A High-Resolution, US-scale Digital Similar of Interacting Livestock, Wild Birds, and Human Ecosystems for Multi-host Epidemic Spread

Abhijin Adiga[a,1,2], **Ayush Chopra**[b,1], **Mandy L. Wilson**[a,1], **S. S. Ravi**[a], **Dawen Xie**[a], **Samarth Swarup**[a], **Bryan Lewis**[a], **Andrew Warren**[a], **John Barnes**[d], **Ramesh Raskar**[b], **and Madhav V. Marathe**[a,c,2]

This manuscript was compiled on March 28, 2025

One Health issues, such as the spread of highly pathogenic avian influenza, present unique challenges at the human-animal-environmental interface. Ongoing H5N1 outbreaks underscore the urgent need for comprehensive modeling efforts that capture the complex interactions between various entities in these interconnected ecosystems. To support such efforts, we develop a methodology to construct a realistic spatiotemporal gridded digital similar of livestock production and processing, human population, and wild birds for the contiguous United States. Our approach involves multi-scale and multi-source data fusion and synthesis using statistical and optimization techniques, followed by extensive verification and validation. This framework, called FIELD, consists of multiple layers and sublayers. It includes farm-level representations of four major livestock types – cattle, poultry, swine, and sheep – with further categorization into subtypes such as dairy cows, beef cows, chickens, turkeys, etc. Abundance data for wild bird species identified in the transmission of avian influenza are included. Gridded distributions of the human population, with demographic and occupational features, capture the placement of agricultural workers and the general population. We apply FIELD to evaluate spillover risk to dairy cows and poultry from wild birds and validate these results using historical H5N1 incidences. The resulting subtype-specific spatiotemporal risk maps identify new hotspots with high risk and map the temporal variations in risk for each region, thus enabling prioritization of surveillance efforts. Our results indicate that some livestock-production-intensive regions exhibit persistent elevated risk across multiple hosts, with increased likelihood of cross-species spillover and zoonotic transmission.

Avian influenza | Digital twin | Epidemiological risk assessment | One Health | Spillover

Significance Statement

This work constructs a high-resolution, multi-layered spatiotemporal representation of the contiguous U.S., integrating livestock populations, processing centers, wild bird abundances, and human demographics using diverse datasets and advanced mathematical techniques. The framework assesses H5N1 spillover risk from wild birds to dairy cattle and poultry, identifying new high-risk areas. While risk varies over time, some livestock-intensive regions exhibit persistently elevated risk across multiple hosts, increasing the likelihood of cross-species spillover and zoonotic transmission. These findings can inform subtype-specific surveillance, resource allocation, and risk assessment for human populations. Additionally, potential future scenarios, such as widespread cattle infection, could significantly alter the risk landscape.

Highly pathogenic avian influenza (HPAI) poses a serious global threat to health, environment and food security. In the Americas alone, the unprecedented spread of H5N1 virus clade 2.3.4.4b has led to severe loss of wildlife (1–4). In the US, the incidence among wild birds is widespread. Large-scale outbreaks in poultry and dairy cattle threaten food production (5–8). There have also been several instances of zoonotic transmissions (9) through exposure to poultry and cattle, which poses a serious pandemic risk (10). Recent works analyzing this phenomenon in poultry and cattle (6, 11, 12) underscore the urgent need for a modeling platform to help researchers understand and respond to the spread of HPAI, accounting for the various agents that shape, and are affected by, this phenomenon.

In recent years, several national-scale, realistic in silico representations of populations, socioeconomic activities, and built infrastructures have been developed at fine spatiotemporal resolutions to study complex phenomena such as epidemiology, emergency response and food security (6, 13–17). Here, we refer to such synthetic datasets as *digital similars*. They have statistical similarity to real data, but differ from "digital twins" (18–21), which are intended as precise "living" replicas of the real-world systems they represent (22, 23). These realistic datasets constructed from diverse datasets and modeling are used for risk assessment and simulation modeling, as evidenced by studies conducted during the COVID-19 pandemic to analyze the dynamics of infectious diseases (24–29).

The first large-scale high-resolution digital similars of socio-technical systems were developed more than two decades ago (16). Subsequently, many products have been proposed to address problems in multiple domains, ranging from simple gridded distributions of populations with demographic attributes, to definitions of activities and interactions with built infrastructures (29, 30). Some of these digital similars focus on modeling the spatial distribution of livestock (6, 11, 13, 14, 31, 32). Of these previous efforts, those on US-scale data sets, including recent ones in the

Author affiliations: [a]Biocomplexity Institute, University of Virginia, Charlottesville, VA; [b] MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA; [c]Dept. of Computer Science, University of Virginia, Charlottesville, VA; [d]Office of Molecular Detection, CDC, Atlanta, GA

MM and AA initiated and outlined the work. AA, AC, MW, and SSR designed and implemented various components of the digital similar. AA, AC, MW, DX, BL, and AW explored and curated various datasets. MM, AA, AC, BL, AW, and SS designed risk models. All authors wrote the paper.

The authors declare no competing interests.

[1]AA, AC and MW contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: mmarathe@virginia.edu, abhijin@virginia.edu

context of HPAI, focus on a single livestock type. Key data challenges stem from the need to explore and fuse diverse, often sparse, data sets, which are misaligned in format and spatial resolution. Methodological challenges in this context include the choice of appropriate objectives, assumptions and constraints in the algorithmic formulations in order to achieve realistic representations that are statistically consistent with the parent data sets (for example, composition and distribution of livestock farms).

Summary of our contributions. This work presents a framework for high-resolution multi-layered spatiotemporal representation of the contiguous US, henceforth referred to as FIELD (Framework for In silico representation of Environment-Livestock-Demographics interface), that captures (i) the distribution of livestock populations and operations for multiple types (like cattle or poultry) and subtypes (like beef or milk cows, chickens, turkeys, etc.), (ii) associated food processing center locations, capacities, and functions, (iii) spatiotemporally-varying wild bird abundances for multiple species affected by H5N1, and (iv) human populations with demographic features and attributes capturing agricultural employment, as illustrated in Figure 1.

Methods and validation. We leverage diverse open datasets (listed in Table 1), and data gaps are addressed by using a combination of statistical tools and mathematical programming. Mapping livestock populations to farms and assigning them to grid cells are cast as optimization problems and solved using integer linear programs. We perform rigorous data quality checks with reference to the source data sets, and verification and validation studies using independent data sets, including known locations of large livestock farms and H5N1 incidence reports. Our work builds on earlier important work that focuses on single type of livestock (e.g., (6, 11, 13)) and extends it to multiple livestock types and subtypes on a national scale. The dataset is made open through an interactive web portal DiTTO (33).

Risk assessment. We demonstrate the utility of FIELD as a comprehensive platform for modeling and assessing the risk of HPAI-like outbreaks at high spatial and temporal resolutions, informing disease surveillance and control strategies. Using spatiotemporal risk maps, we evaluate the spillover risk from H5N1-infected wild birds to dairy cattle and poultry, identifying both known and emerging hotspots for spillover and zoonosis. Our results, validated against historical H5N1 incidences, highlight (i) new regions with persistently elevated risk that have not reported any incidences and (ii) temporal variations in risk at the county level. While risk levels differ by subtype, certain livestock-production-intensive regions exhibit persistent elevated risk across multiple hosts, increasing the likelihood of cross-species spillover and zoonotic transmission. These insights can guide subtype-specific surveillance, resource allocation prioritization, and risk to the human population. Additionally, we explore potential future scenarios, such as widespread cattle infection, which could significantly alter the risk landscape.

Results

The Digital Similar. FIELD provides a unified gridded representation of livestock production and processing operations, the human population, and wild bird populations in the contiguous US. Figure 1 provides a layered view along with a summary of population sizes. Table 1 provides an overview of the data sources used to construct it. Formally, FIELD$(V, \mathcal{L}, \mathcal{P}, \mathcal{B}, \mathcal{H})$ is defined over a grid V overlaid on the study region. Each grid cell $v \in V$ has attributes that capture the details of each of these components. In the current setting, we use a 5×5 arc square-minute grid. Descriptions of the components $\mathcal{L}$, $\mathcal{P}$, $\mathcal{B}$, and $\mathcal{H}$ are provided below.

Livestock $\mathcal{L}\big(\Theta_{\mathcal{L}}, \{\Gamma_\theta \mid \theta \in \Theta_{\mathcal{L}}\}, \{\mathcal{F}_\theta \mid \theta \in \Theta_{\mathcal{L}}\}\big)$. We develop a novel generic approach to construct the livestock layers from agricultural census and grid-level estimates of livestock populations. Figure 3a outlines the methods comprising of statistical methods and optimization techniques. The livestock population comprises four types of animals: $\Theta_{\mathcal{L}} = \{\texttt{cattle}, \texttt{poultry}, \texttt{hogs}, \texttt{sheep}\}$. For each type $\theta \in \Theta_{\mathcal{L}}$, Γ_θ denotes the set of different "subtypes" of animals. For example, $\Gamma_{\texttt{cattle}} = \{\texttt{beef}, \texttt{milk}, \texttt{other}\}$. The full list of subtypes is provided in the supplement. For each type of livestock, the population is partitioned into farms*. The collection of farms for each livestock type θ is denoted by $\mathcal{F}_\theta$. For each farm $f \in \mathcal{F}_\theta$, the population of each subtype γ, denoted by $H_{f\gamma}$, is specified. (We use H for head counts). Also specified is the grid cell v to which this farm is assigned. Figure 3a shows the sequence of steps performed to generate livestock representations from census and gridded data using statistical and optimization techniques.

Processing centers $\mathcal{P}$. This layer provides information regarding livestock-associated food processing centers such as meat processing, dairy processing, and poultry processing units. Each processing unit $p \in \mathcal{P}$ contains attributes such as the location of the unit, type of processing, and the size estimate.

Wild birds $\mathcal{B}\big(\Theta_{\mathcal{B}}, A(\cdot)\big)$. This component captures the spatiotemporal distribution of 39 species of birds identified as significant vectors of avian influenza as derived from eBird (year 2022) and H5N1 incidence data from 2022–2025 (see Table 1 for the data sources). Let $\Theta_{\mathcal{B}}$ denote the set of different species. The abundance of a species $\theta \in \Theta_{\mathcal{B}}$ in grid cell $v \in V$ at time t is denoted by $A(\theta, v, t)$. The data is available at a weekly resolution.

Human population $\mathcal{H}\big(\Delta, \mathcal{E}, \pi(\cdot)\big)$. This component provides a grid-level representation of the human population with emphasis on agricultural workers (workers associated with livestock and its processing). For each cell v, $\pi(v, \delta, \epsilon)$ denotes the size of the subpopulation that belongs to the demographic group $\delta \in \Delta$ defined by attributes including age group, sex, and employment in professional classes specified by $\epsilon \in \mathcal{E}$. An employment group ϵ is defined by occupation and industry attributes, where non-agricultural employees are all binned into one group, namely non-agriculture.

Risk Estimation. To demonstrate the utility of FIELD, we use it to study the risk of H5N1 spillover from wild birds to various livestock populations. To this end, we use the livestock and wild bird abundance layers to conduct simple colocation-based risk assessments at the county and state levels. Given a livestock subtype s, a grid cell i and time t,

*All livestock production operations will be referred to as farms.

Livestock	Heads (≈)	Farms (≈)
Cattle	88M*	732K*
- Beef	29M	622K
- Milk	9.3M	36K
Poultry		
- Layers	400M	240K
- Broilers	1.7B*	43K
- Pullets	144M	35K
- Turkeys	97M	23K
15 other subtypes		
Hogs	74M	61K
Sheep	5M	89K

*B is billion, M is million and K is thousand.

Wild birds:
39 species such as geese, mallards, crows, etc.

Human population:
≈ 300M with ≈ 2M workers in livestock production and processing.

Processing centers:
Poultry: 4978, slaughter: 2086, and dairy: 188

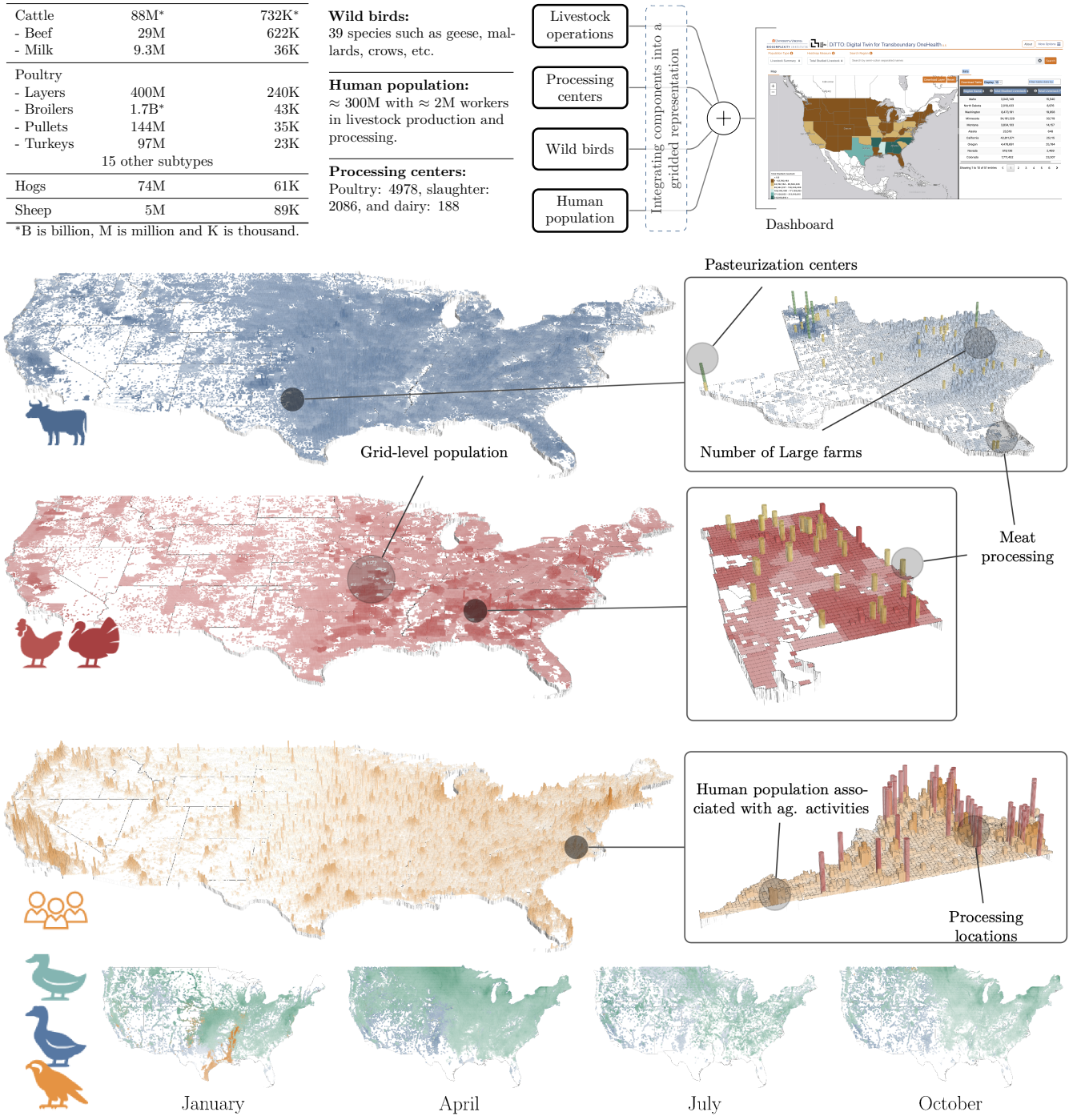


Fig. 1. Overview of the digital similar FIELD. A schematic of the system highlighting the various components and the dashboard through which the data is exposed is provided at the top. Two of the four livestock layers are shown. We have zoomed in on major production regions for the respective livestock. Both population density and counts of farms are depicted. Also shown are livestock and dairy processing centers. For the human population, agricultural workers are highlighted. The spatiotemporal distribution of three wild bird populations is shown in the bottom layer. We provide an interactive visualization dashboard, Digital Twin for Transboundary OneHealth (DiTTO) (33) to make the data available.

the risk of spillover to the population of s is given by

$$R(i,s,t) = P(i,s) \cdot A(i,t) \cdot B_{\mathrm{H5}}(i,t), \qquad [1]$$

where $P(i,s)$ is the livestock population from the livestock layer $\mathcal{L}$, $A(i,t)$ is the wild bird abundance at time t from the corresponding layer $\mathcal{B}$, and $B_{\mathrm{H5}}(i,t)$ is an estimate of the proportion of the wild bird population infected with the disease at time t (6, 34). For each time period, we aggregate the risk across all grid cells of a county to obtain risk $R_c(s,t)$ for a county c. We assign risk percentile ranks to counties, designating them as "Very high" ($\geq$ 95 percentile), "High" (90–94), "Medium" (75–89) and "Low" (0–74). We perform these

risk assessments quarterly. Our results are described below. From a livestock perspective, we focus on milk cattle, turkeys and chicken layers, which have been among the most affected among the livestock subtypes.

Strong predictive accuracy validates risk assessment framework. For both milk cattle and poultry, the risk estimates show strong concordance with observed H5N1 outbreaks. Figure 2c shows the results for milk cattle and turkeys, respectively, for each quarter. In general, across subtypes, a large portion of the H5N1 incidences occur in the very high and high risk counties corresponding to the period of occurrence, demonstrating that wild birds are the primary driver of introduction events in livestock. While our colocation model explains a majority of the incidences, there is scope to improve this model by giving careful consideration to the functional form of the model and accounting for other pathways of spread. The objective here was to demonstrate the importance of the subtype-specific farm abundance and the abundance of wild birds as some of the main driving factors of this phenomenon.

Risk is subtype specific. Figures 2a and 2b show the differences in the spatio-temporal risk across livestock subtypes. Quarterly risk maps in the supplement better illustrate this difference. While wild bird abundance generally influences spillover risk, not all subtypes are affected in the same way as observed in the incidence reports. This could be driven by biology, farming practices (e.g., indoor or outdoor facility), etc. The subtype population also drives the risk factor. However, we observe that some livestock-intensive counties show up as very high risk across subtypes informing cross-species spillover risk, as well as increased exposure to the livestock worker population, which is typically large in such instances, as per our analysis in Figure 3e.

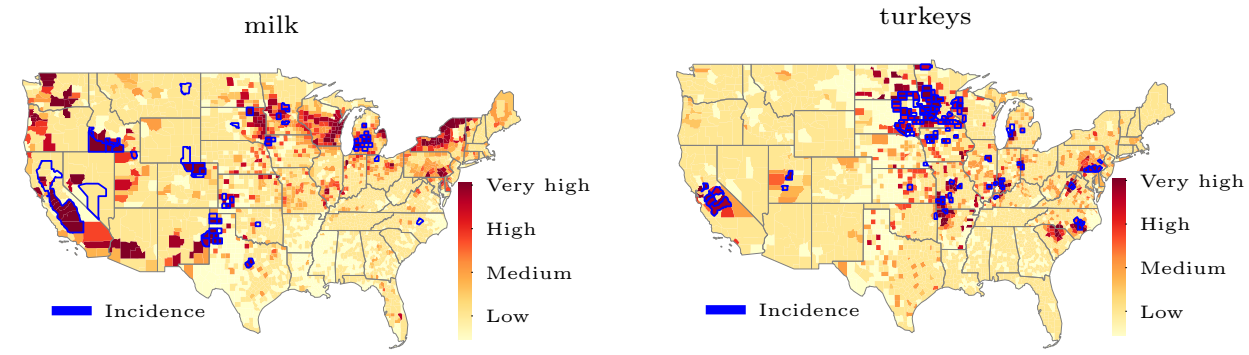
Persistent elevated risk informs surveillance priorities. Figure 2a maps risk persistence in milk cattle across counties, highlighting areas that are at elevated risk for multiple time periods. For both milk cattle and turkeys, there are counties that consistently rank among the top counties in the very high risk category over all four periods (including multiple counties in California, and Weld County in Colorado) that require constant surveillance throughout the year. We also observe counties whose rank fluctuates widely (such as Lancaster County in Pennsylvania), requiring time-dependent surveillance (results tabulated in the supplement). At state level, while California has the highest risk in the number of counties at very high risk throughout the year, for some states, like Colorado, this rank changes over time.

Periods of high risk. County-level temporal analysis of risk reveals spatial clustering. The plots in Figure 2b show, for each county, the quarter of highest risk. For most regions, either the first or the third quarter corresponds to the highest levels of risk. Generally, the risk is spatially clustered, with several county-clusters having the same temporal risk profiles; however, variations in habitats and the wild bird species hosted can lead to differences within a region. Spatial clustering of very high risk counties increases the potential for multiple spillover events, which amplify the probability and size of local outbreaks due to other pathways of spread.

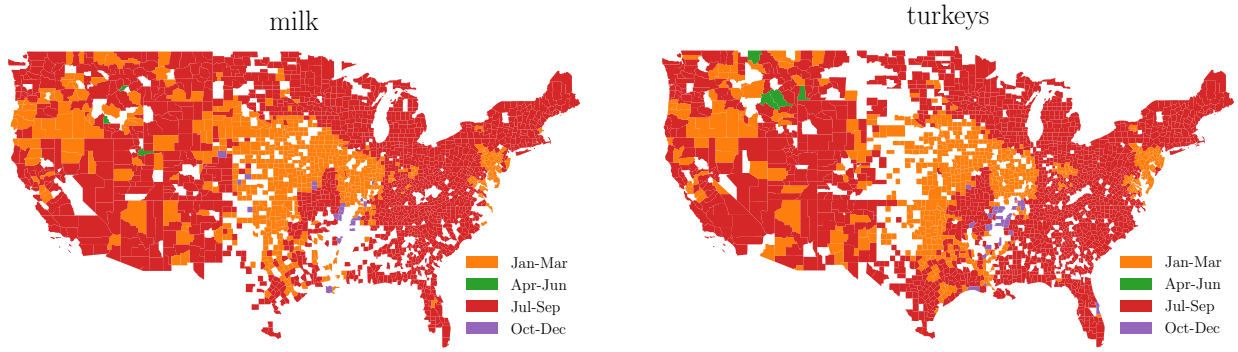
Future scenarios reveal potential shifts in the risk landscape. To evaluate potential future scenarios, we extended our analysis to consider the spillover risk to the entire cattle population not currently experiencing H5N1 outbreaks (e.g., among beef cattle). Figure 2d illustrates how risk hotspots would shift substantially under this scenario, with increased risk in regions like North Dakota, Texas, and Kansas, compared to the milk-centric risk maps. This analysis highlights the importance of early containment and the potential for broader agricultural impacts.

Validation of farm locations. We evaluated the construction process for the livestock layers by comparing the constructed layers with the parent datasets. The algorithms for generating farms and assigning them to grid cells (See Figure 3a) are also evaluated. These analyses are available in the supplement. Here, we focus on the validation of the resulting synthetic representation using independent datasets. Concentrated Agricultural Feed Operations (CAFOs) are large animal feeding operations that are a potential hazard to the environment and health. CAFOs are regulated by the Environmental Protection Agency (EPA), and some state agencies provide location information, among other attributes. We obtained such data from CafoMaps (35). We focus on cattle, hogs, and chickens. For each county–livestock instance for which such data is present, we selected large farms from our farm assignment based on livestock-specific thresholds informed by CAFO size specifications provided by various states (36). We computed the Haversine distance of each CAFO location to the centroid of each grid cell that contains a large farm. We construct a weighted complete bipartite graph $G(A, B)$ for each county–livestock instance. Here, A corresponds to farm locations from our assignments; each farm is assumed to be located at the centroid of the grid cell to which it belongs. The set B corresponds to CAFO locations. For each $u \in A$ and $v \in B$, the weight on the edge (u, v) is the inverse of the distance between the two locations. We compute a maximum weighted perfect matching[†] of this bipartite graph to match each CAFO location to a farm in FIELD. The main objective is to map as many CAFO locations as possible. It is possible that the number of farms is greater than the number of CAFO locations, as not all locations are listed. The results of the matching are analyzed in Figure 3b. We considered two sets of thresholds, the second set corresponding to larger farms compared to the first. A large percentage of CAFO locations were matched in the case of cattle ($> 95\%$) and chickens ($> 83\%$), while in the case of hogs we observe only 50% match. A closer examination of AgCensus data reveals the reason for the low number of matches for hogs: the number of farms specified by the AgCensus dataset for the relevant size categories is less than the number of CAFO locations specified. Among the matched locations, we observe that 90% of the CAFO locations are at most 10 miles from the grid centroid of the corresponding farm from FIELD, which places it in the same grid cell or a neighboring one. We note that a majority of matched CAFO farms corresponding to cattle and hogs are within 20 miles of the matched farm in FIELD.

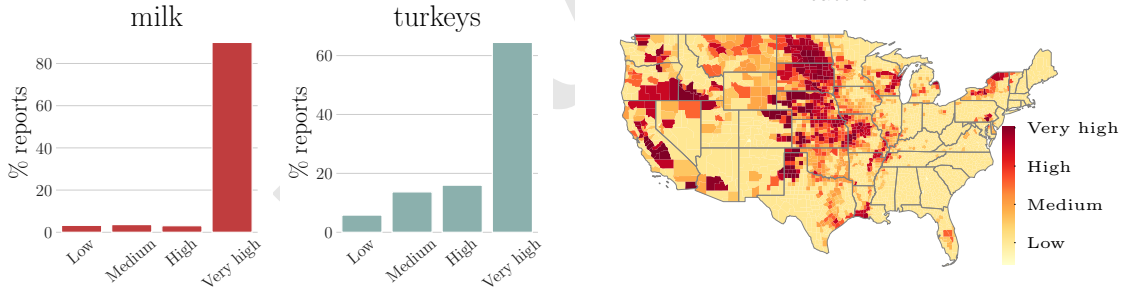
[†] For a complete bipartite graph $G(A, B, E)$, where $|A| = m$ and $|B| = n$, a perfect matching consists of $\min\{m, n\}$ edges.



(a) Persistence of elevated risk



(b) Period of peak risk

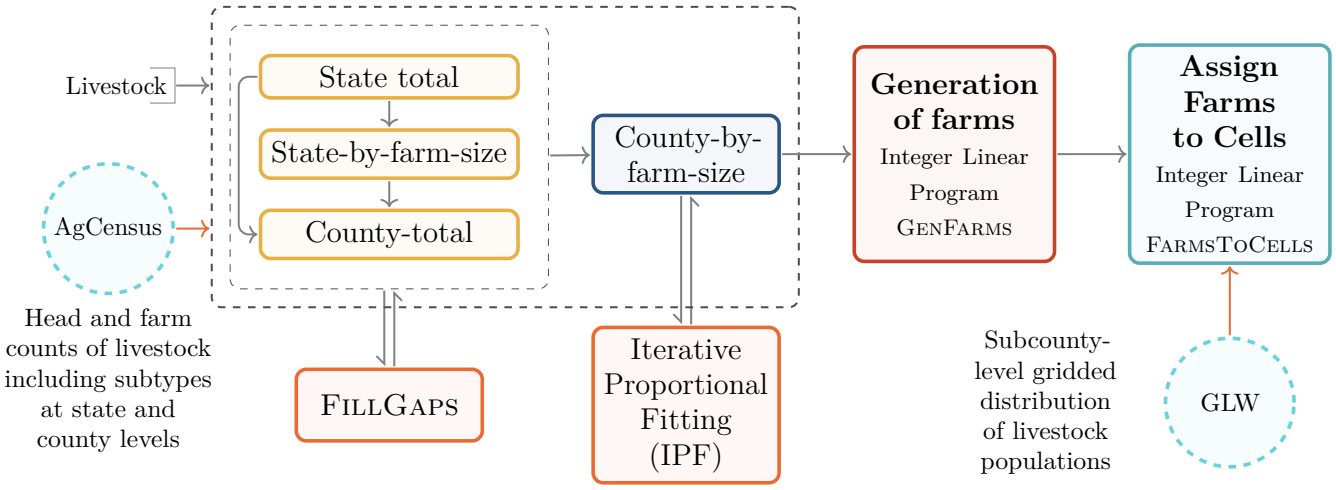


(c) Recall performance

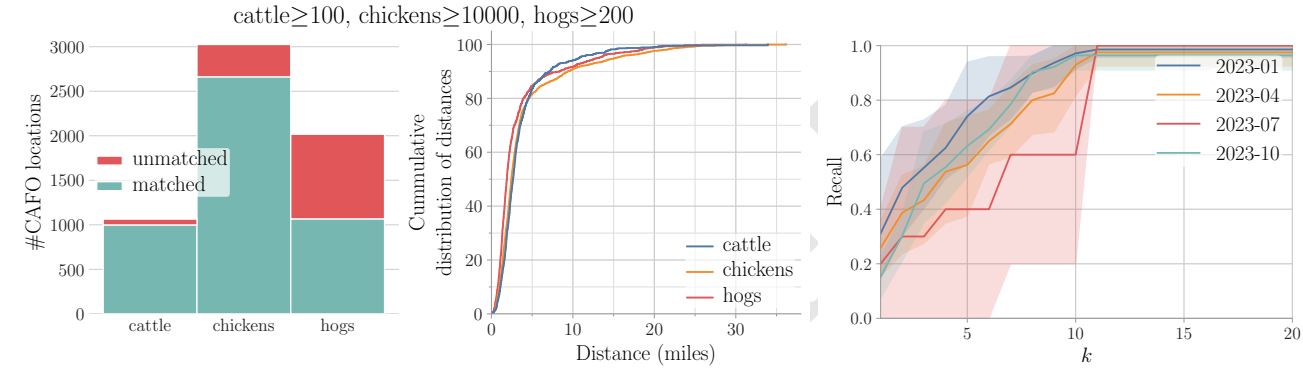
(d) Potential risk to all cattle

Fig. 2. Results corresponding to livestock subtype-specific quarterly risk maps based on county ranks. The risk scores were ranked per quarter and counties were assigned risk profiles: "Very high" ($\geq$ 95%), "High" (90 $-$ 94), "Medium" (75 $-$ 89) and "Low". Here, we show summary plots for milk cattle and turkeys. (a) Persistence of elevated risk. Quarter-specific ranks were combined to obtain a single ranking of counties by sorting counties by number of occurrences of "Very high" risk profiles across quarters followed by the number of occurrences of "high" risk profiles and so on. (b) Period of peak risk: For each county, we plot the period for which the risk is maximum ($\arg\max_{t=1,2,3,4} R_c(s,t)$). (c) Comparing ground-truth incidence reports with risk profiles. Each incidence is mapped to the risk profile of the corresponding county–quarter pair. Then, the incidences are binned into the respective risk profiles. (d) Persistence of elevated risk under a potential scenario of H5N1 spread in cattle (definition same as in (a)).

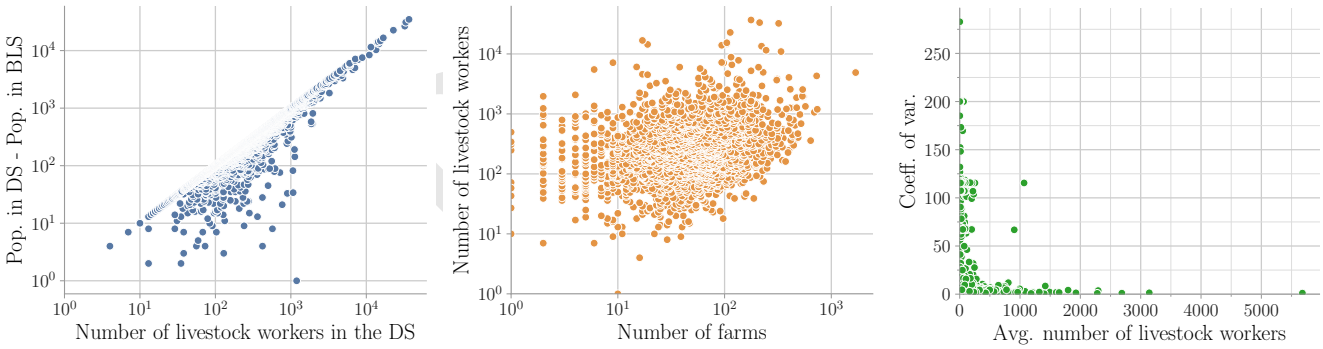
Wild bird abundance and H5N1 incidence. We compare the relative abundance of the chosen bird species, as captured in FIELD, with the occurrence of H5N1 cases at the state level. We recall that we chose to include all species with reference to H5N1 incidence data from 2022–2024. The objective is to ascertain whether there are H5N1 incidences among the most abundant birds in each state where cases have been reported. This exercise establishes the relevance of the included bird population to the colocation-based risk analysis. For each state, we calculate the average abundance for each bird species over the study period (January 1 to March 30, 2023) breaking down the period into four quarters. We then compile H5N1 case counts for each bird species in the same state and time frame. Bird species are grouped into 15 categories (e.g., 'Duck', 'Goose', 'Eagle') based on species similarity to provide more robust comparisons. We employ the **top-k recall** metric to quantify the relationship between bird abundance and H5N1 cases. It is the proportion of species with H5N1 cases that are among the k most abundant species in a state



(a) Schematic of the livestock layer construction



(b) **(c)**



(d) Comparing counts with BLS. **(e)** Workers vs. farms **(f)** Seasonal variations based on BLS.

Fig. 3. (a) A schematic of the livestock layer construction, including data processing, generation of farms and assignment of cells to farms. It takes as input AgCensus data that comprises head and farm counts at various administrative levels and the gridded distribution of the livestock populations from GLW. FillGaps is an integer program that fills gaps in the census data. GenFarms is an integer linear program (ILP) for distributing the livestock populations to farms consistent with the census data. FarmsToCells is an ILP that assigns farms to grid cells with the objective of aligning the population with GLW. Validation of components: (b) Analysis of mapping CAFO locations by livestock type to farms from FIELD. Farms were chosen based on the thresholds stated in the title. The first subplot shows how many CAFO locations were matched. The supplement has an additional plot for a different set of thresholds. The second subplot provides the cumulative distribution of the distances (in miles) between matched pairs of CAFO locations and farms. An additional plot for a different set of thresholds is in the supplement. (c) Analysis of H5N1 cases and bird abundance for the period of January to December of 2023. The plot summarizes the results across eight reporting states for the four quarters. The x-axis corresponds to the number of top species groups considered, while the y-axis corresponds to the count of those groups with H5N1 incidence. (d-f) Analysis of the livestock worker population. All counts presented are at the county level. (d) Comparing livestock worker counts between FIELD and data from the Bureau of Labor Statistics (BLS). (e) Number of livestock workers vs. number of farms at the county level. (f) The variation in worker populations across the year. The coefficient of variation on the x-axis is computed across the four quarters of year 2023.

within the target period and administrative region.

$$\text{Top-k Recall} = \frac{\#\text{Hosts with cases among top k abundant hosts}}{\#\text{Hosts with cases}}.$$

The results presented in Figure 3c summarizes the Top-k recall results for all states and different parts of the year. We observe that, for all quarters, the larger the bird population,

the greater the likelihood of observed H5N1 cases. Secondly, we observe that, in most cases, the top 10 abundant birds cover all reported H5N1 cases (for $k \geq 10$), so the recall value is close to 1. More plots are shown in the supplement for multiple US states with details about specific bird types. These results suggest a robust correlation between bird abundance and H5N1 case occurrences.

Livestock worker population. Here, we compare the livestock worker population in FIELD with counts obtained from the Quarterly Census of Employment and Wages data corresponding to the year 2023 (Table 1, BLS). Details of how livestock workers were identified in each dataset are present in the supplement. The plot in Figure 3d shows the difference between our data and the BLS data in counts of livestock worker population by county and state, respectively, for 2769 counties that are common to both the datasets. We note that, for more than 96% of the counties, the synthetic population counts exceed that of BLS. This is expected as BLS excludes a significant population of farm workers; it only counts workers covered under unemployment insurance. For the remaining counties, the difference is usually very small compared to many of the remaining instances where the counts in FIELD far exceed BLS. It is possible that a significant portion of the farm worker population in these counties did not participate in the census.

Figure 3e shows a comparison between county-level farm counts and the livestock worker population. Due to missing information about mixing livestock in farms, our total farm count is higher than the total mentioned in AgCensus. Hence, we only considered farms with head counts of at least 100. Generally, for counties with higher farm counts, the number of workers is higher. But there is a wide spread in the number of workers for a fixed farm count. Since counts can depend on farm sizes and livestock types, without additional information it becomes almost impossible to compare the two quantities in further detail.

Analysis of BLS shows little variation in the county-level counts of livestock workers across the year (see Figure 3f). With the exception of two outliers, counties with very large coefficients of variation have very few livestock workers. For the two outlier counties, there is at least one quarter with zero count, which could be attributed to missing data.

Discussion

Outline. Our work presents a comprehensive methodology to construct synthetic spatiotemporal datasets of food systems, human population, and wildlife. To this end, we bring together diverse datasets and combine them using combinatorial optimization and statistical methods to develop a multi-layered digital similar called FIELD. This is complemented by extensive verification and validation studies. While the choice of agents and design decisions were influenced by the focus application – the spread of HPAI-like diseases – the utility of such a digital similar extends beyond this domain. The following discussion elaborates on these points while also highlighting the limitations of FIELD.

Related work. Several models have been proposed for subcounty-level disaggregation of livestock populations (13, 14, 31, 37). Using a random forest model, Gilbert et al. (31) develop a global distribution of populations of multiple livestock types. Burdett et al. (13) develop the FLAPS model to simulate populations and locations of individual farms for swine using the AgCensus dataset and a microsimulation model; this work has been used to analyze the spread of porcine deltacoronavirus in the US (37). For mapping concentrated animal feeding operations (CAFOs), several deep learning methods have been proposed to map industrial operations (38, 39). In the context of HPAIs, there is a need for significant extensions to these works given the requirement to simultaneously account for multiple livestock types and wild birds, as well as organizational-level distributions. In a very recent work, Prosser et al. (6) address estimation of transmission risk at the wild waterfowl–domestic poultry interface. They develop a spatiotemporal model combining 10 species-level wild bird abundance models (40) with a commodity-level poultry farm model (41). Humphreys et al. (11) use a variety of datasets, including GLW, to model waterfowl movement and interactions with poultry farms and human populations. Our work is shaped by some of these works.

For many regions, as demonstrated in our risk analysis, colocation of large livestock populations with wild birds is a good explanation for spillover risk. Previous works in the case of poultry (6, 11) highlight this aspect. Our results for dairy are similar in that aspect. However, there are some differences between outbreaks in poultry and dairy cattle. While poultry outbreaks have been largely sporadic, infections in dairy cattle has persisted and spread to neighboring counties in many states (like Colorado and Michigan in the beginning, and later, California). Capturing this would require knowledge of movement of animals (as shown by recent works (8, 12)) and role of agricultural workers from the perspective of human-mediated pathways. Using a stochastic metapopulation model, Rawson et al. (12) attribute the spread in the West Coast states to the network of cattle movements. Capturing the movement of birds could help characterize the spread through spillover from wild birds. The approaches used in BirdFlow (42) model the flight paths of wild birds, but incorporating this into our work would require sample trajectory data.

Limitations. While this dataset offers valuable insights as demonstrated in our work, it is important to acknowledge its limitations and the potential for future improvements. The dataset faces challenges primarily stemming from the nature of its parent datasets. For example, the synthetic dataset GLW is misaligned in time with respect to AgCensus. In addition, as shown in our analysis with respect to farms whose locations are known, not all assigned operations are matched, indicating spatiotemporal misalignment with AgCensus. Unlike some previous works (6, 13), we do not produce coordinate-level assignments for operations. While this might become a limitation for very fine-grained analyses, such as farm-to-farm movement of animals, for such datasets to be useful, additional location- and operation-level information and rigorous validation is required. Our work does not model mixtures of different livestock types (such as cattle and poultry in the same farm), while for some livestock types (like poultry), there is not enough information about farm sizes, leading to a heavy-tailed distribution of populations across farms. Further exploration of cross-tabulation data from AgCensus that is available by request could allow us to improve on these aspects. For wild birds, eBird status and

trends data provides only relative abundance measures, which are subject to observational biases and tend to underestimate true populations. Mapping agricultural workers to farms is a challenging task, as it is a function of farm size, livestock type, and the level of automation employed (43). The comparison exercise with data from BLS highlights the challenges of estimating farm worker counts. Despite these limitations, our dataset provides a valuable foundation for studying complex interactions between livestock, wild birds, and human populations in the context of avian influenza transmission.

Conclusion. Beyond the study of H5N1, FIELD offers valuable applications to other One Health issues and beyond. Its modular nature enables us to leverage subsets of layers depending on the nature of the application. This dataset can be applied to model the spread of other pathogens, such as West Nile virus or Salmonella, which also involve interactions between livestock and human populations (44–46). Such systems have value beyond infectious diseases in domains such as food safety, agricultural economics, environmental damage, pollution, disaster response, biosecurity, and supply chain problems (13, 31, 35, 47). In biodiversity and conservation efforts, the wild bird abundance data can aid in identifying critical habitats and migration corridors, particularly in the context of livestock operations. Spatially explicit synthetic datasets are being extensively developed for such non-epidemiological settings (48–52). Our digital similar can extend such works to account for additional ecosystems, such as livestock, in the respective applications.

Materials and Methods

Datasets. Table 1 summarizes all data used. We used publicly available datasets, which can be categorized into three types: census, synthetic realistic datasets derived from models and data samples, and real location-level datasets. From a spatial unit perspective, some data (e.g., AgCensus and H5N1 incidence data) are specified at various administration levels (county or state), some data (e.g., GLW and eBird) are specified at the grid level, while exact locations are provided for the rest. Some of these datasets have been used for the construction of FIELD, while the rest have been used for subsequent analysis. More details about each dataset are provided in the relevant sections.

Livestock. Two data sources were used to construct the livestock layers: Census of Agriculture (AgCensus) and Gridded Livestock of the World (GLW); these are described in more detail below. An overview of the methodology is shown in Figure 3a. The types of livestock covered in this work are cattle, poultry, sheep and hogs. Additional details including data organization, gaps, and other challenges are described in the supplement.

We use a combination of integer linear programs (ILPs) and iterative proportional fitting (IPF), the latter following Burdett et al. (13, 67, 68). For cattle and poultry, gaps are filled for each subtype, while for hogs and sheep, they are filled for the livestock type. We use the integer program FillGaps (described in the supplement) to fill in missing data for the following types of counts: state-total, state-by-farm-size, and county-total. It takes as input all the known counts, the sum of all the counts, and the bounds on the unknown counts, and distributes the heads that are unaccounted for equitably across all entities for which the counts are missing. The algorithm respects the bounds provided as input. To fill gaps for county-by-farm-size counts, we apply IPF. At each step, the objective is to make use of all available data (in all count types).

The next step is the generation of farms. Given a livestock type and county, the objective is to obtain a grid-level distribution of farms that is consistent with the AgCensus data from the perspective of operations and their sizes, and the GLW data from the perspective of the grid-level distribution of livestock populations. We use a two-step procedure using optimization algorithms: (i) generating farms consistent with the county-by-farm-size counts (either provided for or estimated) and (ii) assigning farms to cells. The GenFarms algorithm for generating farms is described in the supplement. The objective function encodes several minimization criteria. They are stated in order of priority: (i) feasibility: modifies the subtype head count minimally to ensure feasibility of the assignment, (ii) equitable distribution of head counts for each subtype of livestock across farms within a category, (iii) minimize the number of subtypes within a farm, and (iv) align with known county-by-farm-size counts. The FarmsToCells algorithm (described in the supplement) assigns a cell to each farm. The objective of this algorithm is to ensure that the head counts resulting from the assignment and GLW head counts are as closely aligned as possible.

Wild Bird Abundance and Movement. We leveraged data from eBird's Status and Trends products (58) (see Table 1) to construct this component. We recall that this component $\mathcal{B}\left(\Theta_{\mathcal{B}}, A(\cdot), G_{\mathcal{B}}\right)$ captures spatiotemporal abundance and movement of multiple species of birds. The eBird data provides weekly estimates of relative bird abundance across a high-resolution grid (2.96km×2.96km). A relative abundance of 1.0 for a species at a particular location and time would indicate that an average eBird checklist at that place and time would be expected to count one individual of that species. Higher values indicate more individuals would be expected, while lower values indicate the species would be observed less frequently or in smaller numbers. We chose bird species for which H5N1 cases were observed in the period 2022–2024. Out of 40 species, abundance data was available for 36 species. For each of the selected species, we extract relative abundance values along with their associated geographic coordinates for all 52 weeks in the year.

Dairy, Meat, and Egg Processing Plants. We provide a layer for animal product processing plants with attributes such as size, type of processing (dairy, meat, egg, etc.), livestock type, etc. (see Table 1). While coordinate-level data was available for the location of meat processing plants, For dairy plants, the locations at the city level were inferred by manual inspection.

Human Population. We develop a gridded representation of the US population with rich demographic and employment-related attributes. This data is derived from a synthetic population (16, 17, 62) that is developed using diverse datasets such as census data, land use data, activity patterns, building maps, etc. Each individual in the population is associated with a residential location, an occupation identified by the Standard Occupational Classification code (SOCP), and an industry, identified by the North American Industry Classification System (NAICS) code. We identified all occupational and industry codes that include livestock employment; these are listed in the supplement. Individuals whose SOCP or NAICS codes did not belong to this list were assigned a default code 0. For each combination of attributes, the population is aggregated at the grid cell level.

ACKNOWLEDGMENTS. We thank members of the Biocomplexity Institute and UVA Research Computing for their support. We also thank VA PGCOE members, past members of the Office for Pandemic Preparedness and Response Policy, Dan Hanfling, and Cyrus Shahpur, past members of the National Security Council, Shankar Sundaram and Rachel Idowu, and CDC staff members, Eleanor Click, Cećil Viboud, and Margaret (Peggy) Honein, for their thoughtful comments and suggestions. We thank members of the Center for Forecasting and Outbreak Analytics Insight Net, CDC and participants of the workshop titled "Potential Research Priorities to Inform Readiness and Response to Highly Pathogenic Avian Influenza A (H5N1)" (organized by the National Academies of Sciences, Engineering, and Medicine) for their valuable inputs. We are grateful to the eBird Status & Trends team. This work was partially supported by University of Virginia Strategic Investment

Table 1. Datasets explored to construct and validate FIELD. Throughout the paper, each dataset will be referred to by its abbreviation.

Name	Abbrv.	Source	Description
Census of Agriculture	AGCENSUS	(53–55)	Provides location-level Ag data such as number of farms, farm sizes, crop types, and fallowed status. Provides individual-level Ag data such as number of workers on a farm.
Gridded Livestock of the World	GLW	(31, 56, 57)	GLW4.0 provides distribution maps for several livestock types. FAO hosts this website.
eBird Status and Trends	EBIRD	(40, 58)	Weekly data of relative abundance of migratory birds across geospatial regions throughout the year.
Dairy processing	Dairy plants	(59)	Large dairy processing centers and their attributes regulated by USDA AMS.
Meat, poultry, and egg processing	Meat and poultry	(60)	Listing of establishments that produce meat, poultry, and/or egg products regulated by USDA FSIS.
US population	USPOP	(61, 62)	Synthetic digital twin of the US human population
Quarterly Census of Employment and Wages	BLS	(63)	Quarterly counts of livestock workers (NAICS code 112) from Bureau of Labor Statistics
CAFOs in the US	CAFOMAPS	(35, 36)	A map of Concentrated Animal Feeding Operations (CAFOs) in the southern United States covering nine states.
H5N1 outbreaks	H5N1CASES	(9, 64–66)	H5N1 bird flu detections in wild birds, livestock, and humans by state and county

Fund award number SIF160, National Science Foundation (NSF) Expeditions in Computing Grant CCF-1918656, PGCoE CDC-RFA-CK22-2204, DTRA subcontract/ARA S-D00189-15-TO-01-UVA, USDA-NIFA and NSF under the AI Institute: Agricultural AI for Transforming Workforce and Decision Support (AgAID) award No. 2021-67021-35344, USDA-NIFA under the Network Models of Food Systems and their Application to Invasive Species Spread, grant 2019-67021-29933. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agencies. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention.

1. V Caliendo, et al., Transatlantic spread of highly pathogenic avian influenza H5N1 by wild birds from Europe to North America in 2021. *Sci. Reports* **12**, 11729 (2022).
2. M Leguia, et al., Highly pathogenic avian influenza A (H5N1) in marine mammals and seabirds in Peru. *Nat. Commun.* **14**, 5489 (2023).
3. MM Uhart, et al., Massive outbreak of Influenza A H5N1 in elephant seals at Peninsula Valdes, Argentina: increased evidence for mammal-to-mammal transmission (bioRxiv) (2024).
4. W Puryear, et al., Highly pathogenic avian influenza A (H5N1) virus outbreak in New England seals, United States. *Emerg. Infect. Dis.* **29**, 786 (2023).
5. ER Burrough, et al., Highly pathogenic avian influenza A (H5N1) clade 2.3. 4.4 b virus infection in domestic dairy cattle and cats, United States, 2024. *Emerg. infectious diseases* **30**, 1335 (2024).
6. DJ Prosser, et al., Using an adaptive modeling framework to identify avian influenza spillover risk at the wild-domestic interface. *Sci. Reports* **14**, 14199 (2024).
7. LC Caserta, et al., Spillover of highly pathogenic avian influenza H5N1 virus to dairy cattle. *Nature* **634**, 1–8 (2024).
8. TQ Nguyen, et al., Emergence and interstate spread of highly pathogenic avian influenza a (H5N1) in dairy cattle (bioRxiv) (2024).
9. CDC, H5 bird flu: Current situation (https://www.cdc.gov/bird-flu/situation-summary/index.html) (2025) [Accessed 01-2025].
10. MP Koopmans, et al., The panzootic spread of highly pathogenic avian influenza H5N1 sublineage 2.3. 4.4 b: a critical appraisal of one health preparedness and prevention. *The Lancet Infect. Dis.* **24**, e774–e781 (2024).
11. JM Humphreys, et al., Waterfowl occurrence and residence time as indicators of H5 and H7 avian influenza in North American poultry. *Sci. Reports* **10**, 2592 (2020).
12. T Rawson, et al., A mathematical model of H5N1 influenza transmission in US dairy cattle (medRxiv) (2025).
13. CL Burdett, BR Kraus, SJ Garza, RS Miller, KE Bjork, Simulating the distribution of individual livestock farms and their populations in the United States: An example using domestic swine (*Sus scrofa domesticus*) farms. *PloS One* **10**, e0140338 (2015).
14. M Cheng, X Liu, H Sheng, Z Yuan, MAPS: A new model using data fusion to enhance the accuracy of high-resolution mapping for livestock production systems. *One Earth* **6**, 1190–1201 (2023).
15. M van Andel, MJ Tildesley, MC Gates, Challenges and opportunities for using national animal datasets to support foot-and-mouth disease control. *Transboundary Emerg. Dis.* **68**, 1800–1813 (2021).
16. S Eubank, et al., Modelling Disease Outbreaks in Realistic Urban Social Networks. *Nature* **429**, 180–184 (2004).
17. G Harrison, et al., Synthetic information and digital twins for pandemic science: Challenges and opportunities in *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. (IEEE), pp. 23–33 (2023).
18. S Mihai, et al., Digital twins: A survey on enabling technologies, challenges, trends and future prospects. *IEEE Commun. Surv. & Tutorials* **24**, 2255–2291 (2022).
19. Y Wu, K Zhang, Y Zhang, Digital twin networks: A survey. *IEEE Internet Things J.* **8**, 13789–13804 (2021).
20. JA Delgado, NM Short Jr, DP Roberts, B Vandenberg, Big data analysis for sustainable agriculture on a geospatial cloud framework. *Front. Sustain. Food Syst.* **3**, 54 (2019).
21. C Pylianidis, S Osinga, IN Athanasiadis, Introducing digital twins to agriculture. *Comput. Electron. Agric.* **184**, 105942 (2021).
22. M Batty, Digital twins in city planning. *Nat. Comput. Sci.* **4**, 192–199 (2024).
23. G Caldarelli, et al., The role of complexity for digital twins of cities. *Nat. Comput. Sci.* **3**, 374–381 (2023).
24. M Abueg, et al., Modeling the Combined Effect of Digital Exposure Notification and Non-Pharmaceutical Interventions on the COVID-19 Epidemic in Washington State (MedRxiv) (2020).
25. L Ferretti, et al., Quantifying SARS-CoV-2 Transmission Suggests Epidemic Control with Digital Contact Tracing. *Science* **368**, eabb6936 (2020).
26. S Hoops, et al., High Performance Agent-Based Modeling to Study Realistic Contact Tracing Protocols in *2021 Winter Simulation Conference (WSC)*, eds. K Sojung, B Feng, K Smith, S Masoud, Z Zheng. (IEEE), pp. 1–12 (2021).
27. CC Kerr, et al., Covasim: An Agent-Based Model of COVID-19 Dynamics and Interventions. *PLOS Comput. Biol.* **17**, e1009149 (2021).
28. A Aleta, et al., Modelling the Impact of Testing, Contact Tracing and Household Quarantine on Second Waves of COVID-19. *Nat. Hum. Behav.* **4**, 964–971 (2020).
29. J Chen, et al., Effective social network-based allocation of COVID-19 vaccines in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. pp. 4675–4683 (2022).
30. CT Lloyd, A Sorichetta, AJ Tatem, High resolution global gridded data for use in population studies. *Sci. Data* **4**, 1–17 (2017).
31. M Gilbert, et al., Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci. Data* **5**, 1–11 (2018).
32. MC Bruhn, et al., Synthesized population databases: a geospatial database of us poultry farms (Methods Report (RTI Press)) (2012).
33. Biocomplexity Institute, University of Virginia, DiTTO: Digital Twin for Transboundary OneHealth (https://ditto.bii.virginia.edu/) (2025) [Accessed 01-2025].
34. CM Kent, et al., Spatiotemporal changes in influenza A virus prevalence among wild waterfowl inhabiting the continental United States throughout the annual cycle. *Sci. Reports* **12**, 13083 (2022).
35. The University of Iowa, CAFOs in the US (https://cafomaps.org/) (2024) [Accessed 08-19-2024].
36. SRAP project, State CAFO guides (https://sraproject.org/state-cafo-guides/) (2024) [Accessed 10-22-2024].
37. FC Paim, et al., Epidemiology of deltacoronaviruses (δ-CoV) and gammacoronaviruses (γ-CoV) in wild birds in the United States. *Viruses* **11**, 897 (2019).
38. C Handan-Nader, DE Ho, Deep learning to map concentrated animal feeding operations. *Nat. Sustain.* **2**, 298–306 (2019).

39. C Robinson, B Chugg, B Anderson, JML Ferres, DE Ho, Mapping industrial poultry operations at scale with deep learning and aerial imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **15**, 7458–7471 (2022).
40. BL Sullivan, et al., eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142**, 2282–2292 (2009).
41. KA Patyk, et al., Modelling the domestic poultry population in the United States: A novel approach leveraging remote sensing and synthetic data methods. *Geospatial Heal.* **15** (2020).
42. M Fuentes, BM Van Doren, D Fink, D Sheldon, Birdflow: Learning seasonal bird movements from eBird data. *Methods Ecol. Evol.* **14**, 923–938 (2023).
43. JM MacDonald, WD McBride, The transformation of US livestock agriculture scale, efficiency, and risks (Economic Information Bulletin No. 43, Economic Research Service, U.S. Dept. of Agriculture) (2009).
44. Y Xiao, D Clancy, NP French, RG Bowers, A semi-stochastic model for salmonella infection in a multi-group herd. *Math. Biosci.* **200**, 214–233 (2006).
45. MH Myer, JM Johnston, Spatiotemporal Bayesian modeling of West Nile virus: Identifying risk of infection in mosquitoes with local-scale predictors. *Sci. Total. Environ.* **650**, 2818–2829 (2019).
46. K Libera, et al., Selected livestock-associated zoonoses as a growing challenge for public health. *Infect. Dis. Reports* **14**, 63–81 (2022).
47. C Zhu, et al., High spatiotemporal resolution ammonia emission inventory from typical industrial and agricultural province of China from 2000 to 2020. *Sci. The Total. Environ.* **918**, 170732 (2024).
48. R Meyur, et al., Ensembles of realistic power distribution networks. *Proc. Natl. Acad. Sci.* **119**, e2205772119 (2022).
49. S Thorve, et al., High resolution synthetic residential energy use profiles for the United States. *Sci. Data* **10**, 76 (2023).
50. R Yuan, et al., A synthetic dataset of Danish residential electricity prosumers. *Sci. Data* **10**, 371 (2023).
51. C Barrett, et al., Planning and response in the aftermath of a large crisis: An agent-based informatics framework in *2013 Winter Simulations Conference (WSC)*. (IEEE), pp. 1515–1526 (2013).
52. MV Marathe, HS Mortveit, N Parikh, S Swarup, Prescriptive analytics using synthetic information in *Emerging Methods in Predictive Analytics: Risk Management and Decision-Making*. (IGI Global), pp. 1–19 (2014).
53. U National Agricultural Statistics Service, Census of Agriculture (https://www.nass.usda.gov/AgCensus/) (2022) [Accessed 03-Jan-2023].
54. U National Agricultural Statistics Service, Census of Agriculture (full data) (https://www.nass.usda.gov/datasets/qs.census2022.txt.gz.) (2022) [Accessed 11-Jun-2024].
55. ESRI, USDA Census of Agriculture 2017 - cattle production (https://www.arcgis.com/home/item.html?id=53137233a760432bb07c417eb3d758b8) (2017) [Accessed 06-03-2024].
56. TP Robinson, et al., Mapping the global distribution of livestock. *PloS one* **9**, e96084 (2014).
57. FAO, GLW 4: Gridded Livestock Density (https://data.apps.fao.org/catalog/dataset/15f8c56c-5499-45d5-bd89-59ef6c026704) (2020) [Accessed 06-03-2024].
58. D Fink, et al., eBird status and trends, data version: 2022; released: 2023 (2023).
59. USDA AMS, Dairy plants surveyed and approved for USDA grading service (https://apps.ams.usda.gov/dairy/ApprovedPlantList/) (2024) [Accessed 08-2024].
60. U FSIS, Meat, poultry and egg product inspection directory (https://www.fsis.usda.gov/inspection/establishments/meat-poultry-and-egg-product-inspection-directory) (2024) [Accessed 08-2024].
61. A Adiga, et al., Generating a synthetic population of the United States (Network Dynamics and Simulation Science Laboratory, Tech. Rep. NDSSL) (2015).
62. J Chen, et al., Epihiper—a high performance computational modeling framework to support epidemic science. *PNAS nexus* **4**, pgae557 (2025).
63. Bureau of Labor Statistics, Quarterly Census of Employment and Wages (https://www.bls.gov/cew/) (2023) [Accessed 2025-01].
64. USDA APHIS, Detections of highly pathogenic avian influenza in wild birds (https://www.aphis.usda.gov/livestock-poultry-disease/avian/avian-influenza/hpai-detections/wild-birds) (2025) [Accessed 02-2025].
65. USDA APHIS, Confirmations of Highly Pathogenic Avian Influenza in Commercial and Backyard Flocks (https://www.aphis.usda.gov/livestock-poultrydisease/avian/avian-influenza/hpai-detections/commercialbackyard-flocks) (2025) [Accessed 03-2025].
66. World Organization for Animal Health, United States of America - Influenza A viruses of high pathogenicity (https://wahis.woah.org/#/in-event/4451/dashboard) (2025) [Accessed 01-2025].
67. SE Fienberg, An iterative procedure for estimation in contingency tables. *The Annals Math. Stat.* **41**, 907–917 (1970).
68. WE Deming, FF Stephan, On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals Math. Stat.* **11**, 427–444 (1940).

Supplementary Information

A Livestock

A Organization and Preliminary Definitions

The types of livestock covered in this work are shown in Table 2. Two data sources were used to construct the livestock layers: Census of Agriculture (AgCensus) and Gridded Livestock of the World (GLW). These are described in the following sections. The overview of the methodology used to fill gaps, generate farms, and assign farms to grid cells is captured in Figure 3a. The description of the same appears in Sections D and F.

A **livestock type** refers to a class of animals. Examples of livestock types include cattle, poultry, hogs, and sheep. A **livestock subtype** represents a subclass of animals within a livestock type. For example, the livestock type cattle includes subtypes such as beef cows and milk cows. Likewise, the livestock type poultry includes subtypes such as (egg) layers, pullets, turkeys, etc.

B Census of Agriculture (AgCensus)

B.1 Data organization and availability

AgCensus provides counts of heads (i.e., population size) and farms for various livestock types and subtypes. The data is available at three different administrative levels – country, state, and county. The livestock types we consider here are cattle, poultry, hogs, and sheep. We also consider various subtypes for cattle and poultry. These types and subtypes are summarized in Table 2. The data organization is livestock type specific, making it a non-trivial task to extract relevant information. For each administrative level, the total counts are provided. Also provided are counts corresponding to different farm sizes. Farms are binned into categories based on the head counts of the corresponding livestock type, such as 1–24, 25–49, 50–99, 100–199, 200–499, 500–999, and 1000 or more, where each category is specified by the minimum and maximum head count, respectively, in the member farm. Accordingly, we have four types of counts: (i) state-total, (ii) state-by-farm-size, (iii) county-total, and (iv) county-by-farm-size. Table 3 depicts this organization of counts at the county level. The set of farm categories for each subtype is consistent with that of its parent livestock type.

Missing data. In many instances, head counts are redacted. The more refined the count category, the greater the incidence of missing data. There are more instances of missing data (i) at the county level compared to state level, (ii) in the farm-size categories compared to total head counts, and (iii) in subtype counts compared to total livestock counts. This can be observed in Table 2 for head counts aggregated using different types of counts. Operation counts, on the other hand, are always provided. For all livestock types except poultry, farm counts are provided at every administrative level for every farm category.

Notation. We set up some formal notation here to facilitate the description of our framework. Since each livestock type is processed independently, the notation will not carry information about the livestock type; it is assumed that the livestock type is known. The same holds true for the administrative level. Given a livestock type, the number of categories is denoted by ℓ. Then, for $i = 1, \ldots, \ell$, a category is specified by $(W_i^{\min}, W_i^{\max})$, the minimum and maximum population sizes respectively. Given an administrative level, let H denote the total head count and F denote the total farm count. For farm category i, let H_i and F_i denote the head and farm counts, respectively. Let Γ denote the set of different subtypes. The notation is similar to the one developed above; for a subtype $\gamma \in \Gamma$, H_γ and F_γ denote the total counts of the subtype at the target administrative level, and $H_{\gamma k}$ and $F_{\gamma k}$ denote the farm category specific counts for category $k = 1, \ldots, \ell$.

Table 2. Population and operations statistics for various livestock types covered by our synthetic dataset. Also shown are the counts after filling gaps.

livestock	subtype	state tot.	county tot.	filled gaps	heads final	farms state	farms processed
cattle	all	87954742	85973763	85973763	87932032	732123	731981
	beef	29214479	27790671	29207376	29199243	622162	622050
	milk	9309855	7526842	9317802	9317612	36024	35996
	other	49430408	46255380	49422350	49415177	594222	594108
hogs	all	73645928	62541219	73810004	73808393	60809	60731
poultry	chukars	1036946	621024	1048787	1047104	801	800
	ckn-broilers	1737674957	1680674087	1737674725	1737795431	42991	42947
	ckn-layers	375927945	199001866	388508984	389641754	240530	240270
	ckn-pullets	139203843	82678506	144030350	144029544	34874	34829
	ckn-roosters	7656478	6858454	7720552	7720400	42110	42064
	ducks	4341317	3422540	4448858	4448287	34781	34724
	emus	12538	9462	12440	12427	1566	1561
	geese	101823	83174	101521	101320	11940	11911
	guineas	391931	340504	391674	391549	18853	18844
	ostriches	2245	1519	3496	3496	232	232
	partridges	49162	9462	61147	61147	68	68
	peafowl	54947	42795	54679	54669	6930	6928
	pheasants	3187136	1243764	3279830	3266790	2257	2255
	pigeons	212559	160312	302934	285600	2196	2194
	poultry-other	69840	37923	84241	84241	789	789
	quail	9188443	6245272	9294150	9293769	4738	4731
	rheas	1013	382	1122	1122	152	152
	turkeys	97064430	84529090	97312274	97311591	23431	23373
sheep	all	5104328	3664088	5103716	5102574	88853	88795

Table 3. The data format for state-by-farm-size and county-by-farm-size. Some of the head counts data is redacted. The corresponding totals (either state or county) are denoted by H (for "all"), H_{beef}, H_{milk} and H_{other}. Depending on the instance, any of the totals or counts by farm size can be missing.

Cat.	Farm size	all	beef	milk	other
1	1–9	(F_1, H_1)	$(F_{1,\text{beef}}, H_{1,\text{beef}})$	$(F_{1,\text{milk}}, H_{1,\text{milk}})$	$(F_{1,\text{other}}, H_{1,\text{other}})$
2	10–19	(F_2, H_2)	$(F_{2,\text{beef}}, H_{2,\text{beef}})$	$(F_{2,\text{milk}}, H_{2,\text{milk}})$	$(F_{2,\text{other}}, H_{2,\text{other}})$
	⋮			⋯	
i	$W_i^{\min}$–$W_i^{\max}$	(F_i, H_i)	$(F_{i,\text{beef}}, H_{i,\text{beef}})$	$(F_{i,\text{milk}}, H_{i,\text{milk}})$	$(F_{i,\text{other}}, H_{i,\text{other}})$
	⋮			⋯	

B.2 Livestock type-specific information

The relevant head and farm counts were extracted from the full AgCensus dataset by querying out the rows where `statisticcat_desc="INVENTORY"`; these rows are further filtered by `unit_desc="HEAD"` or `unit_desc="OPERATIONS"`, depending on whether head counts or number of operations are being calculated, respectively. The state- and county-level counts were extracted by filtering `agg_level` to `STATE` and `COUNTY`, respectively.

Cattle. The counts were obtained by extracting rows where `commodity_desc="CATTLE"`. The cattle population is categorized into three subtypes, `beef`, `milk`, and `other` (specified by the `class_desc` field). The total count of cattle was obtained by setting `class_desc="INCL CALVES"`. The state- and county-total counts were obtained by extracting rows where `domain_desc="TOTAL"`. The state-by-farm-size and county-by-farm-size counts were obtained by setting `domain_desc`≠`"TOTAL"`. There are several additional conditions that had to be filtered to get the appropriate counts. These conditions only include rows where (i) `domaincat_desc` text includes text "0 HEAD" or "1 OR MORE HEAD"; (ii) (`domain_desc` includes text such as "inventory of milk/beef cows" or "inventory of cows" and (iii) `class_desc` is either "INCL CALVES" or "EXCL COWS").

Poultry. Poultry data has many subtypes. In AgCensus each subtype is organized as separate livestock under the group poultry, i.e., `group_desc="POULTRY"`. We remapped this data by creating

a new livestock called `poultry` and mapping all livestock under it to distinct subtypes. The counts for chickens were obtained by setting `commodity_desc="CHICKENS"`. There are four subtypes corresponding to chickens: layers (`ckn-layers`), broilers (`ckn-broilers`), pullets (`ckn-pullets`), and roosters (`ckn-roosters`). For layers, state-level counts of farms by category is present. However, we have ignored this as the corresponding head counts are absent. The counts of other poultry such as turkeys and ducks was obtained by setting `group_desc="POULTRY"` and `commodity_desc`$\neq$`"CHICKENS"`.

Hogs. The counts were obtained by setting `commodity_desc="HOGS"`. The rest are similar to cattle.

Sheep. The counts were obtained by setting `commodity_desc="SHEEP"` and `class="INCL LAMBS"`. No subtypes were considered. The remaining details are similar to cattle.

Aligning farm categories. Farm category specifications are more refined at the state level than at the county level. For example, at the county level, the largest category is 500 or more, whereas at the state level, there are categories such as '1000-2499' and '2500 or more'. We map all state-level categories to county-level categories by either aggregating the counts in the additional categories or simply removing the categories if they had already been accounted for at the county level.

C Gridded Livestock of the World

The Gridded Livestock of the World (57) dataset provides a gridded distribution of livestock abundance at 5 arc minute resolution. The gridded distribution data was constructed by combining detailed livestock census statistics mined from various sources using random forest models with predictors of the following types: land use, human population, travel times, vegetation, and climate. Unsuitable areas such as water bodies and core urban centers are identified using land cover and human population density information. More details are provided in Gilbert et al. (31).

The data is available in the following format. Each grid cell is identified by a cell ID denoted by a pair of integers, (`x`, `y`). For each grid cell, if a livestock abundance is available, the livestock type and value are provided. Among the livestock types or species provided, we considered cattle, buffaloes, sheep, pigs, chickens, and ducks. Buffaloes were mapped to cattle, and ducks and chickens were mapped to poultry. We did not consider goats and horses. No information on livestock subtypes is provided.

We identified the cells corresponding to the contiguous US and associated them with their respective state and county FIPS codes. The cells are denoted by C_j while the abundance value of a livestock type is denoted by Q_j. The notation does not include livestock type as each type is processed independently.

D Filling gaps

As mentioned above, some head counts are missing in all count types: state-total, state-by-farm-size, county-total, and county-by-farm-size. We use a combination of integer linear programs (ILPs) and iterative proportional fitting (IPF) following Burdett et al. (13) to address these omissions. For cattle and poultry, gaps are filled for each subtype, while for hogs and sheep they are filled for the livestock type.

We use the integer program Algorithm 1 to fill missing data for the following types of counts: state-total, state-by-farm-size, and county-total. It takes as input all the known counts, sum of all the counts, and bounds on the unknown counts, and distributes equitably the heads that are unaccounted for to all entities for which the counts are missing. It respects the bounds provided as input.

Algorithm 1 FillGaps integer program to fill missing gaps in state and county totals and state counts by farm size.

Input: No. of unknown quantities m, their sum total T, and bounds $\big((L_1, U_1), (L_2, U_2), \ldots, (L_m, U_m)\big)$, where $L_i \leq U_i$, $1 \leq i \leq m$.

Output: Assignment of values to the m unknown quantities. (The constraints to be satisfied by the unknown quantities are provided below.)

Variables

x_i, $i = 1, \ldots, m$ // Variables for unknown quantities

$\lambda_0 \geq 0$ // Variable for equitable distribution

Constraints

$L_i \leq x_i \leq U_i$, $1 \leq i \leq m$ // Bounds on unknown quantities based on input data

$\sum_i x_i = T$ // Sum of unknown quantities should be T

$x_i - L_i \leq \lambda_0$ // Bound the difference between the assigned quantities and corresponding lower bounds

Set Objective: Minimize λ_0

return $(x_1, x_2, \ldots, x_m)$

Now, we describe the process used to fill missing data for each count type in the order in which they are processed.

1. **State-total head counts.** The total head count here corresponds to the country head count that is available for every livestock subtype. There are potentially four sources that can be used to derive bounds. If farm counts per farm-size category are given at the state level, an initial set of lower and upper bounds can be derived as follows: $W_i^{\min} F_{i\gamma} \leq H_{i\gamma} \leq W_i^{\max} F_{i\gamma}$. The lower bounds are refined by using the available head counts for each farm size category. Similar refinement can be done from counts per farm size at the county level. The sum of the lower bounds across farm categories provides a lower bound for the state total. Finally, if county totals are provided for some counties of the state, their sum provides another lower bound. We set the final lower bound to be the maximum of bounds obtained as above. This data is fed to FillGaps to obtain estimates for the missing counts.
2. **State-by-farm-size head counts.** The total head count here corresponds to state total which is either available or estimated as above for every livestock subtype. We use the same approach as above by first deriving bounds based on the number of farms in each category and then refining the lower bound using county-by-farm-size head counts. This data is fed to FillGaps to obtain estimates for the missing counts.
3. **County-total head counts.** The total head count here corresponds to the total head count in the state to which the county belongs, which is either available or estimated as above for every livestock subtype. If farm counts are provided for each farm category, then we use it to derive the initial bounds. This is further refined if county-by-farm-size head counts are provided. This data is fed to FillGaps to obtain estimates for the missing counts.
4. **County-by-farm-size head counts using IPF.** To fill gaps in county-by-farm-size counts, we follow the methodolgy of Burdett et al. (13). They apply IPF (67, 68) for the case of hogs to estimate these counts. Here, we give an overview of the method and refer to Burdett et al. (13) for details. The processing is done per state and subtype. In the IPF process, the objective is to impute missing values in a given matrix given row and column totals. In this case, the data matrix consists of county-by-farm-size counts with counties as rows and farm size categories as columns. Note that, at this stage, both county totals and state-by-farm-size counts are available either from data or by estimation. Unknown values in the matrix are seeded with the product of the average size of the corresponding category and the number of farms in that category.

E Generation of farms

Objective. We are given farm categories $i = 1, \ldots, \ell$. Let F_i and H_i denote the number of farms and livestock heads in category i. For $k = 1, \ldots, \ell$, let $F_{\gamma k}$ and $H_{\gamma k}$ denote the number of farms

and heads corresponding to category k with respect to subtype γ (see Table 3). The objective is to find an assignment of farms whose farm counts and composition respects these counts. Note that the categories are pairwise disjoint and cover the entire range. In addition to head counts of subtypes, we are also given counts of the livestock type as well by farm size. Since the IPF procedure does not account for these counts, there could be differences between this data and the counts obtained by aggregating farm subtype counts in our assignment. Our optimization objective is a linear combination of many parameters as discussed below.

(a) To choose an assignment with minimum discrepancy with respect to known counts, we introduce a parameter λ_1.

(b) The parameter λ_2 in the minimization objective represents the largest number of subtypes in a farm.

(c) The minimization objective includes ℓ parameters, denoted by λ_{3i}, $1 \leq i \leq \ell$. The purpose of parameter λ_{3i} is to ensure that the population of all the subtypes in any farm of category i is close to the average value for that farm category. The purpose of minimizing these parameters is to obtain an equitable distribution of the population across all farm categories.

(d) The parameter λ_4 in the minimization objective is used to ensure that the sum of the head counts of subtype γ over all the farms assigned category k is close to the given head count $H_{\gamma k}$.

The optimization objective combines the above parameters into a linear function using appropriate scaling constants. These scaling constants ensure that parameters with larger values have larger penalties. As a consequence, the solver will aggressively minimize parameters with larger values compared to ones with smaller values.

Algorithm 2 GenFarms. Integer linear program to generate farms consistent with input counts of farms and head counts.

Input: County-by-farm-size head and farm counts as in Table 3.
Output: Farms with head counts of each subtype.

Variables

For each farm f in category i, $h_{if\gamma}$ corresponds to the population of subtype γ in that farm For each farm f in category i, $x_{if\gamma k}$ indicates whether (f, i) belongs to category k w.r.t. subtype γ For each farm f in category i, $y_{if\gamma k}$ indicates whether $h_{if\gamma} = 0$ For each farm f in category i, $z_{if\gamma k} = h_{if\gamma}$ if $h_{if\gamma}$ belongs to the category k with respect to subtype γ. Otherwise, it is 0.

Variables for minimization objectives: λ_1 (alignment with known total population within each farm size category), λ_2 (minimize number of subtypes per farm), $\lambda_{3i},\ i \in [1, \ell]$ (equitable distribution of population in each farm size category), and (alignment with subtype population within each farm size category) λ_4.

Constraints

Let M be a suitably large constant `// Population and farm size constraints`

$W_i^{\min} \le \sum_\gamma h_{if\gamma} \le W_i^{\max}$ `// Category farm size constraint`

`// Subtype farm category constraints: Farm counts.`

$i \in [1,\ell], f \in [1, F_i], \gamma \in \Gamma, k \in [1,\ell],\ h_{if\gamma} \ge W_k^{\min} - (1 - x_{if\gamma k}) \cdot M$ `// Lower bound`

$i \in [1,\ell], f \in [1, F_i], \gamma \in \Gamma, k \in [1,\ell],\ h_{if\gamma} \le W_k^{\max} + (1 - x_{if\gamma k}) \cdot M$ `// Upper bound`

$i \in [1,\ell], f \in [1, F_i], \gamma \in \Gamma, k \in [1,\ell],\ h_{if\gamma} \ge 1 - y_{if\gamma k}$ `// Lower bound when subtype count is 0`

$i \in [1,\ell], f \in [1, F_i], \gamma \in \Gamma, k \in [1,\ell],\ h_{if\gamma} \le (1 - y_{if\gamma k}) \cdot M$ `// Upper bound when subtype count is 0`

$i \in [1,\ell], f \in [1, F_i], \gamma \in \Gamma,\ \sum_k x_{if\gamma k} + y_{if\gamma k} = 1$ `// Farm in exactly one category`

$\gamma \in \Gamma, k = [1,\ell],\ \sum_{f,i} x_{if\gamma k} = F_{\gamma k}$ `// Count equals total farms in that category`

`// Subtype farm category constraints: Population counts`

$i = [1,\ell],\ f = [1, F_i],\ \gamma \in \Gamma,\ k = [1,\ell],\ z_{if\gamma k} \le h_{if\gamma}$ `// Upper bound`

$i = [1,\ell],\ f = [1, F_i],\ \gamma \in \Gamma,\ k = [1,\ell],\ z_{if\gamma k} \le x_{if\gamma k} \cdot M$ `//` $x_{if\gamma k} = 0 \Rightarrow z_{if\gamma k} = 0$

$i = [1,\ell],\ f = [1, F_i],\ \gamma \in \Gamma,\ k = [1,\ell],\ z_{if\gamma k} \ge h_{if\gamma} - (1 - x_{if\gamma k}) \cdot M$ `//` $x_{if\gamma k} = 1 \Rightarrow z_{if\gamma k} = h_{if\gamma}$

$\gamma \in \Gamma,\ k = [1,\ell],\ \left| \sum_{f,i} z_{if\gamma k} - H_{\gamma k} \right| \le \lambda_4$ `// Count must be close to` $H_{\gamma k}$

`// The assignment should be such that it is as close a match to the total population distribution in each category,` H_i.

$i = [1,\ell],\ \left| \sum_{f,\gamma} h_{if\gamma} - H_i \right| \le \lambda_1.$

`// Subtype distribution: Minimize number of subtypes per farm`

$i = [1,\ell],\ f = [1, F_i], \sum_{\gamma,k} x_{if\gamma k} \le \lambda_2$

`// Equitable distribution in each category.`

$i = [1,\ell],\ a_i = \sum_{f,\gamma} h_{if\gamma} / F_i,$ `// Average population in each farm category`

$i = [1,\ell],\ f = [1, F_i],\ \left| \sum_\gamma h_{if\gamma} - a_i \right| \le \lambda_{3i}$

Set objective. Minimize $\lambda_1 + (H+1) \cdot \lambda_2 + (|\Gamma|+1)(H+1) \cdot \sum_i \lambda_{3i} + (|\Gamma|+1)(H+1)^2 \cdot \lambda_4$.

return $(h_{if\gamma} \mid i = [1,\ell], f = [1, F_i], \gamma \in \Gamma)$

Implementation notes. The algorithm was run for each county–livestock instance in parallel on a HPC cluster. For faster convergence to a solution, we set the MIP gap (which refers to the percentage difference between the current best feasible solution and the best known bound on the optimal objective value) to 0.1% of the total head count for each instance.

F Farms to cells

Objective. We are given N_f farms with number of heads in each farm f_i denoted by P_i, $i = 1, \ldots, N_f$, and N_c cells with number of heads in each cell C_j denoted by Q_j, $j = 1, \ldots, N_c$. The objective is to assign to each farm f_i a cell C_j such that $\max_j \left| \sum_{i,\delta_i=j} P_i - Q_j \right|$ is minimized, where $\delta_i = j$ if and only if f_i is assigned cell C_j.

Algorithm 3 FARMSTOCELLS. Integer linear program to generate farms consistent with input counts of farms and head counts.

Input: Farms f_i, $i = [1, N_f]$ with total head count P_i, cells C_j, $j = [1, N_c]$ with head count Q_j.
Output: Assignment of each farm to a cell.

Variables

For $1 \leq i \leq N_f$ and $1 \leq j \leq N_c$, x_{ij} indicates whether farm f_i is assigned to cell C_j: $x_{ij} = 1$ if f_i is assigned cell C_j; otherwise, $x_{ij} = 0$ For $1 \leq i \leq N_f$ and $1 \leq j \leq N_c$, $h_{ij} = P_i$ if $x_{ij} = 1$, else 0 Let λ_4 be a positive integer variable that is equal to $\max_j \left| \sum_{i,\delta_i=j} P_i - Q_j \right|$

Constraints

`// Assign farm to a cell`

$1 \leq i \leq N_f, 1 \leq j \leq N_c, \; h_{ij} = x_{ij} \cdot P_i$ `// Contribution of a farm to cell population`

$1 \leq i \leq N_f, \; \sum_j x_{ij} = 1$ `// Each farm belongs to exactly one cell`

$1 \leq j \leq N_c, \; \left| \sum_i h_{ij} - Q_j \right| \leq \lambda_5$ `// Farm assignment should align with cell capacities`

Set objective. Minimize λ_5.

return $(h_{if\gamma} \mid i = [1, \ell], f = [1, F_i], \gamma \in \Gamma)$

Implementation notes. The algorithm was run for each county–livestock instance in parallel on a HPC cluster. For faster convergence to a solution, we set the MIP gap to 0.01% of the total head count for each instance.

G Additional results

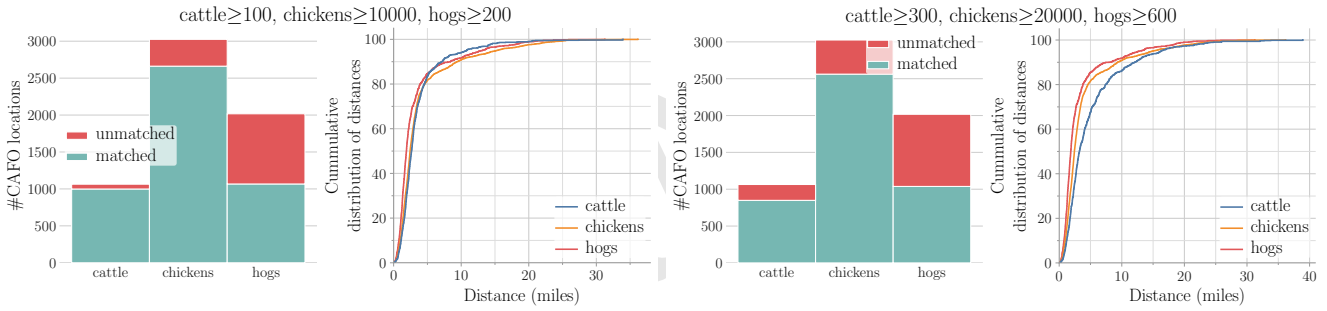


Fig. S1. Analysis of mapping CAFO locations by livestock type to farms from FIELD. The two plots correspond to two sets of thresholds for choosing the farms to compare with. The second set corresponds to larger farms compared to the first. In each plot, the first subplot shows how many CAFO locations were matched. The second subplot provides the cumulative distribution of the distances (in miles) between matched pairs of CAFO locations and farms.

B Wild Birds

Our processing pipeline to extract abundance data involves the following steps:

- Coordinate System Conversion: We transform the data from its original World Eckert IV equal-area projection (ESRI:54012) to a standard geographic coordinate system (EPSG:4326) to ensure compatibility with other geospatial datasets in our study.
- Spatial Sampling: To balance spatial resolution with computational efficiency, we sampled the data at regular intervals consistent with the GLW grid cell dimensions. This sampling strategy preserves the overall spatial patterns while reducing the dataset to a manageable size.
- Temporal Resolution: We maintained the weekly temporal resolution provided by eBird, allowing for detailed analysis of seasonal migration patterns.

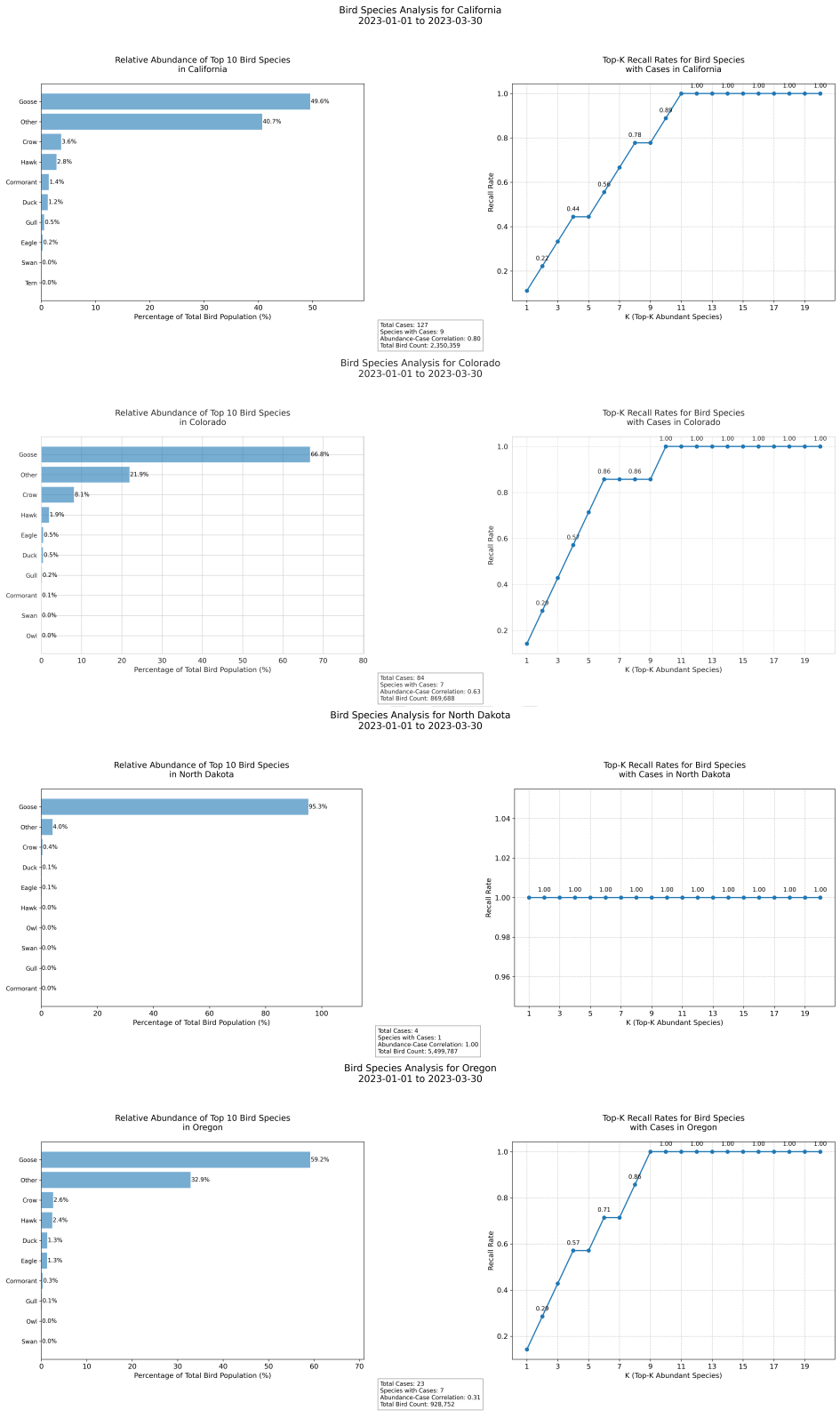


Fig. S2. Bird Abundance and H5N1 incidence data from January to March 2023.

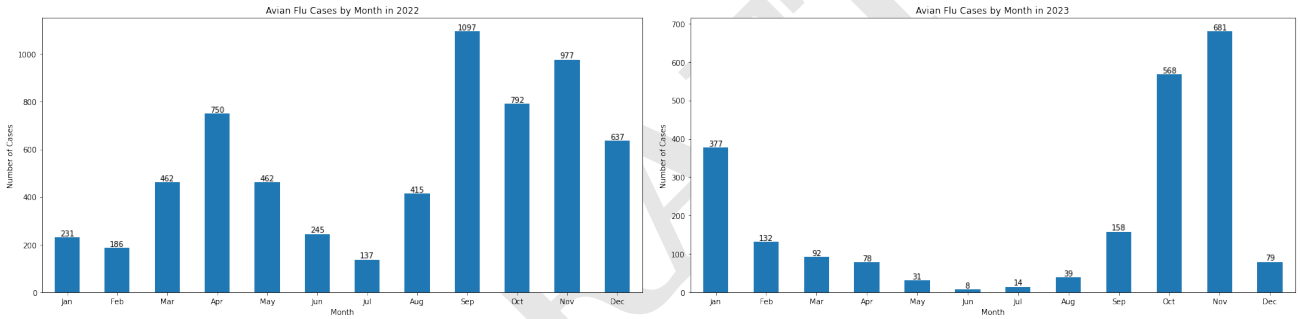


Fig. S3. H5N1 incidence in wild birds across different months in 2022 and 2023. Case abundance is higher in fall and winter months, which can be attributed to breeding and migration patterns and viral transmissibility in colder months.

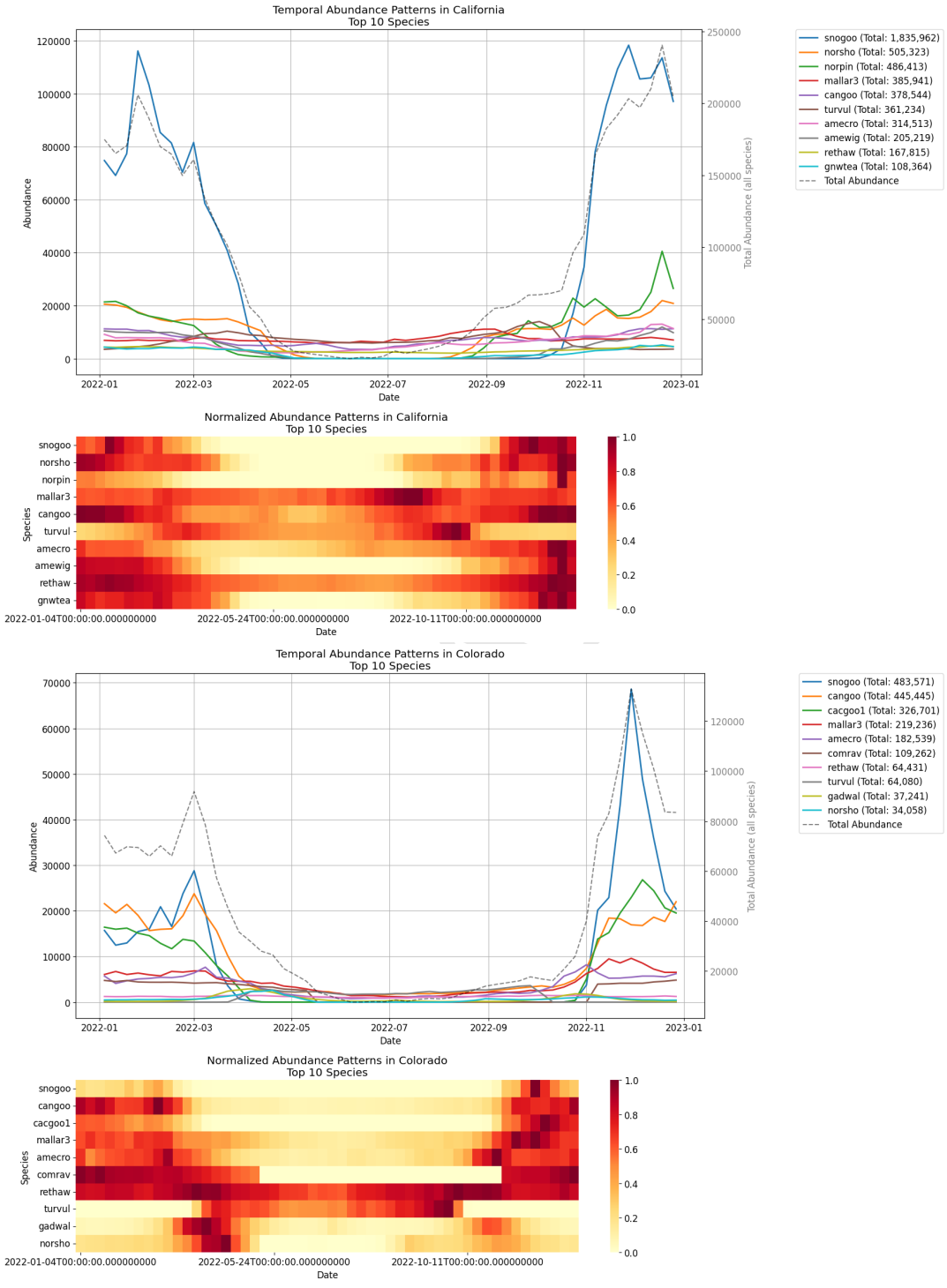


Fig. S4. Temporal Abundance Pattern of Bird Species across states. We observe heterogeneity of species abundance and demographics across different states, throughout the year. Abundance varies across seasons due to migration and breeding in different geographies.

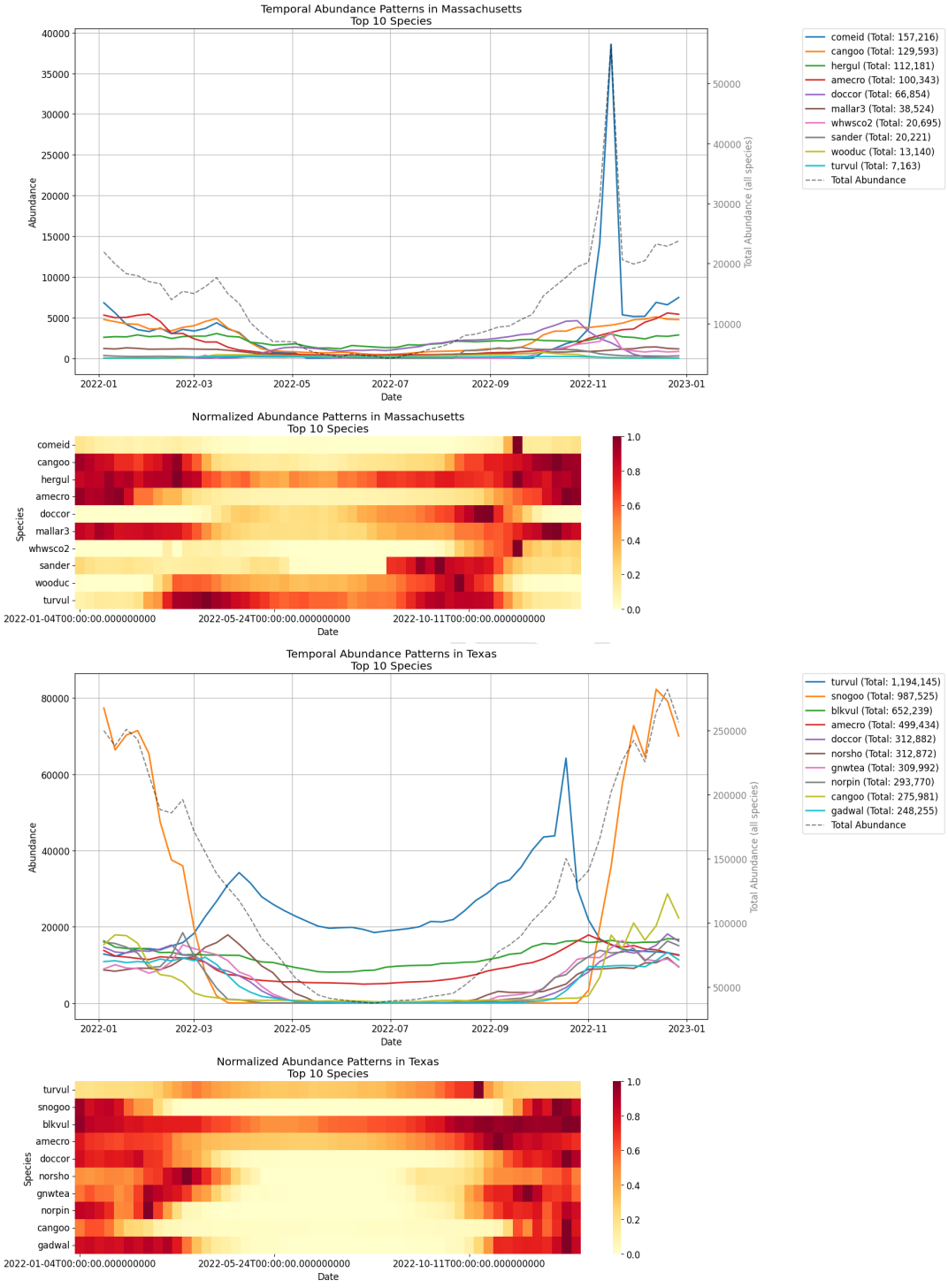


Fig. S5. Temporal Abundance Pattern of Bird Species across states. We observe heterogeneity of species abundance and demographics across different states, throughout the year. Abundance varies across seasons due to migration and breeding in different geographies.

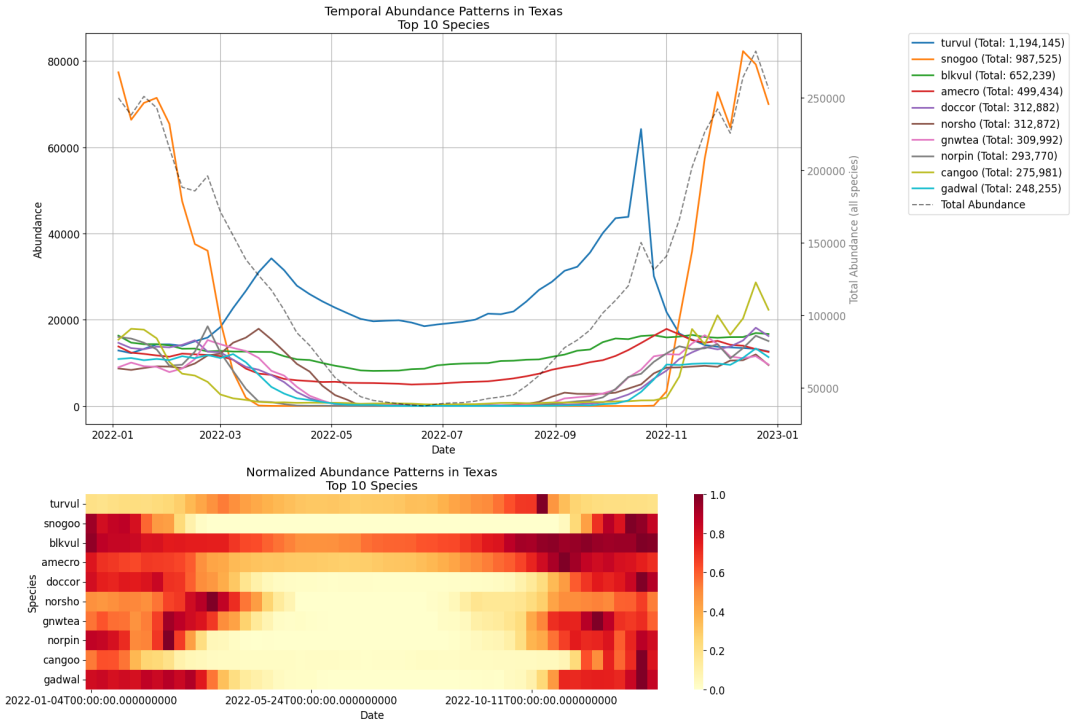


Fig. S6. Temporal Abundance Pattern of Bird Species across states. We observe heterogeneity of species abundance and demographics across different states, throughout the year. Abundance varies across seasons due to migration and breeding in different geographies.

C Additional results for FIELD

A Farm generation and cell assignment verification

Here, we evaluate the construction process for the livestock layers (outlined in Figure 3a) by comparing the constructed layers with the parent datasets. We compared the total head counts from the assigned farms with the corresponding counts from the AgCensus data. The aggregation was done at the state level. The absolute relative difference is plotted in Figure S7a. The differences between the modeled head counts and AgCensus are caused by the assignment of head counts to areas where AgCensus head counts were unreported, and by subsequent adjustment of counts for consistency across farm sizes (Section D). We observe that the relative difference is below 1% in most instances, barring a few outliers. Also, the larger the population of a subtype, the smaller the relative difference. Figure S7b (bottom row) shows the distribution of the livestock population into farms. The largest cattle farms are assigned around 100,000 heads, while chicken farms (corresponding to subtypes layers and broilers) can be assigned up to 5 million heads.

We now analyze the performance of two constrained optimization algorithms (namely, GenFarms and FarmsToCells) used in this work. These algorithms are based on ILP formulations in which optimization objectives and the constraints are chosen carefully based on the available data and the necessary outcomes. Detailed descriptions of these algorithms are provided in the methods and supplement.

We first analyze the performance of GenFarms. In this algorithm, the minimization objective includes the parameter λ_1, which bounds the error in livestock totals by farm size category between the reported value and our assignment. (This error is due to the gap-filling step carried out by the IPF process.) Our results in Figure S8a show that this discrepancy is low across livestock types, which implies that the assignment is close to the known total head counts. Next, we analyze the performance of FarmsToCells in two ways, evaluating how well the assignment of farms aligns with the GLW dataset. In Figure S8b, we plot the parameter λ_5 (see supplement A4), which corresponds to the maximum absolute difference between the head counts corresponding to our assignment and GLW cells. The second plot in Figure S8b measures the agreement of our aggregated head counts with GLW data using Pearson's correlation coefficient. For all livestock types except poultry, the correlation is, on average, around 0.75. However, there are instances which are negatively correlated with GLW. The reason for this behavior is that larger farm sizes make it more difficult to align the head counts with cell capacities. In general, poultry distribution is weakly correlated with GLW.

Livestock worker population

Here, we compare the livestock worker population in FIELD with counts obtained from the Quarterly Census of Employment and Wages data corresponding to the year 2023 (Table 1, BLS). We choose the population associated with livestock-related occupations (SOCP 4520XX) or industry (NAICS 112) as representing livestock workers. Also, considering that this population is not time varying, we analyze BLS for seasonal variations in the livestock worker counts. We note that BLS only counts workers covered under unemployment insurance, due to which farm owners, self-employed workers, and many workers (e.g., undocumented workers) are potentially excluded from the count.

We recall that the livestock worker distribution in FIELD is derived from a digital twin of the US population (Table 1, USPop). This distribution is imputed from the American Community Survey (ACS) 5-year Public Use Microdata Sample (PUMS). The total count of livestock workers in FIELD is 704,126, while the total number of such individuals in the PUMS data is 42,233, which is roughly 6% of the total worker population. This is consistent with the fact that the PUMS represents approximately 5% of the US population.

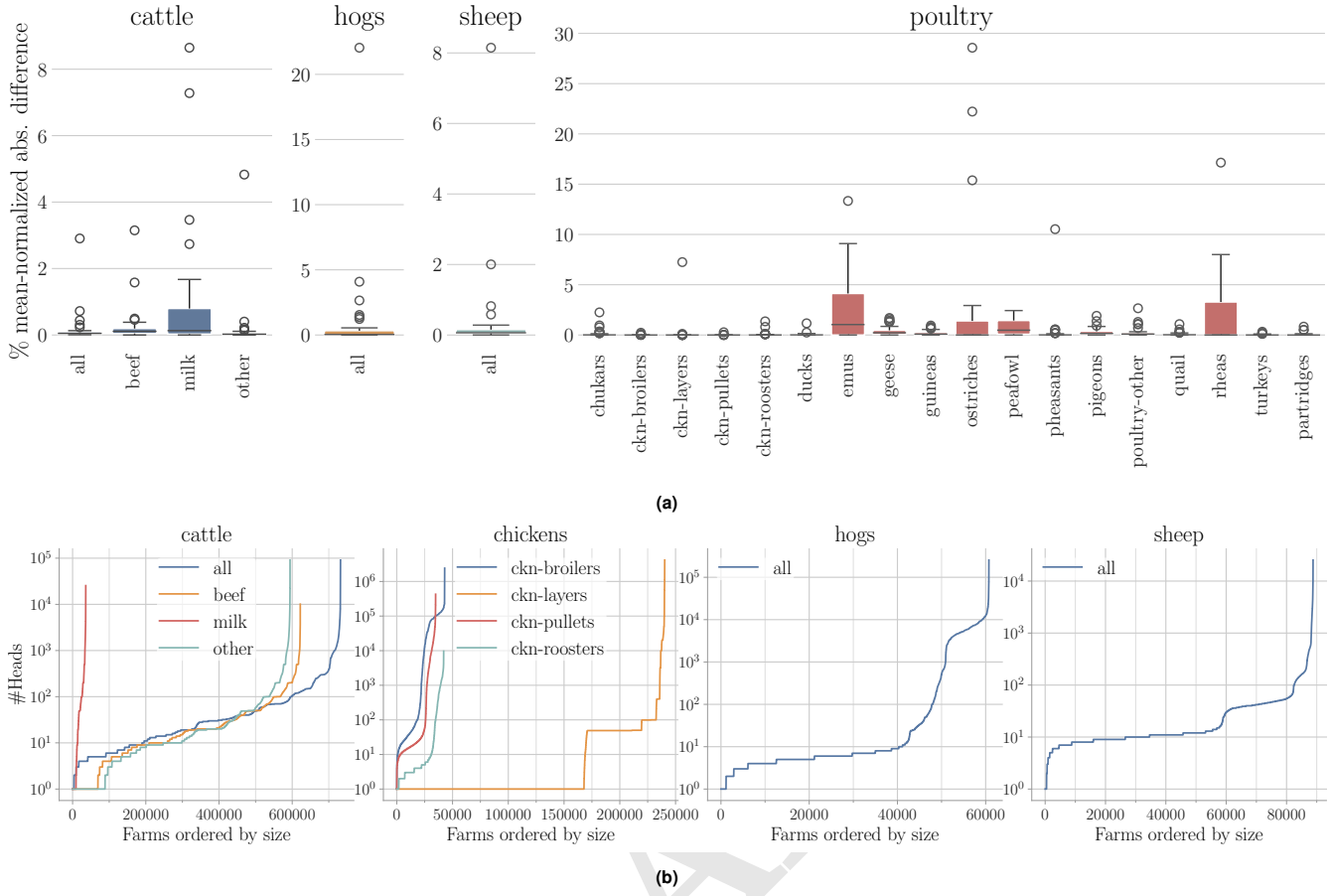


Fig. S7. Alignment of the livestock layers with AgCensus and GLW datasets. (a) Head counts of assigned farms are compared with AgCensus. We have a plot for each livestock type with subtypes on the x-axis and percentage mean-normalized absolute relative difference of state totals from the census and FIELD on the y-axis. (b) The distribution of livestock populations among farms ordered by farm size. The y-axis corresponds to farm size. Separate plots for subtypes are shown for cattle and poultry.

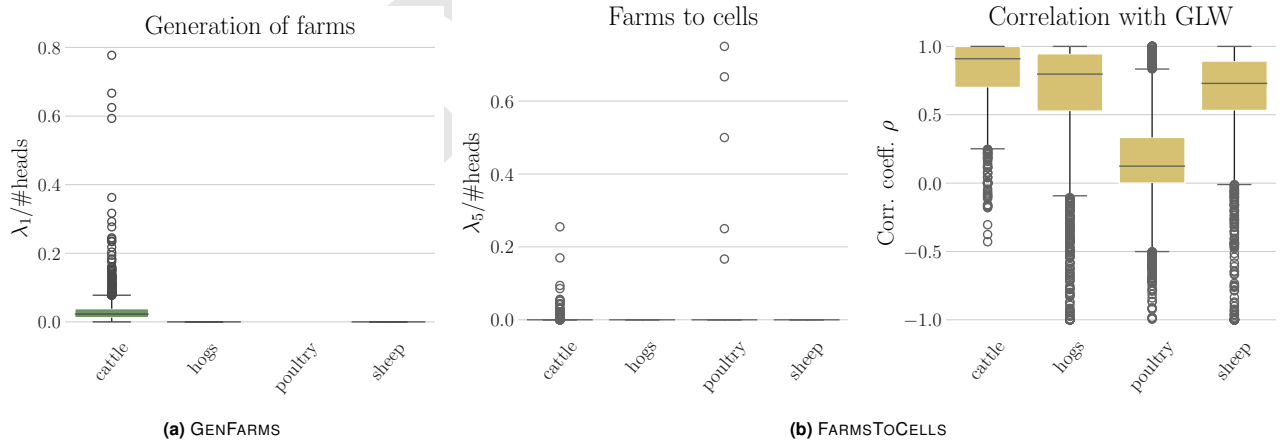


Fig. S8. (a) Analysis of the GenFarms algorithm: We plot the value of the parameter λ_1 (see supplement) relative to the number of heads. This parameter is the maximum absolute difference between the number of heads in AgCensus to that in the generated farms at the county level for each farm size category. Lower is better. Each value in the box plot corresponds to a county. We also plot this value for a restricted set of instances where the county totals are known. (b) Analysis of the FarmsToCells algorithm: Both plots indicate the agreement of the farm assignment with GLW data. The first plot corresponds to parameter λ_5 (see supplement) for county–livestock instances relative to the number of heads. This parameter is the maximum absolute difference between assigned head counts in a cell and its corresponding GLW value. Lower is better. Using the Pearson correlation coefficient, cell-level head counts aggregated from our farm assignment are compared with GLW head counts; higher is better.

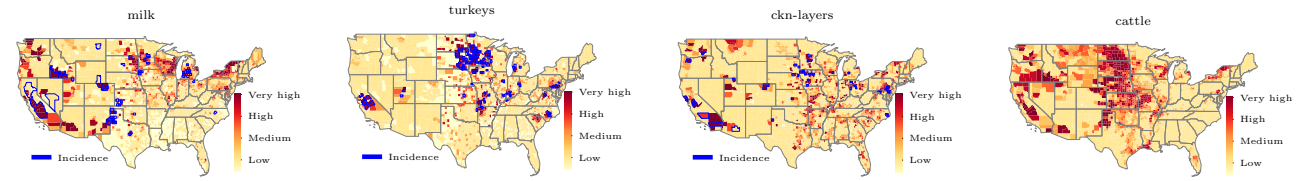
The plot in Figure 3d shows the difference between our data and the BLS data in counts of livestock worker population by county and state, respectively, for 2769 counties that are common to both the datasets. We note that, for more than 96% of the counties, the synthetic population counts

exceed that of BLS. This is expected as BLS excludes a significant population of farm workers as mentioned above. In many cases, the count is zero. There are around 105 counties for which the population in FIELD is less than that of BLS. However, the difference in this case is usually very small compared to many of the remaining instances where the counts in FIELD far exceed BLS. It is possible that a significant portion of the farm worker population in these counties did not participate in the census.

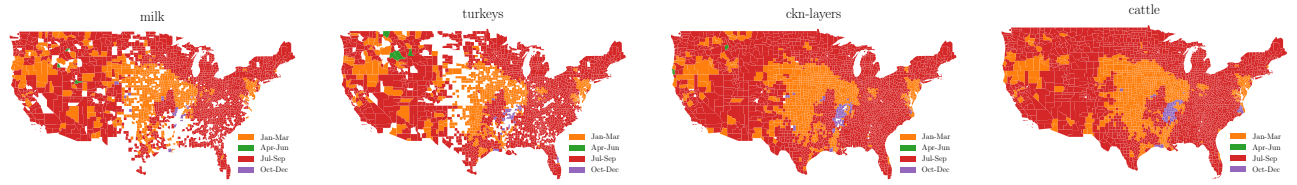
Figure 3e shows a comparison between county-level farm counts and the livestock worker population. Due to missing information about mixing livestock in farms, our total farm count is higher than the total mentioned in AgCensus. Hence, we only considered farms with head counts of at least 100. Generally, for counties with higher farm counts, the number of workers is higher. But there is a wide spread in the number of workers for a fixed farm count. Since counts can depend on farm sizes and livestock types, without additional information it becomes almost impossible to compare the two quantities in further detail.

Analysis of BLS shows little variation in the county-level counts of livestock workers across the year. In Figure 3f, we have plotted a scatter plot of the coefficient of variation for the four quarters of year 2023 with respect to the mean number of workers in the county. With the exception of two outliers, counties with very large coefficients of variation have very few livestock workers. For the two outlier counties, there is at least one quarter with zero count, which could be attributed to missing data.

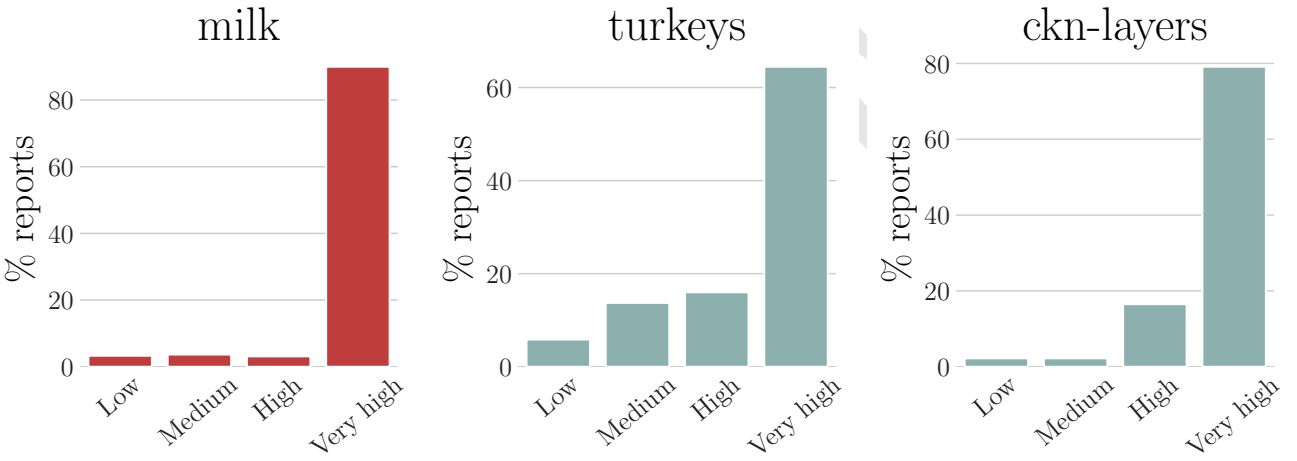
D Additional results for risk estimation



(a) Elevated Risk Level

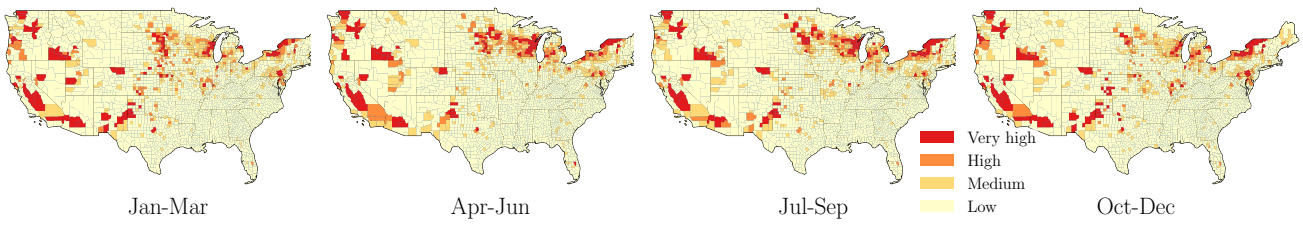


(b) Period of peak risk

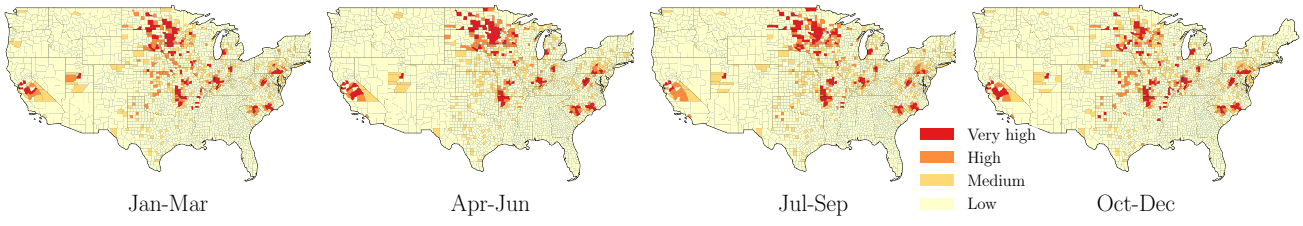


(c) Recall performance.

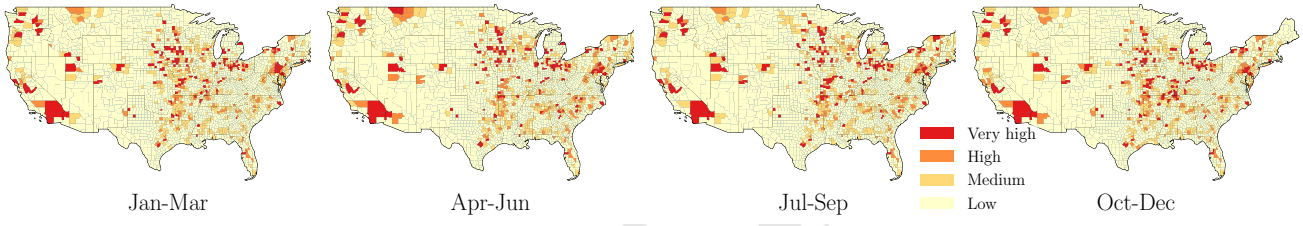
Fig. S9. Summarizing colocation risk. (a) Aggregated risk levels across counties using stable sorting: Counties are ordered by the highest severity of risk among all quarters and then among all counties with the same highest severity, they are ordered by the number of quarters in which this severity occurs. We continue this across all severities. (b) Peak time of risk: The quarter with the highest risk for each county is plotted. (c) Recall performance when colocation risk is compared with ground truth.



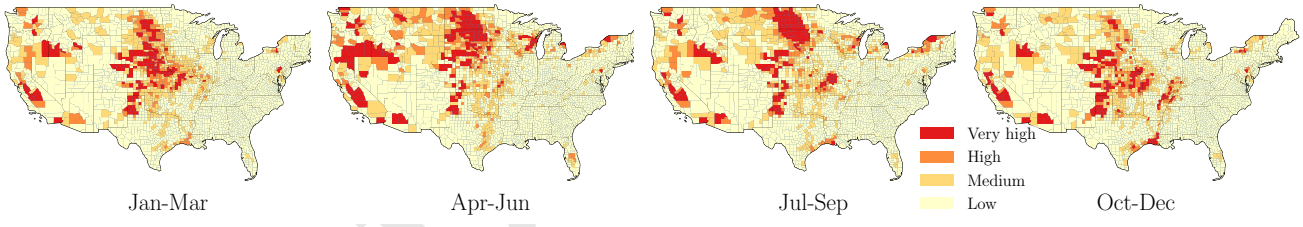
(a) Dairy cattle



(b) Turkeys



(c) Chicken layers



(d) Cattle

Fig. S10. Colocation risk for the four quarters of the year for several affected livestock subtypes.

Table 4. We visualize risk persistence and observed outbreaks in counties that consistently showed very high risk (95th percentile) across time periods in 2024 for dairy cattle. In this table, Avg. Position indicates the mean ranking of risk level across appearances, with lower numbers indicating higher risk (e.g., position 1 means highest risk in that period). Counties appearing in all time periods demonstrate persistent risk factors throughout the year, regardless of seasonal variations. We validate our results against real-world outbreak instances from WHO and CDC reportings.

County	State	Appearances	Avg. Position	Variance	Known Outbreak
Merced	CA	4	1.5	0.2	Yes
Tulane	CA	4	3.0	1.5	Yes
Weld	CO	4	3.5	10.2	Yes
San Joaquin	CA	4	4.2	2.7	Yes
Kern	CA	4	5.8	0.7	Yes
Fresno	CA	4	7.5	6.8	Yes
Maricopa	AZ	4	8.0	3.5	Yes
Kings	CA	4	9.2	2.7	Yes
Stanislaus	CA	4	9.8	5.2	No
Lancaster	PA	4	13.5	58.8	Yes
San Bernardino	CA	4	14.8	6.2	Yes
Box Elder	UT	4	15.5	6.8	No
Yakima	WA	4	16.8	9.2	No

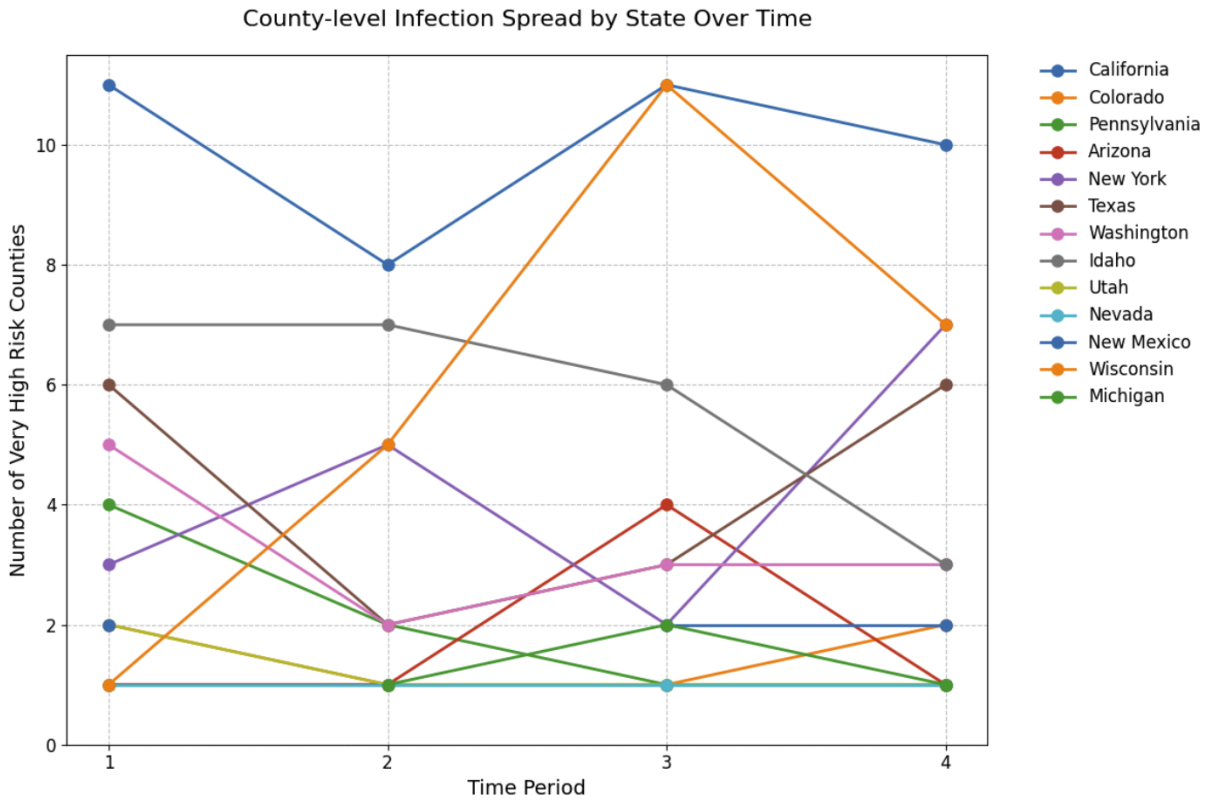


Fig. S11. Time series showing the number of very high risk counties (95th percentile) in each state across four periods in 2024: January-February, March-May, June-August, and September-December. Only states that had at least one very high risk county in three or more periods are displayed. The seasonal variation across states can be explained due to the migration patterns of wild birds. While some states, like Nevada and Utah, had fewer very high risk counties, their persistent presence across multiple periods suggests sustained elevated risk in specific regions throughout the year.

E DiTTO: Web Portal and Data Availability

The synthetic spatiotemporal dataset of interacting livestock and wild bird populations is designed to be easily accessible to and usable by researchers, policymakers, and modelers interested in studying avian influenza dynamics. We provide an interactive visualization dashboard, Digital Twin for Transboundary OneHealth (DiTTO) (shown in Figure S12) to make the data available.

The dashboard user interface is divided into three sections: a navigation bar, where users can indicate which data layers they are interested in viewing by population type and relevant subtypes (under Heatmap Measure), and even to pinpoint specific regions. On the lower left side of the screen is a heatmap where users can view where the selected population type is prevalent; users can view that data at US state resolutions, or click on the map to view county resolution heatmaps for the selected state. On the lower right side of the user interface is a data table where users can view the actual counts across all of the subtypes for the selected population type and region(s). Users can download the datasets in two ways from the web portal: (i) they can click on the Download Table button above the data table to download the queried rows displayed in the data table, or they can click on the Download Layer button on the map to download the complete grid-level layer data for the selected population type.

In short, DiTTO provides the following key features to make the datasets accessible to its users.

- Interactive maps showing the distribution of livestock, farms, wild bird populations, human populations, and processing centers at the state and county levels.
- The ability to search for specific regions for easier comparison.
- Filters for selecting specific regions, time periods (for wild birds), and livestock types.
- Download functionality for either the complete layer or for a subselection of that layer.

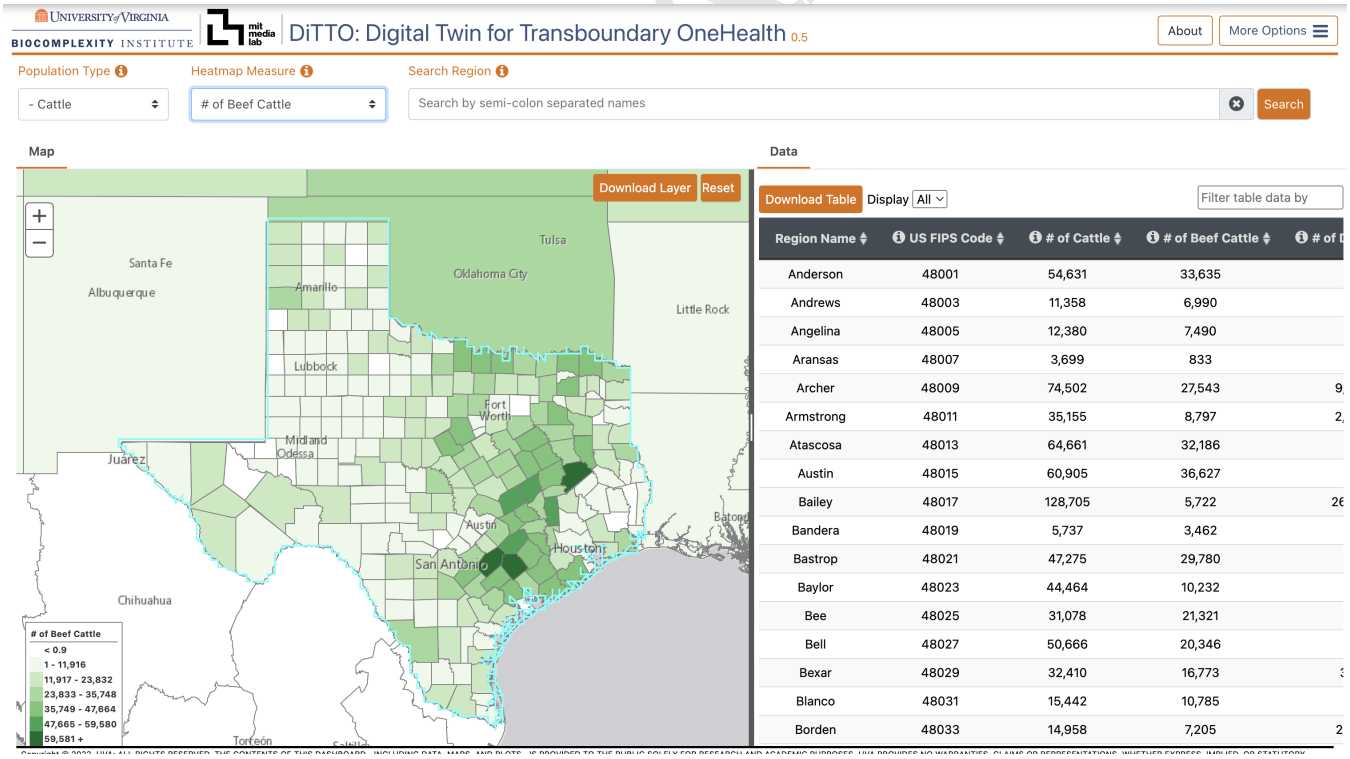


Fig. S12. The interactive visualization dashboard allows users to explore the livestock and avian populations in an interactive, spatiotemporal way. It is available at https://ditto.bii.virginia.edu.