nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23234-5 OPEN

Multiscale influenza forecasting

Dave Osthus [1]✉ & Kelly R. Moran [1,2]

Influenza forecasting in the United States (US) is complex and challenging due to spatial and temporal variability, nested geographic scales of interest, and heterogeneous surveillance participation. Here we present Dante, a multiscale influenza forecasting model that learns rather than prescribes spatial, temporal, and surveillance data structure and generates coherent forecasts across state, regional, and national scales. We retrospectively compare Dante's short-term and seasonal forecasts for previous flu seasons to the Dynamic Bayesian Model (DBM), a leading competitor. Dante outperformed DBM for nearly all spatial units, flu seasons, geographic scales, and forecasting targets. Dante's sharper and more accurate forecasts also suggest greater public health utility. Dante placed 1st in the Centers for Disease Control and Prevention's prospective 2018/19 FluSight challenge in both the national and regional competition and the state competition. The methodology underpinning Dante can be used in other seasonal disease forecasting contexts having nested geographic scales of interest.

[1] Los Alamos National Laboratory, Statistical Sciences Group, Los Alamos, NM, USA. [2] Department of Statistical Science, Duke University, Durham, NC, USA.
✉email: dosthus@lanl.gov

Influenza represents a significant burden to public health with an estimated 9 to 49 million cases each year in the United States (US)[1]. Influenza (flu) related activity is monitored in the US by the Centers for Disease Control and Prevention (CDC) through numerous surveillance efforts. One such effort is the Outpatient Influenza-like Illness Surveillance Network (ILINet). ILINet collects weekly data on influenza-like illness (ILI) from over 2000 healthcare providers from all 50 states, Puerto Rico, the US Virgin Islands, and the District of Columbia. ILI is defined as a temperature greater or equal to 100 °F, a cough or sore throat, and no other known cause, representing symptoms consistent with influenza. ILINet constitutes a significant and necessary effort to understanding the spread and prevalence of flu-like illness in the US in near real-time.

With mature ILI surveillance infrastructure in place in the US, attention has turned in recent years to ILI prediction. The ability to predict the spread of ILI poses a substantial public health opportunity if able to be done accurately, confidently, and with actionable lead times at geographic and temporal scales amenable to public health responsiveness. Since 2013, the CDC has hosted an influenza forecasting challenge called the FluSight challenge to gauge the feasibility of forecasting targets of public health interest in real-time, to galvanize the flu forecasting community around common goals, and to foster innovation and improvement through collaboration and competition[2–4]. The FluSight challenge has been a leading driver of recent model development and flu forecasting advancements[5–21].

Up until the 2016 flu season (i.e., the flu season starting in the fall of 2016 and ending in the spring of 2017), the FluSight challenge's scope encompassed probabilistic forecasting of short-term (1 to 4 week ahead) and seasonal (season onset, peak timing, and peak intensity) targets at two geographic scales: nationally and regionally, where regions correspond to Health and Human Services (HHS) regions. Probabilistic forecasting is carried out by binning the support of the targets (e.g., binning the peak timing target into weeks of the season) and assigning a probability to each bin, representing the probability the eventual outcome will fall in each bin. Probabilistic forecasts are a crucially important component of the FluSight challenge, as it demands not only information on what the forecasting models thinks will happen, but also how confident the forecasting model is in its own prediction. National and regional forecasts give a high-level view of flu activity across the US. Those forecasts provide value to national and regional public health officials, but offer only coarse information for state and local public health practitioners. Thus, starting with the 2017 flu season, the FluSight challenge expanded to a third geographic scale: states and territories (referred to as states). This expansion to a finer geographic scale presents an opportunity to move forecasting to geographic scales better aligned with public health response infrastructure and decision making. It also presents an opportunity to develop and advance methodological forecasting frameworks that can share information across geographic locations, flu seasons, and geographic scales coherently in ways that geographically isolated forecasting models cannot.

Multiscale forecasting in the US requires careful consideration as it presents numerous challenges. For instance, Fig. 1 (as well as Supplementary Fig. 3) shows appreciable state-to-state ILI variability. As an example, Montana's average ILI is about 20% the national average, while the District of Columbia's and Puerto Rico's average ILI is about 250% the national average. Figure 1 also shows evidence of spatial correlation, with states near the Gulf of Mexico having higher than average ILI while most Midwest and Mountain West states have lower than average ILI. Attempts to model the spatial relationships of flu and flu-like illnesses include using network models and US commuter data[5], network models based on Euclidean distance[19], and empirically derived network relationships[21,22]. Though these approaches consider spatial relationships differently, they all support the conclusion that sharing information across geographical units can improve forecasting.

Figure 2 (and Supplementary Fig. 4) shows season-to-season variability, illustrating the common directional effect a flu season can have on nearly all states. 2015, for instance, was a mild flu season in the US with 42 out of 53 states experiencing ILI activity below their state-specific averages (the 53 states are all 50 states, minus Florida with no available data, plus Puerto Rico, the US Virgin Islands, New York City, and the District of Columbia). In contrast, 2017 was an intense flu season with 47 out of 53 states experiencing ILI activity above their state-specific averages. Supplementary Fig. 5 visualizes this information with six intensity levels rather than an above/below average binary. Similar to the findings that sharing information across geographical units can improve flu forecasting, previous work has found that sharing information across seasons can also improve flu forecasting[13]. Similarly positive findings have been identified in other disease modeling contexts using latent random walks[23,24].

Figure 3 shows the average standardized week-to-week volatility across geographic scales (see Supplementary Note 3 for details). Standardized volatility measures how much ILI (states) and wILI (regions and nationally) varies from week-to-week, where wILI is weighted ILI—a state-population weighted version of ILI used to characterize ILI regionally and nationally. High volatility poses a challenge to forecasting as increased volatility can swamp the signal in the (w)ILI data. Figure 3 makes clear that extending forecasts down to the state scale, a more actionable scale for public health officials, comes at the cost of increased volatility. The level of volatility is largely driven by the number of patients seen weekly, as illustrated in Fig. 3. Developing multiscale flu forecasting models that account for decreasing volatility with coarsening geographic scales will be crucial. Some multiscale forecasting models have been developed in the context of norovirus gastroenteritis prediction[25]. In that work[25], showed that modeling at the finest available data scale and aggregating up to coarser scales generally had better predictive performance than models directly operating at the aggregated scales. To our knowledge, such models have not been developed and operationalized for ILI forecasting.

It is in the context of appreciable state-to-state and season-to-season variability, uneven ILINet surveillance participation, and the need to render short-term and seasonal probabilistic forecasts at nested geographic scales that Dante, a probabilistic, multiscale flu forecasting model, was developed. While efforts have been made to address each of these challenges in isolation, no one has yet to tackle all of these challenges simultaneously in the context of influenza forecasting. Jointly addressing all these challenges is the main contribution of this paper.

Dante is composed of two submodels: a fine-scale model for the state scale and an aggregation model for the regional and national scales. The state submodel includes both a data model and a process model. The data model is where Dante learns information about the level of volatility in the ILI time series. The process model is where Dante captures common and specific structure in the data, including a term common to all states and seasons, a state-specific term, a season-specific term, and a state-season specific interaction term. The common term can be thought of as the average profile of the ILI data across space and time, whereas the interaction term captures deviations from common, state-, and season-specific baselines. The aggregation model builds regional and national forecasts using the state forecasts as population weighted building blocks, leading to

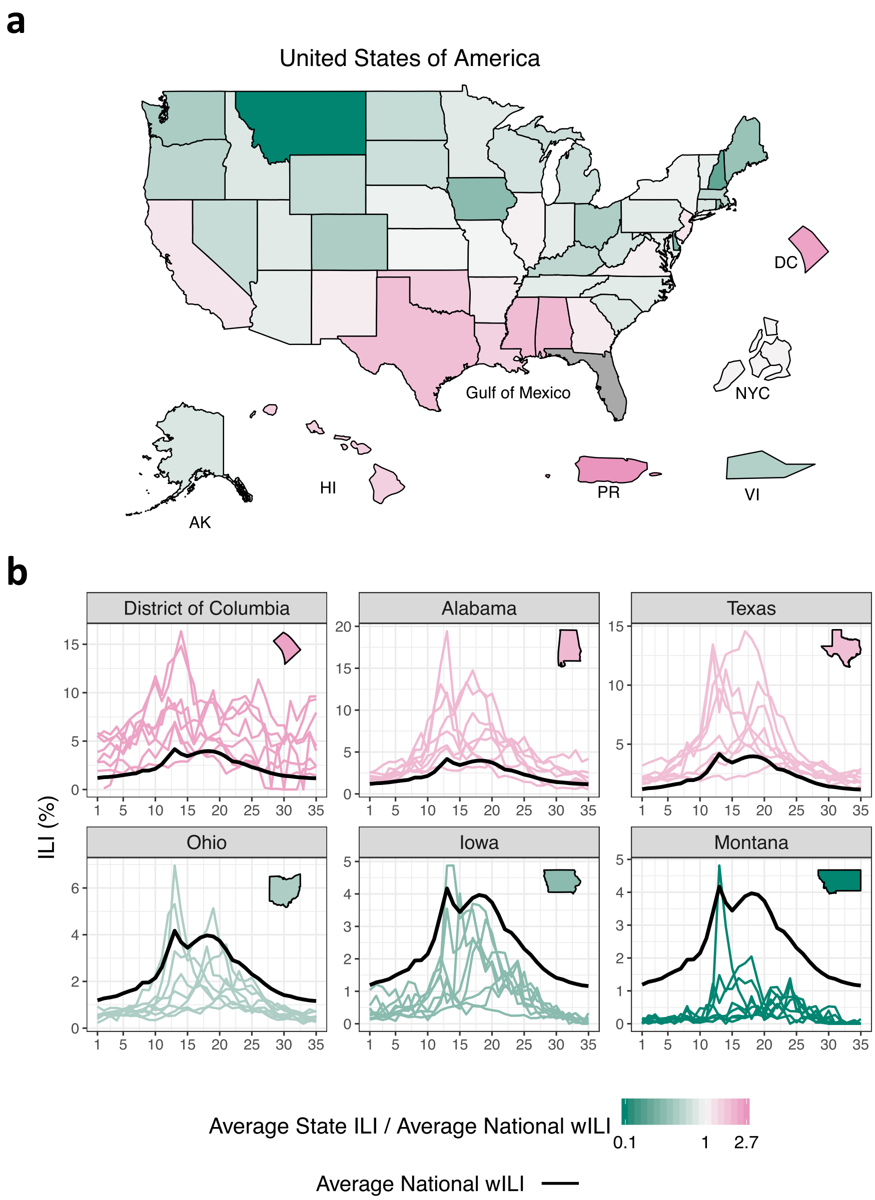

Fig. 1 State-to-state influenza-like illness variability. a Average state influenza-like illness (ILI) relative to average national weighted ILI (wILI). States bordering the Gulf of Mexico tend to have higher ILI than the national average. The geographical sizes of Alaska (AK), Hawaii (HI), Puerto Rico (PR), the US Virgin Islands (VI), New York City (NYC), and the District of Columbia (DC) are not to scale. Data for Florida is unavailable. Averages are based on 2010 through 2017 data. **b** ILI by season (colored lines) for select states. Black line is national average wILI for reference. Appreciable season-to-season and state-to-state ILI variability exists.

forecasts that are coherent across geographic scales. Full details of both submodels are provided in the Methods section.

Results

Dante is compared to a leading flu forecasting model, the Dynamic Bayesian Model (DBM)[13]. Like Dante, DBM models ILI data as the sum of component trajectories. DBM's components include a season/region-specific susceptible-infectious-recovered (SIR) compartmental model, a region-specific statistical discrepancy component capturing deviations from the SIR component common across seasons (e.g., holiday effects), and a regularized season/region-specific discrepancy component to capture ILI structure unable to be captured by the other two components. DBM is fit to each geographic unit separately, thus does not share information across geographic units or geographic scales, but does share information across flu seasons. In contrast, Dante shares information across geographic units, geographic scales, and flu seasons. DBM was the fourth place model and a component model in the second place ensemble model[18] in the prospective national and regional 2017/18 FluSight challenge out of 29 participating models. Dante was the first place model in the 2018/19 national and regional FluSight challenge out of 33 participating models. Dante also came in first out of 14 competing models in the 2018/19 state challenge.

We compare Dante and DBM using forecast skill following the scoring rules of the CDC's FluSight challenge (details in Supplementary Note 8), noting both that this scoring rule is improper[18,26,27] and that a proper scoring rule has been implemented in the FluSight challenge starting with the 2019/20 season. In this paper, forecast skill is defined as the exponentiated average over forecast scores and ranges between 0 and 1, with 1 being the best possible forecast skill. Put another way, skill here is

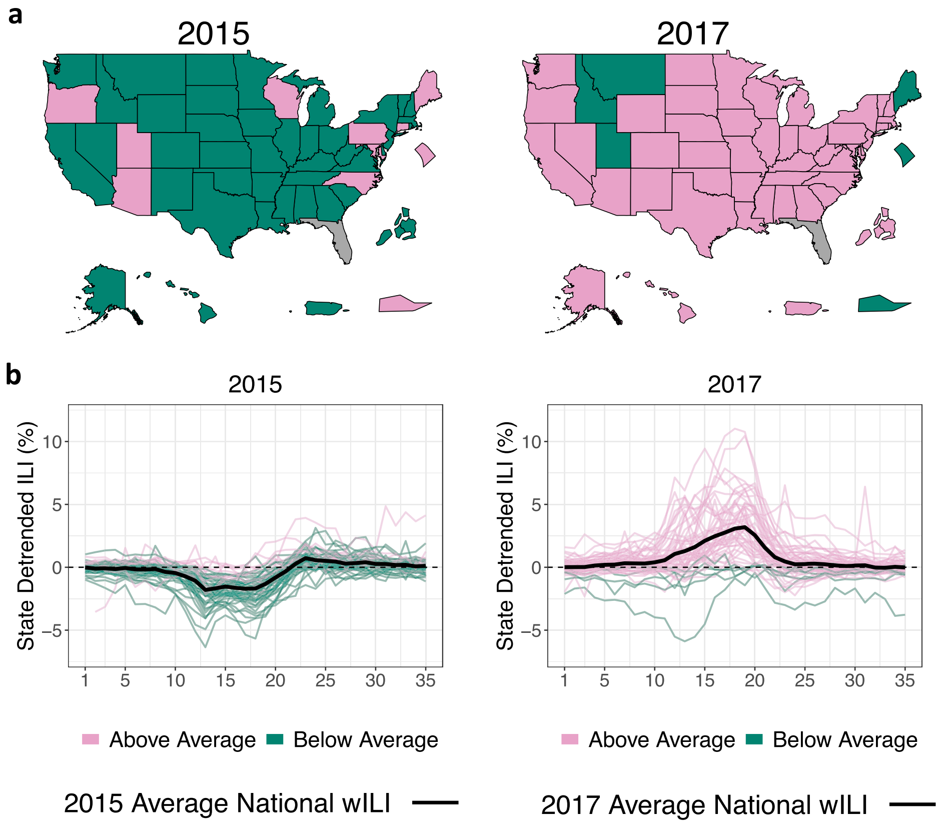

Fig. 2 Seasonal influenza-like illness variability. a Dark green states denote states with ILI less than their state-specific averages while pink states are states with ILI above their state-specific averages. 2015 was a mild flu season for the majority of states relative to their state-specific average ILI, while 2017 was an intense flu season for the majority of states, indicating that season-to-season effects can affect most of the country. Data unavailable for Florida. States displayed outside of the contiguous US are geographically not to scale. **b** State detrended ILI for the 2015 and 2017 flu seasons, where state detrended ILI is ILI for a state/season minus ILI for that state averaged over all seasons. Positive/negative state detrended ILI means ILI for that season was above/below the state-specific average, respectively. Black line is season-specific national average wILI for reference.

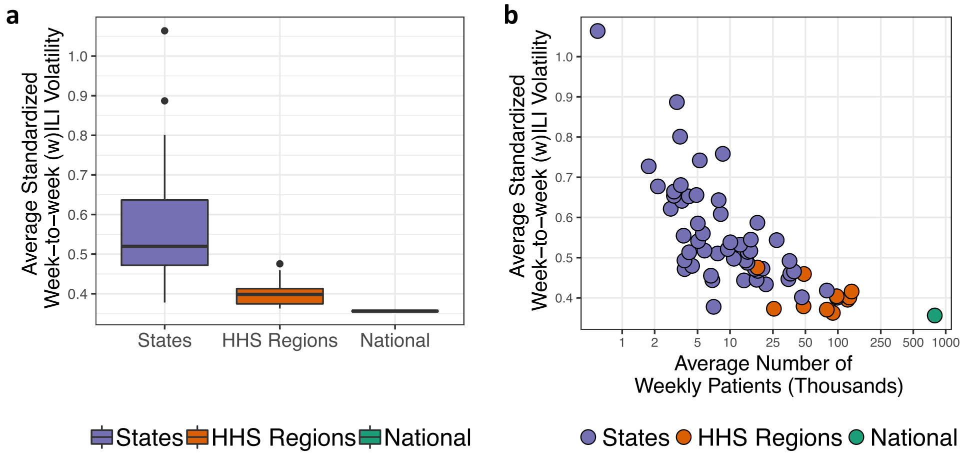

Fig. 3 Standardized volatility by geographic scale. a Average standardized week-to-week influenza-like illness (ILI—states) and weighted ILI (wILI—HHS regions and nationally) volatility for three geographic scales. Volatility decreases as the scales coarsen. Boxplots present median (center line), interquartile range (boxes), 1.5 times the interquartile range (whiskers), and outliers (points) based on $n = 53$, 10, and 1 observations for states, HHS regions, and the nation respectively. **b** Average standardized week-to-week (w)ILI volatility versus the average number of weekly patients on a log scale for each state, HHS region, and nationally. Volatility decreases as the number of weekly patients seen increases, suggesting that volatility is in part a product of ILINet participation.

Table 1 Comparison of forecast performance between Dante and DBM.

Metric	Model	States	HHS Regions	National
Skill	Dante	0.372 (0.060)	0.413 (0.073)	0.439 (0.078)
	DBM	0.337 (0.053)	0.383 (0.066)	0.426 (0.074)
MSE	Dante	3.164	2.441	1.921
	DBM	3.390	2.674	2.244

Dante and DBM average forecast skill and mean squared error (MSE) comparisons across geographic scales. For skill results both the FluSight challenge's forecast skill and the exponentiated, proper log score are reported, with the proper skill in parentheses. Forecast skill and MSE each improve for both models as the geographic scale coarsens. Dante outperformed DBM at all geographic scales under both performance metrics.

the geometric average probability assigned to the observed outcome or, in the case of the CDC's multibin score, values reasonably close to the observed outcome. This differs from the definition of skill in[28] defining an average over forecast scores in relation to a reference forecast. Conceptually, skill is a function of both accuracy (a measure of a point summary of a distributional forecast) and sharpness (a measure of concentration of the distributional forecast). We also score both models using a proper log scoring rule (details in Supplementary Note 8.3). We will show how Dante compares to DBM broadly in terms of skill, and also in terms of its component pieces. Both models were fit in a leave-one-season out fashion, where the data for all seasons not being forecasted along with the data for the season being forecasted up to the forecast date were used for training.

Table 1 shows that Dante outperformed DBM in forecast skill at all geographic scales as calculated by both the improper CDC scoring rule and a proper log scoring rule. Dante also outperformed DBM in terms of accuracy, as measured by mean squared error (MSE) of point predictions (posterior means), at all geographic scales. For both models, forecast skill improves and average MSE decreases as geographic scales coarsen, suggesting that both forecast skill and accuracy degrade as we move to finer scales where volatility is greater. See Supplementary Note 12 for further details and figures comparing the MSE of Dante and DBM.

Figure 4 shows the ratio of forecast skill of Dante to that of DBM for each state, region, and nationally. Dante outperformed DBM for the majority of geographic regions, with the exception of HHS Region 7 and the states Wyoming, Puerto Rico, and Kentucky.

Figure 5 shows forecast skill broken down by targets (left) and flu seasons (right) for each geographic scale. Dante outperformed DBM for all scales and targets, except for peak intensity regionally and onset nationally. Improvement over DBM is largest for the 1-week ahead forecast target. For context, in the 2018/19 FluSight national and regional challenge, Dante placed first for all short-term targets, season onset, and peak intensity (PI), while placing ninth for peak timing (PT). In the state challenge, Dante placed first in all short-term targets and PT, while placing second for PI. Dante also outperformed DBM for all scales and flu seasons, except for 2017 nationally. While forecast skill for DBM and Dante are close for all seasons nationally (sans 2016), Dante consistently and appreciably outperformed DBM for all seasons at the regional and state scales.

Figure 6a provides context as to how Dante is outperforming DBM. Figure 6a displays the ratio of forecast skills at each scale for all short-term forecasts against the difference in the 90% highest posterior density (HPD) predictive interval widths for each of the short-term target's posterior predictive distributions. See Supplementary Note 9 for calculation details. HPD predictive intervals are similar to equal-tailed predictive intervals as they capture the range of probability concentration, but are more appropriate than equal-tailed predictive intervals for distributions that are not unimodal and symmetric. Figure 6a shows that for all scales and short-term forecasts, Dante has smaller HPD interval widths, indicating that Dante's forecasts are more concentrated (i.e., sharper) than DBM. Dante's increased forecast sharpness resulted in higher forecast skill than DBM. This is a promising finding, as sharper forecasts, if well-calibrated, provide more information to public health decision makers.

Figure 6b shows that Dante's forecasts are sharper than DBM's for all short-term targets at all geographic scales. For each short-term target, forecasts for both DBM and Dante become sharper as the geographic scale coarsens. DBM makes sharper short-term forecasts because the (w)ILI DBM is modeling less volatile, i.e., because (w)ILI becomes less volatile as geographic scales coarsen. Dante makes sharper short-term forecasts at coarsening geographic scales as a result of the aggregation model. Dante's 3-week-ahead 90% HPD interval widths nationally and regionally are 1.4 and 2.1%, respectively, about the same as Dante's 1-week-ahead 90% HPD interval widths are regionally and at the state-level, respectively. Said another way, Dante loses about 2 weeks of sharpness in its short-term forecasts for each disaggregating geographic scale.

Though Dante won the 2018/19 FluSight challenge and outperformed a leading flu forecasting model, DBM, in a retrospective comparison, Dante is still a work in progress. Figure 7 shows Dante's 90% empirical coverages for short-term targets and represents an opportunity for future Dante improvement. Overall, empirical coverages for states are close to their nominal coverages. Those empirical coverages drop, however, as geographic scales coarsen. Furthermore, empirical coverages drop with a growing forecast window. The right of Fig. 7 breaks down empirical coverages into the stages relative to the season peak. Empirical coverages are generally good in the post-peak stage of the flu season, arguably when forecasts are least useful. Fig. 7 makes clear that while Dante represents the state of the art, flu forecasting is a field ripe for improvement, advancement, and innovation.

Discussion

The plot of skill score (Fig. 5) and total patients seen and volatility (Fig. 3) suggests that Dante (and DBM) can forecast geographic regions better when the forecasted estimate is based on more data than less. This result suggests that expanded ILI surveillance participation plays a role in improving model forecasts, not just improvements to the models themselves. This idea is not surprising. Disease forecasting has been compared to weather forecasting[29], a field that has continued to make consistent progress through parallel efforts of improved modeling and data collection.

We found that Dante made sharper forecasts, as measured by smaller 90% HPD interval widths than DBM, a model that fits aggregate data directly. Similar findings were noted by[25] when comparing models fit to norovirus data in Berlin stratified by regions and age groups—models fit to the finest scales and subsequently upscaled had sharper predictions than models fit to the aggregated scales directly. This suggests that continued stratification of ILI, such as partitioning state-level ILI by age groups, by flu strain type, or county-level, may provide further sharpening of forecasts at aggregate scales.

Dante's first place finish in the 2018/19 FluSight challenge may come as a surprise given that it is a purely statistical model and uses only ILINet data, while many of its competitors are based in

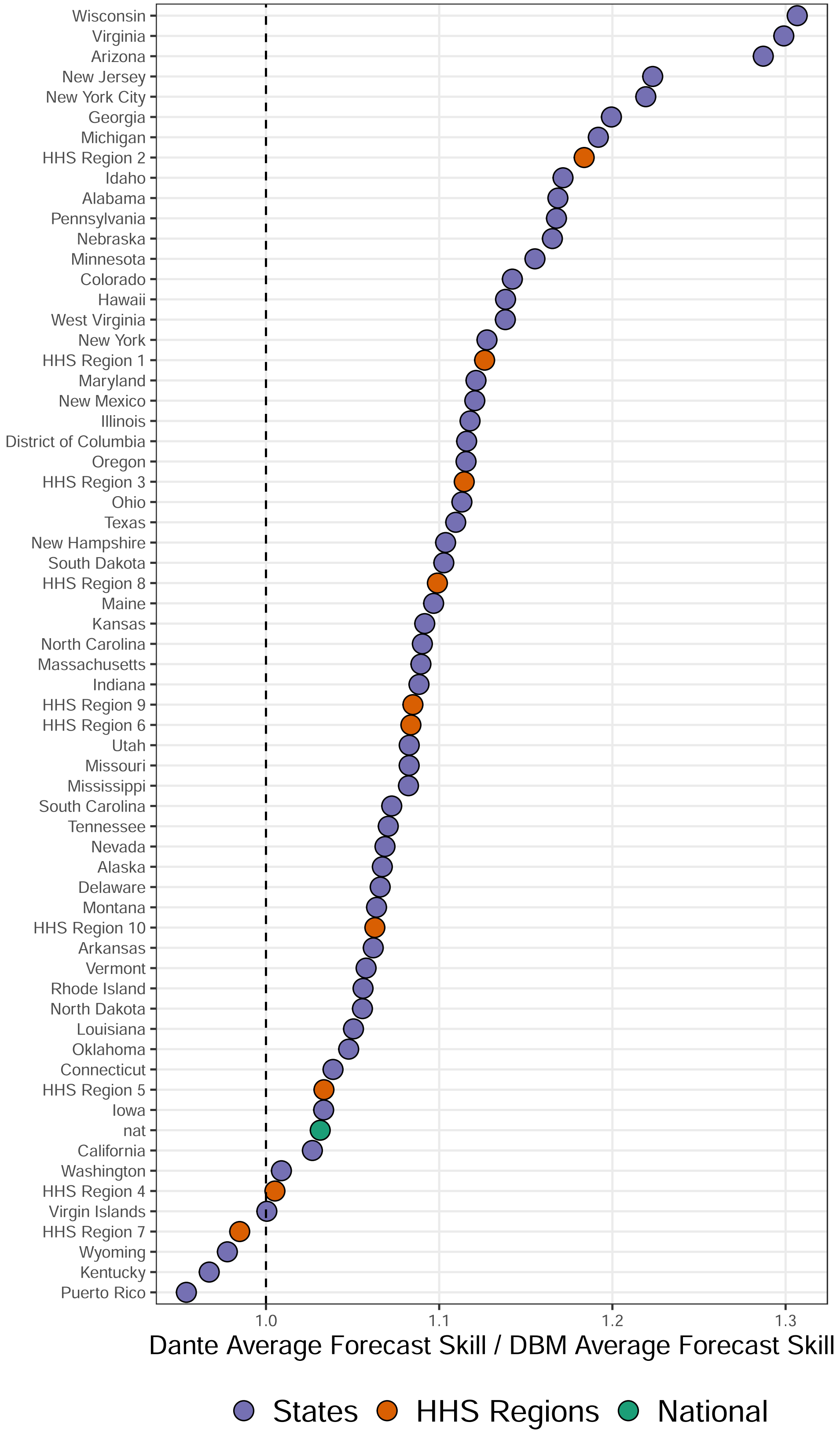

Fig. 4 Dante's forecast skill relative to DBM's by geographic region. Ratio of forecast skill of Dante to that of DBM, for all states, regions, and nationally. Dante had higher forecast skill for all geographic regions except for HHS Region 7, Kentucky, Wyoming, and Puerto Rico.

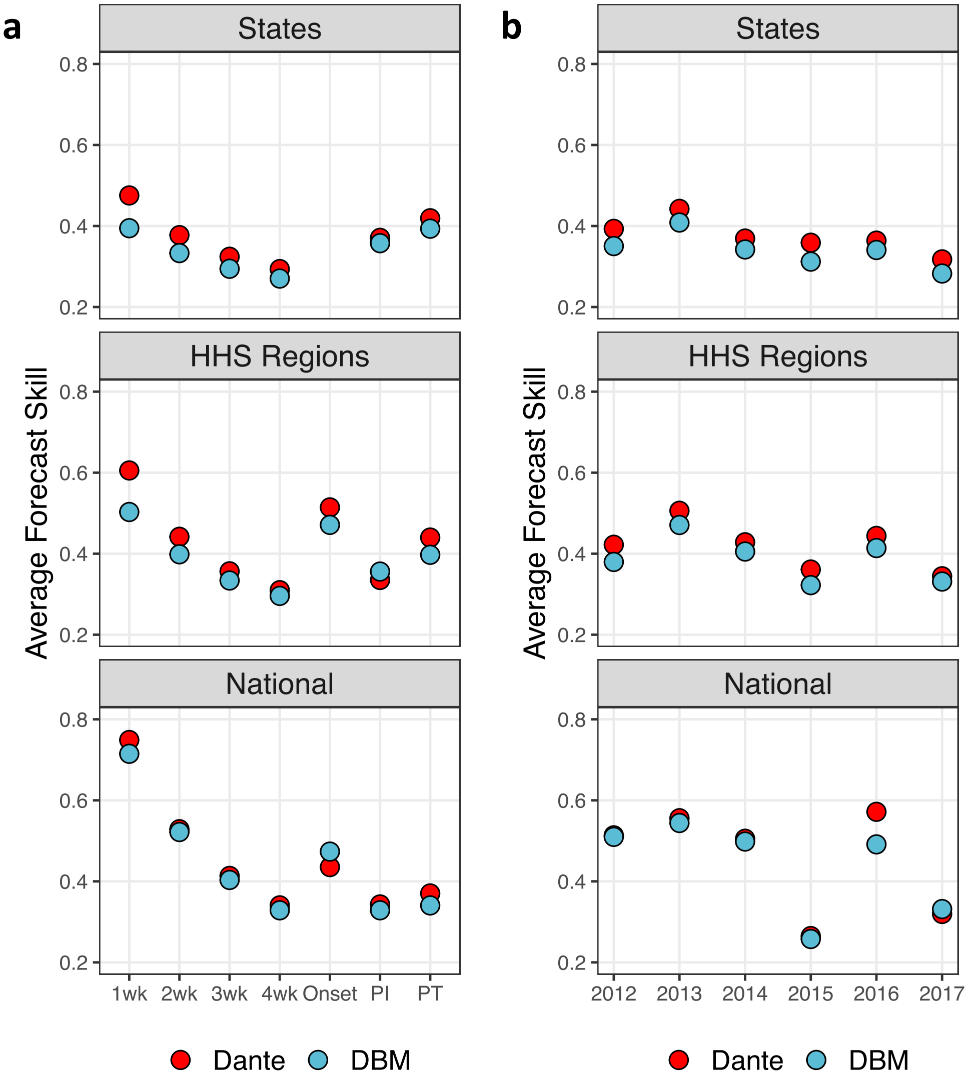

Fig. 5 Dante's and DBM's forecast skill by target and season. a Average forecast skill by scales and targets. PI and PT stand for peak intensity and peak timing, respectively. Dante outperformed DBM for all scales and targets, except for onset nationally and for PI regionally. **b** Average forecast skill by scales and flu seasons. Dante outperformed DBM for all scales and targets, except for 2017 nationally.

full or in part on mechanistic disease transmission models and/or are augmented with alternative data sources (e.g., Google search data). Dante's superior performance suggests that these mechanistic components or alternative data sources may be integrated into those models in a way that is improperly aligned with the truth. For example, DBM includes an SIR model component via a season-specific "I" term but a given season may have multiple circulating influenza strains responsible for the "true" flu component in the ILI data, thus rendering the use of a single "I" term inappropriate.

Revisions made to ILI data after its initial release are referred to as backfill and constitute a meaningful source of uncertainty when prospectively forecasting ILI. In this work, backfill was ignored for the forecasting of both Dante and DBM. As a result, the forecasting results for Dante and DBM are directly comparable to each other in this paper but are not directly comparable to previous FluSight challenge results. The reason backfill was ignored in this paper is because Dante uses state-level ILI data directly, and state-level backfill data is only available starting for the 2017/18 season. Though backfill was not addressed in this paper, Dante's winning 2018/19 FluSight challenge entry did include a backfill model to account for the revision process of real-time ILI data.

Linking understandable processes to observed patterns in the data via models while maintaining high performance is the next frontier in ILI modeling. To do so will require a fuller consideration of the "ILI data generating process." This process non-exhaustively includes a disease transmission process(es) (e.g., things often modeled with a compartmental model(s)), a healthcare visitation process (i.e., a set of processes related to who interacts with the healthcare system and when), an ILINet participation process (i.e., certain providers participate in ILINet while others do not, and the composition of provider networks varies temporally and spatially), and a reporting process (e.g., backfill).

Further stratification is a promising direction for incorporating known facets of the ILI data generating process into Dante in a flexible way. For example, if provider-level ILINet data were available state-level models could be decomposed into models for emergency department (ED) ILI and non-ED ILI. We hypothesize that a systematic difference exists between patients visiting ED and non-ED providers, specifically that the proportion of ED patients with ILI is higher than that of non-ED patients. If so, then provider composition could help explain some state-level variation in ILI magnitude (we expect states with more ED

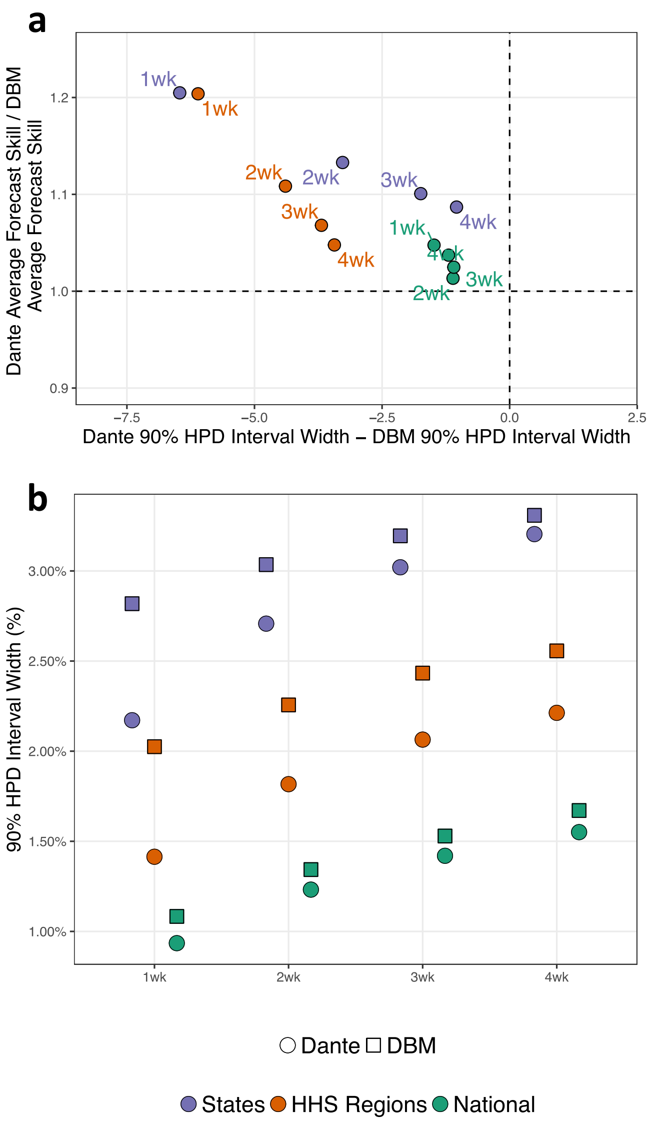

Fig. 6 Dante's and DBM's short-term forecast skill and sharpness by geographic scale. a Ratio of average forecast skills versus difference in 90% highest posterior density (HPD) interval widths for short-term forecasting targets. For all short-term forecasting targets and geographic scales, Dante produced sharper (i.e., smaller 90% HPD interval widths) and higher scoring forecasts than DBM. **b** Average 90% highest posterior density (HPD) interval widths for the short-term forecasting targets. Both Dante and DBM produce sharper (i.e., smaller 90% HPD interval widths) forecasts for coarser geographic resolutions. For all short-term forecasting targets and geographic scales, Dante produced sharper forecasts than DBM.

providers have higher reported ILI). It could also explain part of the holiday-specific spikes in observed ILI (we expect that spikes in ILI activity on holidays are partially due to the provider composition in ILINet changes for that week—more clinics are closed and thus ED providers have a relatively higher contribution).

Modeling across geographic scales in a single, unified model ensures forecasts are simultaneously coherent, a feature that is not present in many FluSight submissions. Ongoing work by our team will provide a model-agnostic tool by which users can modify outputs from a non-unified model so as to attain coherency[30]. Our team is also working to incorporate internet data sources (i.e., nowcasting) into future iterations of Dante. When internet data sources were incorporated into DBM the performance increased, which leads us to be hopeful that Dante will also be improved by the thoughtful incorporation of internet data.

While Dante has utility forecasting seasonal influenza, it would not be particularly useful for forecasting in an emerging outbreak setting. At its core, Dante is learning exploitable structure from data of historical flu seasons. An emerging outbreak, on the other hand, would by its very nature not have sufficient training data from which to learn about expected baseline trajectories. Swapping out the common process model term (μ_t^{all} – see the Methods section) for something resembling the infectious compartment of an SIR model might be one way to steer Dante towards emerging outbreak settings, but Dante's core strength is its ability to exploit historical data.

Methods

Dante is a probabilistic, Bayesian flu forecasting model that is decomposed into two submodels: the fine-scale model (i.e., the state model) and the aggregation model (i.e., the regional and national model).

Dante's fine-scale model. Dante's fine-scale model is itself described in two parts: the data model and the process model.

Dante's data model. Let $y_{rst} \in (0, 1)$ be ILI/100 for state $r = 1, 2, \ldots, R$ in flu season $s = 1, 2, \ldots, S$, for epidemic week $t = 1, 2, \ldots, T = 35$, where $t = 1$ corresponds to epidemic week 40, roughly the first week of October and $T = 35$ most often corresponds to late May. Dante models the observed proportion y_{rst} with a Beta distribution as follows:

$$y_{rst}|\theta_{rst}, \lambda_r \sim \text{ Beta}\left(\lambda_r\theta_{rst}, \lambda_r(1-\theta_{rst})\right), \tag{1}$$

where

$$\text{E}\left(y_{rst}|\theta_{rst}, \lambda_r\right) = \theta_{rst}, \tag{2}$$

$$\text{Var}\left(y_{rst}|\theta_{rst}, \lambda_r\right) = \frac{\theta_{rst}(1-\theta_{rst})}{1+\lambda_r}. \tag{3}$$

In Dante, θ_{rst} is the unobserved true proportion of visits for ILI in state r for season s during week t and $\lambda_r > 0$ is a state-specific parameter that captures the level of noise in the ILINet surveillance system and thus the level of volatility in the ILI time series. In Dante, y_{rst} is modeled as unbiased for the latent state θ_{rst}. The observation y_{rst}, however, is not equal to θ_{rst} due to variability in the measurement surveillance process (i.e., the true proportion of ILI in state r for season s during week t is not going to be perfectly captured by ILINet surveillance). Motivated by Fig. 3, λ_r is likely to be related to ILINet participation as measured by the total number of patients seen weekly in state r. As λ_r increases, the variance of y_{rst} decreases and observations will tend to be closer to θ_{rst}. Because we do not know the relationship between patient count and λ_r a priori, we model λ_r hierarchically, allowing them to be learned from data (details in Supplementary Note 1.1). Specifically, each λ_r is given a central, non-standardized t-distribution prior with support in the interval $[0, \infty)$ and a shared precision parameter that itself has a weakly informative Gamma prior.

Figure 8 shows the posterior mean of λ_r versus the average number of patients seen weekly by each state. A clear linear relationship is observed on a log-log scale, where the variance of y_{rst} goes down (i.e., λ_r increases) as the total weekly seen patients increases. What is particularly striking about Fig. 8 is that Dante has no knowledge of the number of patients seen each week as it is not an input to Dante, illustrating how structure can be learned rather than prescribed with a flexible, hierarchical model, provided sufficient training data.

Dante's Process Model. Dante's process model models the unobserved true proportion of ILI, $\theta_{rst} \in (0, 1)$, as a function of four components:

$$\theta_{rst} = \text{logit}^{-1}(\pi_{rst}), \tag{4}$$

$$\pi_{rst} = \mu_t^{all} + \mu_{rt}^{state} + \mu_{st}^{season} + \mu_{rst}^{interaction}. \tag{5}$$

The two season independent terms in Eq. (5), μ_t^{all} and μ_{rt}^{state}, are modeled as random walks, and the two season dependent terms, μ_{st}^{season} and $\mu_{rst}^{interaction}$, are modeled as reverse-random walks. Random and reverse-random walks allow patterns in the process model to be flexibly learned while capturing week-to-week correlation. Standard random walk priors are used for μ_t^{all} and μ_{rt}^{state}, with the latter being specified hierarchically. This specification involves placing a mean-0 normal prior on the first week of the flu season, and assuming a priori that subsequent weeks are normally distributed and centered at the previous week's value. The reverse random walk term μ_{st}^{season} requires a prior specification for the value in the

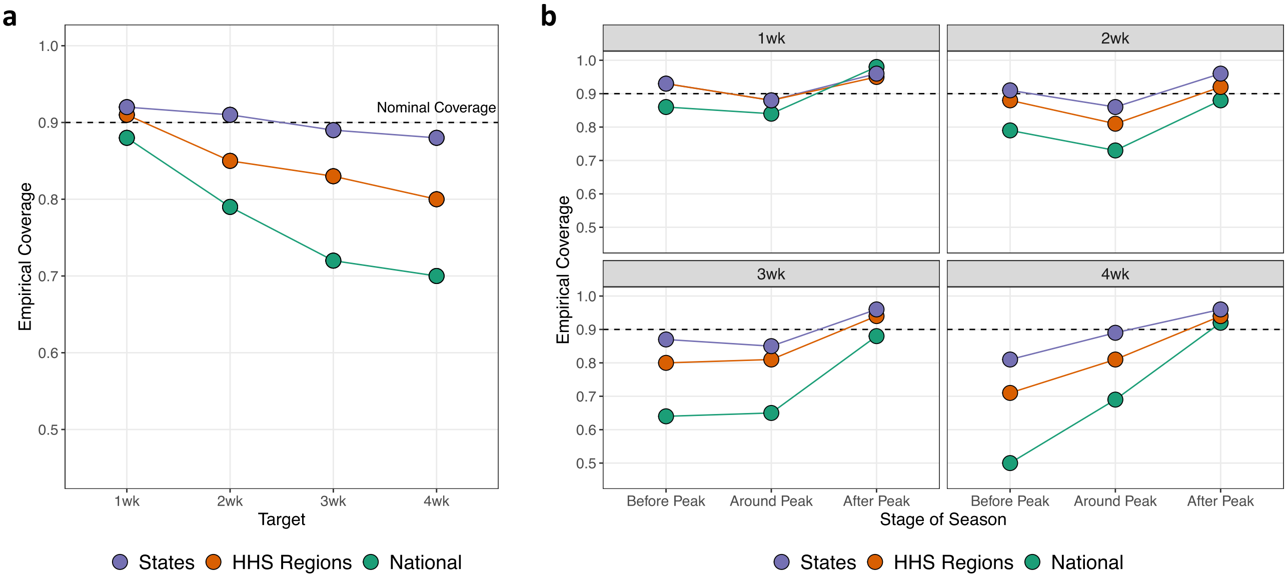

Fig. 7 Dante's empirical coverage. Dante's 90% empirical coverages for short-term targets, broken down by geographic scales. **b** Empirical coverages by target averaged over all seasons, geographic units within scale, and stages of flu season. **a** Empirical coverages broken out by stages of flu season. The "Around Peak" stage is defined as the peak week, plus/minus 2 weeks inclusively. Generally, empirical coverages degrade as the forecast window increases, as geographic scales coarsen, and as we get earlier in the flu season. The disagreement between empirical and nominal coverage for Dante represents an opportunity for iteration and improvement.

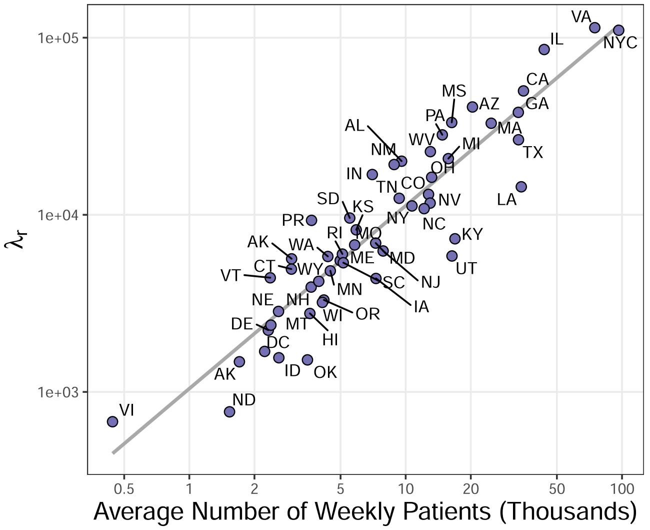

Fig. 8 Dante's posterior means for λ_r. The posterior mean for λ_r versus the average number of patients seen weekly by each state. Both axes are on a log scale. A clear linear relationship is observed. Dante learns this relationship, as it has no explicit knowledge of the average number of patients.

final week of the flu season rather than the first week of the season, which allows the forecasting extrapolation problem to become an interpolation problem. The state-season specific term $\mu_{rst}^{\text{interaction}}$ is the most challenging for Dante to learn because it is only directly informed by state r and season s, and for forecasting purposes there is either no or partial information available from season s. For that reason, an autoregressive term is included in the hierarchical reverse random walk prior specification to help regularize $\mu_{rst}^{\text{interaction}}$ towards 0. Prior choice is an aspect of the model which could still be improved, and it is possible that alternative prior specifications may yield improved performance. More details are available in Supplementary Note 1.2.

Figure 9 illustrates the fits for all model components for Alabama and Iowa for the 2015 and 2017 flu seasons. The component μ_t^{all} is common to every state and season and acts as the anchor for the process model. The shape of μ_t^{all} is similar to the national average ILI trajectory of Fig. 1, capturing the profile for a typical state and season. The component μ_{rt}^{state} captures the state-specific deviation from μ_t^{all} and is common to every season for a given state, but is distinct for each state. As can be seen in Fig. 1, Alabama typically sees ILI above the national average, hence why μ_{rt}^{state} for Alabama is learned to be greater than zero. Iowa, however, typically sees ILI below the national average, explaining why μ_{rt}^{state} is learned to be less than zero for Iowa. The component μ_{st}^{season} captures the season-specific deviation from μ_t^{all} and is common to every state for a given season, but is distinct for each season. This component captures the fact that seasons can have effects that are shared by nearly all states, as illustrated in Fig. 2. The shape of μ_{st}^{season} for 2015 and 2017 has a similar shape to the average residuals for 2015 and 2017, respectively, in Fig. 2. Finally, $\mu_{rst}^{\text{interaction}}$ captures the remaining signal in π_{rst} that cannot be accounted for by μ_t^{all}, μ_{rt}^{state}, and μ_{st}^{season}. The term $\mu_{rst}^{\text{interaction}}$ is distinct for each state and season.

Dante's process model is purposely over-specified. If our interest were purely to fit ILI data, the term $\mu_{rst}^{\text{interaction}}$ alone would suffice. However, there is not enough structure to forecast effectively with only $\mu_{rst}^{\text{interaction}}$. On the other hand, the non-interaction terms in the decomposition of π_{rst} (μ_t^{all}, μ_{rt}^{state}, and μ_{st}^{season}) provide structure for forecasting but not enough flexibility to capture all the signal in the ILI data. Thus, the $\mu_{rst}^{\text{interaction}}$ term provides the flexibility needed to fit the data, but is specified so that it plays as minimal a role as possible so that signal is captured in the non-interaction terms and can drive the shape of forecasts.

Inference for unobserved components of Dante, as well as state-level forecasts of yet-to-be-observed y_{rst} are generated by sampling from the posterior distribution with Markov chain Monte Carlo (MCMC), resulting in a sample of M draws that summarize the posterior distribution (details in Supplementary Note 2). We use the software JAGS (Just Another Gibbs Sampler)[31], as called by the `R` package `rjags`[32] within the programming language `R`[33] to perform the MCMC sampling. We denote each MCMC draw by the index m. Notationally, we denote the mth sample for a yet-to-be-observed y_{rst} as y_{rstm}.

Dante's aggregation model. Dante's regional and national forecasts are computed as linear combinations of state forecasts, where weights are proportional to 2010 US Census population estimates (Supplementary Fig. 6). Let $w_r \in [0, 1]$ be the population of state r divided by the US population such that $\sum_{r=1}^{R} w_r = 1$. For each MCMC draw m, we compute the ILI forecast for aggregated region ρ (indexing all ten HHS regions and nationally) as:

$$y_{\rho stm} = \sum_{r=1}^{R} w_r^{(\rho)} y_{rstm}. \tag{6}$$

For ρ = region X (e.g., ρ = HHS Region 1), $w_r^{(\rho)} = 0$ if state r is not a member of region X and

$$w_r^{(\rho)} = \frac{w_r \, \mathrm{I}(\mathrm{r} \in \text{ regionX})}{\sum_{r=1}^{R} w_r \, \mathrm{I}(\mathrm{r} \in \text{ regionX})} \tag{7}$$

and $\mathrm{I}(r \in \text{region X})$ is an indicator function equal to 1 if state r is in region X and 0 otherwise. By construction, $\sum_{r=1}^{R} w_r^{(\rho)} = 1$ for any ρ. The aggregation model constitutes a bottom-up forecasting procedure and ensures forecasts are coherent

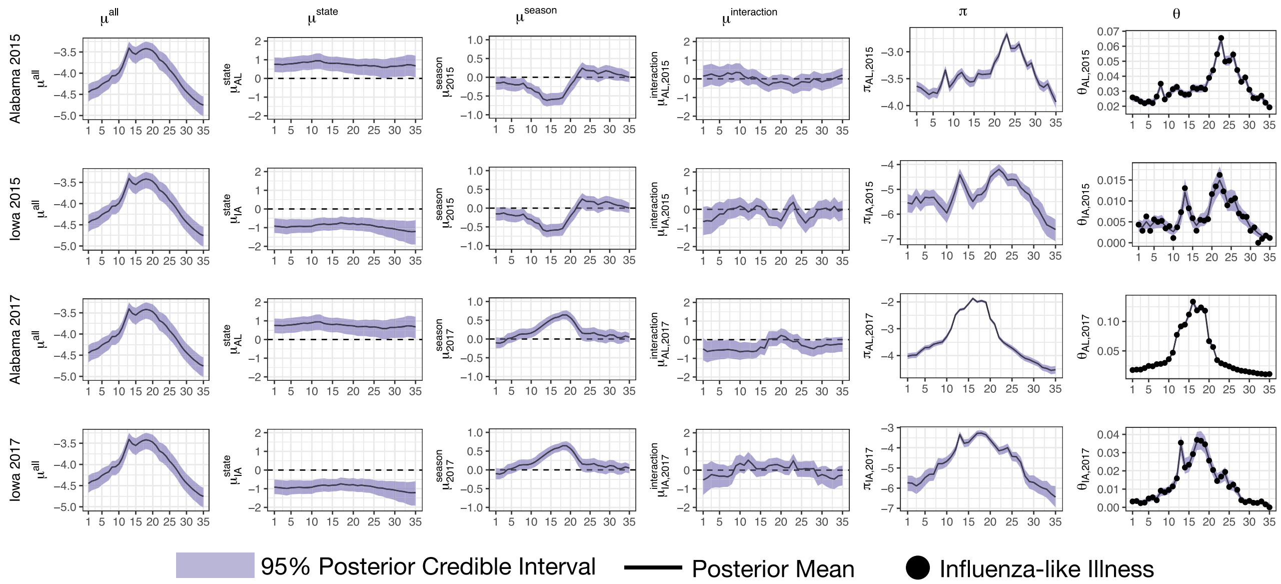

Fig. 9 Posterior summaries of Dante's process model. Posterior summaries for select components, seasons, and states of Dante, fit to seasons 2010 through 2017. Rows, from top to bottom, correspond to Alabama in 2015, Iowa in 2015, Alabama in 2017, and Iowa in 2017. Columns, from left to right, correspond to μ^{all}, μ^{state}, μ^{season}, $\mu^{\text{interaction}}$, π (all from Equation (5)), and θ (Equation (4)). The π column is the sum of the μ^{all}, μ^{state}, μ^{season}, and $\mu^{\text{interaction}}$ columns, accounting for posterior covariances. The θ column is the inverse logit of the π column and is back on the scale of the data. The μ^{all} component is the most structured component, as it is common for all states and seasons (i.e., the same for all rows). The components μ^{state} and μ^{season} are the next most structured components. They describe the state-specific and season-specific deviations from μ^{all}, respectively, and are common for all seasons within a state (μ^{state}) and all states within a season (μ^{season}). The component $\mu^{\text{interaction}}$ is the least structured component of Dante, as it is specific to each season/state (i.e., it is different for each row). Solid lines are posterior means. Ribbons are 95% posterior intervals. In the θ column, points are data, y.

across scales. An example of resulting aggregated forecasts can be found in Supplementary Note 7.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Up-to-date national, regional, and state ILI data are available from the CDC's FluView Interactive application: https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html. Select "ILINet" as the data source, "National," "HHS Regions," and "State" as the regions, and "Select All" for the seasons.

Code availability

The code to run Dante has been provided in the Supplementary Information document. A minimum working example of the code can be found here: https://github.com/lanl/Dante.

Received: 11 October 2019; Accepted: 16 April 2021;
Published online: 20 May 2021

References

1. Disease Burden of Influenza. The Centers for Disease Control and Prevention, 2019. https://www.cdc.gov/flu/about/burden/index.html. Accessed on 6-13-2019.
2. Biggerstaff, M. et al. Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge. *BMC Infec. Dis.* **16**, 357 (2016).
3. Biggerstaff, M. et al. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics* **24**, 26–33 (2018).
4. McGowan, C. J. et al. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Sci. Rep.* **9**, 683 (2019).
5. Pei, S., Kandula, S., Yang, W. & Shaman, J. Forecasting the spatial transmission of influenza in the United States. *Proc. Natl Acad. Sci. USA* **115**, 2752–2757 (2018).
6. Kandula, S. et al. Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. *J. R. Soc. Interface* **15**, 20180174 (2018).
7. Pei, S. & Shaman, J. Counteracting structural errors in ensemble forecast of influenza outbreaks. *Nat. Commun.* **8**, 925 (2017).
8. Kandula, S. & Shaman, J. Reappraising the utility of Google Flu Trends. *PLoS Comput. Biol.* **15**, e1007258 (2019).
9. Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J. & Rosenfeld, R. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Comput. Biol.* **11**, e1004382 (2015).
10. Farrow, D. C. et al. A human judgment approach to epidemiological forecasting. *PLoS Comput. Biol.* **13**, e1005248 (2017).
11. Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J. & Rosenfeld, R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Comput. Biol.* **14**, e1006134 (2018).
12. Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D. & DelValle, S. Y. Forecasting seasonal influenza with a state-space SIR model. *Ann. Appl. Stat.* **11**, 202 (2017).
13. Osthus, D., Gattiker, J., Priedhorsky, R. & Del Valle, S. Y. Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). *Bayesian Anal.* **14**, 261–312 (2019).
14. Osthus, D., Daughton, A. R. & Priedhorsky, R. Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLoS Comput. Biol.* **15**, e1006599 (2019).
15. Hickmann, K. S. et al. Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Comput. Biol.* **11**, e1004239 (2015).
16. Ray, E. L., Sakrejda, K., Lauer, S. A., Johansson, M. A. & Reich, N. G. Infectious disease prediction with kernel conditional density estimation. *Stat. Med.* **36**, 4908–4929 (2017).
17. Ray, E. L. & Reich, N. G. Prediction of infectious disease epidemics via weighted density ensembles. *PLoS Comput. Biol.* **14**, e1005910 (2018).
18. Reich, N. G. et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl Acad. Sci. USA* **116**, 3146–3154 (2019).
19. Ben-Nun, M., Riley, P., Turtle, J., Bacon, D. P. & Riley, S. Forecasting national and regional influenza-like illness for the USA. *PLoS Comput. Biol.* **15**, e1007013 (2019).
20. Yang, S., Santillana, M. & Kou, S. C. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc. Natl Acad. Sci. USA* **112**, 14473–14478 (2015).
21. Lu, F. S., Hattab, M. W., Clemente, C. L., Biggerstaff, M. & Santillana, M. Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. *Nat. Commun.* **10**, 147 (2019).

22. Davidson, M. W., Haim, D. A. & Radin, J. M. Using networks to combine 'big data' and traditional surveillance to improve influenza predictions. *Sci. Rep.* **5**, 8154 (2015).
23. Riebler, A., Held, L. & Rue, H. Modelling seasonal patterns in longitudinal profiles with correlated circular random walks. In *Proc. 26th International Workshop on Statistical Modelling* 503–508. https://d1wqtxts1xzle7.cloudfront.net/39223399/iwsm2011_proceedings.pdf?1444978493=&response-content-disposition=inline%3B+filename%3DProceedings_of_the_26th_International_Wo.pdf&Expires=1619834812&Signature=YDH846Be0amPGITCALn9TccD8CHaoodTy3RKxkWMSbOo0KAHkjPAriaeW-cN6bXT1wyZ8OG5xDR29mWu4ojnaqbTfwrGY5FyvJpPqQ2pOjzgpAhTJ4q3gFsTd3FuV35rYevzDGWWJTR0J9UC3CI2fUv1GlvIevlwYPY8KdpKKwuv9no8MdwUOlGoYnKOZCc0nkuAm1gENZqhgQ09AHQQHq-2mX9mJ8aKjyVr5Jannt0m9nU9dwdr8GyIkWj96VwDB9cKE~dkrFOK9FbwKRt6-Cnjam2l7sVUoC5O5qKTwmiazEFfxm~qk2ouDd6rwHQ9pGpAreLcAf8wvIHmYL8e4A__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA#page=518 (2011).
24. Spencer, S. E. F. et al. The spatial and temporal determinants of campylobacteriosis notifications in New Zealand, 2001–2007. *Epidemiol. Infect.* **140**, 1663–1677 (2012).
25. Held, L., Meyer, S. & Bracher, J. Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Stat. Med.* **36**, 3443–3460 (2017).
26. Bracher, J. On the multibin logarithmic score used in the FluSight competitions. *Proc. Natl Acad. Sci. USA* **116**, 20809–20810 (2019).
27. Reich, N. G. et al. Reply to Bracher: scoring probabilistic forecasts to maximize public health interpretability. *Proc. Natl Acad. Sci. USA* **16**, 20811–20812 (2019).
28. Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
29. Moran, K. R. et al. Epidemic forecasting is messier than weather forecasting: the role of human behavior and internet data streams in epidemic forecast. *J Inf. Dis.* **214**, S404–S408 (2016).
30. Gibson, G. C., Moran, K. R., Reich, N. G. & Osthus, D. Improving probabilistic infectious disease forecasting through coherence. *PLOS Comput. Biol.* **17**, e1007623 (2021).
31. Plummer, M. Jags: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc. 3rd international workshop on distributed statistical computing*. 124 (2003).
32. Plummer, M. *rjags: Bayesian Graphical Models using MCMC*. R package version 4–8. (2018).
33. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2018).

Acknowledgements

This work was funded by the Department of Energy at Los Alamos National Laboratory under contract 89233218CNA000001 through the Laboratory-Directed Research and Development Program, specifically LANL LDRD grant 20190546ECR. The authors thank C.C. Essix for her encouragement and support of this work, as well as the helpful conversations with the Delphi group at Carnegie Mellon University. Approved for unlimited release under LA-UR-19-28977.

Author contributions

D.O. and K.R.M. conceived the study, prepared the data for analysis, designed the computational framework, carried out the implementation of the methods, and wrote the manuscript. D.O. developed the methods and secured the funding for the work. Both authors contributed to and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23234-5.

Correspondence and requests for materials should be addressed to D.O.

Peer review information *Nature Communications* thanks Leonhard Held and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.