PA

Polls, Context, and Time: A Dynamic Hierarchical Bayesian Forecasting Model for US Senate Elections

Yehu Chen[1], Roman Garnett[2] and Jacob M. Montgomery[3]

[1] *Division of Computational and Data Sciences, Washington University in St. Louis, St. Louis, MO, USA.*
E-mail: chenyehu@wustl.edu
[2] *Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO, USA.*
E-mail: garnett@wustl.edu
[3] *Department of Political Science, Washington University in St. Louis, St. Louis, MO, USA.*
E-mail: jacob.montgomery@wustl.edu

Abstract

We present a hierarchical Dirichlet regression model with Gaussian process priors that enables accurate and well-calibrated forecasts for U.S. Senate elections at varying time horizons. This Bayesian model provides a balance between predictions based on time-dependent opinion polls and those made based on fundamentals. It also provides uncertainty estimates that arise naturally from historical data on elections and polls. Experiments show that our model is highly accurate and has a well calibrated coverage rate for vote share predictions at various forecasting horizons. We validate the model with a retrospective forecast of the 2018 cycle as well as a true out-of-sample forecast for 2020. We show that our approach achieves state-of-the art accuracy and coverage despite relying on few covariates.

Keywords: forecasts, Bayesian, Gaussian process, elections

1 Predicting Senate Races

In recent years, there has been an explosion of interest in election prediction models. Primarily, this has been driven by media outlets and popular forecasting websites like fivethirtyeight.com. However, although the most prominent forecasting efforts have been housed in media organizations, these models often build from, or are inspired by, research in political science.

Broadly speaking, academic election forecasting in the U.S. context can be divided into two approaches. First, there are static models that make a single prediction for a given election (e.g., Abramowitz 2008; Fair 1978; Lewis-Beck and Tien 2008). These models are sometimes referred to as "fundamentals" models and primarily rely on economic indicators, incumbency status, and other factors that shape the general context of an election. To the extent, they incorporate polling data at all, they are based on proxies such as presidential approval (e.g., Erikson and Wlezien 2008) or snapshots taken well before Election Day (e.g., Campbell and Wink 1990).

A smaller body of research focuses on dynamic forecasting models that change over the course of the campaign as new polling data arrives. In particular, Linzer (2013) introduces a dynamic Bayesian model forecasting the U.S. presidential election results for all 50 states. This model has served as a basis for presidential forecasts produced by major media outlets including *The Economist* and *Daily Kos*.[1] Another example is Jackman (2005), which presents a somewhat related Bayesian model, although the goal is to aggregate polls rather than to make a prediction *per se*.

While the U.S. presidential election has received the most attention, a smaller body of research has focused on predicting legislative elections. Most examples in this domain do not seek to

1 While *The Economist* presidential model was based on Linzer (2013), their senate forecasting model was not. The details of *The Economist* 2020 U.S. Senate model are not public.

Political Analysis (2023)
vol. 31: 113–133
DOI: 10.1017/pan.2021.42

Published
18 January 2022

Corresponding author
Jacob M. Montgomery

Edited by
Jeff Gill

predict individual races, but rather the aggregate number of seats that swing to a specific party (e.g., Campbell 2018; Lockerbie 2012). Still, some previous work has built models of individual U.S. Senate races based on fundamental factors (Hummel and Rothschild 2014; Klarner 2008; Klarner 2012; Klarner and Buchanan 2006). However, to the best of our knowledge, there are no published models providing dynamic forecasts of individual senate races in political science.

The relative lack of attention to the Senate is probably the result of the intense popular interest in the presidential race. However, it also reflects the fact that predicting individual-level Senate elections is actually a more difficult task than it may first appear. To begin, there is far less polling data for any given Senate election relative to national races, especially early in the cycle. Some states with close races and large media markets may have dozens of polls, but in many others there are very few. In addition, senate races are relatively low salience to voters, especially early in the cycle. Even strong challengers can be unfamiliar to voters until the final weeks. As a consequence, public opinion can be far more dynamic as voters learn about their options in the lead-up to Election Day. In short, senate elections offer fewer polls and what polls exist can be noisy predictors.

A further problem is that local context and other "fundamentals" are often only weakly predictive of candidate performance. Although the general partisan dispositions in each state tends to heavily structure presidential outcomes, results in Senate elections are far less geographically determined. That is, knowing how a party candidate performed in one election is often a poor predictor of performance in subsequent years. A recent examples would be West Virginia where Democrat Joe Manchin won over 60% of the vote in 2012 and Republican Shelly Capito won over 62% just 2 years later. These kinds of dramatic partisan swings occur regularly, making "fundamentals" forecasts difficult. Klarner notes that his fundamentals-based model "has never performed well, being off by three seats in 2006 and five seats in 2008. U.S. Senate elections appear to be influenced by race-specific factors that are difficult to include in forecasting models" (Klarner 2013, p. 45). Meanwhile, Hummel and Rothschild (2014) predicted only 83% of races correctly in-sample and performed similarly out-of-sample.

In combination, this means that for any single cycle it is difficult to provide accurate forecasts in the absence of polling. Yet, polling data itself is relatively sparse and subject to significant trends over the course of the election. And, of course, relying on unvarnished polling data can be inaccurate even where it is not just missing, making simple polling averages suboptimal.

On the other hand, there is one very important advantage to working in this setting relative to national elections; *there are many more observations*. While presidential elections offer only one observation every 4 years, the Senate has roughly 33 election outcomes attached to hundreds of polls every 2 years. In our dataset, which covers only the post-1992 period, we have 501 election results and over 7,900 published polls. This give us some hope that we can train a model that can learn from the past to predict future outcomes and, crucially, correctly calibrate our uncertainty.

Below, we present a hierarchical Dirichlet regression model in a Bayesian framework that enables us to combine polls and fundamentals to accurately forecast election outcomes at various time horizons. This model provides a structured balance between time-dependent opinion polls and state/candidate-level fundamentals. Unlike fundamentals-based models, ours updates throughout the election cycle to reflect recent polling trends. Yet unlike existing dynamic models, ours is trained on a set of historical election outcomes rather a single election cycle. The result is a model that provides uncertainty estimates that arise naturally from the induced posterior based on historical data and therefore provide a better sense of our true uncertainty. Experiments show that our model can achieve high levels of accuracy and correct coverage for various forecasting horizons.

The most important contribution we make in this article is proposing an accurate, dynamic model appropriate for subnational elections at the district level—something the discipline currently lacks.[2] However, we also advance the broader elections forecasting literature in two ways. First, the hierarchical structure we propose combines the unique strengths of poll-based dynamic models and fundamentals-based static models in a single framework. Existing dynamic models are fit to polling data from a single election cycle (Jackman 2005; Linzer 2013). To the extent historical data are used at all, they enter only through informative priors or hyperparameter selection. This makes it more difficult to understand whether the final predictive intervals accurately represents our uncertainty about the outcomes since the likelihood itself is actually fit to polls and not to election results. Our approach differs in that the dynamic component feeds into a higher-level model trained on historical election results. Since this higher-level model also includes fundamental factors (e.g., the partisan orientation of the state), final predictions are better calibrated to reflect our actual uncertainty about unobserved election outcomes and can weight polling and fundamentals to reflect their actual predictive performance at different time horizons.

Second, we introduce a Gaussian process (GP) framework for modeling trends in latent public opinion that is more appropriate for elections with fewer polls—a common feature outside of U.S. presidential races. Our GP approach offers a significant advantage in that we can model polling trends as a linear process where nonlinear deviations are *allowed* given sufficient data. This added structure offers a significant improvement in out-of-sample prediction relative to a random walk (Linzer 2013) while also relaxing strict linearity assumptions when needed. It has the further advantage that it allows us to derive the posterior for these time trends analytically, significantly reducing computation time for any one election. This in turn allows us to fit the full hierarchical model including hundreds of elections.

In the next section, we provide a basic intuition for our modeling framework before providing a more detailed presentation in Section 3. We then test the model using historical data in Section 4 and evaluate a true out-of-sample forecast for the 2020 election cycle in Section 5. We show that our approach achieves state-of-the art accuracy and coverage despite relying on few covariates. We conclude with a discussion of how our model could be improved in future iterations or adjusted for other election settings.

2 Intuition and Related Work

Before introducing the model, we want to focus on the core ideas that inform our approach. First, we suppose that polling for a candidate is a noisy measure of true underlying public opinion, $f(t)$, at any given time t. That is, we assume that there is a true level of underlying support for each candidate that moves smoothly over time and that polling results imperfectly follow these trends.

While modeling smooth latent public opinion is consistent with previous efforts to aggregate polls (Jackman 2005; Linzer 2013; Stoetzer *et al.* 2019), we adopt a strategy that is more appropriate given the sparseness of polling in many senate elections. Our approach assumes a linear trend in the data with mild nonlinear deviations. This provides a sensible compromise between a simple linear model of public opinion and the trend-free smoothing procedures adopted in Jackman (2005) and Linzer (2013) (see also Stoetzer *et al.* 2019; Walther 2015). Indeed, these other approaches can be viewed as special cases of our more general model where no linear trend is included.

Third, our modeling strategy assumes that latent public opinion is only one predictor of election outcomes. That is, latent public opinion is not assumed to translate directly into election outcomes

2 Media outlets like fivethirtyeight.com and *The Economist* provide dynamic predictions. However, the details of these models are not public and we have no way to assess their methodology or build upon their techniques.

as in Linzer (2013). Instead, the model learns the degree to which public opinion accurately predicts elections relative to other "fundamental" factors including state-level voting history, candidate quality, and the like. This approach has two advantages. To begin, it allows us to easily train our model at different time horizons such that public opinion is weighted more heavily as the election approaches and polling becomes more predictive. More fundamentally, however, it allows us to explicitly model the inherent uncertainty in election outcomes that cannot be adequately predicted from polling and and contextual factors. That is, we assume that even if we knew public support for a candidate perfectly, there would still be uncertainty in the outcome due to turnout and other unmodeled factors. Our aim is to use historical data to calibrate our uncertainty and achieve correct coverage rates at various time horizons in a way that reflects this irreducible uncertainty.

Finally, the model is tuned to accurately predict elections not polls. Thus, while polling outcomes are included in the model, the key model parameters are not selected to reduce the error in predicting polls but in predicting vote share. We select hyperparameters intentionally that underpredict individual polling results but provide a better basis for predicting candidate vote share. The result is a parsimonious, but accurate and well calibrated model of elections.[3] The model takes in only four variables: polling data, Cook's partisan voting index (Campbell 2018; MacWilliams 2015), party affiliation of candidates, and candidate quality (Jacobson 1989; Jacobson and Carson 2019). However, it still makes accurate predictions for races at various time horizons while maintaining correct coverage. Indeed, in the 2020 Election our model outperformed the model published in *The Economist* (Economist 2020) and provided comparable (and by some metrics superior) performance to the popular *fivethirtyeight.com* forecasts (Fivethirtyeight 2020).

In the next section, we introduce the model in stages. Section 3.1 provides important background information on Gaussian process regression, an approach that has appeared rarely in political science research. Section 3.2 then applies this framework to the task of projecting latent public support for each candidate. Section 3.3 then explains how this is combined with contextual factors in our Dirichlet regression model of vote share. We then briefly contrast our approach with other forecasting models in the literature in Section 3.4 before turning to our results in Section 4.

3 A Predictive Model of U.S. Senate Elections

Our proposed model has two components as depicted in Figure 1. First, we use candidate-level polling data to predict latent public support for candidate i on Election Day ($t = 0$), which we denote as $f_i(0)$.[4] If we are predicting this before the election ($t < 0$), this quantity is predicted based on all polling data up to the current date as well as an informative prior reflecting the general electoral context. Note that the goal is not to create a point prediction, but to estimate a distribution on $f_i(0)$ that reflects our uncertainty about the trajectory of public opinion over the course of the election as well the inherent uncertainty in polling data itself. We refer to this as our *candidate-level model*.

Second, we then use predicted public support as inputs for an *election-level model*[5] with the goal of predicting the proportion of the vote divided among all candidates in a given race (that is, the entire vote share and not only the winner). We model this with a Dirichlet regression with year-level random effects using a training dataset of elections starting in 1992. Importantly, this Dirichlet regression takes in $f_i(0)$ as an *input* along with contextual factors. Thus, we are able to use historical data to estimate the the degree to which electoral context, public opinion, or some mix of the two are best able to predict vote shares at different time horizons.

3 The model was built using data from 1992–2016 with cross validation. All modeling decisions and hyper-parameter selection was done using only these data. We held out 2018 to serve as a test set for this analysis.
4 Supplementary Appendix A provides a reference table of all notation.
5 Throughout, we use *election* to refer to a race between two or more candidates in a single seat. The overall election cycle (roughly 33 elections in a given year) is referred to as a cycle.

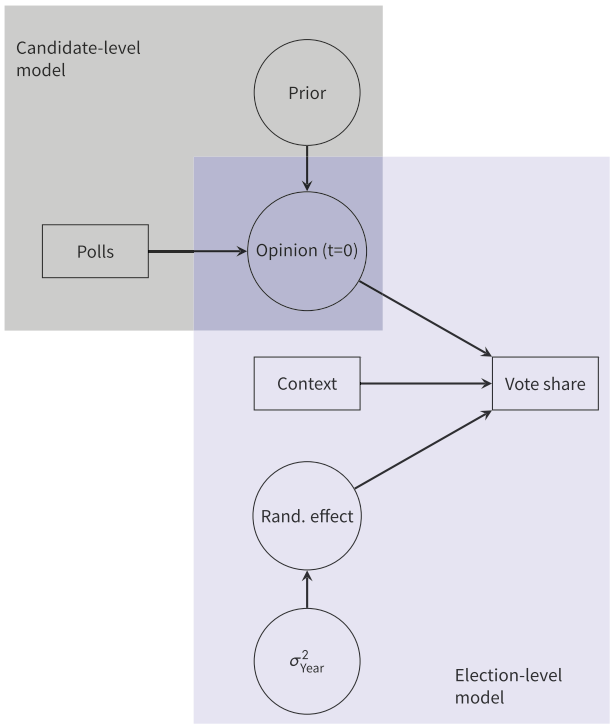

Figure 1. Conceptual outline of the two stage model. The candidate-level model, predicts public opinion on Election Day. The election-level model, predicts vote share as a function of public opinion and contextual factors.

The final output is a prediction for Senate elections that accounts for two levels of uncertainty. We have uncertainty about where latent public opinion *will be* on Election Day given the polling data we have observed so far. But we also have uncertainty as to how well public opinion and contextual factors predict election outcomes based on historical data.

3.1 Background on Gaussian Process Regression

Our model for latent public opinion over time is a linear trend with smooth nonlinear deviations. Here, we subsume both components into a single GP model of latent opinion. GPs offer a flexible Bayesian framework for nonlinear regression widely adopted in machine learning (Rasmussen and Williams 2006). GP models have not been used widely in political science, although they have appeared under the label Bayesian kriging (Gill 2020; Monogan and Gill 2016). However, mathematically they can be considered a Bayesian variant of kernal regularized least squares (KRLS) (Hainmueller and Hazlett 2014; Mohanty and Shaffer 2019).

To define a GP, consider a function $f : \mathcal{X} \rightarrow \mathbb{R}$ on some arbitrary domain $\mathcal{X}$; for our model of latent opinion, $\mathcal{X} = (-\infty, 0]$ is the span of times at which we may wish to predict. The defining property of a Gaussian process is that if $\mathbf{X} \subset \mathcal{X}$ is a finite vector of input locations, then the associated function values $f(\mathbf{X})$ has a multivariate normal distribution. The moments of this distribution are provided by a *mean function* $\mu(x) = \mathbb{E}[f(x) \mid x]$ and covariance function $K(x, x') = \text{cov}[f(x), f(x') \mid x, x']$; evaluating these pointwise provides the mean vector and covariance matrix for any desired vector of function values $f(\mathbf{X})$. Modeling with the GP entails designing the mean and covariance functions to encoding the desired statistical properties of f such as correlations over the domain.

A critical property of GPs is that they enable exact, closed-form inference for regression for observations corrupted by additive Gaussian noise. Let $f \sim \mathcal{GP}(\mu, K)$ have a GP prior and suppose we obtain a vector of observed values $\mathbf{y}$ at locations $\mathbf{X}$, where $y_i = f(x_i) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$. Then the

posterior belief of f given $D = (\mathbf{X}, \mathbf{y})$ is again a GP with updated mean and covariance function:

$$\mu_{f|D}(x) = \mu(x) + K(x, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mu(\mathbf{X}))$$
$$K_{f|D}(x, x') = K(x, x') - K(x, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I})^{-1}K(\mathbf{X}, x').$$

Hence, appealing to the definition above, the posterior predictive distribution of any function value $f(x^*)$ is normal:

$$f(x^*) \mid D, x^* \sim \mathcal{N}(\mu_{f|D}(x^*), K_{f|D}(x^*, x^*)).$$

This final point is important for our application. When modeling the latent opinion with a Gaussian process, our prediction for latent public opinion on Election Day, $f_i(0)$, is a normal distribution that can be directly derived. This in turn becomes a normal prior for public opinion that is passed directly into the election-level model. This allows us to include our uncertainty about where public opinion will be on election day into the *election-level* model while at the same time significantly reducing computation time relative to Linzer (2013).

3.2 Projecting Public Support via GP Regression

We next outline our approach for forecasting voter preferences throughout an election given polling results. Our approach entails building independent GP models for each race conditioned on available polling outcomes.[6] The model includes only polls, hyperparameters, and priors (discussed below).

Denote by C the set of all candidates in all races we wish to reason about. We will consider the unknown proportion of voters preferring candidate $i \in C$ in a given race a function of time, writing $f_i : (-\infty, 0] \to [0, 1]$. Here, the domain of the function is time (measured in days), where the election is defined to occur at time $t = 0$ days. Let $\mathcal{T}_i$ be the set of times when opinion polls for candidate i were conducted.[7]

We model the trend of voter preferences f_i as a sum of an underlying linear trend $a_i + b_i t$, with smooth nonlinear deviations from this trend, $\eta_i(t)$. We place independent Gaussian priors on the intercept a_i (i.e., the prior mean of the voter preferences on Election Day) and slope b_i of the linear trend, and will place an independent, zero-mean GP prior on the nonlinear component η_i. The covariance function K determines the correlation of deviations from the linear trend as a function of time and was taken to be identical across all races. Here, we used the Matérn covariance function with $\nu = \frac{3}{2}$, which models isotropic, once-differentiable functions (Rasmussen and Williams 2006). This covariance function has two hyperparameters that we will estimate from training data: a length scale ρ determining the scale of correlations, and an output scale λ determining the pointwise variance of the process. Intuitively, we can think of ρ as determining the "window" of days over which nonlinear deviations are estimated and λ as controlling the degree of nonlinearity we expect such that higher values lead to more dramatic deviations.

The model can be summarized as:

$$f_i(t) = a_i + b_i t + \eta_i(t) \quad (1)$$
$$a_i \sim \mathcal{N}(\bar{a}_i, \sigma_a^2 = 0.1^2) \quad (2)$$
$$b_i \sim \mathcal{N}(0, \sigma_b^2 = 0.002^2) \quad (3)$$
$$\eta_i(t) \sim \mathcal{GP}(0, K), \quad (4)$$

6 Supplementary Appendix B shows that Senate election results are far less correlated across states than presidential elections. As we discuss in the conclusion, this independence assumption would need to be relaxed for presidential forecasting.

7 We date polls based on the first day they are in the field.

where the covariance function for the nonlinear deviations is

$$K(t,t';\rho,\lambda) = \lambda^2\left(1+\frac{\sqrt{3}d}{\rho}\right)\exp\left(-\frac{\sqrt{3}d}{\rho}\right); \qquad d = |t-t'|.$$

The priors on the linear parameters are constructed to be broad for the slope (so that over a time period of roughly 100 days the linear trend could plausibly assume any possible value) and vaguely informative for the intercept; we will discuss the intercept mean parameter $\bar{a}_i$ shortly.

The above prior choices induce the following joint prior over the voter preference, as shown in (5). Notice that our model provides an automatic marginalization over the linear slope parameters, since the covariance function in our GP model has absorbed the hyperparameters controlling the prior distribution of the linear function parameters.

$$f_i(t) \sim \mathcal{N}\Big(\mu_i(t), V(t,t')\Big), \tag{5}$$

$$\mu_i(t) = \bar{a}_i, \tag{6}$$

$$V(t,t') = \sigma_a^2 + tt'\sigma_b^2 + K(t,t'). \tag{7}$$

Our goal is to infer the latent voter preference trend from opinion poll outcomes, which are by their nature noisy. Our approach is to model the observed poll outcomes as binomially distributed, then approximate each binomial with a Gaussian for mathematical convenience. This will allow closed-form exact inference, yielding a posterior GP belief about underlying voter trends conditioned on available data. As discussed in Section 3.4, this step is an important innovation, allowing us to exactly solve for the posterior predictive distribution of $f_i(t)$.

For a candidate $i \in C$ with S_i conducted opinion polls, let $\mathcal{D}_i = \{t_{is}, n_{is}, x_{is}\}, (s = 1,\ldots,S_i)$ denote the outcomes of all available polls involving that candidate. Here, t_{is} is the time of the poll, n_{is} is the sample size of the poll, and x_{is} is the number of polled people expressing support for the candidate. Dropping subscripts momentarily, consider one such polling outcome $(t,x,n) \in \mathcal{D}$. We make the natural assumption that the number of supporters x is binomially distributed given the sample size n and the true (unknown) voter support f at time t:

$$x \sim \text{Binomial}\big(n, f(t)\big). \tag{8}$$

Unfortunately, it is not possible to condition a GP exactly on observations with a binomial likelihood. However, sample sizes for election polls tend to be large enough (often in the hundreds) that we can safely make a Gaussian approximation to the likelihood by moment matching. Here, we also explicitly consider an additional general noise term σ^2, which designates another level of noise stemming from the polling data. Let $p = x/n$ to be the observed proportion of support in a poll, so (8) could be approximated with

$$p \sim \mathcal{N}\big(f, \hat{p}(1-\hat{p})/n + \sigma^2\big), \tag{9}$$

where we have substituted the estimated $\hat{p}$ for the true unknown proportion $f(t)$ in the variance (in our case, $\hat{p} = p$ *after observation*). This likelihood is now conjugate to our GP prior on f and allows exact inference.

Let us define the vector $\mathbf{p}$ to entail a set of polling outcomes observed at times $\mathbf{t}$, $p_s = x_s/n_s$, and further define $\mathbf{B}$ to be a $S \times S$ diagonal matrix with $B_{ss} = p_s(1-p_s)/n_s + \sigma^2$. This is the approximate noise variance for each of these measurements appearing in (9). Using the results in Section 3.1,

the posterior predictive distribution of the voter preference at any time t is:

$$f(t) \mid D \sim \mathcal{N}\left(\mu_{f|\mathcal{D}}(t), K_{f|\mathcal{D}}(t,t)\right), \tag{10}$$

$$\mu_{f|\mathcal{D}_i}(t) = \mu(t) + V(t,\mathbf{t})\left(\mathbf{V}+\mathbf{B}\right)^{-1}(\mathbf{p}-\boldsymbol{\mu}), \tag{11}$$

$$K_{f|\mathcal{D}}(t,t') = V(t,t') - V(t,\mathbf{t})\left(\mathbf{V}+\mathbf{B}\right)^{-1}V(\mathbf{t},t'), \tag{12}$$

where $\boldsymbol{\mu}$ and $\mathbf{V}$ are the prior mean and covariance of $f(\mathbf{t})$, respectively. Although we may make forecasts for any time t, we are especially interested in public opinion on Election Day, $f(t=0)$. This will also be normal following Equation (10). For notational convenience below, we will again use subscripts for candidates and write $f_i(0) \sim \mathcal{N}(\mu_{f_i}, K_{f_i})$.

The candidate-level model is completed by choosing values for the intercepts $\{\bar{a}_i\}$ and the set of shared hyperparameters $\boldsymbol{\omega} = (\rho, \lambda, \sigma^2)$. Here, σ^2 represents the level of unmodeled noise remaining in the polling data, λ controls the degree to which the time trend deviates from linearity, and ρ represents the "bandwidth" of the smoothing window for these nonlinear deviations.

We chose informative, but wide hyperpriors for $\{a_i\}$ so that projections could be made in races with few or zero polls but that polling data would quickly swamp the prior when plentiful. Since the standard deviation for the hyperprior is set at 0.1, any vote share within ±30 points of the prior should be well supported.[8] To set $\{\bar{a}_i\}$, we ran a simple regression in the training set with normalized vote share as the dependent variable and party, lagged partisan vote index (PVI), and level or prior experience as covariates.[9] While not an accurate model by itself, it proved to be an adequate prior.[10]

For $\boldsymbol{\omega}$, we adopt a leave-one-year-out (loyo) cross-validation approach using the training period from 1992 to 2016. The motivation is to choose hyperparameters that maximize predictive performance for election results even at the expense of choosing parameters that reduce fit for the polling data.

First, we define the search region of output scale and shared noise both to be $[0, 0.05]$. We search length scale with a minimum of 7 days and a maximum of 56 days.[11] Empirically, we generate potential $\boldsymbol{\omega}$'s for the validation procedure from a low-discrepancy Sobol sequence (Sobol 1979) in the search region, since it covers the space more efficiently than a grid. We fit the *complete* model, including the election-level model, for each of 100 values of $\boldsymbol{\omega}$ at each time horizon (τ) leaving out each year in turn.

For example, for choosing the hyperparameters for the model predicting 4 weeks in advance of the election, we used all of the polling data up to day $t = -28$. We then fit the GP models and trained the election-level model described below leaving out each cycle in turn. We then generate out-of-sample predictions for vote shares and choose the hyperparameter setting that maximized the loyo log-likelihood averaged across all cycles.

8 Supplementary Appendix C considers alternative choices for the prior standard deviation σ_a. Using 2016 as a test case, we show that these alternatives do not improve predictive performance.

9 Letting w represent standard regression coefficients, this model was just

$$\text{vote share}_i \sim w_0 + w_1 \text{experienced}_i + w_2 \text{party}_i + w_3 \text{pvi}_i \times \text{party}_i,$$

where *party* was coded as 1 for Republicans, −1 for Democrats, and 0 otherwise. We lag PVI to the previous presidential cycle and experienced is a binary indicator for whether or not the candidate has ever previously held elected office. The entire cycle is left out of the training data when constructing these priors.

10 The few exceptions where the prior proved to be wildly off were for third-party candidates. Future efforts to forecast might create a separate prior structure for third-party candidates.

11 These ranges were determined from earlier exploratory work. Supplementary Appendix D considers an alternative cross-validation strategy and shows it has almost no effect on model predictions.

Table 1. Learned model hyperparameters from leave-one-year-out cross validation.

Time horizon (τ)	Length scale (ρ)	Output scale (λ)	Noise std (σ)
0	38.4	4.61%	2.89%
7	49.1	4.76%	4.61%
14	45.3	3.90%	3.28%
21	54.5	2.96%	4.84%
28	52.2	3.20%	3.67%
42	39.9	1.95%	4.92%
56	44.9	0.74%	3.48%

The chosen hyperparameters for each time horizon are shown in Table 1, and examples of the resulting candidate-level models for one candidate (John McCain in 2016) are shown at various time horizons in Figure 2. (Supplementary Appendix E shows another example for Democrat Katie McGinty [PA-2016].) This approach yields models that begin as linear far from Election Day but become increasingly nonlinear as τ approaches zero. Note also that the uncertainty in $f_i(0)$ narrows considerably in run up to Election Day.

3.3 Election-Level Model

The goal of the candidate-level model (Section 3.2) is to project forward at any time horizon a predictive distribution of latent public support on Election Day, $f_i(0)$. The election-level model in this section takes $f_i(0)$ as an input and combines it with additional contextual factors to generate a predictive distribution. Our method is based on Dirichlet regression, that allows prediction of the election vote shares for multiple candidates.[12]

In our setting, the vast majority of races involve only two credible candidates.[13] Indeed, in the 1992–2018 period, we included more than 2 candidates in only 11 elections.[14] However, we retain the Dirichlet presentation here as being more general and races with third parties can be critical in any given cycle.

Relying on the Dirichlet likelihood contrasts with some work in political science for multi-party elections, which builds on the logistic normal distribution (or t-distribution) applied to log-ratios of the votes (Katz and King 1999) or seemingly unrelated regression (Tomz, Tucker, and Wittenberg 2002). The primary criticism of Dirichlet regression is that it assumes that ratios of outcomes are independent (Aitchison 1982; Katz and King 1999; Philips, Rutherford, and Whitten 2016), which is unrealistic in more standard settings such as multi-party elections. However, while the outcome in U.S. Senate races is always compositional, the meaning of the categories do not correspond across races as these alternative models require. That is, the "third choices" are typically idiosyncratic to each race. So, for instance, the third-party candidate in the 2018 New Mexico race was Libertarian Gary Johnson while in the 2008 Minnesota race it was Independence Party candidate Dean Barkely. In other cases, even the major party candidate labels can be confused. So, for instance, in the 2012 Maine election Cynthia Dill was the official Democratic nominee while Independent Angus King garnered a significant amount of support from Democrats and caucused with the party once he

12 Note that the model we propose assumes that we know which candidates will be on the ballot. We discuss third-party choices below. However, for a typical race this requires that the winner of the major party primaries should be known either because the primary is over or the primary winner can be predicted with high level of certainty.

13 We include third-party candidates only when they are regularly included in public opinion polls in advance of the election. However, this criterion may be *ex ante* difficult to anticipate at the beginning of a cycle. We return to this point in our concluding discussion.

14 This includes races in Arizona (1992), Virginia (1994), Minnesota (2008), Alaska (2010 and 2016), Florida (2012), Maine (2012 and 2018), Maryland (2012), South Dakota (2014), and New Mexico (2018).

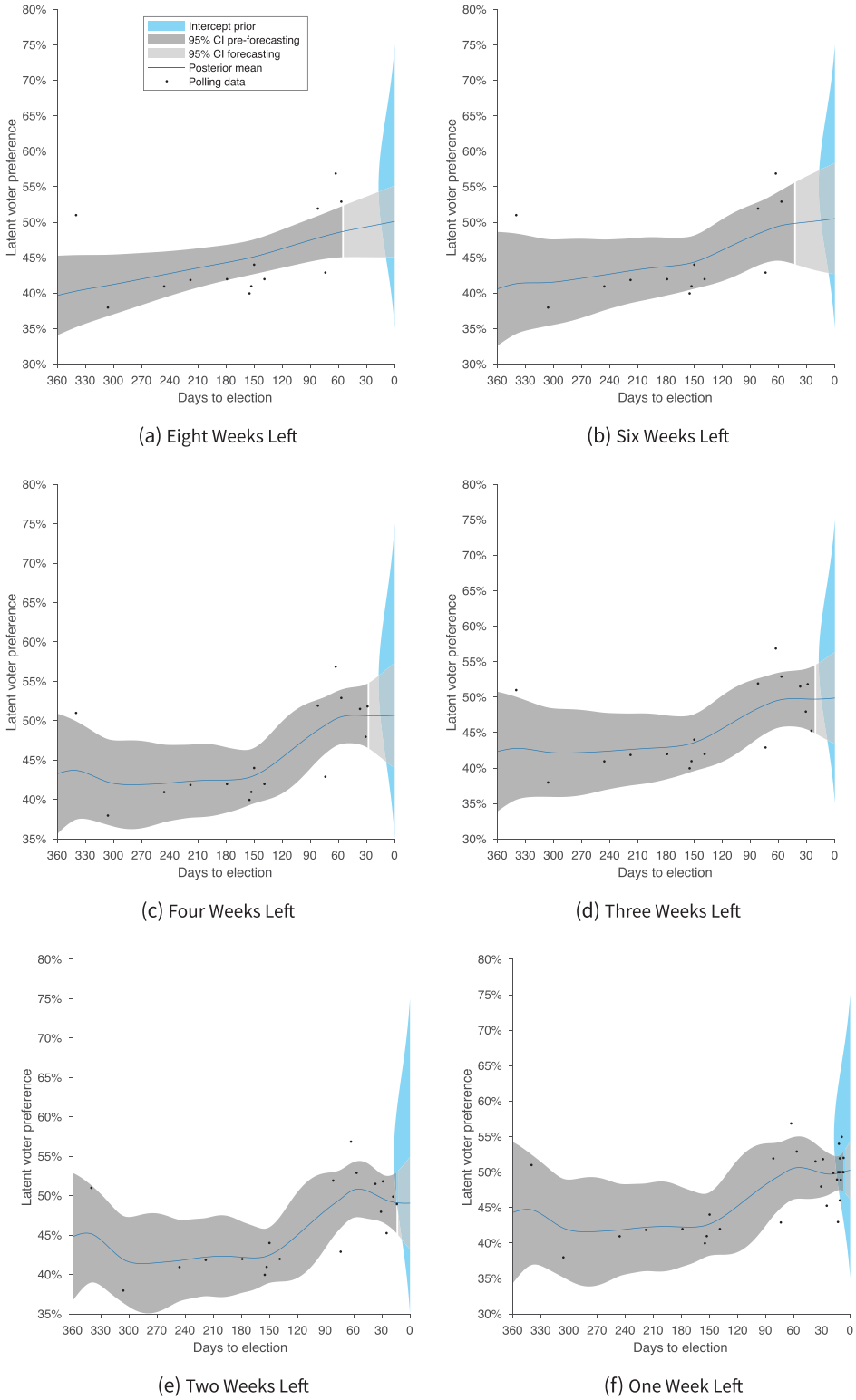

Figure 2. Voter Preference Estimate for John McCain in 2016 Arizona race at various time horizons. Points represent individual polls, the distribution on the right side of each panel is the prior, the dark gray region is the 95% confidence intervals for the estimated latent trend, and the light-gray region is the projected latent trajectory.

joined the senate. Indeed, in some instances the category meanings are unstable even when there are only two choices. For example, the 2016 California race featured two Democrats. We therefore

retain the Dirichlet regression approach despite the independence assumption since modeling the dependence between the choice categories is impossible when the choice set itself changes from observation to observation.[15]

We model the parameters in the Dirichlet distribution as a linear function of voter preferences and "fundamentals." This is similar in nature to other generalized linear models where a linear combination of terms is passed to a link function. In this case, each candidate is represented by a "concentration parameter," α_i, which we model as a linear combination of both $f_i(0)$ and other covariates. The unique feature of the Dirichlet regression is that each race is characterized by a *vector* of concentration parameters, $\boldsymbol{\alpha}$, where we have one α_i for each candidate. When the concentration parameter for candidate i is *relatively* large, she is expected to earn a higher proportion of the vote. Furthermore, the predictive density is more concentrated around this expected value when the individual components of α_j are large.

More formally, consider arbitrary race j with m_j candidates and a specific candidate i. We assume a simple linear model that maps the voter preference $f_i(0)$ to the underlying concentration parameters α_i. Although there are many possible covariates we could include, we found that very few actually improved out-of-sample predictive performance.[16] We therefore include only the lagged PVI generated by Cooks political report (Campbell 2018; MacWilliams 2015), and an indicator for the experience of the candidate where a one indicates the candidate has held elected office before and it is coded zero otherwise (Jacobson 1989; Jacobson and Carson 2019). We also include a year-level random effect to accommodate unmodeled electoral "swings" associated with specific election cycles. PVI and the year random effects are reverse coded by party.

More formally, collect $\boldsymbol{\alpha}_j = (\alpha_{1j} + \tilde{\alpha}, \ldots, \alpha_{m_j j} + \tilde{\alpha})$ from all candidates in the race ($\tilde{\alpha} \geq 0$). The base parameter $\tilde{\alpha}$ here is introduced for two reasons. First, it can reduce variance of samples and thus stabilize the MCMC sampling. Second, $\tilde{\alpha}$ encodes the prior belief on how equally the vote shares should be distributed without any additional information. We assume that the actual vote share vector $\mathbf{y}_j = (y_{1j}, \ldots, y_{m_j j})^\top$ is distributed with a Dirichlet distribution $\mathbf{y}_j \sim \text{Dir}(\mathbf{y}_j; \boldsymbol{\alpha}_j)$, where α_{ij} is a linear function of $f_i(0)$ and contextual predictors. We also need to integrate over the distribution of $f_i(0)$; in our case, the distribution of $f_i(0)$ is a truncated Gaussian. In total, we assume the election outcomes follow the following data generating process:

$$f_i(0) \sim \mathcal{N}\left(\mu_{f_i}, K_{f_i}\right), \quad 0 \leq f_i \leq 1, \tag{13}$$

$$\alpha_{ij} = \theta_1 f_i(0) + \theta_2 \text{party}_{ij} \times \text{pvi}_j + \theta_3 \text{experience}_{ij} + \text{party}_{ij} \times \gamma_{\text{year}}, \quad \alpha_{ij} \geq 0, \tag{14}$$

$$\gamma_{\text{year}} \sim \mathcal{N}(0, \sigma^2_{\text{year}}), \tag{15}$$

$$(y_j, \ldots, y_{m_j})^\top \sim \text{Dirichlet}(\tilde{\alpha} + \alpha_1, \ldots, \tilde{\alpha} + \alpha_{m_j}), \quad \tilde{\alpha} \geq 0. \tag{16}$$

Here, we allow party to be equal to 1 for Democratic candidates, −1 for Republicans, and 0 for independents.[17] This allows for PVI and the year random effects to have opposite effects by party.

The model is completed by placing proper but vague priors across all parameters. The priors for the θ parameters are set to be wide based on the scale of the relevant variable. Specifically, we set

15 Note that in our model, the specific party affiliation of the choices is determined by the values of the covariates, not whether it is the first, second, or third choice. That is, the model is robust to the kinds of idiosyncratic variations in the meaning of categories discussed here.

16 For instance, one alternative would be to include an indicator for incumbency status. While the model does improve at more distant time horizons (e.g., $\tau = 56$) the overall accuracy of the final predictions is no better and by some metrics worse. For the 2018 election discussed below, for example, including an incumbency covariate results in nearly identical root mean squared error but lower predictive accuracy (94.29% vs. 97.14%). However, as we discuss in our concluding discussion, it should be possible to extend the model to consider a wider array of covariates through some form of regularization scheme in future research.

17 We code third-party candidates who regularly caucus with one party as belonging to that party. So, for instance, Sen. Bernie Sanders is coded as a Democrat.

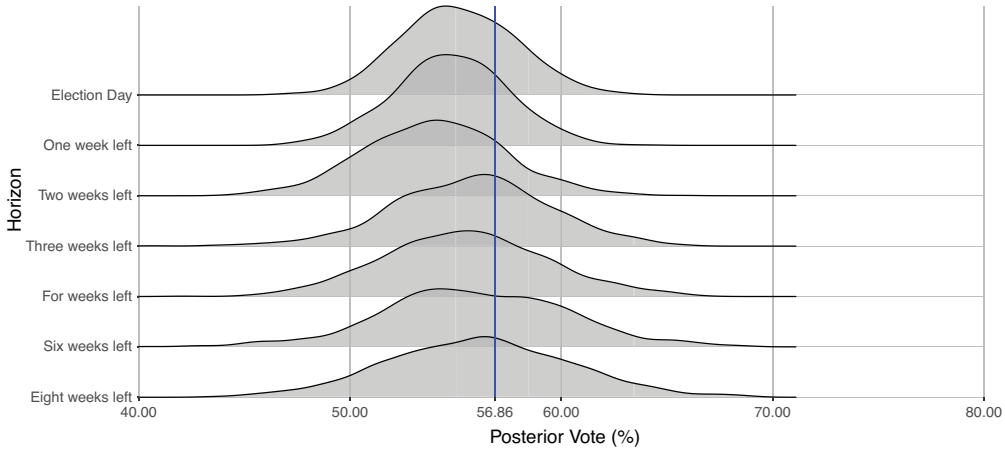

Figure 3. Posterior predictive densities for Sen. John McCain in the 2016 Arizona Election at different time horizons. The vertical line indicates the final vote share. Note that the posterior narrows noticeably as the time horizon shrinks.

independent truncated Gaussian priors on the θ coefficients and the year-level random effects.

$$\theta_1 \sim \mathcal{N}(0, 100^2);\ \theta_1 > 0 \quad \theta_2 \sim \mathcal{N}(0, 10^2);\ \theta_2 > 0,$$
$$\theta_3 \sim \mathcal{N}(0, 10^2);\ \theta_3 > 0 \quad \sigma^2_{\text{year}} \sim \text{Gamma}(1, 0.5).$$

We can combine all of these parameters together in $\boldsymbol{\Theta} = \left(\{\theta\}, \{\gamma_{\text{year}}\}, \sigma^2_{\text{year}}, \tilde{\alpha}\right)$ and let $\mathbf{z}_j$ be the vector of contextual factors for election j. We obtain the posterior $p(\boldsymbol{\Theta} \mid \{\mathbf{y}_j\}, \{\mathbf{z}_j\}, \{f_i(0)\})$ using MCMC estimation. Specifically, we use no-U-turn sampling in Stan (Carpenter *et al.* 2017; Hoffman and Gelman 2014).

With this posterior, the final predictive distribution of future election outcomes with new $\{\mathbf{f}^*_i(0)\}, \{\mathbf{z}^*_j\}$ will be defined by (13)–(16) marginalized by the posteriors:

$$p\left(\mathbf{y}_j \mid \{\mathbf{z}^*_j\}, \{f_i(0)^*\}, \{\mathbf{y}_j\}, \{\mathbf{z}_j\}, \{f_i(0)\}\right)$$
$$= \int p(\mathbf{y}_j \mid \{\mathbf{z}^*_j\}, \{f_i(0)^*\}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta} \mid \{\mathbf{y}_j\}, \{\mathbf{z}_j\}, \{f_i(0)\}) d\boldsymbol{\Theta} \tag{17}$$

A final issue is how to handle the dynamic nature of our forecasting task. While we have the complete set of polls for elections in our training set, when making real-time forecasts we have only the polls up to the current date. Training the model on the complete set of polls (all the way up to Election Day) is likely to lead to higher weight being assigned to polling data and poor predictive performance at remote time horizons. For instance, the coefficients for the Dirichlet regression component in the election model may put too much confidence on the polling. As noted above, this same issue applies to hyperparameter selection for the candidate-level model.

To address this concern, we train the complete model at various time horizons denoted by τ. For any threshold, we discard all data where $|t| < \tau$. Thus, when $\tau = 28$, we ignore all polls in the training data closer than 28 days to the election. This again helps calibrate the model for the levels of accuracy we can expect at various horizons. Table F.1 in Supplementary Appendix F shows the summaries for the posteriors of the model parameters at horizons ranging from $\tau = 0$ to $\tau = 56$ (8 weeks before the election). As expected, the θ parameter associated with $f_i(0)$ increases as Election Day approaches while the fundamental parameter become relatively less important.

Figure 3 shows the posterior prediction for Senator John McCain in 2016 for various time horizons. Note that the outcome (marked with the vertical blue line) is near the center of

the posterior for all horizons, but that the prediction becomes more concentrated as Election Day nears. This reflects both more certainty in $f_i(0)$ and changing weights in the Dirichlet regression.

3.4 Discussion

Before turning to our results, we briefly contrast our approach with existing forecasting models in the literature. Most importantly, we combine a dynamic, poll-based model with an election-level model trained on historical data to make predictions about individual senate races. Some existing poll-based models are dynamic (e.g., Jackman 2005; Linzer 2013) while others create district-level forecasts based on historical election results (e.g., Klarner 2012). However, to the best of our knowledge this is the first published model to explicitly combine these two approaches.[18]

Second, we introduce a GP framework for modeling trends in public opinion. Although related to the random walk model in Linzer (2013), it differs in two crucial ways. To begin with, the GP model allows us to model polling trends as a linear process with nonlinear deviations, which (as we show below) offers significant improvements in predictive performance when polling data is sparse. Further, by adopting the Gaussian approximation to the binomial likelihood in Equation (9), we can exactly derive the posteriors for each candidate. This computational efficiency allows us to build the election-level model and facilitates our loyo cross-validation approach.

Finally, it is also worth considering the computational resources required by the model. Assuming that the hyperparameters have been selected, running the complete model is quite fast. A standard run with 5,000 MCMC iterations takes roughly 5 min on an Intel i7-CPU machine (running three chains in parallel). The GP component is very fast because results can be computed exactly without sampling, usually completing in under one minute. This contrasts with, for instance, a stan implementations of the Linzer (2013) model, which takes approximately 30 min for a given election cycle. Thus, during any one election, the computational load is very reasonable.

The computational bottleneck with our approach is in the loyo cross-validation procedure for choosing hyperparameters. As described above, we ran the loyo validation for the 1992–2016 period with 100 hyperparameter settings at seven forecasting horizons. With three MCMC chains for each model, this results in 27,300 posteriors. Thus, even the 5-min run time is cumulatively computationally intensive requiring the use of a computing cluster. This exercise only needs to be done once in advance of any specific election cycle, but is nonetheless time consuming. We return to this point in our concluding discussion.

4 Empirical Evaluation

In this section, we investigate our model using historical polling data and vote shares in U.S. Senate elections from 1992 to 2018. Throughout our model building process, we held out the 2018 election as a test case and it was not involved in any hyperparameter tuning, variable selection, or other decisions. Therefore, we can assess the model's predictive performance using the 1992–2016 period, but also approximate its true out-of-sample performance using the 2018 data. In the next section, we report predictions for 2020 actually made in advance of Election Day.

4.1 Data and Evaluation Criteria

We obtained opinion polls and election outcomes of all senate election races from 1992 to 2018 from www.fivethirtyeight.com and from *CNN*. On average 16 polls were conducted for each race, although some races such as the 2016 Florida election had over 80. Most of the surveys are conducted 2 weeks to 4 months prior to election, with a median number of respondents of 635.

18 To be sure, Linzer (2013) uses historical data to create informative priors and chooses hyperparameters based on performance in a previous cycle. However, the likelihood is based on polling alone, and the model is fit to only one cycle.

Over 470 entities have conducted these polls, but several active pollsters collectively contribute over half of them, including Rasmussen Reports, Mason-Dixon Polling, Public Policy Polling, SurveyUSA, YouGov, SurveyMonkey, Quinnipiac, and Zogby Interactive. To guarantee credibility, we eliminated polls sponsored by parties or candidates since unfavorable polls from these sources are not released.

We also acquired the partisan voting indices in every election cycle for each election state. The Cook PVI is a measurement of how strongly a U.S. congressional district or state leans toward the Democratic or Republican Party, compared to the nation as a whole. For example, PVI for California in 2018 is 10.76, indicating a strong preference for Democratic candidates, while PVI for the pro-republican state Texas in 2018 is −7.02. For each candidate, we coded partisan affiliation and past experience (whether or not they held office). Where not provided in the *CNN* data, we coded these manually using ballotpedia.com.

To evaluate performance, we examine both the forecasting precision and the validity of our model. Hence, we consider the following measures: the averaged root-mean-squared-error (RMSE) between the expectation of the Dirichlet posterior samples and actual vote shares, the prediction accuracy of winners for election defined by higher winning probability (we calculate the winning probability of each candidate as the proportion of samples with the highest vote shares in Dirichlet posteriors), the coverage rate of actual vote shares in 95% credible intervals for Dirichlet posteriors, the averaged multinomial predictive likelihood and the averaged log-scaled Dirichlet predictive likelihood (LL). RMSE and prediction accuracy focus on the precision of the forecasting ability, while coverage rate focuses on the validity of the claimed credible intervals. The two likelihood measures serve as out-of-sample evaluation criteria for both vote share (Dirichlet) and final outcome (multinomial) that also reflect the uncertainty in the full posterior.

4.2 Baselines

We compare the performance of our combined GP and Dirichlet regression (GP+DR) model against three benchmarks. First, we compare our model to the dynamic Bayesian model in Linzer (2013). This model was developed for predicting state-level results in presidential elections, but we adjusted it for predicting Senate races. Intuitively, this model is a dynamic Bayesian random walk (BRW) model similar to the nonlinear component in the model described above except that latent public opinion is assumed to be a random walk. We use the same informative prior for $\{a\}$ as used above and use the tuning parameters and basic estimation procedures as described in Linzer (2013).[19]

Second, we consider a baseline Dirichlet Regression model that uses a Bayesian linear regression model to forecast voter preferences. We refer the second baseline as LM+DR. To ensure a fair comparison, we choose the same priors for the linear coefficients as those in GP priors. We also chose the σ^2 hyperparameter using the same cross-validation approach described above. This model is, in essence, the same as what we describe above without allowing for deviations from linearity. Finally, we examine the performance of the GP model in isolation excluding the Dirichlet regression portion of the hierarchy. Note that while we frame these as competitors to our favored model, both of these baselines are also novel.

4.3 Results

First, we present results from a loyo cross validation exercise where each election cycle from 1992 to 2016 was held out. This has the advantage that we can use the complete set of election outcomes to validate the model. However, since we followed an identical procedure when choosing our

19 The major deviation from the original model implementation is we do not include a national over time trend since the senate races are far more independent than the state-level presidential races. See Supplementary Appendix B.

hyperparameters above, there may still be some risk of overfitting.[20] We therefore also present results for the 2018 election separately which serves as a stronger out-of-sample test.

We simulate a real forecasting scenario and examine the model's forecasting ability at various horizon τ's. Specifically, we consider horizons of 4 months, 3 months, 6 weeks, 4 weeks, 2 weeks, 1 week and Election Day, where $\tau = 56, 42, 28, 21, 14, 7, 0$. As noted above, Table 1 summarizes hyperparameters learnt for the candidate-level model used throughout this exercise.

Table 2 shows the results for the loyo cross validation exercise for the 1992–2016 period. The results show that GP+DR model on average outperforms the other baselines across metrics. The closest competitor is actually the LM+DR model, which performs quite well in terms of coverage and accuracy. This is explained in part by the fact that the GP model itself is mostly linear at distant horizons and when there is little polling data. However, the nonlinear component in the GP does provide measurable improvements over the linear version in the final lead up to the election when the hyperparameters most enable nonlinear deviations (see Figure 2). In Supplementary Appendix G, we use a paired t-test to show that this improvement in accuracy is statistically significant when $\tau \leq 21$.

We then predict the 2018 cycle, which was not used in our model development or cross validation, and find a nearly identical pattern. (Full results for 2018 are shown in Supplementary Appendix H.) The RMSE for the Election Day forecast was 0.053, 0.055, 0.060, and 0.075 for the GP+DR, LM+DR, BRW, and GP models respectively. Meanwhile the predictive accuracy was 0.951, 0.932, 0.898, and 0.936.[21]

Figure 4 shows the predictions, 95% predictive credible intervals, and outcomes for the 2018 senate elections with $\tau = 7$. The results show that all election outcomes fell within the 95% credible range and that on average the forecast tracked the actual election outcomes very closely. Moreover, the elections where the model is incorrect at a 7-day range are also among the closest contests in that cycle (Arizona and Nevada). Finally, the width of the credible interval can vary significantly depending on the number and recency of polls for that election. For instance, the credible intervals for Wyoming are very large reflecting the fact that we had only one poll. This contrasts with, for instance, Missouri where dozens of polls were reported.

5 Predicting the 2020 Cycle

Finally, we turn to the task of predicting the 2020 senate elections. For this cycle, we again acquired all data from the fivethirtyeight.com website. Following the procedures outlined above, we exclude all partisan polls and and date each poll based on the first day it was fielded. We did not include any third-party candidates,[22] and we exclude the Georgia special election and the Louisiana senate race do to the potential for a runoff after November.[23] We used the same hyperparameters as shown in Table 1, but refit the Dirichlet regression using the complete 1992–2018 training period.

The final predictive densities for the Democratic candidates are shown in Figure 5 (we show only one party since we modeled only two candidates in each state). The model predicted that

20 However, in Supplementary Appendix D, we show that the model's predictions are not strongly sensitive to whether or not the complete set of elections is included during the loyo process.

21 In Supplementary Appendix I, we also compare the 2018 model to the predictions posted on fivethirtyeight.com at different forecasting horizons. The GP+DR model is more accurate at all time horizons (correctly predicting 97% vs. 90% of races on Election Day). However, there is evidence that our model is relatively conservative, with coverage rates of 97% for 80% credible intervals. We emphasize that our model was developed *after* the 2018 cycle.

22 We could have included the third-party candidates in Maine as they were polled regularly and received a significant portion of the final vote. We discuss the issue of selecting third-party candidates *ex ante* to include in the forecast in our concluding discussion.

23 We included the nonspecial Georgia race based on the (incorrect) assumption that the third-party candidate would not deny the winner a majority. One possible extension to this model would be to better accommodate runoffs and elections that take place outside of November. Alternatively, we could simply predict November vote share rather than excluding races with a possible runoff and ignore the potential for a subsequent runoff.

https://doi.org/10.1017/pan.2021.42 Published online by Cambridge University Press

Table 2. Predictive accuracy in the 1992–2016 period.

		Days until Eleciton Day (τ)						
	Model	**56**	**42**	**28**	**21**	**14**	**7**	**0**
RMSE	GP+DR	0.082	0.074	0.068	0.065	0.059	0.056	0.053
	GP	0.102	0.096	0.091	0.088	0.082	0.078	0.075
	LM+DR	0.082	0.073	0.069	0.065	0.061	0.055	0.055
	BRW	0.085	0.081	0.078	0.076	0.072	0.065	0.060
95% Coverage	GP+DR	0.919	0.931	0.932	0.949	0.943	0.961	0.946
	GP	0.800	0.870	0.862	0.860	0.819	0.823	0.747
	LM+DR	0.928	0.919	0.938	0.933	0.933	0.936	0.950
	BRW	0.548	0.523	0.504	0.510	0.514	0.514	0.527
Predictive accuracy	GP+DR	0.892	0.892	0.919	0.915	0.920	0.935	0.951
	GP	0.879	0.895	0.913	0.916	0.921	0.934	0.936
	LM +DR	0.885	0.899	0.908	0.914	0.918	0.934	0.932
	BRW	0.798	0.814	0.827	0.828	0.847	0.859	0.898
APLL Vote Share	GP+DR	1.411	1.552	1.646	1.703	1.803	1.913	1.950
	GP	0.549	1.126	1.171	1.186	1.138	1.214	0.940
	LM+DR	1.410	1.537	1.634	1.682	1.776	1.873	1.902
	BRW	−2.171	−1.606	−1.513	−1.126	−1.048	−0.377	−0.143
APLL Winner	GP+DR	−0.135	−0.116	−0.102	−0.098	−0.092	−0.081	−0.075
	GP	−0.170	−0.117	− 0.101	−0.097	−0.092	−0.078	−0.072
	LM+DR	−0.135	−0.118	−0.105	−0.100	−0.092	−0.082	−0.079
	BRW	−0.371	−0.337	−0.342	−0.289	−0.252	−0.234	−0.172

Cells reports fit statistics at various simulated time horizons using a leave-one-year-out cross validation. RMSE is root mean squared error for the point predictions, while the 95% coverage is the percent of vote shares that fall within the predicted 95% credible intervals. Predictive accuracy measures the percent of races predicted correctly across cycles. Average predicted log-likelihoods (APLL) are predicted using the Dirichlet likelihood (for vote share predictions) and the multinomial likelihood (for winner predictions).

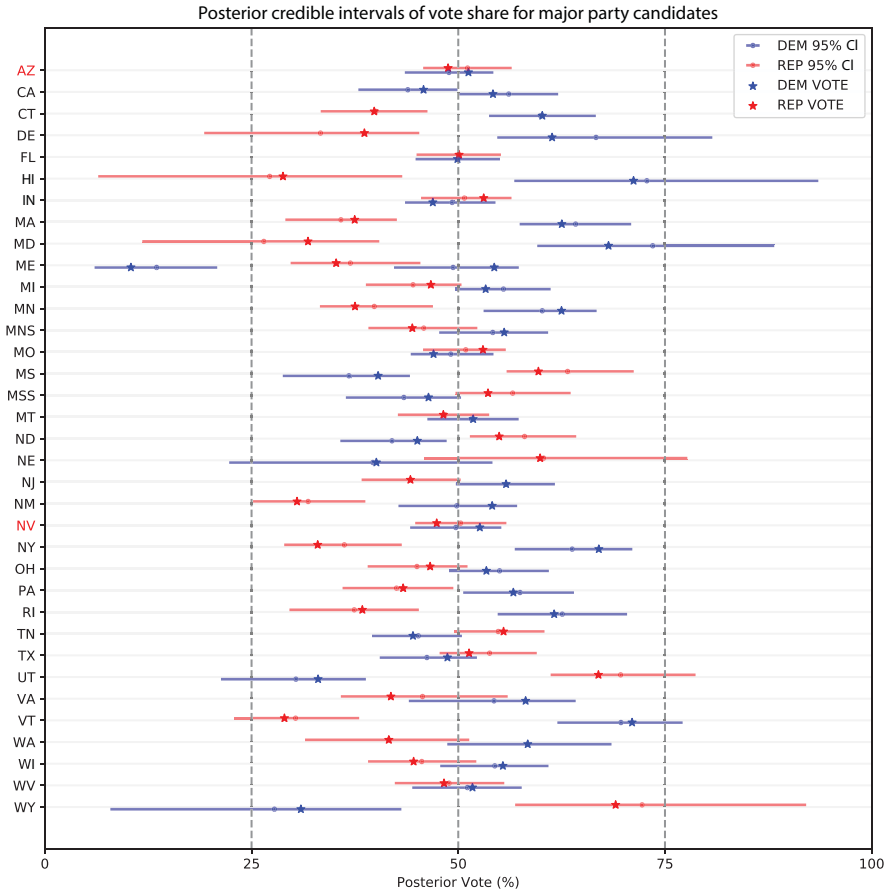

Figure 4. Forecast for 2018 at one week time horizon for major party candidates. Stars indicate actual vote share, while points and confidence intervals reflect posterior mans and 95% credible predictions. The California election had two Democrats and we coded Sen. Angus King of Maine as Democrat. Red font for state names indicates an incorrect prediction. This model also included a forecast for Libertarian candidate Gary Johnson in New Mexico (posterior median 0.185, 95% CI [0.117, 0.247], outcome 0.154). MNS and MSS are special elections held on the usual Election Day.

the Democrats were favored to win in four Republican-held seats (CO, ME, AZ, and NC) and to lose Alabama. However, the election outcomes were predicted to be very close in many states including MS, AK, MT, SC, GA, IA, NC, IA, AZ, ME, and CO (states here are ordered by the degree to which they favor the Democratic candidates).[24]

In all, the forecast was accurate, missing only two election outcomes. One miss was North Carolina, which our model predicted as being a narrow Democratic victory and turned out to be a narrow Republican victory. The only serious miss was Maine, where pre-election polling was dramatically off.[25] Maine was also the only case where the result fell outside of our 95% predictive CI, giving us 96.9% coverage.

We can compare this performance to the *Economist* and fivethirtyeight.com models, although it is important to note that their methods are not public. These results are show in Table 3. Our model outperformed the *Economist* model on all metrics. In addition to NC and ME, *The Economist* also missed Iowa and (the plurality winner in) Georgia. The 95% out-of-sample coverage rate was 90.6% as, in addition to Maine, their model also missed New Jersey and West Virginia.

24 We provide example outputs for candidate-level models in Supplementary Appendix J. Additional details on the predictive posterior for the 18 closest races are shown in Supplementary Appendix K.

25 The last nonpartisan poll showing Sen. Collins winning re-election was reported in September, 2019.

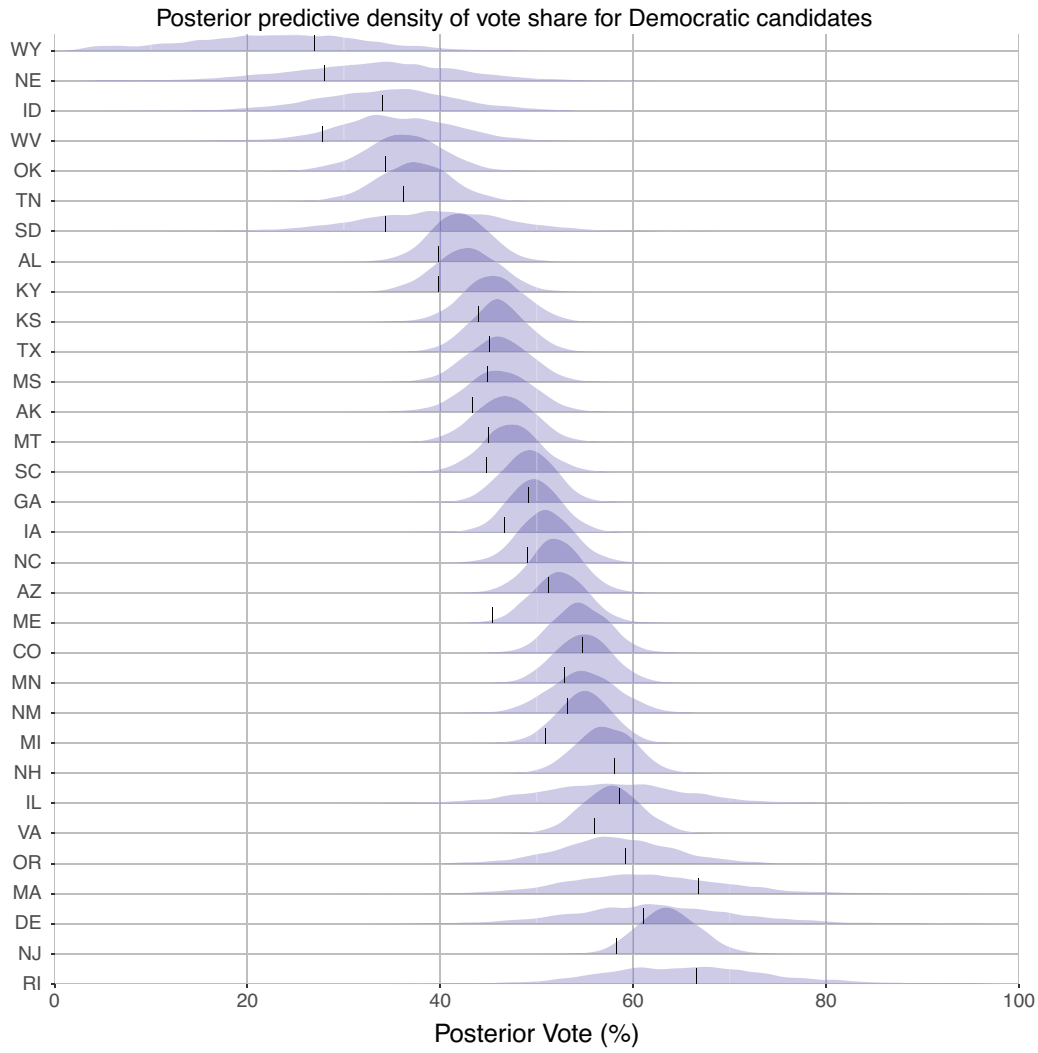

Figure 5. Predicted vote share densities for 2020 on Election Day for Democratic candidates. States are arranged in order of increasing probability of Democrats winning. Solid black vertical lines indicate actual vote shares for Democratic candidates. The plot excludes Louisiana and the Georgia special election.

Table 3. Comparing predictive accuracy in 2020 cycle to prominent media forecasts.

	Model				
	GP+DR	**Economist**	**538 Classic**	**538 Delux**	**538 LITE**
Accuracy (%)	93.75	87.5	90.63	93.75	90.63
RMSE	0.0316	0.0394	0.0421	0.0397	0.0446
80% CI Coverage	0.875		0.859	0.859	0.844
95% CI Coverage	0.969	0.906			

Comparing forecast accuracy based on final model predictions acquired from Fivethirtyeight (2020) and (Economist 2020). The *Economist* provided 95% CIs while *fivethirtyeight.com* provided 80% CIs.

It is not as easy to directly compare performance to the *fivethirtyeight.com* forecasts as they predict non-normalized voter share (not two-party vote share), provide only 80% predictive intervals, and actually produce three predictions. Thus, for instance, the RMSE metric is not on the same scale as our model which predicts the normalized vote share (excluding write-ins,

third-party votes, etc.). However, the results in Table 3 indicate that our model performed at comparable levels of accuracy and coverage as their forecasts, although ours is perhaps slightly conservative in having a 87.5% coverage rate for the 80% CIs. Notably, our model made the same winner prediction for all of the 2020 elections as the "Delux" model, while their other variants missed the plurality winner in Georgia. We also have lower RMSE than all three variations. In all, we consider this to be evidence that our model is at least as accurate as fivethirtyeight.com while having the advantage of being a public and transparent methodology that can be studied and improved upon by other forecasters.

6 Conclusion

In this article, we offer a novel approach to dynamic election prediction that combines both poll-based and fundamentals-based forecasting. Although the model itself is somewhat complex, in the end it includes only a few variables: polling, PVI, experience, and party. The novelty here is not in what factors go into the model, but how they are combined to create accurate, well-calibrated predictions.

Our approach contains two basic stages. The first step is to treat polling data as a probabilistic representation of latent public support for a candidate, where this latent support has a linear and nonlinear trend. By fitting a model to this trend, we can accurately predict forward to where public opinion will be on Election Day. Second, we then incorporate predictions about this latent position into a Dirichlet regression that uses historical data and a few simple features about the election to estimate the degree to which polling can be used to predict elections based on historical data. A final innovation is that we train the data completely at different time horizons to ensure that our final predictions reflect an appropriate level of uncertainty.

While we believe that this model improves upon other Senate forecasting models in the literature, it could be refined in several ways. First, we might better extend it to handle unusual cases like runoff elections or special elections (e.g., the 2020 Georgia special election) or the potential for instant runoffs in states adopting ranked-choice voting. We could, in theory, also extend the model to account for "house effects" of various polling firms or weight more accurate firms more highly in the candidate-level model. Likewise, we could try alternative variables to include in the construction of the candidate-level prior or in the election-level model.[26] However, adding such complications should be done with caution as they may lead to overfitting. Many variables (e.g., money raised or incumbency status) should be reflected in the polling data. Once we have conditioned on latent public support, the list of accurate predictors of outcomes is much smaller. Finally, retrospectively it is relatively easy to identify which third-party candidates should be included in a predictions model since they appear regularly in the polling data and receive a considerable vote share. However, future work might improve upon our efforts by more clearly defining a rule for when to include minor candidates based on *ex ante* conditions.

A further shortcoming is that our model does not allow online-updating of hyperparameters: forecasters have to learn from scratch customized hyperparameters for every new horizon. In Table 1, the learnt length scale and noise standard deviations are somewhat constant across horizons, while the learnt output scales shrink at earlier horizon. When computation capability is limited, practitioners may use the same optimal hyperparameters across horizons and warp the output scale according to the forecasting horizon.

A third extension would be to adjust the model to handle elections at different levels. This model would be relatively straightforward to extend to, for example, gubernatorial races. However, more significant adjustments may be needed for lower (e.g., races for the U.S. House of

26 A simple approach might be to use regularized regression for this task and include the complexity penalty parameter in the loyo cross validation.

Representatives) or higher (presidential) elections. Lower-level races are unusual in that there is even less polling data available for most races, which may require heavier reliance on contextual factors or cycle factors such as generic ballots. Meanwhile, presidential races usually offer many more polls, but the election-level training data is necessarily very sparse at the national level and the state-level outcomes (state-level results) are much more correlated. Researchers wishing to extend this basic approach to those settings should think carefully about how to construct the election-level and candidate-level models to account for these important differences. It will also be important to consider how well our approach to, for instance, cross validation will work given smaller sample sizes.

Finally, it is important to remember that while we have taken steps to gauge the accuracy of the model, there is no way feasible way to assess its true long-term out-of-sample performance until we observe more election outcomes. We created a held-out prediction for 2018 and a true prediction for 2020, but there is always the risk that idiosyncratic features of these election cycles are driving the results. It will be important to re-evaluate the model's performance in future cycles.

Acknowledgments

We are grateful to Harry Enten at CNN for providing data and for David Carlson for help and collaboration in an earlier attempt at modeling elections. We also appreciated the help of the Political Analysis editorial staff as well as the comments from our reviewers.

Funding

YC and RG were supported by the National Science Foundation (NSF) under award number IIS–1845434.

Data Availability Statement

Replication code for this article has been published in Code Ocean, a computational reproducibility platform that enables users to run the code, and can be viewed interactively at https://doi.org/10.24433/CO.4154884.v1 (Chen, Garnett, and Montgomery 2021a). A preservation copy of the same code and data can also be accessed via Harvard Dataverse at https://doi.org/10.7910/DVN/GNHESM (Chen, Garnett, and Montgomery 2021b).

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2021.42.

Bibliography

Abramowitz, A. I. 2008. "Forecasting the 2008 Presidential Election with the Time-for-Change Model." PS: Political Science & Politics 41 (4): 691–695.

Aitchison, J. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society: Series B (Methodological)* 44 (2): 139–160.

Campbell, J. E. 2018. "The Seats-in-Trouble Forecasts of the 2018 Midterm Congressional Elections." PS: Political Science & Politics 51 (S1): 12–16.

Campbell, J. E., and K. A. Wink. 1990. "Trial-heat Forecasts of the Presidential Vote." *American Politics Research* 18 (3): 251–269.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32.

Chen, Y., R. Garnett, and J. M. Montgomery. 2021a. "Replication Data for: Polls, Context, and Time: A Dynamic Hierarchical Bayesian Forecasting Model for US Senate Elections." Code Ocean. https://doi.org/10.24433/CO.4154884.v1

Chen, Y., R. Garnett, and J. M. Montgomery. 2021b. "Replication Data for: Polls, Context, and Time: A Dynamic Hierarchical Bayesian Forecasting Model for US Senate Elections." https://doi.org/10.7910/DVN/GNHESM, Harvard Dataverse, V1.

Economist . 2020. "Forecasting the US Elections." https://projects.economist.com/us-2020-forecast.

Erikson, R. S., and C. Wlezien. 2008. "Leading Economic Indicators, the Polls, and the Presidential Vote." PS: Political Science & Politics 41 (4): 703–707.
Fair, R. C. 1978. "The Effect of Economic Events on Votes for President." *The Review of Economics and Statistics* 60(2): 159–173.
Fivethirtyeight 2020. "Democrats are favored to win the Senate." https://projects.fivethirtyeight.com/2020-election-forecast/senate/.
Gill, J. 2020. "Measuring Constituency Ideology Using Bayesian Universal Kriging." *State Politics & Policy Quarterly* 21(1): 80–107.
Hainmueller, J., and C. Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22: 143–168.
Hoffman, M. D., and A. Gelman. 2014. "The No-U-Turn sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15 (1): 1593–1623.
Hummel, P., and D. Rothschild. 2014. "Fundamental Models for Forecasting Elections at the State Level." *Electoral Studies* 35: 123–139.
Jackman, S. 2005. "Pooling the Polls Over an Election Campaign." *Australian Journal of Political Science* 40 (4): 499–517.
Jacobson, G. C. 1989. "Strategic Politicians and the Dynamics of US House Elections, 1946–86." *The American Political Science Review* 83 (3): 773–793.
Jacobson, G. C., and J. L. Carson. 2019. *The Politics of Congressional Elections*. Lanham, MD: Rowman & Littlefield.
Katz, J. N., and G. King. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review* 93(1): 15–32.
Klarner, C. E. 2008. "Forecasting the 2008 US House, Senate and Presidential Elections at the District and State Level." PS: Political Science and Politics 41 (4): 723–728.
Klarner, C. E. 2012. "State-Level Forecasts of the 2012 Federal and Gubernatorial Elections." *PS, Political Science & Politics* 45 (4): 655–662.
Klarner, C. E. 2013. "2012 Presidential, US House, and US Senate Forecasts." *PS, Political Science & Politics* 46 (1): 44–45.
Klarner, C. E., and S. Buchanan. 2006. "Forecasting the 2006 Elections for the United States Senate." PS: Political Science and Politics 39 (4): 849–855.
Lewis-Beck, M. S., and C. Tien. 2008. "The Job of President and the Jobs Model Forecast: Obama for '08?" PS: Political Science & Politics 41(4): 687–690.
Linzer, D. A. 2013. "Dynamic Bayesian Forecasting of Presidential Elections in the States." *Journal of the American Statistical Association* 108 (501): 124–134.
Lockerbie, B. 2012. "Economic Expectations and Election Outcomes: The Presidency and the House in 2012." *PS, Political Science & Politics* 45 (4): 644–647.
MacWilliams, M. C. 2015. "Forecasting Congressional Elections Using Facebook Data." *PS, Political Science & Politics* 48 (4): 579.
Mohanty, P., and R. Shaffer. 2019. "Messy Data, Robust Inference? Navigating Obstacles to Inference with bigKRLS." *Political Analysis* 27 (2): 127–144.
Monogan, J. E., and J. Gill. 2016. "Measuring State and District Ideology with Spatial Realignment." *Political Science Research and Methods* 4 (1): 97–121.
Philips, A. Q., A. Rutherford, and G. D. Whitten. 2016. "Dynamic Pie: A Strategy for Modeling Trade-Offs in Compositional Variables Over Time." *American Journal of Political Science* 60 (1): 268–283.
Rasmussen, C. E., and C. K. Williams. 2006. *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.
Sobol, I. M. 1979. "On the Systematic Search in a Hypercube." *SIAM Journal on Numerical Analysis* 16 (5): 790–793.
Stoetzer, L. F., M. Neunhoeffer, T. Gschwend, S. Munzert, and S. Sternberg. 2019. "Forecasting Elections in Multiparty Systems: A Bayesian Approach Combining Polls and Fundamentals." *Political Analysis* 27 (2): 255–262.
Tomz, M., J. A. Tucker, and J. Wittenberg. 2002. "An Easy and Accurate Regression Model for Multiparty Electoral Data." *Political Analysis* 10(1): 66–83.
Walther, D. 2015. "Picking the Winner (s): Forecasting Elections in Multiparty Systems." *Electoral Studies* 40: 1–13.