eLIFE elifesciences.org

RESEARCH ARTICLE

Figures and figure supplements

Conjunction of factors triggering waves of seasonal influenza

Ishanu Chattopadhyay *et al*

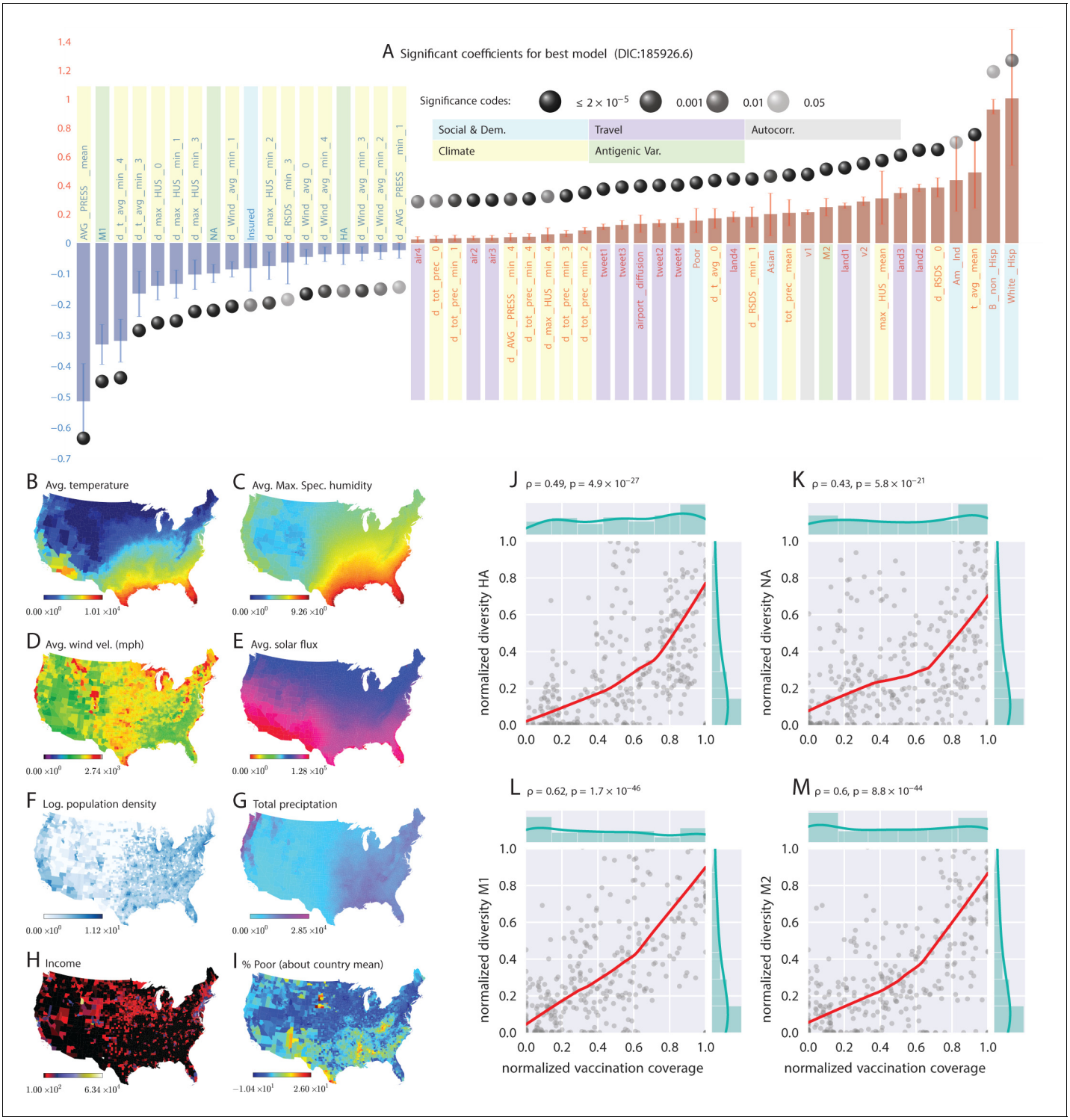


Figure 1. Putative determinants of seasonal influenza onset in the continental US and Poisson mixed-effect regression analysis (Approach 2). Plate A shows the significant variables along with their computed influence coefficients from the mixed-effect Poisson regression analysis (the best model chosen from 126 different regression equations with different variable combinations). The statistically significant estimates of fixed effects are grouped into several classes: climate variables, economic and demographic variables, auto-regression variables, variables related to travel, and those related to antigenic diversity (see the last entry in ***Table 5*** for the detailed regression equation used. The complete list of all models considered is given in Table S-D7). The fixed-effect regression coefficients plotted in Plate A are shown on a logarithmic scale, meaning that the absolute magnitude of predictor-specific effect is obtained by exponentiating the parameter value. A negative coefficient for a predictor variable suggests that the influenza rate falls as this factor increases, while a positive coefficient predicts a growing rate of infection as the parameter value grows. The integrated influence of

Figure 1 continued on next page

Figure 1 continued

individual predictors, under this model, is additive with respect to the county-specific rate of infection. For example, a coefficient of -0.6 for parameter AVG_PRESS_mean tells us that the average atmospheric pressure has a negative association with the influenza rate. As the mean atmospheric pressure for the county grows, the probability that the county would participate in an infection initiation wave falls. As $\exp(-0.6) = 0.54$, the rate of infection drops by 46% when atmospheric pressure increases by one unit of zero-centered and standard-deviation-normalized atmospheric pressure. Similarly, an increase in the share of a white Hispanic population predicts an increase in influenza rate: A coefficient of 1.3 translates into a $\exp(1.3) \times 100\% - 100\% = 267\%$ rate increase, possibly, because of the higher social network connectivity associated with this segment of population. Plates B - I enumerate the average spatial distribution of a few key significant factors considered in Poisson regression: (**B**) average temperature; (**C**) average maximum specific humidity; (**D**) average wind velocity in miles per hour; (**E**) average solar flux; (**F**) logarithm of population density (people per square mile); (**G**) total precipitation; (**H**) income, and; (**I**) percent of poor as deviations about the country average. Plates J-M show the strong dependence between our estimated antigenic diversity (normalized, see Definition in text) corresponding to the HA, NA, M1, and M2 viral proteins, and the cumulative fraction of the inoculated population (normalized between 0 and 1), where both sets of variables are geo-spatially and temporally stratified. Pearson's correlation tests shown in Plates J-M were performed under null hypothesis that there the two quantities (plotted along axes X and Y) are statistically independent ($H_0 : \rho = 0$).

DOI: https://doi.org/10.7554/eLife.30756.003

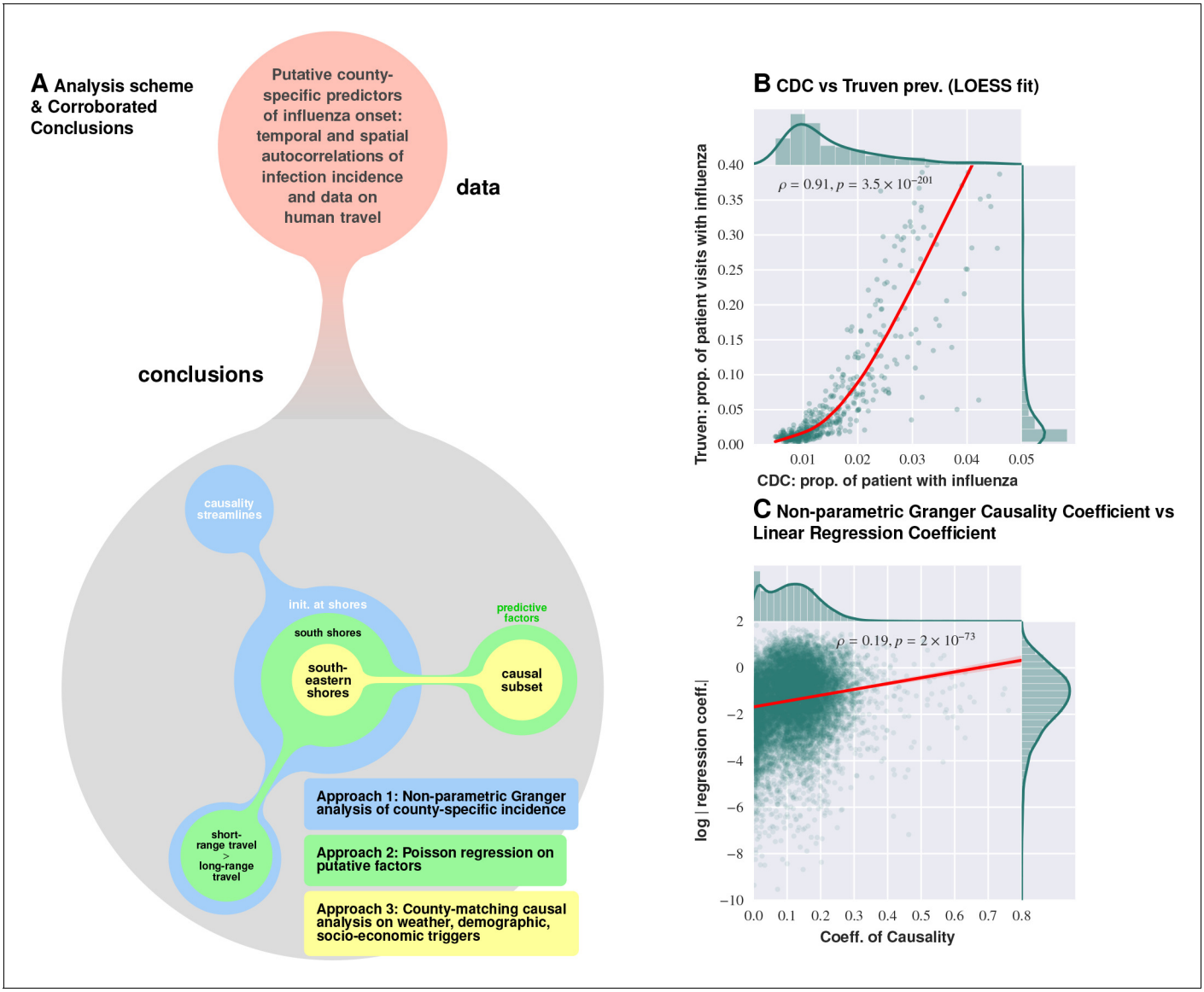


Figure 1—figure supplement 1. Logical flow and cross-corroboration of conclusions. Plate A: Diverse data sets processed via multiple techniques to reach convergent, and reinforcing, conclusions. Approach 1 (the Granger-causality analysis) shows that the epidemic tends to begin near water bodies, and that short-range travel is more influential compared to air travel for propagation. Approach 2 (Poisson regression) identifies significant predictive factors, suggesting that the epidemic begins near the southern shores of the US, and corroborates the result on short- vs. long-range travel. Approach 3 (county-matching) points to south eastern shores of the continental US as where the epidemic initiates, and identifies a validated subset of predictive factors. Plate B shows that influenza prevalence as reported by Truven dataset positively correlates with CDC reports. Plate C illustrates that our conclusions regarding a putatively causal influence between neighboring counties, inferred using different techniques (mixed-effect regression vs. non-parametric Granger-causality), match up positively. Pearson's correlation tests shown in Plates B and C are performed under null hypothesis that there the two types of quantities (plotted along axes X and Y) are statistically independent ($H_0 : \rho = 0$).

DOI: https://doi.org/10.7554/eLife.30756.004

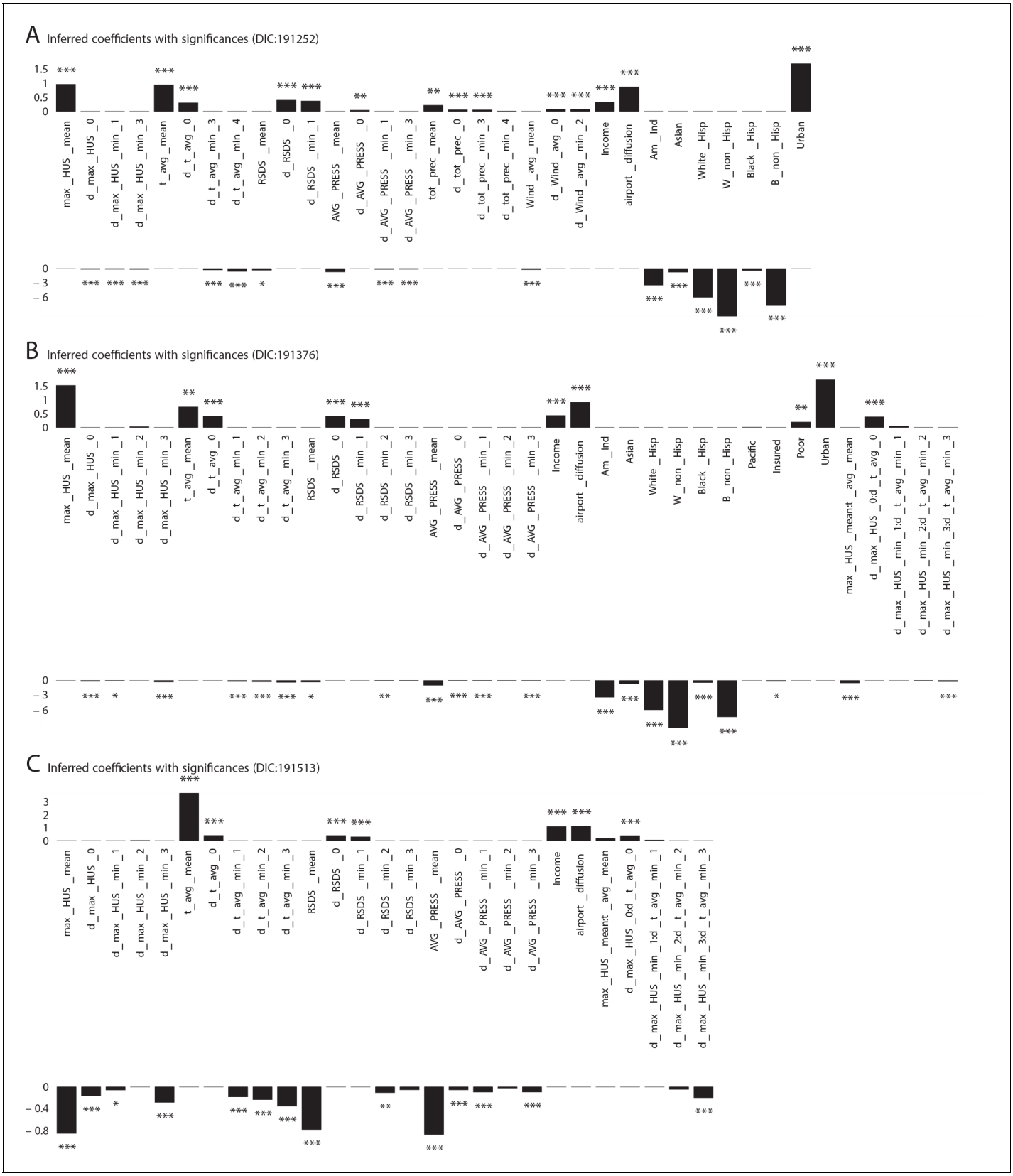


Figure 1—figure supplement 2. Significant influencing variables obtained with mixed effect regression with different models as tabulated in ***Table 1*** of main text (three more models with DIC larger than that of the best model shown in ***Figure 1*** plate A).

DOI: https://doi.org/10.7554/eLife.30756.005

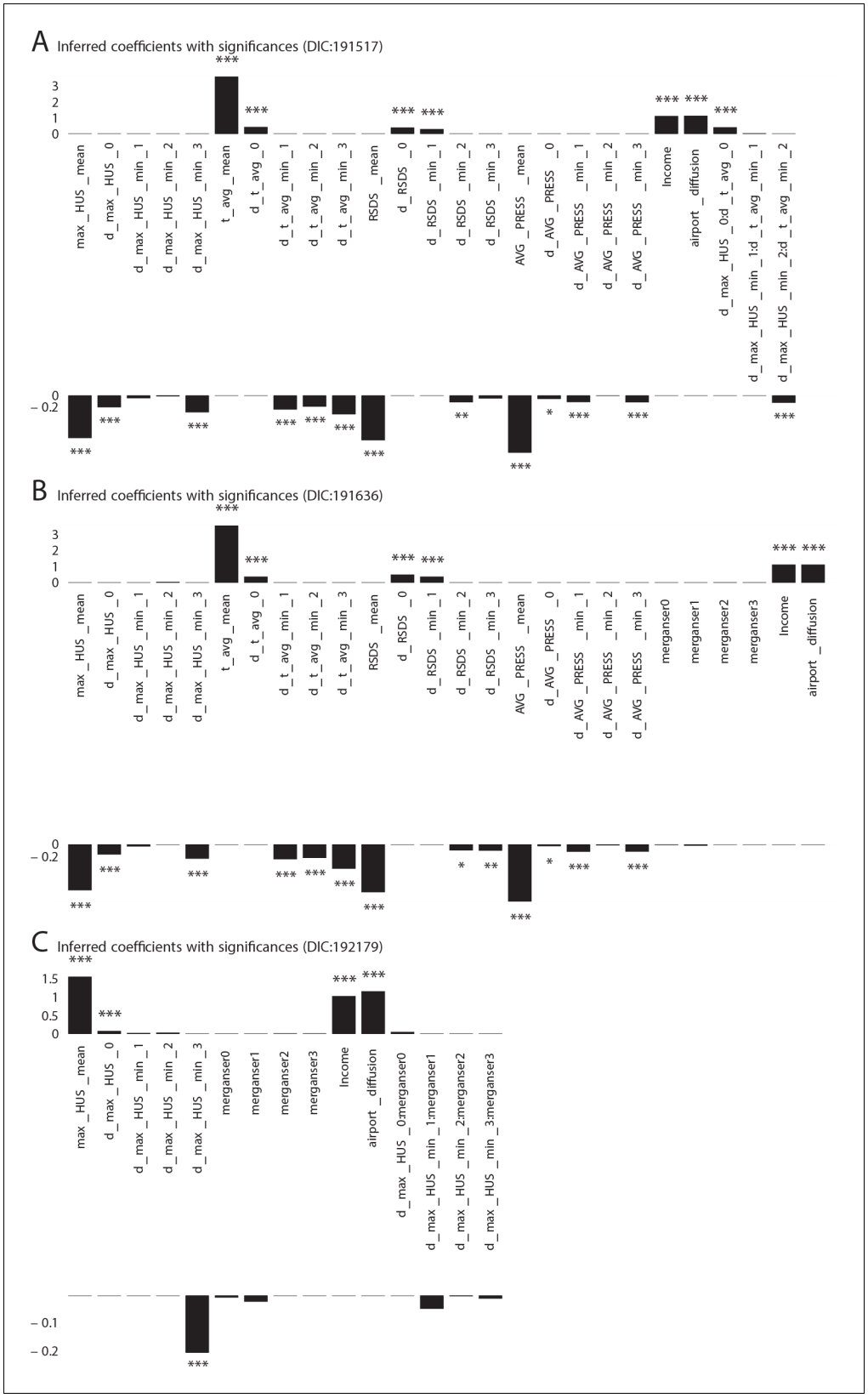


Figure 1—figure supplement 3. Additional Cases: Significant influencing variables obtained with mixed effect regression with different models as tabulated in ***Table 1*** of main text (three more models with DIC larger than that of the best model shown in ***Figure 1*** plate A).

DOI: https://doi.org/10.7554/eLife.30756.006

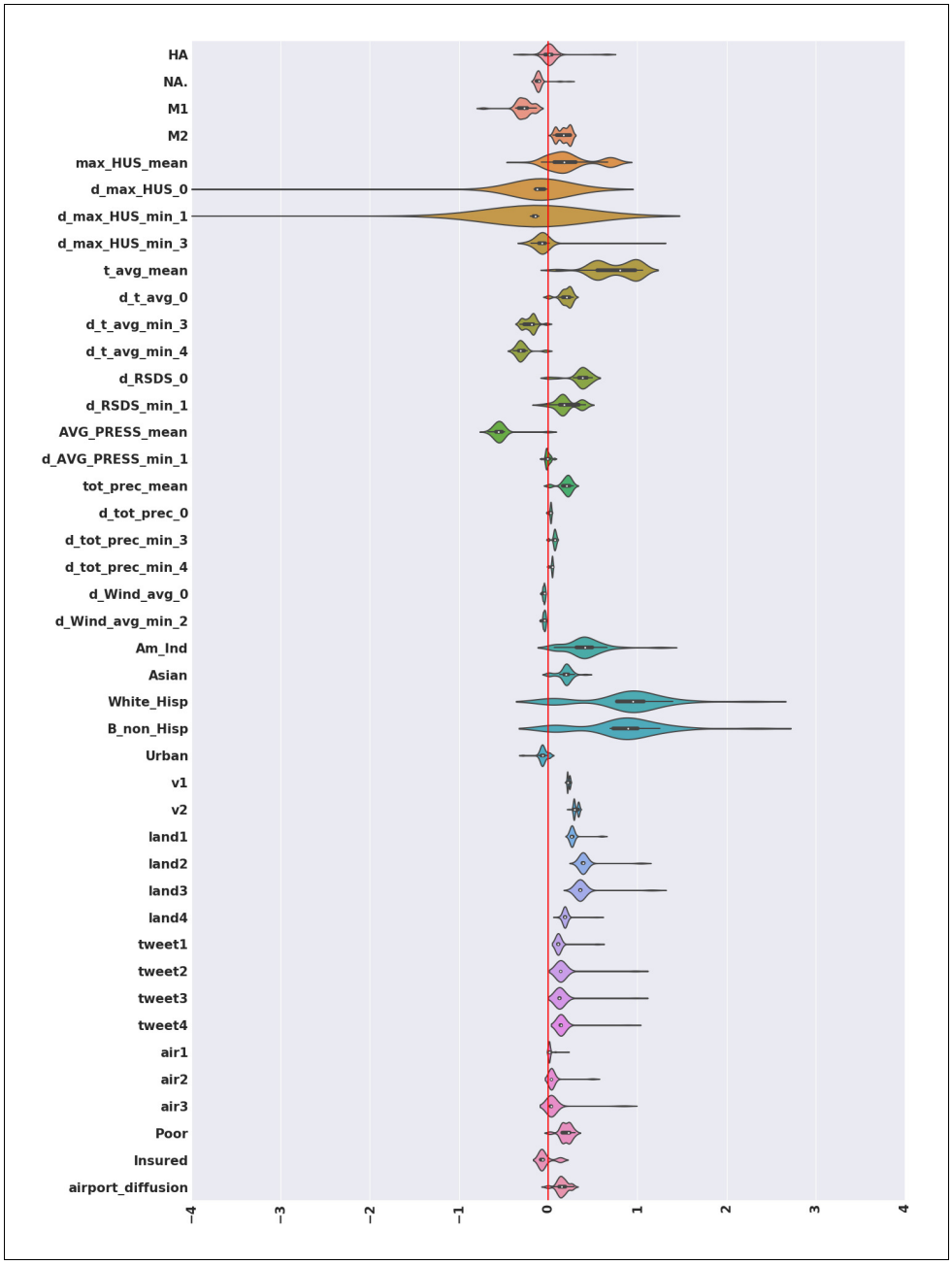


Figure 1—figure supplement 4. Violin plots for the coefficients inferred for variables that turn out to be significant in the best model, computed considering the complete set of models we investigated. We note that with the exception of the antigenic variation of the surface protein hemagglutinin (HA), and some derivatives of maximum absolute humidity, significant mass of the individual violin plots fall either entirely on the positive or entirely on the negative half-plane; implying that the significant factors rarely flip sign. Thus, while the coefficients inferred change as we explore different models, the direction of influence remains mostly unchanged.

DOI: https://doi.org/10.7554/eLife.30756.007

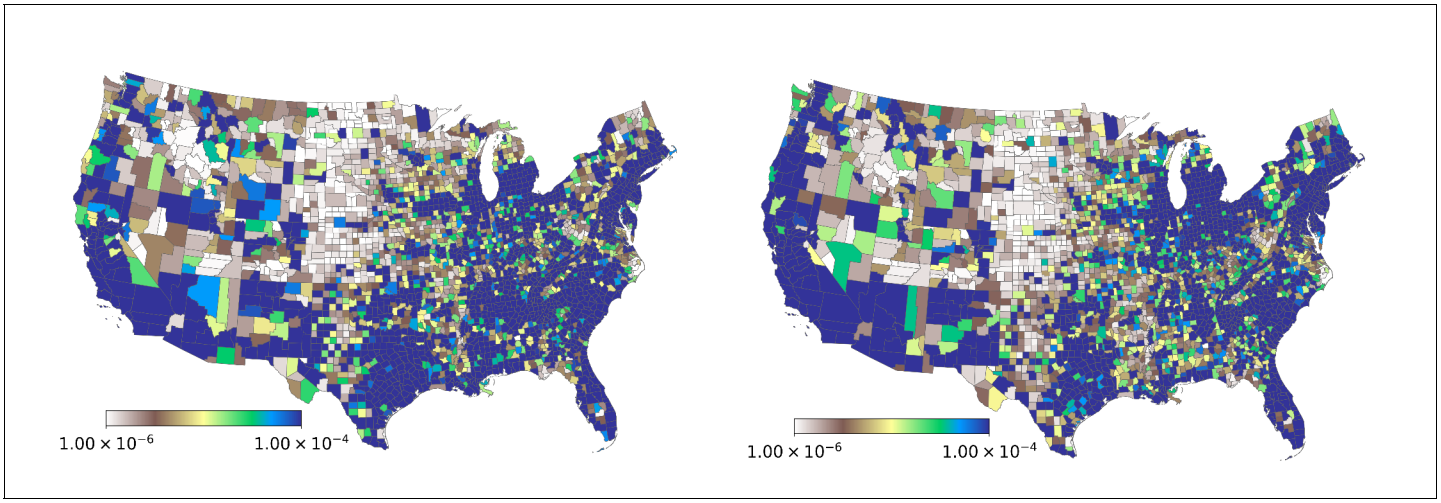


Figure 1—figure supplement 5. Spatial variation in the probability of patient visits corresponding to any ICD9-CM code (plate on left), and for diagnoses corresponding to influenza-like diseases (plate on right).
DOI: https://doi.org/10.7554/eLife.30756.008

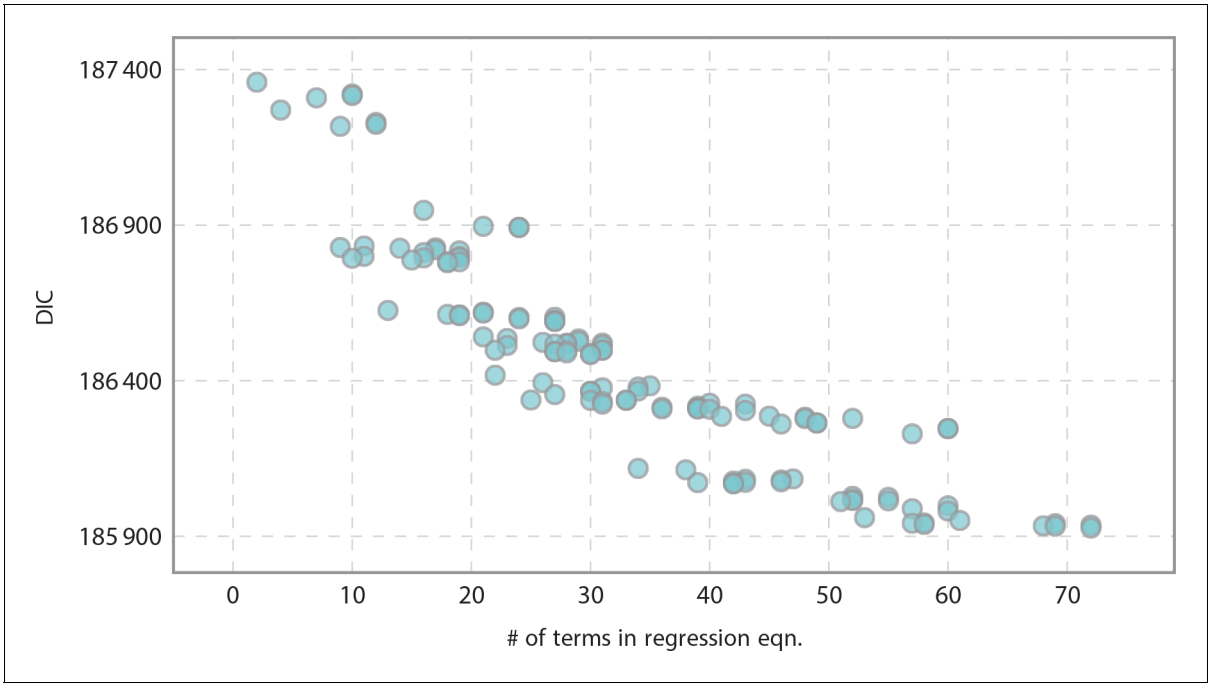


Figure 1—figure supplement 6. Informativeness of model vs model complexity as related to the number of terms in the regression equation. As expected, we yielded more informative models as we increased complexity.
DOI: https://doi.org/10.7554/eLife.30756.009

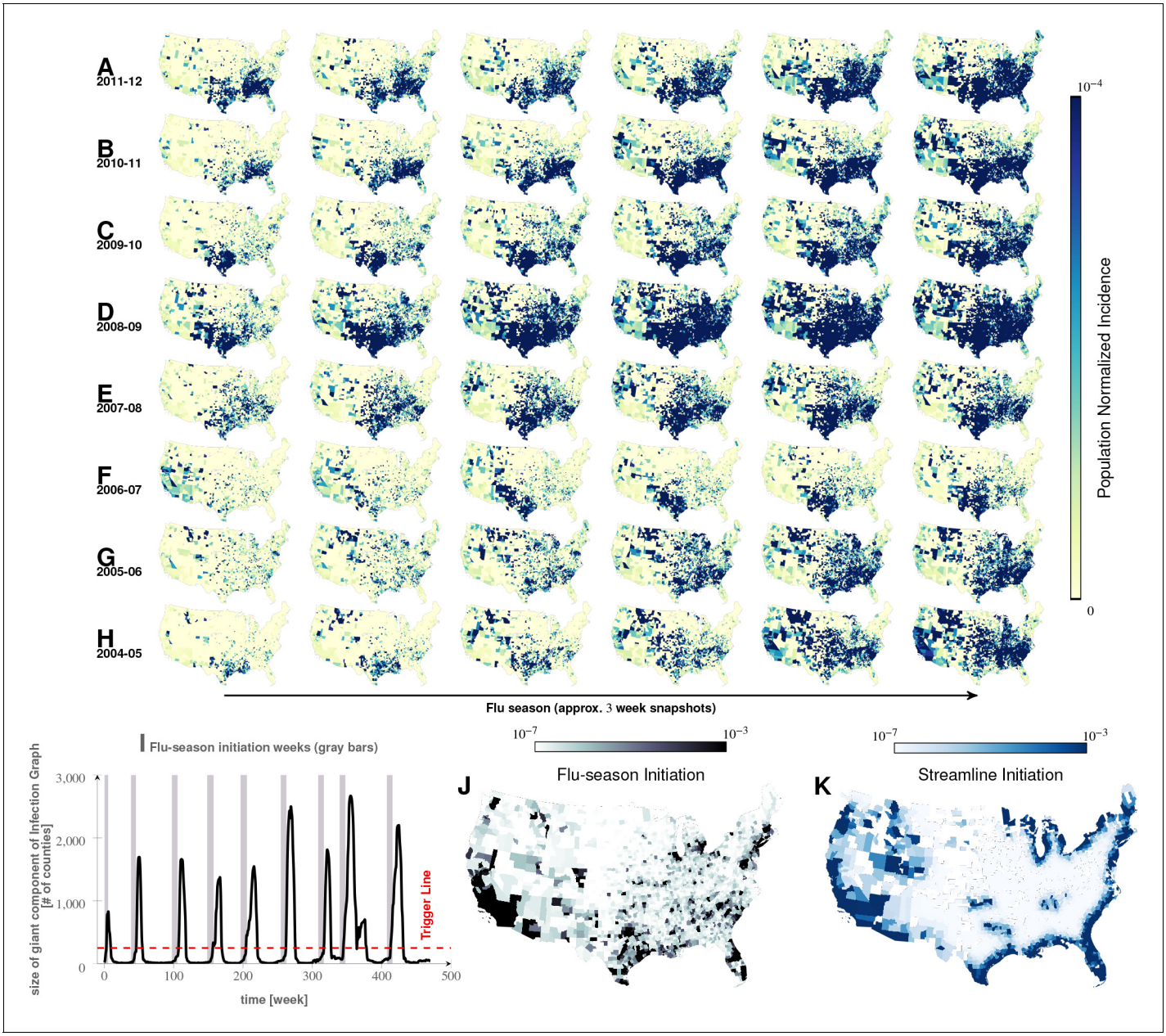


Figure 2. Characteristics of seasonal influenza in the continental US An analysis of county-specific, weekly reports on the number of influenza cases for a period of 471 weeks spanning January 2003 to December 2013 (Plates A-H) for recurrent patterns of disease propagation. In particular, the weeks leading up to that in which an epidemic season peaks (determined by significant infection reports from the maximum number of counties for that season) demonstrate an apparent flow of disease from south to north, which cannot be explained by population density alone (also see movie in Supplement). Plate I illustrates the near-perfect time table for a seasonal epidemic. Plates J and K compare the county-specific initiation probabilities of an influenza season, and the causality streamlines.

DOI: https://doi.org/10.7554/eLife.30756.010

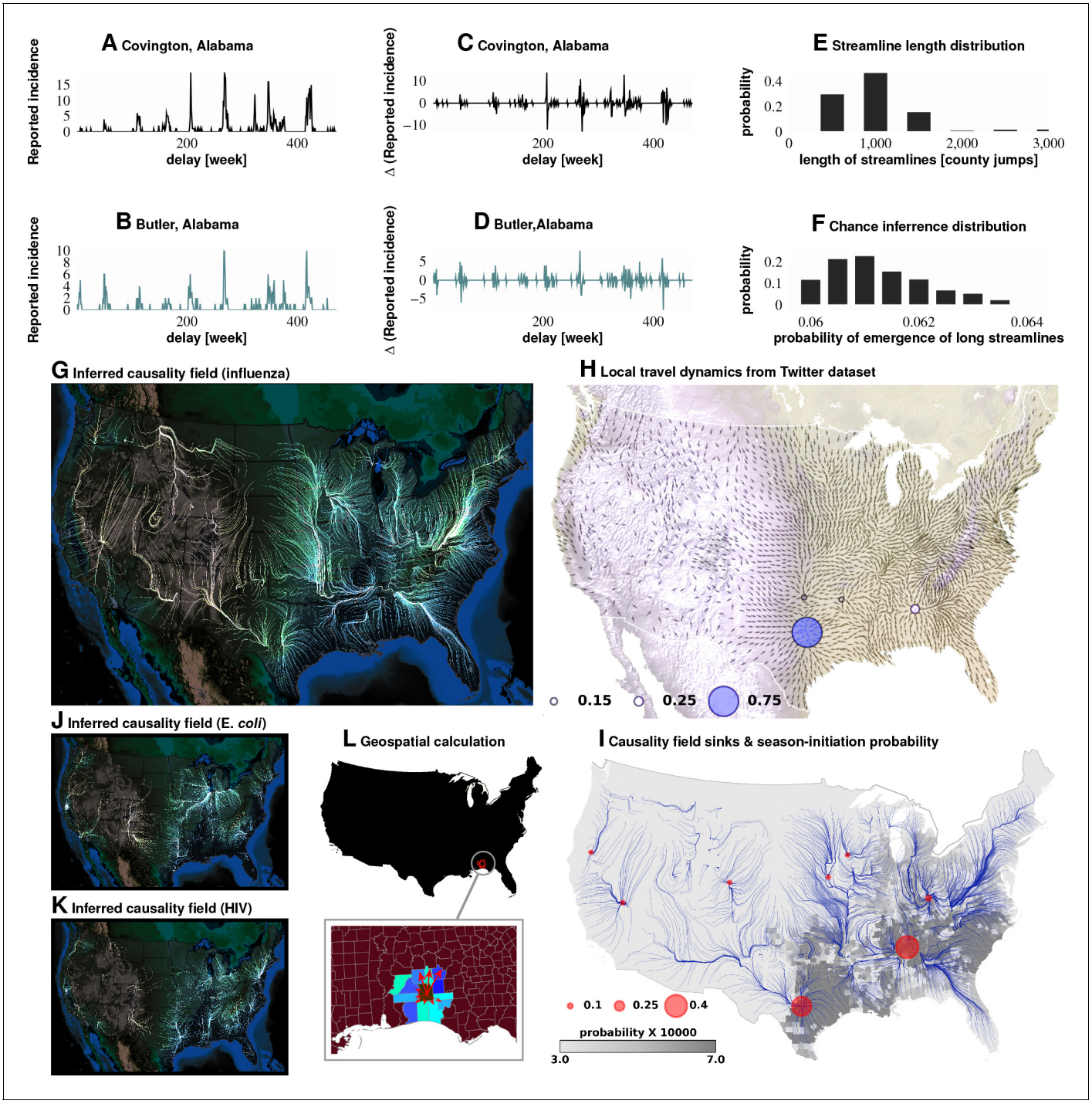


Figure 3. Computation of causality field, Approach 1 Plates A and B: Incidence data from neighboring counties in Alabama, US. Plates C and D: Transformation to difference-series, *i.e.*, change in the number of reported cases between weeks. We imposed a binary quantization, with positive changes mapping to '1,' and negative changes mapping to '0.' From a pair of such symbol streams, we computed the direction-specific coefficients of Granger causality (see Supplement). For each county, we obtained a coefficient for each of its neighbors, which captured the degree of influence flowing outward to its respective neighbors (Plate L). We computed the expected outgoing influence by considering these coefficients as representative of the vector lengths from the centroid of the originating county to centroids of its neighbors. Viewed across the continental US, we then observed the emergence of clearly discernible paths outlining the 'causality field' (Plate G). The long streamlines shown are highly significant, with the probability of chance occurrence due to accidental alignment of component stitched vectors less than 10^{-185}; while each individual relationship has a chance occurrence probability of ~6% (Plates E and F). Plate H: Spatially-averaged travel patterns (see text in Materials and methods) and the sink distribution between expected travel patterns. These patterns (Plate H), along with the inferred causality field (Plate I), match up closely, with sinks

Figure 3 continued on next page

Figure 3 continued

showing up largely in the Southern US, explaining the central role played there. In Plate H, the size of the blue circles indicate the percentage of movement streamlines (computed by interpreting the locally averaged movement directions as a vector field) that sink to those locations. In Plate I, the size of the red circles indicate the percentage of causality streamlines that sink to the indicated locations. We note that $\sim 75\%$ of the movement streamlines sink in counties belonging to the Southern states, which matches up well with the sinks of the causality streamlines. In Plates J and K show spatial analysis results for two different infections (HIV and *E. coli*, respectively) and which exhibit very different causality fields, negating the possibility of boundary effects.

DOI: https://doi.org/10.7554/eLife.30756.015

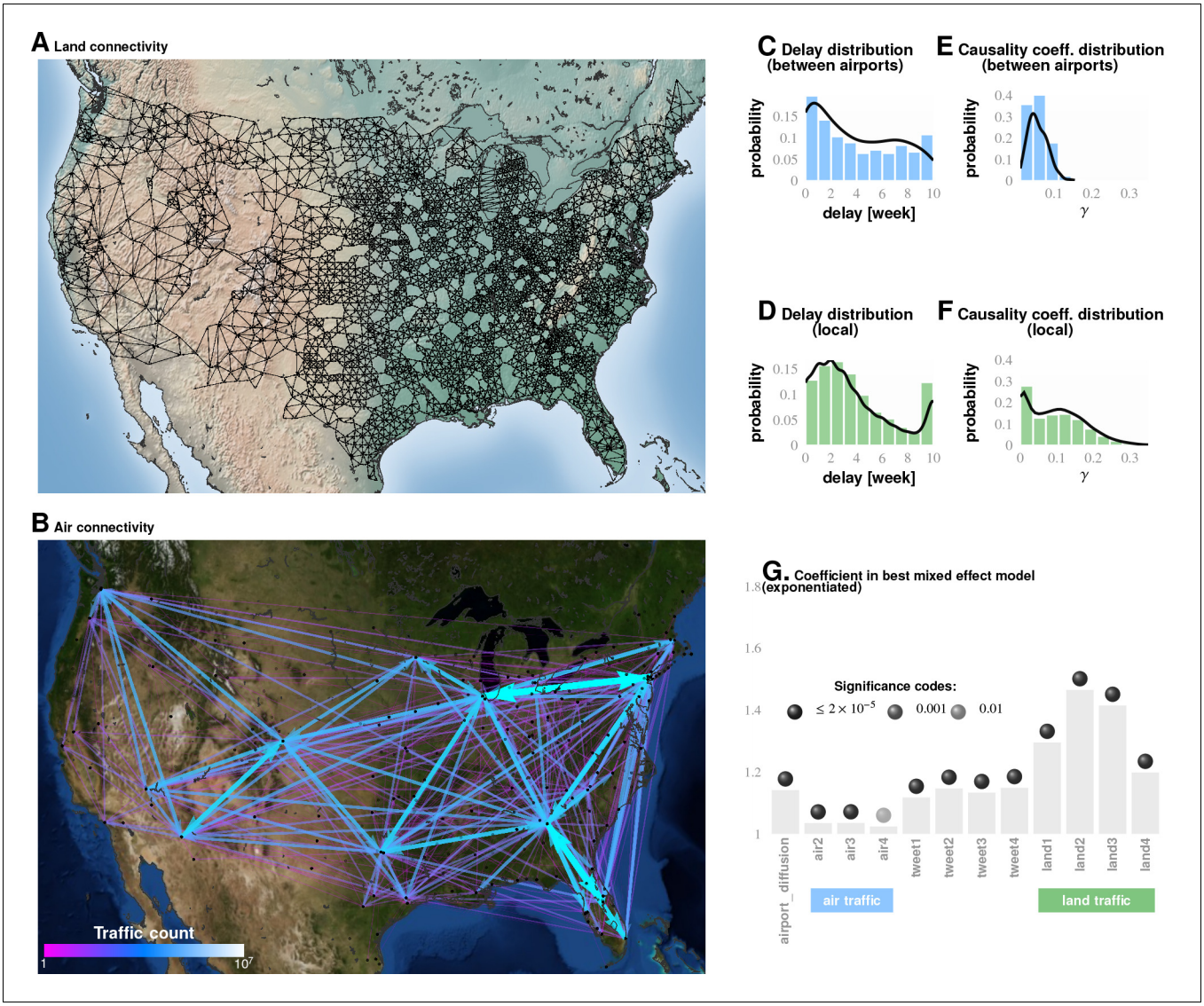


Figure 4. Comparing influence of short- and long-distance travel on infection propagation Plate A shows land connectivity visualized as a graph with edges between neighboring counties. Plate B shows air connectivity as links between airports, with edge thickness proportional to traffic volume. Plate C shows the delay in weeks for the propagation of Granger-causal influence between counties in which major airports are located, and Plate E shows the distribution of the inferred causality coefficient between those same counties. Plates D and F show the delay and the causality coefficient distribution respectively, which we computed by considering spatially neighboring counties. The results show that local connectivity is more important. We reached a similar conclusion using mixed-effect Poisson regression, as shown in Plate G: The inferred coefficients for land connectivity are significantly larger than those for air connectivity, tweet-based connectivity, or exponential diffusion from the top 30 largest airports. The coefficients shown in Plate G are exponentiated, allowing us to visualize probability magnitudes (see 'Model Definition').

DOI: https://doi.org/10.7554/eLife.30756.016

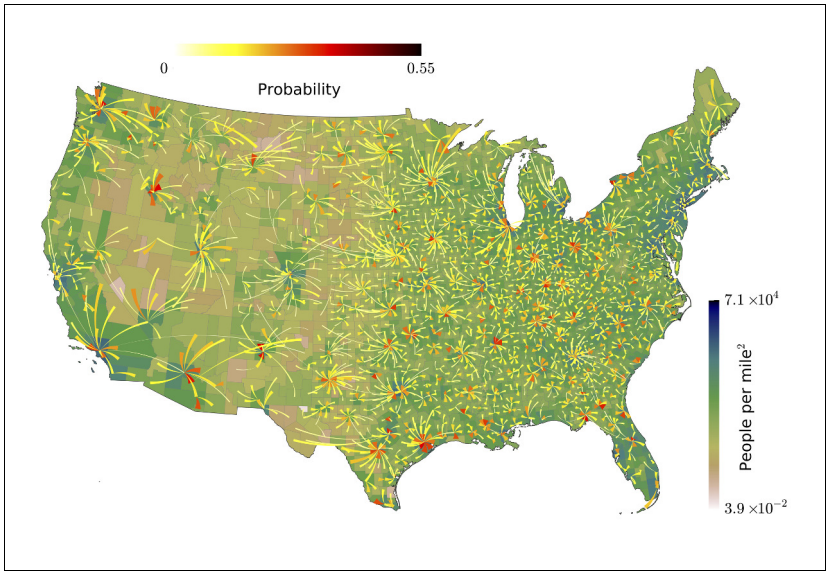


Figure 4—figure supplement 1. Our analysis of the Twitter movement matrix indicates that people most frequently travel between neighboring counties, preferentially towards higher-population-density areas, which shows that the maximum-probability movement patterns follow the local gradient of increasing population density.
DOI: https://doi.org/10.7554/eLife.30756.017

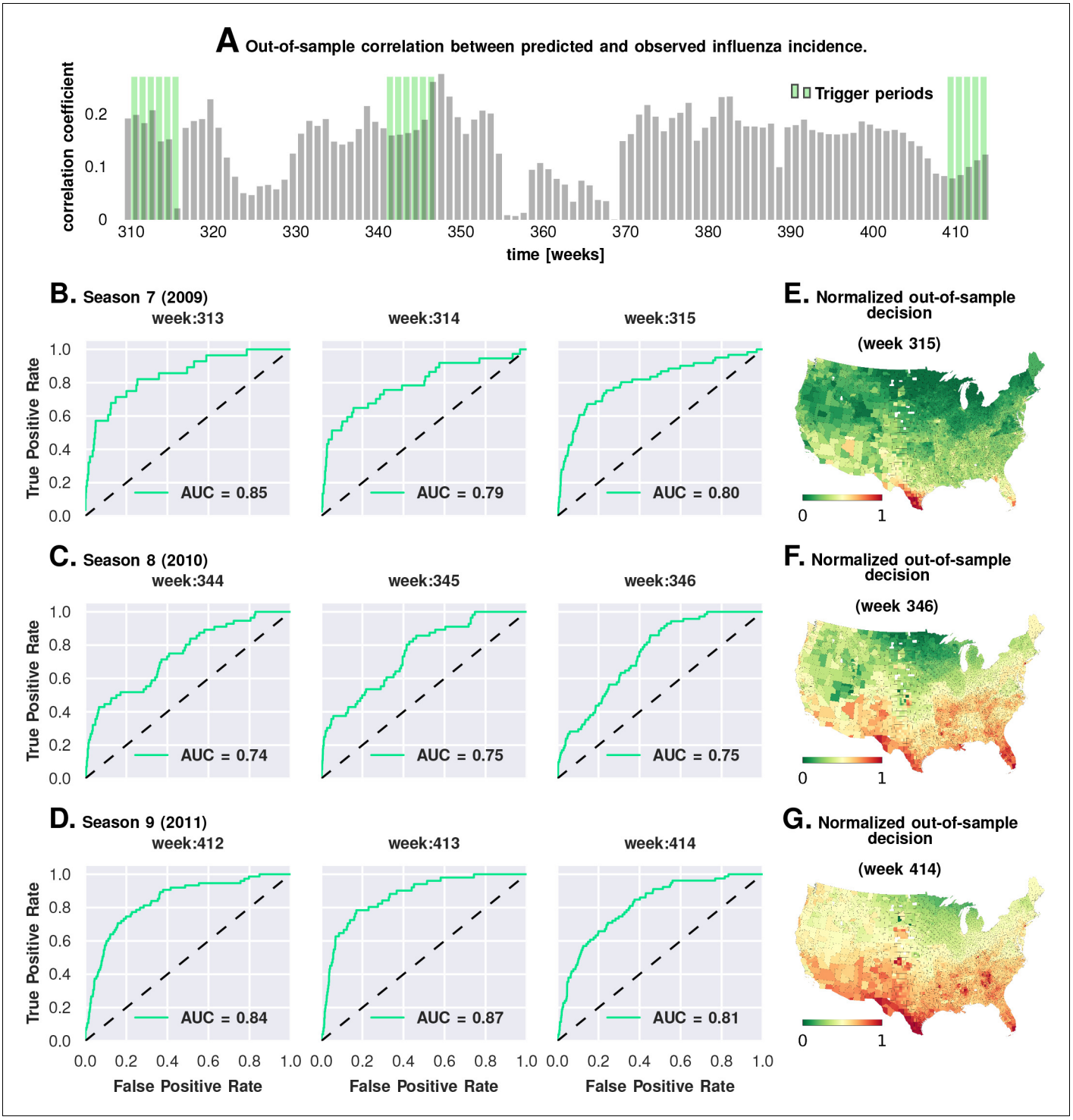


Figure 5. Prediction performance with training data from the first six seasons and validation on the last three. Plate A shows the correlation between the observed incidence and the model-predicted response. We show significant positive correlation, particularly within the trigger periods, between the model predictions and the actual held-out data. This gives us confidence to construct ROC curves for each week. Plates B-D show the ROC curves for the last three weeks of each of the three seasons in the out-of-sample period (potentially, these computations can be repeated for all possible partitions of study weeks into training and test samples). Plates E-G illustrate that the normalized decision variable, which is the normalized response from the model, identifies the South and Southeastern counties as the trigger zones.

DOI: https://doi.org/10.7554/eLife.30756.018

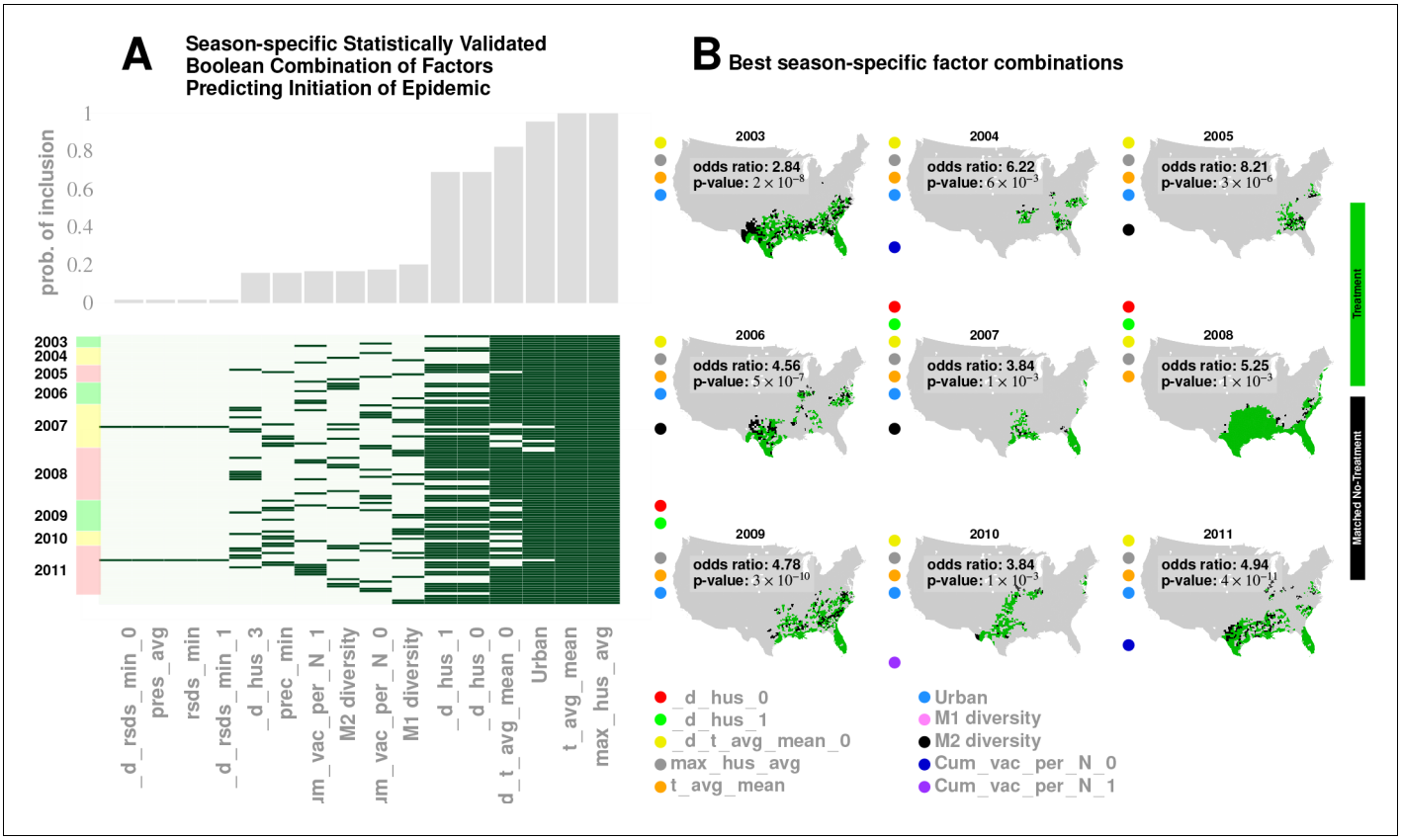
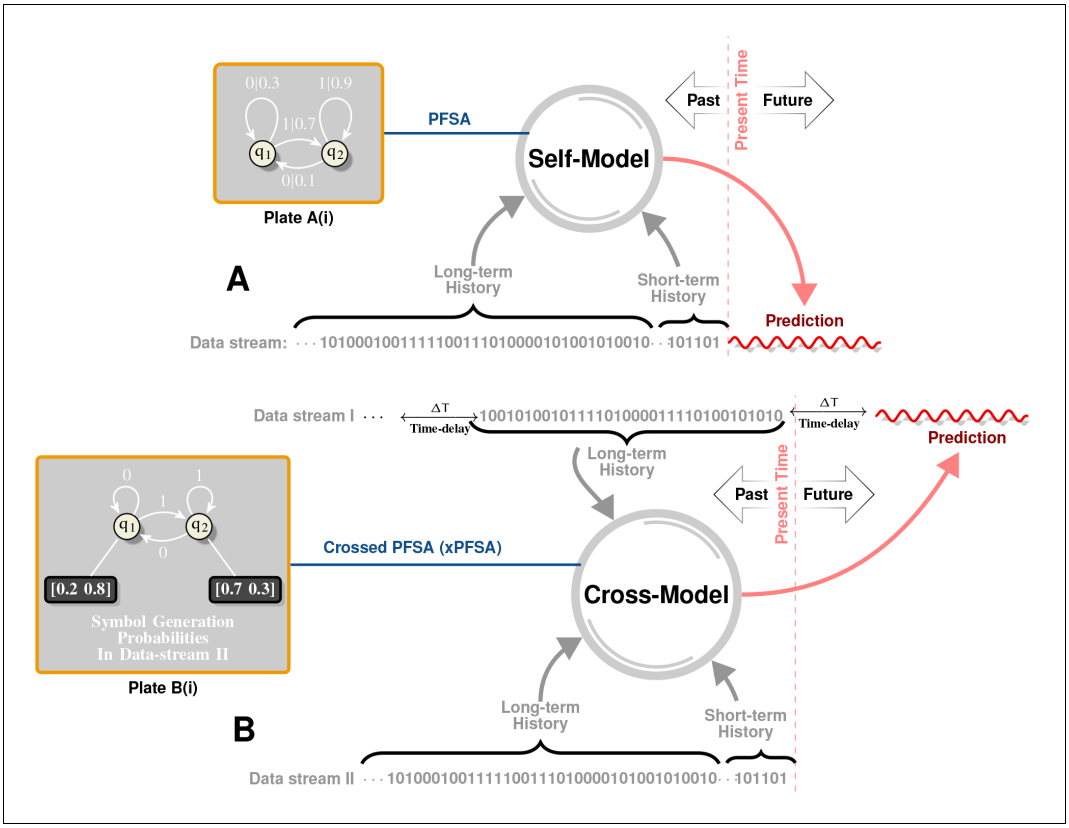


Figure 6. Results for our analysis involving county-matching (Approach 3). Plate A illustrates the factor combinations that turn out to be significant over the nine seasons. Notably, for each season, we have multiple, distinct factor sets that turn out to be significant ($p<0.05$) and yield a greater-than-unity odd ratio. Plotting the probability with which different factors are selected when we look at season-specific county matchings (the top panel in Plate A), we see a corroboration of the conclusions drawn in Approach 2. We find that specific humidity and average temperature, along with their variations, are almost always included. We do see some new factors that fail to be significant in the regression analysis, *e.g.*, degree of urbanity and vaccination coverage. While vaccination coverage is indeed included as a factor in our best performing model, in Approach two it failed to achieve significance, perhaps due to its strong dependence on antigenic variation (see ***Figure 1J–M***). Degree of urbanity is indeed significant for some of the regression models we considered (see Supplementary Information), but was not significant for the model with the smallest DIC. Note that 'Treatment' here is defined as a logical combination of weather factors. A treatment is typically a conjunction of several weather variables. For example, the treatment shown in top left panel of Plate B involves a conjunction of: (1) a drop in average temperature during the week of infection; (2) a drop in temperature during the week of infection; (3) a higher-than-average specific humidity; (4) a higher-than-average temperature, and; (5) a high degree of urbanity. With respect to the 'treatment,' we can divide counties into three groups: (1) 'treated counties,' shown in green; (2) at least one matching county for each of the treated counties (matching counties are very close to the treated counties in all aspects but in treatment, which we called 'control' counties), shown in black, and; (3) other counties, shown in grey. The counties in the 'treatment' and 'control' groups are further subdivided into those counties that initiated an influenza wave and those that have not, resulting in four counts arranged into a two-by-two contingency table. We then used the Fisher exact test to test for association between treatment and influenza onset. Panels in Plate B show both the treated and control sets for the 9 seasons for a subset of chosen factors. The results are significant, as shown in ***Tables 2*** and ***3***. The variable definitions are given in ***Table 4***. Notably, some of the variables found significant in the regression analysis are not included above, and some which are not found to be significant in the best regression model show up here. This is not to imply that they are not predictive or lack causal influence. The matched treatment approach, as described above, is not very effective if we use more than $\sim 10-15$ factors simultaneously to define the treated set (for the amount of data we have); this results in a contingency table populated with zero entries.

DOI: https://doi.org/10.7554/eLife.30756.019



Appendix 1—figure 1. Intuitive Description of Self and Cross Models. *Plate A* illustrates the notion of self-models. Historical data is first represented as a symbol sequence (denoted as 'Data stream' in Plate A) using space-time discretization and magnitude quantization. For example, we may use a spatial discretization of $\pm 3^{\circ}$ in both latitudes and longitudes, a temporal discretization of 1 week, and a binary magnitude quantization that maps all magnitudes below 4.0 to symbol 0, and all higher magnitudes to symbol 1. This symbol stream then represents a sample path from a hidden, quantized stochastic process. A self-model is a generative model of this data stream, which captures symbol patterns that causally determine (in a probabilistic sense) future symbols. Specifically, our inferred self-model (see Plate A(i) for an example) is a probabilistic, finite state automata (PFSA). *Plate B* illustrates the notion of cross-models. Instead of inferring a model from a given stream to predict future symbols in the same stream, we now have two symbol streams (Data Stream I and Data Stream II), and the cross-model is essentially a generative model that attempts to predict symbols in one stream by reading historical data in another. Notably, as shown in Plate B(i), the cross-model is syntactically not exactly a PFSA (arcs have no probabilities in the cross-model, but each state has an output distribution). We call such models "crossed probabilistic finite state automata,' or XPFSA. Once these models are inferred, they may be used to predict the future evolution of the data streams. Thus, the self-model in Plate A may be initialized with its unique stationary distribution, after which a relatively short observed history would dictate the current distribution on the model states. This, in turn, would yield a distribution over the symbol alphabet in the next time step. For a cross-model, we would be able to obtain future symbol distribution in the second stream, given a short history in the first stream. Note that the cross-model from I→ II is not necessarily the same as the cross-model in the other direction.

DOI: https://doi.org/10.7554/eLife.30756.033