Identifying Environmental Risk Factors and Mapping the Distribution of West Nile Virus in an Endemic Region of North America

A. Hess¹, J. K. Davis¹, and M. C. Wimberly¹

¹Department of Geography and Environmental Sustainability, University of Oklahoma, Norman, OK, USA

Abstract Understanding the geographic distribution of mosquito-borne disease and mapping disease risk are important for prevention and control efforts. Mosquito-borne viruses (arboviruses), such as West Nile virus (WNV), are highly dependent on environmental conditions. Therefore, the use of environmental data can help in making spatial predictions of disease distribution. We used geocoded human case data for 2004–2017 and population-weighted control points in combination with multiple geospatial environmental data sets to assess the environmental drivers of WNV cases and to map relative infection risk in South Dakota, USA. We compared the effectiveness of (1) land cover and physiography data, (2) climate data, and (3) spectral data for mapping the risk of WNV in South Dakota. A final model combining all data sets was used to predict spatial patterns of disease transmission and characterize the associations between environmental factors and WNV risk. We used a boosted regression tree model to identify the most important variables driving WNV risk and generated risk maps by applying this model across the entire state. We found that combining multiple sources of environmental data resulted in the most accurate predictions. Elevation, late-season humidity, and early-season surface moisture were the most important predictors of disease distribution. Indices that quantified interannual variability of climatic conditions and land surface moisture were better predictors than interannual means. We suggest that combining measures of interannual environmental variability with static land cover and physiography variables can help to improve spatial predictions of arbovirus transmission risk.

Plain Language Summary West Nile virus, a disease transmitted by mosquitoes, is often considered a tropical disease. However, within the United States, the most cases per capita are reported in the northern Great Plains, with South Dakota having the highest rate. In this study, we combined human case data with satellite-based Earth observations and other environmental data to highlight the places in South Dakota with high West Nile virus risk and identify their environmental characteristics. Areas in northeastern South Dakota, specifically in the northern James River Valley, had the highest West Nile virus risk. These high-risk zones were associated with low-lying pasturelands and poorly drained soils in areas with high year-to-year variability in climate and land surface moisture.

1. Introduction

Understanding the geographic distribution of disease and mapping disease risk are important in supporting public health efforts to reduce the burden of infectious disease (Hay et al., 2013; Ostfeld et al., 2005; Pigott et al., 2015). Locating high-risk areas and identifying populations at risk can allow health officials to implement targeted disease prevention, control, and elimination efforts. In the case of mosquito-borne diseases, targeted interventions can improve the effectiveness of disease prevention and vector control campaigns. By knowing where disease hot spots are located, public authorities can apply insecticides for mosquito control primarily in high-risk areas and therefore efficiently decrease disease transmission (Bousema et al., 2012) while reducing concerns associated with the development of resistance against insecticides (Hemingway & Ranson, 2000). Additionally, targeting community outreach and public education encourages communities at risk to engage in prevention behaviors, as people who are aware of the risk in their neighborhood are more likely to eliminate potential breeding sites in their homes, apply insect repellent, dress appropriately to avoid bites, and avoid the outdoors during mosquito feeding hours (Mitchell et al., 2018). Such spatially targeted strategies at the community level are only possible if we can identify risk areas at a sufficiently high spatial resolution.
In this study, we focus on understanding and predicting fine-grained spatial patterns of human West Nile virus (WNV) risk. WNV is a mosquito-borne pathogen of global importance, as it is the most widely distributed encephalitic flavivirus (May et al., 2011) and the most widespread cause of arboviral neurological disease in the world (Chancey et al., 2015). It is primarily a zoonotic disease transmitted between birds as the main reservoir hosts and mosquitoes as vectors. Infections in horses and humans can occur as spillovers, but these organisms are dead-end hosts. First isolated in Uganda, the virus has spread across the globe (Chancey et al., 2015), including in the United States. After its first introduction in the United States in 1999, it dispersed rapidly and now covers a large geographical area and a wide ecological range (Reisen, 2013). In the United States, WNV has caused more than 45,000 reported human cases, including 22,913 cases of severe neuroinvasive disease and 2,138 deaths, as reported by the Centers for Disease Control and Prevention (CDC, 2016, 2017). Surprisingly, the highest incidence rates of WNV neuroinvasive disease within the country are reported in the Northern Great Plains, with South Dakota having the highest disease rate (Burakoff et al., 2018). Detailed WNV risk maps for high-risk states such as South Dakota will help public health authorities to implement strategies that can reduce transmission rates by allowing them to target those communities at highest risk.

Environmental conditions can have a large influence on the ecology of mosquito-borne diseases, such as WNV. For WNV to occur in humans, the environmental conditions need to be suitable for vector mosquitoes to breed, become infected from suitable avian hosts, survive long enough to become infectious, and finally transmit the virus to a susceptible human host. Vector species, host species, and the virus itself are sensitive to fluctuations in climate and landscape features. High temperature has been found to increase the risk of WNV transmission (Davis et al., 2017; Lockaby et al., 2016; Marcantonio et al., 2015), whereas the effects of humidity (Lebi et al., 2013) and precipitation (Wimberly et al., 2014) are more variable. Increasing temperature generally leads to higher mosquito growth rates and shortens the gonotrophic cycle (the time required to produce eggs after a blood meal) and the extrinsic incubation period (the time required for an infected mosquito to become infectious; Hartley et al., 2012; Kilpatrick et al., 2008; Reisen et al., 2006). However, high temperature can also decrease the longevity of female mosquitoes (Ciota et al., 2014). In addition to these temperature effects, humidity influences the survival of adult mosquitoes and their activity pattern, and precipitation influences the availability of breeding sites. Landscape features, such as topography, soils, and vegetation, also affect WNV transmission, as these variables determine the available habitats for mosquitoes and birds, and WNV risk is highest in areas with suitable habitat for mosquitoes and birds to complete the transmission cycle (Chuang et al., 2012). Depending on the geographic location and the particular vector and host species, these high-risk zones can include rural agricultural areas (Chuang et al., 2012; Reisen, 2013), wetlands (Sánchez-Gómez et al., 2017), and urban areas with small forest patches (Lockaby et al., 2016).

Because mosquito-borne disease transmission is sensitive to variations in habitat and climate, geospatial environmental data can be used to develop disease risk maps. In this approach, data on disease occurrence at specific locations are combined with predictor variables from geospatial data sets to develop predictive models (Peterson, 2014). These models are then applied across the entire geographic domain to produce predictions of disease risk at unsampled locations. Considerable advances in using environmental data for assessing disease risk have been made in recent years (Kraemer et al., 2016). The increasing availability of geospatial data sets such as gridded meteorological data, land cover maps, and digital elevation models now facilitates the analysis and prediction of disease risk across large areas ranging from landscapes to the globe (Ostfeld et al., 2005; Peterson, 2014). Numerous studies have used geospatial data sets and spatial modeling techniques to predict spatial patterns of vector-borne diseases, including tick-borne diseases (Wimberly, Baer, & Yabesley, 2008), WNV (Chuang et al., 2012; Peterson et al., 2008; Tran et al., 2014; Young et al., 2013), malaria (Benali et al., 2014; Midekisa et al., 2014), and chikungunya (Fischer et al., 2013).

One important source of environmental data is satellite remote sensing. Sensors carried by Earth-observing satellites provide information on reflected, emitted, and backscattered radiation that can be used to identify objects and phenomena on the Earth’s surface. Today, we have unprecedented access to freely available, long-term archives of high-quality satellite images with global coverage at spatial resolutions from tens to thousands of meters (Hansen & Loveland, 2012; Wulder et al., 2016). As a result, there has been increasing usage of satellite remote sensing data in large-scale environmental monitoring applications, including disease modeling. Satellite data can be used to compute a variety of spectral indices that are sensitive to environmental factors such as vegetation greenness, water, and land surface temperature. These indices characterize a wide range of environmental conditions and thus may be more effective at identifying the
specific vector and host habitats associated with disease transmission than other data sources such as classified land cover maps and coarser-grained meteorological data sets. Additionally, historical satellite archives allow us to observe long-term changes in land surface characteristics and measure their interannual and intraannual variability, as opposed to viewing land cover as a static condition. Thus, the use of satellite-derived spectral information can help to improve disease risk maps by providing novel sources of environmental information that are relevant to vector and host ecology.

With the ever-increasing availability of high-quality geospatial data, researchers face the challenge of deciding what data to include for their specific applications. The objective of this study is to explore the use of different types of geospatial environmental data for predicting the spatial patterns WNV disease risk in South Dakota. We compared three broad classes of geospatial data commonly used in previous studies: (1) land cover and physiography data, describing the physical structure, vegetation, and hydrology of the land surface at a single point in time; (2) climate data, derived from gridded and interpolated meteorological data; and (3) spectral indices derived from satellite sensors, measuring a variety of environmental conditions on the land surface. We sought to determine which of these data types produced the best fitting spatial model of human WNV cases and which specific environmental variables had the strongest influence on WNV patterns when all data types were included in a combined model. This approach allowed us to identify the best model for predicting the spatial distribution of WNV risk in South Dakota and yielded broader insights that can help to inform similar mapping efforts for different diseases in other locations.

2. Methods

The overall workflow involved acquiring and processing human WNV case data, exploratory analysis of the WNV case data, acquiring and processing environmental data, fitting boosted regression tree (BRT) models of WNV cases using the environmental variables, and applying these models to generate risk maps (Figure 1). These steps are described in detail in the following subsections.

2.1. WNV Case and Control Data

From the South Dakota Department of Health we received 1,381 serologically confirmed case records of West Nile fever, West Nile neuroinvasive disease, and WNV infected blood donor detections with spatial information for the years 2004–2017. Cases of West Nile fever and West Nile neuroinvasive diseases were diagnosed by clinicians and confirmed by laboratory techniques and meet the CDC case definition (CDC, 2015). In addition to symptomatic WNV cases, we also included cases of WNV-positive blood donors reported to the South Dakota Department of Health. After the first 2 years of introduction in 2002–2003, the annual number of WNV cases declined and the geographic distribution of cases shifted (Wimberly et al., 2013). These first 2 years were excluded from our analysis as the disease was not yet endemic, and the patterns were atypical from the following years.

Of the case records between the years 2004 and 2017, 1,257 contained geocoded coordinates of the home addresses of confirmed cases, an additional 121 data points contained at least zip code information, and only 3 cases had no spatial information. For those cases where only zip code information was available, we assigned the case data point to a random location within the given zip code tabulation area. The availability of spatial coordinates at the household level allowed us to assign a precise spatial location to most cases and avoided problems associated with summarizing cases within arbitrary spatial boundaries such as zip code tabulation areas and counties.

We used the technique of Chuang et al. (2012) to generate a set of 50,000 control points to represent the background distribution of the human population. We used census block level population information from the 2010 census to generate control points proportional to the population of each census block. We assigned these control points to random locations within each block. This population-based weighting controlled for the uneven population distribution and prevented a bias that would otherwise have skewed our results toward overpredicting risk in highly populated places.

2.2. Exploratory Analysis

To study the distribution of WNV cases in relation to the population distribution, we created a map of smoothed incidence rates. We applied a kernel smoothing technique to create a gridded layer where every grid cell contained density estimates for case and control points. The density estimates for each cell were
based on the numbers of case and control points within a set radius given by a smoothing parameter and their distances to the cell. As smoothing parameter, we chose a bandwidth of 0.5°. We used the R package kernSmooth to apply the kernel smoothing functions by Wand and Jones (1995). Then we created an approximate smoothed incidence rate by dividing the pixel values of the smoothed case grid by the control grid.

2.3. Environmental Variables

We compiled a variety of data sets that contain environmental information relevant to disease amplification and transmission. A list of the data used for the disease risk models is provided in Table 1. We grouped the data into three main types. Land cover and physiography data included a variety of standard national-level data products characterizing land cover and land use, topography, soils, and wetlands. Climate data included indices derived from a gridded meteorological data set. Spectral data included indices calculated from high temporal resolution remotely sensed surface reflectance data in the optical and infrared wavelengths. The specific variables derived from these data sets are listed in Table 2.

We retrieved land cover and physiography data for South Dakota from various national data sets. Land cover variables were derived from the 2011 National Land Cover Database (NLCD) at a 30-m resolution. From the initial NLCD land cover gridded data set, we created new data sets containing presence/absence information for each of the five land cover classes: developed areas (developed low intensity, developed medium, and developed high), forest (deciduous forest, evergreen forest, and mixed forest), grassland (grassland/herbaceous), pasture (pasture/hay), and cropland (cultivated crops). Data on wetland distribution were downloaded from the National Wetlands Inventory (NWI). We first converted the shapefiles to raster format and then created individual presence/absence layers for each of five wetland types: freshwater emergent wetland, freshwater pond, riverine, freshwater forested/shrub wetland, and lakes. We extracted data on

---

Figure 1. Workflow for West Nile virus (WNV) modeling and risk mapping. The analysis is divided into four major sections. The processing of case and control data, an exploratory analysis of case and control data, the processing of environmental variables, and the statistical modeling to produce WNV risk maps. Bold text represents the four main sections; italic text represents data processing steps, boxes, and arrows represent the main data flows. BRT = boosted regression tree.
ponding frequency from the Soil Survey Geographic Database. Ponding frequency is reported as percentage of the map unit that is subject to water ponded on the soil surface. Lastly, we downloaded data from the National Elevation Data Set at a 30-m resolution.

Climatic variables were derived from the North American Land Data Assimilation System atmospheric forcing data (Mitchell et al., 2004). We retrieved hourly temperature, precipitation, and specific humidity data for the years 2004–2017 from the Goddard Earth Sciences Data and Information Services Center. We calculated daily mean values for temperature (°C) relative humidity (%), and vapor pressure deficit (kPa), as well as daily total values for precipitation (mm) from the initial hourly data. Then we calculated yearly mean values for two parts of the WNV transmission season, as defined by Wimberly, Hildreth, et al., (2008). The early WNV season (May and June) encompasses the period of virus amplification in the mosquito-bird system prior to the occurrence

| Table 1 | Environmental Data Sets Used as Explanatory Variables |
|---------|--------------------------------------------------------|
| Data set | Spatial resolution/map scale | Temporal resolution | Time frame | Source |
| MODIS MCD43A4 (collection 6) | 500 m | Daily, based on 16-day moving window | 2004–2017 | https://urs.earthdata.nasa.gov/ |
| NLDAS | 1/8° (at 40° north, 13.9-km latitude, and 10.7 longitude) | Hourly | 2004–2017 | https://daac.gsfc.nasa.gov |
| NLCD | 30 m | NA | 2011 | https://datagateway.nrcs.usda.gov/ |
| NWI | 1:24,000 | NA | 2017 | https://www.fws.gov/wetlands/Data/Data-Download.html |
| SSURGO | 10 m | NA | 2018 | https://datagateway.nrcs.usda.gov/ |
| NED | 1 arcsec (approximately 30 m) | NA | n.d. | https://datagateway.nrcs.usda.gov/ |

Note. MODIS = Moderate Resolution Imaging Spectroradiometer; NLDAS = North American Land Data Assimilation System; NLCD = National Land Cover Database; NWI = National Wetlands Inventory; SSURGO = Soil Survey Geographic; NED = National Elevation Data Set; NA = not applicable; n.d. = no date.

| Table 2 | List of Environmental Variables and Their Abbreviations |
|---------|--------------------------------------------------------|
| Variable | Description | Data source |
| NDVI_a | Normalized Difference Vegetation Index (Bands: red, nir) | MODIS MCD43A4 Collection 6 |
| NDWI_a | Normalized Difference Water Index (Bands: green, nir) | MODIS MCD43A4 Collection 6 |
| NDWI2_a | Normalized Difference Water Index 2 (Bands: nir, swir2) | MODIS MCD43A4 Collection 6 |
| MNDWI_a | Modified Normalized Difference Water Index (Bands: green, swir2) | MODIS MCD43A4 Collection 6 |
| TcWet_a | Tasseled cap transformation wetness | MODIS MCD43A4 Collection 6 |
| TcGreen_a | Tasseled cap transformation greenness | MODIS MCD43A4 Collection 6 |
| TcBright_a | Tasseled cap transformation brightness | MODIS MCD43A4 Collection 6 |
| Prec_a | Total precipitation | NLDAS |
| RH_a | Relative humidity | NLDAS |
| VPD_a | Vapor pressure deficit | NLDAS |
| Temp_a | Daily mean temperature | NLDAS |
| Crop | Cultivated cropland | NLCD |
| Pasture | Pasture/hay | NLCD |
| Develop | Developed land | NLCD |
| Forest | Forest | NLCD |
| Grass | Grassland | NLCD |
| Lakes | Lakes | NWI |
| Ponds | Ponds | NWI |
| River | Riverine | NWI |
| ForestWetland | Forested/shrub wetland | NWI |
| Emergent | Emergent wetlands | NWI |
| Elevation | Elevation | NED |
| PondFr | Ponding frequency | SSURGO |

*aWe calculated four seasonal summaries for each of these variables. EarlyMean: May/June interannual mean, EarlySD: May/June interannual standard deviation, LateMean: May/June interannual mean, and LateSD: July/August interannual standard deviation. MODIS = Moderate Resolution Imaging Spectroradiometer; NLDAS = North American Land Data Assimilation System; NLCD = National Land Cover Database; NWI = National Wetlands Inventory; NED = National Elevation Data Set; SSURGO = Soil Survey Geographic.
of most human WNV cases. The late WNV season is when most mosquito-human transmission takes place and the majority of human cases occur. We calculated the interannual mean and standard deviation of each early-season variable and late-season variable.

We acquired spectral data from the Moderate Resolution Imaging Spectroradiometer (MODIS) MCD43A4 Collection 6 nadir bidirectional reflectance distribution function adjusted surface reflectance data set for the years 2004–2017. The imagery has a 500-m spatial resolution and comes as daily data summarized over a 16-day moving window. Only pixels using the full bidirectional reflectance distribution function inversion were used. On a pixel basis, we first calculated different environmental indices presenting different aspects of environmental conditions on the land surface. We calculated the normalized difference vegetation index (Tucker, 1979) to study the impact of vegetation on disease risk. Additionally, we calculated three water indices: the normalized difference water index which is sensitive to open surface water (McFeeters, 1996), a different normalized difference water index which is sensitive to vegetation liquid water (Gao, 1996), and the modified normalized difference water index which is useful for identifying open water in areas with high background noise (Xu, 2006). Additionally, we used the tasseled cap transformation, which converts spectral layers into interpretable bands conveying information on vegetation greenness, vegetation and soil wetness, and land surface brightness (Lobser & Cohen, 2007). The spectral indices and tasseled cap transformation were calculated in the R package RStoolbox. We then calculated the median of these indices in each year for the early and late West Nile transmission seasons defined previously. Finally, we calculated the interannual mean and standard deviation for these yearly indices.

We converted all data sets to the same spatial resolution. We chose 300 m as the common resolution, as a multiple of the NLCD and National Elevation Data Set data at 30-m resolution. For rescaling continuous variables, we used bilinear interpolation. For rescaling categorical variables, we aggregated pixels using a majority filter. To account for the uncertainty of the exact location of infection as given by the patient’s home address and to include environmental information in further vicinity of the case locations, we applied a moving window of 5 × 5 pixels. Examples of the gridded environmental layers are shown in Figure 2.

2.4. Statistical Modeling

We fitted an ensemble of BRT models to predict the spatial distribution of WNV transmission risk and to study the influence of the different environmental factors, using the gbm package in the statistical software R, version 3.4.4 (R Core Team, 2018).

BRT models are widely used in applications to predict species distribution (Elith et al., 2008; Elith & Graham, 2009) and increasingly also to predict spatial patterns in disease transmission (Hay et al., 2013). BRT models are useful in these applications due to their high predictive accuracy and their ability to capture nonlinear relationships between predictor and response variables and interactions among predictor variables. We created four different models based on different sets of environmental data, (1) land cover and physiography data, (2) climate data, (3) spectral data, and (4) a combination of all data sets.

To fit the BRT models, we created a data set of all 1,374 disease case points and a set of 12,366 control points (9 times the number of all case points) that were randomly sampled from our previously generated control data set. Pixel information was extracted from the environmental data sets, at each sampled case and control point location. We initially fit the BRT model using the same approach as Elith et al. (2008) with a tree complexity of 5, a learning rate of 0.01, and a bag fraction of 0.5. We compared the cross-validated predictive performance of the model under different parameter combinations with tree complexity ranging from 3 to 6, learning rate from 0.01 to 0.001, and bag fraction from 0.5 to 0.7 to select the optimal parameters for each model.

To evaluate the predictive accuracy of each model, the data were split into subsets for training (80%) and validation (20%). To account for variation in the model results that arise from control point sampling, we fitted each model 20 times, using a different subset of control points and a different random assignment of training and validation subsets each time. For each model run, we used all available case points and sampled new control points (with replacement) from the complete control data set. We then used the fitted BRT model object to produce spatial predictions across the entire state. We calculated the mean predicted value for each grid cell from all 20 predictions for each pixel to characterize the spatial distribution of WNV transmission risk.
We converted the case probabilities to percentiles to visualize the relative risk of WNV transmission across the state. To quantify the relative influence of different environmental variables, we ranked their relative importance during the model fitting process. Relative importance was measured as the normalized reduction of squared error attributable to each variable. We then created partial dependence plots for the most important variables to examine the relationship between environmental variables and disease risk. For accuracy assessment, we used the predicted values and the validation data set to calculate values for the area under the receiver operating characteristic curve (AUC) for each model run and calculated a mean AUC for the 20 model runs.

Figure 2. Environmental variables mapped across South Dakota. (a) Elevation, (b) 14-year standard deviation of relative humidity in July/August, (c) 14-year mean of relative humidity in July/August, (d) 14-year standard deviation of tasseled cap wetness in May/June, (e) percent pasture, (f) 14-year standard deviation of precipitation in May/June, (g) 14-year mean of precipitation for July/August, and (h) ponding frequency.
3. Results

3.1. Spatial Patterns

Figure 3 shows the spatially smoothed distribution of WNV cases, control points, and incidence rates for South Dakota. We observed the largest number of cases (Figure 3a) in areas with high population densities (Figure 3b). However, the incidence rate, calculated as cases divided by population, was relatively low in most urban areas and highest in the northern James River Valley located in northeastern South Dakota. There was another, smaller, area of high incidence just north of the Black Hills in northwestern South Dakota.

Figure 3. (a) Spatially smoothed West Nile virus case points, (b) population-weighted control points, and (c) incidence rates of human West Nile virus disease in South Dakota for 2004–2017.
All three risk maps in Figure 4 showed a similar pattern in highlighting the James River Valley stretching from northeastern to southeastern South Dakota and area of high risk. Also, all models showed an east-west gradient with areas east of the Missouri River having higher risk (Figure 4).

The model based on the land cover and physiography data (Figure 4a) clearly highlighted the whole James River Valley with the areas of highest relative risk all located in the valley. The risk gradually decreased with increasing elevation above and distance from the river. The areas west of the Missouri River predominantly had relative risk below the median. The climate-driven model (Figure 4b) showed a patchy pattern, with high risk occurring mostly in the northern parts of the James River Valley and the plateaus to the east and west of the valley, as well as a few isolated hot spots in western South Dakota. The model based on spectral data (Figure 4c) mostly highlighted the northern James River Valley and the parts of the Missouri Coteau region.

Figure 4. West Nile virus risk patterns based on three different environmental data sets: (a) land cover and physiography data, (b) climate data, and (c) spectral data.
located in the west. Also, as a result of the different resolutions of the environmental data sets, risk patterns are resolved at different scales. The models based on land cover and physiography data and the spectral data show finer-grained spatial patterns, whereas the climate data-driven model presents a comparatively smooth pattern.

The spatial pattern of relative WNV risk in South Dakota predicted by the model including all available environmental variables is presented in Figure 5. Based on this model, the relative risk of WNV was highest in the northern part of the James River Valley, with more areas of high risk in eastern South Dakota and one additional hot spot at the edge of the Prairie Coteau.

The combined model performed best, with the highest mean AUC value of 0.727. The combined model's AUC was followed by lower AUC values from the models driven by land cover and physiography data (0.711), climate data (0.719), and spectral data (0.679). Despite the relatively narrow range of AUC values for the best three models, the maps produced by the models showed considerable differences in spatial patterns of relative WNV risk as described previously.

### 3.2. Environmental Data Analysis

We interpreted the influences of environmental variables by quantifying the relative importance of each variable in the different BRT models (Figure 6) and by visualizing the partial dependence between relative risk and the most influential variables (Figure 7). Elevation was the most important variable in the model based on land cover and physiography data (Figure 6a), with low elevations more favorable for disease transmission. Of the climatic variables, the standard deviation of precipitation during early mosquito season (May/June) had the strongest impact on disease risk (Figure 6b). We observed the highest risk at very high and very low interannual variation of early-season precipitation. Interannual variations in moisture-related metrics, including the tasseled cap wetness index, normalized difference vegetation index, and modified normalized difference water index, were the most important predictors in the model based on spectral data (Figure 6c). Higher interannual variation of these moisture-sensitive indices was associated with higher disease risk.

The combined model was mainly driven by elevation, followed by metrics of relative humidity and precipitation in the late season (RH_LateSD, RH_LateMean, and Prec_LateMean), interannual variation in precipitation and surface moisture availability in early season (TcWet_EarlySD, Prec_EarlySD), percent pasture, and soil conditions associated with surface water ponding (PondFr; Figure 6d). In general, climatic variables proved to be
more influential than spectral variables. Of all spectral variables, interannual variability in surface moisture proved to be of highest importance.

4. Discussion

This study applied classified land cover and physiography data, climate data, and spectral data to map the relative risk of WNV transmission in South Dakota. All models highlighted particular aspects of environmentally driven spatial heterogeneity in WNV risk, and they generated predicted risk maps with more

Figure 6. Relative importance of environmental variables in predicting the relative risk of West Nile virus. We used four sets of environmental variables to fit a BRT model: (a) land cover and physiography data, (b) climate data, (c) spectral data, and (d) all data sets combined. Explanations of the variable codes are provided in Table 2.

Figure 7. Boosted regression tree partial dependence of risk on the top ranked environmental variables from the combined model containing all data sources. Black lines represent the fitted function for environmental variables and relative West Nile virus risk. Blue lines represent a smoothed approximation. Variables are explained in Table 2.
Spatial detail than the smoothed incidence rates. The models also showed some broad-scale similarities at the state level. These similarities reflect broad-scale associations between the land cover, physiography, climate, vegetation, and surface moisture variables that were measured in the various data sets. In particular, there is a strong east-west variation across the state reflected in an elevation gradient, a climatic gradient, and a transition from predominantly cropland and pasture in the east to predominantly grasslands in the west. Despite these large-scale collinearities, each of the data sets provided unique information and captured a different aspect of the environmental influences on WNV risk distribution.

Increasing the complexity of the model by combining all three data sets produced the best fitting model and helped the quality of the spatial prediction. The combined model identified spatial differences in transmission risk and highlighted a major high-risk area in the northern James River Valley with one distinct hot spot on the Prairie Coteau. We found that these high-risk areas were characterized by distinctive hydrological conditions during the early WNV season and climatological conditions during late WNV season. The combined model also helped to lessen the influence of spatial artifacts visible in the other models, such as the rings in the climate model that are a by-product of processing the climate data and the very strong signature of elevation in the land cover and physiography model. It is notable that the four most important variables in the combined model came from all three of the data sources, emphasizing that these different types of data provide complementary information to support disease predictions.

The model based on land cover and physiography data, which was driven primarily by elevation, highlighted the whole James River Valley as a high-risk area, with lower risk on the plateaus to the east and west of the valley. The James River Valley is a glacial valley that is dominated by relatively flat lowlands with poorly drained soils, shallow water tables, and extensive wetlands. In South Dakota, the main mosquito vector is *Culex tarsalis* Coquillett (Vincent, 2018). This mosquito is associated with small unconnected semipermanent ponds with standing water (Mercer et al., 2005; Skaff & Cheruvelil, 2016), which are abundant in the northern James River Valley. Our study also found that such catchment areas with temporary ponding are associated with high densities of WNV cases. In contrast, the highest elevation in the state is in the Black Hills in western South Dakota, which has a different ecosystem than in the rest of the state, with lower temperatures, higher forest cover, and few suitable breeding sites for *Culex tarsalis*. Therefore, we suggest that geospatial data on land cover, physiography, and soils can identify locations that are likely to contain the temporary water bodies that serve as breeding sites for the main vector species, *Culex tarsalis*.

The model based on climatic variables highlighted the northeastern part of the state, emphasizing the distinct climatology in the northern part of the James River Valley and adjacent uplands. Both late-season interannual variability and the long-term mean of relative humidity were important, as well as the late-season long-term mean of precipitation. Mosquitoes are small, ectothermic animals, and are sensitive to ambient climatic conditions, especially humidity. Mosquitoes desiccate at low humidity and avoid high humidity (Thomson, 1938), and it is not surprising that humidity is associated with disease transmission. In addition, drought conditions negatively affect the nesting success of bird communities, including passerine birds, in the Northern Great Plains (George et al., 1992). Variability in precipitation and relative humidity can therefore also affect the abundance of avian hosts. Climatic conditions were important mostly during late season, when the vector mosquito *Culex tarsalis* is feeding primarily on humans (Kent et al., 2009), and we observe the majority of the human WNV cases. We assume that climatic conditions are related to increased WNV cases in late season because during this time of increased feeding on humans, a longer life span of mosquitoes and their increased host-seeking activity will increase the chances of transmitting the disease. However, late-season transmission to humans is also contingent upon high levels of mosquito-bird transmission earlier in the season that amplify the virus and increase the mosquito infection rate.

The spectrally driven model highlights the same northeastern cluster as the climate-driven model. However, the distinctive land surface dynamics of this area were characterized by interannual variability in early-season surface moisture availability. In contrast to the importance of climate during the late season, during the early-season the availability of surface moisture was most important. As mentioned earlier, *Culex tarsalis* mosquitoes depend on small, temporary water bodies. Mosquito-bird interaction and the potential for virus transmission are influenced by the availability of species-rich environments, such as wetlands (Reisen et al., 2013). We suggest that surface moisture conditions may be most important for early-season mosquito population growth and disease amplification in the mosquito-bird system, setting the stage for large
outbreaks later in the season if conditions are favorable for mosquito survival and increased feeding on humans. We found that areas with high interannual variability of surface moisture were at highest risk. This result does not mean that the risk in those areas is consistently high but rather that those areas have the potential to produce abundant breeding sites in some years, which can lead to high mosquito numbers and high rates of disease amplification.

Contrary to our expectation, temperature did not have a strong influence in our models. Temperature does have a major effect on year-to-year variation in case occurrence (Davis et al., 2017; Wimberly et al., 2014), with WNV outbreaks tending to occur in warmer years, but it does not explain the spatial patterns of disease risk. Within South Dakota, the areas with the highest incidence rates are located in the north, where the mean annual temperature is lowest. Even on a larger scale, areas of the United States with warmer temperature are not the ones most prone to WNV. South Dakota is one of the coldest states but has the highest incidence rate in the country (Burakoff et al., 2018). We found that hydrology and humidity-related variables were more strongly associated with spatial variation in WNV risk. This finding is relevant to projections of disease distribution under different climate change scenarios. Increasing temperatures might lead to increasing frequency and severity of WNV transmission in areas already prone to outbreaks. However, other climatic variables such as precipitation and relative humidity, as well as changing land cover and hydrological conditions, may be more important predictors of future shifts in the geographic pattern of WNV transmission.

This study had several important limitations. First, the locations where the WNV transmission took place are unknown, and our data only include the home residence of each case. Therefore, our map is based on the assumption that most people become infected in close proximity to their homes. Additionally, we did not include other nonenvironmental factors such as access to health care, underreporting of cases with weak symptoms, and underlying variability in the blood donation rate. These variables most likely explain some of the spatial distribution of WNV in South Dakota, but they were not available as geospatial data sets. Instead, this study focused on predicting WNV using environmental data sets, and the results therefore highlight the specific and environmental conditions for which local residents are at highest risk of WNV. These various sources of uncertainty are one reason why the AUC values we reported did not exceed 0.73. However, our aim was not to deliver an accurate classification of where every WNV case occurred within the state but rather to present the distribution of relative risk across the state given environmental differences. Despite these limitations, our study was able to identify the underlying environmental factors that influenced risk and to map the environmentally driven relative risk across the state.

Our research suggests that combining multiple sources of data that characterize the hydrological and climatological characteristics and their variability over long time frame can improve estimates of the spatial dimension of disease risk across larger scales. Although models with only spectral indices did not demonstrate the best model performance, adding spectral variables to the combined model improved the spatial predictions. Our recommendation for future research is to include spectral data in combination with additional environmental data to identify areas of high risk. Locating these areas will assist public health education efforts to reach the local audiences at highest risk and enable more targeted and efficient disease prevention activities.

Also, targeted vector control strategies can also help to reduce the usage of insecticides, which would have financial benefits, reduce public concerns about insecticide use, and help slow the development of insecticide resistance.

## References

Benali, A., Nunes, J. P., Freitas, F. B., Sousa, C. A., Novo, M. T., Lourenço, P. M., et al. (2014). Satellite-derived estimation of environmental suitability for malaria vector development in Portugal. Remote Sensing of Environment, 145, 116–130. https://doi.org/10.1016/j.rse.2014.01.014

Boussama, T., Griffin, J. T., Sauverein, R. W., Smith, D. L., Churcher, T. S., Takken, W., et al. (2012). Hitting hotspots: Spatial targeting of malaria for control and elimination. PLoS Medicine, 9(1), e1001165. https://doi.org/10.1371/journal.pmed.1001165

Burakoff, A., Lehman, J., Fischer, M., Staples, J. E., & Lindsey, N. P. (2018). West Nile virus and other nationally notifiable arboviral diseases—United States, 2016. Morbidity and Mortality Weekly Report, 67(1), 13–17. https://doi.org/10.15585/mmwr.mm6701a3

CDC (2015). Arboviral diseases, neuroinvasive and non-neuroinvasive—2015 case definition. Retrieved May 24, 2018, from https://www.cdc.gov/mmwr/conditio.../arboviral-diseases-neuroinvasive-and-non-neuroinvasive-case-definition/2015/

CDC (2016). Final cumulative maps & data for 1999–2016. Retrieved May 24, 2018, from https://www.cdc.gov/westnile/statsmaps/cumMapsData.html

## Acknowledgments

This work was supported by the grant NNX15AF74G from the NASA Applied Sciences Health and Air Quality Program. Human case data were obtained through a data-sharing agreement with the South Dakota Department of Health (SDDOH). We thank Lon Kightlinger, Nick Hill, and Joshua Clayton of the SDDOH for facilitating access to these data. The other data used in the article are accessible through the University of Oklahoma’s public access institutional repository and information exchange https://shareok.org/.

---

HESS ET AL. 407
Reisen, W. K., Fang, Y., & Martinez, V. M. (2006). Effects of temperature on the transmission of West Nile virus by Culex tarsalis (Diptera: Culicidae). *Journal of Medical Entomology*, 43(2), 309–317. https://doi.org/10.1093/jmedent/43.2.309

Reisen, W. K., Lothrop, H. D., & Thiemann, T. (2013). Host selection patterns of Culex tarsalis (Diptera: Culicidae) at wetlands near the Salton Sea, Coachella Valley, California, 1998–2002. *Journal of Medical Entomology*, 50(5), 1071–1076. https://doi.org/10.1603/ME13078

Sánchez-Gómez, A., Amela, C., Fernández-Carrión, E., Martínez-Avilés, M., Sánchez-Vizcaíno, J. M., & Sierra-Moros, M. J. (2017). Risk mapping of West Nile virus circulation in Spain, 2013. *Acta Tropica*, 169, 163–169. https://doi.org/10.1016/j.actatropica.2017.02.022

Skaff, N. K., & Cheruvelli, K. S. (2016). Fine-scale wetland features mediate vector and climate-dependent macroscale patterns in human West Nile virus incidence. *Landscape Ecology*, 31(7), 1615–1628. https://doi.org/10.1007/s10980-016-0346-1

Thomson, R. C. M. (1938). The reactions of mosquitoes to temperature and humidity. *Bulletin of Entomological Research*, 29(02), 125–140. https://doi.org/10.1017/S0007485300002618

Tran, A., Sudre, B., Paz, S., Rossi, M., Desbrosse, A., Chevalier, V., & Semenza, J. C. (2014). Environmental predictors of West Nile fever risk in Europe. *International Journal of Health Geographics*, 13(1), 26. https://doi.org/10.1186/1476-072X-13-26

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2), 127–150. https://doi.org/10.1016/0034-4257(79)90013-0

Vincent, G. P. (2018). Surveillance of South Dakota mosquito abundance, infection rate, and insecticide susceptibility. (Doctoral dissertation). Retrieved from https://openprairie.sdstate.edu/etd/2461. Brookings, SD: South Dakota State University

Wand, M. P., & Jones, M. C. (1995). Kernel Smoothing. London: Chapman & Hall.

Wimberly, M. C., Baer, A. D., & Yabsley, M. J. (2008). Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *International Journal of Health Geographics*, 7(5), 573. https://doi.org/10.1186/1476-072X-7-573

Wimberly, M. C., Giacomo, P., Kightlinger, L., & Hildreth, M. (2013). Spatio-temporal epidemiology of human West Nile virus disease in South Dakota. *International Journal of Environmental Research and Public Health*, 10(11), 5584–5602. https://doi.org/10.3390/ijerph10115584

Wimberly, M. C., Hildreth, M. B., Boyte, S. P., Lindquist, E., & Kightlinger, L. (2008). Ecological Niche of the 2003 West Nile virus epidemic in the Northern great Plains of the United States. *PLoS One*, 3(12), e3744–e3747. https://doi.org/10.1371/journal.pone.0003744

Wimberly, M. C., Lamsal, A., Giacomo, P., & Chuang, T. W. (2014). Regional variation of climatic influences on West Nile virus outbreaks in the United States. *American Journal of Tropical Medicine and Hygiene*, 91(4), 677–684. https://doi.org/10.4269/ajtmh.14-0239

Wulder, M. A., White, J. C., Loveland, T. R., Woodcock, C. E., Belward, A. S., Cohen, W. B., Fosnight, E. A., et al. (2016). The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment*, 185, 271–283. https://doi.org/10.1016/j.rse.2015.11.032

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033. https://doi.org/10.1080/01431160600589179

Young, S. G., Tullis, J. A., & Cothren, J. (2013). A remote sensing and GIS-assisted landscape epidemiology approach to West Nile virus. *Applied Geography*, 45, 241–249. https://doi.org/10.1016/j.apgeog.2013.09.022