esa

ECOSPHERE

MACROSYSTEMS ECOLOGY

Big data–model integration and AI for vector-borne disease prediction

Debra P. C. Peters[1,†] D. Scott McVey,[2] Emile H. Elias,[1] Angela M. Pelzel-McCluskey,[3] Justin D. Derner,[4] N. Dylan Burruss,[5] T. Scott Schrader,[1] Jin Yao,[1] Steven J. Pauszek,[6] Jason Lombard,[3] and Luis L. Rodriguez[6]

[1]*US Department of Agriculture, Agricultural Research Service, Jornada Experimental Range Unit, and Jornada Basin Long Term Ecological Research Program, New Mexico State University, Las Cruces, New Mexico 88003 USA*
[2]*US Department of Agriculture, Agricultural Research Service, Center for Grain and Animal Health Research, Arthropod-Borne Animal Diseases Research Unit, Manhattan, Kansas 66506 USA*
[3]*US Department of Agriculture, Animal and Plant Health Inspection Service, Veterinary Services, Fort Collins, Colorado 80526 USA*
[4]*US Department of Agriculture, Agricultural Research Service, Rangeland Resources and Systems Research Unit, Cheyenne, Wyoming 82009 USA*
[5]*Jornada Experimental Range, New Mexico State University, Las Cruces, New Mexico 88003 USA*
[6]*US Department of Agriculture, Agricultural Research Service, Plum Island Animal Disease Center, Orient Point, New York 11957 USA*

Citation: Peters, D. P. C., D. S. McVey, E. H. Elias, A. M. Pelzel-McCluskey, J. D. Derner, N. D. Burruss, T. S. Schrader, J. Yao, S. J. Pauszek, J. Lombard, and L. L. Rodriguez. 2020. Big data–model integration and AI for vector-borne disease prediction. Ecosphere 11(6):e03157. 10.1002/ecs2.3157

Abstract. Predicting the drivers of incursion and expansion of vector-borne diseases as part of early-warning strategies (EWS) is a major challenge for geographically extensive diseases where spread is mediated by spatial heterogeneity in climate and other environmental drivers. Geospatial data on these environmental drivers are increasingly available affording opportunities for application to a predictive disease ecology paradigm provided the data can be synthesized and harmonized with fine-scale, highly resolved data on vector and host responses to their environment. Here, we apply a multi-scale big data–model integration approach using human-guided machine learning to objectively evaluate the importance of a large suite of spatially distributed environmental variables (>400) to develop EWS for vesicular stomatitis (VS), a common viral vector-borne vesicular disease affecting livestock throughout the Americas. Two temporally and phylogenetically distinct events were used to develop disease occurrence–environment relationships in incursion (2004) and expansion years (2005), and then to test those relationships (2014, 2015) at two scales: (1) local and (2) landscape to regional. Our results show that VS occurrence at a local scale of individual landowners was related to conditions that can be monitored (rainfall, temperatures, streamflow) or modified (vegetation). On-site green vegetation during the month of occurrence and higher rainfall four months prior combined with either cool daytime (expansion) or nighttime (incursion) temperatures one month prior were indicators of VS occurrence. Distance to running water (incursion) and host density based on neighboring ranches (expansion) with infected animals were also important in individual years. At landscape-to-regional scales, conditions that favor specific VSV biological vectors were indicated, either black flies in incursion years or biting midges in expansion years. Changes in viral genetic lineage were less important to patterns in VS occurrence than factors affecting the host–vector–environment interactions. In combination with our onset map based on latitude, elevation, and long-term annual precipitation, this year- and scale-specific information can be used to develop strategies to minimize effects of future VS events. This big data approach coupled with expert knowledge and machine learning can be applied to other emerging diseases for improvement in understanding, prediction, and management of vector-borne diseases.

Key words: artificial intelligence; continental scale; expert knowledge; insect vectors; livestock; phylogeography; regional scale; RNA virus; vesicular stomatitis virus.

Received 25 September 2019; revised 16 March 2020; accepted 6 April 2020. Corresponding Editor: David D. Breshears.

† **E-mail:** deb.peters@usda.gov

Introduction

Vector-borne diseases that spread over regional to continental scales pose threats to public health, food security, and national and international trade (Tomley and Shirley 2009). Predicting the factors whose variation explains patterns of disease as part of early warning strategies (EWS) for arthropod-borne diseases (arboviruses) is challenging because different hosts (livestock, humans, or wildlife), vectors, and dispersal mechanisms can be involved as a disease system encounters a spatially heterogeneous environment through time (Racloz et al. 2012, Altizer et al. 2013, Stewart-Ibarra and Lowe 2013, Parham et al. 2015). Arboviruses with endemic and epidemic life cycles, such as West Nile, Rift Valley fever, and Venezuelan equine encephalitis, are directly and indirectly influenced by environmental variability at multiple spatial and temporal scales that often result in different processes governing dynamics as spatial extent changes (Woolhouse et al. 2013). Focusing on one or a few factors or processes at the exclusion of others can lead to devastating and surprising consequences for human and livestock health (Jacquot et al. 2017, Mayer et al. 2017, Sule et al. 2018). For many diseases, there is often insufficient information about the environmental factors and processes that interact to govern dynamics leading to multi-scale patterns in spread (Escobar and Craft 2016). It is thus necessary to first develop data-intensive, process-based approaches using a disease system where datasets are readily available for the full suite of potential environmental factors across its spatial domain (Michael et al. 2017). Developing scale-dependent, analytics-based approaches will facilitate the implementation of more cost-effective early warning mitigation strategies for geographically extensive diseases (Han and Drake 2016).

Traditional approaches to identifying processes governing spread of disease have had limited success in examining multiple potential environmental factors as the spatiotemporal conditions change. Local-scale disease processes examined in response to environmental conditions are traditionally studied in the laboratory; however, these results cannot be extrapolated to other spatial extents and field conditions unless stability and stationarity in fine-scale pattern–process relationships are assumed (Althouse et al. 2012, 2016, Michael et al. 2017). Likewise, studies of multi-scale correlations cannot easily infer the importance of different processes (Cohen et al. 2016), and case studies at regional scales often fail to provide a mechanistic understanding of disease dynamics at finer scales (Ginsberg et al. 2009, Messina et al. 2014, Walsh and Haseeb 2015, Faria et al. 2017). Climate-driven models can examine the role of multi-scale climatic variation on disease dynamics (Morin and Comrie 2013), but not the relative effects of the full suite of environmental factors that vary across geographic space. Few studies have quantified the multi-scale relationships among more than one environmental factor and the habitat of a vector-borne disease in attempts to elucidate processes (Lo Iacono et al. 2018), or to develop a generalized strategy that can be applied to other diseases globally (National Academies of Sciences Engineering and Medicine 2016). Recent advances in data science and online availability of geo-referenced, multi-scale climate and environmental datasets combined with the development of trans-disciplinary approaches provide opportunities to develop EWS that are both context-dependent at a local scale and generalizable across the geographic extent of a disease (Han and Drake 2016).

Recently, Peters et al. (2018) developed a big data–model integration (BDMI) approach guided by expert knowledge to identify and evaluate the relative importance of a large and diverse suite of all known environmental factors and life-history variables to patterns in vector-borne pathogen incursion and expansion. This framework uses a trans-disciplinary team to coherently integrate: (1) fine-scale, process-based data and understanding of vector and host responses to a

pathogen and to the local environment, (2) georeferenced disease incidence and virus phylogenetic relationships, and (3) fine-scale patterns in climate, land surface properties, and host density for a multi-decadal temporal period across the continental extent of a disease. This approach differentiates drivers (e.g., climate, topography) from specific variables within each driver that are chosen for their ecological meaning in the system under study (e.g., daily minimum temperature, annual precipitation, elevation). The approach considers potential effects of many environmental variables on disease processes, both individually and their interactions, within and across spatiotemporal scales when little is known about the ecology of a particular system. The approach focuses on pattern–process relationships in the neighborhood of an individual premise leading up to the time of disease. Both key local (potential vertical transmission, horizontal transmission, possible insect overwintering, contact spread between animals, insect transmission to animals) and spatial processes (dispersal of hosts and vectors) are used to identify major drivers that typically influence disease systems. The relative influence of one or more variables within each driver is assessed in a multi-model analysis based on expected relationships from literature and team expertise with responses in one or more processes, ultimately leading to patterns in disease occurrence. This approach allows comprehensive examination of a large number of potential variables across a range of scales, but then restricts analysis to those variables that are biologically meaningful. A trans-disciplinary team is used to develop the conceptual model of the system, identify the variables, and then harmonize, analyze, and interpret the data. Here, the utility of this approach is illustrated using incursions into North America by vesicular stomatitis New Jersey virus (VSNJV), a model system for other vector-borne diseases caused by RNA viruses.

The VS disease system

VSNJV is a vector-borne, zoonotic RNA virus in the family *Rhabdoviridae* that causes readily observed vesicular lesions on wildlife and domestic livestock. In some species (ruminants, pigs), these lesions are clinically indistinguishable from foot-and-mouth disease (FMD), one of the most devastating exotic diseases in livestock that was eradicated from the United States in 1929. VS is the most reported vesicular disease affecting livestock (domestic horses, cattle, pigs) throughout the Americas (Rodríguez 2002), and although VS is less severe than FMD, there is no cure and no vaccine. Economic costs of VS through loss of milk and meat production and through regulatory repercussions (e.g., quarantines, limited movement, and sale of animals and animal products) can be significant (e.g., losses totaled> $14 M in the United States in 1995).

Despite decades of documented patterns of occurrence (Rodríguez et al. 2000) and multiple epidemiological studies, there is limited understanding about factors governing patterns in the spread of VS through time. A VS event originating in Mexico has occurred every decade in the Western United States since 1906 with a combined spatial extent over> 1.1 M km^2 from 2004 to 2016 (Fig. 1a, b). Major disease cycles occur at ca. 10-yr intervals that consist of the following: (1) failed incursion years (defined as initial disease spread into the United States) where disease is limited to southern states bordering Mexico and stops after one year (e.g., 2012), (2) successful incursion years with initially limited spatial distribution (e.g., 2004; Fig. 1c) followed by widespread expansion in subsequent years (defined as proliferation and spread) (e.g., 2005; Fig. 1d), and (3) extinction years where few or no cases occur in a limited geographic area following expansion (e.g., 2006). Recent analyses show that each cycle of events has been caused by single distinct viral genetic lineages that originated in southern Mexico (Rodríguez et al. 2000, Rainwater-Lovett et al. 2007, Velazquez-Salinas et al. 2014).

Landscape-scale patterns of disease have been related to one or a few factors, such as elevation, monthly precipitation, or stream flow (Rodríguez et al. 1996, McCluskey et al. 2003, Elias et al. 2019). Host density and environmental conditions affecting the life cycle and dispersal of relevant VSV biological vectors (e.g., black flies, biting midges, sand flies) are believed to play a major role in VS occurrence based on field observations and laboratory experiments (Walton et al. 1987, Kramer et al. 1990, Cupp et al. 1992). The Western United States is particularly rich in climate, stream flow, vegetation, and

(a)

VS Outbreak Year	States	Counties Infected	Number of Premises Infected
2004	CO, NM, TX	43	294
2005	AZ, CO, ID, MT, NE, NM, TX, UT, WY	71	445
2006	WY	3	13
2009	NM, TX	3	5
2010	AZ	1	2
2012	CO, NM, TX	12	36
2014	AX, CO, NE, TX	32	435
2015	AZ, CO, NE, NM, SD, TX, UT, WY	79	823
2016	none	0	0

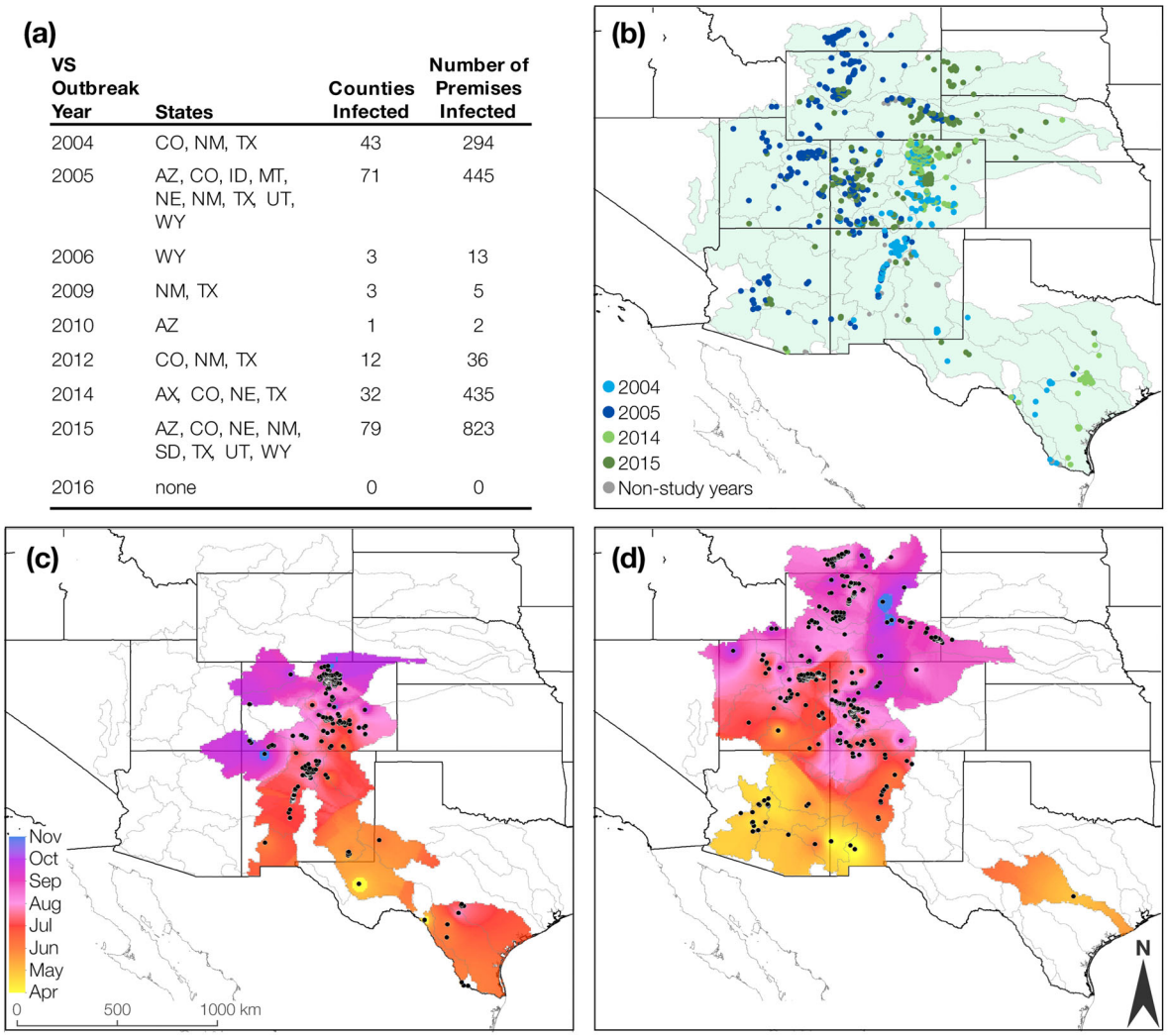


Fig. 1. Location of VS confirmed cases in the United States (a) all cases from 2004-2015 by year, states, counties, and premises infected, (b) 2004-05 (shades of blue) and 2014-15 (shades of green), (c) 2004, and (d) 2005 cases showing general trend from south to north by colored areas developed using inverse distance measures with kriging (data from www.aphis.usda.gov; USDA-APHIS-VS databases).

other biophysical data relevant to different components of the disease transmission system (Table 1). However, a multi-scale analysis of the potential suite of factors that could affect insect population dynamics and disease transmission across the spatial extent and temporal domain of this disease has not been conducted, and these factors have not been examined under natural conditions. Thus, the range of conditions for each environmental variable where the VS disease occurs has not been quantified, and the mechanisms mediating disease emergence and expansion remain elusive.

Furthermore, VS can be a model system for other vector-borne diseases because of the following: (1) the complexity and importance of the reportable disease, (2) the epidemiological knowledge of VS, yet lack of ecological understanding about the natural cycle and spread of this disease, (3) the accessibility of co-located disease occurrence, viral phylogenetic, and environmental data through time, and (4) the availability

Table 1. Source, temporal, and spatial resolution of input data.

Input data	Source of data	Temporal resolution	Spatial resolution	Variables‖	Component or Process†
VSNJV case occurrence‡	AM Pelzel-McCluskey (USDA-APHIS-VS databases)	Daily data (2003-2016)	point	NA	Host
VSNJV lineage‡	LL Rodriguez	Daily data (2003-2015)	point	NA	Virus
Animals§	https://quickstats.nass.usda.gov/	2002, 2007, 2012 data	county	Horse (density)	Dispersal (D, C, H)
Animal premises§	https://quickstats.nass.usda.gov/	2002, 2007, 2012 data	county	Farm and ranch (density)	Dispersal (D, C, H)
Pedology¶	http://www.soilinfo.psu.edu/index.cgi?soil_data&conus&data_cov&fract [NRCS]	Static maps [STATSGO]	900 m	Soil properties: % clay, AWC#	Biting midge (V)
Hydrology	https://www.sciencebase.gov/catalog/item/51360134e4b03b8ec4025bfa [USGS]	Static maps	30 m	Location of water bodies	Black fly (V)
	https://waterdata.usgs.gov/nwis/sw [USGS]	Daily data (2003-2016)	30 m	Stream flow	Black fly (V)
	http://giovanni.gsfc.nasa.gov/giovanni/ [NASA]	Monthly data (2003-2016)	12 km	Runoff (cm); Soil moisture (%)	Black fly (V); biting midge (V)
Topography¶	http://www2.jpl.nasa.gov/srtm/ [NASA]	Static DEM	900 m	Elevation (m)	OW, V
Climatology¶	http://www.prism.oregonstate.edu/normals/ [OSU]	Daily, monthly data (2003-2016); long-term average data (1981-2010)	4 km	Minimum, maximum temperature (°C); precipitation (cm)	OW, V; V, H
Climatology¶	http://climate.colostate.edu/~drought [NOAA]	Monthly data (2002-2015)	12 km	Evaporative demand drought index (EDDI)	V, H
Land surface properties¶,††	https://lpdaac.usgs.gov/node/78 [NASA]	Monthly imagery; MODIS (2003-2016)	5.6 km	Vegetation greenness (NDVI)	V, H

Note: Agencies in square brackets [ ] are the U.S. state or federal government agency with data.

† Predominant process(es) expected to be important. Abbreviations are: D, dispersal; C, contact transmission; H, horizontal transmission; V, vertical transmission; OW, overwintering (other processes are either less important or there is insufficient data on importance).

‡ Response variable.

§ Host factors: Linear extrapolation was used to estimate values in years without sampling.

¶ Environmental drivers.

Available water holding capacity.

‖ Variable classes used in Fig. 3.

†† See Table 2 for temporal variables.

of scientific and technological expertise. In addition, occurrences of the disease are known to be associated with the presence of the virus because our case definition is clinical signs of disease confirmed by laboratory detection of the virus or immunological evidence of recent infection.

We had two objectives: (1) to identify the factors governing spatial variability in VS occurrence at the landscape-to-regional scale and (2) to assess how local-scale environmental conditions differ between years when animals become infected compared to years without infection, considering incursion years when neighboring premises are not infected separately from expansion years when neighboring premises contain infected animals. Addressing these objectives will inform the development of scale-dependent early-warning strategies for VS in the Western United States given that some of the conditions for disease spread at a larger scale can be modified at the local scale.

We used our multi-scale framework to test hypotheses at two spatial scales. (1) At the *landscape-to-regional scale*, we hypothesized that spatial patterns are determined by either (a) viral genetic determinants alone, such that

relationships developed between VS occurrence and explanatory variables in 2004, an incursion year, can be successfully applied to explain patterns in an expansion year with similar virus phylogeny (2005) (virus hypothesis) or (b) environmental variables such that relationships developed in one incursion (e.g., 2004) or expansion year (e.g., 2005) can explain patterns in another incursion (or expansion, respectively) year with a different phylogeny (incursion–expansion hypothesis). To test these alternative hypotheses, we compared variables related to occurrence patterns in two separate incursion-expansion events separated by a decade (2004–2005, 2014–2015) in which viral phylodynamics have been characterized.

(2) At the *local scale* of individual premises, we focused on vector–environment interactions because information on host immunity was unavailable. We expected a sequence of conditions would precede infection that are related to processes that increase the abundance of competent and infected insect vectors; different conditions in incursion and expansion years should indicate that different vectors are involved in each year. We tested this hypothesis using a detailed temporal analysis (1–4 weeks and 1–12 months prior to infection) to compare the importance of factors to VS occurrence at individual premises.

Methods

Approach to variable selection and harmonization

The trans-disciplinary team consists of experts in this disease, ecologists, and ecoinformatics experts, including people with experience in big data analytics and machine learning (Fig. 2). Our approach focuses on pattern–process relationships in the neighborhood of an individual premise leading up to the time of disease. The spatiotemporal resolution of a neighborhood is assumed to depend on the relationship between a variable and the vector or host process with the potential to influence infection. Both key local (potential vertical transmission, horizontal transmission, possible insect overwintering, contact spread between animals, insect transmission to animals) and spatial processes (dispersal of hosts and vectors) were used to identify six environmental drivers a priori (pedology, hydrology, topography, climate, drought, land surface properties) that typically influence disease ecology systems based on the literature (Table 1). Within each driver, we used expert knowledge in the VS system to select one or more variables expected to influence patterns or responses in one or more disease processes (Table 1), ultimately leading to patterns in VS occurrence. This approach allows comprehensive examination of a large number of potential variables, but then restricts analysis to those variables that are biologically meaningful (Tables 1, 2). We then identified a data source for each variable to cover the spatial and temporal extent of our study location, the contiguous watersheds in the Western United States where VS occurred between 2004 and 2015 (www.aphis.usda.gov) (Fig. 1b).

After variables and corresponding datasets were identified (Table 1), harmonization was used to facilitate synthesis and integration (Fig. 2). The VS disease occurrence data were converted from a geographic coordinate system to an equal area projection system (e.g., Albers Equal Area Conic) to ensure cell sizes remained the same (1 km^2) throughout the large spatial extent (~1.1 M km^2) of the study area. All other variables were harmonized to the projection, geographic origin, and cell size of this VS occurrence base map. The type of the native structure of the source data (raster, vector, polygon) determined the harmonization procedure and format of the analysis. For raster data (e.g., gridded PPT at a 4 km × 4 km resolution), harmonization consisted of resampling maps to 1 km × 1 km. Vector data (e.g., points, lines) were converted to raster, harmonized to the base layer, and then translated into distance maps. Polygons were rasterized by calculating average properties and then harmonized to the base layer. All spatial data were manipulated with ArcGIS v.10.3 to assist in the harmonization procedure. The variable selection procedure resulted in 472 raster layers, which were collated into a harmonized data cube to enable calculations and predictions to be carried out in geographical space.

For *landscape-to-regional-scale analysis*, the R package MaxentVariableSelection (Jueterbock 2015) was used to control model complexity, avoid collinearity among predictor variables, and optimize parameters for analysis. Variables were

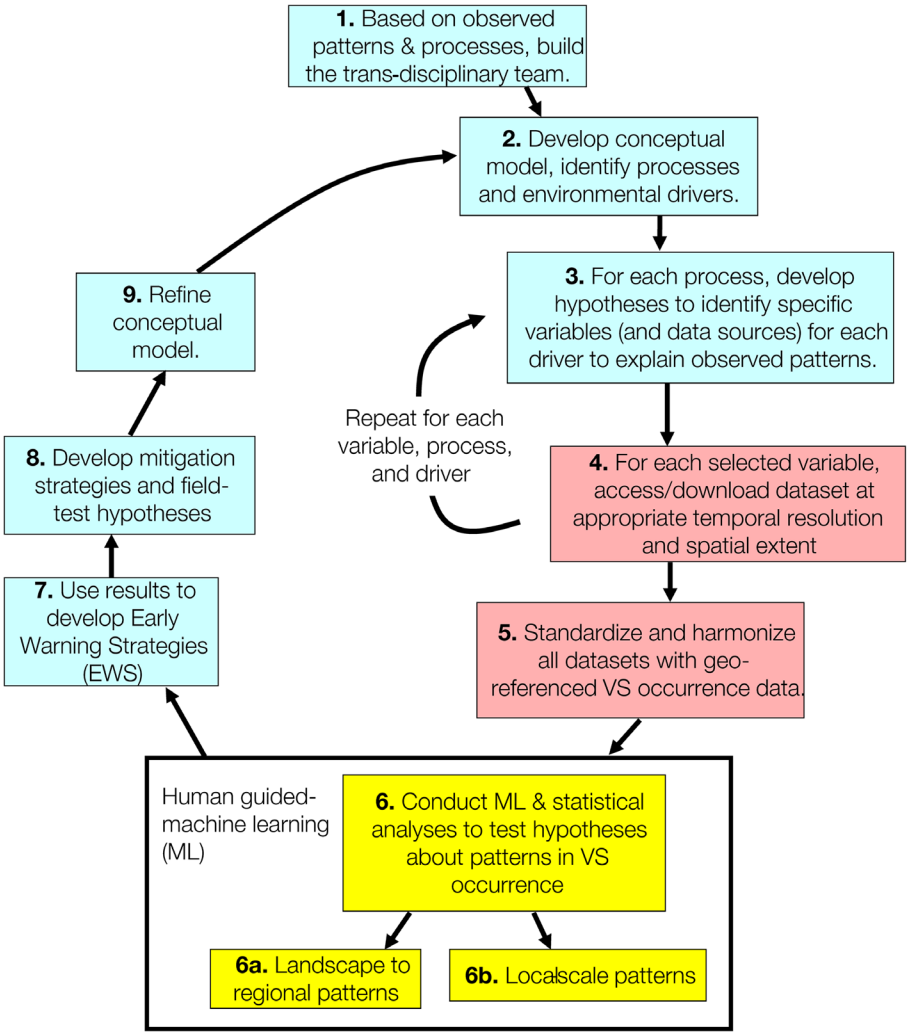


Fig. 2. Workflow showing the major steps in our approach to predictive disease ecology that includes an integration of big data with models, expert knowledge, and machine learning. Expert knowledge from disease experts and ecologists are shown in blue boxes, ecoinformatics expertise is shown in pink boxes, and human-guided machine learning is shown in yellow boxes.

removed from the analysis whose contribution to the model was <5% and whose correlation with another variable was >0.7. The resulting model typically had <20 uncorrelated variables, and <10 variables with contribution >5%. For the *local-scale analysis*, a tabular dataset was generated from the raster maps such that fine-scale relative temporal relationships (weeks or months prior to a VS occurrence) were preserved between occurrence and environmental variables, and arbitrary classifications (e.g., month, season) at broad spatial scales were masked. Cell size (4 km^2) was selected to characterize the environment surrounding a VS occurrence. All spatial data exported to tabular data were extracted from raster maps (118 variables) by calculating the mean of values extracted from a 100-point grid centered on a VS occurrence location. The number of candidate variables for multivariate analysis was reduced to 20 using an iterative, human learning procedure to identify those variables with the strongest univariate relationship to VS occurrence to avoid collinearity among predictors >70%. All tabular data processing and analyses

Table 2. Derived temporal variables for landscape- and regional-scale (noted by a superscript "1") and local-scale (noted by a superscript "2") analyses.†

Variables with temporal resolution (from Table 1)	Weekly[2] (n = 4)	Monthly[1,2] mean (n = 12)	Seasonal[1] mean (n = 4)	ΔSeasonal[1] mean (n = 4)	Annual[1,2] mean (n = 1 or 2)
Surface water properties‡					
Runoff		X	X	X	
Soil moisture		X	X	X	
Streamflow	X	X			
Air temperature‡					
Minimum	X	X	X	X	
Maximum	X	X	X	X	
Precipitation‡	X	X	X	X	
Drought index (EDDI)‡		X	X	X	
Vegetation greenness (NDVI)‡		X	X	X	
Horse (density)§					X¶
Properties with horses (density)§					X¶

Notes: "Monthly" refers to January to December. "Seasonal" refers to winter, spring, summer, and fall.
† Each of 4 yr analyzed separately (2004, 2005, 2014, and 2015).
‡ Environmental variables.
§ Biotic factors.
¶ Linear extrapolation used for non-sample years; temporal analysis included prior and current year.

were conducted in SAS v.9.4: SAS Institute, Cary, NC USA.

Identification of variables and data sources

VS occurrence data.—Records of VS occurrence (2004 through 2016) were obtained from USDA Animal Plant Health Inspection Service (APHIS). The USDA-APHIS policy of mandatory reporting of VS occurrence by doctors of veterinary medicine resulted in accurate identification of VSNJV-infected animal, onset date, and premise location. Occurrence data represent the clinical onset of VS lesions. Since only occurrences were reported, premises where VSV did not occur (i.e., absence data) were unavailable for analysis. We focused on two VS events that represented 96% (n = 1550) of recorded occurrences from 2003 to 2015: 2004–2005 for model development, and 2014–2015 for validation and hypothesis testing.

Virus phylogenetic data.—Viral genetic variability evaluated with a phylogenetic analysis using partial genomic sequences (P gene hypervariable region) have been previously described for the 2004–2006 and 2014–2015 outbreaks (Rainwater-Lovett et al. 2007, Velazquez-Salinas et al. 2014). Phylogenetic analyses were based on near full-length genomic sequences of representative viral strains from the two outbreaks. Alignments were conducted using the ClustalW algorithm implemented in MEGA v7.0.18 (Kumar et al. 2016). Models of nucleotide substitution were evaluated using the Model Testing tool in CLC Genomics Workbench where various statistical analyses were assessed (hLRT, BIC, AIC, and AICc) and the GTR + G model was implemented in a maximum likelihood (ML) phylogenetic reconstruction (www.qiagenbioinformatics.com).

Livestock density.—To capture host population densities, two livestock variables were acquired: number of horses and number of horse premises from tabular county census data in available years (2002, 2007, 2012) (Table 1). Broad-scale and long-term patterns were evaluated using the mean of yearly census data while fine-scale patterns were evaluated by linearly interpolating annual values. Lastly, county livestock density (number of horses or premises with horses per km^2; Table 2) was calculated as the average or yearly number of animals or premises divided by county area. Only the horse data for VS had sufficient numbers for analysis in all four years; data for other animals were too sparse for multi-year analyses. These averages were rasterized and included in the raster dataset and yearly values were incorporated into the tabular dataset.

Since proximity to VS occurrences and horse density are likely to facilitate transmission, we calculated the number of neighbors with VSV using the following procedure for the local analysis. For each VS occurrence, the number of horses

with VSV in the neighborhood was counted during the occurrence year, defined as 1 March through February, and during the prior year. A neighborhood was defined by centering a 36 km × 36 km grid (9 × 9 cells) on an occurrence location.

Pedology.—Two variables were used to capture soil properties that influence small pools of standing water and favorable conditions for biting midges. Available water content (AWC) and percentage clay content in the top 20 cm were extracted from STATSGO (Table 1) and summarized by calculating weighted means by area from all soil components in each map unit. The polygon data were then rasterized and incorporated into the raster and tabular datasets.

Hydrology.—Three variables were used to capture hydrological conditions expected to lead to favorable environments for black flies. Daily stream flow data (streamflow: Table 1) were used to estimate the presence of water in ephemeral watercourses and to quantify the rate of flow (in cubic feet per second [cfs]). Distance to nonzero monthly, annual, and mean annual flow data, measured as Euclidean distance, were incorporated into the raster and tabular dataset. Euclidean distance to major rivers and lakes (USGS 2017) was included in the raster dataset only. Lastly, mean monthly stream flow calculated from the closest gauges (mean = 18.7 km) for the 11 months prior to a VS incident was incorporated into the tabular dataset.

Surface runoff and soil moisture were included (Table 1) to identify favorable conditions for black fly reproduction. Daily runoff (in kg/m^2) and soil moisture (in kg/m^2) values averaged for each month, season, and year were included in the raster dataset. Monthly averages for the prior 11 months were calculated and incorporated into the tabular dataset.

Topography.—Elevation (in m) was used to capture ecosystem variability resulting from topoclimatic processes. Elevation derived from digital elevation models (Table 1) was incorporated into the raster and tabular datasets.

Climatology.—Daily and monthly total precipitation (mm) and average maximum and minimum temperature (°C) data were used to calculate seasonal, annual, and 30-yr means to explore macro-scale climatic relationships with disease occurrence. In addition, weekly averages, calculated from daily data, and monthly averages relative to VS incident date (i.e., prior weeks or months) were extracted for each VS occurrence and incorporated into the tabular dataset. Severe departure from normal climatic conditions resulting in drought was captured by the evaporative demand drought index (EDDI: Table 1). Seasonal, annual, and long-term values were calculated with the longest available dataset following the same procedure used for the PRISM data and incorporated into the raster and tabular datasets.

Vegetation.—Variation in vegetation growth may be important to both host and vectors. To capture variability in green vegetation biomass, we included normalized difference vegetation index from MODIS (NDVI: Table 1).

Normalized environmental variables

In addition to raw data values, normalized values were calculated to limit sensitivity to differences in magnitude. Deviations from seasonal long-term means were calculated using 30-yr means for climate from the PRISM database for temperature and precipitation. For other variables, such as livestock density, streamflow, EDDI, NDVI, soil moisture, and runoff, the long-term averages were calculated over the duration of the time period of VSNJV incidence data (2004–2015).

Hypothesis testing

Objective 1: Landscape-to-regional-scale analysis.—We used Maxent (Phillips et al. 2006, Phillips and Dudík 2008) to model annual distributions of VS and to calculate relative occurrence rates (RORs) across the study area. Maxent has been used extensively to create and evaluate species distributions in covariate space, often represented geographically, across a broad range of biological applications using presence-only data (Elith et al. 2011). This machine-learning approach, based on the principle of maximum entropy, enables fitting of complex non-linear relationships and performs similarly or better than traditional general linear modeling approaches, particularly when only occurrence data are available (Elith et al. 2006).

For each Maxent analysis, 10 replicates were conducted using a 90% random subsample of the occurrence data for model training which were

then averaged together to create the final model. Because Maxent settings were optimized using the R package MaxentVariableSelection (Jueterbock et al. 2016), we were able to compare multiple models of annual VS occurrence using AIC. Model performance was assessed using the corrected Akaike information criterion (AICc) which provides a relative measure of model quality considering fit and complexity (Akaike 1998). Model selection criteria, such as AIC, are designed to minimize overfitting by penalizing excess complexity (Boxand Draper 1987, Burnham and Anderson 2004) which can result in poor model transferability (Chatfield 1995, Sarle 1995). Underfitting can also reduce transferability, especially when indirect predictor variables are incorporated in a model (Wenger and Olden 2012). Multi-model inference was applied in the development of annual models and only the best performing model was projected into the validation years. To ensure that models were sufficiently parsimonious but not underfit, we assessed the best performing model's transferability by non-randomly partitioning our data into temporally distinct subsets (years), which were then used for training (2004 and 2005) and validation (2014 and 2015). Our results provide insight into how the model will perform (transferability) under subsequent environmental scenarios (Vaughan and Ormerod 2005). We followed previous recommendations by interpreting the Maxent output as relative occurrence rate (ROR) (Merow et al. 2013). We evaluated variable importance using jackknife plots, variable response curves, and frequency distribution plots to test hypotheses about VSNJV occurrence.

To evaluate support for our virus or incursion–expansion hypotheses, variability in annual VSNJV occurrence patterns was explored by developing separate models for VS occurrences in 2004 (incursion) and 2005 (expansion). The 2004 and 2005 models were then cross-evaluated by projecting each onto 2014 and 2015 environmental conditions. The ability of each model to predict subsequent events and event specificity (e.g., whether or not an incursion model [2004] can predict a future phylogenetically unrelated incursion event [2014] or a phylogenetically related expansion event [2005]) was then evaluated using the mean predicted ROR values at occurrence and randomly selected background locations and by visually inspecting the geographic representation of ROR. Degree of niche similarity between models was determined using Schoener's *D* and Warren's *I*.

Objective 2: Local-scale analysis.—A tabular dataset was developed from the raster maps (118 variables) such that fine-scale relative temporal relationships (weeks or months prior to a VS occurrence) were preserved between occurrence and environmental variables. Cell size (4 km^2) was selected to characterize the environment surrounding a VS occurrence and included only the weekly and monthly data associated with the 1393 unique combinations of sampling locations and "occurrence" month-week. To capture the degree of deviation from normal conditions, we analyzed the standardized values:

$$\text{standardized value}_t = (V_t - \overline{V}) / \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

where V represents a variable at time interval t and x represents each value of the dataset. Standardization enabled comparison among years at the same location, reduced problems associated with input scale, reduced issues with multicollinearity, and allowed interpretation in the normal manner since regression coefficients are identical. All spatial data were exported to tabular data by calculating the mean of values from a 100-point grid centered on a VS occurrence location.

This process was performed for two purposes. *First*, to compare individual contributions of explanatory variables in order to explain temporal variation of VSNJV occurrence. We conducted logistic regression of VSNJV occurrence (1 or 0) against one explanatory variable at a time to select 20 explanatory variables with highest maximum rescaled R^2. Correlated explanatory variables (Pearson's $r \geq 0.7$) were prevented from co-occurring in the multivariate model. *Second,* daily occurrence data were converted to Julian month, and least-squares regression was used to evaluate the relationship between county occurrence and latitude, elevation, and long-term precipitation. Predictor variables were selected based on the highest R^2. The resulting model was used to generate a map of estimated onset dates for the entire study area at 1-km^2 resolution using the

rasterized environmental data. All local-scale statistical analyses were conducted in SAS v.9.4.

Results and Discussion

Landscape-to-regional-scale analysis

Spatial distribution models based on human-guided machine learning and constructed using the geo-referenced, harmonized maps showed that of the 472 possible variables (Table 1), only four environmental drivers and host density were needed to explain patterns in VS occurrence in 2004 and 2005 (Table 3). Remarkably, these drivers represent all but one of the possible classes of environmental drivers included in the analysis; only large-scale drought is missing when both years are combined (Fig. 3, Table 3). Hydrology, vegetation, and climate were important in both years, and elevation was important in 2004 while soil properties were important in 2005. These few drivers and the variables contained within them provide both technological and biological complexity that yield new insight into factors governing spatial patterns in VS occurrence.

First, there is high diversity in data sources and types in these variables that required a big data approach of creating derived data products using decisions about spatial or temporal aggregation and harmonization of online national datasets prior to analysis. Detailed climate data in time and space, digital elevation models, digital soil maps and accompanying properties, remotely sensed imagery for vegetation, and distance calculations from each premise to the nearest stream were handled differently before analysis and often required domain technical expertise for interpretation. Although most previous studies typically included one or a few of these variables, and most studies focused on climate or elevation (e.g., Rodríguez et al. 1996, McCluskey et al. 2003), our findings clearly show the importance of including all of these variables in a trans-disciplinary approach to vector-borne diseases across large spatial extents.

Second, these few drivers are associated with different processes, multiple levels of biological organization (host, vectors), and different vectors that reflect different variables (Table 1). High horse density, proximity to streams with water, and high green vegetation during the summer were important to VS occurrence in both the incursion (2004) and expansion years (2005) (Tables 3, 4). These variables may define general habitat characteristics for VS based on horses as known hosts, and black flies, a known vector, that are biologically bound to flowing streams for oviposition, hatching, and larval development (Adler and McCreadie 1997). In the arid and semiarid Western United States, an increase in green vegetation during pre-monsoonal summer months is often found spatially distributed along streams and other water bodies.

Because 2004 and 2005 have similar viral phylogenies (Rainwater-Lovett et al., 2007), we were able to test whether spatial patterns are determined either by viral genetic determinants alone (viral hypothesis) or whether environmental variables are needed to explain different patterns in occurrence in incursion (e.g., 2004) vs. expansion years (e.g., 2005) (incursion–expansion hypothesis). Results show that different environmental variables were needed to explain patterns of VS occurrence in each year (Table 3), in support of the incursion–expansion hypothesis. In 2004, high horse density (0.82 animals/km^2), locations at moderately high elevations (mean = 1642 m) with low spring, prior winter, and prior fall precipitation, and close proximity to streams (17 km) were the most important variables to patterns in VS occurrence (Tables 3, 4). Elevation has been implicated as an explanatory variable for landscape-scale variation in VS occurrence in Mexico and may be a surrogate for a combination of vector–environment interactions (Rodríguez et al. 1996) or may contribute to stream flow dynamics that affect the black fly life cycle. Hydrological patterns are also related to patterns in VS occurrence in the Western United States (Elias et al. 2019). The combination of these factors and the limited spatial distribution of VS occurrences in 2004 are consistent with the hypothesis that black flies might be the principal vector dispersing VS northward along streams or flowing water canals in incursion years.

In 2005, summer conditions of low precipitation, cooler than average temperatures, and high green vegetation along with higher than average rainfall in the fall, close proximity to streams, and soils with moderately high available water holding capacity (AWC) were the most important variables associated with VS cases (Tables 3,

Table 3. Significant variables in VS occurrence patterns in incursion (2004) and expansion (2005) years.

Variable	Temporal unit	2004† Percentage contribution (rank)	2004 Permutation importance	Sum (rank)	2005‡ Percentage contribution (rank)	2005 Permutation importance§	Sum (rank)
Horse density (animals/km^2)	5-yr interval survey data	32 (1)	18	50 (1)	8 (7)	3	11
Elevation (m)	static DEM	16 (2)	16	36 (3)			
Distance to nearest stream (km)	static maps	15 (5)	6	21 (5)	10 (5)	8	18 (5)
AWC (cm)	static maps				8 (6)	3	11
Vegetation greenness	summer				11 (4)	15	26 (3)
	[dev], summer	7 (7)	7	14			
Precipitation (mm/month)	spring	14 (3)	19	26 (4)			
	summer			0	29 (1)	28	57 (1)
	winter	14 (4)	33	47 (2)			
	fall	7 (6)	1	8			
	[dev], fall				13 (3)	31	44 (2)
Temperature – maximum (°C)	[dev], summer				15 (2)	10	25 (4)
	[dev], winter				6 (8)	1	7

Notes: The percentage contribution (with rank of importance in parentheses) and permutation importance of each variable from MaxEnt analysis in explaining spatial variation in each year are shown. AWC, available water holding capacity of the soil; dev, deviation from long-term mean.

† Beta multiplier = 3.0 optimized from the R package and run in MaxEnt; (rank).

‡ Beta multiplier = 2.5 optimized from the R package and run in MaxEnt; (rank).

§ The permutation contribution is the decrease upon removal of a variable from the model. A large decrease indicates that the model depends heavily on that variable. Values are normalized to sum to 100.

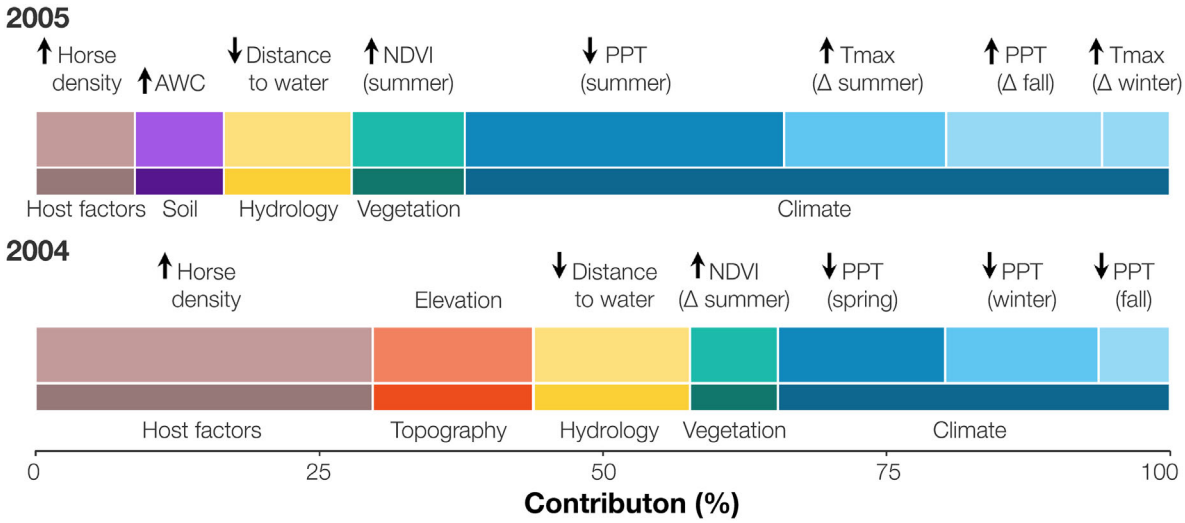


Fig. 3. Significant variables in VS occurrence patterns by driver class in incursion (2004) and expansion (2005) years. The relative percentage contribution of each variable in explaining spatial variation in each year is shown.

4) These conditions would promote an additional insect vector that requires small patches of shallow water or waste-enhanced mud along the edges of bodies of water for successful breeding habitat (e.g., biting midges; *Culicoides* spp.) (Mullens and Rodriguez 1988). Numbers of competent female midges peak in late summer and early fall (Mullens and Rodriguez 1988, Mullens

Table 4. Mean predicted relative occurrence rates for important variables (from MaxEnt analysis in Table 1) in occurrence (o) and background (bg) locations for incursion (2004, 2014) and expansion (2005, 2015) years.

Variable	Temporal unit	2004		2005		2014		2015	
		o	bg	o	bg	o	bg	o	bg
Horse density	5-yr interval survey data	0.82	0.36	0.49	0.36	1.11	0.36	0.66	0.36
Elevation	static DEM	1641.94	1384	1536.32	1384	1375.13	1384	1657.15	1384
Distance to nearest stream	static maps	17.35b	28.80	16.05b	28.80	22.57a	28.80	17.87b	28.80
AWC	static maps	12.10	10.67	12.51	10.67	13.17	10.67	12.18	10.67
Vegetation greenness	summer	0.45	0.38	0.43	0.40	0.51	0.38	0.49	0.41
	[dev], summer	0.02b	0.003	0.02b	0.02	0.03b	0.0006	0.04a	0.03
Precipitation	spring	40.70	41.50	34.85	43.97	51.56	36.11	65.86	69.60
	summer	54.63a	54.39	27.17c	47.91	55.28a	55.85	49.28b	52.49
	winter	15.40c	26.20	24.90a	36.0	16.5c	17.72	20.9b	27.45
	fall	37.63a	59.08	27.91b	29.51	37.38a	41.95	31.22b	48.18
	[dev], fall	8.89a	19.58	0.84c	-9.98	3.28b	2.45	1.40c	8.68
Temperature, maximum	[dev], summer	-1.25c	-1.06	-0.36a	0.08	-0.67b	-0.06	-0.32a	-0.11
	[dev], winter	0.26	-0.12	1.58	0.79	-1.01	-0.31	1.23	0.83

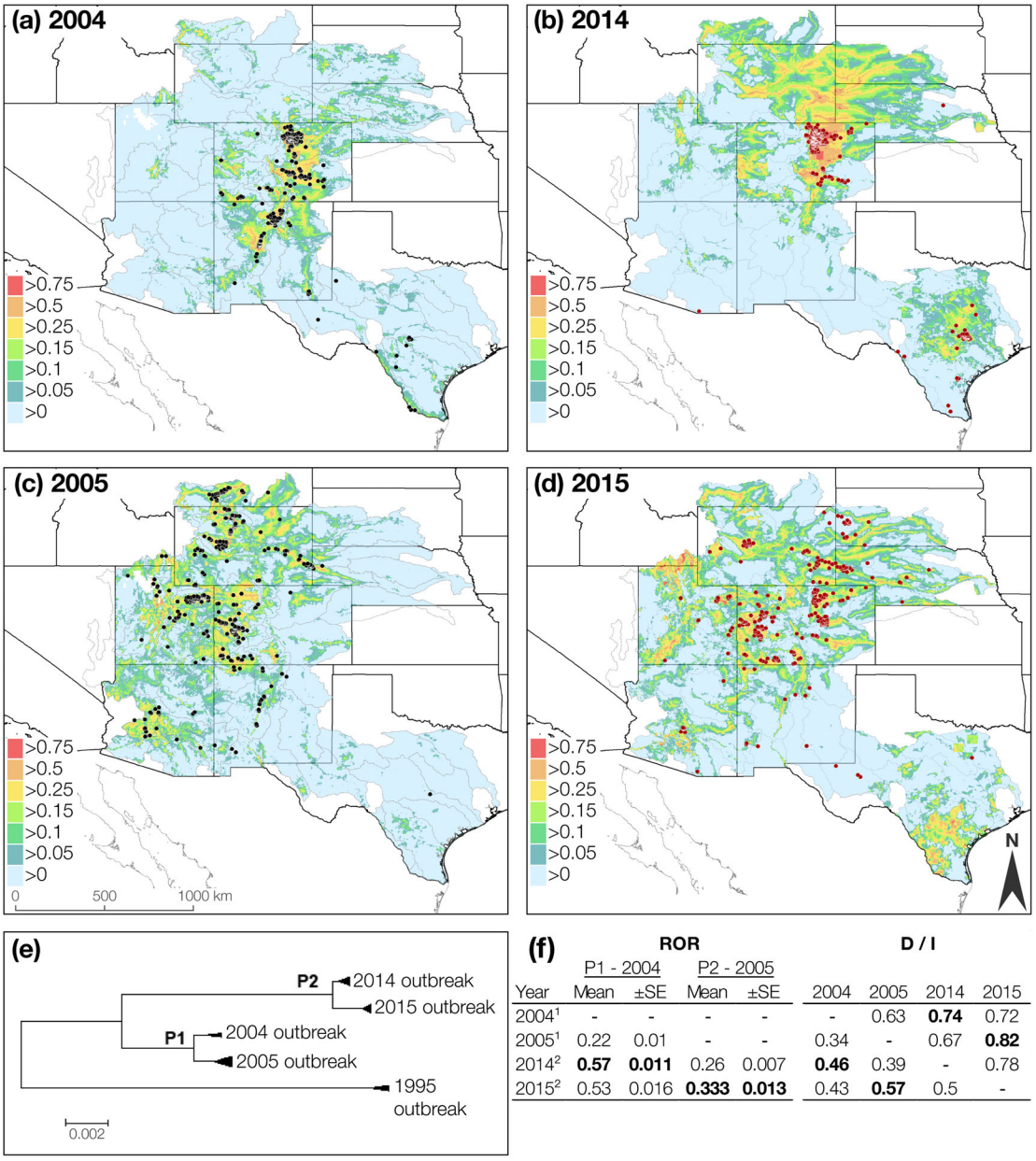
Note: Variables are defined in Table 1.

1989, Pfannenstiel and Ruder 2015), and have been associated with variations in local climate (Stallknecht et al. 2015). Thus, we hypothesize that these differences in important variables between years is a shift from an insect vector (black flies) in incursion years where moving water is needed for successful breeding and transport of eggs and larvae, and could explain patterns in disease concentrated along rivers, to multiple insect vectors (e.g., black flies and midges) in expansion years that would allow widespread expansion of disease throughout the geographic area.

We further tested our hypotheses by comparing distribution models constructed from 2004 (or 2005) environmental data and VS occurrence data with a second event a decade later (2014 incursion; 2015 expansion) caused by a phylogenetically different viral lineage (R. M. Palinski, S. J. Pauszek, N.D. Burruss et al., *unpublished data*). Our results show that the two incursion years (2004, 2014) were most similar, and the two expansion years were most similar in environmental conditions (Fig. 4f). The model created using 2004 environmental data (i.e., incursion model; Fig. 4a) had the highest ROR (0.57; Fig. 4f) and the greatest degree of niche overlap (Schoener's $D = 0.46$, Warren's $I = 0.74$) with the 2014 (Fig. 4b) environmental data in predicting VS occurrences in that year more so than with data from the same viral lineage in an expansion year (2005; Fig. 4c) or with a different viral lineage (2015; Fig. 4d). Using the 2005 model (Fig. 4c), the highest ROR (0.33) and greatest niche overlap (Schoener's $D = 0.74$, Warren's $I = 0.82$) were found with 2015, the other expansion year, yet with a different viral lineage (Fig. 4d). Both incursion years (2004, 2014) had lower winter precipitation, higher summer and fall precipitation, and cooler long-term summer maximum temperatures compared with both expansion years (2005, 2015) (Fig. 5). These results indicate that changes in viral genetic lineage were less important to patterns in VS occurrence than factors affecting the host–vector–environment interaction in incursion (2004, 2014) and expansion (2005, 2015) years. Thus, vector–environment interactions had a stronger control on patterns in disease occurrence than viral phylogeny.

Local-scale conditions preceding VS incursion and spread

Similar to the large-scale analysis, there was some overlap in on-site environmental conditions that represent incursion and expansion years, but there were also distinct differences (Fig. 6). These results provide further support for two vectors: black flies predominating in the incursion phase when few infected animals were present, and biting midges being the dominant vector in the expansion phase when far more infected animals were present. This relationship with the host has been observed previously as



(f)

	ROR				D / I			
	P1 - 2004		P2 - 2005					
Year	Mean	±SE	Mean	±SE	2004	2005	2014	2015
2004[1]	-	-	-	-	-	0.63	**0.74**	0.72
2005[1]	0.22	0.01	-	-	0.34	-	0.67	**0.82**
2014[2]	**0.57**	**0.011**	0.26	0.007	**0.46**	0.39	-	0.78
2015[2]	0.53	0.016	**0.333**	**0.013**	0.43	**0.57**	0.5	-

Fig. 4. Distribution of VS occurrences compared with MaxEnt modeled distributions at the landscape-to-regional scale. Model parameters are treated as traits to calculate niche overlap. (a) 2004 modeled values (colors) with 2004 VS occurrence point data, (b) 2004 modeled values (colors) using 2014 environmental data showing 2014 VS occurrence point data points for validation, (c) 2005 modeled values (colors) with 2005 VS occurrence point data, (d) 2005 modeled (colors) using 2015 environmental data showing 2015 VS occurrence point data for validation, and (e) phylogenetic tree reconstructed using isolates from recent US outbreaks. The subtree labeled P1 contains the isolates from 2004-2005 indicated by black dots in (a) and (c) while the subtree labeled P2 contains the isolates from 2014-2015 indicated by red dots in (b) and (d). (f) Assessment of each model's ability to predict future events was evaluated with mean (and standard error [SE]) predicted relative occurrence rates (ROR) at VS locations during subsequent years of infection. Degree of niche similarity between models based on Schoener's *D* (lower left triangle) and Warren's *I* (upper right triangle).

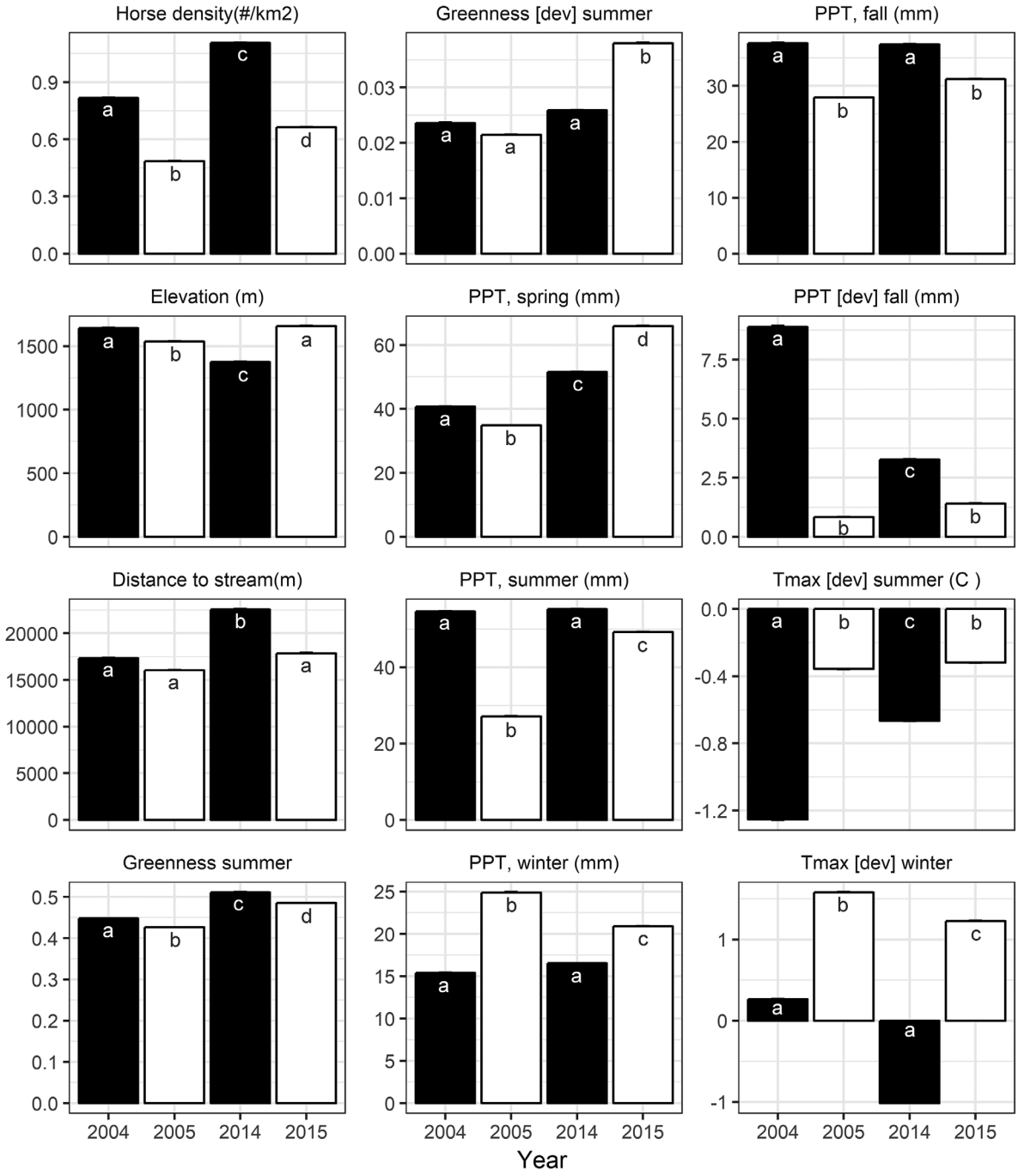


Fig. 5. Comparisons of mean environmental conditions between incursion (2004, 2014; filled bars) and expansion years (2005, 2015; open bars) at VS occurrence locations: horse density, elevation, distance to nearest stream, vegetation greenness in summer, normalized vegetation greenness in summer, spring precipitation, summer precipitation, winter precipitation, fall precipitation, normalized fall precipitation, normalized maximum summer temperature, and (normalized maximum winter temperature. Letters indicate significant differences between years ($P < 0.05$).

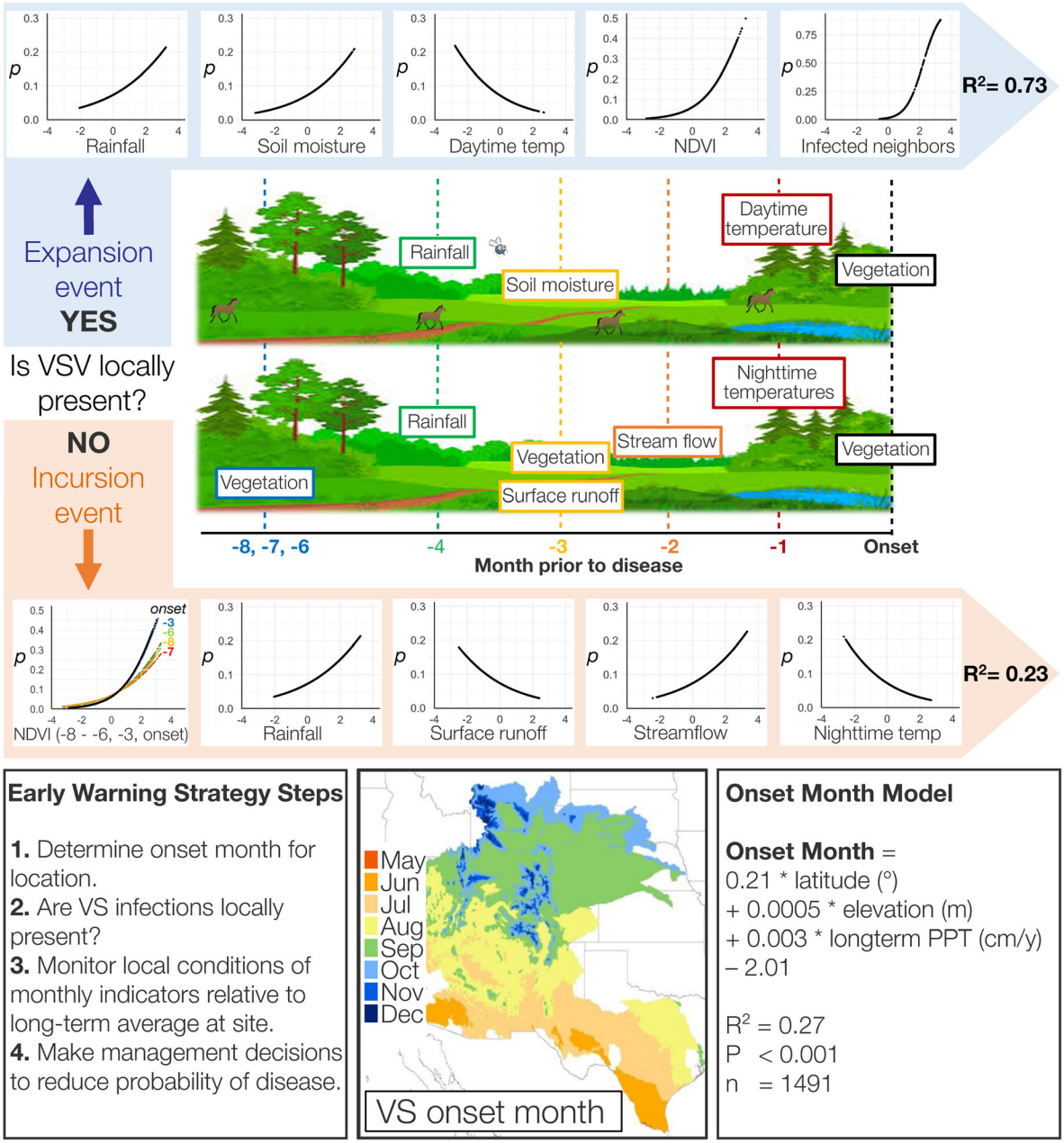


Fig. 6. The sequence of events that precedes the occurrence of VS at scale of an individual premise in incursion (lower) and expansion years (upper panel). Mean VS onset month calculated for any location in the region based on latitude, elevation, and long-term precipitation (R^2 = 0.27, P < 0.001, n = 1491). Given this date, a livestock or equine owner can monitor their local conditions to determine the likelihood that VS will occur in each month of that year. Univariate plots illustrate the relationship between predicted probability of occurrence (P) and standardized environmental conditions.

horses (and other livestock) within 0.4 km of running water (which presumably placed them at higher exposure to black flies) had a greater risk of contracting VS (Hurd et al. 1999, McCluskey et al. 1999). On-site green vegetation during the month of occurrence and higher rainfall four

months prior combined with either cool daytime (expansion) or nighttime (incursion) temperatures one month prior were common indicators of VS occurrence in both years (Fig. 6). The remainder of the variables were related to the putative vector: either black flies in incursion years (stream flow [2 months prior], surface runoff [3 months prior]) or midges in expansion years (soil moisture [3 months prior]) (Fig. 6). Our results are consistent with studies on black flies showing that numbers of viable eggs and larvae that survive the winter are dependent on dynamics of early season stream flow, water and air temperature, atmospheric humidity, and other factors (Mullens and Rodriguez 1988).

Developing early warning strategies for VS and other vector-borne diseases

Our multi-scale findings showed a temporal sequence of early warning indicators supported by a spatial analysis of important vectors that veterinarians and livestock owners can use as early warning strategies for VS occurrence throughout the Western United States. These indicators do not depend on knowledge of environmental variables that vary geographically, but rather depend on conditions relative to average conditions on a premise. Recently, an EWS was proposed using computer-based searches that relies on the public response triggered by the occurrence of the initial outbreaks (Wang et al. 2018). Our EWS consist of a statistical relationship to first estimate onset month of VS for a premise (Fig. 6), and then a set of indicators is selected based on disease phase (incursion, expansion) and vector identity (black fly, biting midges) (Fig. 6). Predictions that do not consider disease phase or vector identity are likely to be too general for use at finer spatial and temporal scales. Proactive measures to reduce exposure to vectors can then be implemented in advance, such as reducing on-site vegetation in years with cool summer temperatures, relocating susceptible animals away from streams and housing them in structures guarded from biting insects (Hurd et al. 1999) or implementing aggressive vector habitat mitigation strategies in locations with a high probability for VS expansion.

Prior to this analysis, specific knowledge about vector–environment interactions in incursion vs. expansion phases, and the identity of these early warning indicators were not available for VS. The hypotheses generated by this analysis are being used to guide field studies to sample vectors for VSV across environmental gradients in the Western United States. Collection of new data about vector–environment relationships, in particular in different years of a VS event, will improve future model predictions and ultimately aid in the refinement of the conceptual model under development for the VS system (Peters et al. 2018).

This big data approach coupled with human and machine learning can be applied to other emerging diseases for improvement in understanding, prediction, and management of vector-borne diseases (National Academies of Sciences Engineering and Medicine 2016). Translation of this knowledge can be made to improving animal and human welfare, and aiding food security to assist in development of early warning strategies that are currently based primarily on climate (Muñoz et al. 2016) or environmental predictions based on only a few variables (Lo Iacono et al. 2018).

Acknowledgments

This work was supported by USDA-ARS CRIS Projects at the Jornada Experimental Range (#6235-11210-007), Plum Island Animal Disease Center (#8064-32000-058-00D), Center for Grain and Animal Health Research (#8064-32000-058-00D, #3020-32000-008-00D), and the Rangeland Resources and Systems Research Unit (#3012-21610-001-00D). Funding was provided by the National Science Foundation to New Mexico State University for the Jornada Basin Long Term Ecological Research Program (DEB 12-35828, 18-32194) and DEB 14-40166. We thank Mr. Darren James and Dr. Geovany Ramirez of NMSU for additional analyses in support of this study. We thank Palantir Technologies for their involvement in a pilot study that assisted in identifying challenges in working with the many diverse formats of the spatial data. We thank the USDA Office of the Chief Scientist for support of DPCP.

Debra Peters, Scott McVey, Emile Elias, Angela Pelzel-McCluskey, Justin Derner, and Luis Rodriguez designed the study and wrote the manuscript. Dylan Burruss, Scott Shrader, and Jin Yao performed spatial and/or temporal data harmonization, geo-referencing, and analysis. Debra Peters, Scott McVey, Emile Elias, Angela Pelzel-McCluskey, Justin Derner, Luis Rodriguez, Steven Pauszek, and Jason Lombard provided expertise to interpret data analysis and edited the

manuscript. All authors were critical to project coordination, and identification of data sources and interpretation of results from their specific expertise.

Literature Cited

Adler, P. H., and J. W. McCreadie. 1997. Insect life: the hidden ecology of black flies: sibling species and ecological scale. American Entomologist 43:153–162.

Akaike, H. 1998. Information theory and an extension of the maximum likelihood principle. Pages 199–213 *in* Selected papers of Hirotugu Akaike. Springer, New York, New York, USA.

Althouse, B. M., J. Lessler, A. A. Sall, M. Diallo, K. A. Hanley, D. M. Watts, S. C. Weaver, and D. A. T. Cummings. 2012. Synchrony of sylvatic dengue isolations: a multi-host, multi-vector SIR model of dengue virus transmission in Senegal. PLOS Neglected Tropical Diseases 6:e1928.

Althouse, B. M., N. Vasilakis, A. A. Sall, M. Diallo, S. C. Weaver, and K. A. Hanley. 2016. Potential for Zika virus to establish a sylvatic transmission cycle in the Americas. PLOS Neglected Tropical Diseases 10:e0005055.

Altizer, S., R. S. Ostfeld, P. T. J. Johnson, S. Kutz, and C. D. Harvell. 2013. Climate change and infectious diseases: from evidence to a predictive framework. Science 341:514–519.

Box, G. E., and N. R. Draper. 1987. Empirical model-building and response surfaces. Wiley, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference. Sociological Methods & Research 33:261–304.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society A (Statistics in Society) 158:419–466.

Cohen, J. M., D. J. Civitello, A. J. Brace, E. M. Feichtinger, C. N. Ortega, J. C. Richardson, E. L. Sauer, X. Liu, and J. R. Rohr. 2016. Spatial scale modulates the strength of ecological processes driving disease distributions. Proceedings of the National Academy of Sciences of the United States of America 113:E3359–E3364.

Cupp, E. W., C. J. Maré, M. S. Cupp, and F. B. Ramberg. 1992. Biological transmission of vesicular stomatitis virus (New Jersey) by *Simulium vittatum* (Diptera: Simuliidae). Journal of Medical Entomology 29:137–140.

Elias, E., D. S. McVey, D. P. C. Peters, J. D. Derner, A. Pelzel-McCluskey, T. S. Schrader, and L. Rodriguez. 2019. Contributions of hydrology to Vesicular Stomatitis virus emergence in the western USA. Ecosystems 22:416–433.

Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17:43–57.

Escobar, E. E., and M. E. Craft. 2016. Advances and limitations of disease biogeography using ecological niche modeling. Frontiers in Microbiology 7:1174.

Faria, N. R., et al. 2017. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature 546:406–410.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature 457:1012–1014.

Han, B. A., and J. M. Drake. 2016. Future directions in analytics for infectious disease intelligence. Science and Society. EMBO Reports 17:785–789.

Hurd, H., B. McCluskey, and E. Mumford. 1999. Management factors affecting the risk for vesicular stomatitis in livestock operations in the western United States. Journal of the American Veterinary Medical Association 215:1263–1268.

Jacquot, M., K. Nomikou, M. Palmarini, P. Mertens, and R. Biek. 2017. Bluetongue virus spread in Europe is a consequence of climatic, landscape and vertebrate host factors as revealed by phylogeographic inference. Proceedings of the Royal Society B: Biological Sciences 284:20170919.

Jueterbock, A. 2015. R package MaxentVariableSelection: selecting the best set of relevant environmental variables along with the optimal regularization multiplier for Maxent niche modeling. https://cran.r-project.org/web/packages/MaxentVariableSelection/index.html

Jueterbock, A., I. Smolina, J. A. Coyer, and G. Hoarau 2016. The fate of the Arctic seaweed *Fucus distichus* under climate change: an ecological niche modeling approach. Ecology and Evolution 6:1712–1724.

Kramer, W. L., R. H. Jones, F. R. Holbrook, T. E. Walton, and C. H. Calisher. 1990. Isolation of arboviruses from *Culicoides* midges (Diptera: Ceratopogonidae) in Colorado during an epizootic of vesicular stomatitis New Jersey. Journal of Medical Entomology 27:487–493.

Kumar, S., G. Stecher, and K. Tamura. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution 33:1870–1874.

Lo Iacono, G., A. A. Cunningham, B. Bett, D. Grace, D. W. Redding, and J. L. N. Wood. 2018. Environmental limits of Rift Valley fever revealed using

ecoepidemiological mechanistic models. Proceedings of the National Academy of Sciences of the United States of America 115:E7448.

Mayer, S. V., R. B. Tesh, and N. Vasilakis. 2017. The emergence of arthropod-borne viral diseases: a global prospective on dengue, chikungunya and Zika fevers. Acta Tropica 166:155–163.

McCluskey, B. J., B. J. Beaty, and M. D. Salman. 2003. Climatic factors and the occurrence of vesicular stomatitis in New Mexico, United States of America. Reviews in Science and Technology 22:849–856.

McCluskey, B. J., H. S. Hurd, and E. L. Mumford. 1999. Review of the 1997 outbreak of vesicular stomatitis in the western United States. Journal American Veterinary Medical Association 215:1259–1262.

Merow, C., M. J. Smith, and J. A. Silander. 2013. A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. Ecography 36:1058–1069.

Messina, J. P., et al. 2014. Global spread of dengue virus types: mapping the 70 year history. Trends in Microbiology 22:138–146.

Michael, E., B. K. Singh, B. K. Mayala, M. E. Smith, S. Hampton, and J. Nabrzyski. 2017. Continental-scale, data-driven predictive assessment of eliminating the vector-borne disease, lymphatic filariasis, in sub-Saharan Africa by 2020. BMC Medicine 15:176.

Morin, C. W., and A. C. Comrie. 2013. Regional and seasonal response of a West Nile virus vector to climate change. Proceedings of the National Academy of Sciences of the United States of America 110:15620–15625.

Mullens, B. A. 1989. A quantitative survey of *Culicoides variipennis* (Diptera: Ceratopogonidae) in dairy wastewater ponds in southern California. Journal Medical Entomology 26:559–565.

Mullens, B. A., and J. L. Rodriguez. 1988. Colonization and response of *Culicoides variipennis* (Diptera: Ceratopogonidae) to pollution levels in experimental dairy wastewater ponds. Journal of Medical Entomology 25:441–451.

Muñoz, A. G., M. C. Thomson, L. Goddard, and S. Aldighieri. 2016. Analyzing climate variations at multiple timescales can guide Zika virus response measures. GigaScience 5:1–6.

National Academies of Sciences Engineering and Medicine. 2016. Big data and analytics for infectious disease research, operations, and policy: proceedings of a workshop. Page 98 *in* Big data and analytics for infectious disease research, operations, and policy: proceedings of a workshop. The National Academies Press, Washington, D.C., USA.

Parham, P. E., et al. 2015. Climate, environmental and socio-economic change: weighing up the balance in vector-borne disease transmission. Philosophical Transactions of the Royal Society B: Biological Sciences 370. https://doi.org/10.1098/rstb.2013.0551

Peters, D. P. C., et al. 2018. An integrated view of complex landscapes: a big data-model integration approach to trans-disciplinary science. BioScience 68:653–669.

Pfannenstiel, R. S., and M. G. Ruder. 2015. Colonization of bison (*Bison bison*) wallows in a tallgrass prairie by *Culicoides* spp (Diptera: Ceratopogonidae). Journal of Vector Ecology 40:187–190.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. Ecography 31:161–175.

Racloz, V., R. Ramsey, S. Tong, and W. Hu. 2012. Surveillance of dengue fever virus: a review of epidemiological models and early warning systems. PLOS Neglected Tropical Diseases 6:e1648.

Rainwater-Lovett, K., S. J. Pauszek, W. N. Kelley, and L. L. Rodriguez. 2007. Molecular epidemiology of vesicular stomatitis New Jersey virus from the 2004–2005 US outbreak indicates a common origin with Mexican strains. Journal of General Virology 88:2042–2051.

Rodríguez, L. L. 2002. Emergence and re-emergence of vesicular stomatitis in the United States. Virus Research 85:211–219.

Rodríguez, L. L., T. A. Bunch, M. Fraire, and Z. N. Llewellyn. 2000. Re-emergence of vesicular stomatitis in the western United States is associated with distinct viral genetic lineages. Virology 271:171–181.

Rodríguez, L. L., W. M. Fitch, and S. T. Nichol. 1996. Ecological factors rather than temporal factors dominate the evolution of vesicular stomatitis virus. Proceedings of the National Academy of Sciences of the United States of America 93:13030–13035.

Sarle, W. S. 1995. Stopped training and other remedies for overfitting. Interface Foundation of North America, Fairfax Station, Virginia, USA.

Stallknecht, D. E., A. B. Allison, A. W. Park, J. E. Phillips, V. H. Goekjian, V. F. Nettles, and J. R. Fischer. 2015. Apparent increase of reported hemorrhagic disease in the Midwestern and Northeastern USA. Journal of Wildlife Diseases 51:348–361.

Stewart-Ibarra, A. M., and R. Lowe. 2013. Climate and non-climate drivers of Dengue epidemics in Southern Coastal Ecuador. American Journal of Tropical Medicine and Hygiene 88:971–981.

Sule, W. F., D. O. Oluwayelu, L. M. Hernández-Triana, A. R. Fooks, M. Venter, and N. Johnson. 2018. Epidemiology and ecology of West Nile virus in sub-Saharan Africa. Parasites & Vectors 11:414.

Tomley, F. M., and M. W. Shirley. 2009. Livestock infectious diseases and zoonoses. Philosophical Transactions of the Royal Society B: Biological Sciences 364:2637–2642.

USGS 2017. North America Rivers and Lakes. U.S. Geological Survey National Water Information System. https://waterdata.usgs.gov/nwis.

Vaughan, I. P., and S. J. Ormerod. 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology 42:720–730.

Velazquez-Salinas, L., S. J. Pauszek, S. Zarate, F. J. Basurto-Alcantara, A. Verdugo-Rodriguez, A. M. Perez, and L. L. Rodriguez. 2014. Phylogeographic characteristics of vesicular stomatitis New Jersey viruses circulating in Mexico from 2005 to 2011 and their relationship to epidemics in the United States. Virology 449:17–24.

Walsh, M., and M. A. Haseeb. 2015. Modeling the ecologic niche of plague in sylvan and domestic animal hosts to delineate sources of human exposure in the western United States. PeerJ 3:e1493.

Walton, T. E., P. A. Webb, W. L. Kramer, G. C. Smith, T. Davis, F. R. Holbrook, C. G. Moore, T. J. Schiefer, R. H. Jones, and G. C. Janney. 1987. Epizootic vesicular stomatitis in Colorado, 1982: epidemiologic and entomologic studies. American Journal of Tropical Medicine and Hygiene 36:166–176.

Wang, J., T. Zhang, Y. Lu, G. Zhou, Q. Chen, and B. Niu. 2018. Vesicular stomatitis forecasting based on Google Trends. PLOS ONE 13:e0192141.

Wenger, S. J., and J. D. Olden. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3:260–267.

Woolhouse, M. E., K. Adair, and L. Brierley. 2013. RNA viruses: a case study of the biology of emerging infectious diseases. Microbiology Spectrum 1:83–97.