International Journal for Parasitology xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

Prediction of hookworm prevalence in southern India using environmental parameters derived from Landsat 8 remotely sensed data

Alexandra V. Kulinkina [a,b,*], Rajiv Sarkar [c], Venkata R. Mohan [d], Yvonne Walz [e], Saravanakumar P. Kaliappan [c], Sitara S.R. Ajjampur [c], Honorine Ward [c,f], Elena N. Naumova [c,g], Gagandeep Kang [c]

[a] Department of Public Health and Community Medicine, Tufts University School of Medicine, Boston, MA, USA
[b] Partners In Health, Neno, Malawi
[c] Division of Gastrointestinal Sciences, Christian Medical College, Vellore, Tamil Nadu, India
[d] Department of Community Health, Christian Medical College, Vellore, Tamil Nadu, India
[e] Institute for Environment and Human Security, United Nations University, Bonn, Germany
[f] Division of Geographic Medicine and Infectious Diseases, Tufts Medical Center, Boston, MA, USA
[g] Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA, USA

ARTICLE INFO

Article history:
Received 20 June 2019
Received in revised form 30 September 2019
Accepted 3 October 2019
Available online xxxx

Keywords:
Hookworm
Soil-transmitted helminths
Spatial modelling
India
Remote sensing

ABSTRACT

Soil-transmitted helminth infections propagate poverty and slow economic growth in low-income countries. As with many other neglected tropical diseases, environmental conditions are important determinants of soil-transmitted helminth transmission. Hence, remotely sensed data are commonly utilised in spatial risk models intended to inform control strategies. In the present study, we build upon the existing modelling approaches by utilising fine spatial resolution Landsat 8 remotely sensed data in combination with topographic variables to predict hookworm prevalence in a hilly tribal area in southern India. Hookworm prevalence data collected from two field surveys were used in a random forest model to investigate the predictive capacity of 15 environmental variables derived from two remotely sensed images acquired during dry and rainy seasons. A variable buffer radius (100–1000 m) was applied to the point-prevalence locations in order to integrate environmental conditions around the village centroids into the modelling approach and understand where transmission is more likely. Elevation and slope were the most important variables in the models, with lower elevation and higher slope correlating with higher transmission risk. A modified normalised difference water index was among other recurring important variables, likely responsible for some seasonal differences in model performance. The 300 m buffer distance produced the best model performance in this setting, with another spike at 700 m, and a marked drop-off in R^2 values at 1000 m. In addition to assessing a large number of environmental correlates with hookworm transmission, the study contributes to the development of standardised methods of spatial linkage of continuous environmental data with point-based disease prevalence measures for the purpose of spatially explicit risk profiling.

1. Introduction

Soil-transmitted helminth (STH) infections cause a significant health burden in low-income countries (Montresor et al., 2002; Pullan et al., 2014). Of particular importance are roundworms (*Ascaris lumbricoides*), whipworms (*Trichuris trichiura*) and hookworms (*Necator americanus* or *Ancylostoma duodenale*) (Bethony et al., 2006). These infections are associated with anaemia, malnutrition, stunting and impaired cognitive development (Montresor et al., 2002; Bethony et al., 2006), propagating poverty and slowing economic growth in the affected countries. Interactions with other infectious diseases such as malaria, tuberculosis and HIV have also been documented (Thigpen et al., 2011; Webb et al., 2012; Salgame et al., 2013).

Humans acquire STH infections by ingesting parasite eggs (*A. lumbricoides* and *T. trichiura*) or having parasite larvae found in soil penetrate through the skin (hookworm) (Bethony et al., 2006). Because parts of the parasite lifecycle occur in the environment (i.e. soil), environmental conditions are important determinants of transmission for all STH species. For example, temperatures in

* Corresponding author at: Tufts University School of Medicine, 145 Harrison Ave, Boston, MA, USA.
E-mail address: alexandra.kulinkina@tufts.edu (A.V. Kulinkina).

https://doi.org/10.1016/j.ijpara.2019.10.001
0020-7519/

the range of 20–30 °C, adequate soil moisture and relative atmospheric humidity provide the most suitable conditions for STH larval survival and development (Brooker and Michael, 2000; Brooker et al., 2006a). Among other environmental factors contributing to STH transmission are soil type, rainfall and altitude (Brooker and Michael, 2000). Poverty and inadequate water, sanitation and hygiene (WASH) conditions also play an important role in environmental contamination with parasite eggs through unsafe faecal management (de Silva et al., 2003; Bethony et al., 2006).

STH infections are neglected tropical diseases (NTDs) that tend to affect the poorest populations and lack funding priority (Morel, 2003; Bethony et al., 2006; King, 2015). Many of the NTDs present with primarily non-fatal chronic symptoms, which have historically caused underestimation of their effects on economic development. In 2001, the World Health Assembly put several NTDs on the global priority agenda, which has enabled many low-income countries to deploy large-scale preventive chemotherapy campaigns (WHO, 2001). Although treatment with albendazole has shown to be effective in reducing STH prevalence and worm burden in the short term (Sunish et al., 2015), rapid reinfection calls into question the long-term sustainability of the approach. Hookworm prevalence of 30% and 55% of the pre-treatment levels three and six months post-treatment, respectively, have been documented (Jia et al., 2012). Hence, elimination of these infections is unlikely through preventive chemotherapy alone, without complementary vaccine development (Loukas et al., 2006), information, education and communication (IEC) and improvements in WASH conditions (Ziegelbauer et al., 2012; Freeman et al., 2013; Campbell et al., 2014; Strunz et al., 2014; Coffeng et al., 2015).

Control efforts for STH and other NTDs are hindered by a lack of accurate prevalence data. Hence, spatial predictive models play an important role in identifying transmission hotspots and high-risk communities in order to allocate limited resources (Utzinger et al., 2003). Because many NTDs are environmentally mediated, increasing use of remotely sensed (RS) data to characterise environmental conditions of transmission hotspots has also helped achieve better understanding of the spatial distribution of these diseases (Reiss et al., 2013). In a brief review of modelling studies that utilised RS data to predict STH transmission risk (Table 1), most were conducted at large spatial extents (e.g. national), utilised coarse spatial resolution RS data (e.g. 1 km, 8 km and 50 km), and included relatively few RS environmental predictors (e.g. normalised difference vegetation index (NDVI), land surface temperature (LST) and elevation).

The present study builds upon the existing sub-national predictive modelling approaches using geocoded community level hookworm prevalence data from two surveys (Kaliappan et al., 2013; Sarkar et al., 2017) conducted in tribal areas of Tamil Nadu, India. We utilised fine resolution Landsat 8 RS data in combination with topographic variables, expanded the number of RS environmental predictors to 15, and tested a variable radius (100–1000 m) for aggregating continuous RS data for point-prevalence locations (Reiss et al., 2013; Walz et al., 2015).

The use of buffer distances around point-prevalence locations for extraction of environmental variables is rare in the literature, although highly relevant when using fine resolution RS data for small spatial extent modelling. In just three prior studies, a buffer radius of 1000 m has been found relevant for hookworm modelling (Reiss et al., 2013) and 1000 m (Kulinkina et al., 2018) or 5000 m (Walz et al., 2015) for schistosomiasis modelling. The justification of this approach lies in that point locations where infection status is measured or to which prevalence values are aggregated, are not necessarily representative of where transmission occurs. Hookworm is most likely to be transmitted in areas that are contaminated by faeces (e.g. open defecation fields), have suitable environmental conditions, and where people walk barefoot. Although these areas most likely represent the immediate surroundings of the households (Brooker et al., 2006b; Pullan et al., 2010), this phenomenon is highly dependent on cultural habits. We investigated this phenomenon in the present study, using a variable buffer radius that encompasses a wider area where community members live, work and play.

2. Materials and methods

2.1. Study area

The study was conducted in Jawadhu Hills, a tribal area located in the northern part of Tamil Nadu, India (Fig. 1). Approximately 80,000 people live in these hills, organised into ~250 agrarian communities of 15–100 households. The area has poor access to roads and WASH facilities (Kaliappan et al., 2013). Most households have no access to a toilet (>99%), obtain drinking water from a communal tap stand (89%) and keep animals in or near the home (84%) (Kaliappan et al., 2013; Sarkar et al., 2017). The area is endemic for STH, with predominantly hookworm infections. According to a baseline survey conducted in 2011–2012, average STH prevalence among children in the 6–15 years age group was >30% in Jawadhu Hills (Kaliappan et al., 2013) and was significantly higher than the regional average of 7.8% measured among school children of similar ages in 2008–2009 (Kattula et al., 2014).

2.2. Ethial approval

This manuscript presents the results of secondary data analysis that did not require any additional field data collection. Both field studies from which data were obtained for the models conducted in this manuscript were approved by the Christian Medical College (CMC) Institutional Review Board (IRB), India. Written informed consent was obtained from adult participants (aged ≥18 years). For children <18 years of age, a parent/legal guardian provided written consent; additionally, children aged 8–17 years provided oral assent (Kaliappan et al., 2013; Sarkar et al., 2017).

2.3. Data sources

Data for this study were obtained from satellite RS sources and field studies. Surface reflectance, thermal and elevation data were obtained from RS sources. From these, vegetation- and moisture-related indices, LST and topographic predictor variables were derived. The outcome variable, aggregated community-level hookworm prevalence (%), was obtained from previously published field studies. Methods used to derive the analysis variables are described below.

2.3.1. Outcome: Hookworm prevalence data

Hookworm prevalence data were obtained from two studies carried out by the CMC. Survey 1 was a prevalence survey and survey 2 was a baseline survey for a cluster randomised community intervention trial. Survey 1 was conducted between November 2011 and April 2012 in 37 geocoded villages. A total of 1237 individuals participated, with a median (inter-quartile range (IQR)) of 30 (20–41) participants per village. Average hookworm prevalence in this survey, determined by analysing five stool samples per participant, was 37.9% and ranged between 16.7% and 77.5% (Kaliappan et al., 2013). Survey 2 was conducted between October 2013 and November 2014 in 45 villages (with an overlap of 13 villages that were included in both studies). A total of 2082 individuals participated, with a median (IQR) of 46 (44–48) participants per village. Average hookworm prevalence in this survey, determined by analysing three stool samples per participant, was lower at

Please cite this article as: A. V. Kulinkina, R. Sarkar, V. R. Mohan et al., Prediction of hookworm prevalence in southern India using environmental parameters derived from Landsat 8 remotely sensed data, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.10.001

Table 1
Summary of environmental modelling of soil-transmitted helminth prevalence.

Infections	Spatial extent	Environmental predictors	Reference
Hookworm	Sub-national (South Africa) Prevalence data obtained from various types of surveys	Elevation, temperature, rainfall, soil type	Mabaso et al. (2003)
Hookworm	Sub-national (South Africa) Prevalence aggregated at household level	NDVI, soil properties, population density	Saathoff et al. (2005)
Hookworm *Schistosoma mansoni*	Sub-national (Côte d'Ivoire) Prevalence aggregated at school level	Elevation, slope, rainfall, LST, NDVI, land cover, soil type, distance to waterbodies	Raso et al. (2006)
Hookworm	Sub-national (Brazil) Prevalence aggregated at household level	Watershed, NDVI, population density	Pullan et al. (2008)
Hookworm	Regional (Kenya/Tanzania) Prevalence aggregated at school level	Elevation, LST, NDVI, distance to waterbodies	Brooker and Clements (2009)
Hookworm	National (Ghana) Prevalence aggregated at school level	LST, NDVI, distance to waterbodies	Soares Magalhães et al. (2011)
Ascaris lumbricoides Trichuris trichiura Hookworm	National (Bolivia) Prevalence data obtained from various types of surveys	Climate variables, elevation, LST, NDVI, EVI, land cover, soil properties, population density	Chammartin et al. (2013)
A. lumbricoides T. trichiura Hookworm	National (China) Prevalence data obtained from various types of surveys	Rainfall, climate zones, elevation, LST, NDVI, rainfall, land cover, soil type, soil properties, distance to waterbodies, population density	Lai et al. (2013)
Hookworm	Sub-national (Tanzania) Prevalence aggregated at household level	Elevation, slope, EVI, LST, rainfall, population density, sanitation coverage	Reiss et al. (2013)
A. lumbricoides T. trichiura Hookworm	National (Brazil) Prevalence aggregated at municipality level	Climate variables, elevation, LST, NDVI, EVI, potable water and sanitation coverage, population density	Scholte et al. (2013)
A. lumbricoides T. trichiura Hookworm	National (Philippines) Prevalence data aggregated at sub-district (barangay) level	Rainfall, LST, NDVI, distance to waterbodies	Soares Magalhães et al. (2015)
A. lumbricoides T. trichiura Hookworm	Regional (Bangladesh/India/ Nepal/ Pakistan) Prevalence data obtained from various types of surveys	LST, NDVI, land cover, elevation, climate zones and variables, socioeconomic data, waterbodies, soil and water properties, population density	Lai et al. (2019)

NDVI, normalised difference vegetation index; LST, land surface temperature; EVI, enhanced vegetation index.

18.7% and ranged between 2.1% and 44.2% (Sarkar et al., 2017). Community-level prevalence (inclusive of adults and children) from these surveys (Fig. 1) was used as the outcome variable in the present modelling analysis.

2.3.2. Predictors: environmental RS data

Landsat 8 data were obtained from United States Geological Survey (USGS) Earth Explorer (http://earthexplorer.usgs.gov/) as level 2 data products, which had been atmospherically corrected. These products contained spectral bands (1–9) as surface reflectance values with spatial resolution of 30 m from the Operational Land Imager (OLI) and thermal bands (10 and 11) with spatial resolution of 100 m from the Thermal InfraRed Sensor (TIRS). All available images that encompassed the study area (path 143 row 51) from April 2013 through December 2014 were screened for quality. A total of five images that were minimally affected by clouds (<10% of the pixels) with the following acquisition dates that coincided with field data collection of survey 2 were downloaded: 14 Feb 2014; 24 Mar 2014; 27 May 2014; 15 Aug 2014 and 02 Oct 2014 (Supplementary Table S1).

Although images were not available for the time period of survey 1, the 2014 images were used for analysis with both surveys, based on the assumption that climate patterns were similar for the 2011–2014 time period. This assumption was validated using publicly available meteorological data for the nearest city of Chennai, India (Supplementary Fig. S1). Of the five available images, two images were used in the analysis in order to test the potential effect of different environmental conditions on model performance: the image acquired on 27 May 2014 represented dry and hot conditions and the image acquired on 02 Oct 2014 represented cooler and more humid/rainy conditions (Supplementary Fig. S1).

ASTER Global Digital Elevation Model (GDEM v2) data were obtained from USGS Global Data Explorer (gdex.cr.usgs.gov) with a spatial resolution of 30 m. A moving window (3x3) majority filter was applied to the elevation data to eliminate image artefacts (Walz, Y., 2014. Remote sensing for disease risk profiling: a spatial analysis of schistosomiasis in West Africa. PhD Thesis, University of Würzburg, Germany) using the Spatial Analyst extension in ArcGIS software (version 10.2.2).

2.4. Data processing

RS data pixels affected by clouds or cloud shadows were masked using the quality assurance band. Spectral bands were used to compute five vegetation indices (normalised difference vegetation index (NDVI), enhanced vegetation index (EVI), soil-adjusted vegetation index (SAVI), modified soil-adjusted vegetation index (MSAVI) and normalised difference moisture index (NDMI)) and two water indices (normalised difference water index (NDWI) and modified normalised difference water index (MNDWI)) (Supplementary Table S2). Thermal bands were used to derive LST in R software (version 3.4.3). Elevation data were used to derive slope in ArcGIS software (version 10.2.2). Data processing steps are further described in Supplementary Fig. S2 and the distribution of RS parameter values over the study area are shown in Supplementary Fig. S3.

Please cite this article as: A. V. Kulinkina, R. Sarkar, V. R. Mohan et al., Prediction of hookworm prevalence in southern India using environmental parameters derived from Landsat 8 remotely sensed data, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.10.001

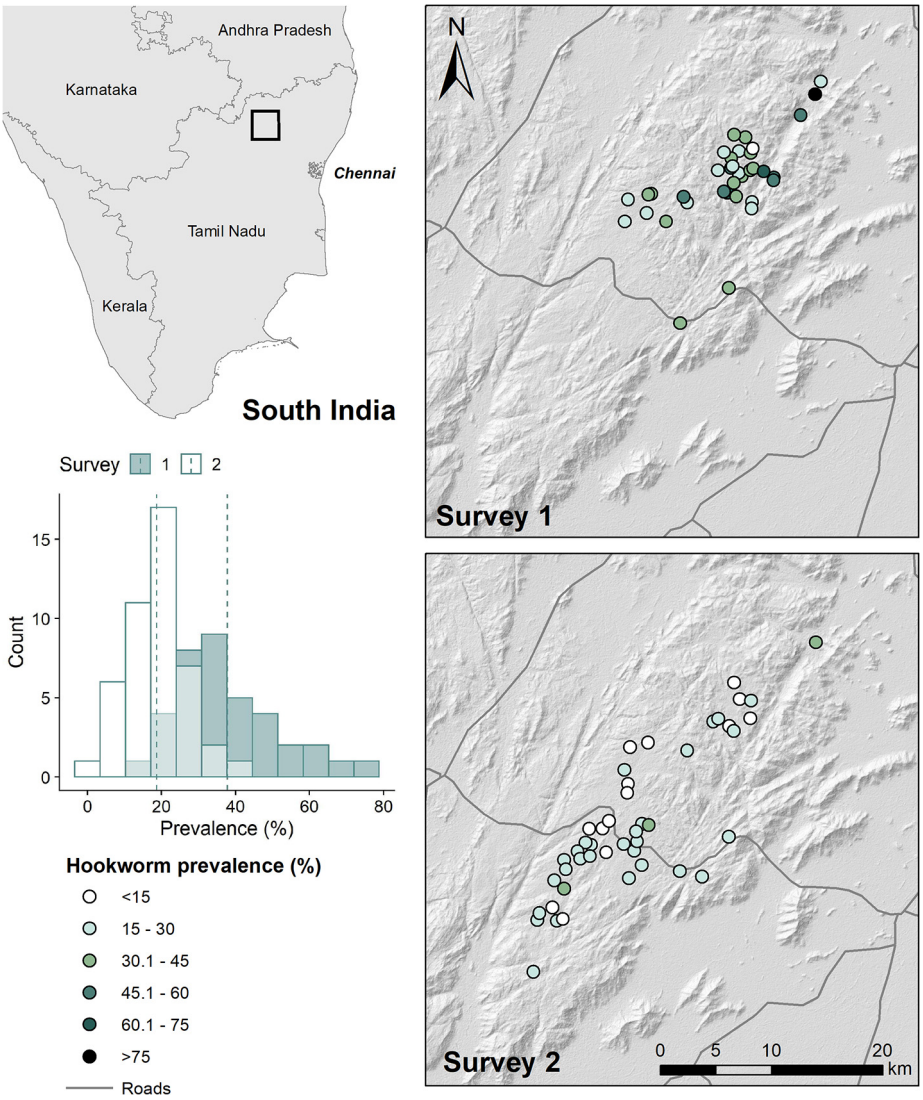

Fig. 1. Map of Jawadhu Hills in southern India and spatial distribution of hookworm prevalence from two surveys created using the following data sources: hillshade relief was derived from the ASTER Global Digital Elevation Model (v2); stool samples for estimating hookworm prevalence were collected by Christian Medical College study teams. The histogram shows the distribution of prevalence values with mean values denoted by dashed vertical lines. Survey 1 was conducted between November 2011 and April 2012; survey 2 was conducted between October 2013 and November 2014.

2.5. Variable extraction and aggregation

Aggregated community-level point-prevalence of hookworm (% positive samples) was used as the outcome variable. A total of 15 environmental predictor variables were derived from RS data and resampled to a matching 30 m spatial resolution for analysis (Table 2). While the environmental predictors were represented by continuous raster data, prevalence was represented by point data, necessitating extraction and aggregation of the raster data. A variable buffer radius (100–1000 m) around each point-prevalence location was used by extracting the median pixel value from the buffer area to be matched to each prevalence measure.

2.6. Data analysis

Exploratory analyses included graphical variable summaries conducted in R and ArcGIS software. The final analysis consisted of non-parametric random forest models conducted with all 15 environmental predictor variables (Table 2). The random forest approach was chosen because it can deal with continuous outcome data, multicollinear predictor variables and low numbers of training samples. It is the recommended machine learning method for generating predictions (Kampichler et al., 2010) that has been successfully applied in similar studies (Walz et al., 2015; Kulinkina et al., 2018).

The goal of this analysis was to compare the performance of RS data acquired in dry and wet conditions and extracted using different buffer distances. Furthermore, three versions of the outcome were tested: model 1 (M1) used survey 1 data only; model 2 (M2) used survey 2 data only; and combined model (CM) used a combination of the two surveys. In total, 60 variants of the random forest model were conducted: 10 buffer distances over which RS variables were extracted (100–1000 m) * two RS images (representing dry and wet conditions) * three sets of prevalence data (M1, M2 and CM).

The explanatory power of random forest models was compared using root-mean-square error (RMSE) and R^2 values (Li et al., 2017). The relative importance of predictor variables was assessed using the increasing node purity ("IncNodePurity") metric (Grömping, 2009; Hastie et al., 2009). Subsequently, random forest

Please cite this article as: A. V. Kulinkina, R. Sarkar, V. R. Mohan et al., Prediction of hookworm prevalence in southern India using environmental parameters derived from Landsat 8 remotely sensed data, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.10.001

Table 2
Summary of 15 continuous environmental predictor variables.

Data source	Variable	Spatial resolution (m)
Landsat 8 (OLI)	Blue band reflectance	30
Landsat 8 (OLI)	Green band reflectance	30
Landsat 8 (OLI)	Red band reflectance	30
Landsat 8 (OLI)	Near infrared (NIR) band reflectance	30
Landsat 8 (OLI)	Short-wave infrared (SWIR) band reflectance	30
Landsat 8 (TIRS)	Land surface temperature (LST) (°C)	100
Landsat 8 (OLI)	Normalised difference vegetation index (NDVI)	30
Landsat 8 (OLI)	Enhanced vegetation index (EVI)	30
Landsat 8 (OLI)	Soil-adjusted vegetation index (SAVI)	30
Landsat 8 (OLI)	Modified soil-adjusted vegetation index (MSAVI)	30
Landsat 8 (OLI)	Normalised difference moisture index (NDMI)	30
Landsat 8 (OLI)	Normalised difference water index (NDWI)	30
Landsat 8 (OLI)	Modified normalised difference water index (MNDWI)	30
GDEM v2	Elevation (m)	30
GDEM v2	Slope (°)	30

OLI, Operational Land Imager; TIRS, Thermal InfraRed Sensor; GDEM, Global Digital Elevation Model.

models were applied back to the data cube of predictor variables to derive continuous predicted prevalence surfaces. Lastly, the median predicted values were plotted against observed prevalence values as scatter plots. The quality of the prediction was assessed using Spearman's rank correlation (r value) between model predicted and observed values, compared with the line of equality, as well as plotting the average of the observed and predicted values against their difference and assessing the proportion of predicted observations within the 95% limits of agreement (q value) (Bland and Altman, 2003).

3. Results

3.1. Exploratory analyses

As part of exploratory analyses, distributions of median RS parameter values from two images and 10 buffer distances were compared using boxplots (Supplementary Fig. S4). This analysis showed significant variability in parameter values across the two images. Individual band values (blue, green, red, NIR and SWIR) and vegetation indices (NDVI, EVI, SAVI, MSAVI and NDMI) were significantly lower in the image acquired in hot and dry conditions (27 May 2014) compared with cooler and more humid conditions (02 Oct 2014). There were no apparent differences between the two images in NDWI values; whereas the MNDWI values were significantly higher in October than May. LST values were slightly lower in May (26–28 °C) than in October (28–29 °C). Median variable values varied little across the 10 buffer distances; however, RS variables tended to exhibit slightly more variability over shorter than longer buffer distances.

Hookworm prevalence values in the subset of 13 villages that were sampled in both surveys were also explored. In this subset, average prevalence values in survey 1 were approximately two times higher than in survey 2 (35.6% versus 17.4%). Only four villages exhibited similar prevalence values during both survey rounds. No notable spatial trends were detected by mapping these villages (Supplementary Fig. S5).

3.2. Random forest models

Results of the random forest models showed better fit for locations in survey 1 compared with survey 2, as shown by higher R^2 values (Fig. 2). The model fit for survey 1 models (M1) applied to both RS images was similar. Spikes in R^2 values at 300 m and 700 m and a marked drop off at 1000 m indicated that the fit depended on buffer distance. Similar trends were observed in RMSE values (Fig. 2), with an increase in RMSE at 900 m and 1000 m buffer distances. The trends in model fit exhibited by survey 2 models (M2) were slightly different, with only one peak per image. The best data fit was achieved at 300 m using the 02 Oct 2014 image and at 700 m using the 27 May 2014 image according to R^2 values. In the combined models (CM), R^2 values generally peaked at 300–400 m and declined with increasing buffer distance.

To explore the best performing models further, variable importance was extracted from a total of six models (two RS images * one buffer distance (300 m) * three prevalence datasets). In survey 1 models (M1), the most important variables were elevation and slope, regardless of RS image used (Supplementary Fig. S6). In survey 2 models (M2), the near infrared (NIR) band and MNDWI were the only visibly more important variables when the 27 May 2014 RS image was used. When using the 02 Oct 2014 RS image, all variables had low IncNodePurity values. In the combined models (CM), elevation and slope re-gained their importance (Supplementary Fig. S6). Of the seasonal RS variables, the most important across all models were the NIR band, LST, NDMI and MNDWI.

As expected, M1 predicted prevalence (median pixel value extracted over the 300 m buffer distance) had the highest Spearman's rank correlation coefficient compared with survey 1 results: r = 0.66 (P < 0.05) with the 27 May 2014 image and r = 0.55 (P < 0.05) with the 02 Oct 2014 image. M2 and CM models produced lower correlation coefficients, which remained consistently significant when compared with survey 1 results but were largely non-significant when compared with survey 2 results (Supplementary Fig. S9). Bland-Altman plots illustrated similar trends in q values, with higher r values generally corresponding to higher q values.

Predicted surfaces were also derived for the same six models described above. Elevation was negatively associated and slope was positively associated with hookworm risk in M1 and CM models conducted with both RS images, as represented by low risk areas in the middle of the image (high elevation and low slope) and high risk along the edges of the hills (moderate elevation and high slope) (Fig. 3; Supplementary Fig. S5). No clear associations were produced by the M2 models.

4. Discussion

In this study, we utilised publicly available environmental data from the Landsat 8 satellite, in combination with topographic variables, to predict hookworm prevalence at a sub-national spatial extent (an area of approximately 50 km^2). Furthermore, we used a larger number of RS environmental predictors than previous modelling studies and conducted 60 iterations of the model with three prevalence datasets, two RS images acquired under different climatic conditions and 10 buffer distances used for environmental variable extraction.

Survey 1 models achieved consistently higher R^2 values than survey 2 models. This is not surprising, considering a smaller sample size, higher average prevalence and a wider prevalence range.

Please cite this article as: A. V. Kulinkina, R. Sarkar, V. R. Mohan et al., Prediction of hookworm prevalence in southern India using environmental parameters derived from Landsat 8 remotely sensed data, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.10.001

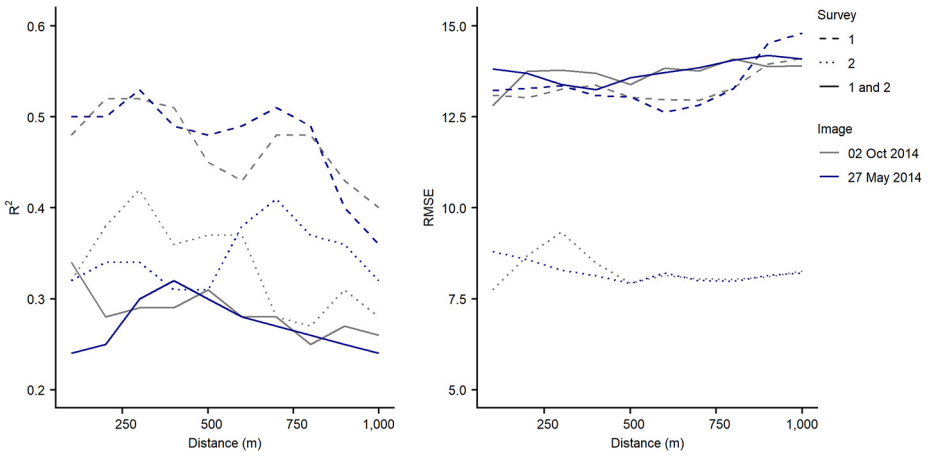

Fig. 2. Random forest model R^2 and root–mean–square error values (*Y*-axis) for 10 buffer distances (*X*-axis), two Landsat 8 images (line colour) and three versions of the model, using data from survey 1, survey 2 and both surveys combined (line type).

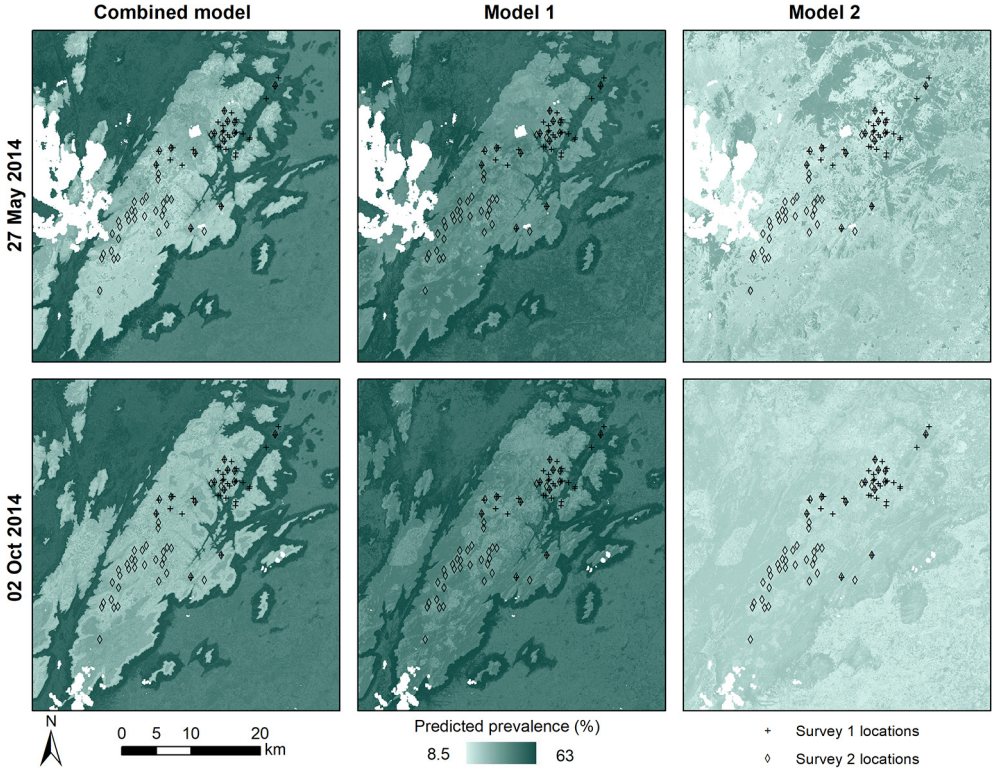

Fig. 3. Predicted hookworm prevalence from six random forest models using two Landsat 8 remote sensing images and three datasets (model 1 used survey 1 data, model 2 used survey 2 data, and combined model used data from both surveys). White areas represent pixels in the remote sensing image affected by clouds that were excluded from analysis.

There are several potential contributing factors to survey 1 exhibiting higher prevalence values than survey 2. First, survey 1 was conducted earlier, when preventive chemotherapy was less prevalent. The effects of treatment are shown by generally lower prevalence values in the subset of 13 villages that participated in both surveys. Second, survey 1 was conducted in a different subset of villages, which were more remote than those used in survey 2 and were therefore less likely to receive treatment. Third, fewer stool samples were collected during survey 2, meaning that fewer chances to detect less severe infections were available, likely underestimating the true prevalence (Knopp et al., 2008; Kaliappan et al., 2013).

Elevation was by far the most important variable in survey 1 models, and was negatively associated with hookworm risk, consistent with the findings of many but not all of the reviewed studies (Mabaso et al., 2003; Brooker and Clements, 2009; Reiss et al., 2013; Scholte et al., 2013). Some found no association between elevation and hookworm risk (Chammartin et al., 2013; Lai et al., 2013), while one study found a positive association, where high prevalence communities were situated at elevations ≥400 m (Raso et al., 2006). Slope was consistently positively associated with hookworm risk, a variable that has not been explored in many other studies. NDMI and MNDWI were among other recurring

Please cite this article as: A. V. Kulinkina, R. Sarkar, V. R. Mohan et al., Prediction of hookworm prevalence in southern India using environmental parameters derived from Landsat 8 remotely sensed data, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.10.001

important variables in dry conditions. Similar to the findings of another study that utilised Landsat 8 data for schistosomiasis risk modelling in Ghana (Kulinkina et al., 2018), we found that NDWI was less sensitive in detecting waterbodies compared with MNDWI, frequently misclassifying developed surfaces (i.e. roads and settlements) as waterbodies (Supplementary Fig. S3).

The buffer distance of 300 m produced the best model performance. In the present study, >99% of the households did not have access to a toilet and practised open defecation, suggesting defecation fields as a likely source of environmental contamination within the 300 m radius of the village centroid. In fact, during survey 2, a total of 133 defecation fields were mapped in the vicinity of the 45 study villages. Of these, 43 (32%) were within 100 m, 104 (78%) within 200 m and 127 (95%) within 300 m of the village centroids. There was another notable spike in model R^2 values at around 700 m from the village centroids, indicating that perhaps a secondary source of contamination may be present (e.g. agricultural areas or defecation fields of neighbouring villages).

Overall, even in survey 1 models, predicted prevalence values deviated substantially from the observed values, signifying some overprediction in the low prevalence range and underprediction in the high prevalence range. The high Spearman's rank correlation value is driven by the two clusters visible in the scatter plot (Supplementary Fig. S6). The cluster with high observed values and higher predicted values consists of villages located in the high risk (low elevation and high slope) area that forms a diagonal canyon in the top right corner of the study area. The cluster of lower predicted and observed prevalence values consists of villages located in the lower risk (high elevation and low slope) area in the centre of the study area (Fig. 3).

Cloud cover presented a substantial challenge in RS data acquisition. In a prior study (Kulinkina et al., 2018), RS images were available only in the dry season. In this study, we were able to locate one rainy season image to use for comparison. However, only five good quality images were available over the entire duration of field data collection, with some still substantially affected by cloud cover (up to 10% of the pixels). Fortunately, the locations of the study villages were largely unaffected by pixels with unacceptable data quality according to the quality assurance information, but we cannot say with certainty that other, less significant, problems with data quality were not present.

Despite the stated limitations, our study makes important contributions to the modelling approaches of hookworm transmission at small spatial extents. First, we found that villages that are located on steeper slopes at moderate elevations are at higher risk of hookworm transmission than villages located in higher and flatter locations. Second, it justifies the use of a buffer radius for extracting environmental variables to link with point-prevalence data. In the Jawadhu Hills, the radius of 300 m was most appropriate, and consistent with most (95%) of the defecation fields being located within 300 m of the village centroids. However, it is a unique tribal setting and these findings should be validated in a more generic location and for other STH species. The importance of elevation, slope and MNDWI variables in the models suggests that the dynamic between elevation, slope and runoff in hilly areas with poor sanitation and high STH transmission warrants further study.

Acknowledgements

The authors wish to acknowledge the Christian Medical College, India study teams who conducted field data collection and laboratory analysis. This study was funded in part by the National Institutes of Health, (D43 TW009377-01A1) and Tufts University School of Medicine, United States (Natalie V. Zucker research grant). The STH cluster randomised trial was supported by the India Alliance, India through an Early Career Fellowship to R. Sarkar (IA/E/12/1/500750). We also appreciate the helpful comments and suggestions of the International Journal for Parasitology reviewers and editor.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpara.2019.10.001.

References

Bethony, J., Brooker, S., Albonico, M., Geiger, S.M., Loukas, A., Diemert, D., Hotez, P.J., 2006. Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. Lancet 367, 1521–1532.

Bland, J.M., Altman, D.G., 2003. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet. Gynecol. 22, 85–93.

Brooker, S., Alexander, N., Geiger, S., Moyeed, R.A., Stander, J., Fleming, F., Hotez, P.J., Correa-Oliveira, R., Bethony, J., 2006a. Contrasting patterns in the small-scale heterogeneity of human helminth infections in urban and rural environments in Brazil. Int. J. Parasitol. 36, 1143–1151.

Brooker, S., Clements, A.C.A., 2009. Spatial heterogeneity of parasite co-infection: determinants and geostatistical prediction at regional scales. Int. J. Parasitol. 39, 591–597.

Brooker, S., Clements, A.C.A., Bundy, D.A.P., 2006b. Global epidemiology, ecology and control of soil-transmitted helminth infections. Adv. Parasitol. 62, 221–261.

Brooker, S., Michael, E., 2000. The potential of geographical information systems and remote sensing in the epidemiology and control of human helminth infections. Adv. Parasitol. 47, 245–288.

Campbell, S.J., Savage, G.B., Gray, D.J., Atkinson, J.A.M., Soares Magalhães, R.J., Nery, S.V., McCarthy, J.S., Velleman, Y., Wicken, J.H., Traub, R.J., Williams, G.M., Andrews, R.M., Clements, A.C.A., 2014. Water, sanitation, and hygiene (WASH): a critical component for sustainable soil-transmitted helminth and schistosomiasis control. PLoS Negl. Trop. Dis. 8, e2651.

Chammartin, F., Scholte, R.G.C., Malone, J.B., Bavia, M.E., Nieto, P., Utzinger, J., Vounatsou, P., 2013. Modelling the geographical distribution of soil-transmitted helminth infections in Bolivia. Parasites Vectors 6, 152.

Coffeng, L.E., Bakker, R., Montresor, A., de Vlas, S.J., 2015. Feasibility of controlling hookworm infection through preventive chemotherapy: a simulation study using the individual-based WORMSIM modelling framework. Parasites Vectors 8, 541.

de Silva, N.R., Brooker, S., Hotez, P.J., Montresor, A., Engels, D., Savioli, L., 2003. Soil-transmitted helminth infections: updating the global picture. Trends Parasitol. 19, 547–551.

Freeman, M.C., Clasen, T., Brooker, S., Akoko, D.O., Rheingans, R., 2013. The impact of a school-based hygiene, water quality and sanitation intervention on soil-transmitted helminth reinfection: a cluster-randomized trial. Am. J. Trop. Med. Hyg. 89, 875–883.

Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. Am. Stat. 63, 308–319.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Jia, T.-W., Melville, S., Utzinger, J., King, C.H., Zhou, X.-N., 2012. Soil-transmitted helminth reinfection after drug treatment: a systematic review and meta-analysis. PLoS Negl. Trop. Dis. 6, e1621.

Kaliappan, S.P., George, S., Francis, M.R., Kattula, D., Sarkar, R., Minz, S., Mohan, V.R., George, K., Roy, S., Ajjampur, S.S.R., Muliyil, J., Kang, G., 2013. Prevalence and clustering of soil-transmitted helminth infections in a tribal area in southern India. Trop. Med. Int. Health 18, 1452–1462.

Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., Arriaga-Weiss, S., 2010. Classification in conservation biology: a comparison of five machine-learning methods. Ecol. Inform. 5, 441–450.

Kattula, D., Sarkar, R., Ajjampur, S.S.R., Minz, S., Levecke, B., Muliyil, J., 2014. Prevalence and risk factors for soil transmitted helminth infection among school children in south India. Indian J. Med. Res. 139, 76–82.

King, C.H., 2015. It's time to dispel the myth of "asymptomatic" schistosomiasis. PLoS Negl. Trop. Dis. 9, e0003504.

Knopp, S., Mgeni, A.F., Khamis, I.S., Steinmann, P., Stothard, J.R., Rollinson, D., Marti, H., Utzinger, J., 2008. Diagnosis of soil-transmitted helminths in the era of preventive chemotherapy: effect of multiple stool sampling and use of different diagnostic techniques. PLoS Negl. Trop. Dis. 2, e331.

Kulinkina, A.V., Walz, Y., Koch, M., Biritwum, N.-K., Utzinger, J., Naumova, E.N., 2018. Improving spatial prediction of *Schistosoma haematobium* prevalence in southern Ghana through new remote sensors and local water access profiles. PLoS Negl. Trop. Dis. 12, e0006517.

Lai, Y.-S., Zhou, X.-N., Utzinger, J., Vounatsou, P., 2013. Bayesian geostatistical modelling of soil-transmitted helminth survey data in the People's Republic of China. Parasites Vectors 6, 359.

Lai, Y.-S., Biedermann, P., Shrestha, A., Chammartin, F., à Porta, N., Montresor, A., Mistry, A.F., Utzinger, J., Vounatsou, P., 2019. Risk profiling of soil-transmitted helminth infection and estimated number of infected people in South Asia: a

systematic review and Bayesian geostatistical analysis. PLoS Negl. Trop. Dis. 13, e0007580.
Li, J., Alvarez, B., Siwabessy, J., Tran, M., Huang, Z., Przeslawski, R., Radke, L., Howard, F., Nichol, S., 2017. Application of random forest and generalised linear model and their hybrid methods with geostatistical techniques to count data: predicting sponge species richness. Environ Model Softw 97, 112–129.
Loukas, A., Bethony, J., Brooker, S., Hotez, P., 2006. Hookworm vaccines: past, present, and future. Lancet Infect. Dis. 6, 733–741.
Mabaso, M.L.H., Appleton, C.C., Hughes, J.C., Gouws, E., 2003. The effect of soil type and climate on hookworm (*Necator americanus*) distribution in KwaZulu-Natal, South Africa. Trop. Med. Int. Health 8, 722–727.
Montresor, A., Crompton, D.W.T., Gyorkos, T.W., Savioli, L., 2002. Helminth Control in School-Age Children: A Guide for Managers of Control Programmes. World Health Organization, Geneva.
Morel, C.M., 2003. Neglected diseases: underfunded research and inadequate health interventions. EMBO Rep. 4, S35–S38.
Pullan, R.L., Bethony, J.M., Geiger, S.M., Cundill, B., Correa-Oliveira, R., Quinnell, R.J., Brooker, S., 2008. Human helminth co-infection: analysis of spatial patterns and risk factors in a Brazilian community. PLoS Negl. Trop. Dis. 2, e352.
Pullan, R.L., Kabatereine, N.B., Quinnell, R.J., Brooker, S., 2010. Spatial and genetic epidemiology of hookworm in a rural community in Uganda. PLoS Negl. Trop. Dis. 4, e713.
Pullan, R.L., Smith, J.L., Jasrasaria, R., Brooker, S.J., 2014. Global numbers of infection and disease burden of soil-transmitted helminth infections in 2010. Parasites Vectors 7, 37.
Raso, G., Vounatsou, P., Singer, B.H., N'Goran, E.K., Tanner, M., Utzinger, J., 2006. An integrated approach for risk profiling and spatial prediction of *Schistosoma mansoni*-hookworm coinfection. Proc. Natl. Acad. Sci. U.S.A. 103, 6934–6939.
Reiss, H., Clowes, P., Kroidl, I., Kowuor, D.O., Nsojo, A., Mangu, C., Schüle, S.A., Mansmann, U., Geldmacher, C., Mhina, S., Maboko, L., Hoelscher, M., Saathoff, E., 2013. Hookworm infection and environmental factors in Mbeya Region, Tanzania: a cross-sectional, population-based study. PLoS Negl. Trop. Dis. 7, e2408.
Saathoff, E., Olsen, A., Sharp, B., Kvalsvig, J.D., Appleton, C.C., Kleinscmidt, I., 2005. Ecologic covariates of hookworm infection and reinfection in rural KwaZulu-Natal, South Africa: a geographic information system-based study. Am. J. Trop. Med. Hyg. 72, 384–391.
Salgame, P., Yap, G.S., Gause, W.C., 2013. Effect of helminth-induced immunity on infections with microbial pathogens. Nat. Immunol. 14, 1118–1126.
Sarkar, R., Rose, A., Mohan, V.R., Ajjampur, S.S.R., Veluswamy, V., Srinivasan, R., Muliyil, J., Rajshekhar, V., George, K., Balraj, V., Grassly, N.C., Anderson, R.M., Brooker, Kang, G., 2017. Study design and baseline results of an open-label cluster randomized community-intervention trial to assess the effectiveness of a modified mass deworming program in reducing hookworm infection in a tribal population in southern India. Contemp. Clin. Trials Commun. 5, 49–55.
Scholte, R.G.C., Schur, N., Bavia, M.E., Carvalho, E.M., Chammartin, R., Utzinger, J., Vounatsou, P., 2013. Spatial analysis and risk mapping of soil-transmitted helminth infections in Brazil, using Bayesian geostatistical models. Geospat. Health 8, 97–110.
Soares Magalhães, R.J., Biritwum, N.-K., Gyapong, J.O., Brooker, S., Zhang, Y., Blair, L., Fenwick, A., Clements, A.C.A., 2011. Mapping helminth co-infection and co-intensity: geostatistical prediction in Ghana. PLoS Negl. Trop. Dis. 5, e1200.
Soares Magalhães, R.J., Salamat, M.S., Leonardo, L., Gray, D.J., Carabin, H., Halton, K., McManus, D.P., Williams, G.M., Rivera, P., Saniel, O., Hernandez, L., Yakob, L., McGarvey, S.T., Clements, A.C.A., 2015. Mapping the risk of soil-transmitted helminthic infections in the Philippines. PLoS Negl. Trop. Dis. 9, e0003915.
Strunz, E.C., Addiss, D.G., Stocks, M.E., Ogden, S., Utzinger, J., Freeman, M.C., 2014. Water, sanitation, hygiene, and soil-transmitted helminth infection: a systematic review and meta-analysis. PLoS Med. 11, e1001620.
Sunish, I.P., Rajendran, R., Munirathinam, A., Kalimuthu, M., Kumar, V.A., Nagaraj, J., Tyagi, B.K., 2015. Impact on prevalence of intestinal helminth infection in school children administered with seven annual rounds of diethyl carbamazine (DEC) with albendazole. Indian J. Med. Res. 141, 330–339.
Thigpen, M.C., Filler, S.J., Kazembe, P.N., Parise, M.E., Macheso, A., Campbell, C.H., Newman, R.D., Steketee, R.W., Hamel, M., 2011. Associations between peripheral *Plasmodium falciparum* malaria parasitemia, human immunodeficiency virus, and concurrent helminthic infection among pregnant women in Malawi. Am. J. Trop. Med. Hyg. 84, 379–385.
Utzinger, J., Müller, I., Vounatsou, P., Singer, B.H., N'Goran, E.K., Tanner, M., 2003. Random spatial distribution of *Schistosoma mansoni* and hookworm infections among school children within a single village. J. Parasitol. 89, 686–692.
Walz, Y., Wegmann, M., Leutner, B., Dech, S., Vounatsou, P., N'Goran, E.K., Raso, G., Utzinger, J., 2015. Use of an ecologically relevant modelling approach to improve remote sensing-based schistosomiasis risk profiling. Geospat. Health 10, 398.
Webb, E.L., Ekii, A.O., Pala, P., 2012. Epidemiology and immunology of helminth-HIV interactions. Curr. Opin. HIV AIDS 7, 245–253.
WHO, 2001. Schistosomiasis and soil-transmitted helminth infections. World Health Assembly Resolution 54, 19.
Ziegelbauer, K., Speich, B., Mäusezahl, D., Bos, R., Keiser, J., Utzinger, J., 2012. Effect of sanitation on soil-transmitted helminth infection: systematic review and meta-analysis. PLoS Med. 9, e1001162.

Please cite this article as: A. V. Kulinkina, R. Sarkar, V. R. Mohan et al., Prediction of hookworm prevalence in southern India using environmental parameters derived from Landsat 8 remotely sensed data, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.10.001