Distribution Agreement

Signature:

Jessica Belle

Date

Model approaches to evaluating potential mechanisms of parasite diffusion

By

Jessica H. Belle
Master of Science in Public Health

Environmental Health - Epidemiology

____

Justin V. Remais, PhD
Committee Chair

____

Paige Tolbert, PhD
Committee Member

Model approaches to evaluating potential mechanisms of parasite diffusion

By

Jessica H. Belle

B.A.
Augustana College
2009

Thesis Committee Chair: Justin V. Remais, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Environmental Health - Epidemiology
2013

Abstract

Model approaches to evaluating potential mechanisms of parasite diffusion
By Jessica H. Belle

Environmental parasites, including the trematode that causes schistosomiasis, are capable of dispersing across a landscape through environmental pathways, in particular hydrological networks. Yet our understanding of how environmental dispersion of parasites influences patterns of human disease is limited. In this thesis, several conceptualizations of environmental dispersion from putative sources of transmission were formulated and tested as to their ability to explain patterns of schistosomiasis incidence in Sichuan province, China. The dispersion models explored included: Euclidean dispersion; dispersion limited to downstream movement, where risk is determined by the nearest source; and dispersion occurring downstream of each source with exponentially decreasing likelihood, where the contributions of all upstream sources are summed. Each conceptualization of dispersion was used to generate exposure estimates for each location across a grid covering Sichuan Province. Statistical models were constructed to examine associations between each exposure estimate and the spatial distribution of cases reported to China's National Infectious Disease Reporting system in Sichuan province for the period of January 1, 2005 through December 31, 2011. A zero-inflated negative binomial modeling framework was used, and model fit was evaluated based on the Akaike information criterion (AIC). The models including dispersion of any kind performed better than the model with no dispersion. The model including dispersion occurring with exponentially decreasing likelihood downstream of each source, with a median dispersal distance of 1,200 m, performed slightly better than the other dispersion models based on AIC. However, the differences in the AIC values between the different dispersion models were small. The residuals from each of the models were also examined for evidence of spatial auto-correlation, however the distributions of the residual values were highly skewed, and calculation of a global Moran's I index was not possible. This paper makes methodological contributions to the literature despite the modest conclusions drawn, namely through the representation of anisotropic dispersion from a potential source in a regression framework where it could be directly related to spatial patterns of disease.

Model approaches to evaluating potential mechanisms of parasite diffusion

By

Jessica H. Belle

B.A.
Augustana College
2009

Thesis Committee Chair: Justin V. Remais, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Environmental Health - Epidemiology
2013

Acknowledgements

This work was supported in part by the NSF/NIH Ecology of Infectious Disease Program (award 0622743), by the NSF Division of Chemical, Bioengineering, Environmental, and Transport Systems (award 1249250), by the National Institute for Allergy and Infectious Disease (grant K01AI091864), and by the Emory University Global Health Institute Faculty Distinction Fund.

1. Introduction

China's national schistosomiasis control program, established in the 1950s, in conjunction with a 1992-2001 World Bank initiative, achieved substantial reductions in the national burden of morbidity and mortality attributable to schistosomiasis (Spear, Seto et al. 2011, Zhang, Zhu et al. 2012). However, concerns have been raised regarding the ability of current strategies to effectively move the country from a low transmission scenario to disease elimination. This is due to potential re-emergent disease in areas where transmission had previously been considered interrupted (Spear, Seto et al. 2011), as well as the finding that the World Bank program failed to reduce the spatial distribution of disease using a combination of snail control and mass administration of praziquantel (Zhang, Zhu et al. 2012, Zhang, Zhu et al. 2012). The re-emergence of schistosomiasis, in particular, has raised questions concerning the ability of *Schistosoma japonicum* to diffuse across a landscape into formerly controlled areas (Gurarie and Seto 2009) (Remais, Akullian et al. 2010).

The *S. japonicum* life cycle involves transmission from a mammalian host to an intermediate snail host and back to a mammalian host. The stages infectious to humans and snails are aquatic, and the *Oncomelania* snail is amphibious, spending a large portion of its lifespan underwater. Thus, river and irrigation networks determine, in part, the spatial distribution of the parasite and its intermediate host, which strongly mediates the epidemiology of human disease (Spear, Seto et al. 2004). However, accurately distinguishing between areas where risk of infection is present and areas where it is not is challenging. This generally requires large-scale surveys, and becomes increasingly costly as the prevalence and infection intensity of human disease falls. Recently, there has been interest in identifying environmental factors determining infection risk. These could be monitored using satellite imagery, and would allow the use of targeted mapping to identify areas where sustained, endemic transmission is likely to occur (Lustigman, Prichard et al. 2012) (McManus, Li et al. 2009).

In Sichuan, individual determinants of disease have been identified, including occupational status as a fisherman or farmer (Seto, Lee et al. 2007), age (Seto, Lee et al. 2007), and education (Spear, Seto et al. 2004). At the village level, other important determinants have been identified, including presence of infected bovines and water buffalo (Raso, Li et al. 2009), density of snails (Spear, Seto et al. 2004, Seto, Lee et al. 2007), number of infected snails (Spear, Seto et al. 2004), presence of irrigation channels (Maszle, Whitehead et al. 1998, Spear, Seto et al. 2004, Lowe, Xi et al. 2005), application of manure to crops, and crop type (Spear, Seto et al. 2004). Many such factors relate to water resources and to irrigated agriculture in particular.

Yet given a location with the potential to support sustained local transmission, additional questions arise as to the possibility of transmission risk radiating away from the site, through various dispersion mechanisms. Experimental studies have investigated the capability of cercariae and snails to move downstream, finding for instance that a small proportion of cercariae can be carried as far as a few kilometers downstream and remain viable during transit (Radke, Ritchie et al. 1961). Other work (Maszle, Whitehead et al. 1998, Lowe, Xi et al. 2005, Akullian, Lu et al. 2012) has corroborated these findings. For instance, Akullian et al. (Akullian, Lu et al. 2012) investigated snail dispersion and found that the snails, on average, move downstream at a rate of a few meters a day. Similarly, although they do not explicitly consider direction, studies of parasite mitochondrial DNA sequences have shown that distance influences the likelihood that parasite populations will interbreed, but that interbreeding between spatially distributed populations of parasites is more common in the middle and lower reaches of the Yangtze river (Zhao, Jiang et al. 2012). It was speculated that this difference may be the result of flood events, and observational studies in the middle and lower reaches of the Yangtze found that flooding was associated with expansions in the range of the *Oncomelania* snail as well as human cases of schistosomiasis in either previously controlled areas or new habitat (Zhou, Dandan et al. 2002, Wu, Zhang et al. 2008). This range expansion has been estimated to occur on a spatial scale of a few kilometers; however, the experimental studies seem to indicate that, in the absence of flooding, dispersion is likely to occur at smaller spatial scales. Small-scale studies in Sichuan using modeling have demonstrated that inter-village connectivity, and importation of parasites from upstream sources, may play a role in maintaining endemic schistosomiasis in individual villages (Xu, Gong et al. 2006, Spear 2012). However, the extent to which these forces may shape patterns of human disease at the province level has not been as thoroughly investigated.

The downstream dispersion of parasites and snails represents potential anisotropic determinants of the spatial distribution of schistosomiasis. However, the role of upstream/downstream relationships in contributing to spatial patterns of human disease is poorly understood. In order to investigate this process further, an analysis was conducted examining the role of anisotropic factors in determining patterns of reported schistosomiasis cases. A series of alternative models were explored that considered various ways in which dispersion may occur: Euclidean-based dispersion, where direction and overland distance is ignored; hydrological dispersion assuming downstream dispersion only; and cumulative hydrological dispersion where disease risk from multiple upstream sources accumulates downstream, but the effect of each source diminishes exponentially with increasing distance. The relative capabilities of these models to predict the existing pattern of incident human disease were examined with respect to the role of upstream locations in supporting downstream transmission.

2. Methods

2.0 Study site

Sichuan province is located in the western part of China, and includes only the upper reaches of the Yangtze River. Flooding occurs in this area, but the overall volume of precipitation associated with extreme rainfall events is lower than in the middle and lower reaches of the Yangtze (Su, Gemmer et al. 2008); thus, floods in Sichuan tend not to be as large or frequent as they are downstream. The diffusion of parasites and snails in the upper reaches of the Yangtze may occur somewhat differently than in the middle and lower reaches. Diffusion may not be as strongly associated with flood events, and may occur at smaller spatial scales. However, as evidenced by laboratory, modeling, and genetic studies, this does not preclude the possibility of diffusion of parasites, snails, and human disease over measurable spatial scales in Sichuan.

2.1 Disease data

Schistosomiasis incidence data were obtained from China's National Infectious Disease Reporting (NIDR) system via a query of the NIDR database at the China Centers for Disease Control and Prevention in Chengdu, Sichuan. This query specified all cases of schistosomiasis involving patients residing in Sichuan province between January 1st, 2005 and Dec. 31st 2011. Cases were initially classified as acute, chronic, or unknown, and were diagnosed using either purely clinical means or via laboratory confirmation, the later generally indicating Kato-Katz (Katz, Chaves et al. 1972). All cases were treated identically in the analysis regardless of classification.

Data were geocoded using NAC locator (NAC Geographic Products 2011) and residential addresses, and the number of cases were counted in each cell of a 30 arc-second (~1 km^2) square grid in ArcGIS (ESRI 2012). To obtain rates, population data obtained from the 2010 Landscan dataset (Bright, Coleman et al. 2011) were used as a denominator. Locations with a population count equal to zero were excluded from the analysis; 241,781 locations were excluded as a result.

2.2 Source identification and exposure estimation

Exposure in a given grid cell was defined by its distance from cells that could potentially serve as 'sources' of disease transmission. A grid cell was classified as a potential 'source' whenever three criteria were met: 1) the cell contained land used for irrigated agriculture; 2) the cell elevation was <2,150 m; and 3) the cell was >26,500 m from the border of Sichuan province (to reduce so-called edge-effects; figure 1). Land used for irrigated agriculture was identified from the Globcover 2005 10 arc-second (~300 m^2) regional land use dataset for Central Asia (Arino, Gross et al. 2007). Locations below 2,150 m were identified from the HydroSHEDS 3 arc-second (~90 m^2) void-filled elevation dataset (Lehner, Verdin et al. 2008). The elevation cut-off was based on previous work examining seroprevalence across elevation (Steinmann, Zhou et al. 2007, Steinmann, Zhou et al. 2007).

Given classified source grid cells, exposure at all grid cells was calculated considering the proximity of each cell to sources and assuming five different dispersion scenarios. Scenarios are roughly based on the biology of the parasite and intermediate host (table 1). The first dispersion scheme considered simple *Euclidean distance* in meters, calculated using ArcGIS (ESRI 2012), between each cell and the nearest cell classified as a source.

The other four exposure measures were based on dispersion schemes that accounted for hydrological distance. The *Hydrological distance* scheme used a path-distance function (ESRI 2008) that applied an infinite cost to movement against the direction of water flow, as determined by the drainage directions available from HydroSHEDS (Lehner, Verdin et al. 2008). Overland distance accounting for changes in elevation information was included. Both the *Euclidean* and *Hydrological distance* models assume that disease risk associated with dispersion is only determined by the distance to the *nearest* source. In contrast, the dispersion models below consider *all* potential nearby sources.

Three additional *Cumulative hydrological distance* dispersion functions were explored (table 1), each involving a different decay coefficient. A first-order decay function: e^{-kd} was used to assess the influence of multiple upstream ‘sources’ on downstream risk, where increasing distance, *d*, yielded diminishing influence as determined by the decay constant, *k*. Multiple values of *k* represented a range of potential scales of downstream sharing of risk, with each value being informed by experimental studies investigating cercarial and snail dispersion. The first *k* corresponds to a spatial scale assuming that the average lifetime dispersion distance of either snails or cercariae is equal to 300 m. This value corresponds to the median dispersion distance of cercariae over the course of their active lifetime (Lowe, Xi et al. 2005), as well as to rough estimates of the average annual dispersion of the snail hosts (Wu, Zhang et al. 2008, Akullian, Lu et al. 2012). The second *k* corresponds to a spatial scale where the median dispersion distance is equal to 1,200 m, a value based on a maximum cercarial lifetime dispersion distance (equal to 0.05 in kernel function) of ~3,000 m (Maszle, Whitehead et al. 1998, Akullian, Lu et al. 2012), as well as rough estimates of annual snail dispersion during flood years (Wu, Zhang et al. 2008). The third *k*, with an average dispersion distance equal to 2,600 m, corresponds to extrapolated snail lifetime dispersion estimates (Wu, Zhang et al. 2008), or combined snail and cercarial dispersion (Akullian, Lu et al. 2012).

For the *Cumulative distance models*, the decayed hydrological distances from all sources contributing to a grid cell were summed to produce an estimate of the risk from all nearby sources, weighted by their proximity to the cell. Following calculation, the five different exposure measures were entered into statistical models of counts of cases in grid cells across the study region.

2.3 Statistical analysis

Logistic, Poisson, negative binomial, zero-inflated poisson, and zero-inflated negative binomial models were fit to incidence rate data for each of the exposure definitions outlined in table 1 using R (Zeileis, Kleiber et al. 2007, Team 2008). Additionally, these models were fit for a *Source* exposure definition, which used the binary classification of a location as a source or non-source, with no dispersion. A number of different categorizations, schemes, and transformations of the five different continuous exposure measures were explored under the different modeling frameworks, including deciles, quartiles, and log transforms of the continuous data. The fit and ability of each of the six models to explain the observed distribution of disease was assessed and compared using the Akaike Information Criterion (AIC) (Akaike 1974), as well as through examination of the residual errors (Arcaya, Brewster et al. 2012). These were tested for evidence of spatial autocorrelation using Moran's I (Tiefelsdorf 1998, Leung, Mei et al. 2003), and mapped for visual comparisons in ArcGIS. (ESRI 2012).

3. Results

3.1 Disease data

Out of the 1,750,384 possible ~1 km grid locations containing at least one person, 261 had at least one case over the 6-year study period.. Within locations with at least one case, the median number of cases per person was 0.0027, and ranged from 0 to 2.33. The vast majority of the locations with reported cases had low case densities, with only a few grid cells having more than 0.02 cases per person. Cases were highly clustered in space, and were primarily reported in the plain around Chengdu and in the Yangtze headwaters in the Yi Autonomous prefecture (figure 2).

3.2 Source identification and exposure estimation

A total of 50,267 locations within Sichuan province were identified as sources based on the definition described above (containing land used for irrigated agriculture at ≤2,150 m). Euclidean distances from all grid cells in the province to the nearest source grid cell ranged from 0 to 523.1 km, with the median at 70.4 km. The spatial distribution of the Euclidean distances, as categorized into quartiles, is shown in figure 3. As expected, lower values are found surrounding areas with sources present. The distribution of distances is right-skewed, and the values are concentrated at the lower end of the spectrum. The other four exposure variables are similarly right-skewed.

There were no locations containing cases that were found more than 134,800 m from a source using *Euclidean distance.* In order to ensure that all models would be able to converge while still being comparable to one another, these locations were removed prior to the analysis. This resulted in the exclusion of 370,528 locations, and reduced the total number of locations used in the analysis to 1,379,856.

The *Hydrological distance* was calculated using a moving window around each source location, effectively considering only sources ≤ 26,000 km from the grid cell of interest. Approximately 70% of grid cells in the total province were > 26,000 km from a source, and were considered 'not calculated' for the variable hydrological distance. This includes the 19.9% of all locations with at least one case for which *Hydrological distance* was not calculated*.* For the analysis, locations with not calculated data for hydrological distance were assigned values above the maximum recorded value, and were placed in a separate category, where the OR for 'not calculated' values was allowed to vary independently of the recorded values. The spatial distribution of this variable and of areas that were not calculated is presented in figure 4. This distribution is somewhat similar to the distribution observed for the *Euclidean distance* exposure variable. However, asymmetry is evident in this figure, a product of the cost-distance function that limits dispersion to the direction of water flow.

The *Cumulative hydrological distance* models all exhibited a similar right-skewed distribution of values, clustered around sources, as expected (figure 5). The first model, which incorporates a median dispersion distance from the source of only 300 m, generates a spatial distribution very similar to the original distribution of sources shown in figure 1. However, the second *Cumulative hydrological distance* model, incorporating a median dispersion distance of 1,200 m, shows a marked directional spreading of higher values out from the original source definitions, in the same direction as in the hydrological distance measure. The last *Cumulative hydrological distance* model, with a median dispersion distance of 2,600 m, shows this pattern yet more clearly, with higher values extending even further in the direction of water flow.

3.3 Statistical Analysis

In a simple univariate analysis, locations defined as sources were 29.1 [24.7, 34.2] times more likely to contain cases than other locations in the province, which supports the definition of sources used in the dispersion analysis. Of the model fit options considered, the zero-inflated negative binomial model with continuous variables categorized into quartiles produced the best overall fit to the data for all six exposure models. Zero-inflated negative binomial regression detected significant associations between the presence of a case in a grid cell and each of the six exposure models. Locations at increasing distances from the nearest source were less likely to harbor cases, while locations with larger values for the *Cumulative hydrological distance* measures were increasingly likely to harbor cases. All models that defined exposure based on an explicit dispersion mechanism offered an improved fit, based on a comparison of AIC, over the *Source* model. Based on AIC values, the best model was *Cumulative hydrological distance* at a spatial scale of 1,200 m, but *Cumulative hydrological distance* at the 2,600 m scale came in a close second. Apart from the *Source* model, the *Cumulative hydrological distance* model at the 300 m scale also performed poorly according to this criterion (table 2).

The odds ratios output for different quartiles, as compared to the 0 values in each of the models, are highly significant, with the risk associated with residence in a location dropping sharply with increasing *Euclidean and Hydrological distance*, and increasing sharply with increasing *Cumulative hydrological distance*. However, the goal of this analysis was to evaluate a set of models for their fit to observed data, rather than to produce unbiased effect and error estimates for the exposure measures in each model. Potential confounders were not included, and therefore these values may be biased.

An examination of the residuals for evidence of spatial autocorrelation via Moran's I indicated that residuals of all models exhibited significant spatial autocorrelation (table 2). The values from the global Moran's I tests ranged from 0.0002 to 0.0028 (z-scores between 1.98 and 8.66), with the *Cumulative hydrological distance* model at the 300 m scale producing the lowest (least clustering) value, and the model relating the *Cumulative hydrological distance* measure at the 1,200 m scale to the case counts in each location producing the highest. However, visual examination of the raw residual values revealed that their statistical distribution was highly right-skewed and non-normal, with the majority of the values near zero and slightly negative (figure 6). This implies that the Moran's I test, which assumes that the data is normally distributed, is not appropriate for use with this data and may produce unreliable results (Waller and Gotway 2004). For this reason, no substantive conclusions, other than that the inclusion of dispersion from a source, in this instance, produces a better-fitting model than categorization of locations as containing sources or not, can be drawn at this time.

4. Discussion

This study focused on the extent to which the spatial process of downstream diffusion, when incorporated into statistical models, influenced the association between at-risk locations and observed patterns of schistosomiasis incidence within Sichuan province. Using AIC as a metric, we showed that incorporating the effects of diffusion processes into the modeling framework improves the relative capability of a model to accurately represent the data. However, it is difficult to draw any substantive conclusions beyond that, since the relative differences in AIC values between the set of dispersion models were relatively small, and examination of the residuals produced inconclusive results.

The analysis also did not account for other spatial or membership processes that may have influenced disease patterns (Arcaya, Brewster et al. 2012). This becomes particularly important in the context of the data used here. China's NIDR system is a passive, hospital-based reporting system, and there may have been reporting bias resulting in higher or lower numbers of reported cases in some areas. It is likely that underreporting not only exists, but exists differentially in space, and could thus have led to case misclassification. Case reporting may have been influenced by membership processes such as the hospital the patient visited, access to care, the diagnostic capabilities of the hospital, or township of residence. While it is difficult to conceptualize how these might vary substantially with upstream/downstream relationships, they may vary with the classification of land as being used for irrigated agriculture or not, particularly since hospitals are generally less available and well-equipped in rural areas. Without sensitivity analysis, it is difficult to predict the extent to which this would have influenced the results. Unfortunately, detailed information on the extent to which underreporting occurs for this disease in Sichuan, and what influences it, is not available at this time.

A further source of uncertainty in this study is the quality of the geospatial data. The land use, elevation, and flow direction rasters used to identify sources and calculate the exposures, as well as the population data, were originally drawn from satellite imagery. While every effort was made to ensure that the satellite images used to create these rasters came from years relevant to the study period, had good coverage over the study area, and had been vetted by other researchers, there is a certain amount of misclassification error inherent in the process of creating these datasets (Atkinson and Graham 2006, Arino, Gross et al. 2007). This is most likely to have been an issue with the identification of source locations. However, this type of uncertainty would most likely have influenced the significance of the relationship between the locations of cases sand sources, rather than the relative differences between models that incorporated dispersion from the source and the *Source* model.

Additionally, there is uncertainty associated with the location of cases. Had the analysis been conducted at the township level, as opposed to using a grid, this uncertainty would have been greatly mitigated, as the codes referencing the township of residence included in the database are fairly accurate and standardized. However, this would have come at the expense of using larger, less regular cells in which to count the cases, and may have required resampling the exposure and population data. The township boundaries are also based on political considerations that may have little to do with the hydrology of the landscape and so may have introduced bias in the form of the modifiable areal unit problem (Waller and Gotway 2004). In the face of this issue, a smaller unit of aggregation was chosen in order to limit the extent to which changes in data support and aggregation would influence the results.

Finally, the use of the global Moran's I statistic as a way of measuring the ability of each of the models to explain the existing patterns of human disease proved to be unreliable, given the distribution of the residuals in this dataset. A better option may have been to use a Monte Carlo test, in whcih the zero-inflated negative binomial distribution was simulated as the null model rather than the Gaussian distribution that the Moran's I statistic is built on (Waller and Gotway 2004). Alternatively, an autocorrelation term could have been built into the model instead of being assessed as an outcome. However, this term would not have been directly comparable between models, leading to difficulties in its interpretation between models.

5. Conclusion

While the results of this study were somewhat ambiguous, it was possible to establish that including parasite dispersion in a model of the spatial distribution of schistosomiasis cases improved model fit to the observed distribution of disease reporting patterns across an area, as compared to ignoring dispersion entirely. Furthermore, we show that our definition of a location as a source was a highly significant predictor of schistosomiasis risk in Sichuan Province in a simple univariate analysis.

The methods used here also represent, at the very least, a new epidemiological approach to the problem of characterizing dispersion across a landscape. We were able to represent an anisotropic definition of of exposure based on environmental dispersion. Our exposure definitions accounted for the potential movement of schistosome parasites or their intermediate hosts across a landscape using distance or cost-distance based measures, and allowed for the evaluation of a variety of different conceptualizations of dispersion. The use of these exposure metrics in a regression framework where they were related to the spatial distribution of disease was demonstrated, and yet the analysis had a number of limitations. Future directions for this work would include thorough sensitivity analyses and adoption of a Monte Carlo test for spatial autocorrelation that would be more appropriate for the distribution of these data (Waller and Gotway 2004). A modeling approach could also be taken to validate the exposure metrics in a wider range of situations.

Works Cited

Akaike, H. (1974). "A new look at the statistical model identification." Automatic Control, IEEE Transactions on **19**(6): 716-723.

Akullian, A. N., D. Lu, J. Z. McDowell, G. M. Davis, R. C. Spear and J. V. Remais (2012). "Modeling the Combined Influence of Host Dispersal and Waterborne Fate and Transport on Pathogen Spread in Complex Landscapes." Water Quality, Exposure and Health: 1-10.

Arcaya, M., M. Brewster, C. M. Zigler and S. Subramanian (2012). "Area variations in health: A spatial multilevel modeling approach." Health & Place.

Arino, O., D. Gross, F. Ranera, L. Bourg, M. Leroy, P. Bicheron, J. Latham, A. Di Gregorio, C. Brockman and R. Witt (2007). GlobCover: ESA service for global land cover from MERIS. Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International, IEEE.

Atkinson, P. M. and A. J. Graham (2006). Issues of Scale and Uncertainty in the Global Remote Sensing of Disease. Advances in Parasitology. A. G. Simon I. Hay and J. R. David, Academic Press. **Volume 62:** 79-118.

Bright, E. A., P. R. Coleman, A. N. Rose and M. L. Urban (2011). LandScan 2010. Oak Ridge, TN, Oak Ridge National Laboratory.

ESRI (2008). ArcGIS Desktop. Release 9.3. Redlands, CA, Environmental Systems Reseach Institute.

ESRI (2012). ArcGIS Desktop. Release 10.1. Redlands, CA, Environmental Systems Research Institute.

Gurarie, D. and E. Y. Seto (2009). "Connectivity sustains disease transmission in environments with low potential for endemicity: modelling schistosomiasis with hydrologic and social connectivities." Journal of the Royal Society Interface **6**(35): 495-508.

Katz, N., A. Chaves and J. Pellegrino (1972). "A simple device for quantitative stool thick-smear technique in Schistosomiasis mansoni." Revista do Instituto de Medicina Tropical de São Paulo **14**(6): 397.

Lehner, B., K. Verdin and A. Jarvis (2008). New global hydrography derived from spaceborne elevation data. AGU. Transactions Eos.

Leung, Y., C.-L. Mei and W.-X. Zhang (2003). "Statistical test for local patterns of spatial association." Environment and Planning A **35**(4): 725-744.

Lowe, D., J. Xi, X. Meng, Z. Wu, D. Qiu and R. Spear (2005). "Transport of Schistosoma japonicum cercariae and the feasibility of niclosamide for cercariae control." Parasitology International **54**(1): 83-89.

Lustigman, S., R. K. Prichard, A. Gazzinelli, W. N. Grant, B. A. Boatin, J. S. McCarthy and M.-G. Basáñez (2012). "A Research Agenda for Helminth Diseases of Humans: The Problem of Helminthiases." PLoS Negl Trop Dis **6**(4): e1582.

Maszle, D. R., P. G. Whitehead, R. C. Johnson and R. C. Spear (1998). "Hydrological studies of schistosomiasis transport in Sichuan Province, China." Sci Total Environ **216**(3): 193-203.

McManus, D. P., Y. Li, D. J. Gray and A. G. Ross (2009). "Conquering 'snail fever': schistosomiasis and its control in China." Expert Rev Anti Infect Ther **7**(4): 473-485.

NAC Geographic Products, I. (2011). "NAC Locator - A Universal Geocoding Solution for the Entire World." Retrieved 08/2012, 2012, from http://nactag.info/map.asp.

Radke, M. G., L. S. Ritchie and W. B. Rowan (1961). "Effects of water velocities on worm burdens of animals exposed to Schistosoma mansoni cercariae released under laboratory and field conditions." Experimental Parasitology **11**(4): 323-331.

Raso, G., Y. Li, Z. Zhao, J. Balen, G. M. Williams and D. P. McManus (2009). "Spatial distribution of human Schistosoma japonicum infections in the Dongting Lake Region, China." PLoS One **4**(9): e6947.

Remais, J., A. Akullian, L. Ding and E. Seto (2010). "Analytical methods for quantifying environmental connectivity for the control and surveillance of infectious disease spread." J R Soc Interface **7**(49): 1181-1193.

Seto, E. Y., Y. J. Lee, S. Liang and B. Zhong (2007). "Individual and village-level study of water contact patterns and Schistosoma japonicum infection in mountainous rural China." Trop Med Int Health **12**(10): 1199-1209.

Spear, R. C. (2012). "Internal versus external determinants of Schistosoma japonicum transmission in irrigated agricultural villages." J R Soc Interface **9**(67): 272-282.

Spear, R. C., E. Seto, S. Liang, M. Birkner, A. Hubbard, D. Qiu, C. Yang, B. Zhong, F. Xu, X. Gu and G. M. Davis (2004). "Factors influencing the transmission of Schistosoma japonicum in the mountains of Sichuan Province of China." Am J Trop Med Hyg **70**(1): 48-56.

Spear, R. C., E. Y. Seto, E. J. Carlton, S. Liang, J. V. Remais, B. Zhong and D. Qiu (2011). "The challenge of effective surveillance in moving from low transmission to elimination of schistosomiasis in China." Int J Parasitol **41**(12): 1243-1247.

Steinmann, P., X. N. Zhou, Y. L. Li, H. J. Li, S. R. Chen, Z. Yang, W. Fan, T. W. Jia, L. H. Li, P. Vounatsou and J. Utzinger (2007). "Helminth infections and risk factor analysis among residents in Eryuan county, Yunnan province, China." Acta Trop **104**(1): 38-51.

Steinmann, P., X. N. Zhou, B. Matthys, Y. L. Li, H. J. Li, S. R. Chen, Z. Yang, W. Fan, T. W. Jia, P. Vounatsou and J. Utzinger (2007). "Spatial risk profiling of Schistosoma japonicum in Eryuan county, Yunnan province, China." Geospat Health **2**(1): 59-73.

Su, B., M. Gemmer and T. Jiang (2008). "Spatial and temporal variation of extreme precipitation over the Yangtze River Basin." Quaternary International **186**(1): 22-31.

Team, R. C. (2008). "R: A language and environment for statistical computing." Vienna, Austria: R Foundation for Statistical Computing: 1-1731.

Tiefelsdorf, M. (1998). "SOME PRACTICAL APPLICATIONS OF MORAN'S I'S EXACT CONDITIONAL DISTRIBUTION." Papers in Regional Science **77**(2): 101-129.

Waller, L. A. and C. A. Gotway (2004). Applied spatial statistics for public health data, John Wiley & Sons.

Wu, X. H., S. Q. Zhang, X. J. Xu, Y. X. Huang, P. Steinmann, J. Utzinger, T. P. Wang, J. Xu, J. Zheng and X. N. Zhou (2008). "Effect of floods on the transmission of schistosomiasis in the Yangtze River valley, People's Republic of China." Parasitol Int **57**(3): 271-276.

Xu, B., P. Gong, E. Seto, S. Liang, C. Yang, S. Wen, D. Qiu, X. Gu and R. Spear (2006). "A spatial-temporal model for assessing the effects of intervillage connectivity in schistosomiasis transmission." Annals of the Association of American Geographers **96**(1): 31-46.

Zeileis, A., C. Kleiber and S. Jackman (2007). "Regression models for count data in R."

Zhang, Z., R. Zhu, M. P. Ward, W. Xu, L. Zhang, J. Guo, F. Zhao and Q. Jiang (2012). "Long-term impact of the World Bank Loan Project for schistosomiasis control: a comparison of the spatial distribution of schistosomiasis risk in China." PLoS Negl Trop Dis **6**(4): e1620.

Zhang, Z. J., R. Zhu, R. Bergquist, D. M. Chen, Y. Chen, L. J. Zhang, J. G. Guo, F. Zhao and Q. W. Jiang (2012). "Spatial comparison of areas at risk for schistosomiasis in the hilly and mountainous regions in the People's Republic of China: evaluation of the long-term effect of the 10-year World Bank Loan Project." Geospat Health **6**(2): 205-214.

Zhao, Q. P., M. S. Jiang, H. F. Dong and P. Nie (2012). "Diversification of Schistosoma japonicum in Mainland China revealed by mitochondrial DNA." PLoS Negl Trop Dis **6**(2): e1503.

Zhou, X., L. Dandan, Y. Huiming, C. Honggen, S. Leping, Y. Guojing, H. Qingbiao, L. Brown and J. B. Malone (2002). "Use of landsat TM satellite surveillance data to measure the impact of the 1998 flood on snail intermediate host dispersal in the lower Yangtze River Basin." Acta Tropica **82**(2): 199-205.

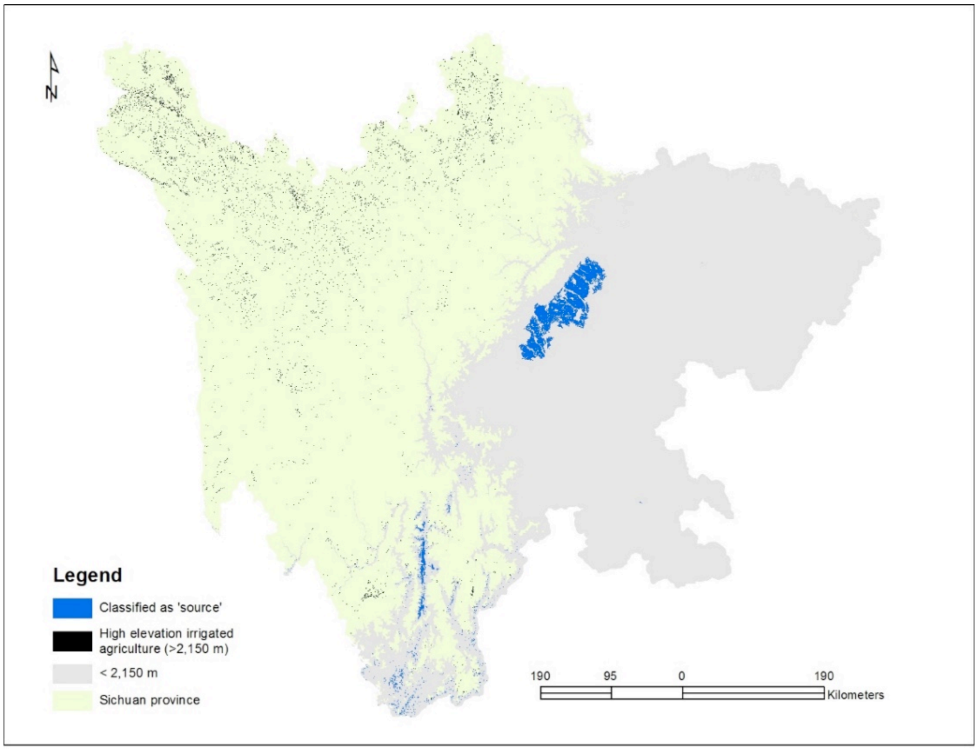


Figure 1: Spatial distribution of grid cells classified as 'sources', high elevation land used for irrigated agriculture (>2150 m, and thus not considered sources), and low elevation (<2,150 m) cells with land use other than irrigated agriculture.

Dispersion Scheme	Value used at cells	Spatial Scale (average dispersion distance)	Resolution of distance calculations (arc-seconds)	Threshold value** (m)
Euclidean distance	Minimum	-	Calculated using vector data	-
Hydrological distance	Minimum	-	15	26,500
Cumulative hydrological distance	Sum	300m	3	3,100
Cumulative hydrological distance	Sum	1,200m	15	26,500
Cumulative hydrological distance	Sum	2,600m	15	26,500
** Equal to the distance at which the exponential kernel decay function with the largest spatial scale calculated at that resolution decays to below 0.001 plus the value of one additional pixel, rounded up to the nearest 100m. Also equal to half of an edge in the moving window used to calculate these values.				

Table 1: Relevant spatial scales and dispersion schemes loosely identified using biological data on parasite and host dispersion.

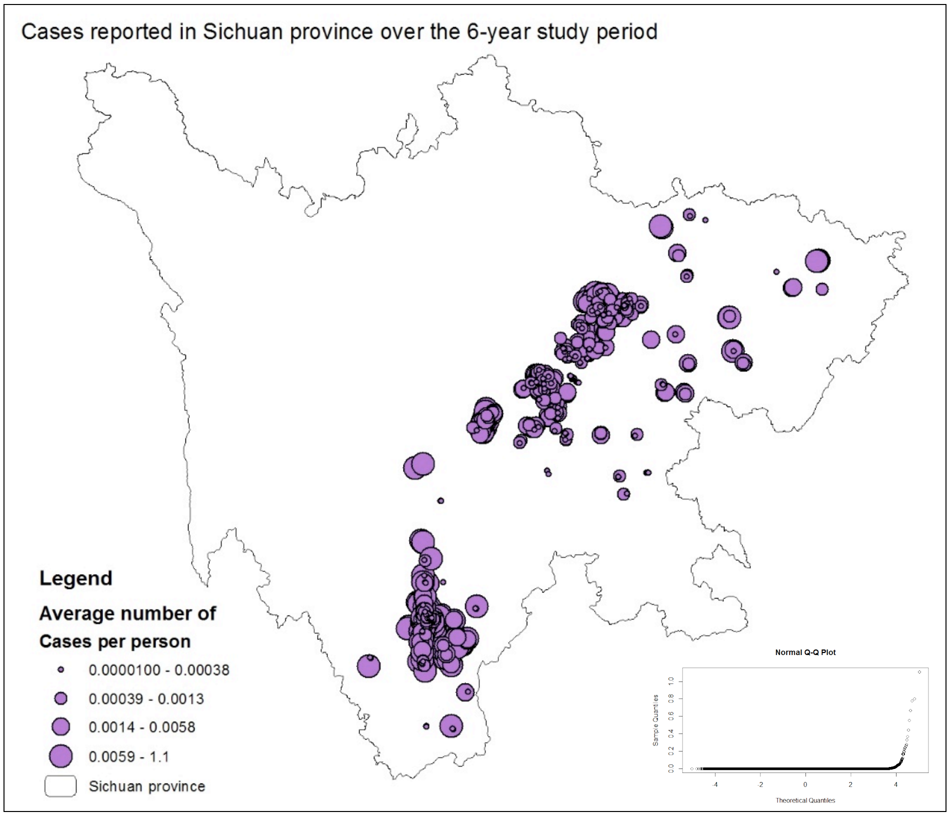


Figure 2: Spatial distribution of cases per person for the period from 2005-2011 in Sichuan province, China. For reference, a Q-Q plot showing the distribution of the average numbers of cases per person has been included in the lower right-hand corner.

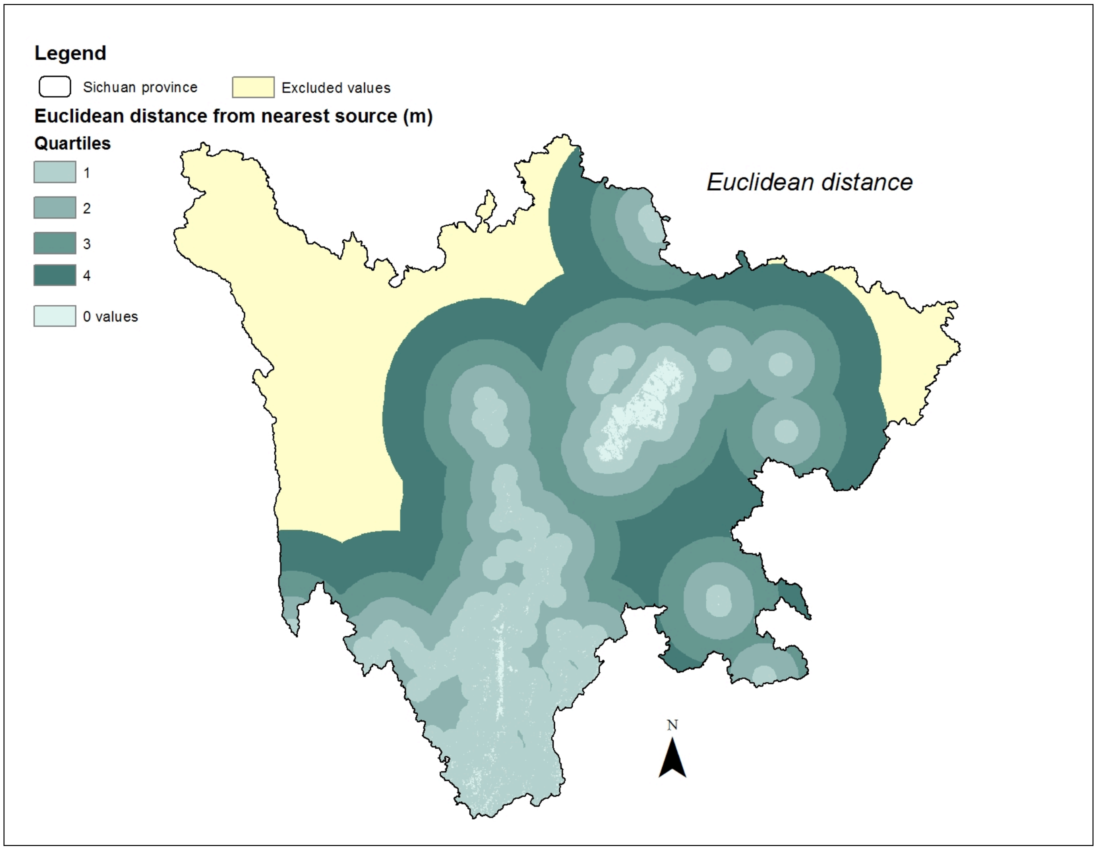


Figure 3: Spatial distribution of quartiles of Euclidean distance from each location to the nearest 'source' in Sichuan province. Some of the data was excluded from the analysis to ensure that all models would converge.

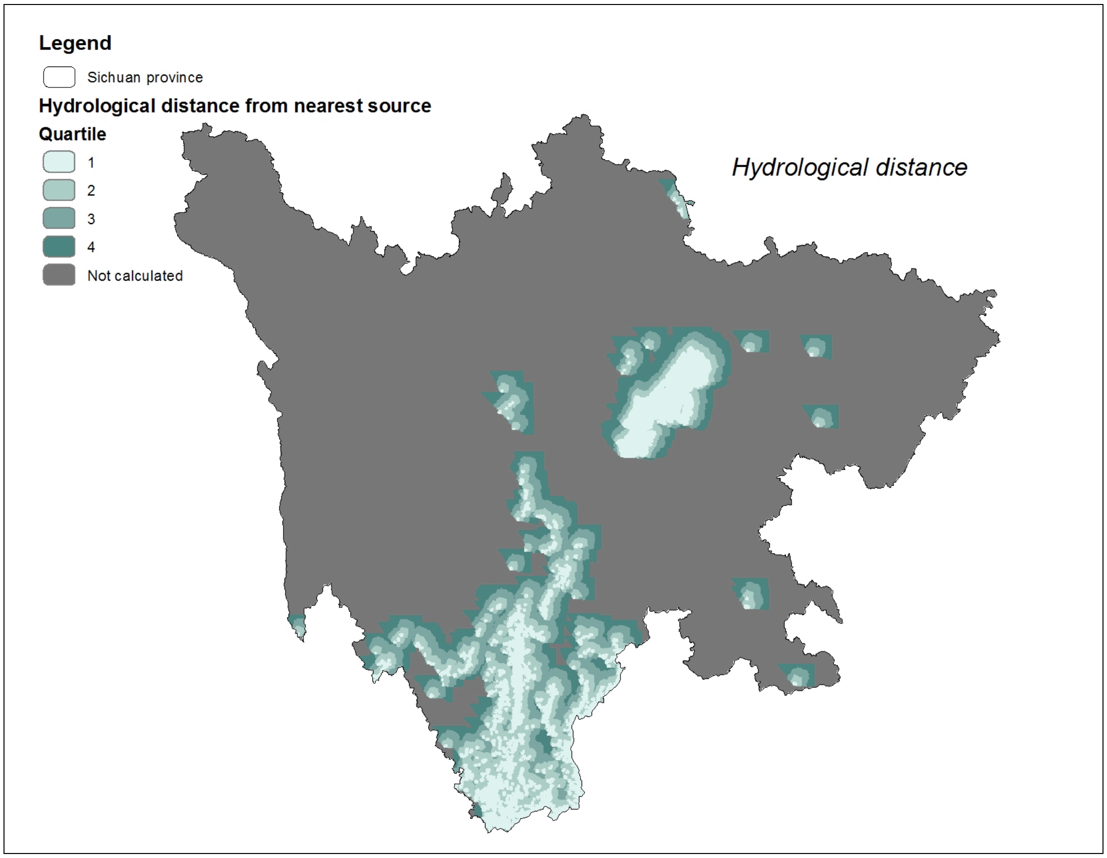


Figure 4: Spatial distribution of quartiles and areas not calculated for hydrological distance of each location to the nearest upstream source.

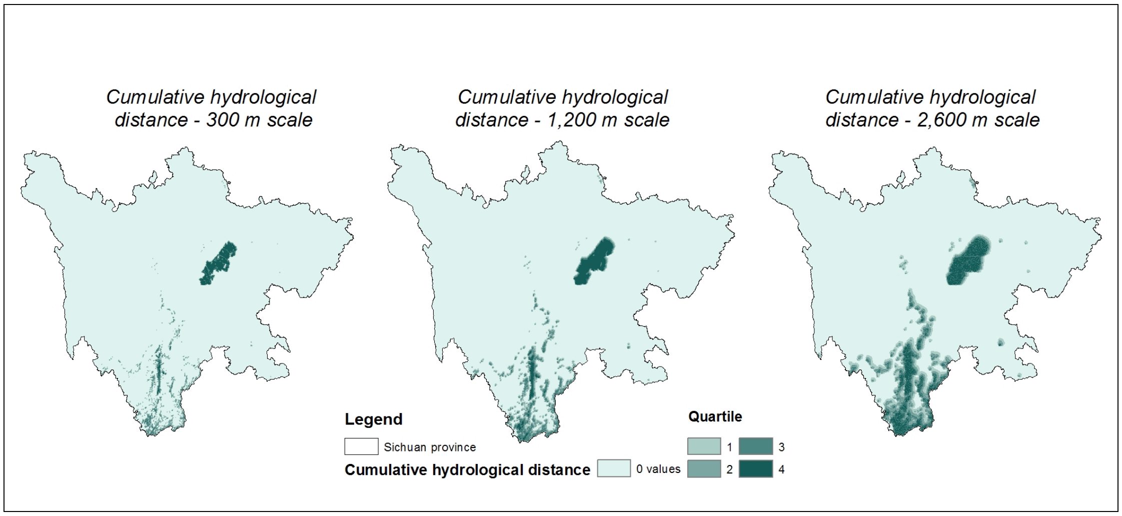
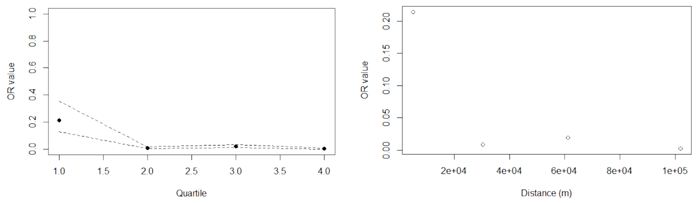
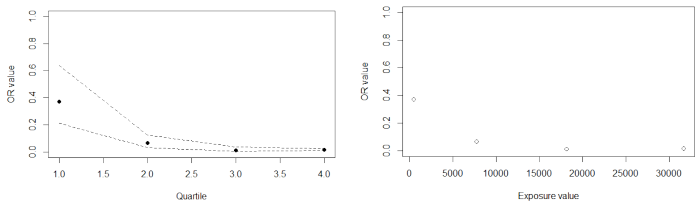
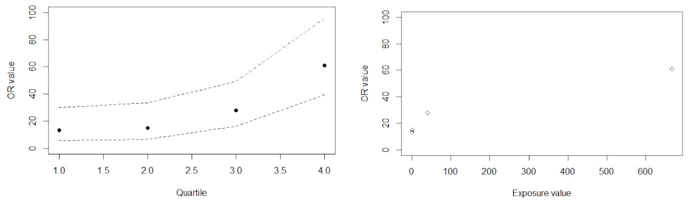
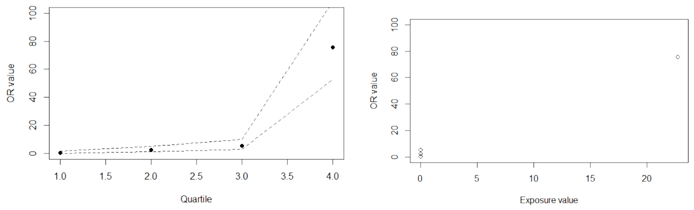

Cumulative hydrological distance - 300 m scale
Cumulative hydrological distance - 1,200 m scale
Cumulative hydrological distance - 2,600 m scale
Legend
Sichuan province
Cumulative hydrological distance
0 values
Quartile
1
2
3
4

Figure 5: Spatial distribution of quartiles of cumulative hydrological distances for all three scales of shared downstream risk.

Dispersion Scheme	AIC	OR values – Shown on both quartile and value-based scales
	Moran's I (Z-score)	
Binary definition of location as containing a source or not	5,264.13	27.94 [18.36, 42.51]
	0.0008 (3.74)	
Euclidean distance	5,061.28	
	0.0015 (5.63)	
Hydrological distance	5,025.70	
	0.0014 (5.29)	
Cumulative hydrological distance – 300 m scale	5,097.89	
	0.0002 (1.98)	
Cumulative hydrological distance – 1,200 m scale	4,951.26	
	0.0028 (8.66)	

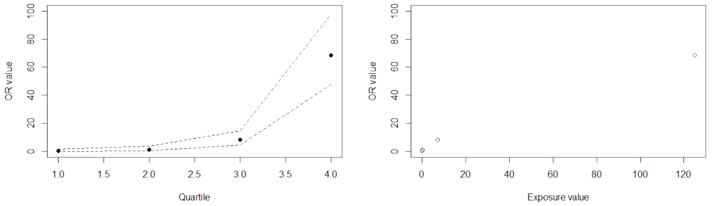
First order decayed accumulated hydrological distance – 2,600 m scale	4,977.48	
	0.0019 (6.05)	

Table 2: OR and AIC values for each of the exposure models