# Extended Methods for

## Dynamic Risk Maps Predict Highly Pathogenic Avian Influenza Hotspots Across North America

**Authors:**
M. Bakheet[1,2], O. Babasola[1,2], C. Næsborg-Nielsen[1,2], S. Subedi[1,2,3], MHM Mubassir [1,2,3], S. Peng[1,2], T. Rajamand[1,2,3]and J. Bahl[1,2,3,4]
[1]Center for Ecology of Infectious Diseases, University of Georgia, Athens, GA, United States
[2]Department of Infectious Diseases, University of Georgia, Athens, GA, United States
[3]Institute of Bioinformatics, University of Georgia, Athens, GA, United States
[4]Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA, United States

**Corresponding author: justin.bahl@uga.edu**

### Extended Materials and Methods

Study area and period

All analyses were performed on a North America grid in EPSG:4326 covering -170° to -50° longitude and 10° to 85° latitude. Country boundaries for the United States, Canada, and Mexico (Natural Earth) defined the analysis mask. To ensure consistency, all spatial data were reprojected to WGS84 and standardized to a common grid of ~1 km resolution. The study period covers 1991-2025 to capture long-term migratory patterns as well as the recent H5N1 clade 2.3.4.4b expansion.

Data sources and processing

We assembled ecological, demographic, and outbreak data from multiple sources (Extended Data Table 1).

Wild bird observations were obtained from eBird[50] and the Global Biodiversity Information Facility (GBIF)[51], livestock densities from the FAO Global Livestock Environmental Assessment Model (GLIMS)[52], human population from GPWv4[53], climate predictors from WorldClim v2.1[54], and confirmed outbreaks from FAO EMPRES-i[55], WOAH WAHIS[56], and CFIA NEOC[57]. Additional predictors included MODIS land cover and vegetation indices[58], NASA distance-to-coast, and WorldClim digital elevation models (DEM)[54].

We filtered waterfowl and other susceptible migratory species, excluding taxa with fewer than 150 georeferenced records. This yields 126 species for analysis (Extended Data Table S2). To minimize multicollinearity among environmental predictors, we screened all 19 bioclimatic variables and retained seven with low inter-correlation: mean diurnal temperature range (BIO2), temperature seasonality (BIO4), temperature annual range (BIO7), precipitation of driest month

(BIO14), precipitation seasonality (BIO15), precipitation of driest quarter (BIO17), and precipitation of coldest quarter (BIO19).

Species distribution modeling

To generate species distribution models (SDM)[59] for each species-month, occurrence data were paired with background (pseudo-absence) points sampled from environmentally complete cells. Equal number of presences-absences were used, and each point was labeled as 1 (presence) or 0 (absence). We trained Random Forest, XGBoost, logistic regression implemented in scikit-learn, LightGBM, and pyimpute. Predictors were z-standardized after mean AUC guided model choice. Hyperparameters were tuned by five-fold cross-validation, with the mean AUC guiding model choice.

The best algorithm for each species-month combination produced probability surface at 0.5° resolution. The resulting monthly species richness maps for waterbird communities across North America (Extended Fig. 1) reveal consistent ecological hotspots from January through December. High-richness areas recur in the Prairie Pothole Region, the Great Lakes basin, along the Atlantic Flyway (including the Delmarva Peninsula), and within key coastal and wetland complexes of the Gulf of Mexico and Pacific Flyway. These spatial and temporal patterns reflect the dynamic migratory strategies of waterbirds and delineate regions of persistent ecological importance for avian diversity that underpin subsequent risk-mapping analyses.

Predictions were then combined into monthly species richness layers (Extended Data Fig. 3) and a Temporal Co-occurrence Index (TCI) as follows

$$TCI = \frac{\mu}{\sigma} \times 100$$

where μ and σ are the pixel-wise mean and standard deviation of richness across 12 months. High TCI values identify where many species co-occur frequently through time.

Entropy-based metrics

To capture temporal dynamics in wild bird activity and community structure, we extended beyond species richness and the TCI by computing several entropy-based diversity measures that describe both the magnitude and evenness of species' seasonal occurrence. These metrics quantify the temporal distribution of habitat suitability across the 12 months of the year, where $p_m$ denotes the proportion of annual suitability attributed to month $m$ (where $\sum_{m=1}^{12} p_m = 1$). These metrics quantify how evenly habitat suitability is distributed across months, values close to 1 indicate long, even activity seasons, while lower values indicate strong seasonal peaks concentrated in few months.

1. Shannon Entropy[60]

Shannon entropy measures the uncertainty or information content of the monthly suitability distribution.

$$D_1 = exp(H)$$

This converts entropy into the effective number of equally active months, providing a more intuitive diversity scale. For instance, a value of $D_1 = 5$ indicates that the temporal activity is equivalent to having five months contributing equally to the annual suitability. Hill $q = 1$ therefore emphasizes common months while remaining sensitive to evenness.

1. Hill Number of Order 1 (Hill q = 1)[61]

The Hill diversity index of order 1 is the exponential of Shannon entropy**:**

$$H_{q=1} = -\sum_{m=1}^{12} p_m \log p_m$$

2. Inverse Simpson Diversity (Hill q = 2)[61]

This index emphasizes the dominant months in the temporal distribution and is less influenced by rare or weakly active periods:

$$D_2 = \frac{1}{\sum_{m=1}^{12} p_m^2}$$

Higher values indicate broader and more evenly distributed temporal activity, whereas lower values reflect dominance by a few peak months.

Together, these complementary metrics describe different facets of temporal diversity: Shannon entropy quantifies evenness on a logarithmic scale, Hill $q = 1$ translates that information into an intuitive count of effective months, and Hill $q = 2$ highlights the dominance structure of seasonal activity. Collectively, they provide a multidimensional view of host community dynamics relevant to viral persistence and transmission potential.

Static predictors: ingestion and scaling

Poultry, cattle, and human density rasters were log-transformed as log(x+1) and min-max scaled to [0,1], after reprojection/resampling. Land cover (EPSG:4326 NA-wide raster) was reduced to the six most frequent classes plus an "other" bin. Each class was encoded as a binary dummy layer.

Presence data and monthly split

Confirmed HPAI case records were imported from CSV files using a flexible field-detection routine that automatically recognized coordinate and date information even when column names or formats varied among data sources. Values were validated to ensure they fell within geographic bounds (-180 ≤ longitude ≤ 180; -90 ≤ latitude ≤ 90) and transformed to the WGS84 coordinate reference system.

Date fields were similarly standardized using a multi-format parser that could interpret ISO-8601 strings (e.g., "2024-03-15"), ordinal dates ("20240315"), or delimited text ("03/15/2024", "15-Mar-2024"). When multiple temporal fields were present (e.g., *observation_date*, *report_date*), the observation date was prioritized. From each valid date, the month (m) was extracted to preserve temporal resolution.

For each month, presence points were filtered to that period; if no confirmed cases were reported, all valid detections within the year were retained as a conservative fallback to maintain spatial coverage. This flexible detection and harmonization process ensured consistent geotemporal alignment of all case records, allowing seamless integration with monthly environmental predictors and species suitability layers for model training.

Pseudo-absence sampling and feature extraction

For month (m), we extracted predictors at all presence points and sampled an equal number of pseudo-absence locations, constrained to cells with complete predictor coverage. The final design matrix for month (m) consisted of: Response (factor cases $\in$ {absence, presence}), and Predictors (poultry, cattle, population; ten species layers for month (m); and land-cover dummies).

In our occurrence dataset, confirmed HPAI presences and pseudo-absence locations were distributed across a study area restricted to ecologically and anthropogenically relevant zones (Extended Data Fig. 2). Rather than using the entire continent, we defined the sampling mask by combining wild-bird diversity hotspots, areas of high poultry and livestock density, and regions of substantial human settlement. To reduce contamination of the background by undetected infection, we applied a 15 km exclusion buffer around each confirmed outbreak and generated pseudo-absences by stratified random sampling within the remaining mask, matching the environmental space occupied by presence locations. We used equal numbers of presence and pseudo-absence points to balance model calibration and minimize spatial sampling bias in downstream species distribution modelling.

To quantify continental-scale seasonality in HPAI exposure potential, we derived monthly Hill-1 entropy surfaces that integrate predictions from 126 waterbird species distribution models (Extended Data Fig. 3). For each month, these entropy-based risk maps summarize both the richness and evenness of waterbird communities, yielding a composite measure of potential viral exposure. The resulting sequence of monthly panels reveals pronounced north-south oscillations in inferred HPAI risk that track large-scale migratory dynamics along the major North American flyways, with risk concentrating in southern wintering and staging areas in boreal winter and shifting toward northern breeding regions in summer.

We quantified the marginal effects of key anthropogenic and ecological predictors on HPAI risk using mean partial dependence plots (PDPs) for human density, poultry density, cattle density, and the temporal co-occurrence index (TCI) aggregated across all months (Extended Data Fig. 4). For each variable, we computed the average change in predicted risk while holding other covariates at their empirical distributions, and summarized month-to-month variability as ±1 SD around the

mean partial dependence. Across the annual cycle, risk generally increased with higher human density and TCI values, whereas poultry and cattle density exhibited nonlinear relationships: predicted risk rose from low to intermediate densities and then declined at the highest densities, consistent with saturation effects and changing agro-ecological context in heavily intensified production zones.

Spatial cross-validation

To limit spatial leakage, we constructed spatial folds via k-means clustering on presence coordinates. Absences were assigned to the nearest presence cluster (1-NN). When (k=1) (data-sparse months), a single-model path is used (no CV AUC reported for that month).

Random forest training and AUC

For each entropy formulation and month, we trained a probabilistic random forest model configured with 1500 trees, a variable selection parameter of $m_{try} = \lfloor \sqrt{p} \rfloor$, and permutation-based importance. Models were fitted in probability mode to generate continuous risk estimates. Predictive performance was evaluated using fold-wise AUC computed on held-out validation data via the pROC package, with results aggregated across folds to assess monthly model stability.

Probability calibration (Platt scaling)

Random forest probabilities were calibrated independently for each month using a simple Platt scaling approach[62]. Approximately 20% of the training data, stratified by class, were randomly selected as a calibration subset. A logistic regression of the form

$$L(y) = a + bP_{raw}$$

was then fitted to relate raw random forest probabilities to observed presence-absence outcomes. The resulting coefficients were used to transform model outputs according to

$$P_{cal} = L^{-1}(a + bP_{raw}),$$

producing calibrated probability surfaces that better reflected empirical prevalence. In months where spatial cross-validation yielded a single fold (k=1), the calibration model was instead trained on a subsample of the full dataset to preserve consistency.

Species attribution (ablation Δ-probability)

For each month and focal species, we quantified its spatial contribution to overall HPAI risk by computing Δ-probability (Δ-prob) maps. This was achieved by first neutralizing the species-specific predictor layer, replacing it with its spatial mean to preserve the marginal distribution while removing spatial structure, and then re-predicting risk using the same random forest model and Platt calibrator. The difference between the calibrated baseline and neutralized predictions,

$$\Delta = P_{cal,\ baseline} - P_{cal,\ neutralized},$$

represents the marginal effect of that species on predicted outbreak probability. Positive Δ values indicate locations where the species increases risk, while negative values denote areas of reduced influence.

Dominant driver (index and magnitude)

To identify which species most strongly structured local HPAI risk through time, we mapped the dominant waterfowl contributor to Hill-1 based risk for each grid cell and month (Extended Data Fig. 5). For each focal species, we generated a "species-neutralized" prediction by replacing its modeled probability surface with a regional baseline and computing the resulting Δ-probability between the full and neutralized scenarios. Monthly Δ-probability rasters across all focal species were then stacked to derive two summary layers: a dominant index, assigning each pixel to the species with the maximum positive Δ (set to 0 where all $\Delta \leq 0$), and a dominant magnitude, representing the corresponding maximum Δ intensity (also set to 0 when no positive values occurred). These layers respectively indicate which species contributed most to local HPAI risk and the strength of that contribution. Seasonal dominance patterns closely track migration, with Mallard and Northern Pintail largely governing interior continental risk (particularly across the Prairie Pothole Region and central flyways), other dabbling ducks (e.g., Blue-winged Teal, Gadwall) emerging in southern and spring staging areas, and geese species increasingly dominating northern and coastal staging regions during peak migration and breeding periods.

To assess overall species influence on predictive performance, we conducted a leave-one-species-out (LOSO) ΔAUC analysis (Extended Data Fig. 6). For each cross-validation fold, we retrained the random forest model while omitting a single focal species predictor and recomputed the test AUC. The reduction in AUC relative to the full model provided a quantitative measure of that species' contribution to model skill. This sensitivity analysis revealed that excluding Mallard, Canada Goose, Northern Pintail, or Gadwall produced the largest declines in AUC, indicating that these species carry disproportionate weight in explaining spatial and temporal patterns of HPAI risk. Removing other dabbling ducks led to more moderate reductions, consistent with their secondary, context-dependent roles in structuring exposure across the North American landscape.

To quantify regional species-level contributions to HPAI risk, we aggregated species-specific risk contributions within each ecological hotspot and expressed them as proportional shares of total modeled risk (Extended Data Fig. 7). For each hotspot polygon, we integrated the species-resolved risk surfaces across all grid cells (and relevant months) and normalized by the summed contribution of the ten focal species to obtain region-specific composition profiles. These profiles are shown as stacked bar plots, where bar height reflects total risk and segment height reflects each species' proportional contribution. The Prairie Pothole Region and Mississippi Alluvial Valley exhibit relatively balanced stacks spanning many species, consistent with diffuse, community-level pressure. In contrast, Delmarva, the GA-SC coastal hotspot, and the Pacific Northwest display higher concentration in a subset of species, and the associated sensitivity bars mirror this

pattern by assigning disproportionately high marginal influence to the top four species in these regions.

Robustness and data-scarcity safeguards

To ensure robustness under varying data availability, several safeguards were implemented throughout the modeling workflow. Species with incomplete monthly suitability layers were automatically excluded, and the pipeline proceeded if at least three focal species remained. Months lacking a valid training dataset were skipped and written as NA rasters, with corresponding warnings logged. For data-scarce months in which spatial cross-validation produced only a single fold (k = 1), a single random forest model was fitted and calibrated.

Phylogeographic Analysis of HPAI H5N1 Spread

To quantify cross-species and inter-regional transmission dynamics of HPAI H5N1 (clade 2.3.4.4b), we analyzed a dataset of 14,263 full-length hemagglutinin (HA) gene sequences obtained from GISAID[63] (January 2021 - May 2025). Each sequence was annotated with host category (wild birds “WB”, domestic birds “DB”, wild mammals "WM”, domestic mammals "DM”, humans "H”) and geographic origin at the state or provincial level.

Sequences were aligned using MAFFT v7.505[64] under default parameters, and a maximum-likelihood phylogeny was inferred with IQ-TREE v2.2.6 [65] using the GTR+F+Γ4 substitution model and ultrafast bootstrap approximation (1000 replicates) for branch support. A time-scaled phylogeny was subsequently reconstructed with TreeTime v0.10.1 [66], applying a strict molecular clock and a coalescent skyline prior to calibrate branch lengths in calendar years.

Transmission events were inferred following the discrete trait framework of Leke et al.[70], which counts state-to-state changes along internal branches as proxy transmission events between defined categories (here, host class × geographic unit). Trait-transition matrices were computed from annotated internal node reconstructions, and normalized carrier-to-recipient transmission counts were summarized for each ($host_1$ to $host_2$) pair. Statistical significance of associations among host and location traits was assessed using a log-linear likelihood-ratio test (LRT), evaluating whether observed transmission frequencies deviated from independence expectations.

The reconstructed time-scaled phylogeny captured strong host and geographic structure, revealing extensive viral exchange among wild and domestic birds across North America. We identified >3,000 inferred transmission events, with a highly significant interaction among host categories ($LRT = 3039.8$, $df = 544$, $p < 0.001$).

We quantified cross-host transmission patterns using a discrete-trait phylogenetic reconstruction of host state along a time-scaled HA phylogeny (n = 14,263 sequences), treating wild birds (WB), domestic birds (DB), wild mammals (WM), domestic mammals (DM), and humans (H) as discrete categories. For each internal branch, we inferred state transitions and summarized the resulting

counts in a carrier-recipient transmission matrix (Extended Data Fig. 8), where rows represent the inferred source host and columns represent the recipient host. Color intensity in the corresponding heatmap reflects the relative frequency of transitions from each carrier to each recipient. These matrices show that wild birds act as the dominant source of cross-species spread, accounting for most inferred transmissions to domestic birds and sporadic seeding events into mammalian hosts. In contrast, mammal-to-mammal and human-linked transitions are rare, indicating limited onward transmission within mammalian hosts and supporting the view that the epidemic is primarily maintained within avian reservoirs.

We characterized the geographic structure of transmission using state- and province-level discrete trait reconstructions from the time-scaled HA phylogeny (n = 14,263 sequences), and summarized the top 20 locations acting as viral sources and sinks (Extended Data Fig. 9). For each inferred state transition, we recorded the origin and destination region and tallied these counts across the tree; "source" counts reflect transitions originating in a given region, whereas "sink" counts reflect transitions terminating in that region. Geographically, the reconstructions revealed repeated viral exchanges among the Prairie, Great Lakes, and Atlantic regions, consistent with established migratory flyway routes. Western and central regions such as Alberta, Minnesota, and British Columbia emerged as major exporters, contributing disproportionately to onward dissemination toward downstream recipient areas in the Midwest, Great Lakes, and northeastern United States. These movements were primarily driven by wild-bird lineages that frequently seeded new introductions into domestic poultry populations, where onward spread was limited, and spillover into mammals appeared as sporadic, largely terminal events without evidence of sustained circulation. Collectively, these patterns underscore the dominant role of wild birds as the principal maintenance and dissemination host of H5N1 in North America, with domestic birds acting mainly as epidemiological sinks. A significant host interaction effect in the discrete trait model (LRT = 3039.8, $p < 0.001$) further highlights the tight ecological coupling between migratory waterfowl and poultry systems that underpins the persistence and continental spread of the epidemic.

Phylogeography and BEAST Configuration

To investigate the spatial and temporal dynamics of highly pathogenic avian influenza (HPAI) H5N1 clade 2.3.4.4b across North America, we analyzed a representative subset of 1,900 hemagglutinin (HA) gene sequences drawn from a larger dataset of 14,263 sequences collected between January 2021 and May 2025 from both wild and domestic hosts. Each sequence contained county-level geographic metadata, which was integrated with monthly entropy-based risk maps. For each county, the mean risk value was extracted and classified into three discrete ecological risk states high, medium, or low providing a spatially explicit framework for phylogeographic inference. A time-scaled phylogeny was first inferred using TreeTime, which was then used as the fixed starting tree for Bayesian phylogeographic analysis in BEAST v1.10.4 [72].

Phylogeographic inference was performed under an asymmetric continuous-time Markov chain (CTMC) model for transitions among the three risk states (High, Medium, Low). Temporal

calibration was based on precise sampling dates (year-month-day) under a strict molecular clock, with a lognormal prior on the substitution rate (mean = $1 \times 10^{-3}$ subs/site/year, SD = 0.33). Nucleotide substitution followed a GTR + Γ4 model with all parameters estimated.

The Bayesian Skygrid[71] coalescent model (20 grid points spanning 2020-2025) was used to accommodate flexible changes in effective population size. Bayesian stochastic search variable selection (BSSVS) was applied to identify well-supported transitions among risk states, with the CTMC rate reference prior set to exponential(mean = 1.0). Uniformization and complete Markov-jump history logging were enabled to quantify directional transition counts and state occupancy through time.

Each MCMC chain was run for 50 million steps, sampling every 10,000 iterations, with 10% burn-in. Convergence (ESS > 200) was confirmed in Tracer v1.7.2 [72]. The maximum clade credibility (MCC) tree was summarized in TreeAnnotator (posterior probability ≥ 0.8) and visualized using FigTree and ggtree.

The resulting MCC tree revealed frequent transitions between medium- and high-risk nodes, consistent with sustained viral circulation within ecological hotspots and recurrent introductions into lower-risk regions. Temporal scaling placed the tMRCA of North American H5N1 lineages in early 2021 (95% HPD: 2020.8-2021.3), corresponding to the onset of the continental HPAI wave. Tips were colored according to the observed risk category, while internal branches and nodes were shaded by the most probable reconstructed category (Extended Data Fig. 10). The resulting MCC tree revealed frequent transitions between medium- and high-risk nodes, consistent with sustained viral circulation within ecological hotspots and recurrent movement along major exposure corridors, as well as intermittent introductions from high- into lower-risk regions. Temporal scaling placed the tMRCA of North American H5N1 lineages in early 2021 (95% HPD: 2020.8-2021.3), coinciding with the onset of the continental HPAI wave and the emergence of persistent diversification within high-risk areas from mid-2021 onward.

Continuous diffusion and spatial reconstruction

To infer geographic diffusion dynamics, we reanalyzed the HA dataset under a continuous phylogeographic framework using a relaxed random walk (RRW) model[73], treating latitude and longitude as continuous traits. The diffusion process was modeled with a Cauchy-distributed RRW to accommodate occasional long-distance dispersal events, as recommended for avian influenza datasets. Molecular clock and substitution priors were kept identical to those used in the discrete-trait analyses, and the RRW variance parameter was assigned a gamma prior (shape = 0.5, scale = 0.5). Independent BEAST runs were combined in LogCombiner after confirming convergence (all key parameters with ESS > 200), and marginal likelihood estimation via stepping-stone sampling supported an asymmetric CTMC + Skygrid coalescent model over symmetric or constant-size alternatives. Continuous diffusion patterns were summarized and visualized in SPREAD3 v0.9.7 [74], yielding time-resolved maps of inferred dispersal routes (Extended Data Fig. 11).

Model Validation and Spatial Interpretation

Spatial reconstructions of Markov jump counts and continuous diffusion paths revealed major viral movement corridors along the Pacific, Central, and Atlantic flyways, with recurrent re-entry into the Prairie Pothole and Mississippi Valley regions that overlapped high-entropy ecological risk zones from the mapping framework. Continuous-trait reconstructions confirmed strong north-south migration-linked spread and additional east-west exchanges between the Mississippi and Atlantic flyways (Extended Data Fig. 11). Collectively, these results indicate that viral movement aligns more closely with ecological barriers and high-risk hotspots than with unrestricted open corridors: high-risk regions function as persistent reservoirs and diversification hubs, whereas transitions into low-risk areas are typically transient. Although Bayes factor support for individual directional transitions was moderate, the consistency of the spatial patterns across models and analyses points to strong ecological constraints on H5N1 dispersal across North America.

**Data availability**

All genome sequences and associated metadata utilized in the study are available in GISAID. The GISAID Acknowledgement Table is provided in Supplementary Data Table 3 and is accessible via the persistent DOI (https://doi.org/10.55876/gis8.260114dw). The other data supporting the findings of this study are available within the paper and its supplementary information files.

**Reference:**


50. Sullivan, B. L. *et al.* eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142**, 2282–2292 (2009).
51. GBIF.org. GBIF Occurrence Download. https://doi.org/10.15468/dl.cnpwng (2025).
52. Gilbert, M. *et al.* Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci. Data* **5**, 1–11 (2018).
53. Ciesin, I. Gridded population of the world, version 4 (GPWv4): Population count. *Palisades NY NASA Socioecon. Data Appl. Cent. SEDAC* (2016).
54. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
55. Food and Agriculture Organization of the United Nations (FAO). EMPRES-i+ Global Animal Disease Information System. https://empres-i.fao.org (2024).
56. World Organisation for Animal Health (WOAH). World Animal Health Information System (WAHIS). https://wahis.woah.org (2024).
57. Canadian Food Inspection Agency (CFIA). National Emergency Operations Centre (NEOC) GIS Avian Influenza Dashboard. https://inspection.canada.ca (2024).
58. Friedl, M., Gray, J. & Sulla-Menashe, D. MODIS/Terra+ Aqua land cover dynamics yearly L3 global 500m SIN grid V061. *NASA EOSDIS Land Process. Distrib. Act. Arch. Cent. DAAC Data Set* MCD12Q2-061 (2022).

59. Miller, J. Species distribution modeling. *Geogr. Compass* **4**, 490–509 (2010).
60. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
61. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
62. Platt, J. & others. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **10**, 61–74 (1999).
63. GISAID (database): sequences and metadata used in this study.
64. MAFFT v7.505.
65. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
66. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
67. Lyu, L. *et al.* Characterizing spatial epidemiology in a heterogeneous transmission landscape using the spatial transmission count statistic. *Commun. Med.* **5**, 165 (2025).
68. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Syst. Biol.* **67**, 901–904 (2018).
69. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
70. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
71. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
72. Bielejec, F. *et al.* SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. *Mol. Biol. Evol.* **33**, 2167–2169 (2016).

**Extended Data Table 1.** Data sources used in modelling.

| Category | Variable(s) | Resolution | Source(s) |
|---|---|---|---|
| Wild-bird occurrence | Species observations (1990-2025) | Point | eBird, GBIF |
| Livestock density | Poultry, cattle (2018) | 1 km | FAO GLIMS |
| Human population | GPWv4 (2020) | 1 km | CIESIN |
| Climate | Temperature & precipitation normals | 1 km | WorldClim v2.1 |
| Outbreak locations | Confirmed HPAI events (2015-2025) | Point | FAO EMPRES-i, WOAH WAHIS, CFIA NEOC GIS |
| Land cover | MODIS MCD12Q1 (2019) | 500 m (aggregated) | NASA LP-DAAC |
| Vegetation dynamics | MODIS NDVI & EVI (2001-2024) | 500 m (aggregated) | NASA LP-DAAC |

**Extended Data Table 2.** Water bird species (Family Anatidae) and Raptors retained for the HPAI spread models after data filtering.

| **Species 1** | **Species 2** | **Species 3** |
|---|---|---|
| Accipiter cooperii | Aythya valisineria | Scolopax minor |
| Accipiter striatus | Botaurus lentiginosus | Charadrius semipalmatus |
| Actitis macularius | Branta bernicla | Charadrius vociferus |
| Aechmophorus occidentalis | Branta canadensis | Chlidonias niger |
| Aix sponsa | Branta hutchinsii | Clangula hyemalis |
| Anas acuta | Bubulcus ibis | Coragyps atratus |
| Anas crecca | Bucephala albeola | Cygnus buccinator |
| Anas fulvigula | Bucephala clangula | Cygnus columbianus |
| Anas platyrhynchos | Bucephala islandica | Cygnus olor |
| Anas rubripes | Buteo jamaicensis | Dendrocygna autumnalis |
| Anhinga anhinga | Buteo lagopus | Egretta caerulea |
| Anser albifrons | Buteo lineatus | Egretta thula |
| Anser caerulescens | Buteo platypterus | Egretta tricolor |
| Aquila chrysaetos | Buteo swainsoni | Elanus leucurus |
| Aramus guarauna | Cairina moschata | Eudocimus albus |

| | | |
|---|---|---|
| Ardea alba | Calidris alba | Fregata magnificens |
| Ardea herodias | Calidris alpina | Fulica americana |
| Arenaria interpres | Calidris himantopus | Gallinago delicata |
| Aythya affinis | Calidris mauri | Gavia immer |
| Aythya americana | Calidris melanotos | Gavia stellata |
| Aythya collaris | Calidris minutilla | Grus canadensis |
| Aythya marila | Calidris pusilla | Haematopus palliatus |
| Ictinia mississippiensis | Cathartes aura | Haliaeetus leucocephalus |
| Larus californicus | Numenius americanus | Hydroprogne caspia |
| Larus delawarensis | Numenius phaeopus | Somateria mollissima |
| Larus fuscus | Nyctanassa violacea | Spatula clypeata |
| Larus glaucescens | Nycticorax nycticorax | Spatula cyanoptera |
| Larus glaucoides | Oxyura jamaicensis | Spatula discors |
| Larus heermanni | Pandion haliaetus | Sterna forsteri |
| Larus marinus | Pelecanus erythrorhynchos | Sterna hirundo |
| Larus occidentalis | Pelecanus occidentalis | Sternula antillarum |
| Limnodromus griseus | Phaethon aethereus | Sula dactylatra |
| Limnodromus scolopaceus | Platalea ajaja | Sula leucogaster |
| Limosa fedoa | Plegadis chihi | Thalasseus maximus |
| Lophodytes cucullatus | Plegadis falcinellus | Tringa flavipes |
| Mareca americana | Pluvialis squatarola | Tringa melanoleuca |
| Mareca strepera | Podiceps auritus | Tringa semipalmata |
| Megaceryle alcyon | Podiceps grisegena | Tringa solitaria |
| Melanitta americana | Podiceps nigricollis | Urile penicillatus |
| Melanitta deglandi | Podilymbus podiceps | - |
| Melanitta perspicillata | Porzana carolina | - |
| Mergus merganser | Rallus crepitans | - |
| Mergus serrator | Rallus limicola | - |
| Morus bassanus | Recurvirostra americana | - |
| Mycteria americana | Rynchops niger | - |

**Extended Data Table 2.** GISAID Acknowledgment table.

| GISAID Identifier | Digital Object Identifier | Number of individual viruses | Data Collection range | Number of Countries/ territories |
|---|---|---|---|---|
| EPI_SET_260114dw | https://doi.org/10.55876/gis8.260114dw | 14,363 | 14,363 2021-12-16 to 2025-03-31 | 5 |