Nowcasting Reported COVID-19 Hospitalizations Using De-Identified, Aggregated Medical Insurance Claims Data

Xueda Shen[1] Aaron Rumack[2] Bryan Wilder[2] Ryan J. Tibshirani[1]

[1]University of California, Berkeley [2]Carnegie Mellon University

Abstract

We propose, implement, and evaluate a method for nowcasting the daily number of new COVID-19 hospitalizations, at the level of individual US states, based on de-identified, aggregated medical insurance claims data. Our analysis proceeds under a hypothetical scenario in which, during the Delta wave, states only report data on the first day of each month, and on this day, report COVID-19 hospitalization counts for each day in the previous month. In this hypothetical scenario (just as in reality), medical insurance claims data continues to be available daily. At the beginning of each month, we train a regression model, using all data available thus far, to predict hospitalization counts from medical insurance claims. We then use this model to nowcast the (unseen) values of COVID-19 hospitalization counts from medical insurance claims, at each day in the following month. Our analysis uses properly-versioned data, which would have been available in real-time at the time predictions are produced (instead of using data that would have only been available in hindsight). In spite of the difficulties inherent to real-time estimation (e.g., latency and backfill) and the complex dynamics behind COVID-19 hospitalizations themselves, we find altogether that medical insurance claims can be an accurate predictor of hospitalization reports, with mean absolute errors typically around 0.4 hospitalizations per 100,000 people, i.e., proportion of variance explained around 75%. Perhaps more importantly, we find that nowcasts made using medical insurance claims are able to qualitatively capture the dynamics (upswings and downswings) of hospitalization waves, which are key features that inform public health decision-making.

1 Introduction

Timely access to public health data is critical to enable informed decision making during infectious disease outbreaks. However, setting up and maintaining public health reporting pipelines can be a burden on the health system itself. For example, beginning in May 2020, hospitals in the US have been required to report data on COVID-19 hospitalizations to the Department of Health and Human Services (HHS) (Department of Health and Human Services, 2023). This data has been critical for understanding the state of the pandemic and the current load on the health system. But the frequent reporting required for up-to-date situational awareness (daily, throughout most of the pandemic) has been quite difficult to implement and maintain. In an effort to achieve compliance, the Centers for Medicare and Medicaid Services (CMS) issued regulations in August 2020 that threatened to expel hospitals from the Medicare program, and apply monetary penalties, if they failed to comply with daily COVID-19 reporting requirements. This was strongly and openly opposed by the American Hospital Association (AHA) (Nickels, 2020), but the regulations remained in place for nearly three years, lasting until the conclusion of the COVID-19 Public Health Emergency in May 2023.

An intriguing alternative lies in data streams which already exist and are maintained for other purposes, yet are relevant for inferring disease activity. One example is medical insurance claims, which are filed by a healthcare provider to seek reimbursement from an insurance company for medical services performed. In this paper, we examine the use of *de-identified, aggregated* medical insurance claims data as a complement to public health reporting over the course of the pandemic. Specifically, we investigate the following: if hospital reporting on COVID-19 would have been reduced in frequency from daily to monthly during the Delta wave, could we use signals derived from medical insurance claims to accurately *nowcast* COVID-19 hospitalization counts during the interim periods?

Medical insurance claims have long been utilized in public health policy analysis, ranging from economic implications of healthcare (e.g., recent examples include Panczak et al. (2018); Li and Yang (2021); Sakai et al. (2019); Zheng and Peng (2021); Durizzo et al. (2022); Mori et al. (2022)), to examinations of treatment effectiveness and satisfaction (e.g., Nakayama et al. (2017); Jung et al. (2020); Geng et al. (2021); Yao et al. (2022); Song et al. (2023)). All of these works, however, are *retrospective* in nature: they seek to develop an understanding of a particular phenomenon using data that would not have been available in real-time, but only in hindsight. In contrast, our goal to carry out an analysis that reflects *real-time* estimation, so that we can understand (to the best extent we can) how our models would perform if they were to be operationalized in the future, for true prospective nowcasting. This requires us to use properly-versioned data at all times in our analysis, which would have been available in real-time, at the time nowcasts are produced. This is true of all data sources in question, but it is especially crucial for medical insurance claims signals: these signals are subject to heavy revisions, as claims can be filed long after a service was performed (we describe this more concretely later in the paper), altering previously-computed signal values. Therefore, using finalized (rather than properly-versioned) values of medical insurance claims signals for modeling and prediction can present a misleading picture of nowcasting performance. Figure 1 displays a clear example of this.

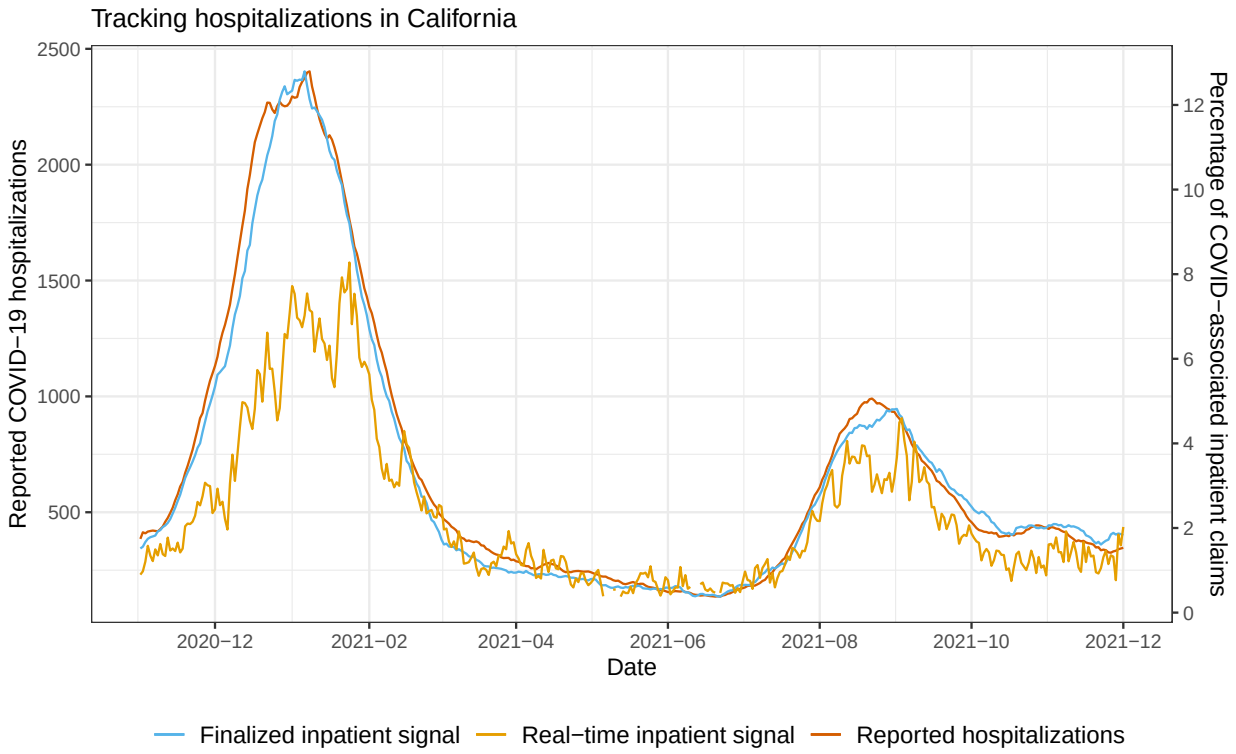

Figure 1: Reported daily COVID-19 hospitalizations, plotted alongside a signal derived from medical claims that measures daily COVID-associated inpatient admissions, for the state of California over a 1 year period during the pandemic. (All data is smoothed using a 7-day trailing average.) The y-axis on the right side of the plot is for the inpatient claims signal, of which two versions are shown: a real-time signal, representing what would have been available at each reference date along the x-axis, and a finalized signal, representing what would have been available 30 days after the reference date. We can see that the finalized signal tracks reported hospitalizations well, but the concordance with the real-time signal is much worse. (The real-time signal is also missing at some dates around the summer of 2021, which is why no values are plotted there. We will discuss this later in the methods section.)

The importance of data versioning for epidemic tracking and prediction tasks is emphasized in Reinhart et al. (2021); McDonald et al. (2021), and the importance of leveraging existing healthcare data streams for epidemic surveillance, as a complement to traditional public health reporting, is motivated in Rosenfeld and

Tibshirani (2021). Infectious disease nowcasting, using data from healthcare pipelines and also from a variety of other auxiliary data sources, has received increased attention over the last decade or so: see, e.g., Viboud et al. (2014); Smolinski et al. (2015); Yang et al. (2015); Farrow (2016); Santillana et al. (2016); Bastos et al. (2019); Jahja et al. (2019); Yang et al. (2019); Ackley et al. (2020); Brooks (2020); Leuba et al. (2020); Radin et al. (2020); Jahja et al. (2021); Li and White (2021); Menkir et al. (2021); Bergström et al. (2022); Miller et al. (2022); Seaman et al. (2022); Salazar et al. (2022); Wolffram et al. (2023); Lison et al. (2024). The goal in much of this work is to produce high-resolution, up-to-date estimates of disease activity for a fast-moving pathogen like influenza or COVID-19. The nowcasting problem is essentially to project finalized values from preliminary or partial measurements, making use of backfill or delay distributions, and possibly proxy signals which are exogenous to the main data reporting pipeline. The action of backfill or delay is naturally modeled via convolution, and Bayesian methods have been popular here, given their ability to streamline estimation and uncertainty quantification.

Our paper complements this line of work by considering the effectiveness of nowcasting based *purely* on proxy signals (for us, medical claims signals): we consider a hypothetical scenario in which hospital reporting on COVID-19 would have been set up at a coarser frequency, and under this hypothetical there is no partial or preliminary reporting data in the interm periods between reports (lasting a month, or longer) on which to base nowcasts. Finally, we examine intentionally simple statistical models based on regression, which would be relatively easy for a public health office to operationalize in practice.

2 Methods

This section describes the data that we use, the hypothetical scenarios (for hospital reporting cadence) that we consider, and the nowcast and backcast models that we build and evaluate.

2.1 Data

Throughout, we restrict our attention to nowcasting state-level hospitalizations. This is because the medical claims signals that we describe below are not generally available at each US county. That said, in principle, the same ideas we describe in what follows could be applied to nowcast hospital reports in large counties (subject to enough claims data being available in order to form robust signals). We also restrict our attention to nowcasting daily reported hospitalizations between April 1, 2021 and August 1, 2023, though we use data back until November 1, 2020 for training models.

State-level, daily COVID-19 hospitalization reports are obtained from the HHS, accessed via the Delphi Epidata API (Farrow et al., 2015; Reinhart et al., 2021). We use $Y_{\ell,s}$ to denote the 7-day trailing average of *finalized* reported new COVID-19 hospitalization counts corresponding to location ℓ and time s. Averaging over 7-day trailing windows is mainly used as a smoother, and to account for weekday-weekend differences. Reported hospitalization counts are subject to revision (these are typically minor in comparison to revisions for claims signals), hence we also introduce notation to work with versioned data henceforth: we denote by $Y_{\ell,s}^{(t)}$ the 7-day trailing average of hospitalization counts for location ℓ and time s, but whose data version is *as of* time t. In this context, we often refer to s as the *reference date* and t the *issue date*. We use analogous notation and nomenclature for all versioned data in this paper.

De-identified, aggregated medical insurance claims data are provided to us by Change Healthcare, which covers a sizeable fraction of healthcare providers in the US, around 25% of all commercial medical insurance claims. This is large enough for us to derive signals that reflect COVID-associated outpatient and inpatient activity for each U.S. state. That said, to be clear, our claims data set can still only lead to an approximation of the total counts from HHS hospitalization reports, since the latter covers effectively the entire population (all hospitals except psychiatric and rehabilitation hospitals are subject to reporting requirements), whereas we only have medical insurance claims from insurance systems that work with Change Healthcare.

In this work, we consider the following two signals:

- $O_{\ell,s}$: the *finalized* percentage of outpatient claims in a 7-day trailing window with a confirmed COVID diagnostic code, corresponding to location ℓ and time s.
- $I_{\ell,s}$: the *finalized* percentage of inpatient claims in a 7-day trailing window with a COVID-associated diagnostic code, corresponding to location ℓ and time s.

As before, we use $O_{\ell,s}^{(t)}$ and $I_{\ell,s}^{(t)}$ to denote the *versioned* outpatient and inpatient claims signals, respectively, corresponding to issue date t. The use of versioned data is extremely important when working with claims-based signals because medical insurance claims are often submitted and/or processed late, many days (and even months) after a given date of service. This process is generally referred to as *backfill*, and it has quite a pronounced effect on both outpatient and inpatient signals. We can look back at Figure 1 for an example of the inpatient signal in California. What is labeled as the "real-time inpatient signal" in the figure is $I_{\ell,s}^{(s)}$ in our notation introduced here, for s ranging over the time values along the x-axis (and ℓ = California). What is labeled as the "finalized inpatient signal" is actually $I_{\ell,s}^{(s+30)}$; note that 30 is just a large number chosen for simplicity, and it is not necessarily the case that all claims would have been filed after 30 days.

Sometimes backfill is so severe that no claims for reference date s are filed at all until a later issue date t, which we refer to as *latency*. If the latency is large enough, in particular, if no claims are available until a full 7 days after a given reference date, then the real-time claims signal value will be missing. This happens at a few dates in Figure 1 in May and June of 2021.

In defining the inpatient and outpatient claims signals, the use of a ratio of COVID-associated claims to all claims, instead of a count of COVID-associated claims, is important for two reasons. First, it adjusts for the unknown market share (unknown to us) of Change Healthcare in each given location. Second, this ratio tends to be more robust to backfill than a pure count, as both the numerator and denominator get updated as new claims are filed (whereas a count would only be revised upward, and as we will see later in Figure 2, especially for the inpatient signal, only a small fraction of the total volume of claims are available in the first few days after a given date of service).

Like the HHS hospitalization reports, various signals derived from medical insurance claims are available in the Delphi Epidata API. The precise diagnostic codes used in defining the outpatient and inpatient signals are given in the API documentation: `https://cmu-delphi.github.io/delphi-epidata`; for convenience, we have relayed these definitions in Appendix A, and we have made all data (including properly-versioned data) used in our analysis available for download at: `https://github.com/cmu-delphi/hhs-nowcasting`.

2.2 Hypothetical scenarios

We consider two hypothetical scenarios in the analysis in what follows. Recall that our analysis examines nowcasting daily reported hospitalizations between April 1, 2021 and August 1, 2023, and we have training data all the way back until November 1, 2020.

Scenario 1: monthly updates. Our first hypothetical scenario spans nowcasting dates from April 1, 2021 to November 30, 2021. In this scenario, on the first day of each month during this period, we receive daily hospitalization counts for the previous month, and we nowcast and backcast the (unobserved) hospitalization counts for each following day in the given month, using the outpatient and inpatient claims signals described above. Note that the period in this scenario covers the Delta wave in the U.S. The start of this hypothetical scenario is chosen as April 1, 2021 so that we have enough data (reported hospitalizations and claims signals) to train our initial regression models for nowcasting. Recall that our data extends back through November 1, 2020, thus our initial training set includes the winter wave of 2020, which helps the initial models capture the relationship between reported hospitalization counts and claims signals.

To make this all more precise, let us introduce some notation. We drop reference to the location ℓ here and in what follows, whenever convenient (whenever it is not needed for the given explanation). Let t_0 be a date that marks the start of a month during the period of April 1, 2021 to November 30, 2021. Then on each day $t = t_0, t_0 + 1, \ldots, t_1 - 1$, where t_1 marks the first day of the next month, we have access to:

- $\{Y_s^{(t_0)}\}_{s \le t_0 - 1}$, hospitalization counts through day $t_0 - 1$, with versions as of day t_0; and
- $\{(I_s^{(t)}, O_s^{(t)})\}_{s \le t}$, outpatient and inpatient signals through day t, with versions as of day t.

On each such day t, we use the data we have to train a regression model, call it f_t, to predict hospitalizations from claims signals. We use this to make nowcasts:

$$\hat{Y}_t^{(t)} = f_t\big(I_t^{(t)}, O_t^{(t)}\big),$$

and lag-k backcasts:

$$\hat{Y}_{t-k}^{(t)} = f_t\big(I_{t-k}^{(t)}, O_{t-k}^{(t)}\big),$$

for each $k = 1, \ldots, 10$. (Clearly, a nowcast is equivalent to a lag-0 backcast.) This is repeated for each of the 9 months in the period spanned by this monthly-update scenario. The details of how regression models are trained and evaluated will be given in the next subsection.

Scenario 2: no updates. Our second hypothetical scenario spans nowcasting dates from December 1, 2021 to August 1, 2023. In this scenario, we receive reported hospitalizations up through November 30, 2021 with versions as of December 1, 2021, and we receive *no further* hospitalization counts after that. As before, we nowcast and backcast the (unobserved) hospitalization counts in the remaining period using the outpatient and inpatient claims signals. Note that the period in this scenario covers the Omicron wave in the U.S., and that this is second scenario is generally far more challenging than the first scenario.

The data received and nowcasts and backcasts made in the no-update scenario can be written in precise notation, in fact, exactly as introduced in the description for the monthly-update scenario above, but now we fix t_0 at December 1, 2021, and make nowcasts and backcasts at each t from t_0 through the end of the period, August 31, 2023. The details of how regression models are trained and evaluated in this scenario will again be covered in the next subsection.

2.3 Regression model

As our basic working model, we predict hospitalizations using a linear combination of a set L^I of lags of the inpatient signal and a set L^O of lags of the outpatient signal,

$$f_t\big(I_s^{(t)}, O_s^{(t)}\big) = \beta_{t,0} + \sum_{j \in L^I} \beta_{t,j}^I \, I_{s-j}^{(t)} + \sum_{j \in L^O} \beta_{t,j}^O \, O_{s-j}^{(t)}. \tag{1}$$

where the coefficients $\beta_{t,0}$, $\beta_{t,j}^I$, and $\beta_{t,j}^O$ are estimated, i.e., the model f_t is fit, by training on historical data available at time t, either separately for each location ℓ (recall, the notational dependence on the location has been dropped for now), or in a way that pools data across locations. Details will be given below. First, we describe how we select the lag sets L^O, L^I for the inpatient and outpatient features.

2.3.1 Selecting feature lags

In order to select the lag sets L^O, L^I for the working model (1), we restrict our attention to training data *before* the first nowcast date of April 1, 2021 (so as to be careful not to overfit to the data in our nowcasting period, and reserve this for proper evaluation of our models in the two scenarios outlined previously).

To guide lag selection, we consider three metrics, which measure predictive power, feature stability, and data availability. Let $X_s^{(t)}$ denote a generic feature in our usual notation for versioned data, i.e., $X_s^{(t)} = I_s^{(t)}$ for the inpatient feature or $X_s^{(t)} = O_s^{(t)}$ for outpatient feature, we consider, as a function of lag j:

- correlation of $X_{t-j}^{(t)}$ with finalized hospitalizations Y_t, which is a measure of predictive power;
- correlation of $X_{t-j}^{(t)}$ with its own finalized value X_{t-j}, which is a measure of stability;
- the fraction of total claims used to compute the finalized signal X_{t-j} that are observed by time t.

Each metric is computed over time (between November 1, 2020 and March 31, 2021), for each location, and the results are displayed in Figure 2, averaged over all locations (all states). The shaded regions display ± 1 standard error bands, computed over the state averages.

We can see that lag 6, for each of the inpatient and outpatient signals, offers a nice balance across the three metrics: maximum predictive power, and high stability and availability. We therefore include lag 6 in each of the sets L^I and L^O. To improve robustness and capture longer-range dependencies, we also include lags 13 and 20 in each of L^I and L^O, which is roughly consistent with choices of feature lags in other basic epidemic prediction models (e.g., McDonald et al. (2021)). The choice of 7-day spacing here is also motivated by the desire to limit correlations between features in (1); recall, the inpatient and outpatient signals are computed using a trailing 7-day window of claims data.

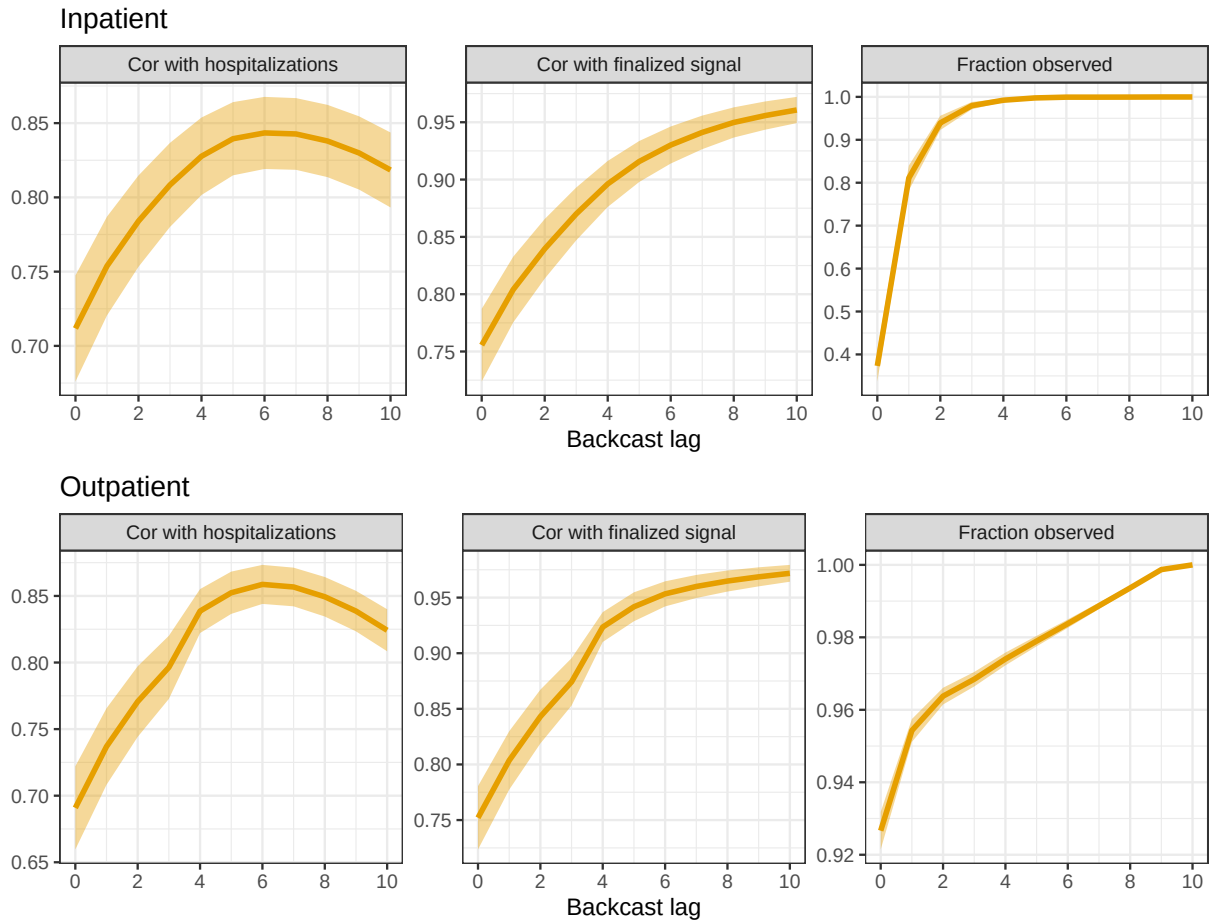

Figure 2: Analysis to help guide lag selection for inpatient and outpatient features in the working model (1). The rows correspond to different features—inpatient or outpatient, and the columns to different metrics, as explained precisely in the main text. The shaded regions show ± 1 standard error bands for each metric, over the state averages.

Lastly, we note that these lag sets have not really been directly optimized for maximum nowcasting and backcasting accuracy. We have avoided doing this for simplicity, and a more direct lag optimization over the training data (pre April 1, 2021) would likely only improve our predictive accuracy in general.

2.3.2 Training the regression model

The coefficients $\beta_{t,0}$, $\beta^I_{t,j}$, $\beta^O_{t,j}$ in (1) are fit by solving the following weighted least squares problem at time t:

$$\underset{\beta_{t,0},\, \beta^I_{t,j},\, \beta^O_{t,j}}{\text{minimize}} \quad \sum_{s<t_0} w_\gamma(t_0 - s) \cdot \left(Y_s^{(t_0)} - \beta_{t,0} - \sum_{j \in L^I} \beta^I_{t,j}\, I_{s-j}^{(t)} - \sum_{j \in L^O} \beta^O_{t,j}\, O_{s-j}^{(t)} \right)^2, \tag{2}$$

where we use exponentially decaying weights:

$$w_\gamma(u) = \exp(-\gamma u) \tag{3}$$

for a tuning parameter $\gamma \geq 0$. The index t_0 in (2) marks the latest observation boundary for hospitalization reports before time t: this is either the start of the month in scenario 1 (where we receive monthly updates), or December 1, 2021 in scenario 2 (where we receive no further updates). Hence, in other words, we fit the coefficients in (2) by minimizing the weighted mean squared error of our working regression model, over all

dates at which response values (reported hospitalization counts) are available. Note carefully that in (2) we only use properly-versioned data that would have been available at t. If any such data, response or feature values, are missing at time t then the corresponding summand is simply omitted from (2).

As a default, we fit the model by solving (2) separately for each location ℓ (recall, we have suppressed the dependence on ℓ in the notation for simplicity). We call this the *state-level* model. Later, we discuss schemes for pooling training data across locations. Next, we focus on the decay parameter γ in (2), (3).

2.3.3 Selecting the decay parameter by cross-validation

Before describing how we select γ in the exponential weights (3) that are used in the weighted least squares problem (2), we pause to discuss the question: why is it useful to use decaying observation weights in the first place? The reason is because the the features—lagged versions of the inpatient and outpatient medical insurance claims signals, and response—reported hospitalizations, need not be jointly stationary, i.e., their relationship may be changing over time.

Looking back at Figure 1, we can already see clear evidence of nonstationarity in the relationship between the inpatient signal and reported hospitalizations. If we were to simply regress reported hospitalizations on this inpatient signal alone, then it looks like a regression coefficient of about 1 would be appropriate for the period April–July 2021, but it would be far too small for the first hospitalization wave starting in December 2020 (and to a lesser extent, too small for the second wave starting in August 2021).

By allowing γ itself change over time, we can adapt to the degree of nonstationarity at any point in time, i.e., we can adapt to the amount of past training data that is relevant for the current prediction. Indeed, we will select γ in a dynamic, time-varying fashion using cross-validation (CV). Given the sequential nature of our prediction problem, we use a version of CV that is purely forward-looking, and is sometimes referred to as *time series cross-validation* in the literature (Hyndman and Athanasopoulos, 2021). This works as follows. Let t_0 denote the index that marks the most recent observation boundary for hospitalization reports, and let t_{-1}, t_{-2} denote the indices that mark the previous two observation boundaries before t_0. For each γ in a grid Γ of tuning parameter values, we carry out the following procedure:

- for each $t \in [t_{-2}, t_{-1})$:
  - fit a regression model by solving:

$$\underset{\beta_{t,0},\, \beta^I_{t,j},\, \beta^O_{t,j}}{\text{minimize}} \quad \sum_{s<t_{-2}} w_\gamma(t_{-2}-s) \cdot \Big(Y_s^{(t_{-2})} - \beta_{t,0} - \sum_{j \in L^I} \beta^I_{t,j}\, I^{(t)}_{s-j} - \sum_{j \in L^O} \beta^O_{t,j}\, O^{(t)}_{s-j} \Big)^2;$$

  - produce backcasts $\hat{Y}_s^{(t)}$, $s \in [t_{-2}, t]$;
- for each $t \in [t_{-1}, t_0)$:
  - fit a regression model by solving:

$$\underset{\beta_{t,0},\, \beta^I_{t,j},\, \beta^O_{t,j}}{\text{minimize}} \quad \sum_{s<t_{-1}} w_\gamma(t_{-1}-s) \cdot \Big(Y_s^{(t_{-1})} - \beta_{t,0} - \sum_{j \in L^I} \beta^I_{t,j}\, I^{(t)}_{s-j} - \sum_{j \in L^O} \beta^O_{t,j}\, O^{(t)}_{s-j} \Big)^2;$$

  - produce backcasts $\hat{Y}_s^{(t)}$, $s \in [t_{-1}, t]$;
- compute the mean absolute error (MAE) of all backcasts made at all times $t \in [t_{-2}, t_0)$, against reported hospitalization counts $Y_t^{(t_0)}$, $t \in [t_{-2}, t_0)$.

In other words, this procedure uses the last 2 months of data as a validation set for tuning γ. Ultimately, we choose $\gamma \in \Gamma$ that minimizes the MAE computed in the last step above. To be clear, after choosing γ in this way, we then fit the model by solving (2) and use this to make backcasts at each time $t \in [t_0, t_1)$, where t_1 is the observation boundary after t_0.

In our experiments, we take the set Γ to be a grid of 25 evenly-spaced values between 0 and $\gamma_{\max}$, where the latter is defined as the value of γ such that the effective sample size of the weight sequence,

$$\frac{\big(\sum_{s<t_0} |w_\gamma(t_0-s)|\big)^2}{\sum_{s<t_0} |w_\gamma(t_0-s)|^2},$$

equals 30. This is chosen as a heuristic upper bound the "reasonable" range of γ values, motivated by the idea using at least (effectively) 1 month of training data helps to avoid fitted models which are too volatile.

2.3.4 Stabilizing predictions by geo-pooling

To borrow strength across locations, we consider fitting the regression model at each time t by pooling data across locations, which we refer to as the *geo-pooled* model. To be precise, instead of solving (2) per location ℓ, here we instead solve:

$$\underset{\beta_{t,0},\,\beta^I_{t,j},\,\beta^O_{t,j}}{\text{minimize}} \quad \sum_{\ell} \sum_{s<t_0} w_\gamma(t_0 - s) \cdot \left(\tilde{Y}^{(t_0)}_{\ell,s} - \beta_{t,0} - \sum_{j \in L^I} \beta^I_{t,j}\, I^{(t)}_{\ell,s-j} + \sum_{j \in L^O} \beta^O_{t,j}\, O^{(t)}_{\ell,s-j} \right)^2. \tag{4}$$

Note that in (4), the training set includes data from all locations ℓ. Importantly, in (4), the response $\tilde{Y}^{(t)}_{\ell,s}$ is the *hospitalization rate* in location ℓ at reference date s, as of time t, which is defined as the hospitalization count per 100,000 people, i.e.,

$$\tilde{Y}^{(t)}_{\ell,s} = 10^5 \cdot \frac{Y^{(t)}_{\ell,s}}{N_\ell},$$

where N_ℓ is the population of location ℓ. The use of rates rather than counts in the pooled regression (4) is critical, because otherwise the coefficients would have different meanings in different locations, and pooling across locations would not make sense. After solving (4), in order to use the fitted model to make predictions of hospitalization counts at a given location ℓ, we would then need to rescale the predictions by $N_\ell/10^5$.

Finally, we also consider a *mixed* model, which linearly combines the state-level $\hat{Y}^{(t),\text{state}}_{\ell,s}$ and geo-pooled $\hat{Y}^{(t),\text{pooled}}_{\ell,s}$ predictions, via

$$\hat{Y}^{(t),\text{mixed}}_{\ell,s} = \alpha\, \hat{Y}^{(t),\text{state}}_{\ell,s} + (1-\alpha)\, \hat{Y}^{(t),\text{pooled}}_{\ell,s}, \tag{5}$$

for a mixing parameter $\alpha \in [0,1]$. We select α, separately per state, using the same cross-validation strategy described previously, where we tune α over a grid of 50 evenly-spaced values between 0 and 1. (We tune over γ separately for each of the state-level and geo-pooled models, and then tune over α.)

2.3.5 Prediction intervals via quantile tracking

In addition to considering point predictions from the models described above, we use residuals from these models, and *quantile tracking* (Angelopoulos et al., 2023), a method from the online conformal literature, to generate prediction intervals. We defer all details and results to Appendix E.

3 Results

This section examines our backcasting results in scenario 1 (monthly-update) and scenario 2 (no-update), through quantitative backcast error analysis and qualitative inspection of the nowcast dynamics over time.

3.1 Scenario 1: backcast error analysis

Figure 3 shows the MAE of all backcasts, as a function of lag $k = 0, \ldots, 10$, from the state-level, geo-pooled, and mixed models over the monthly-update period (which, recall spans April 1, 2021 to November 30, 2021). To be clear, for each model, this is computed by averaging the absolute error of backcasts made to *finalized* hospitalization counts, per state. The left panel displays the average of these state MAE values, and as well as standard error bands. The right panel is similar but normalizes the state MAE values by state population times 10^5, thus considering MAE on the scale of hospitalization rates (rather than counts).

The geo-pooled model is clearly worse in terms of MAE than the mixed and state-level models. On the counts scale (left panel), the state-level model has slightly better MAE than the mixed model; on the rates scale (right panel), the opposite is true. This is because the mixed model generally performs better for smaller states, where predictions tend to be more volatile and shrinking toward the geo-pooled model helps to reduce variance. This is not reflected in the MAE plot on the left panel, since poor backcast performance on small states contributes little when measured in terms of counts.

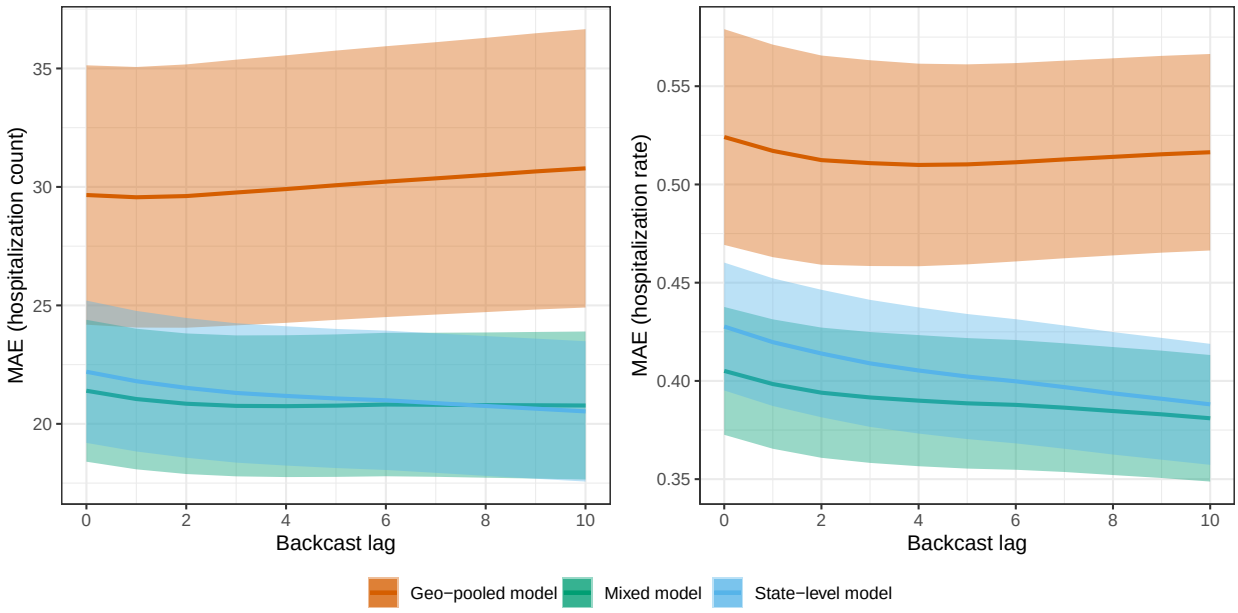

Figure 3: MAE as a function of backcast lag for the state-level, geo-pooled, and mixed models in scenario 1, the monthly-update period. The shaded regions show ±1 standard error bands, over the state MAE values.

In absolute terms, the backcasts made by the state-level and mixed models are generally quite accurate. For example, the nowcasts (backcasts at lag 0) made by the state-level and mixed models have MAEs of 0.428 and 0.405, respectively, on the scale of hospitalization rates. These correspond to proportions of variance explained (PVEs) of 71.6% and 75.4%, respectively. Even the geo-pooled model, which is relatively quite a bit worse, produces nowcasts with an MAE of 0.524 on the scale of hospitalization rates, which corresponds to a still a respectable PVE of 61.1%.

3.2 Scenario 1: illustrative nowcast examples

We examine nowcasts made by the state-level and mixed models during the monthly-update period, in four states: California (CA), Kentucky (KY), Vermont (VT), and New York (NY). The first three are chosen to demonstrate the qualitative behavior of nowcasts in states of different sizes, with CA having the largest population, KY having roughly median population, and VT having the smallest. NY is chosen because it represents somewhat of failure case for the robustness of nowcasts from the state-level model.

California, Kentucky, Vermont. Figure 4 displays state-level and mixed model nowcasts for CA, KY, and VT. To be clear, in each panel of the figure, the nowcasts use *real-time* inpatient and claims signals as predictive features, and the models behind these nowcasts are trained using hospitalization data up to the latest observation boundary (marked by dotted vertical lines). The shaded bars in the figure display *finalized* reported hospitalizations, the target of ultimate interest.

In CA and KY, where the Delta wave is prominent (roughly July–November), we can see that both the state-level and mixed model nowcasts track the dynamics of the Delta wave. For example, looking at July 1, 2021 in CA, hospitalization reports from the previous months (which is all that would have been available at that time) show no indication of an upswing to come. Still, the nowcasts during July present a clear upward trend, which means that policy-makers would know from these nowcasts that a wave is underway. As we see it, evaluating nowcasts for their qualitative shape (in comparison to hospitalization waves) is important—just as much as (if not more so than) numeric error analysis, as in the last subsection.

Being much smaller, VT has on the order of ten reported COVID-19 hospitalizations daily, not several hundreds, as in KY or CA. Accordingly, we can see in Figure 4 that the (finalized) reported hospitalizations curve in VT is much noisier—we do not really see a clear Delta wave (instead, an increasing but quite noisy

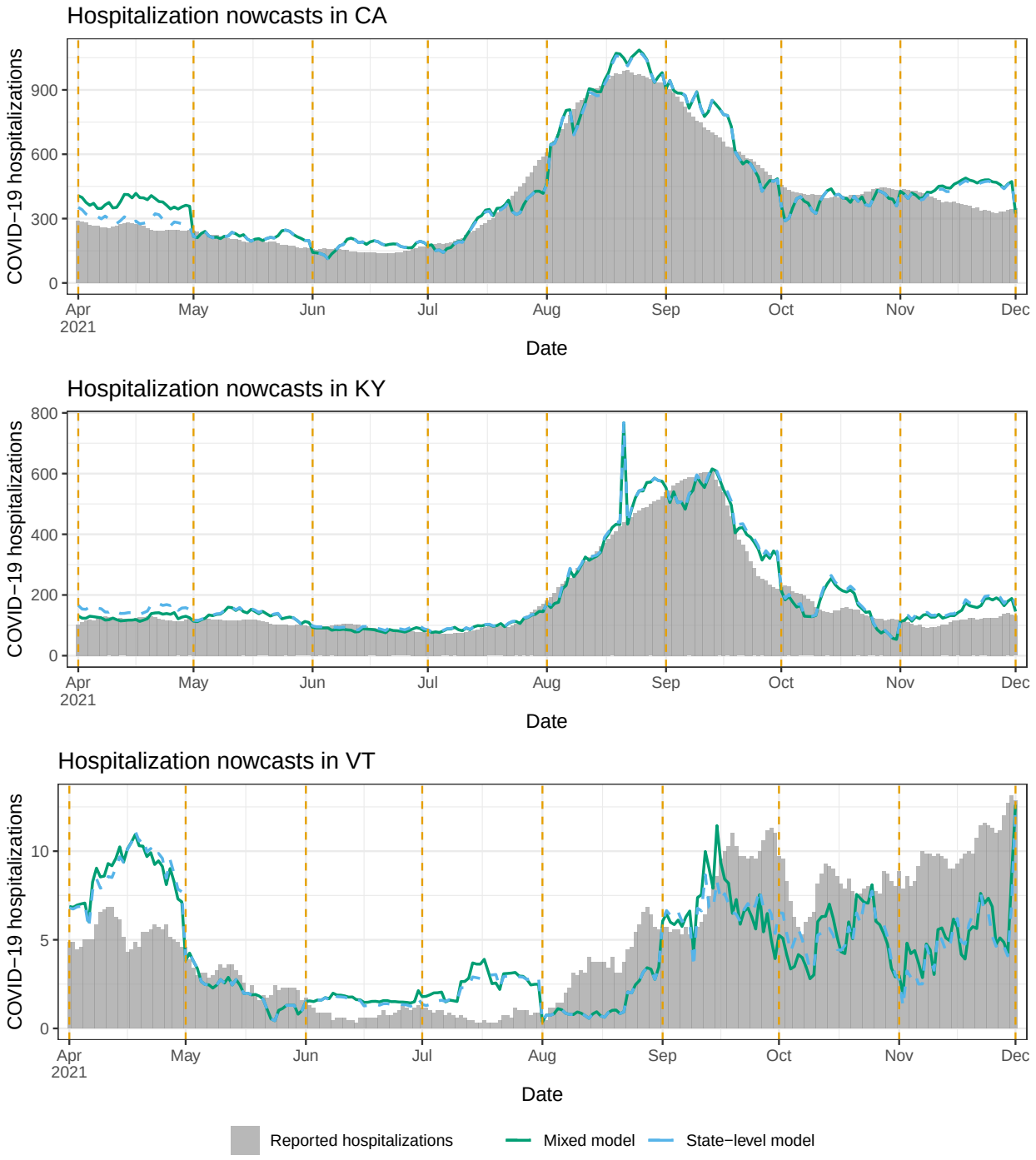

Figure 4: Nowcasts from the state-level and mixed models in scenario 1, the monthly-update period, for CA, KY, and VT. The dotted vertical lines mark the observation boundaries—when hospitalization reports for the previous month are received, and models are retrained.

trend from August through December). The claims signals are also more noisy in a small state like VT, since they are ratios of small counts. Therefore, both the response and features are more volatile in the prediction problem for VT, and not surprisingly, the nowcasts here appear qualitatively worse. That said, the nowcasts still roughly track the gross trends in reported hospitalizations: higher in April and May, lower in June and July, and growing again in August onward.

Generally, the discussion above translates to the rest of the US: nowcasts tend to be in good qualitative agreement with hospitalization waves in larger states, and less so in smaller states. Appendix C provides a full set of nowcast and backcast plots in all 50 states, from the mixed model, under scenario 1. Occasionally, we find that the trend in the predictions disagrees with that in hospitalization reports in a given month, but this gets corrected at the next observation boundary, once new training data is available and the regression model is refit. We also find that backcasts at lag 5 or 10 tend to be smoother than nowcasts.

New York. Figure 4 displays state-level and mixed model nowcasts for NY. This is a notable example of a failure in robustness of the state-level model: its nowcasts for a good part of the month of June are actually *negative*, and are truncated at zero for visualization purposes. This is due to a large and systematic revision that occurred on June 8 for the outpatient signal, where this signal's values for reference dates in May were revised upward. Moreover, when the state-level model is fit to data through the end of May, the regression coefficient on the largest outpatient feature lag ends up being negative, and consequently the upward revision of past feature values on June 8 introduces a strong downward bias in the subsequent predictions. Supporting analysis for this explanation is given in Appendix B.

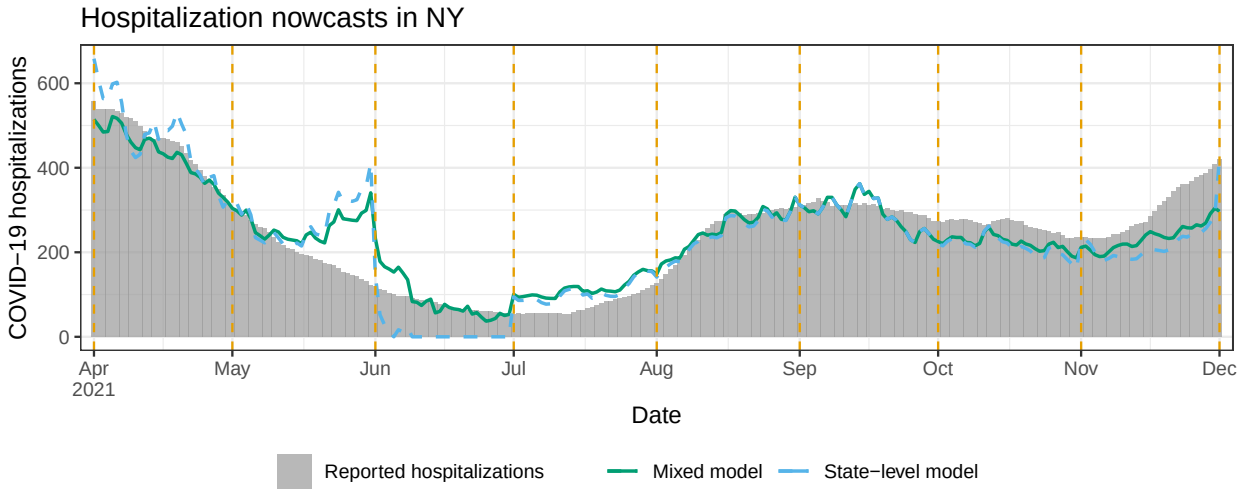

Figure 5: Nowcasts from the state-level and mixed models in scenario 1, the monthly-update period, for NY. The dotted vertical lines mark the observation boundaries, as in Figure 4.

As we can see in the figure, the mixed model produces nowcasts that are much more stable in the month of June. However, the mixed model and even the geo-pooled model are themselves not immune to large and erratic revisions in the input features at prediction time. These revisions can result in erratic nowcasts and backcasts. To mitigate this, we could turn to models that fuse information across a wider variety of auxiliary signals (where ideally, some of these signals would have less severe backfill than claims-based signals), an idea that we return to in the discussion section.

3.3 Scenario 2: backcast error analysis

Figure 6 shows the MAE of all backcasts, as a function of lag $k = 0, \ldots, 10$, from the state-level, geo-pooled, and mixed models over the no-update period (spanning December 1, 2021 to August 31, 2023). The format is as in Figure 3. We see two notable differences compared to the results in the monthly-update period. First, as expected, each model performs worse than it did in Figure 6. On the scale of hospitalization rates (right

panel), the range of MAEs has jumped from (roughly) 0.4–0.53 in the monthly-update period to 0.62–0.72 in the no-update period, and correspondingly, the PVEs have dropped from (roughly) 61–75% to 10–37%.

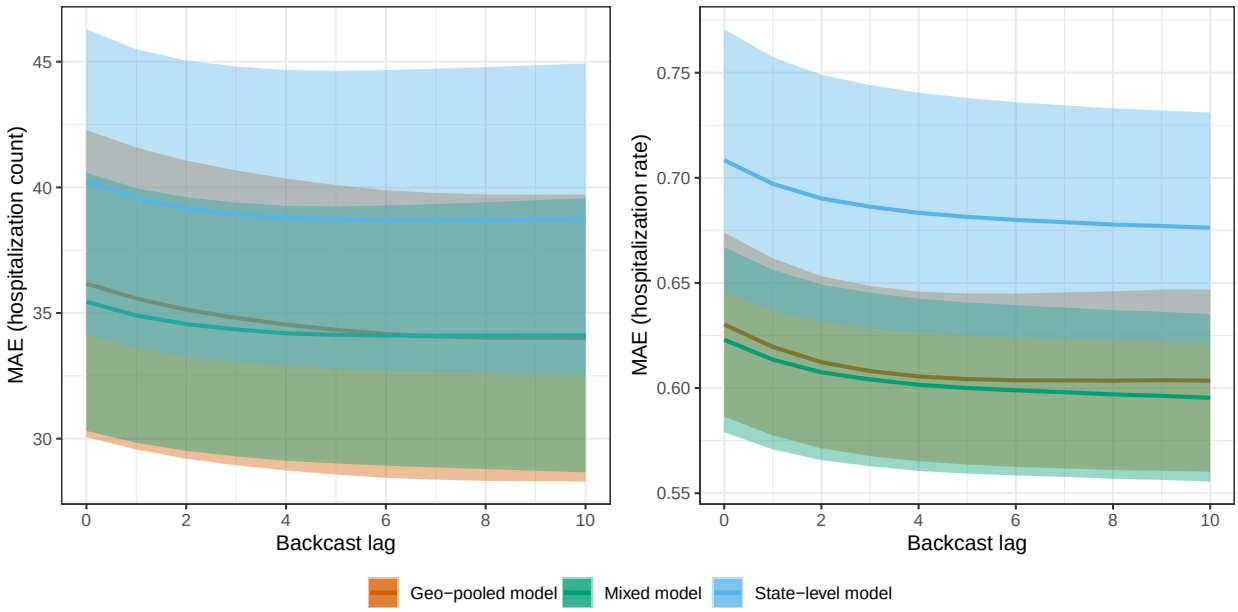

Figure 6: MAE as a function of backcast lag for the state-level, geo-pooled, and mixed models in scenario 2, the no-update period. The shaded regions show ±1 standard error bands, over the state MAE values.

Second (and perhaps a bit more surprising), we see in the no-update period that the state-level model performs clearly worse than the geo-pooled and mixed models when MAE is measured either on the counts or rates scale. The mixed model MAE is somewhat better than the geo-pooled model MAE with respect to counts, while the two are basically the same with respect to rates. It is encouraging to see that the mixed model performs competitively across both monthly-update and no-update scenarios: this model—equipped with the CV procedure for tuning α—is able to effectively adapt the strengths of local and global training approaches (underlying the state-level and geo-pooled models) to the task at hand, in order to yield accurate predictions in an average-case sense.

3.4 Scenario 2: illustrative nowcast examples

We examine nowcasts made by the geo-pooled and mixed models for CA, KY, and VT. NY, which served as failure case in the monthly-update period, is not shown here, because KY itself provides such an example: as we will see, the mixed model lacks robustness over a part of the Omicron wave (meanwhile, the mixed model performs fine for NY throughout the no-update period—as shown in the appendix).

Figure 7 displays the nowcasts for CA, KY, and VT, with the same general format as in Figure 4. CA is a clear success case: both geo-pooled and mixed models capture the dynamics of the Omicron wave faithfully, and also track the summer 2023 and winter 2023 waves, despite (recall) receiving no reported hospitalizations past December 1, 2022. VT, as in the monthly-update period, is a challenging case because it corresponds to much noisier prediction problem, and the nowcasts here look qualitatively worse overall. However, the mixed model nowcasts still pick up the Omicron wave. KY represents a failure case: the mixed model nowcasts for all of February are *negative* and truncated at zero for the visualization. What this is actually demonstrating is a failure of the state-level model, and simultaneously, a failure of tuning of the mixing parameter α. The mixed model here ends up placing a large weight on state-level predictions (not shown), which are themselves volatile for reasons similar to what happens in NY during the monthly-update period—large revisions to the input features at prediction time.

The volatility of the mixed model nowcasts in KY during Omicron provides an important perspective: in the MAE sense, we found in Figure 6 that the mixed model (with its CV tuning for α) successfully navigates

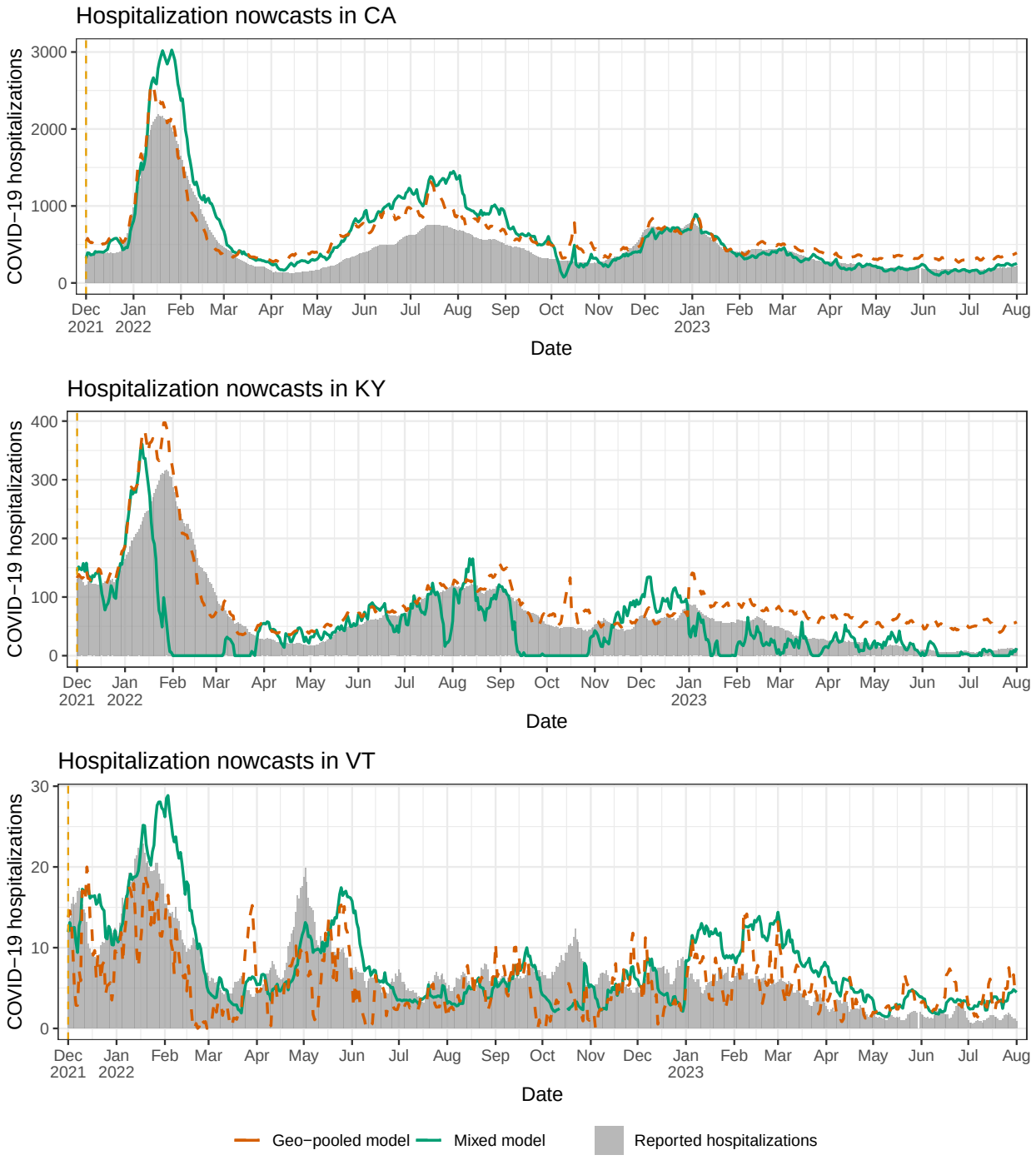

Figure 7: Nowcasts from the geo-pooled and mixed models in scenario 2, the no-update period, for CA, KY, and VT. The dotted vertical line (on the left side of the plot) marks the sole observation boundary in this period—hospitalization reports for all dates prior to this are available, but no reports are received after that.

the strengths and weaknesses of the geo-pooled and state-level models; and yet for specific states and periods of time, the mixed model can still lack robustness. To be clear, such fragility is not limited to the no-update period, and it could have happened in NY in the monthly-update period, had the CV tuning procedure for α not downweighted the state-level predictions so heavily. Improving robustness is a direction for future work, and we revisit this topic in the discussion section.

Lastly, in Appendix D, we again provide a full set of nowcast and backcast plots in all 50 states, from the mixed model, under scenario 2. For larger states, we find that the nowcasts generally trace out the Omicron wave, but for smaller states, this happens less consistently and the nowcasts look more noisy. Backcasts at lag 5 or 10 tend to smooth out the nowcasts, though not dramatically.

3.5 Ablation study

To examine the importance of some of our modeling choices (as described in the methods section), we carry out an ablation study in which we remove a particular component of the model, use a simpler alternative in its place, and evaluate the result in terms of MAE. In particular, we consider four ablated models:

1. Unweighted, all past: we fit the regression model (2) without observation weights, on all past reported hospitalization data available (considering summands $s < t$ in (2)).
2. Unweighted, two months: we fit the regression model (2) without observation weights, on the reported hospitalization data available for the latest two months (considering summands $s \in [t_{-2}, t_0)$ in (2)).
3. Weighted, inpatient only: we fit the regression model (2) using only the lags of the inpatient signal.
4. Weighted, outpatient only: we fit the regression model (2) using only the lags of the outpatient signal.

Tables 1 and 2 compare the performance of the state-level model to these ablated models, for scenarios 1 and 2, respectively. In each scenario, for each model, we compute its MAE over the range of time values in the scenario, per backcast lag $k = 0, \ldots, 10$ and state, then we report the average and standard error of these MAE values across all backcast lags and states. The state-level and inpatient-only models perform the best overall, and make up the two best models in either scenario. Meanwhile, the all-past model is slightly worse and the two-month model is significantly worse, which emphasizes the importance of decaying weights (and CV tuning) as part of the original model design. The outpatient-only model is competitive in scenario 1 but not in scenario 2. This could be a reflection of distribution-drift in the relationship between the outpatient signal and hospitalization reports post Omicron, which in scenario 2 is compounded by our inability to retrain the regression model in order to account for such drift.

4 Discussion

This paper demonstrates that relatively simple regression models and medical insurance claims data can be used to provide accurate real-time estimates of reported COVID-19 hospitalizations, in different hypothetical scenarios in which hospitalization reporting is dramatically reduced in frequency or shut down entirely. Of course, we are not advocating for nowcasts from claims data to replace traditional public health surveillance. However, leveraging statistical models which use auxiliary data for nowcasting, so that we may then reduce reporting frequency and thus reduce the burden that reporting entails, may be a favorable tradeoff for public health agencies to consider.

The claims data used in our work has been available (and was published in the Delphi Epidata repository) in real-time, for essentially the entirety of the COVID-19 pandemic. In principle, medical insurance claims cover a broad swath of health conditions and COVID-19 is certainly not the only worth target for nowcasting systems built on top of claims data. Nowcasts of influenza hospitalizations, for example, would also be useful to public health decision-makers, and could be generated via analogous regression models using proxy claims signals. Whether a method like ours could be used operationally in a future pandemic is a complex issue and depends on many factors (including continued real-time access to medical insurance claims data). That said, one can argue that methods which tend to be the most reliable in times of emergency are the ones that have

MAE (hospitalization count)	
Model	MAE (Standard error)
All-past	23.74 (1.03)
Two-month	51.79 (2.83)
State-level	21.18 (0.88)
Inpat-only	22.11 (0.88)
Outpat-only	23.17 (1.14)

MAE (hospitalization rate)	
Model	MAE (Standard error)
All-past	0.44 (0.01)
Two-month	0.86 (0.02)
State-level	0.41 (0.01)
Inpat-only	0.42 (0.01)
Outpat-only	0.43 (0.01)

Table 1: MAE results from the state-level and ablated models, averaged over all backcast lags and all states, in scenario 1.

MAE (hospitalization count)	
Model	MAE (Standard error)
All-past	47.65 (3.00)
Two-month	90.01 (5.82)
State-level	38.98 (1.78)
Inpat-only	33.52 (1.46)
Outpat-only	47.81 (2.11)

MAE (hospitalization rate)	
Model	MAE (Standard error)
All-past	0.72 (0.02)
Two-month	1.58 (0.05)
State-level	0.70 (0.02)
Inpat-only	0.70 (0.03)
Outpat-only	0.89 (0.02)

Table 2: MAE results from the state-level and ablated models, averaged over all backcast lags and all states, in scenario 2.

been developed and tested already. This motivates us to continue working on proof-of-concept systems like the one in this paper.

Medical insurance claims are certainly not the only relevant auxiliary data stream for tracking COVID-19 hospitalizations, and this analysis may be repeated with a number of other auxiliary signals. In operational systems in public health, robustness is arguably more important than average-case performance, and failure examples of the proposed nowcasting methods, as seen in the results section (NY in scenario 1, and KY in scenario 2) would likely be concerning to public health decision-makers. These failure examples were driven by large revisions to the claims signals which occurred at particular points in time. Any model which makes predictions based on a linear combination of a set of features can suffer from erratic behavior if these features are subject to large fluctuations at prediction time. Combining multiple nowcasts built from different auxiliary signals is a way to improve robustness, especially when some of these signals are more stable and less subject to heavy revisions. Signals derived from electronic medical records (EMR), medical device data, and internet search queries all typically have less backfill compare to medical insurance claims signals.

Model combination methods—which also go by names such as: aggregation, ensembling, or fusion—have been quite successful in influenza nowcasting systems in pre-pandemic years (e.g., Farrow (2016); Jahja et al. (2019)). Ensembles of COVID-19 forecasts have likewise demonstrated considerable robustness compared to the constituent forecasters (e.g., Cramer et al. (2022); Ray et al. (2023)). An important direction for future work is to incorporate similar ideas into the nowcasting settings considered in this paper, in an effort to move toward more reliable and robust systems.

Acknowledgements

We would like to thank members of the Delphi research group for valuable feedback, and Change Healthcare and Optum/United Health Group for their invaluable data partnership and their collaboration. This work was supported by Centers for Disease Control and Prevention (CDC) grant no. 75D30123C15907.

References

A. F. Ackley, S. Pilewski, V. S. Petrovic, L. Worden, E. Murray, and T. C. Porco. Assessing the utility of a smart thermometer and mobile application as a surveillance tool for influenza and influenza-like illness. *Health Informatics Journal*, 26(3):2148–2158, 2020.

A. N. Angelopoulos, E. J. Candès, and R. J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, 2023.

L. S. Bastos, T. Economou, M. F. C. Gomes, D. A. M. Villela, F. C. Coelho, O. G. Cruz, O. Stoner, T. Bailey, and C. T. Codeço. A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine*, 38(22):4363–4377, 2019.

F. Bergström, F. Günther, M. Höhle, and T. Britton. Bayesian nowcasting with leading indicators applied to COVID-19 fatalities in Sweden. *PLOS Computational Biology*, 18(12):e1010767, 2022.

L. C. Brooks. *Pancasting: Forecasting epidemics from provisional data.* PhD thesis, Carnegie Mellon University, 2020.

E. Y. Cramer, E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. C. Rivadeneira, A. Gerding, T. Gneiting, K. H. House, Y. Huang, D. Jayawardena, A. H. Kanji, A. Khandelwal, K. Le, A. Mühlemann, J. Niemi, A. Shah, A. Stark, Y. Wang, N. Wattanachit, M. W. Zorn, Y. Gu, S. Jain, N. Bannur, A. Deva, M. Kulkarni, S. Merugu, A. Raval, S. Shingi, A. Tiwari, J. White, S. Woody, M. Dahan, S. Fox, K. Gaither, M. Lachmann, L. A. Meyers, J. G. Scott, M. Tec, A. Srivastava, G. E. George, J. C. Cegan, I. D. Dettwiller, W. P. England, M. W. Farthing, R. H. Hunter, B. Lafferty, I. Linkov, M. L. Mayo, M. D. Parno, M. A. Rowland, B. D. Trump, S. M. Corsetti, T. M. Baer, M. C. Eisenberg, K. Falb, Y. Huang, E. T. Martin, E. McCauley, R. L. Myers, T. Schwarz, D. Sheldon, G. C. Gibson, R. Yu, L. Gao, Y. Ma, D. Wu, X. Yan, X. Jin, Y.-X. Wang, Y. Chen, L. Guo, Y. Zhao, Q. Gu, J. Chen, L. Wang, P. Xu, W. Zhang, D. Zou, H. Biegel, J. Lega, T. L. Snyder, D. D. Wilson, S. McConnell, R. Walraven, Y. Shi, X. Ban, Q.-J. Hong, S. Kong, J. A. Turtle, M. Ben-Nun, P. Riley, S. Riley, U. Koyluoglu, D. DesRoches, B. Hamory, C. Kyriakides, H. Leis, J. Milliken, M. Moloney, J. Morgan, G. Ozcan, C. Schrader, E. Shakhnovich, D. Siegel, R. Spatz, C. Stiefeling, B. Wilkinson, A. Wong, Z. Gao, J. Bian, W. Cao, J. L. Ferres, C. Li, T.-Y. Liu, X. Xie, S. Zhang, S. Zheng, A. Vespignani, M. Chinazzi, J. T. Davis, K. Mu, A. P. y. Piontti, X. Xiong, A. Zheng, J. Baek, V. Farias, A. Georgescu, R. Levi, D. Sinha, J. Wilde, N. D. Penna, L. A. Celi, S. Sundar, S. Cavany, G. España, S. Moore, R. Oidtman, A. Perkins, D. Osthus, L. Castro, G. Fairchild, I. Michaud, D. Karlen, E. C. Lee, J. Dent, K. H. Grantz, J. Kaminsky, K. Kaminsky, L. T. Keegan, S. A. Lauer, J. C. Lemaitre, J. Lessler, H. R. Meredith, J. Perez-Saez, S. Shah, C. P. Smith, S. A. Truelove, J. Wills, M. Kinsey, R. Obrecht, K. Tallaksen, J. C. Burant, L. Wang, L. Gao, Z. Gu, M. Kim, X. Li, G. Wang, Y. Wang, S. Yu, R. C. Reiner, R. Barber, E. Gaikedu, S. Hay, S. Lim, C. Murray, D. Pigott, B. A. Prakash, B. Adhikari, J. Cui, A. Rodríguez, A. Tabassum, J. Xie, P. Keskinocak, J. Asplund, A. Baxter, B. E. Oruc, N. Serban, S. O. Arik, M. Dusenberry, A. Epshteyn, E. Kanal, L. T. Le, C.-L. Li, T. Pfister, D. Sava, R. Sinha, T. Tsai, N. Yoder, J. Yoon, L. Zhang, S. Abbott, N. I. Bosse, S. Funk, J. Hellewel, S. R. Meakin, J. D. Munday, K. Sherratt, M. Zhou, R. Kalantari, T. K. Yamana, S. Pei, J. Shaman, T. Ayer, M. Adee, J. Chhatwal, O. O. Dalgic, M. A. Ladd, B. P. Linas, P. Mueller, J. Xiao, M. L. Li, D. Bertsimas, O. S. Lami, S. Soni, H. T. Bouardi, Y. Wang, Q. Wang, S. Xie, D. Zeng, A. Green, J. Bien, A. J. Hu, M. Jahja, B. Narasimhan, S. Rajanala, A. Rumack, N. Simon, R. J. Tibshirani, R. Tibshirani, V. Ventura, L. Wasserman, E. B. O'Dea, J. M. Drake, R. Pagano, J. W. Walker, R. B. Slayton, M. Johansson, M. Biggerstaff, and N. G. Reich. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022.

Department of Health and Human Services. COVID-19 guidance for hospital reporting and FAQs for hospitals, hospital laboratory, and acute care facility data reporting, 2023. URL https://www.hhs.gov/sites/default/files/covid-19-faqs-hospitals-hospital-laboratory-acute-care-facility-data-reporting.pdf.

K. Durizzo, K. Harttgen, F. Tediosi, M. Sahu, A. Kuwawenaruwa, P. Salari, and I. Günther. Toward mandatory health insurance in low-income countries? An analysis of claims data in Tanzania. *Health Economics*, 31(10):2187–2207, 2022.

D. C. Farrow. *Modeling the past, present, and future of influenza.* PhD thesis, Carnegie Mellon University, 2016.

D. C. Farrow, L. C. Brooks, R. J. Tibshirani, and R. Rosenfield. Delphi Epidata API, 2015. URL https://github.com/cmu-delphi/delphi-epidata.

J. Geng, X. Chen, J. Shi, H. Bao, Q. Chen, and H. Yu. Assessment of the satisfaction with public health insurance programs by patients with chronic diseases in China: A structural equation modeling approach. *BMC Public Health*, 21(1), 2021.

R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice.* OTexts, third edition, 2021.

M. Jahja, D. Farrow, R. Rosenfeld, and R. J. Tibshirani. Kalman Filter, sensor fusion, and constrained regression: Equivalences and insights. In *Advances in Neural Information Processing Systems*, 2019.

M. Jahja, A. Chin, and R. J. Tibshirani. Real-time estimation of COVID-19 infections: Deconvolution and sensor fusion. *Statistical Science*, 37(2):207–228, 2021.

Y.-S. Jung, Y.-E. Kim, D.-S. Go, R. Munkhzul, J. Jung, and S.-J. Yoon. Associations between private health insurance and medical care utilization for musculoskeletal disorders: Using the Korea Health Panel Survey Data for 2014 to 2015. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 57:004695802098146, 2020.

S. I. Leuba, R. Yaesoubi, M. Antillon, T. Cohen, and C. Zimmer. Tracking and predicting U.S. influenza activity with a real-time surveillance network. *PLOS Computational Biology*, 16(11):1–14, 11 2020.

S. Li and Y. Yang. An empirical study on the influence of the basic medical insurance for urban and rural residents on Family Financial Asset Allocation. *Frontiers in Public Health*, 9, 2021.

T. Li and L. F. White. Bayesian back-calculation and nowcasting for line list data during the COVID-19 pandemic. *PLOS Computational Biology*, 17(7):e1009210, 2021.

A. Lison, S. Abbott, J. Huisman, and T. Stadler. Generative Bayesian modeling to nowcast the effective reproduction number from line list data with missing symptom onset dates. *PLOS Computational Biology*, 20(4):e1012021, 2024.

D. J. McDonald, J. Bien, A. Green, A. J. Hu, N. DeFries, S. Hyun, N. L. Oliveira, J. Sharpnack, J. Tang, R. Tibshirani, V. Ventura, L. Wasserman, and R. J. Tibshirani. Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction? *Proceedings of the National Academy of Sciences*, 118(51):e2111453118, 2021.

T. F. Menkir, H. Cox, C. Poirier, M. Saul, S. Jones-Weekes, C. Clementson, P. M. de Salazar, M. Santillana, and C. O. Buckee. A nowcasting framework for correcting for reporting delays in Malaria surveillance. *PLOS Computational Biology*, 17(11):e1009570, 2021.

A. C. Miller, L. A. Hannah, J. Futoma, N. J. Foti, E. B. Fox, A. D'Amour, M. Sandler, R. A. Saurous, and J. A. Lewnard. Statistical deconvolution for inference of infection time series. *Epidemiology*, 33(4):470–479, 2022.

T. Mori, J. Komiyama, T. Fujii, M. Sanuki, K. Kume, G. Kato, Y. Mori, H. Ueshima, H. Matsui, N. Tamiya, and T. Sugiyama. Medical expenditures for fragility hip fracture in Japan: a study using the Nationwide Health Insurance Claims Database. *Archives of Osteoporosis*, 17(1), 2022.

T. Nakayama, Y. Imanaka, Y. Okuno, G. Kato, T. Kuroda, R. Goto, S. Tanaka, H. Tamura, S. Fukuhara, S. Fukuma, M. Muto, M. Yanagita, and Y. Yamamoto. Analysis of the evidence-practice gap to facilitate proper medical care for the elderly: Investigation, using databases, of utilization measures for National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB). *Environmental Health and Preventive Medicine*, 22(1), 2017.

T. P. Nickels. RE: CMS-3401-IFC, 2020. URL https://www.aha.org/system/files/media/file/2020/11/aha-comment-cms-aug-25-interim-final-rule-on-covid-19-data-reporting-letter-11-2-20.pdf.

R. Panczak, V. von Wyl, O. Reich, X. Luta, M. Maessen, A. E. Stuck, C. Berlin, K. Schmidlin, D. C. Goodman, M. Egger, K. Clough-Gorr, and M. Zwahlen. Death at no cost? Persons with no health insurance claims in the last year of life in Switzerland. *BMC Health Services Research*, 18(1), 2018.

J. M. Radin, N. E. Wineinger, E. J. Topol, and S. R. Steinhubl. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: A population-based study. *The Lancet Digital Health*, 2 (2):e85–e93, 2020.

E. L. Ray, L. C. Brooks, J. Bien, M. Biggerstaff, N. I. Bosse, J. Bracher, E. Y. Cramer, S. Funk, A. Gerding, M. A. Johansson, A. Rumack, Y. Wang, M. Zorn, R. J. Tibshirani, and N. G. Reich. Comparing trained and untrained probabilistic ensemble forecasts of COVID-19 cases and deaths in the United States. *International Journal of Forecasting*, 39(3):1366–1383, 2023.

A. Reinhart, L. Brooks, M. Jahja, A. Rumack, J. Tang, S. Agrawal, W. A. Saeed, T. Arnold, A. Basu, J. Bien, Á. A. Cabrera, A. Chin, E. J. Chua, B. Clark, S. Colquhoun, N. DeFries, D. C. Farrow, J. Forlizzi, J. Grabman, S. Gratzl, A. Green, G. Haff, R. Han, K. Harwood, A. J. Hu, R. Hyde, S. Hyun, A. Joshi, J. Kim, A. Kuznetsov, W. L. Motte-Kerr, Y. J. Lee, K. Lee, Z. C. Lipton, M. X. Liu, L. Mackey, K. Mazaitis, D. J. McDonald, P. McGuinness, B. Narasimhan, M. P. O'Brien, N. L. Oliveira, P. Patil, A. Perer, C. A. Politsch, S. Rajanala, D. Rucker, C. Scott, N. H. Shah, V. Shankar, J. Sharpnack, D. Shemetov, N. Simon, B. Y. Smith, V. Srivastava, S. Tan, R. Tibshirani, E. Tuzhilina, A. K. V. Nortwick, V. Ventura, L. Wasserman, B. Weaver, J. C. Weiss, S. Whitman, K. Williams, R. Rosenfeld, and R. J. Tibshirani. An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51):e2111452118, 2021.

R. Rosenfeld and R. J. Tibshirani. Epidemic tracking and forecasting: Lessons learned from a tumultuous year. *Proceeding of the National Academy of Sciences*, 118(51):e2111456118, 2021.

M. Sakai, S. Ohtera, T. Iwao, Y. Neff, G. Kato, Y. Takahashi, and T. Nakayama. Validation of claims data to identify death among aged persons utilizing enrollment data from Health Insurance Unions. *Environmental Health and Preventive Medicine*, 24(1), 2019.

P. M. D. Salazar, F. Lu, J. A. Hay, D. Gómez-Barroso, P. Fernández-Navarro, E. V. Martínez, J. Astray-Mochales, R. Amillategui, A. García-Fulgueiras, M. D. Chirlaque, A. Sánchez-Migallón, A. Larrauri, M. J. Sierra, M. Lipsitch, F. Simón, M. Santillana, and M. A. Hernán. Near real-time surveillance of the SARS-CoV-2 epidemic with incomplete data. *PLOS Computational Biology*, 18(3):e1009964, 2022.

M. Santillana, A. T. Nguyen, T. Louie, A. Zink, J. Gray, I. Sung, and J. S. Brownstein. Cloud-based electronic health records for real-time, region-specific influenza surveillance. *Scientific Reports*, 6(1):1–8, 2016.

S. R. Seaman, P. Samartsidis, M. Kall, and D. De Angelis. Nowcasting COVID-19 deaths in England by age and region. *Journal of the Royal Statistical Society: Series C*, 71(5):1266–1281, 2022.

M. S. Smolinski, A. W. Crawley, K. Baltrusaitis, R. Chunara, J. M. Olsen, O. Wójcik, M. Santillana, A. Nguyen, and J. S. Brownstein. Flu Near You: Crowdsourced symptom reporting spanning 2 influenza seasons. *American Journal of Public Health*, 105(10):2124–2130, 2015.

S. O. Song, E. Han, K. J. Son, B.-S. Cha, and B.-W. Lee. Age at mortality in patients with type 2 diabetes who underwent kidney transplantation: An analysis of data from the Korean National Health Insurance and Statistical Information Service, 2006 to 2018. *Journal of Clinical Medicine*, 12(9), 2023.

C. Viboud, V. Charu, D. Olson, S. Ballesteros, J. Gog, F. Khan, B. Grenfell, and L. Simonsen. Demonstrating the use of high-volume electronic medical claims data to monitor local and regional influenza activity in the US. *PLOS ONE*, 9(7):1–12, 07 2014.

D. Wolffram, S. Abbott, M. an der Heiden, S. Funk, F. Günther, D. Hailer, S. Heyder, T. Hotz, J. van de Kassteele, H. Küchenhoff, S. Müller-Hansen, D. Syliqi, A. Ullrich, M. Weigert, M. Schienle, and J. Bracher. Collaborative nowcasting of COVID-19 hospitalization incidences in Germany. *PLOS Computational Biology*, 19(8):e1011394, 2023.

C.-Y. Yang, R.-J. Chen, W.-L. Chou, Y.-J. Lee, and Y.-S. Lo. An integrated influenza surveillance framework based on national influenza-like illness incidence and multiple hospital electronic medical records for early prediction of influenza epidemics: Design and evaluation. *Journal of Medical Internet Research*, 21(2):e12341, 2019.

S. Yang, M. Santillana, and S. C. Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.

Q. Yao, H. Li, and C. Liu. Use of social health insurance for hospital care by internal migrants in China—Evidence from the 2018 China migrants dynamic survey. *Frontiers in Public Health*, 10, 2022.

L. Zheng and L. Peng. Effect of major illness insurance on vulnerability to poverty: Evidence from China. *Frontiers in Public Health*, 9, 2021.

A More details on claims signals

For the outpatient signal, any claim that has a primary ICD-10 code of U07.1, B97.21, or B97.29 is counted as confirmed COVID. For the inpatient signal, any claim with a primary ICD-10 code of U07.1, U07.2, B97.29, J12.81, Z03.818, B34.2, or J12.89 is counted as COVID-associated.

These definitions are consistent with the definitions of analogous outpatient and inpatient signals in the Delphi Epidata API. However, the signals that we use in our analysis differ from those in the Delphi Epidata API, at the time that this paper was written, in a way that relates to smoothing: the signals in the API are smoothed with a more sophisticated smoothing approach (which also explicitly adjusts for weekday/weekend differences), whereas the signals in this paper are smoothed via 7-day pooling (which implicitly accounts for weekday/weekend differences).

B Scenario 1: further investigation for NY

We present further analysis as to why state-level model for NY produces negative nowcasts over a portion of the month of June 2021, in scenario 1 (recall Figure 5 in the main paper). Figure 8 displays two versions of the outpatient signal over the month of May 2021, corresponding to issue dates of June 7 and June 8. We can see that a systematic upward revision of all signal values occurs on June 8. The nowcasting model, fit by training on data through the end of May (which sees the revision on June 8), places a sizeable negative weight on the largest lag of the outpatient feature; once the upward revision occurs on June 8, the nowcasts on that and subsequent days display a significant downward trend. This is confirmed by Figure 9, where we decompose the nowcasts made in June into the contributions from each feature (simply the coefficient times the feature value). After June 7, we see that `out_20`, the largest lag of the outpatient figure, is responsible for driving the downward trend.

C Scenario 1: full set of backcasts

Figures 10–26 display backcasts at lag 0 (i.e., nowcasts), 5, and 10, for all 50 US states and DC, made by the mixed model in scenario 1, the monthly-update period. The format follows Figure 4 in the main paper.

D Scenario 2: full set of backcasts

Figures 27–43 display backcasts at lag 0 (i.e., nowcasts), 5, and 10, for all 50 US states and DC, made by the mixed model in scenario 2, the no-update period. The format follows Figure 7 in the main paper.

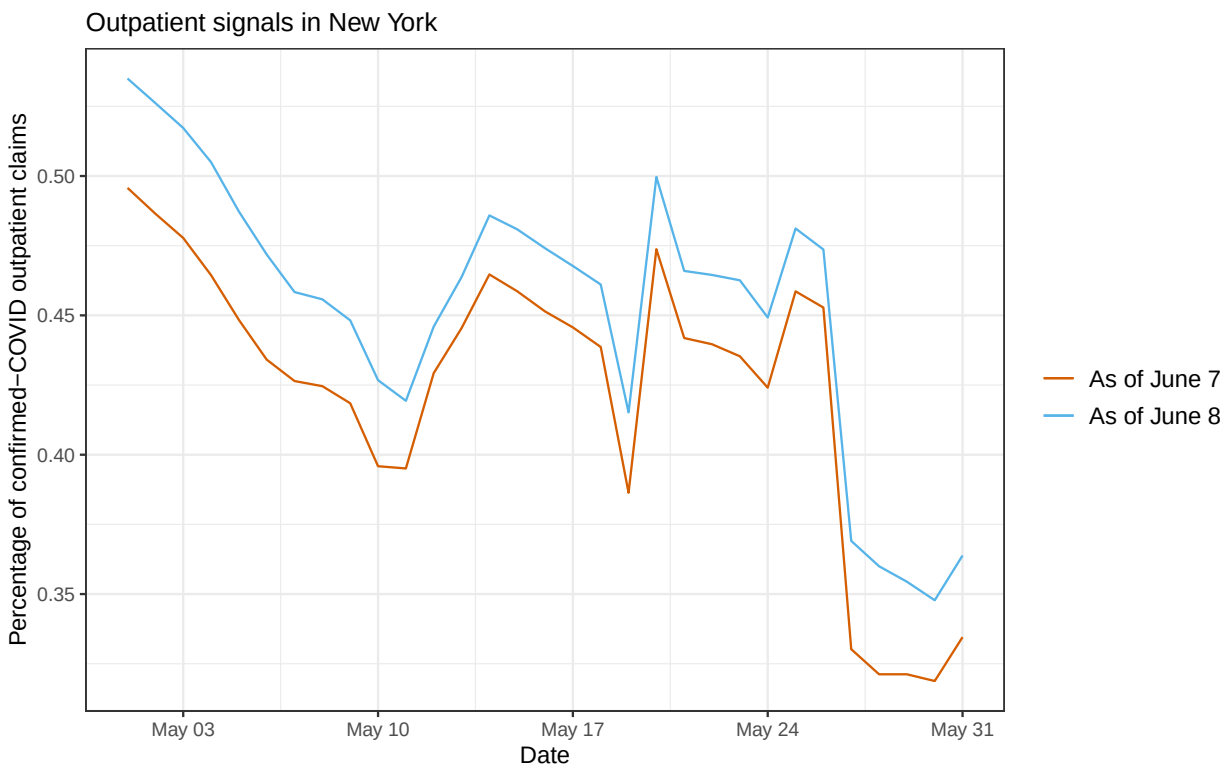

Figure 8: Two versions of the outpatient signal during May 2021, as of June 7 and June 8.

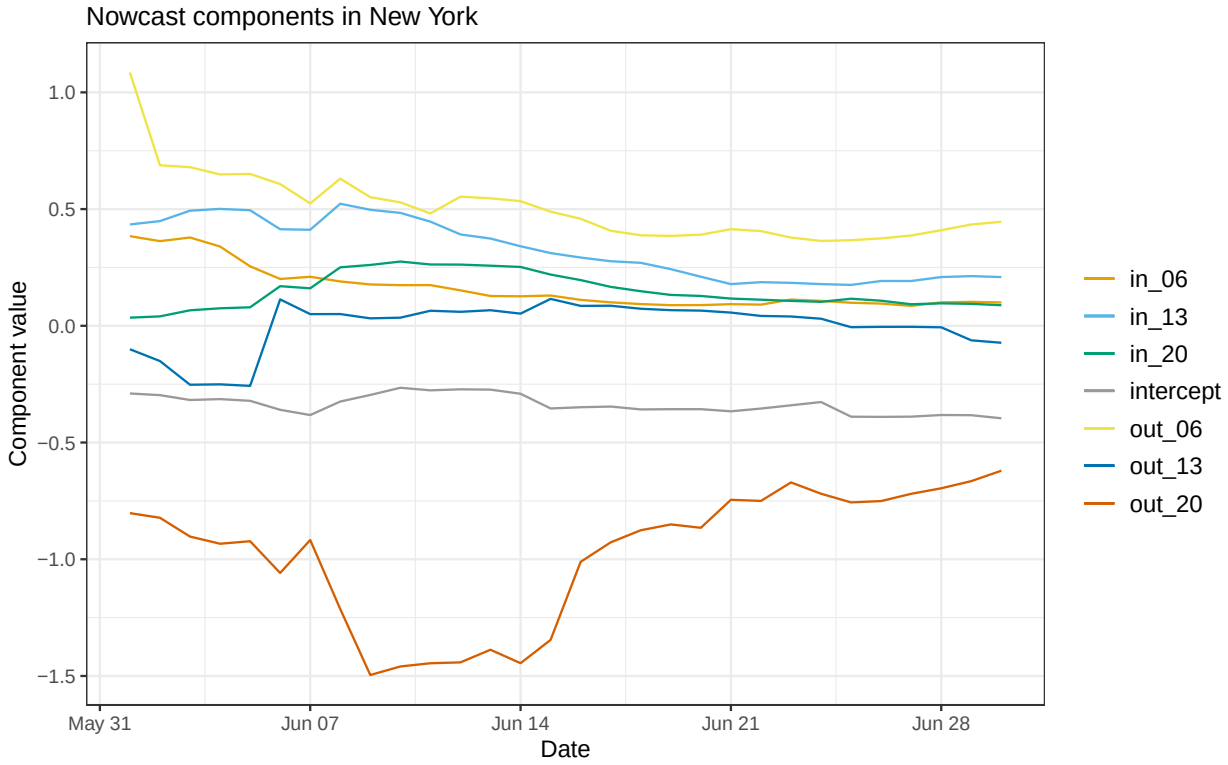

Figure 9: Contributions of each feature to the nowcasts during June 2021.

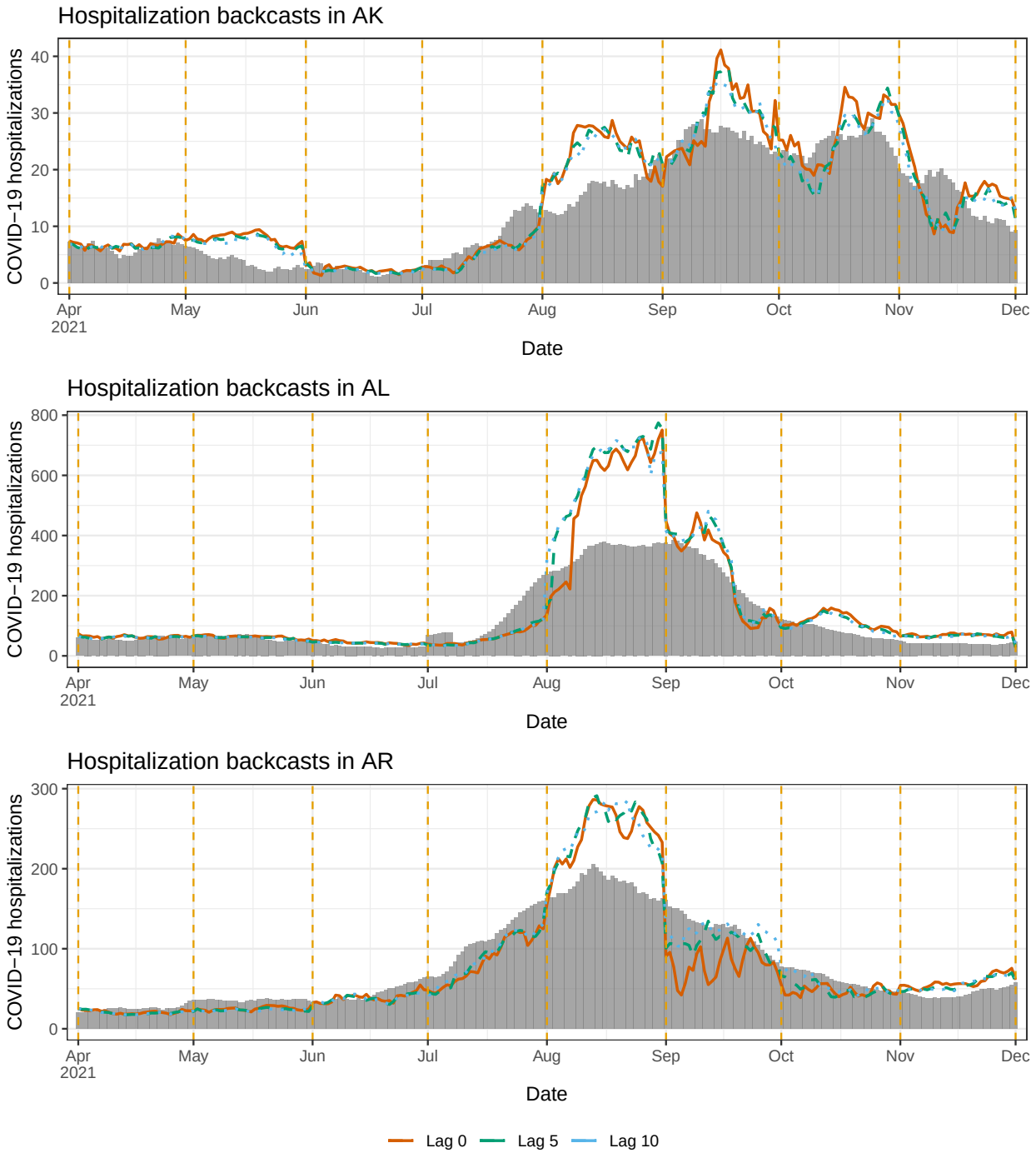

Figure 10: Backcasts from the mixed model in scenario 1, for AL, AK, AZ.

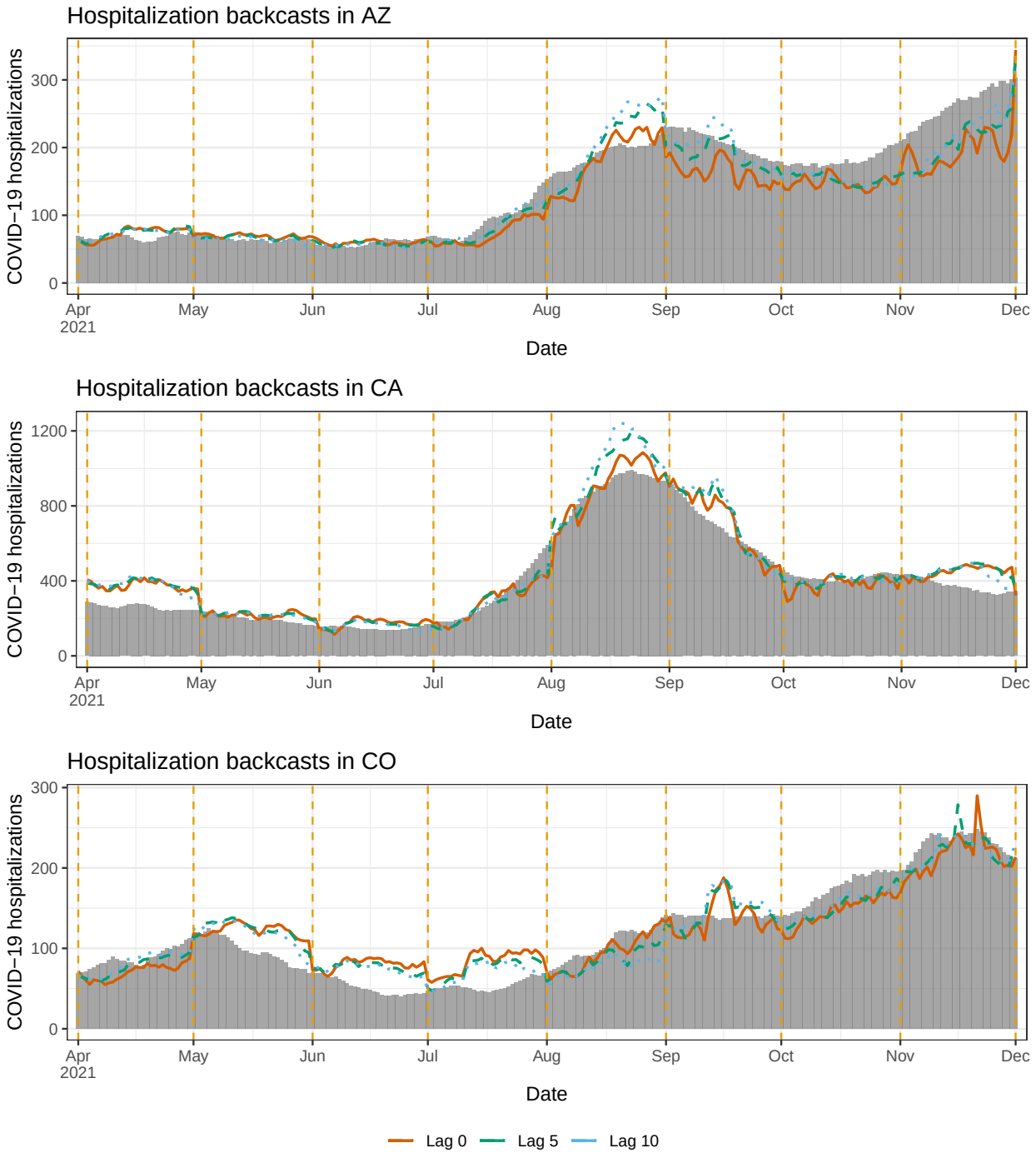

Figure 11: Backcasts from the mixed model in scenario 1, for AR, CA, CO.

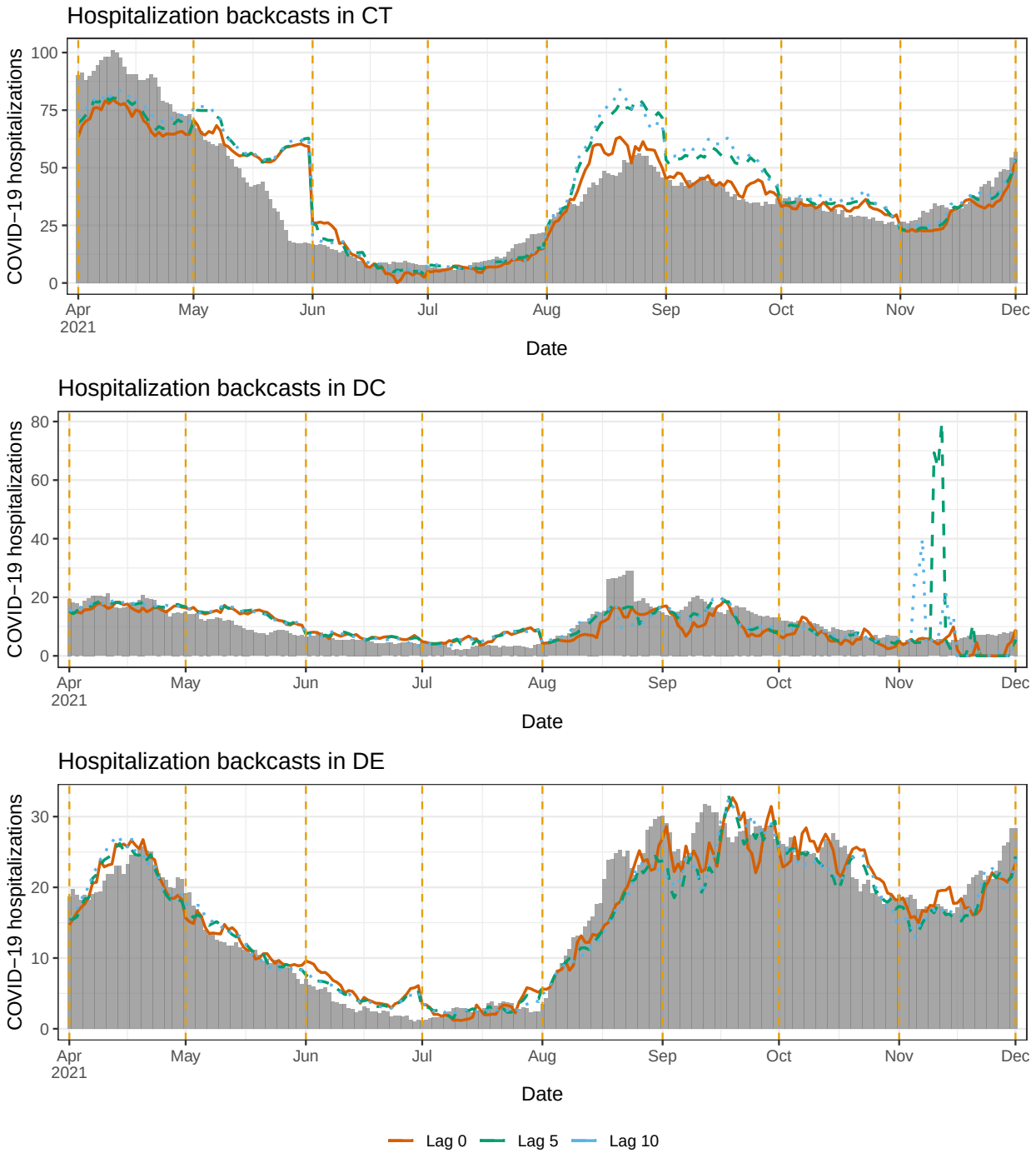

Figure 12: Backcasts from the mixed model in scenario 1, for CT, DC, DE.

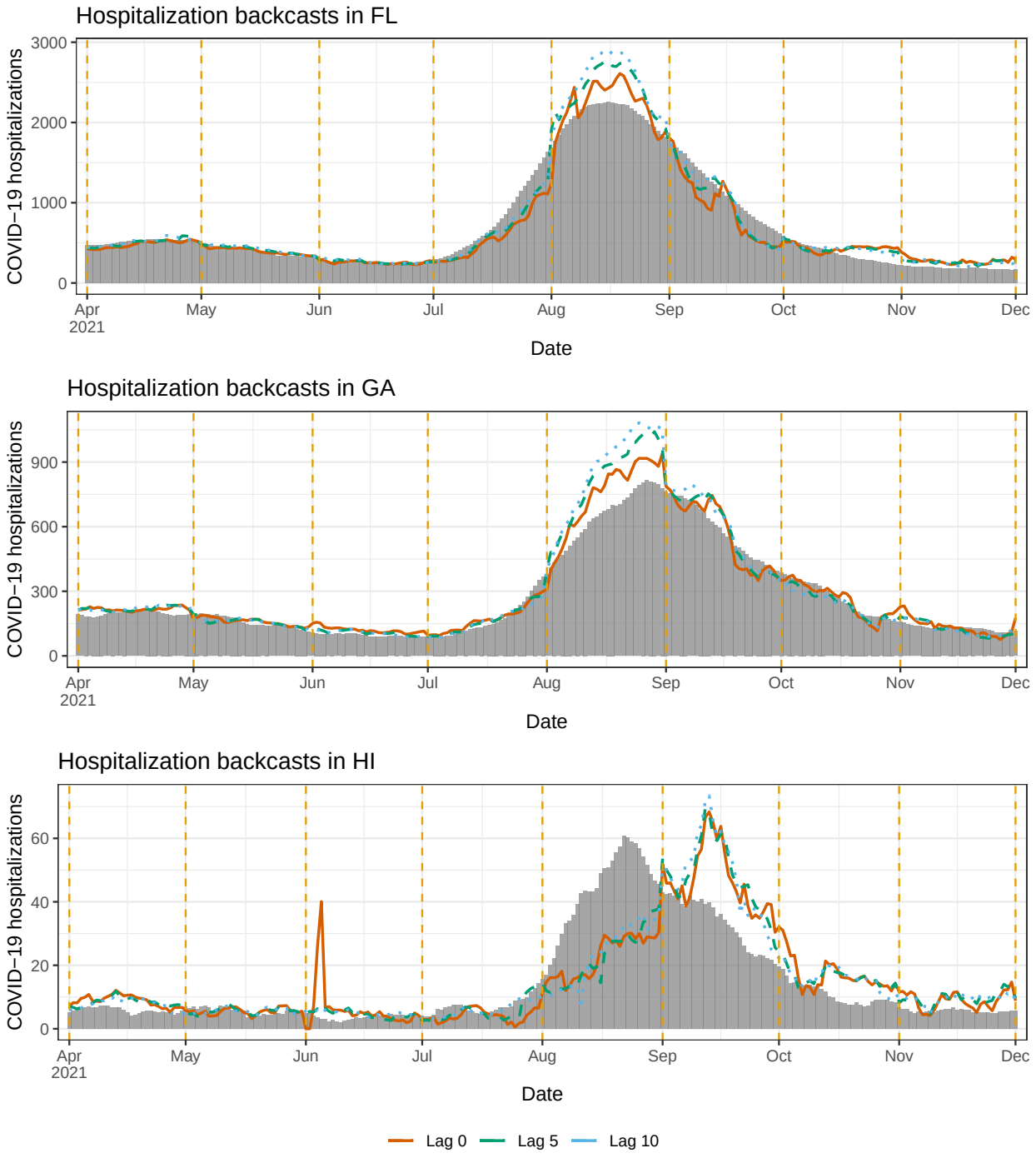

Figure 13: Backcasts from the mixed model in scenario 1, for FL, GA, HI.

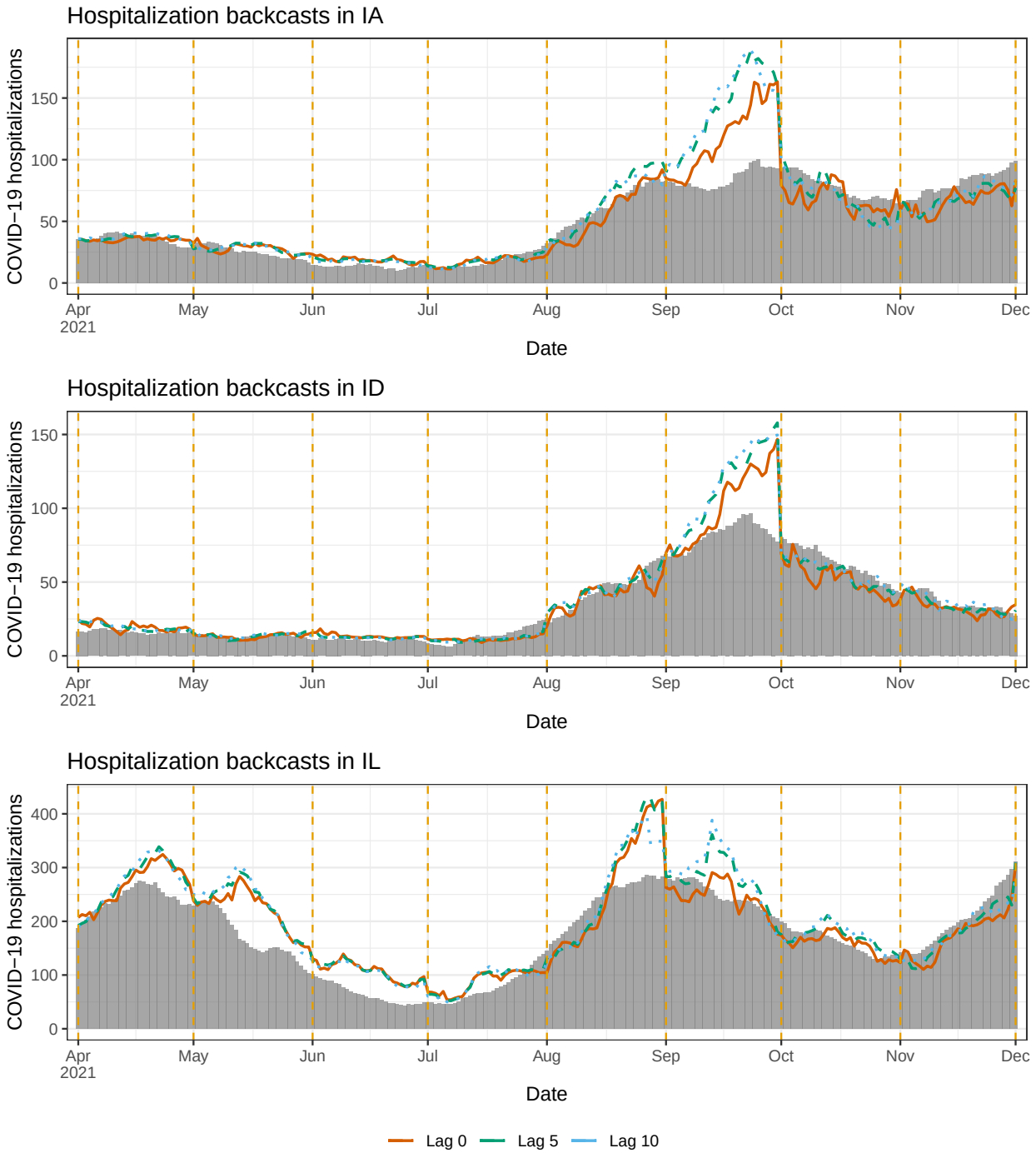

Figure 14: Backcasts from the mixed model in scenario 1, for IA, ID, IL.

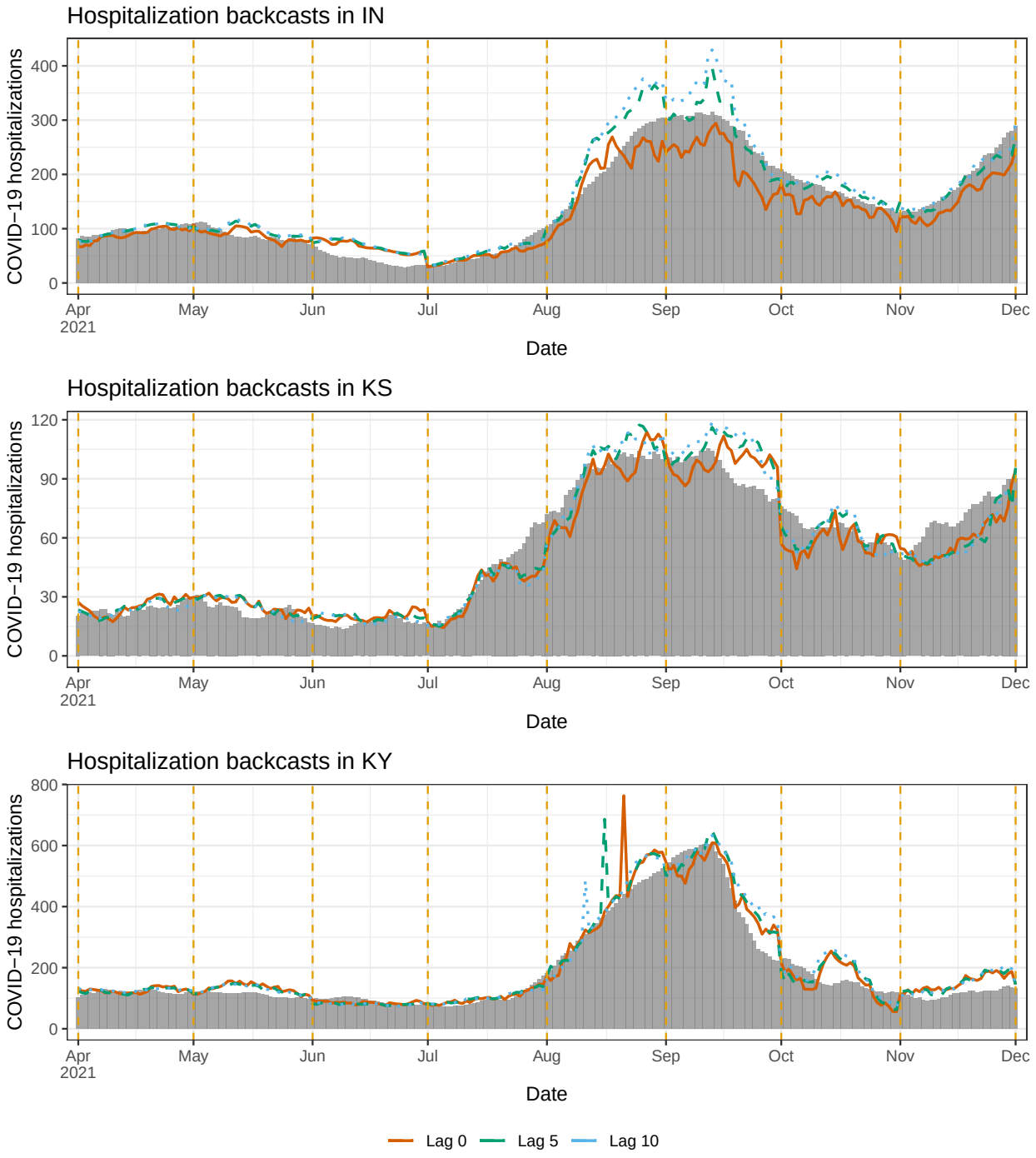

Figure 15: Backcasts from the mixed model in scenario 1, for IN, KS, KY.

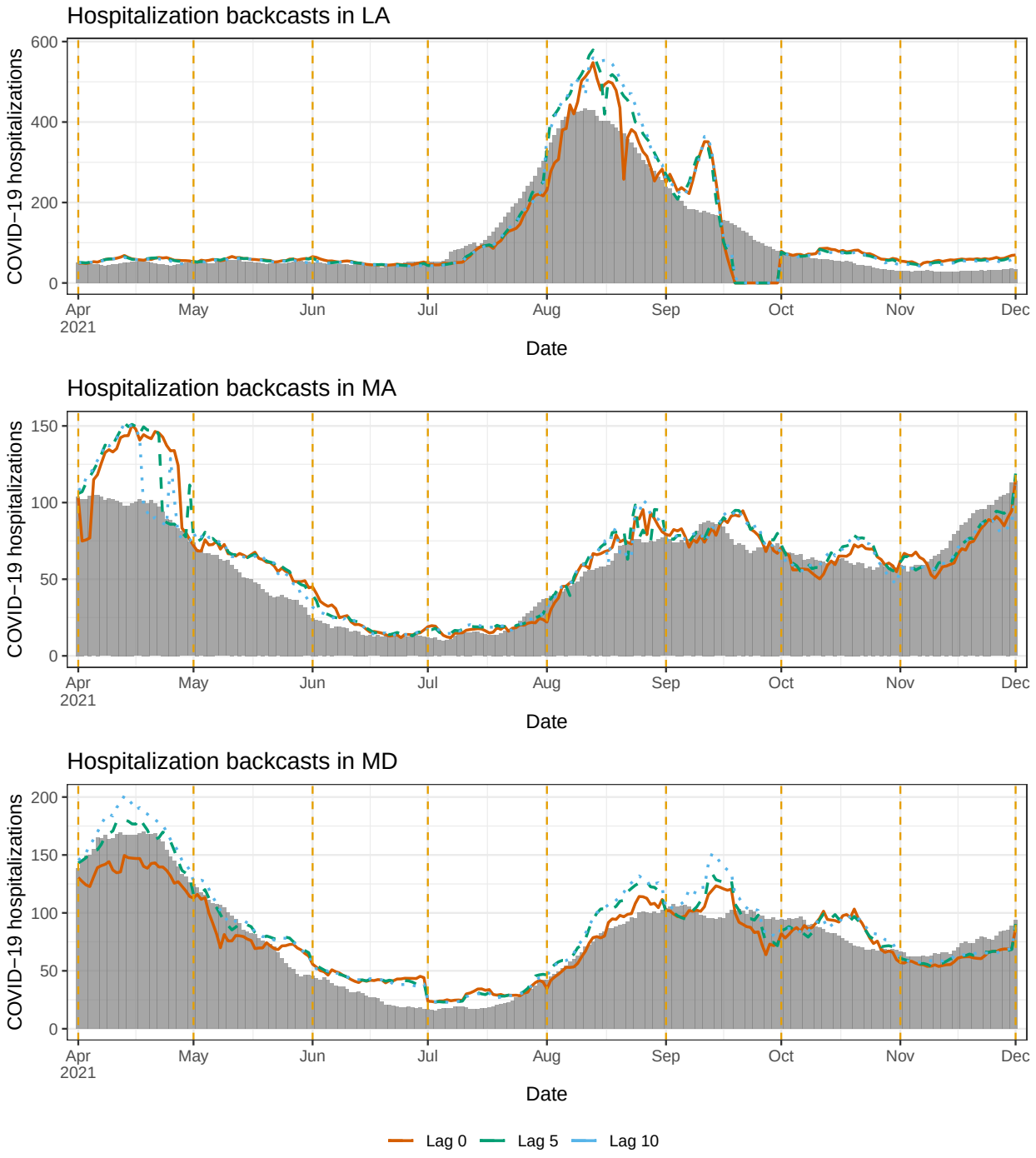

Figure 16: Backcasts from the mixed model in scenario 1, for LA, MA, MD.

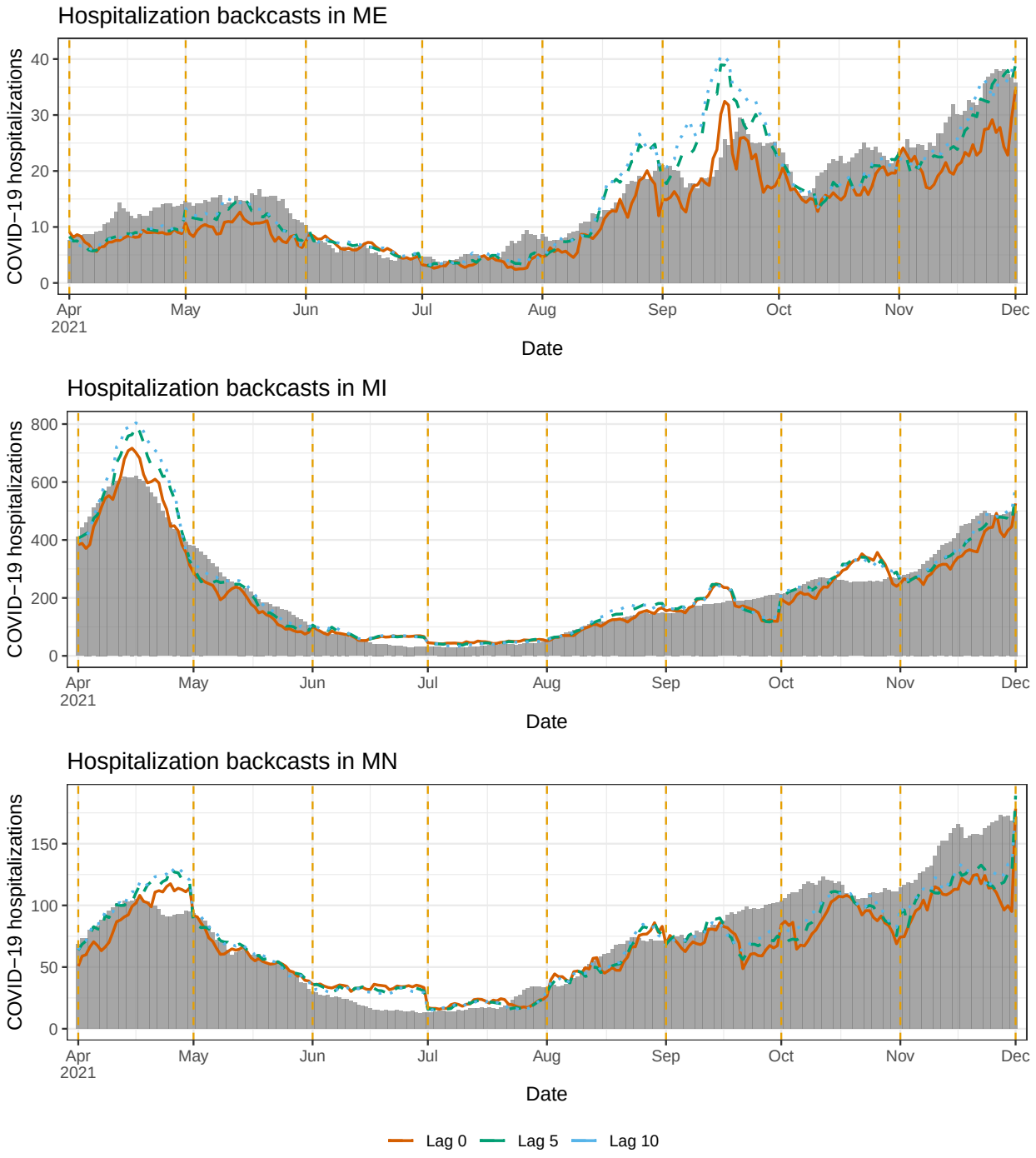

Figure 17: Backcasts from the mixed model in scenario 1, for ME, MI, MN.

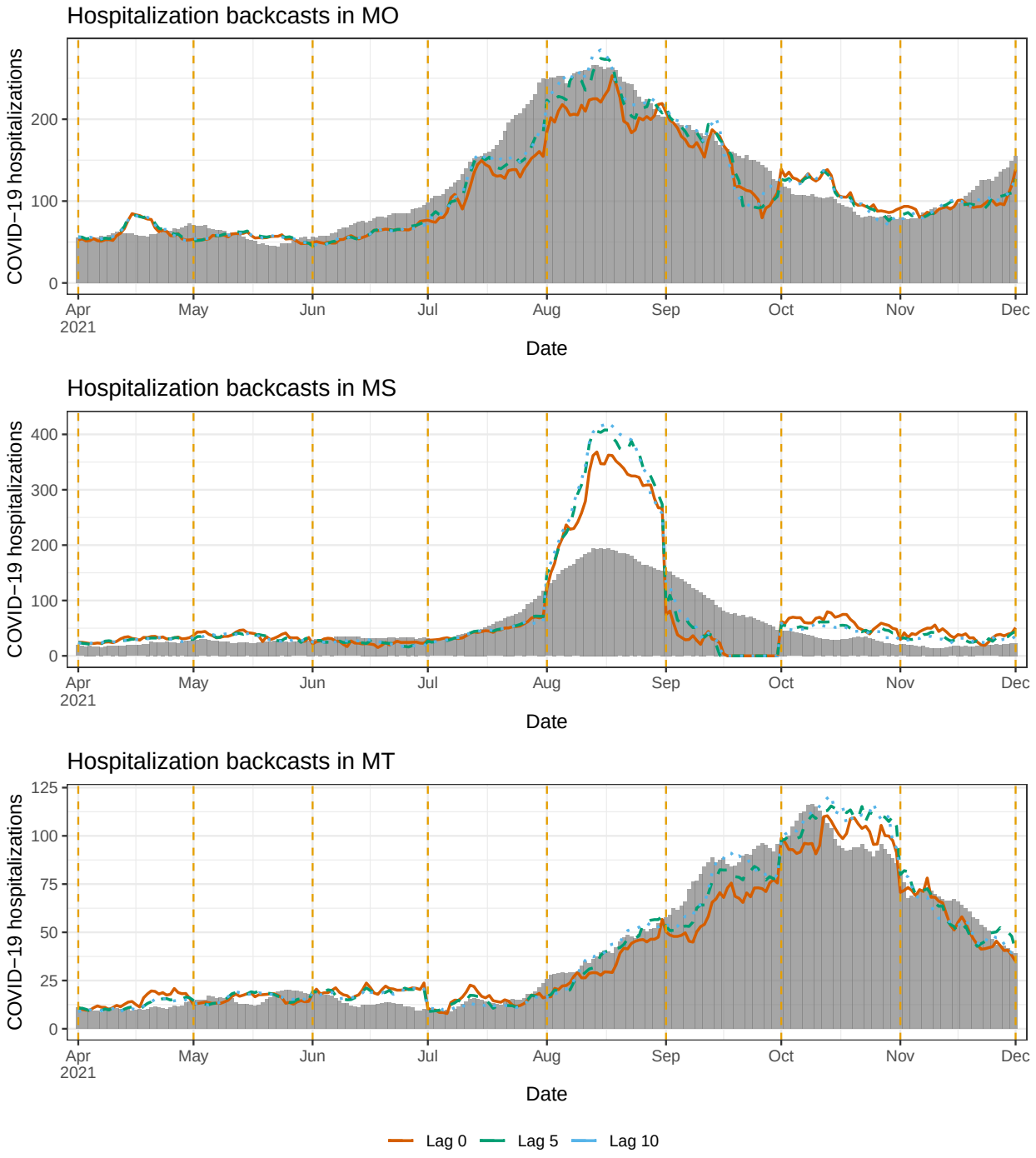

Figure 18: Backcasts from the mixed model in scenario 1, for MO, MS, MT.

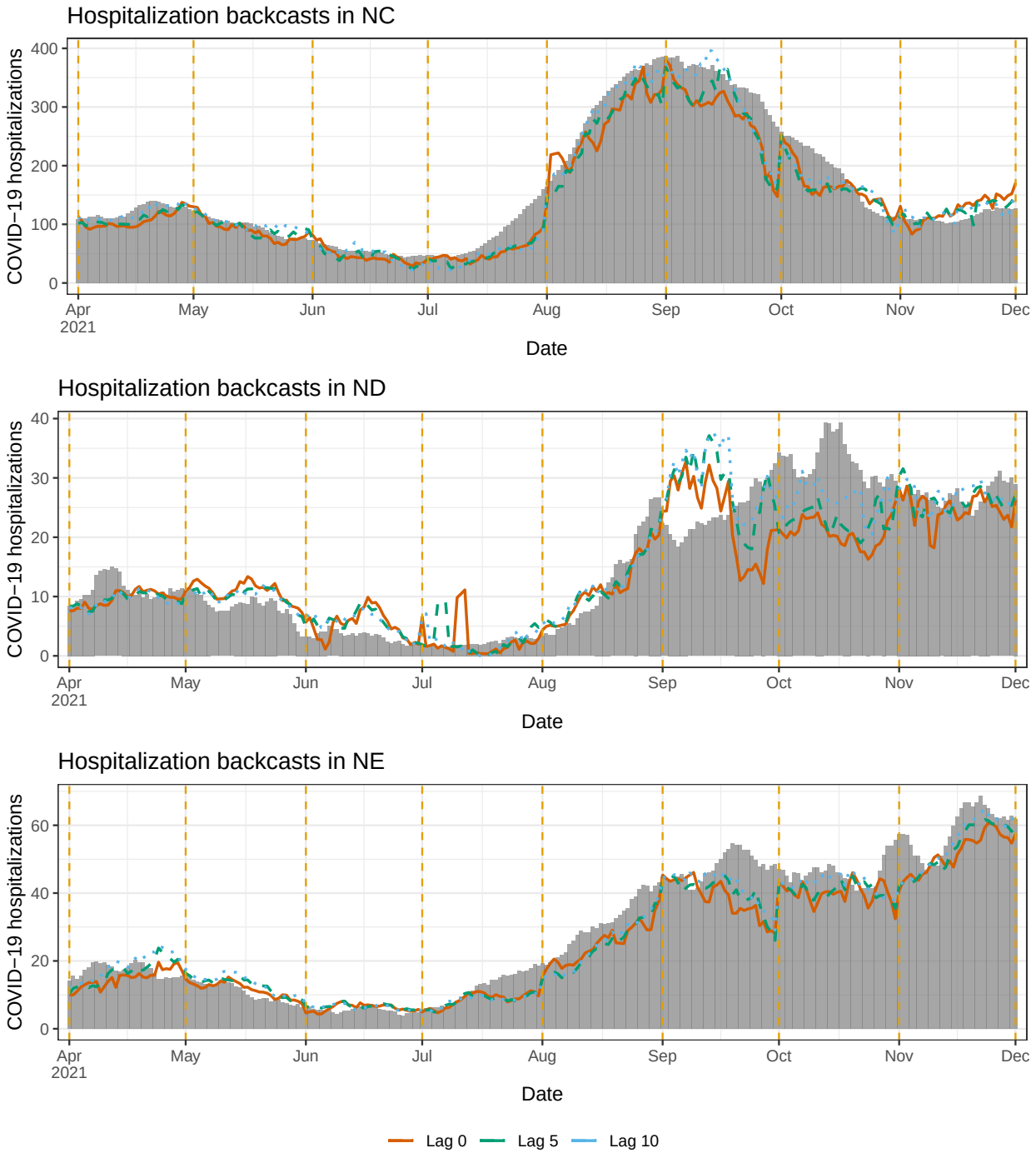

Figure 19: Backcasts from the mixed model in scenario 1, for NC, ND, NE.

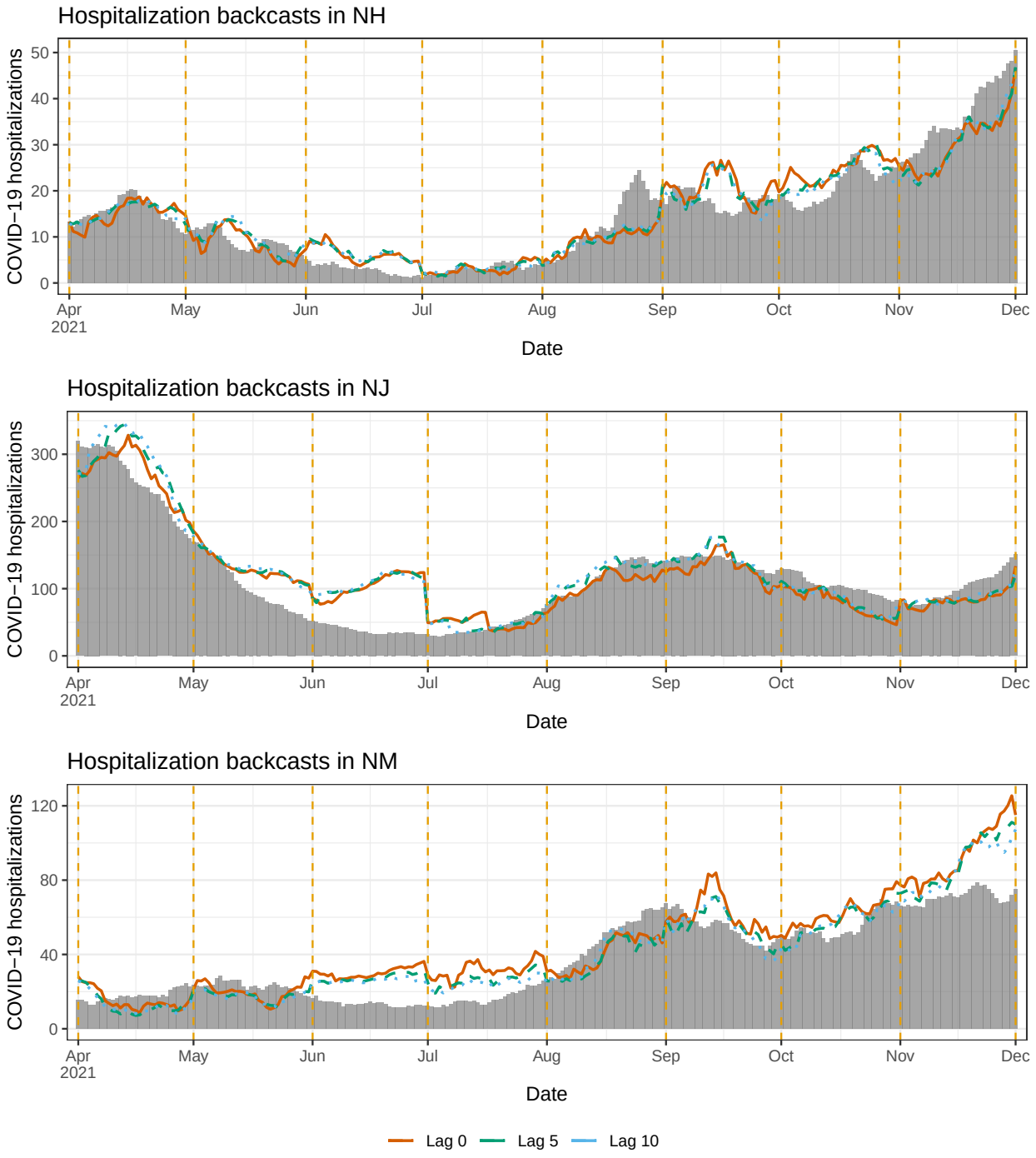

Figure 20: Backcasts from the mixed model in scenario 1, for NH, NJ, NM.

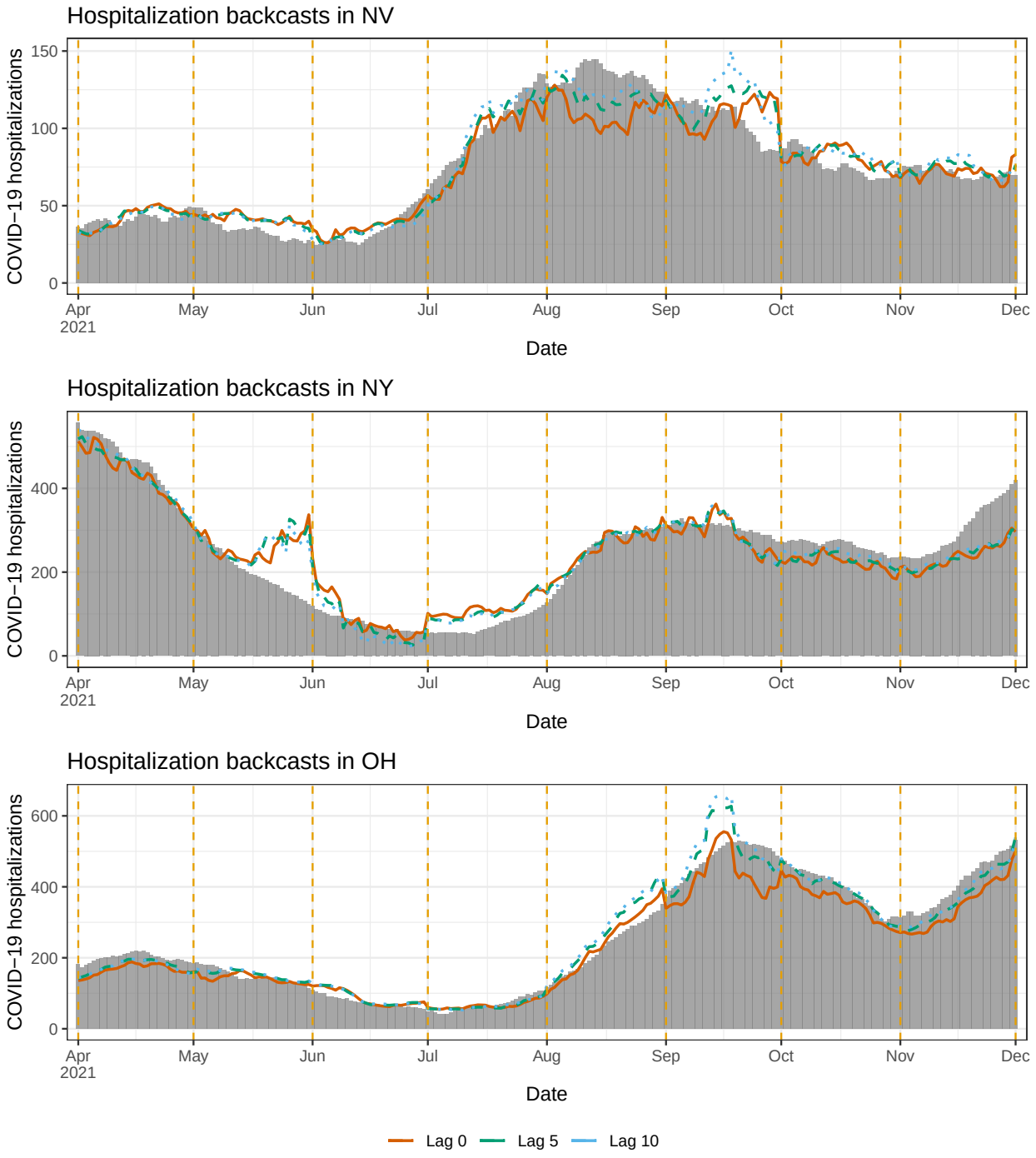

Figure 21: Backcasts from the mixed model in scenario 1, for NV, NY, OH.

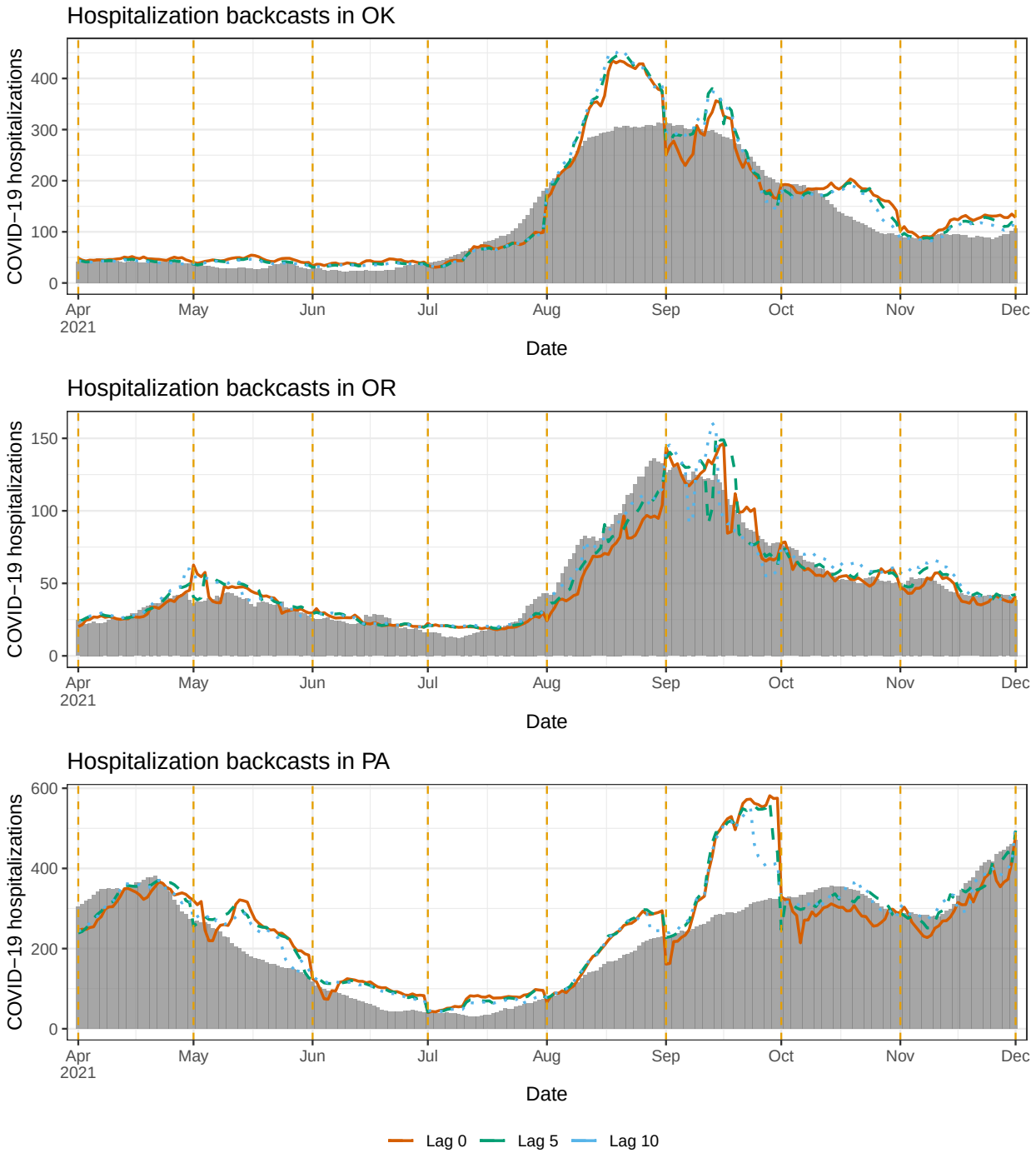

Figure 22: Backcasts from the mixed model in scenario 1, for OK, OR, PA.

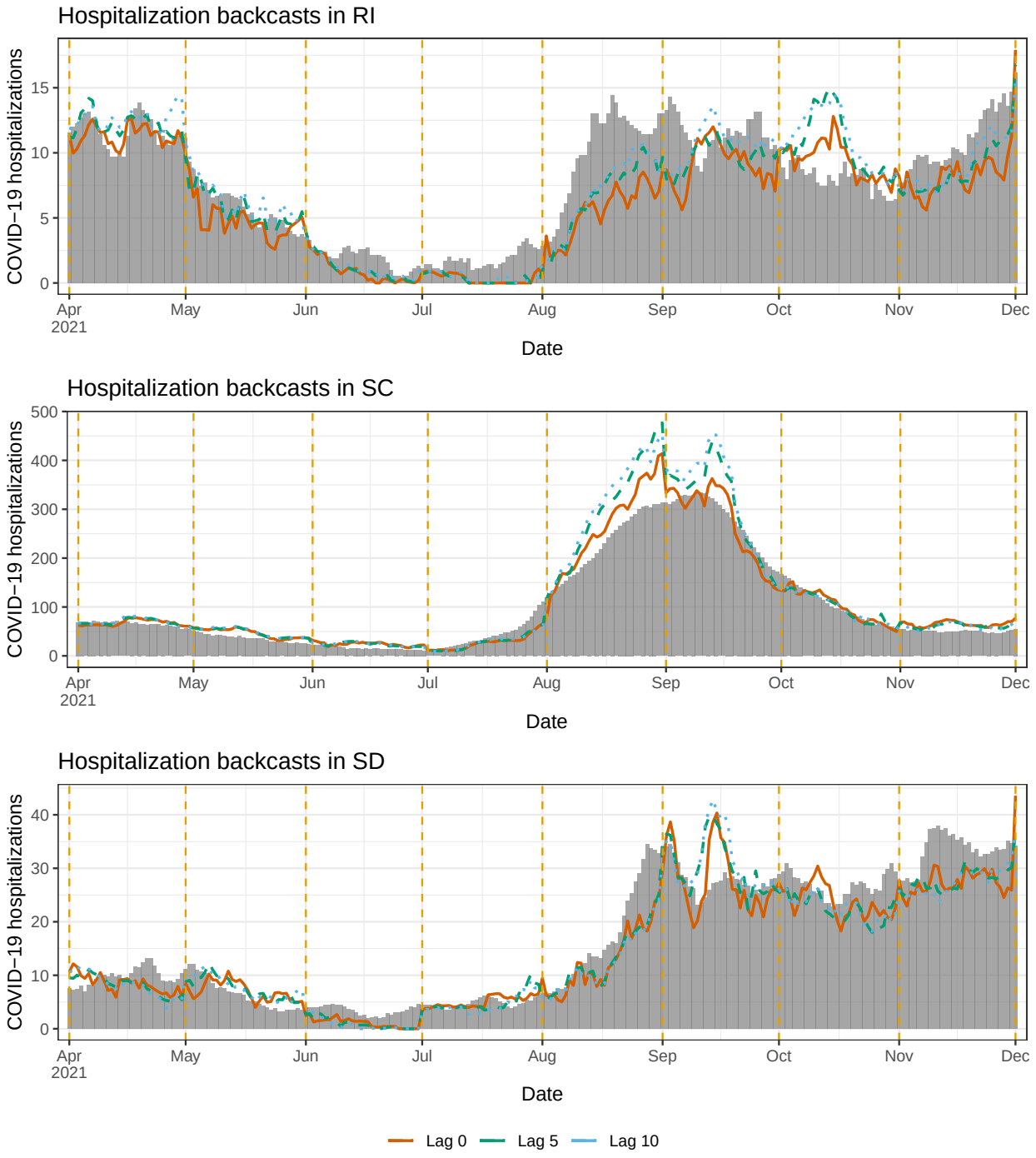

Figure 23: Backcasts from the mixed model in scenario 1, for RI, SC, SD.

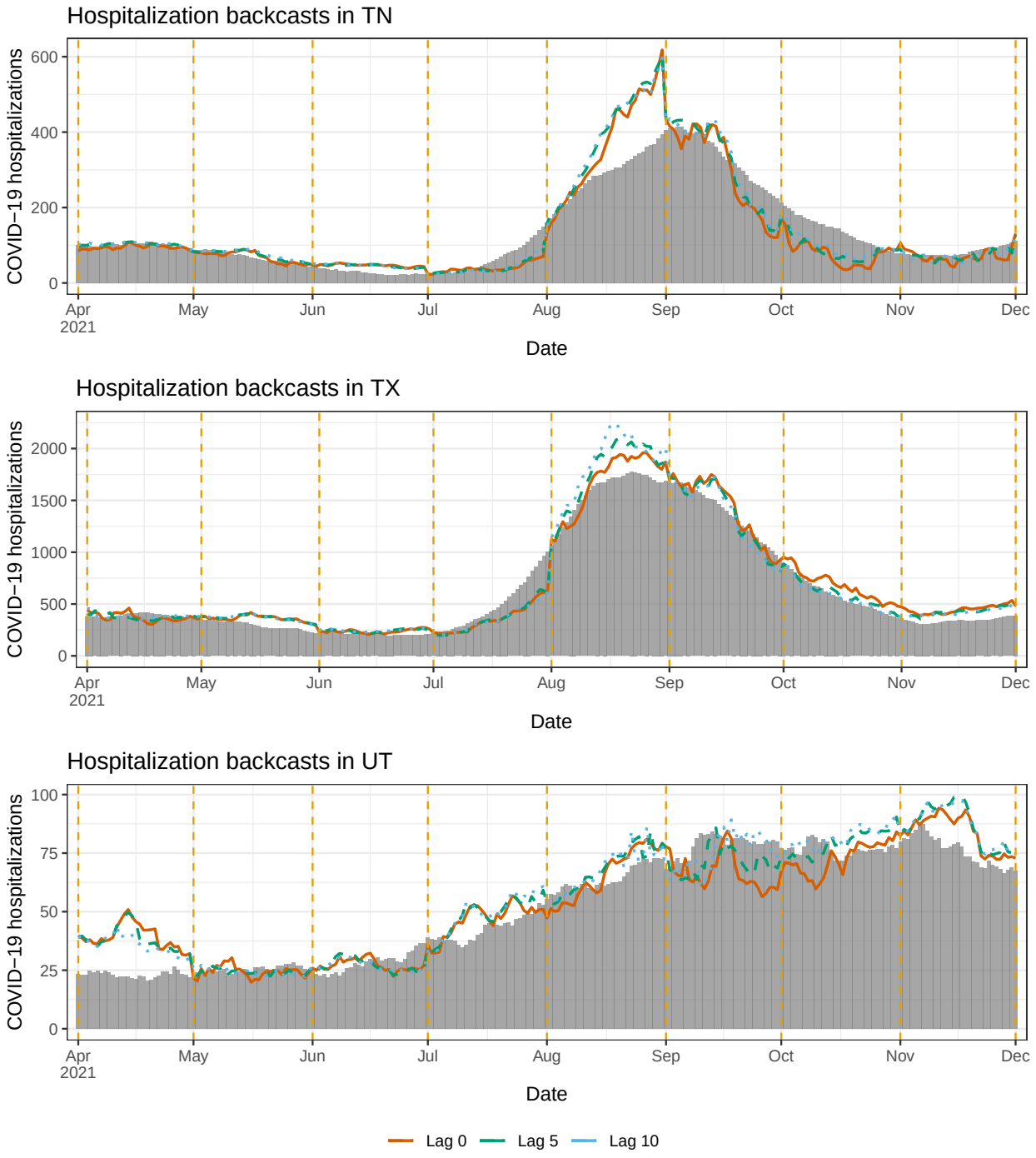

Figure 24: Backcasts from the mixed model in scenario 1, for TN, TX, UT.

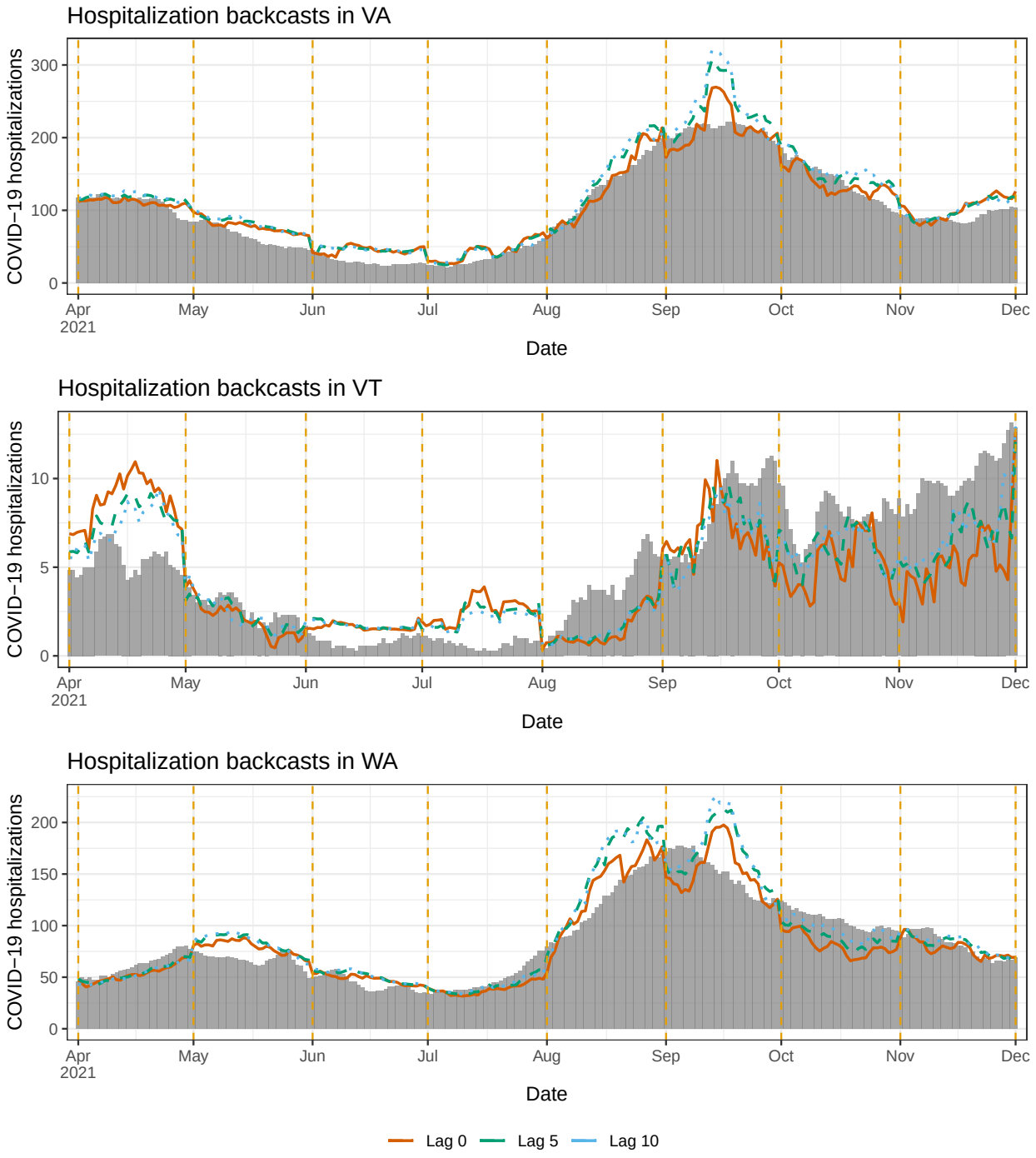

Figure 25: Backcasts from the mixed model in scenario 1, for VA, VT, WA.

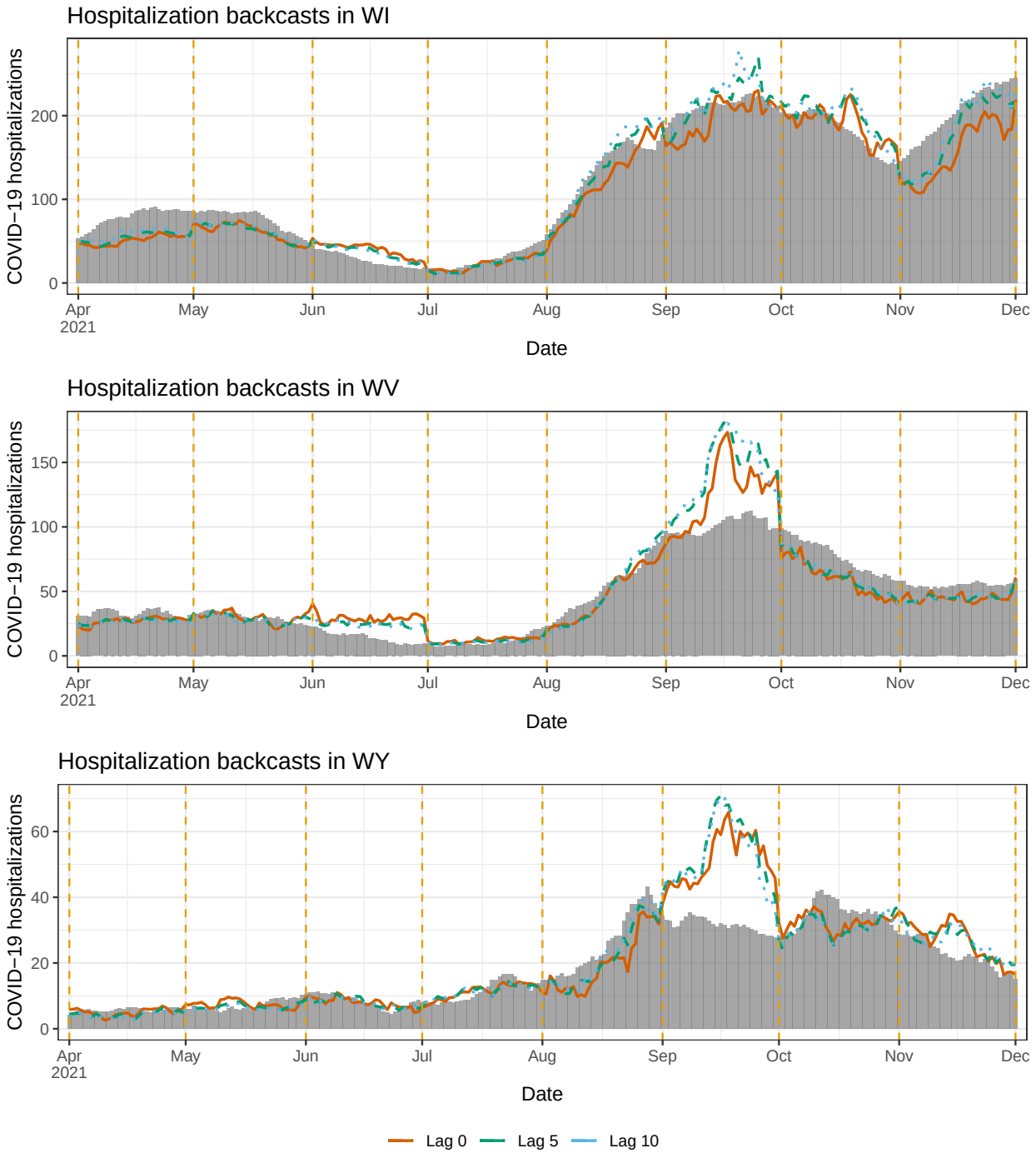

Figure 26: Backcasts from the mixed model in scenario 1, for WI, WV, WY.

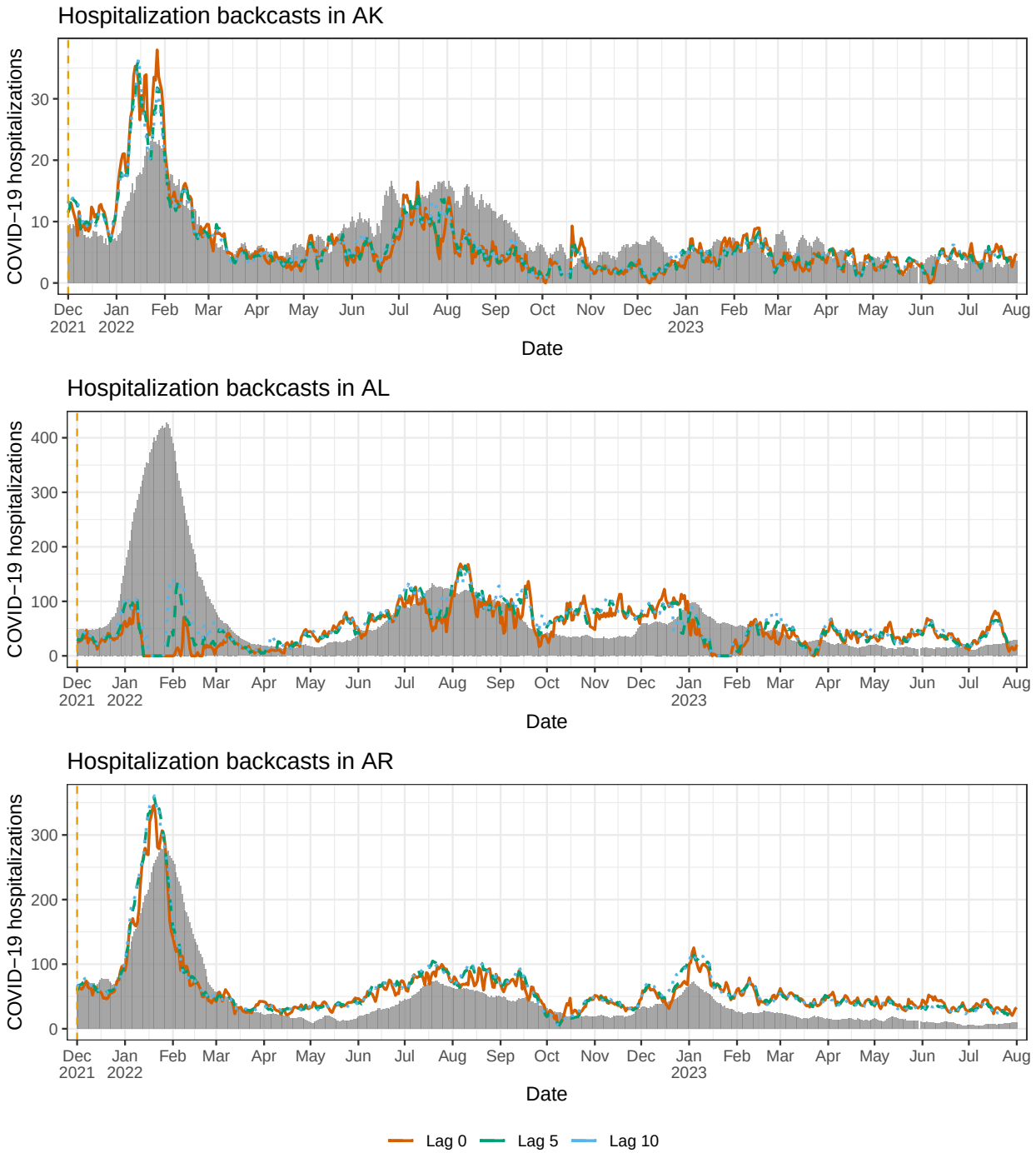

Figure 27: Backcasts from the mixed model in scenario 2, for AL, AK, AZ.

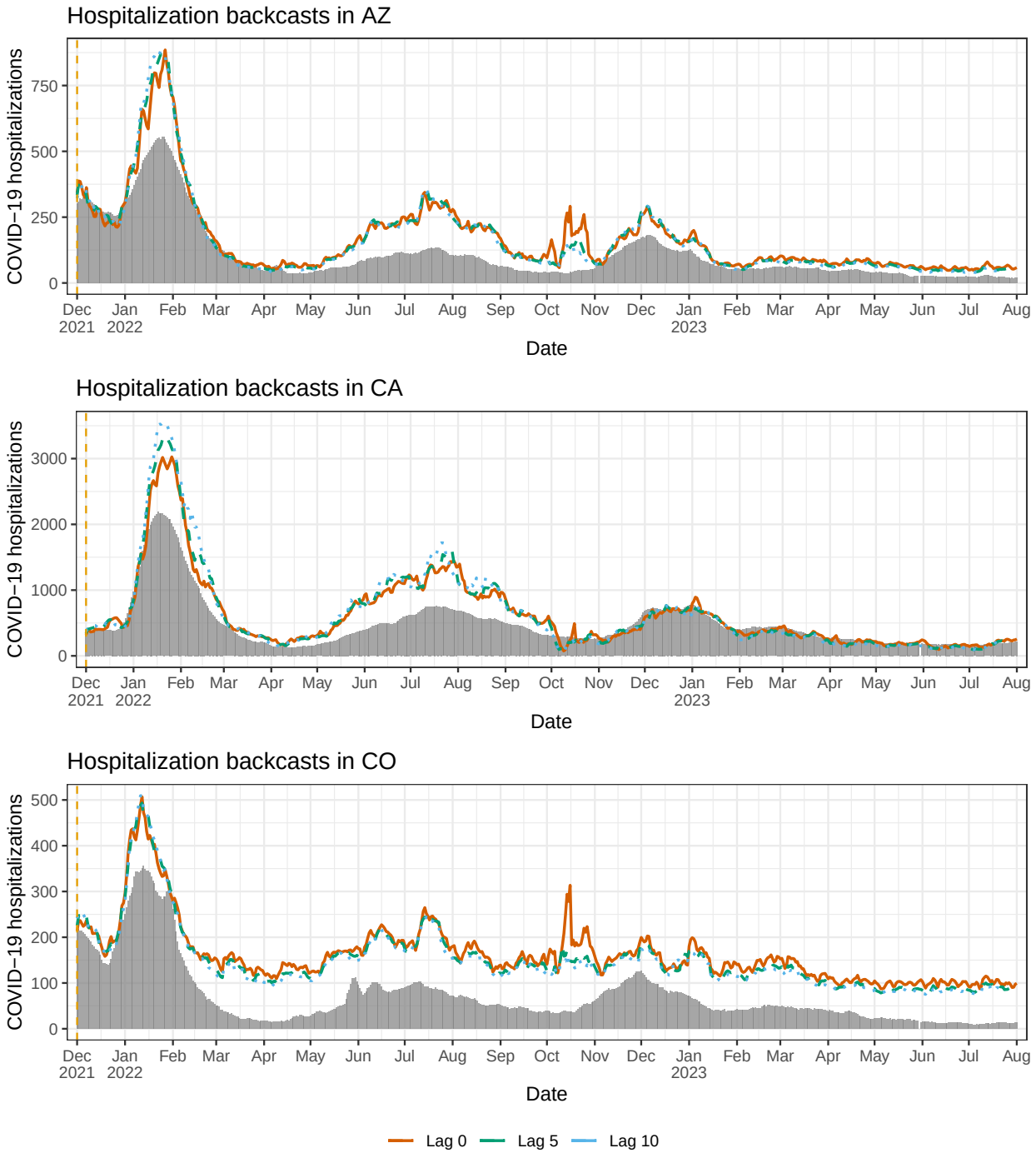

Figure 28: Backcasts from the mixed model in scenario 2, for AR, CA, CO.

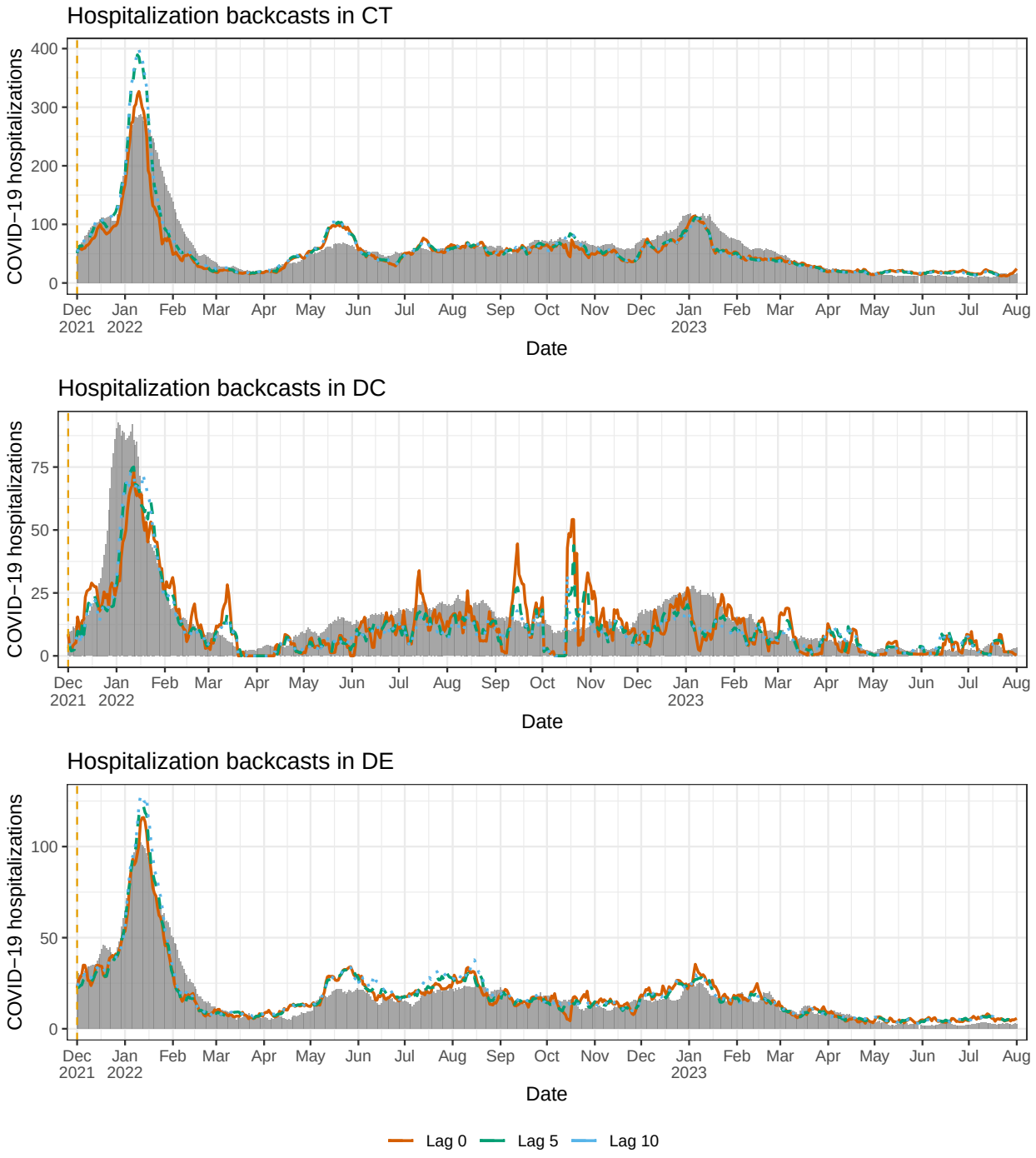

Figure 29: Backcasts from the mixed model in scenario 2, for CT, DC, DE.

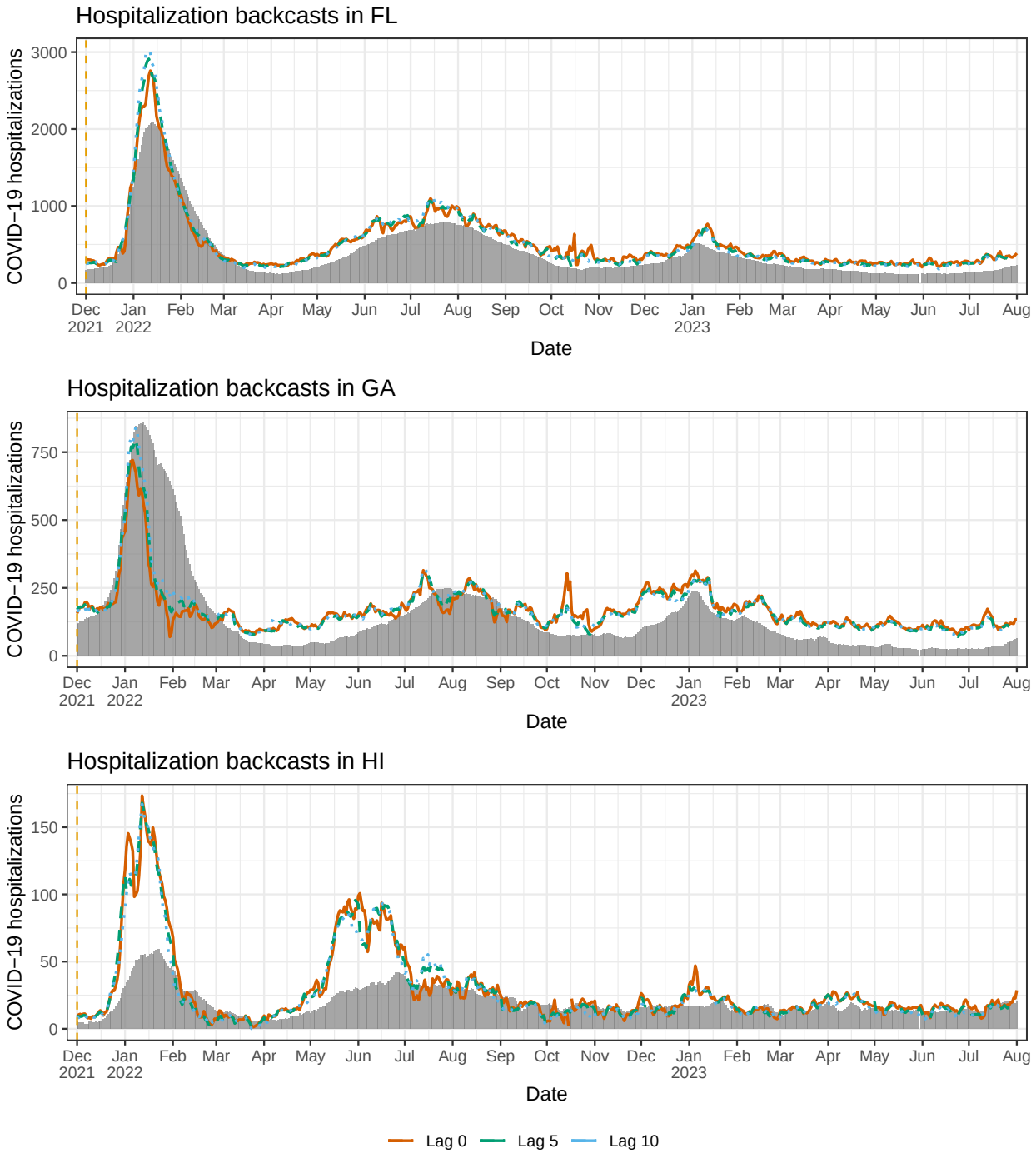

Figure 30: Backcasts from the mixed model in scenario 2, for FL, GA, HI.

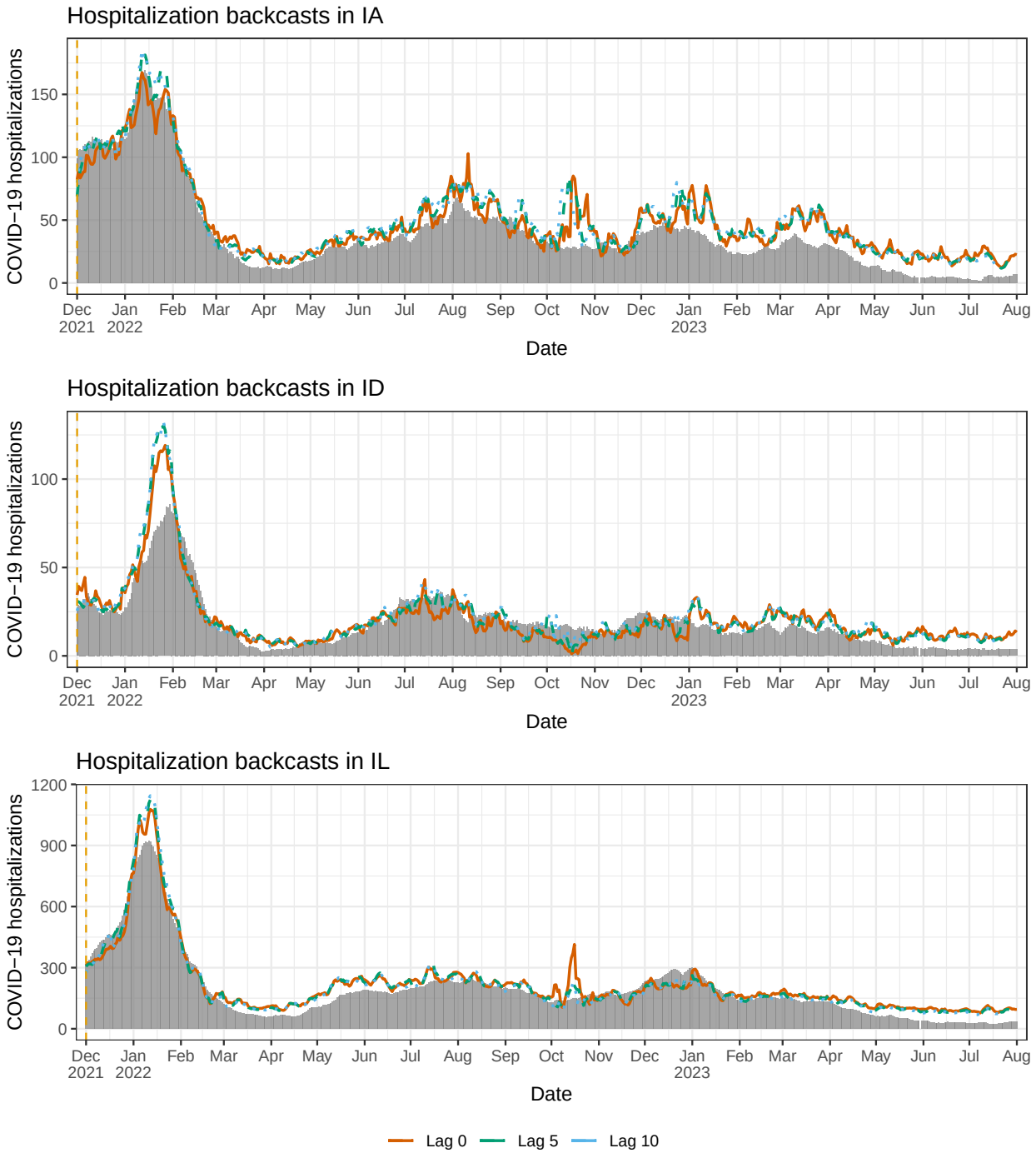

Figure 31: Backcasts from the mixed model in scenario 2, for IA, ID, IL.

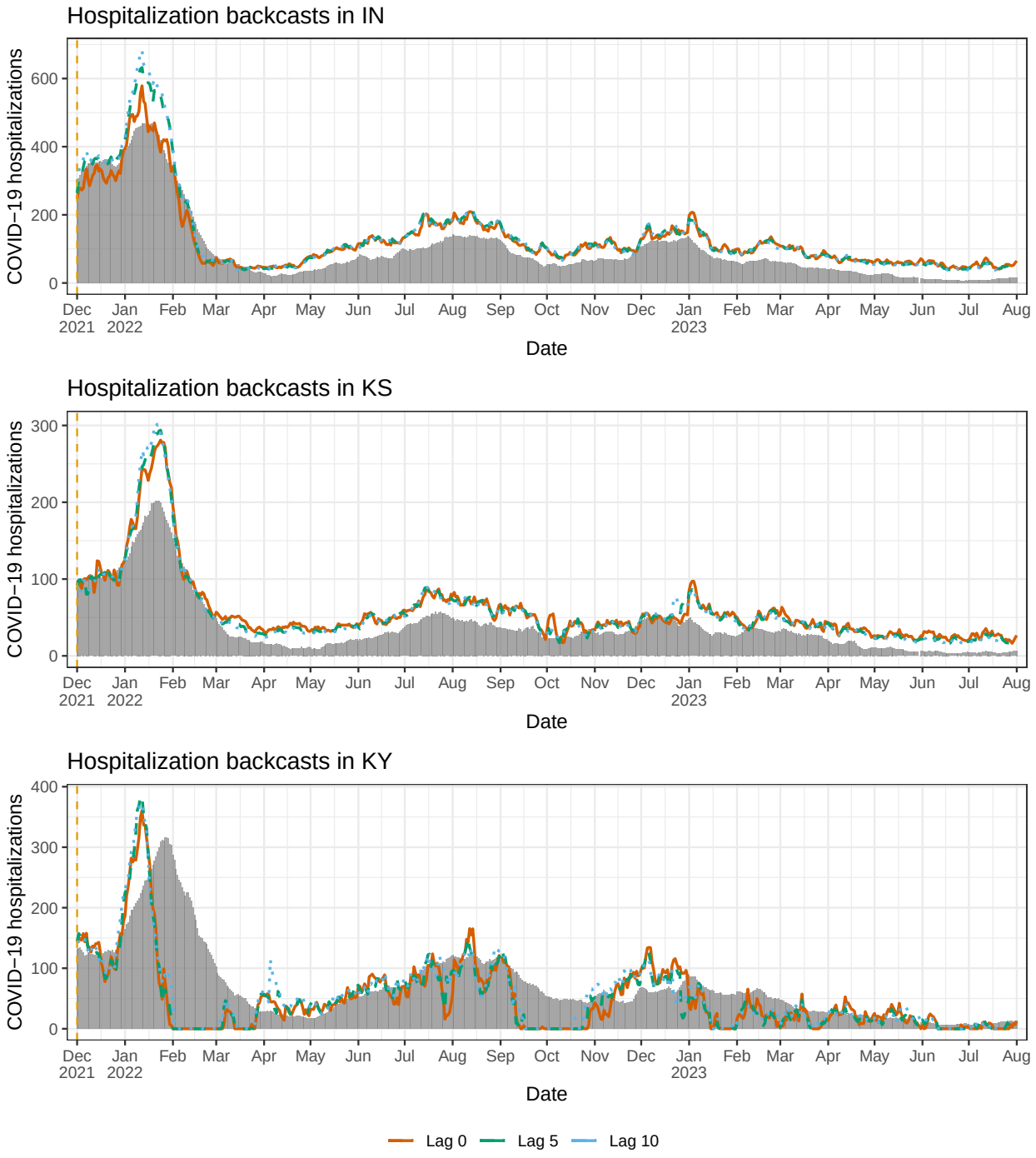

Figure 32: Backcasts from the mixed model in scenario 2, for IN, KS, KY.

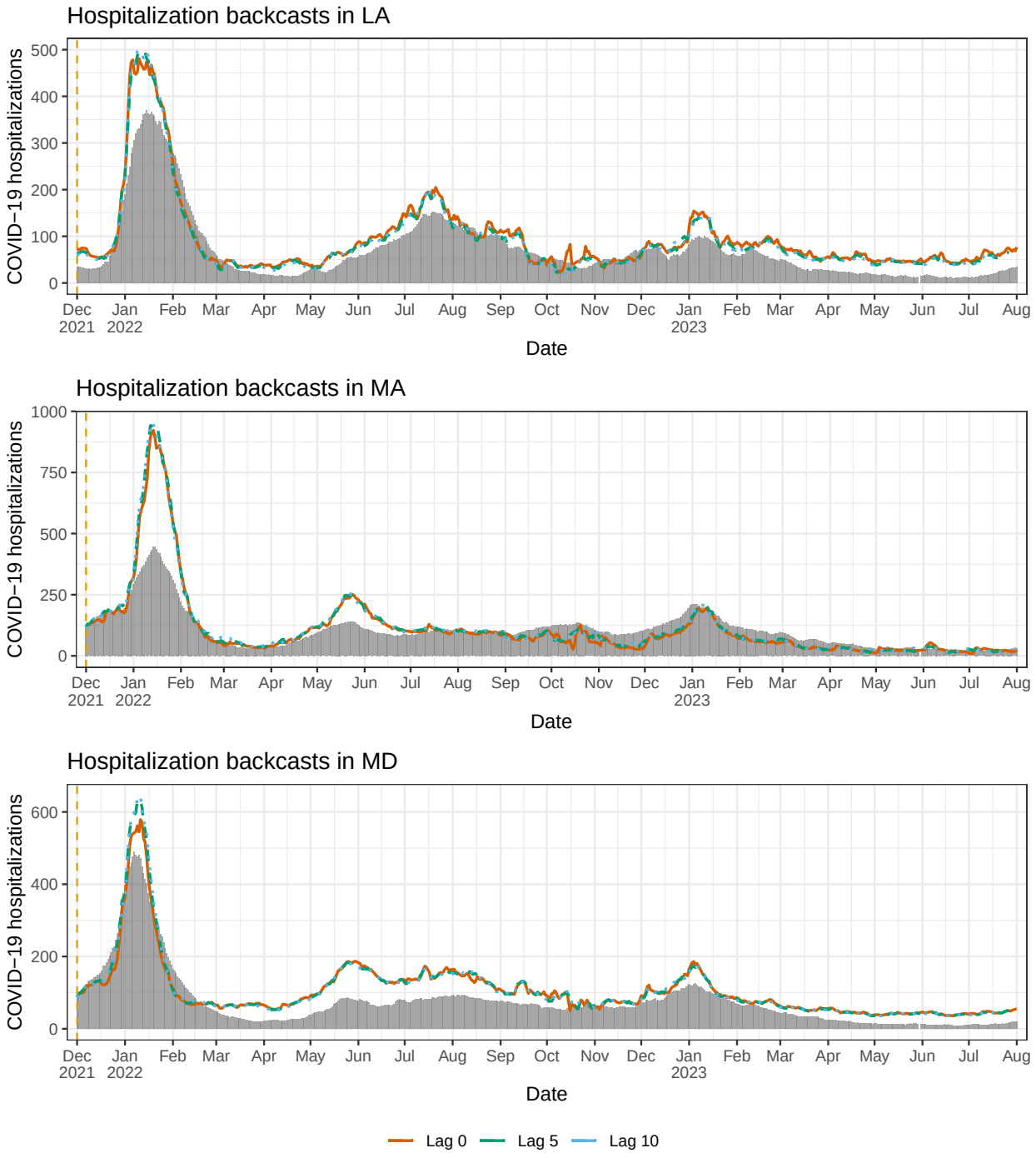

Figure 33: Backcasts from the mixed model in scenario 2, for LA, MA, MD.

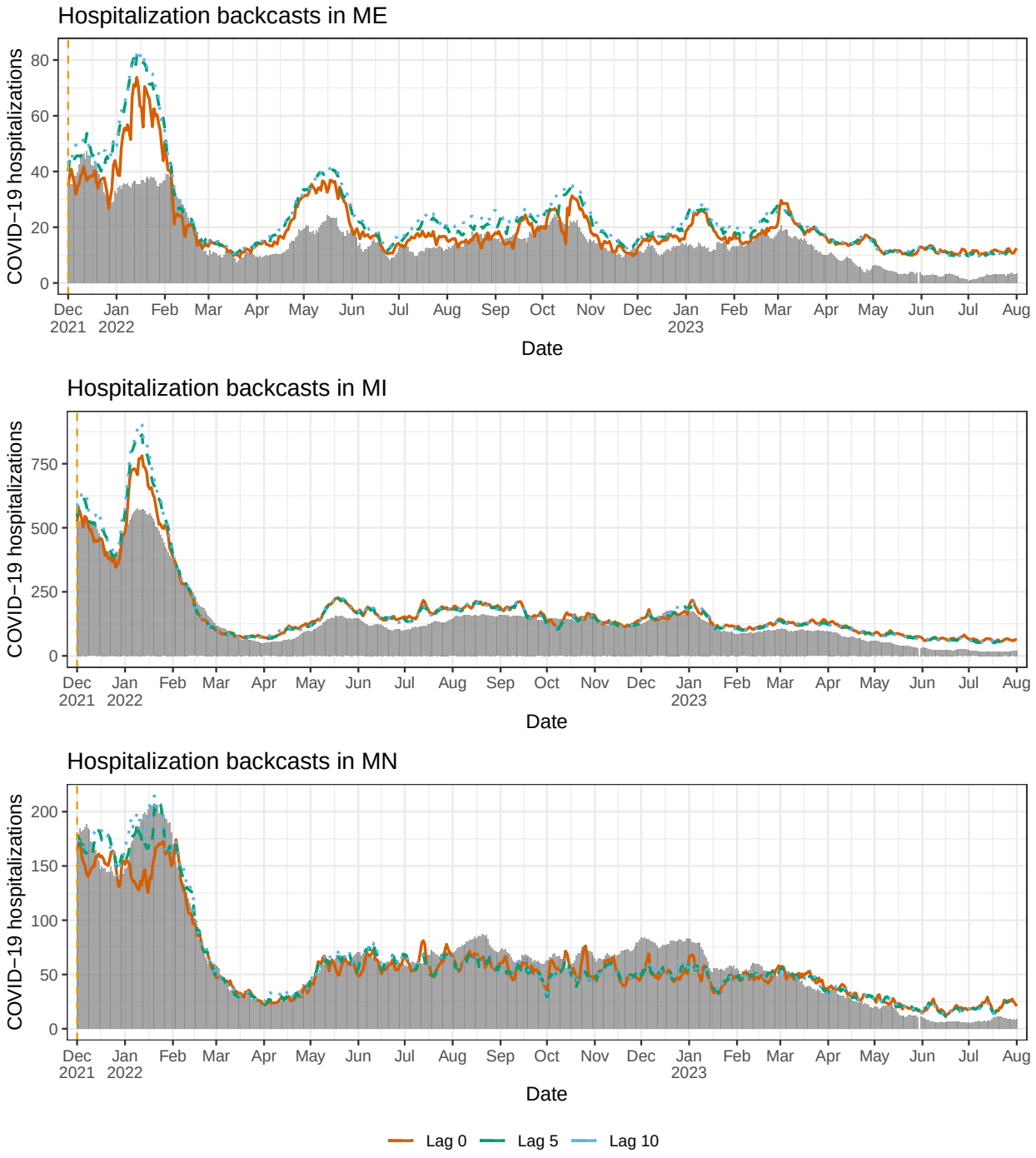

Figure 34: Backcasts from the mixed model in scenario 2, for ME, MI, MN.

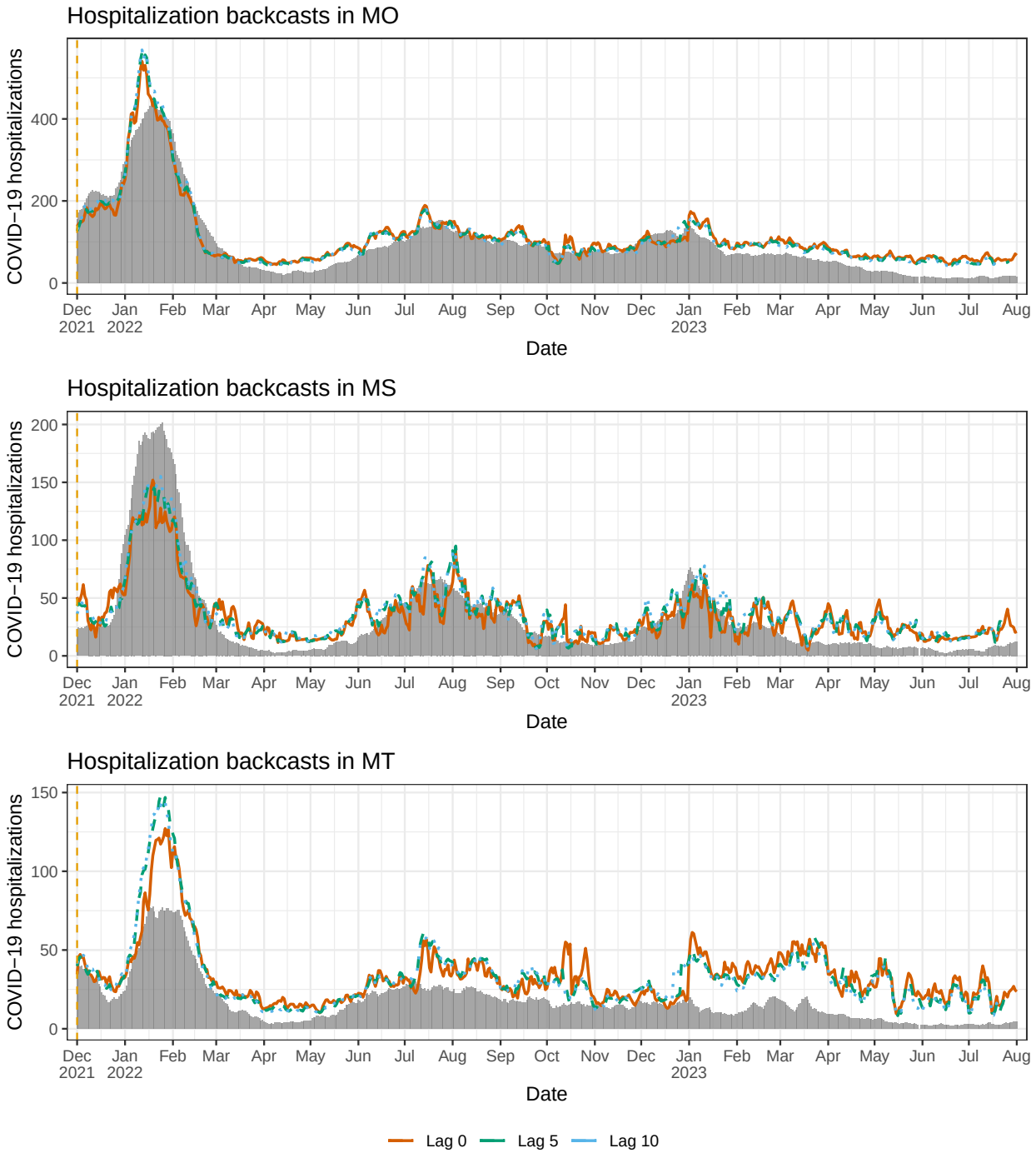

Figure 35: Backcasts from the mixed model in scenario 2, for MO, MS, MT.

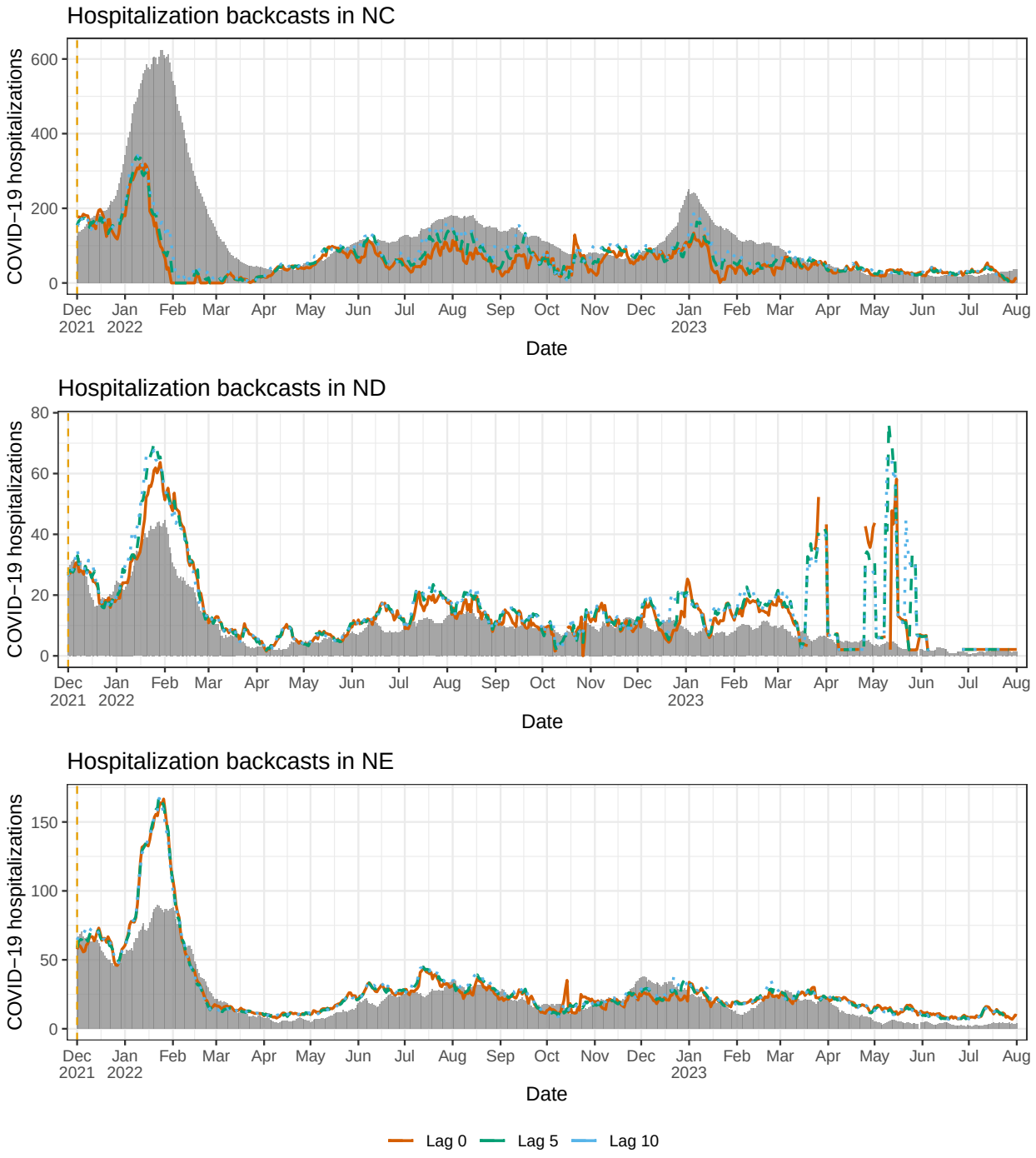

Figure 36: Backcasts from the mixed model in scenario 2, for NC, ND, NE.

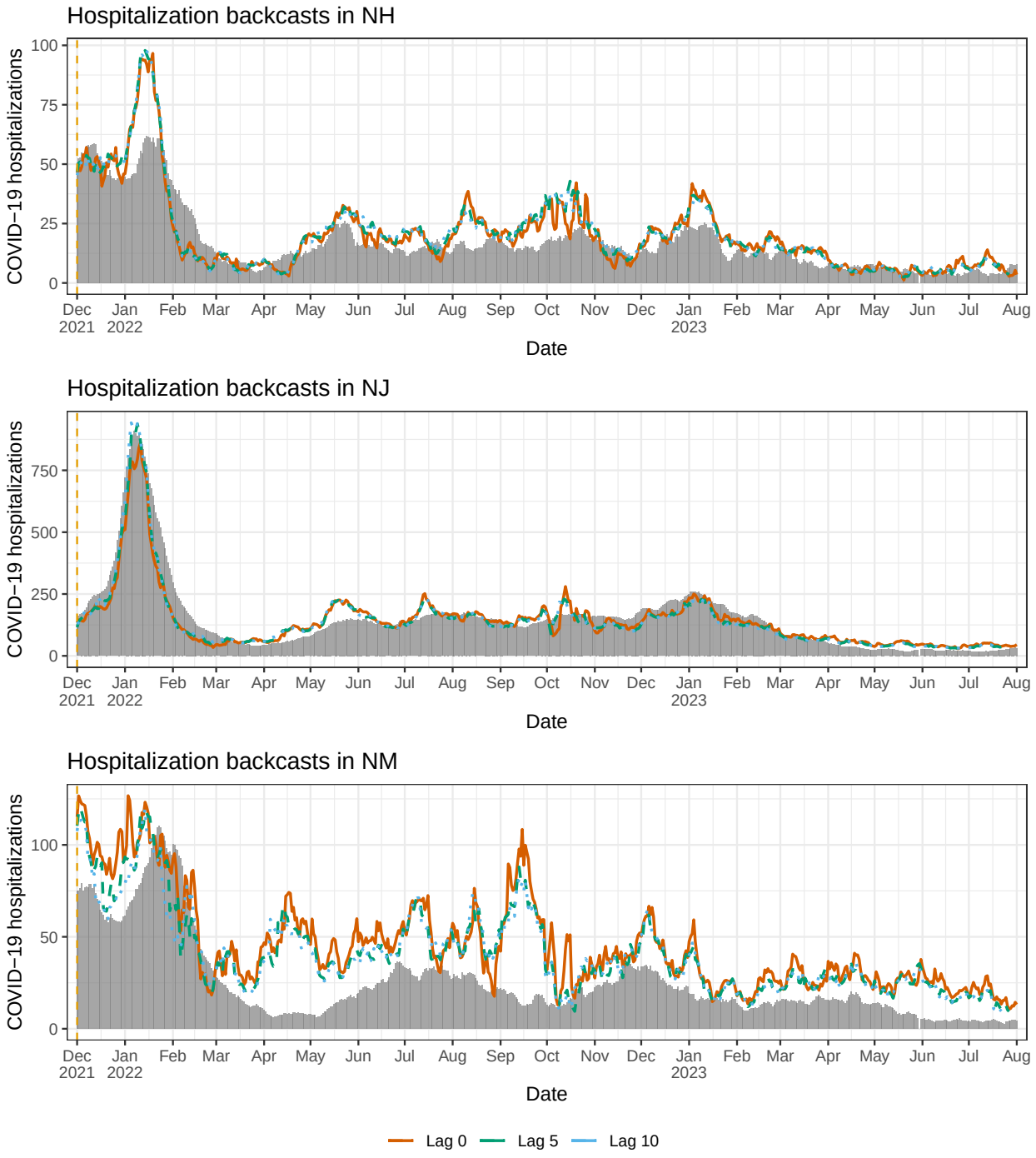

Figure 37: Backcasts from the mixed model in scenario 2, for NH, NJ, NM.

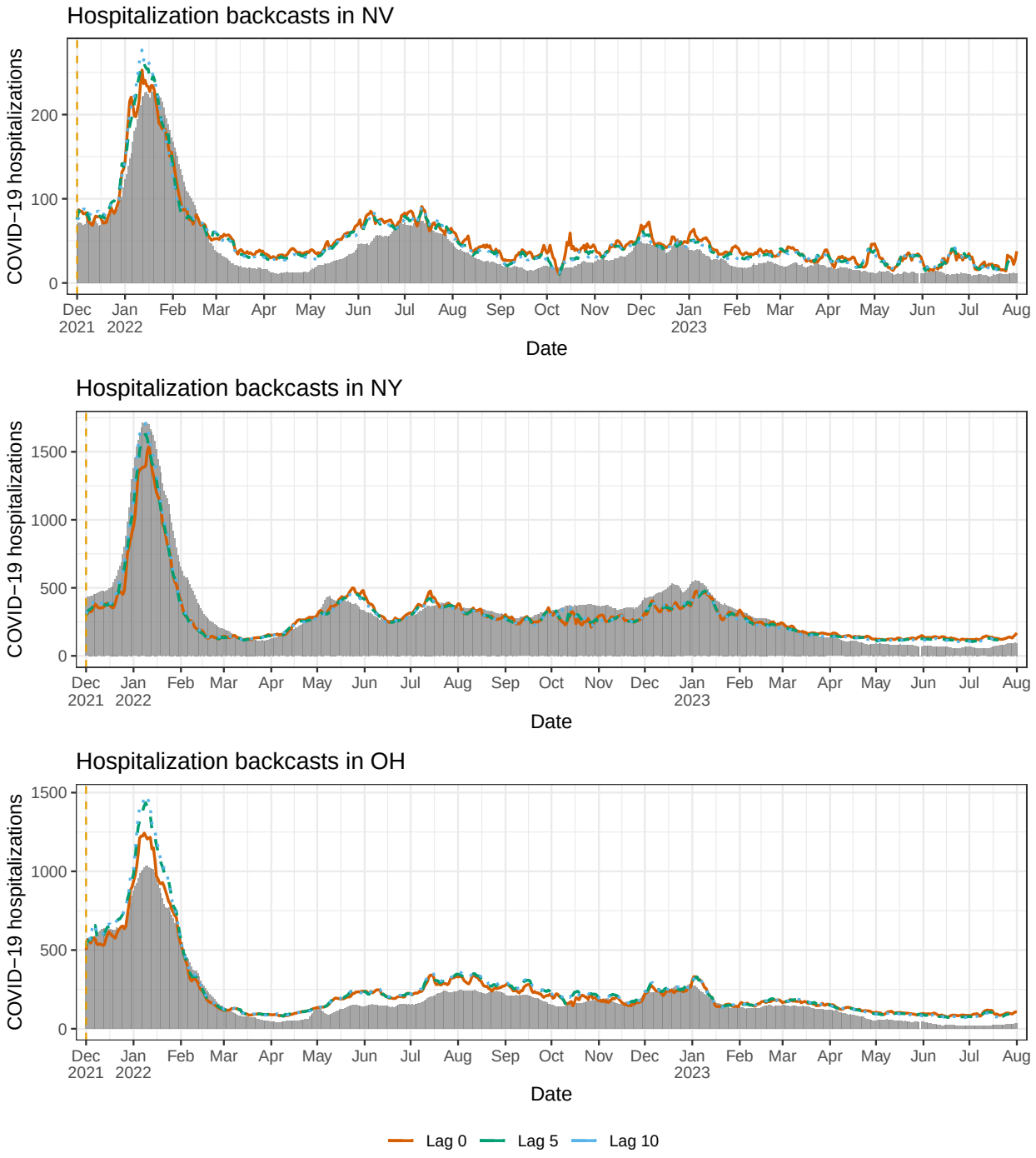

Figure 38: Backcasts from the mixed model in scenario 2, for NV, NY, OH.

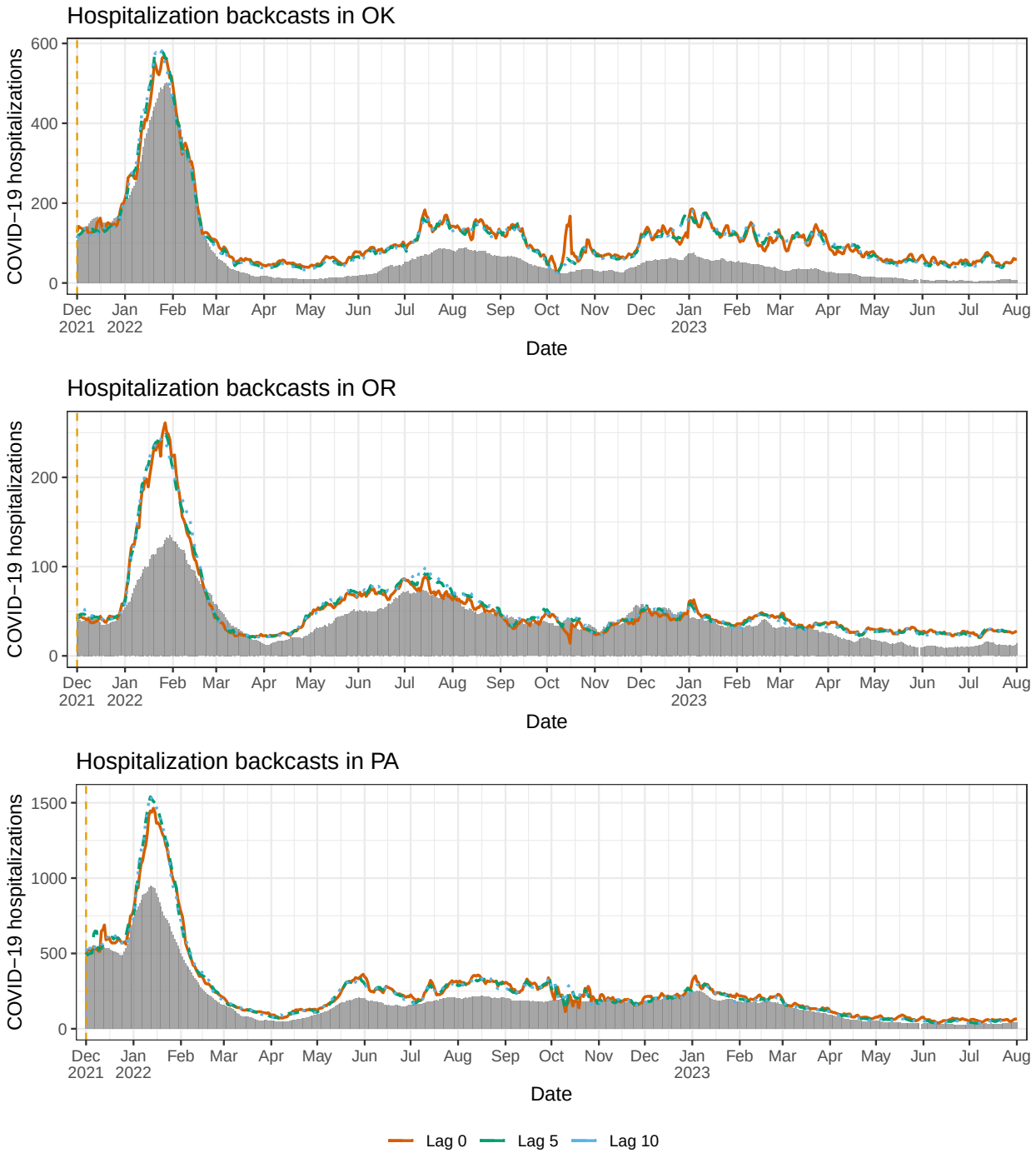

Figure 39: Backcasts from the mixed model in scenario 2, for OK, OR, PA.

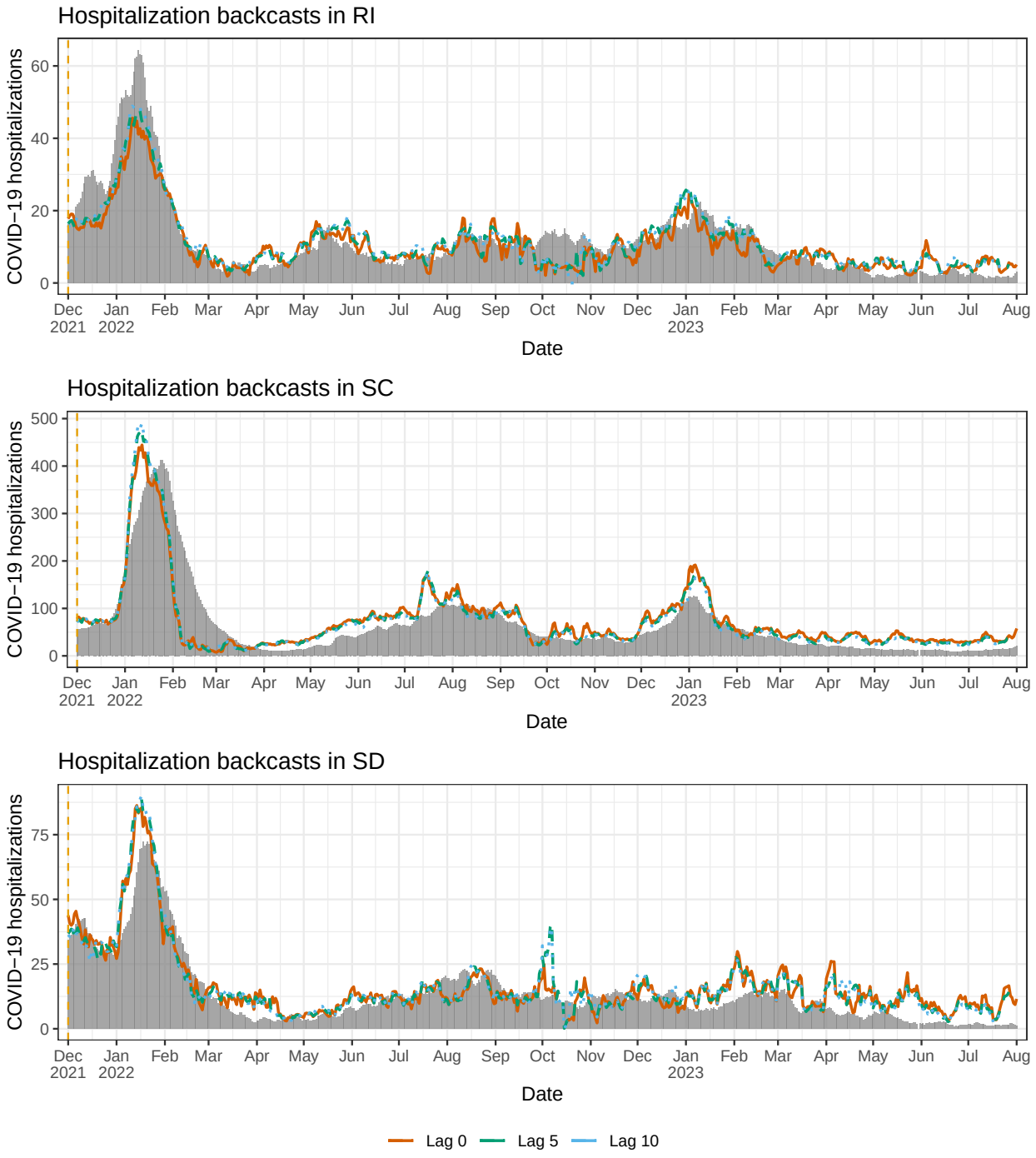

Figure 40: Backcasts from the mixed model in scenario 2, for RI, SC, SD.

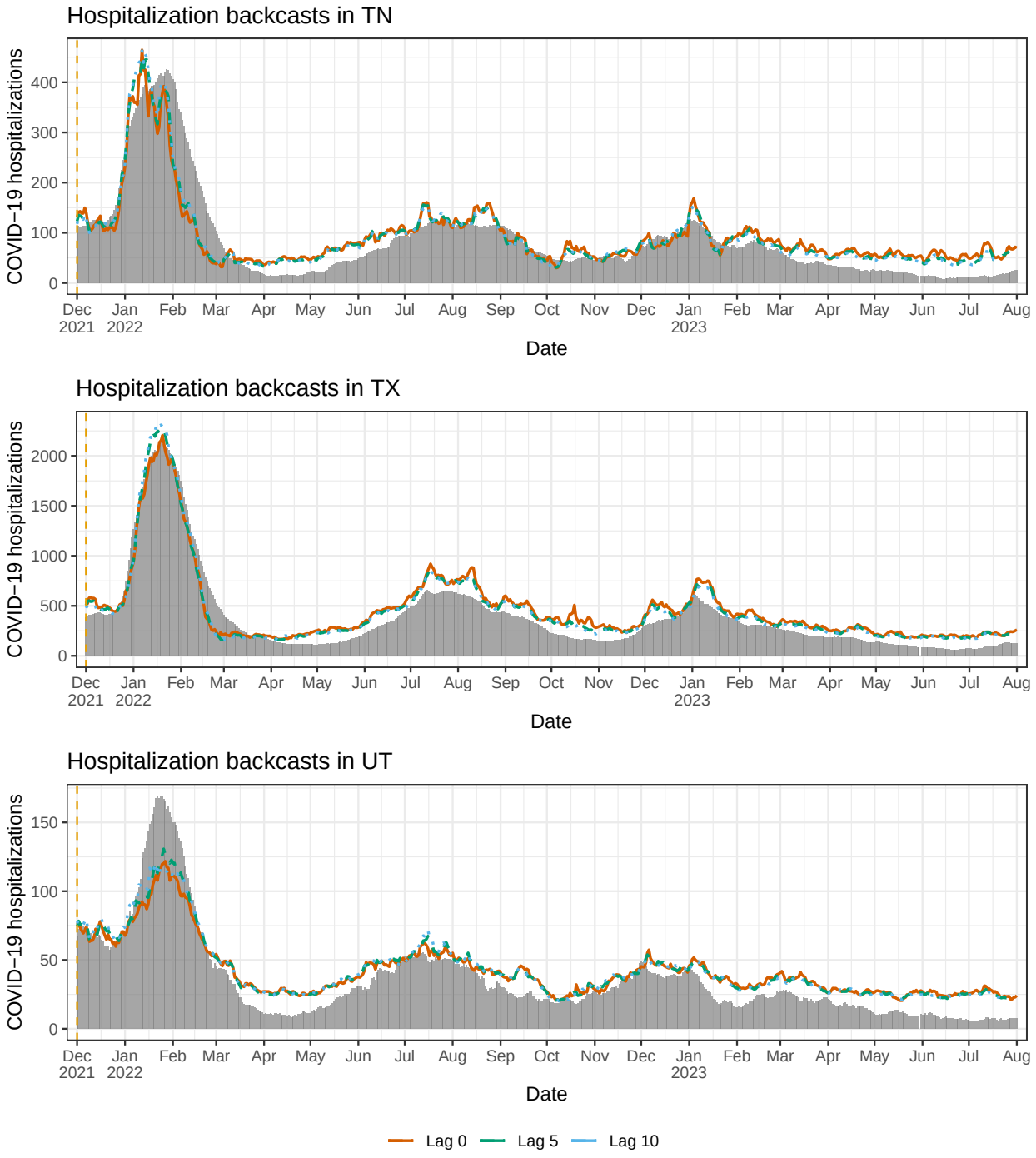

Figure 41: Backcasts from the mixed model in scenario 2, for TN, TX, UT.

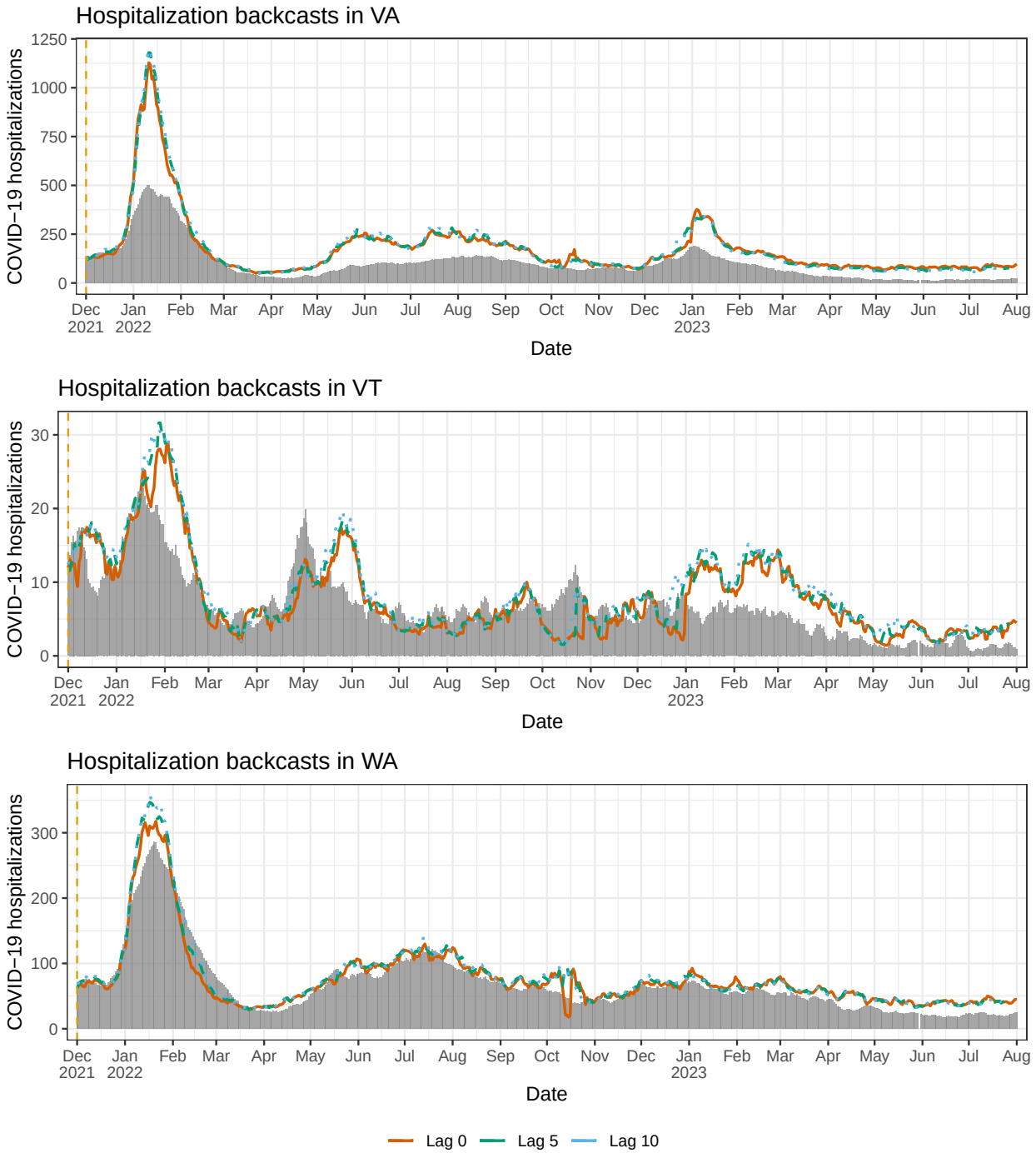

Figure 42: Backcasts from the mixed model in scenario 2, for VA, VT, WA.

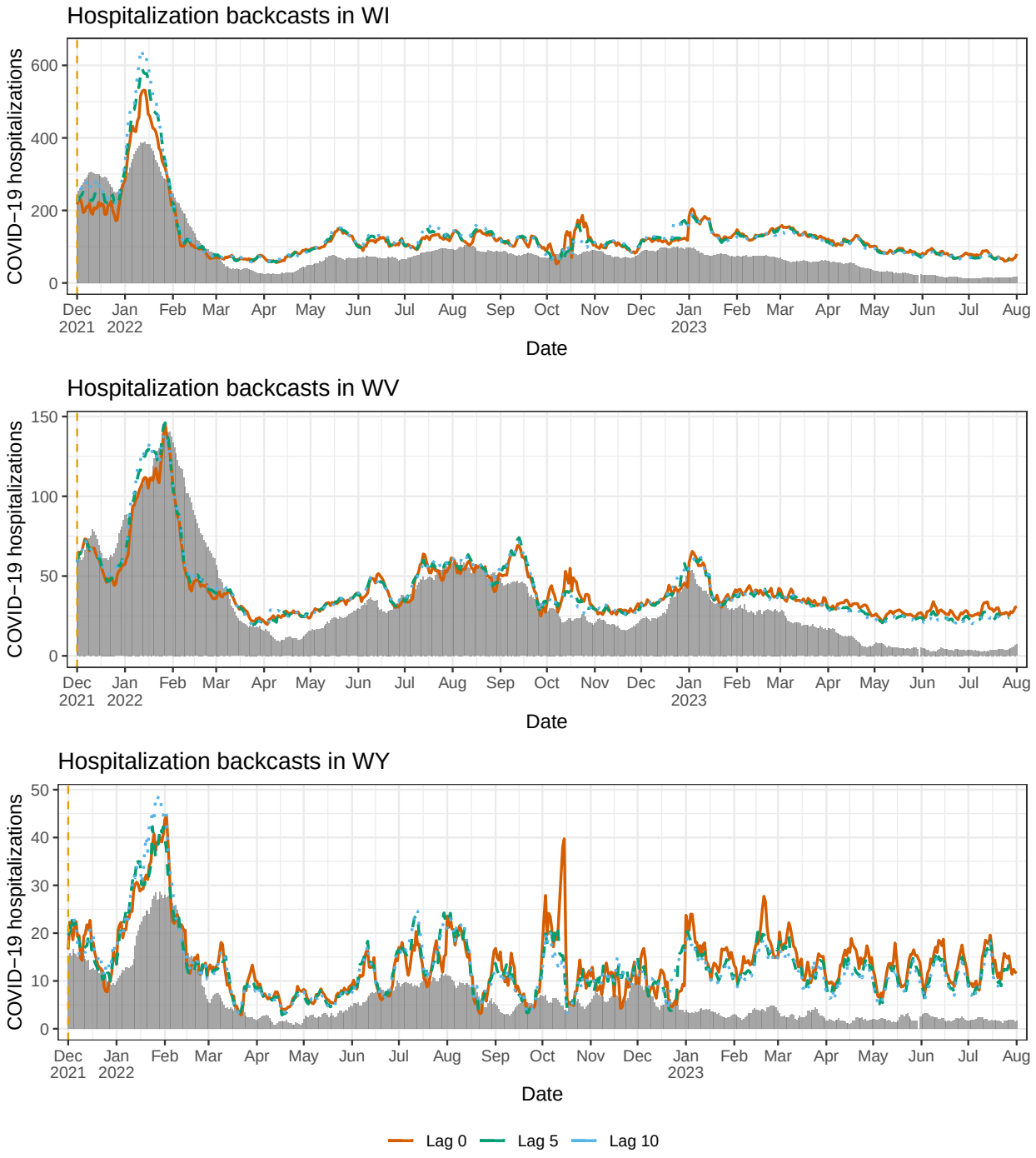

Figure 43: Backcasts from the mixed model in scenario 2, for WI, WV, WY.

E Quantile tracking

In this section, we consider and evaluate a nonparametric method for producing prediction intervals around the nowcasts and backcasts generated by the regression models described in the main paper. The method we use requires feedback (it uses prediction errors in the past to adjust intervals in the future) and so it is only applicable in scenario 1, which is therefore our focus in what follows.

E.1 Methods

Quantile tracking (Angelopoulos et al., 2023) is an algorithm for producing online prediction intervals, which attains long-run coverage for arbitrary bounded data sequences (including highly nonstationary time series). Abstractly, given a predicted value $\hat{y}_t$ of an unobserved target value y_t at time t, we begin with a specification of a score function s (assumed to be negatively-oriented, so that lower values correspond to better accuracy), and construct a prediction set $\mathcal{C}_t$ at time t via

$$\mathcal{C}_t = \{y : s(\hat{y}_t, y) \leq q_t\}.$$

Here q_t denotes a quantile parameter that is trained online by the quantile tracking algorithm, which, given a nominal coverage level $1-\alpha$, runs online gradient descent on a carefully-constructed optimization problem defined by summing the level $1-\alpha$ quantile losses of $s(\hat{y}_t, y) - q_t$ over the sequence $t = 1, 2, 3, \ldots$.

In our application, we do not receive new target values after we make each prediction, unlike the standard online learning setup used in Angelopoulos et al. (2023), so we adapt their method so that q_t is adjusted only at the observation boundaries in scenario 1. Moreover, to allow our intervals to have varying-width between observation boundaries, we use a scaled score that divides the residual by the value of the prediction. Finally, in order to allow our intervals to be asymmetric around the predicted value, we define a separate *lower* and *upper* score, s^l and s^u, respectively:

$$s^l(\hat{y}_t, y_t) = \frac{\hat{y}_t - y_t}{\max\{\hat{y}_t, 1\}},$$
$$s^u(\hat{y}_t, y_t) = \frac{y_t - \hat{y}_t}{\max\{\hat{y}_t, 1\}},$$

and effectively run two instances of quantile tracking, with parameters q_t^l and q_t^u, respectively, each at the level $1-\alpha/2$. Combining the results from these two procedures then gives us the lower and upper endpoints of the prediction intervals.

We are now ready to describe our adaptation of quantile tracking under batched updates. In the same notation as that used to describe fitting the regression models in the main text, let t_0 denote the the latest observation boundary, and t_1 denote the next observation boundary. Suppose we are making backcasts at lag k (with $k = 0$ representing nowcasts). For each $t \in [t_0, t_1)$, we:

- produce a point prediction $\hat{Y}_{t-k}^{(t)}$ using one of the regression models described in the main text;
- produce a prediction interval via

$$\left[\hat{Y}_{t-k}^{(t)} - q_{t_0}^{l,k} \cdot \max\left\{\hat{Y}_{t-k}^{(t)}, 1\right\},\ \hat{Y}_{t-k}^{(t)} + q_{t_0}^{u,k} \cdot \max\left\{\hat{Y}_{t-k}^{(t)}, 1\right\}\right].$$

Then, at $t = t_1$, we:

- look back and compute one-sided coverage errors,

$$\mathrm{err}_{t_0:t_1}^{l,k} = \sum_{s=t_0}^{t_1-1} \mathbb{1}\left\{s^l\left(\hat{Y}_{s-k}^{(s)}, Y_{s-k}^{(t_1)}\right) > q_{t_0}^{l,k}\right\},$$
$$\mathrm{err}_{t_0:t_1}^{u,k} = \sum_{s=t_0}^{t_1-1} \mathbb{1}\left\{s^u\left(\hat{Y}_{s-k}^{(s)}, Y_{s-k}^{(t_1)}\right) > q_{t_0}^{u,k}\right\};$$

- update each of $q_{t_0}^{l,k}$, $q_{t_0}^{u,k}$ by taking appropriate gradient descent steps,

$$q_{t_1}^{l,k} = q_{t_0}^{l,k} + \eta \cdot \big(\mathrm{err}_{t_0:t_1}^{l,k} - \alpha/2\big),$$
$$q_{t_1}^{u,k} = q_{t_0}^{u,k} + \eta \cdot \big(\mathrm{err}_{t_0:t_1}^{u,k} - \alpha/2\big),$$

  where $\eta > 0$ is a small fixed learning rate.

To prevent crossing of quantiles, we enforce $q_{t_1}^{l,k} \geq 0$ and $q_{t_1}^{u,k} \geq 0$ (by clipping them at zero if a gradient descent update were to make them negative).

E.2 Results

We now examine the performance of the quantile tracker, applied to the state-level model in scenario 1. To gain an understanding of how quantile tracking performs against off-the-shelf alternatives, we consider:

- parametric intervals: we apply the standard method for obtaining prediction intervals associated with a well-specified linear regression model with independent Gaussian errors;
- sample quantiles: we obtain intervals by simply computing $q_{t_0}^{l,k}$ and $q_{t_0}^{u,k}$ via sample quantiles of all past scores (in lieu of quantile tracking), where the empirical distribution of past scores is weighted with the same decaying weights that are used to train the regression model.

We consider two nominal coverage levels, 0.6 and 0,8, and evaluate all methods in terms of both coverage of their prediction intervals, and interval score (a metric that combines coverage and sharpness), which for a predicted interval $[l, u]$ at the nominal level $1 - \alpha$, and target value y, is defined as

$$\mathrm{IS}_\alpha\big([l,u], y\big) = (u - l) + \frac{2}{\alpha}(l - y) \cdot \mathbb{1}\{y < l\} + \frac{2}{\alpha}(y - u) \cdot \mathbb{1}\{y > u\}.$$

Figures 44 and 45 display interval score and coverage, averaged over all dates in scenario 1. The setup is altogether similar to that in Figure 3 in the main text, except with either interval score or coverage replacing absolute error. At both nominal levels, quantile tracking clearly dominates the other methods, and it is the only method achieving anywhere close to the nominal coverage level.

Figure 46 visualizes the intervals formed by quantile tracking and alternative methods in CA, KY, VT, as in Figure 4 in the main text. The nominal coverage level is 0.8. We can see that quantile tracking generally accounts for uncertainty much better than the other methods, avoiding exceedingly narrow intervals. Further intervals for PA, OR, AK are shown in Figure 47. These examples show the ability of quantile tracking to "react" to the past month and resize its intervals accordingly, demonstrating better adaptivity than the other methods.

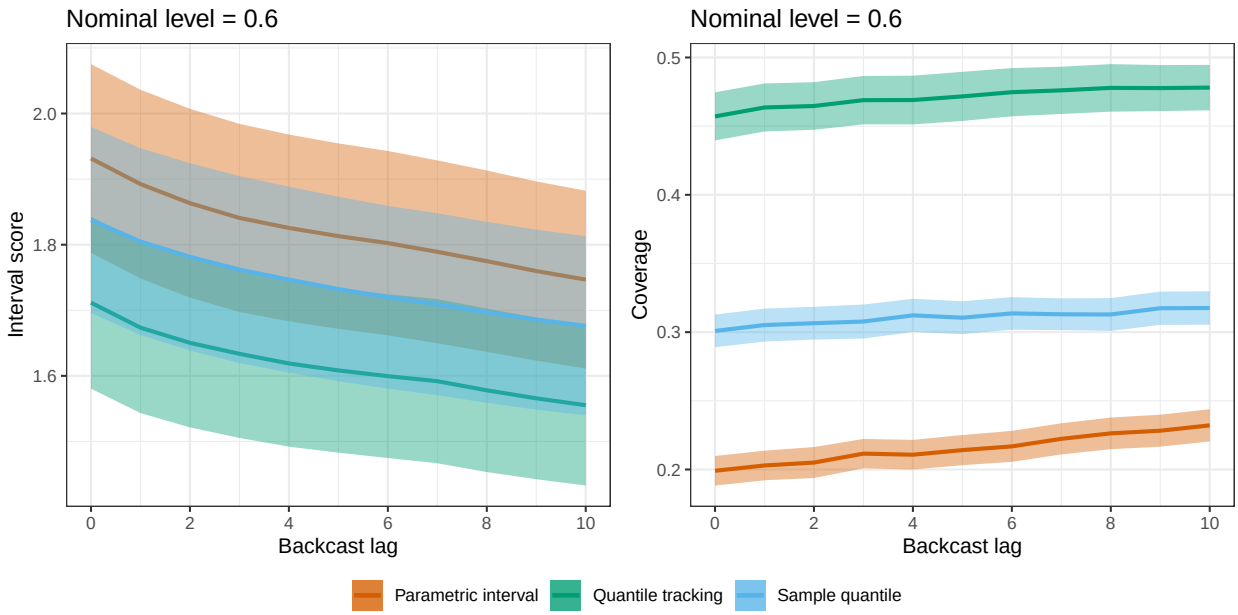

Figure 44: Interval score and coverage as a function of backcast lag for quantile tracking, parametric (linear regression) modeling, and sample quantiles in scenario 1. The shaded regions show ± 1 standard error bands, over the state values.

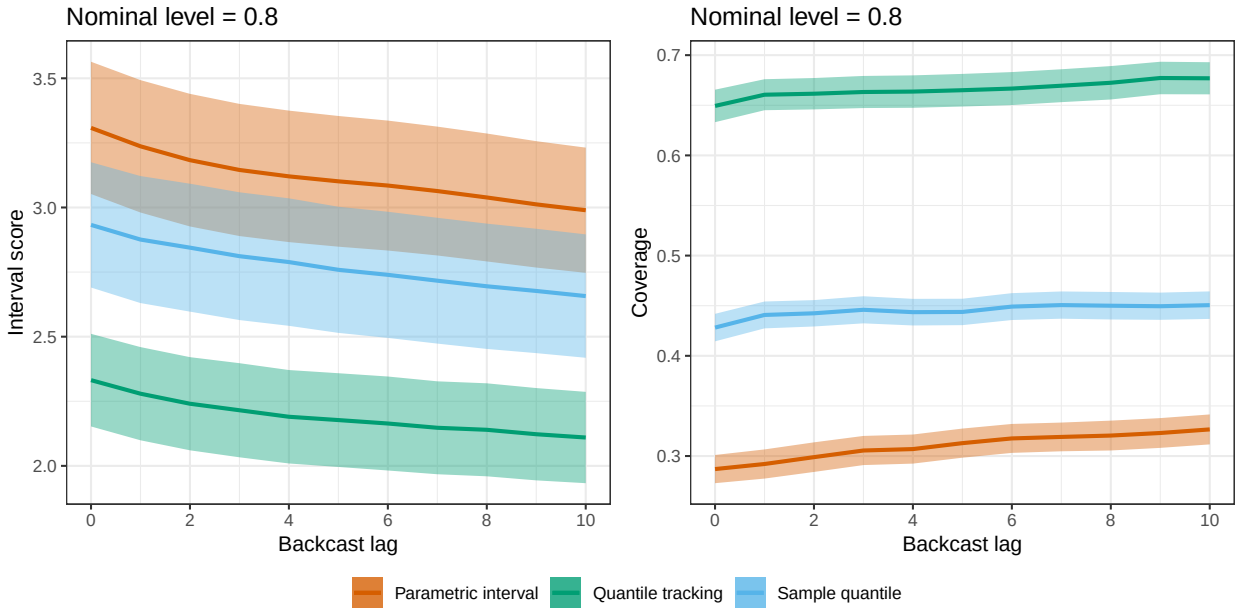

Figure 45: As in Figure 44, but at the nominal level 0.8.

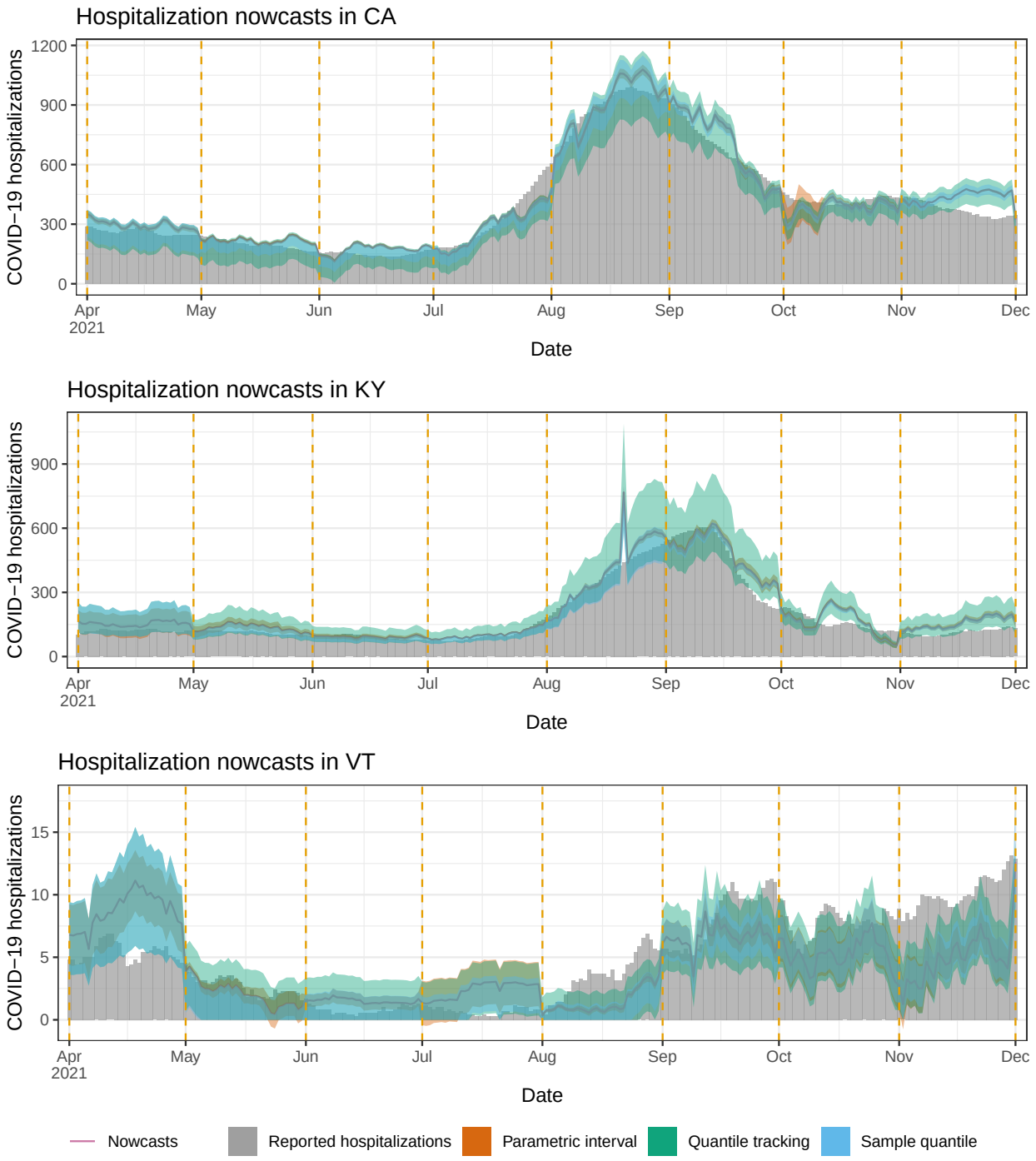

Figure 46: Nowcasts and prediction intervals from quantile tracking, parametric (linear regression) modeling, and sample quantiles in scenario 1, for CA, KY, and VT.

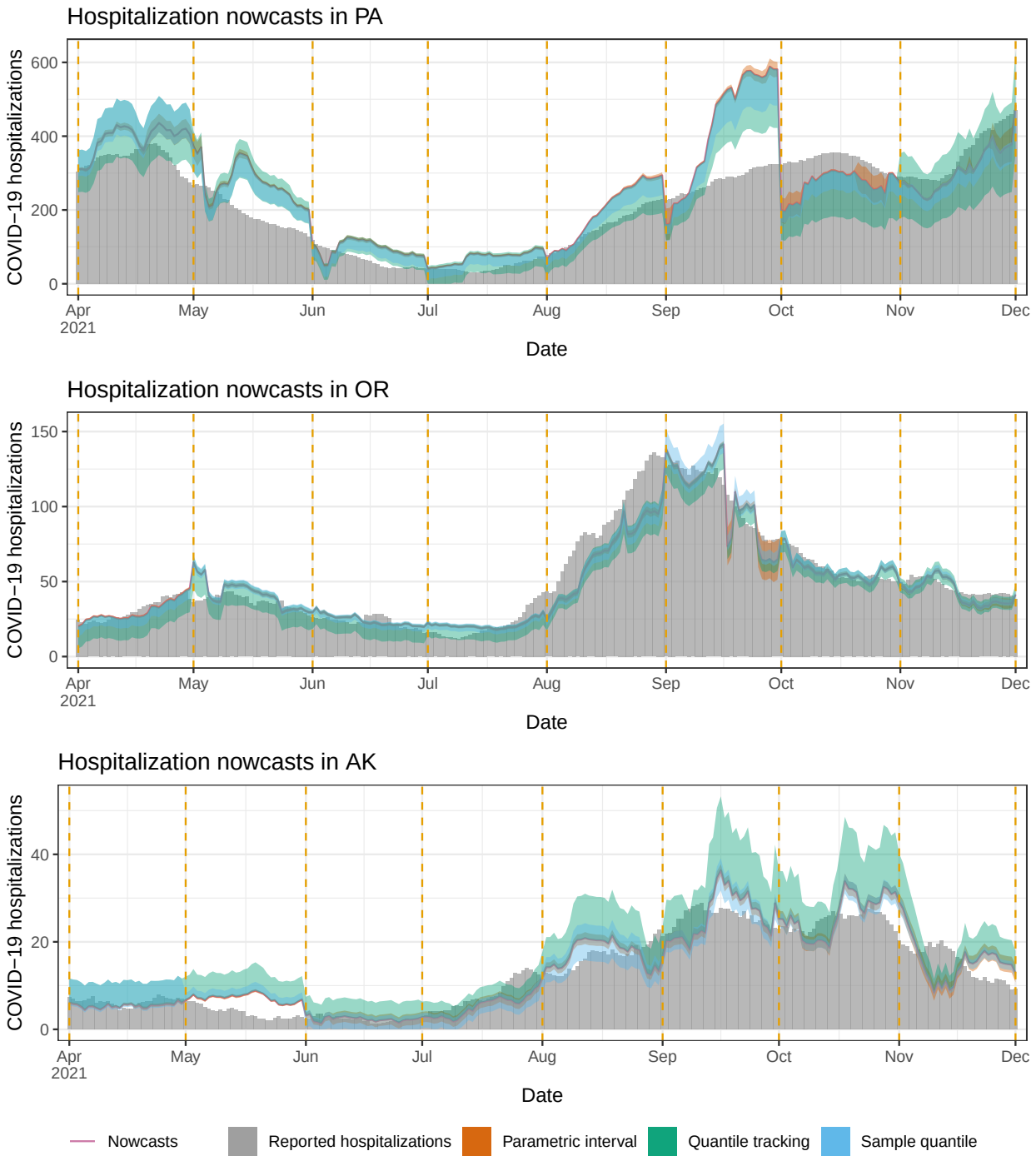

Figure 47: Nowcasts and prediction intervals from quantile tracking, parametric (linear regression) modeling, and sample quantiles in scenario 1, for PA, OR and AK.