PGSS model, tractable updates of posterior and predictive distributions are available in DLMs. We use the first order polynomial DLM (West and Harrison ((1997), Chapter 2.1)) that has the following form:

$$\begin{aligned} N_t &= \mu_t + v_t, \qquad & v_t &\sim N(0, V_t), \\ \mu_t &= \mu_{t-1} + w_t, \qquad & w_t &\sim N\big(0, V_t W_t^*\big), \end{aligned}$$

where the parameters for the observational and state errors, V_t and W_t^*, are modelled using discount factors, β and δ, respectively (West and Harrison ((1997), Chapter 10.8)). Our experiments with the web traffic data showed that using $\delta \in \{0.95, 0.75\}$ for the state evolution and $\beta = 0.95$ for the stochastic volatility provide reasonable coverage to the Fox News streams. The discount factor β for the stochastic volatility has less impact on the posterior

results and forecasting performance. The initial values are set such that the comparison with the sa-PGSS models is fair. Using the general terminology of DLMs, we set the initial hyperparameters to $m_0 = C_0 = N_0$ (using the discarded, latest observation as the prior mean and variance of μ_t) and $n_0 = 2$ and $S_0 = N_0$ to mimic the initial priors that we used for the PGSS models (Note that n_t is the degree-of-freedom parameter and can be interpreted as the size of the prior relative to the number of actual observations). For the low count example, as $N_0 = 0$ (as the initial variance will be undefined), we replaced $N_0 = 0.1$ to initialize the DLMs which otherwise would be undefined.

Performance measures. Allocating advertisements across multiple webpages requires a decision supported by an associated loss function. While the posterior mean and median are associated with squared error and ℓ_1 loss functions, respectively, different applied settings may require alternative loss functions. For instance, in the Fox News example the objective may be to maximize advertising revenue, though often, revenue maximization is not the only goal. There are typically additional advertising campaign constraints, such as serving particular ads to particular target demographics. These constraints may require the development of multicriteria loss functions to account for both revenue and campaign objectives. With this in mind, we focus on the following multiple predictive measures and hold our methodology open to any specific loss analysis.

In assessing the model performance in analyzing web traffic data, we consider several performance measures that can be tied to respective loss functions, such as the mean absolute percent error (MAPE), the mean squared error (MSE), the mean absolute deviation (MAD), scaled mean standard deviation (sMSD), and the log marginal likelihood. In addition, we also show how the posterior model probability estimates can be used for assessing/monitoring the online model fit performance. Many of these measures are used to assess the performance of count time series models in the literature. We refer the reader to the discussions on predictive performance and loss functions in the recent works of Berry and West (2020) and Berry, Helman and West (2020).

MAPE is a standard measure of predictive performance and is defined as

$$\text{MAPE}_t = \frac{100}{t} \sum_{s=1}^{t} \frac{|N_s - f_s|}{N_s},$$

where f_t is the point forecast of N_t at $t-1$; in our study, f_t is the posterior median of the one-step-ahead predictive distribution $p(N_t|\mathcal{D}_{t-1})$ for simplicity, although the optimal point forecast for the standard of MAPE can also be considered (e.g., Berry, Helman and West ((2020), Section 4.3)).

We remark here that MAPE estimates at time t would not be well defined if $N_s = 0$ for any $s \leq t$. This problem is more severe when the web traffic counts are small, as is the case with one of our examples. In reporting the MAPE estimates for the low count example, we replaced the denominator by 1 when $N_s = 0$. Alternatively, we computed the mean absolute distance (MAD) for all cases which can be obtained by replacing the denominator of the MAPE estimate by 1 for all s. The scaled mean standard deviation (sMSD) can also be used to assess model performance and can be obtained by using the sample average at t, $\frac{1}{s}\sum_{u=1}^{s} N_u$, instead of N_s. In addition, we also computed the commonly used mean squared error (MSE) estimates. For all measures, MAPE, MAD, MSE, and sMSD, a lower number indicates a better prediction performance.

The posterior model probability $p(\mathcal{M}|\mathcal{D}_t)$ for model $\mathcal{M} \in \{$sa-PGSS, PGSS-random, PGSS-deterministic, DLM($\delta = 0.75$), DLM($\delta = 0.95$)$\}$ is used to monitor the online model fit. Particularly, $p(\mathcal{M}|\mathcal{D}_t)$ can help us identify when and why a particular model outperforms others. For instance, in count data with sudden bursts and/or structural breaks, $p(\mathcal{M}|\mathcal{D}_t)$ can

provide simple-to-interpret visual guidance. In addition, one can also consider $p(\mathcal{M}|\mathcal{D}_t)$ to assess different choices of hyperparameters, computational methodologies, and particle sizes.

In order to compute $p(\mathcal{M}|\mathcal{D}_t)$, the marginal likelihood is needed and analytically available through (the sum of) (8) in the PGSS family of models. For instance, in the sa-PGSS model, the log marginal likelihood can be computed as a mixture, as in

$$\begin{aligned}&\log p(N_t|\mathcal{D}_{t-1})\\&\quad=\int \log p(N_t|\mathcal{D}_{t-1},\gamma_t,\alpha_{t-1},\beta_{t-1})\,p(\gamma_t,\alpha_{t-1},\beta_{t-1}|\mathcal{D}_{t-1})\,d(\gamma_t,\alpha_{t-1},\beta_{t-1})\\&\quad=\frac{1}{S}\sum_{i=1}^{S}\log p\big(N_t|\mathcal{D}_{t-1},\gamma_t^i,\alpha_{t-1}^i,\beta_{t-1}^i\big),\end{aligned}$$

where the density of $(N_t|\mathcal{D}_{t-1},\gamma_t,\alpha_{t-1},\beta_{t-1})$ is the negative binomial distribution given by equation (11). In the above, the particle set, $(\gamma_t^i,\alpha_{t-1}^i,\beta_{t-1}^i)$, is obtained by augmenting $(\gamma_{t-1}^i,\alpha_{t-1}^i,\beta_{t-1}^i,\vartheta^i)$ with γ_t^i through $p(\gamma_t|\gamma_{t-1}^i,\vartheta^i)$. We remark here that the state variable, θ_t, is Rao–Blackwellized which reduces the overall computational burden significantly. For DLMs that are based on Gaussian likelihoods, the above marginal likelihoods will be replaced by the t-distribution. The marginal likelihoods can also be used to assess the overall model fit to data.

Computational details and performance. The computations for all three examples are implemented in Ox (Doornik (2007)) on a laptop computer with Intel Core i7-7500U CPU 2.70GHz, 2.90GHz, RAM-8GB specifications. Table 1 summarizes the actual time (in seconds) of sampling the online joint posterior distribution, $p(\theta_t,\gamma_t,\vartheta|\mathcal{D}_t)$, using PF and MCMC methods. For instance, in the Fox News example (with high counts) the time for completing the update of $N=5000$ particles from time period t to $t+1$ for the PF algorithm (without any explicit parallelization) is 0.27 seconds, on average, and 0.328 at maximum for $t=0:T-1$. We note here that this performance stays approximately the same as the dimension of the series (T) increases. Conversely, the estimation of the online posterior at time $t=T$, using an MCMC algorithm with an independent Metropolis–Hastings step, is approximately equal to 65.85 seconds. The details of the MCMC algorithm (5000 iterations after a 500 burn-in period) can be found in the Appendix. We note here that, for both Fox News examples, each time period is 30 seconds long, and the MCMC approach far exceeds this threshold. The computational results for the low count web traffic stream were almost identical. Our experiments indicated that, as the dimension of T got larger, the computational burden for the MCMC method exponentially increased while the PF algorithm stayed around the same. These results provide further evidence that any model that requires the implementation of MCMC methods in a sequential setting, such as Markov switching models, is not a suitable alternative for analyzing web traffic data that is observed in short time intervals.

The online updating for DLMs with known discount factors takes less than 0.0001 seconds. However, we note here that, in our numerical examples, the DLM discount factor is

TABLE 1
Summary of computational performance in seconds

	Simulation ($T=99$)	Fox News ($T=99$)
PF (Avg)	0.264	0.270
PF (Max)	0.297	0.328
MCMC	64.947	65.858

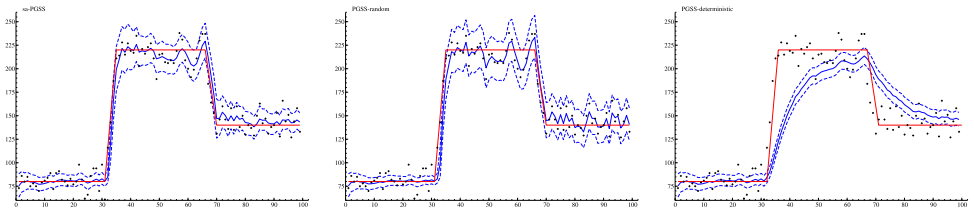
not reestimated for each time period which would require the sequential evaluation of the marginal likelihoods. The DLMs and PGSS models with static discount factors are unable to quickly adapt to structural changes, due to nondynamic nature of their respective discount factors. The attractive feature of the sa-PGSS model is that it achieves such adoption without the need for heavy computational burden (each update takes less than half a second). In practical applications, these online updates typically become inputs of a large multivariate network model (Chen et al. (2018)) where computational efficiency becomes of utmost importance. A comparable model in the domain of DLMs would involve treating the discount factor as dynamic. To the best of our knowledge, there are no DLM extensions that can achieve this while preserving computational feasibility.

6.1. *Simulated web traffic data with structural breaks.* We first start our analysis of web traffic data by simulating a stream that exhibits structural breaks. The simulation is designed such that these breaks are visually clear with significant jumps when observed retrospectively. We note here that our analysis is designed with the lens of sequential learning and forecasting in high-frequency time intervals to mimic a hypothetical web traffic scenario. For instance, Markov switching models may, in fact, work to detect changes when the full data series is available for a retrospective analysis, but Markov switching models do not offer the same sequential parameter updates that are available in our model for online monitoring and forecasting. As such, we designed a simulation study that can highlight the adaptability of our proposed sa-PGSS model in analyzing web traffic data where automated machine monitoring without the need for human intervention is key.

The data are generated from a nonhomogeneous Poisson model via $N_t \sim \text{Po}(\theta_t^*)$ independently, where

$$\theta_t^* = \begin{cases} 80 & t \in 1{:}32, \\ 115, 150, 185 & t = 33, 34, 35, \text{ resp.}, \\ 220 & t \in 36{:}67, \\ 200, 180, 160 & t = 68, 69, 70, \text{ resp.}, \\ 110 & t \in 71{:}100. \end{cases} \tag{19}$$

In the simulation design, two relatively slow-to-build structural breaks are represented at time points $t = 33$ and $t = 68$ with structural shifts occurring shortly after (see the red lines on Figures 3(a), 3(b), and 3(c)). The overall pattern of the simulated set roughly mimics that of the Fox News example with steeper and clearer breaks. The simulation design allows us to investigate the flexibility of the sa-PGSS model in adopting to sudden surges in the count



(a) sa-PGSS (b) PGSS-random (c) PGSS-deterministic

FIG. 3. *Online posterior distributions of θ_t with mean (solid) and 95% credible intervals (dashed) and the true values of θ_t^* (solid, piecewise linear) with the observed counts (+). The three figures, (a), (b), and (c) correspond to sa-PGSS, PGSS-random and PGSS-deterministic models, respectively. Compared with the posterior of sa-PGSS in (a), that of PGSS-random in (b) is volatile and overly adaptive in $t \geq 71$. The posterior of PGSS-deterministic in (c) is too persistent for the changes of true Poisson rates.*

data without the need for more complex models that are computationally expensive and thus are not suitable for fast online learning/forecasting.

Figures 3(a), 3(b), and 3(c) display the online state posterior distributions with the respective 95% credible intervals for all three PGSS models where the straight red line represents the level of the true state variable, θ_t. The posterior uncertainty, provided by the sa-PGSS model in Figure 3(a), exhibits a fairly quick adaptive behavior to the sudden changes on the level. In Figure 3(b) the PGSS-random model also seems to provide flexible coverage at first glance, with some excessive overfitting right around the second state change at $t = 68$. In contrast, the posterior coverage provided by the PGSS-deterministic model from Figure 3(c) clearly shows the shortcomings of the base PGSS model with a deterministic discount factor, as evidenced by its inability to adopt to sudden changes in the level.

To further investigate the online fit performances around and at the inflation points, we computed the cumulative mean squared error (MSE) estimates over time via

$$\mathrm{MSE}_t = \frac{1}{t}\sum_{s=1}^{t}\big(E[\theta_s|\mathcal{D}_s] - \theta_s^*\big)^2,$$

where θ_s^* represents the true value of the Poisson rate at time s given in equation (19). The overall pattern of the MSEs for all three models are shown in Figure 4. Right after the first change point, the PGSS-deterministic model provides the worst coverage with respect to the other two models with random discount factors. The encouraging finding here is that the sa-PGSS model consistently outperforms the PGSS-random model strictly after the first change-point (around $t = 33$). This may be explained by plotting the estimated paths of discount factor γ for both models. For lower values of γ, the PGSS model tends to over-fit the data (i.e., posterior mean estimates will follow recently observed data too closely).

Figures 5 and 6 show the estimated paths of the posterior means and the respective 95% credible intervals of the discount factors of the sa-PGSS (via γ_t) and PGSS-random (via γ) models. For the sa-PGSS model, the initial γ_t estimates are high (close to 1) followed by a steep drop right after the first change point ($t = 33$). Another drop can be observed at the second change point ($t = 68$) beyond which the discount factor gradually increases back to higher levels. The path of γ for the PGSS-random model tells a similar story during the first 32 time points with a steep decrease at the change point. However, after the second change point, the PGSS-random model is unable to push γ back to the region of 0.9 to 1, unlike the sa-PGSS model. We believe that this sheds light on the dominance previously observed in the MSE estimates from Figure 4, as the PGSS-random model is unable to recover the

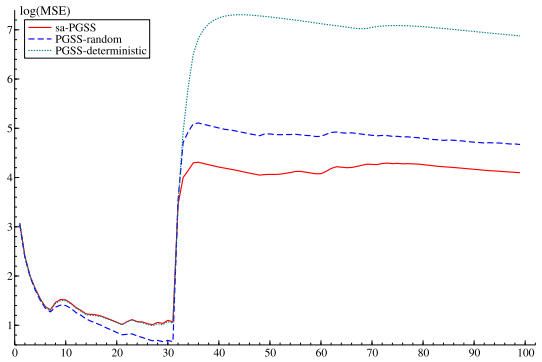


FIG. 4. *The mean squared errors (log-scale) for the sa-PGSS (solid), PGSS-random (dashed), and PGSS-deterministic (dotted) models. The MSE estimates for the PGSS-deterministic model are extremely large with respect to the other two models and after $t = 33$ (the first switching point) are beyond the borders of the figure in this scale.*

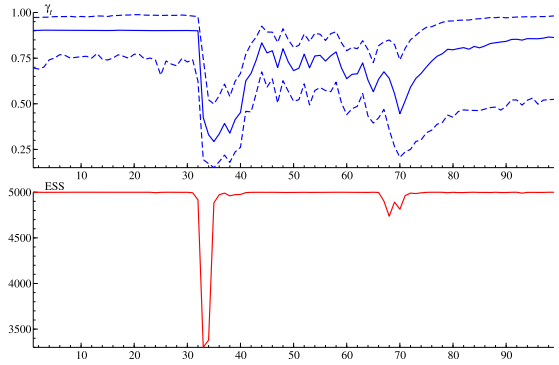


FIG. 5. *The means and credible intervals of the posterior of dynamic discount factor,* $p(\gamma_t|\mathcal{D}_t)$ (*top*), *and the ESS over time* (*bottom*). *Both discount factor and ESS are lowered when the true Poisson rates started to change. The posterior of discount factors starts to increase after* $t \geq 71$.

appropriate value of γ, especially after the second break point. The dynamic nature of γ_t in the sa-PGSS model allows the posterior distribution to shift between high values (when θ_t's are similar or close to identical) and low values (when θ_t's are not similar which occurs at the breaks). These structural breaks can also be identified by the sudden dips in the ESS estimates from Figure 5, once again occurring at $t = 33$ and $t = 68$. Severe and sudden drops in ESS estimates can be used as a formal monitoring tool for identifying structural breaks in automated machine learning settings.

In terms of online model fit and predictive performance (marginal likelihoods, model probabilities, and MAPE estimates), the sa-PGSS model outperforms the other two models after structural breaks occur. Figure 7 shows the posterior model probabilities for the PGSS models and the DLMs, where equal prior probabilities are assumed. One noteworthy observation is that the PGSS-random is found to be the best model during the initial 30 observations. This is expected since there is no need for a dynamically changing discount factor until around $t = 30$, as the simulation design implies that γ should be equal to 1 in this epoch (e.g., $\theta_{1:30}$ are the same). After the first change point, sa-PGSS quickly becomes the dominant model and continues to outperform others due to its ability to rapidly adapt to the new level of the generated counts. In a similar vein the results from the MAPE estimates from Figure 8 also confirm the findings implied by the model probabilities where the sa-PGSS model consistently outperforms the other models after the first change point. It is worth mentioning here that, in the

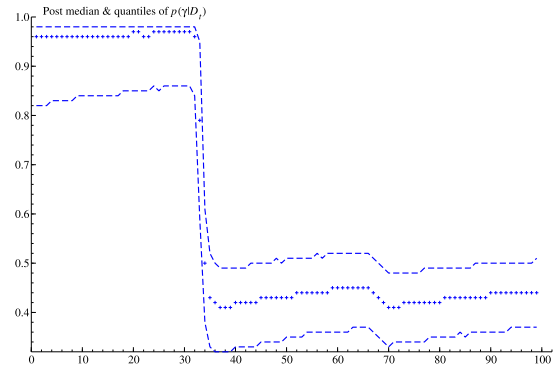


FIG. 6. *The median* (+ *symbols*) *and credible intervals of the posterior of constant discount factor,* $p(\gamma|\mathcal{D}_t)$. *Unlike the postrior results of sa-PGSS in Figure 5, once the discount factor is lowered, it remains to be around* 0.4–0.5 *and never increases.*

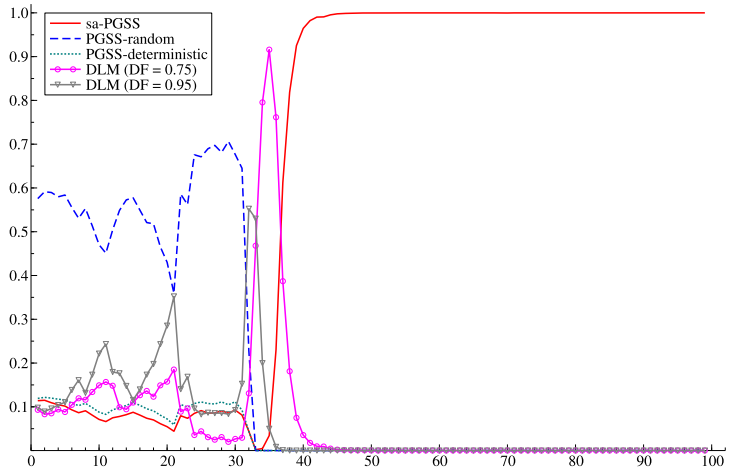


FIG. 7. *The posterior model probabilities of sa-PGSS* (*solid*), *PGSS-random* (*dashed*), *PGSS-deterministic* (*dotted*), *DLM*($\delta = 0.75$) (*circle*), *and DLM*($\delta = 0.95$) (*triangle*) *with equal prior probabilities*.

first 30 time points, the difference between the three models is small. The difference becomes visually clearer at the two time points of structural change. A summary of the MAPE, MAD, sMSD, MSE, and marginal log-likelihood estimates are shown in Table 2 where the sa-PGSS and PGSS-random models outperform the DLMs and the PGSS-deterministic models.

In summary, our goal was to develop a highly adaptable PGSS model suitable for sequential parameter learning and online demand forecasting of web traffic data with structural breaks. In doing so, we focused on developing a fast and efficient particle based algorithm while avoiding traditional MCMC methods that are found to increase computational burden significantly. The summary of results discussed previously based on the simulated study confirms that our proposed sa-PGSS model performs well in terms of online model fit and predictive performances such as MAPE, MAD, sMSD, MSE, and log-marginal likelihood when compared against other suitable modeling strategies.

6.2. *Web traffic data from the Fox News website* (*high count example*). In what follows, we first analyze one of the flows from the Fox News website with relatively high web traffic

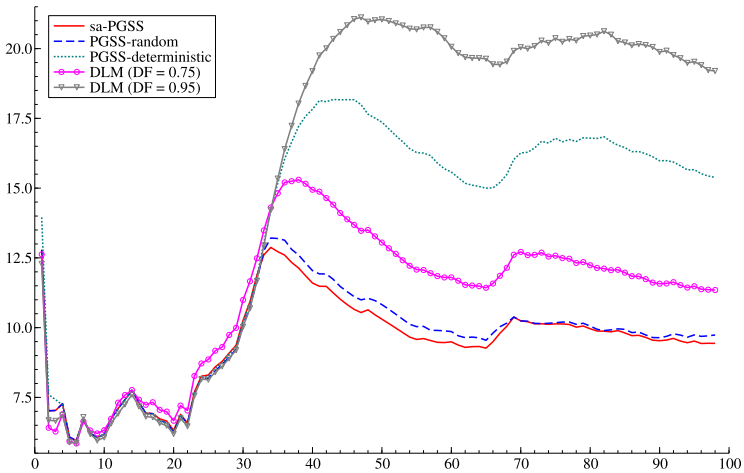


FIG. 8. *The mean absolute percentage errors* (*MAPEs*) *of sa-PGSS* (*solid*), *PGSS-random* (*dashed*), *PGSS-deterministic* (*dotted*), *DLM*($\delta = 0.75$) (*circle*), *and DLM*($\delta = 0.95$) (*triangle*) *over time*.

TABLE 2
Overall summary of predictive performances in the simulation study, Fox News high-count and Fox News low-count data

	PGSS			DLM	
	sa	Random	Deterministic	$\delta = 0.75$	$\delta = 0.95$
Simulation					
MAPE	9.4	9.7	15.3	11.3	19.1
MAD	1309.6	1373.4	2440.8	1653.3	3207.5
sMSD	11.6	12.2	21.4	14.9	27.5
MSE	335.6	381.9	1228.2	576.5	1919.1
log-ML	−421.0	−435.6	−690.4	−439.0	−486.5
Fox News (high-count)					
MAPE	9.4	9.0	14.9	10.1	18.6
MAD	1504.0	1423.2	2591.9	1635.0	3422.7
sMSD	13.2	13.0	23.0	14.8	29.8
MSE	388.4	364.1	1171.6	501.6	2061.7
log-ML	−436.3	−431.0	−664.8	−439.9	−494.3
Fox News (low-count)					
MAPE	53.5	52.5	51.4	61.4	54.9
MAD	136.3	137.3	135.3	139.4	137.7
sMSD	58.5	58.6	57.9	59.9	59.1
MSE	348.4	355.5	345.4	325.8	342.6
log-ML	−194.2	−193.5	−193.7	−196.5	−198.7

counts (roughly between 70–270), as described in Section 2. In doing so, we highlight here the three main aspects that we envision web traffic monitoring and forecasting models should possess. The first one is to be able to adequately quantify the uncertainty exhibited by web traffic counts in high-frequency intervals, in other words sequentially observed count data. A natural choice for analyzing such count data is to use a model with a Poisson distributed likelihood, such as the PGSS family of models. Another approach would be to use Gaussian based models as an approximation, such as the DLMs that may be viable alternatives when web traffic counts are large. However, when web flows are relatively low, say between zero to 20 (as is the case with our next example), Gaussian based models are unable to quantify the uncertainty characteristics of web traffic data accurately. The forecasting (predictive) densities for DLMs are also Gaussian by definition and are, therefore, symmetric, unlike those of the PGSS models that are negative binomial (11) that are able to exhibit skeweness and are able to represent zeros. The second main aspect we would expect the models to posses is the ability to provide fast and efficient monitoring and prediction mechanisms that are suitable for analyzing web traffic in high-frequency intervals (30 seconds in our illustrations). The class of PGSS models and DLMs fit this profile, whereas more complex models, such as Markov switching models with unknown number of states, do not, due to the additional computational burden that would be required for sequential analysis, especially in 30 second long time intervals. The third, and the last, feature we would envision web traffic models to have is the rapid adaptability to sudden structural changes while preserving computational feasibility. Our proposed sa-PGSS model was designed to exhibit such adaptability as an extension of the current PGSS models in the literature. A Markov switching model with an unknown number of states (as the data is observed sequentially) may provide such adaptability but would fail to provide the necessary computational feasibility. Table 1 provides support in favor of our proposed family of PGSS models with respect to their computational capability.

With the three previously discussed main features, we turn our attention to the analysis of web traffic data. Chen et al. (2018) present a thorough analysis of the Fox News data set uti-

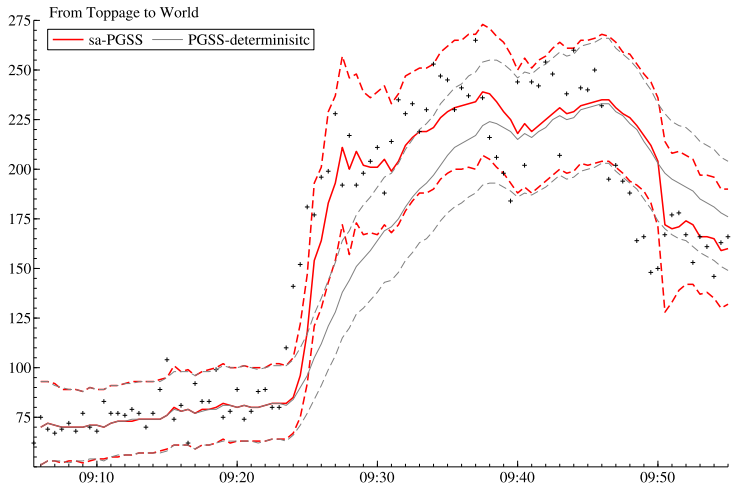
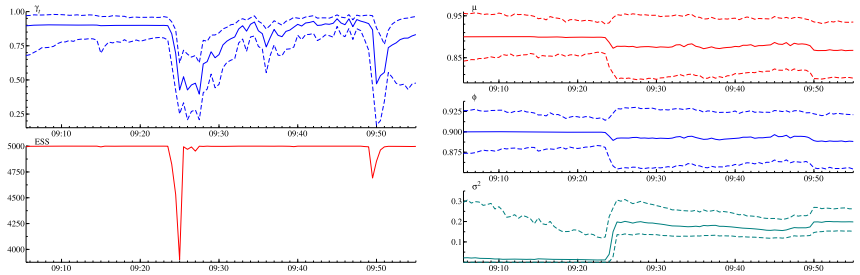


FIG. 9. *One-step-ahead predictive distributions of N_t for the sa-PGSS (solid/dashed lines) and PGSS-deterministic (thinner lines) models, revisiting the dataset used in Figure 1. The sa-PGSS model is able to change its predictive location flexibly in* 9:25–9:30 *and* 9:50–9:55, *while it makes stable predictions that are almost identical to those of the PGSS-deterministic model in* 9:05–9:25 *and* 9:40–9:45.

lizing the PGSS-deterministic model, and in this section we benchmark the sa-PGSS model's forecasting performance and model fit to PGSS-random, PGSS-deterministic models, and DLMs. We consider one particular flow from the top (main) page of the website to the category titled "World" between 9:05 and 9:55 a.m. on February 23, 2015. The first observation at 9:05 is omitted from the series, as it is set equal to the hyperparameter of the initial state prior α_0. The total length of the time series is $T = 99$. Our goal is to monitor and forecast the number of visitors navigating to the "World" Section 30 seconds in advance, allowing advertising impressions to be optimally allocated across sections.

We first discuss the performance of the sa-PGSS model against the PGSS-deterministic of Chen et al. (2018) that is used to model web traffic data and DLMs. Figure 9 shows the one-step-ahead predictions and credible intervals for the sa-PGSS model and the PGSS-deterministic model where the sa-PGSS model yields significantly better predictions when there is a surge in the web traffic around 9:25. Forecasts from the PGSS-deterministic model do not quickly adapt to such a sudden change in counts, highlighting the need for more flexible discounting strategies. The estimation paths of γ_t and ESS are shown in Figure 10(a),



(a) Posterior of γ_t and ESS. (b) Posteriors of ϑ

FIG. 10. *Left: Online posterior median and* 95% *credible intervals for γ_t (top) and the ESS (bottom). The drop of discount factors can be seen only in the time of sudden changes in observed counts. Right: Online posterior of AR(1) parameters, that is, $p(\text{logit}^{-1}(\mu)|\mathcal{D}_t)$, $p(\phi|\mathcal{D}_t)$ and $p(\sigma^2|\mathcal{D}_t)$. The informative prior chosen for this analysis is affected only by the sudden burst in* 9:25.

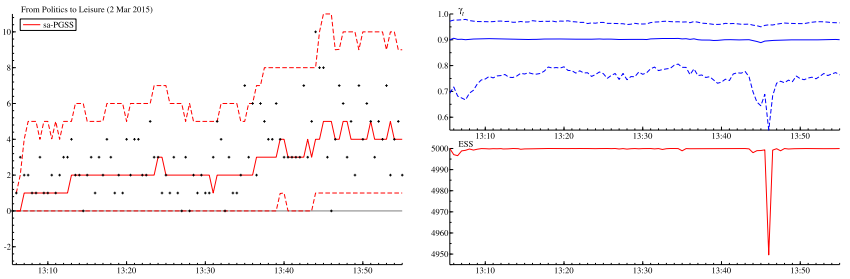
where the drop in the posterior mean of γ_t and the ESS coincide with the sudden shifts in the web traffic counts, a property that the PGSS-deterministic model fails to capture. The performance summary from Table 2 confirms that the sa-PGSS model has better predictive performance with respect to the PGSS-deterministic model and the DLMs, no matter what performance measure is used. Even though the web traffic counts are relatively large where DLMs can be considered as viable approximations, based on all performance measures (MAPE, MAD, sMSD, MSE, and log-marginal likelihoods) the sa-PGSS and PGSS-random yield better forecasting performances. Such a finding further shows support in favor of using models with Poisson likelihoods and negative binomial predictive densities for modeling and forecasting web traffic data, even when the counts are relatively large.

Our experiments with both the simulated and real web traffic flows lead us to believe that the PGSS-random model also provides a reasonable fit to data in the absence of sudden surges in traffic. However, it also does not provide any significant computational advantage over the sa-PGSS, even if γ is treated as a discrete random variable, as was the case in all of our numerical analysis. For more realistic cases where γ is treated as a continuous random variable (similar to the sa-PGSS model), the sa-PGSS model would yield superior computation performance, as it will not involve any Metropolis–Hastings type sampling with accept-reject steps. However, PGSS-deterministic always underperforms with respect to sa-PGSS and PGSS-random models. As a rule of thumb, we suggest the use of the sa-PGSS model, as it can adopt to sudden changes in web traffic and is not computationally more expensive than the PGSS-random model.

Next, we discuss the implications of the hyperpriors used in modeling the dynamic discount factor of the sa-PGSS model. The AR model structure and informative prior distributions mimic a constant discount factor when called for by the data, enabling us to sharply estimate the dynamic discount factor γ_t in stable epochs. The effect of using informative priors on the hyperparameters, $\vartheta = (\mu, \phi, \sigma^2)$, in the AR model of γ_t can be observed in Figure 10(b). The posterior distribution paths of μ, ϕ, and σ^2 are quite stable, especially during the first 20 minutes before the sudden surge in the traffic. While the prior is informative, it is sufficiently diffuse, placing moderate prior mass on lower values of γ_t, as in Figure 2. Observe the drop in location parameter μ and increased variance σ^2 from 9:20-9:23. The changes in posterior distributions for μ, ϕ, and σ^2 translate to a drop in γ_t from 9:20-9:23 but with increased uncertainty as in Figure 10(a). Our analysis shows that, despite the relatively strong priors on the AR parameters $\vartheta = (\mu, \phi, \sigma^2)$, the posterior distributions quickly respond to changes in the level of the web traffic data.

6.3. *Web traffic data from the Fox News website* (*low count example*). In this section we consider a Fox News flow with relatively low number of counts (zero to 10) to further investigate the implications of using different modeling strategies. More specifically, we are interested in investigating the performance of the sa-PGSS model with respect to the DLMs, PGSS-random, and PGSS-deterministic models in cases where the web traffic counts are stable and exhibit no clear structural changes. The flow we consider is from the "Politics" topic to the "Leisure" topic that is recorded between 13:05–13:33 p.m. on March 2, 2015. Notably, in contrast to the previous subsection, the observed web traffic counts in this example are small but not zero-inflated. We view this as an example where the Poisson likelihood and the Negative binomial predictive density of the PGSS family of models are more suitable for quantifying the uncertainty characteristics of web traffic data, such as skewness and the existence of zero counts, features Gaussian models are unable to capture.

Figure 11(a) displays the predictive distributions of the sa-PGSS models against the observed counts of this flow. The data process is stable in the first 30 minutes, with a subtle and upward level shift occurring around 13:40 p.m. Clearly, this shift is not as evident as



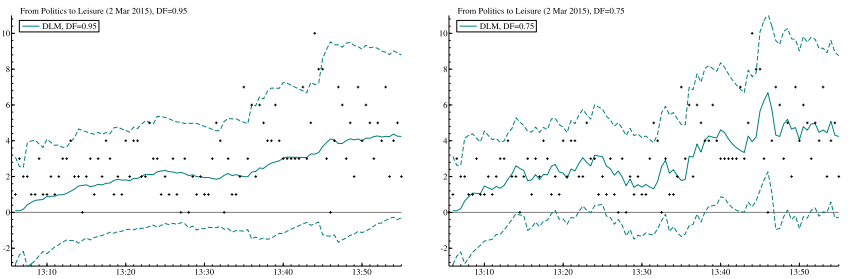
(a) Forecasting $p(N_{t+1}|\mathcal{D}_t)$. (b) Posteriors of γ_t and ESS.

FIG. 11. *Left*: *One-step-ahead predictive median and* 95% *credible intervals of web traffic* (*low counts*). *Right*: *Online posterior median and* 95% *credible intervals for* γ_t *and ESS.*

those observed in the previous flow (from the main page to the topic "World"). As a consequence, the mean estimates of the posterior distribution of γ_t are fairly stable, as shown in Figure 11(b). However, it is clear that the subtle shift around 13:40 p.m. is captured by the spike in the lower uncertainty bounds. In addition, the effective sample size drops sharply at the suspected change point. As before, this is a desirable quality of the sa-PGSS models that can be used for automated machine monitoring of web traffic.

Next, we investigate the implications of using state space models with Gaussian likelihoods, such as DLMs, in analyzing web traffic data with relatively smaller counts. We assume that the counts, N_t's, are continuous valued and are modeled using the first-order polynomial DLMs with two discount factors, δ for the state evolution and β for the stochastic volatility. As is the case with the PGSS family, the sequential state posterior updating and one step ahead prediction can be done analytically via forward filtering; see West and Harrison ((1997), Table 10.4).

Even tough DLMs are computationally comparable to the PGSS family of models, they are unable to accurately represent the characteristics of low web traffic counts. Figures 12(a) and 12(b) show the predictive distributions of the first-order polynomial DLMs with two different discount factor estimates for the state evolution. These discount factors were selected such that the DLM results were comparable to the performance of the PGSS family of models. We remark here that the predictive 95% credible intervals using the DLMs can contain negative values which is not realistic in modeling web traffic data. When compared with the predictions provided by the sa-PGSS model shown in Figure 11(a), the predictive distributions of DLMs are more biased toward negative values, due to the symmetry of the predictive



(a) DLM with $\delta = 0.95$. (b) DLM with $\delta = 0.75$.

FIG. 12. *One-step-ahead predictive median and* 95% *credible intervals of web traffic* (*low counts*) *using a first order polynomial dynamic linear model with discount factor for the state evolution of* $\delta = 0.95$ (*left*) *and* $\delta = 0.75$ (*right*).

density functions of t-distributions. Another issue is that, when δ is relatively high (0.95), the predictions provided by the DLM model are smoother but not entirely accurate. Conversely, when δ is smaller (0.75), the predictions are less smooth, indicating that not much learning from data occurs with concerns of overfitting. The summary shown in Table 2 indicates that the DLMs do not perform as well as the PGSS family of models based on MAPE, MAD, sMSD, and log-marginal likelihood metrics. Their performance is shown to be better in only one (MSE) out of the five performance metrics. In this numerical example the count values are small (most are in the range of 0–3). We believe that the noninteger predictions obtained using the DLMs are producing slightly favorable MSE estimates when compared to those obtained using the integer-valued predictions of the negative binomial predictive densities. One surprising finding is that the PGSS-deterministic model provides the best performance. This can be attributed to the stableness of the web traffic that mostly fluctuates between 1–5.

7. Discussion. In this paper we introduced strategies for modeling, monitoring, and forecasting web traffic data from a simulated study and the Fox News website where the flows between pages are observed in high-frequency time intervals. We analyzed two sets of flows, which we termed the high-count and low-count examples. In doing so, we considered the family of Poisson gamma state space models that have been used to model count-valued time series in the literature. In addition, we considered an extension of the PGSS class of models where the dynamic discount factor whose temporal evolution is modeled with an autoregressive process. Modeling the discount factor as a dynamic state variable is methodologically novel, as current approaches treat the discount factor as either a fixed tuning parameter, a random (static) parameter, or a deterministically time-varying quantity. We discussed how the sa-PGSS models can be estimated using particle based methods designed for fast sequential learning, monitoring, and forecasting and showed its computational superiority against MCMC based methods. More specifically, we developed a particle learning algorithm that harnesses closed-form conditional sufficient statistics to rapidly estimate dynamic state variables and static parameters. We found that the PL algorithm is $\sim$250 times faster than comparable MCMC methods when the time series has 100 observations (see Table 1). Our experiments showed that, as the time series lengthens, the relative speed gap between our PF algorithm and MCMC significantly widens.

In our analysis of web traffic data, we argued that any approach considered for modeling should possess three main features. First, the model should be able to accurately quantify the uncertainty exhibited by web traffic data (count-valued). Second, the model should be able to provide fast and efficient monitoring and prediction mechanisms in a sequential manner, suitable for high frequency time intervals (sequential). Third, the model should be able to rapidly adapt to sudden bursts or structural changes while preserving computational feasibility (state shifting, computational feasibility). These features are summarized on Table 3 where sa-PGSS, PGSS, DLM, and Markov switching (MS) models are shown. In summary, our experiments with the Fox News data showed that the current class of PGSS models (with

TABLE 3
List of web traffic models

Models	Count-valued	Sequential	State shifting	Computational Feasibility
sa-PGSS	✓	✓	✓	✓
PGSS	✓	✓	✗	✓
DLM	✗	✓	✗	✓
MS	✗	✗	✓	✗

static and random discount factors) are unable to quickly adapt to sudden shifts and state changes. The random-PGSS model provides comparable results in terms of forecasting, when there are no sudden shifts in data, but does not offer any significant computational advantage over the sa-PGSS model. The PGSS models with deterministic γ (both static and dynamic) consistently provide the worst forecasting performance in scenarios where web traffic exhibit fluctuations and state shifts. Furthermore, the DLMs are not only unable to accurately quantify the uncertainly exhibited by web traffic data, but they also fail to yield competitive forecasting performance with respect to the class of PGSS models. In our view, Markov switching models with unknown number of states (as the data is observed sequentially) can certainly provide the required level of adaptability to state shifts but are simply not suitable for online learning, monitoring, and forecasting web traffic data in high-frequency time intervals (30 second time intervals in our analysis), due to their computational burden in sequential settings.

Even though we focused on the analysis of web traffic data, it is important to note that our approach is general and can be applied to many other settings where monitoring and forecasting of relatively high-frequency count data with potential bursts are of interest. For instance, structural changes in count-valued time series are pervasive in the digital economy. Demand for Uber rides may exhibit a sudden burst at uncommon times, due to the conclusion of a sporting event or concert. Surges in web traffic on Facebook, Instagram, Twitter, and Google—while often due to predictable intraday variations—are occasionally driven by unanticipated news events. At call centers and online help desks for insurance companies, unexpected natural disasters and severe weather may result in dramatic increases in the number of customers requiring service. Many of these e-commerce platforms continually allocate resources in a sequential manner to meet consumer demand. When business operations of these web platforms depend on accurate short-term predictions and monitoring of consumer demand, the ability to quickly identify structural breaks and adjust forecasts of customer counts is critical. Our proposed family of PGSS models can be a reasonable alternative for monitoring and forecasting of sequential demand data typically observed in ridesharing economies (Uber, Lyft), online advertising (Facebook, Google, ...), customer call centers (Liberty Mutual, GEICO), and rapid-delivery online retailing (Amazon's Prime Now and Fresh services), among others. In all of these examples, resource allocation decisions are made on short time intervals, and balancing model complexity and computational speed is imperative.

Many modern applications involve analysis of multiple time series that exhibit auto and cross-sectional correlations. For example, Uber rides requested at nearby locations likely exhibit rich temporal and cross-series structure. Not only are the time series of pick-up requests spatially related, but the pick-up locations themselves may have defining characteristics that explain variation in the number of requests. These applied challenges call for a multivariate extension of the sa-PGSS model that includes covariates; however, extending the sa-PGSS model to a multivariate setting with covariates presents significant technical difficulties beyond the scope of our current paper. While we recognize the current limits of the sa-PGSS model, we believe that it offers significant promise for scaling online learning, monitoring, and forecasting of count data to higher dimensions.

APPENDIX: MARKOV CHAIN MONTE CARLO ALGORITHM FOR THE SA-PGSS MODEL

In what follows, we present a summary of steps of the MCMC algorithm that is an alternative for the proposed PF algorithm. The goal is to generate samples from the full joint posterior distribution of state as well as static parameters, $p(\theta_{1:t}, \gamma_{1:t}, \vartheta | \mathcal{D}_t)$, in a sequential manner. This can be achieved via the following steps:

1. Sampling $\theta_{1:t}$

Given $\gamma_{1:t}$ and ϑ, sampling from the conditional posterior $p(\theta_{1:t}|\gamma_{1:t}, \vartheta, \mathcal{D}_t)$ can be done by forward filtering and backward sampling. First, we compute $(a_{1:t}, b_{1:t})$ by forward filtering. Next, we sample from $\theta_t \sim \mathrm{Ga}(a_t, b_t)$. Recursively, at each $s < t$, we sample θ_s, based on the distributional relation $\theta_s = \gamma_s\theta_{s+1} + \mathrm{Ga}((1-\gamma_s)a_s, b_s)$.

2. Sampling ϑ

The conditional posterior of $p(\vartheta|\theta_{1:t}, \gamma_{1:t}, \mathcal{D}_t)$ is given in Section 4 where the normal-inverse gamma distribution for the transformed parameters are shown. Same approach can be followed here.

3. Sampling $\gamma_{1:t}$

This is the hardest part of the MCMC algorithm to implement. We take the single-mover sampler approach and consider the sampling of each γ_s for $s = 1{:}t$. The conditional posterior is written as (e.g., for $0 < s < t$)

$$p(\gamma_s|\theta_{1:t}, \gamma_{1:t\setminus s}, \vartheta, \mathcal{D}_t) \propto p(g_{s+1}|g_s, \vartheta)p(g_s|g_{s-1}, \vartheta)\prod_{u=s}^{t} p(\theta_u|\theta_{u-1}, \gamma_{1:u}, \mathcal{D}_{u-1}), \tag{20}$$

where the transition density of states is that of the scaled-beta distribution,

$$\begin{aligned} &p(\theta_u|\theta_{u-1}, \gamma_{1:u}, \mathcal{D}_{u-1}) \\ &\quad= \frac{1}{\mathrm{Be}(\gamma_u a_{u-1}, (1-\gamma_u)a_{u-1})} \\ &\qquad\times \left(\frac{\gamma_u}{\theta_{u-1}}\right)^{\gamma_u\alpha_{u-1}} \theta_u^{\gamma_u\alpha_{u-1}-1}\left(1 - \frac{\gamma_u}{\theta_{u-1}}\theta_u\right)^{(1-\gamma_u)\alpha_{u-1}-1} \end{aligned} \tag{21}$$

Note that γ_s is involved implicitly in $p(\theta_u|\theta_{u-1}, \gamma_{1:u}, \mathcal{D}_{u-1})$ for not only $u = s$ but also $u > s$ through the sufficient statistics α_u, that is, sequentially updated by, for example, $\alpha_{s+1} = \gamma_s\alpha_s + N_s$.

The sampling from equation (20) is the key in the implementation of the MCMC algorithm. The common approach is to use a random-walk Metropolis–Hastings step where t tuning parameters are required in addition to many iterations, making this is an unattractive solution. As an alternative, an independent Metropolis–Hastings step with a Gaussian proposal density can be considered which requires the computation of the gradient and the Hessian of the density in (20) in the log scale. To speed up the estimation, we propose to sample from

$$q(g_s) \propto p(g_{s+1}|g_s, \vartheta)p(g_s|g_{s-1}, \vartheta)$$

and accept the generated particle g_t^{new} with acceptance probability

$$P\big[g_s^{\mathrm{old}} \to g_t^{\mathrm{new}}\big] = \max\left\{1, \prod_{u=s}^{t} \frac{p(\theta_u|\theta_{u-1}, \gamma_{1:u\setminus s}, \gamma_s^{\mathrm{new}}, \mathcal{D}_{u-1})}{p(\theta_u|\theta_{u-1}, \gamma_{1:u\setminus s}, \gamma_s^{\mathrm{old}}, \mathcal{D}_{u-1})}\right\}.$$

Acknowledgments. The authors would like to thank the Editor, the Associate Editor, and the anonymous referees for helpful comments.

Funding. The first author was supported by the Japan Society for the Promotion of Science (JSPS KAKENHI) grant number 17K17659.

REFERENCES

AKTEKIN, T., POLSON, N. and SOYER, R. (2018). Sequential Bayesian analysis of multivariate count data. *Bayesian Anal*. **13** 385–409. MR3780428 https://doi.org/10.1214/17-BA1054

AKTEKIN, T. and SOYER, R. (2011). Call center arrival modeling: A Bayesian state-space approach. *Naval Res. Logist*. **58** 28–42. MR2796402 https://doi.org/10.1002/nav.20436

AKTEKIN, T., SOYER, R. and XU, F. (2013). Assessment of mortgage default risk via Bayesian state space models. *Ann. Appl. Stat*. **7** 1450–1473. MR3127954 https://doi.org/10.1214/13-AOAS632

BERRY, L. R., HELMAN, P. and WEST, M. (2020). Probabilistic forecasting of heterogeneous consumer transaction–sales time series. *Int. J. Forecast*. **36** 552–569.

BERRY, L. R. and WEST, M. (2020). Bayesian forecasting of many count-valued time series. *J. Bus. Econom. Statist*. **38** 872–887. MR4154894 https://doi.org/10.1080/07350015.2019.1604372

CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553. MR1311096 https://doi.org/10.1093/biomet/81.3.541

CARVALHO, C. M., JOHANNES, M. S., LOPES, H. F. and POLSON, N. G. (2010a). Particle learning and smoothing. *Statist. Sci*. **25** 88–106. MR2741816 https://doi.org/10.1214/10-STS325

CARVALHO, C. M., LOPES, H. F., POLSON, N. G. and TADDY, M. A. (2010b). Particle learning for general mixtures. *Bayesian Anal*. **5** 709–740. MR2740154 https://doi.org/10.1214/10-BA525

CHEN, X., BANKS, D. and WEST, M. (2019). Bayesian dynamic modeling and monitoring of network flows. *Netw. Sci*. **7** 292–318.

CHEN, X., IRIE, K., BANKS, D., HASLINGER, R., THOMAS, J. and WEST, M. (2018). Scalable Bayesian modeling, monitoring, and analysis of dynamic network flow data. *J. Amer. Statist. Assoc*. **113** 519–533. MR3832205 https://doi.org/10.1080/01621459.2017.1345742

DAVIS, R., HOLAN, S., LUND, R. and RAVISHANKER, N. (2015). *Handbook of Discrete-Valued Time Series*. CRC Press/CRC, Boca Raton.

DOORNIK, J. A. (2007). *Object-Oriented Matrix Programming Using Ox*, 3rd ed. Timberlake Consultants Press and Oxford, London.

FEARNHEAD, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. *J. Comput. Graph. Statist*. **11** 848–862. MR1951601 https://doi.org/10.1198/106186002321018821

FREELAND, R. K. and MCCABE, B. P. M. (2004). Analysis of low count time series data by Poisson autoregression. *J. Time Series Anal*. **25** 701–722. MR2089191 https://doi.org/10.1111/j.1467-9892.2004.01885.x

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal*. **15** 183–202. MR1263889 https://doi.org/10.1111/j.1467-9892.1994.tb00184.x

FRÜHWIRTH-SCHNATTER, S. and WAGNER, H. (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika* **93** 827–841. MR2285074 https://doi.org/10.1093/biomet/93.4.827

GAMERMAN, D., REZENDE DOS SANTOS, T. and FRANCO, G. C. (2013). A non-Gaussian family of state-space models with exact marginal likelihood. *J. Time Series Anal*. **34** 625–645. MR3127211 https://doi.org/10.1111/jtsa.12039

GLYNN, C., TOKDAR, S. T., HOWARD, B. and BANKS, D. L. (2019). Bayesian analysis of dynamic linear topic models. *Bayesian Anal*. **14** 53–80. MR3910038 https://doi.org/10.1214/18-BA1100

GORDON, N. J., SALMOND, D. J. and SMITH, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F* (*Radar and Signal Processing*) **140** 107–113. IET.

GRAMACY, R. B. and POLSON, N. G. (2011). Particle learning of Gaussian process models for sequential design and optimization. *J. Comput. Graph. Statist*. **20** 102–118. MR2816540 https://doi.org/10.1198/jcgs.2010.09171

HARVEY, A. C. and FERNANDES, C. (1989). Time series models for count or qualitative observations. *J. Bus. Econom. Statist*. **7** 407–417.

LOPES, H. F. and POLSON, N. G. (2016). Particle learning for fat-tailed distributions. *Econometric Rev.* **35** 1666–1691. MR3511035 https://doi.org/10.1080/07474938.2015.1092809

LOPES, H. F. and TSAY, R. S. (2011). Particle filters and Bayesian inference in financial econometrics. *J. Forecast*. **30** 168–209. MR2758809 https://doi.org/10.1002/for.1195

PITT, M. K. and SHEPHARD, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc*. **94** 590–599. MR1702328 https://doi.org/10.2307/2670179

PRADO, R. and LOPES, H. F. (2013). Sequential parameter learning and filtering in structured autoregressive state-space models. *Stat. Comput*. **23** 43–57. MR3018349 https://doi.org/10.1007/s11222-011-9289-1

PRADO, R. and WEST, M. (2010). *Time Series*: *Modeling*, *Computation and Inference*. CRC Press, Boca Raton.

SINGPURWALLA, N. D., POLSON, N. G. and SOYER, R. (2018). From least squares to signal processing and particle filtering. *Technometrics* **60** 146–160. MR3804244 https://doi.org/10.1080/00401706.2017.1341341

SMITH, R. L. and MILLER, J. E. (1986). A non-Gaussian state space model and application to prediction of records. *J. Roy. Statist. Soc. Ser. B* **48** 79–88. MR0848053

STORVIK, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.* **50** 281–289.

UHLIG, H. (1994). On singular Wishart and singular multivariate beta distributions. *Ann. Statist.* **22** 395–405. MR1272090 https://doi.org/10.1214/aos/1176325375

UHLIG, H. (1997). Bayesian vector autoregressions with stochastic volatility. *Econometrica* **65** 59–73. MR1433685 https://doi.org/10.2307/2171813

WEST, M. and HARRISON, P. J. (1986). Monitoring and adaptation in Bayesian forecasting models. *J. Amer. Statist. Assoc.* **81** 741–750.

WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1482232