Supporting information: Identifying key aspects to enhance predictive modeling for early identification of schistosomiasis hotspots to guide mass drug administration

BY YEWEN CHEN [1], FANGZHI LUO [1], LEONARDO MARTINEZ[2], SUSAN JIANG[1], AND YE SHEN[1]

[1]*Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, Georgia, USA*

[2]*Department of Epidemiology, School of Public Health, Boston University, Boston, USA*

Corresponding author: Ye Shen, email: yeshen@uga.edu

Contents

List of Tables

List of Figures

S1. Schistosoma *mansoni* in Tanzania and Kenya

Schistosoma infection data for the 295 study villages in Kenya and Tanzania were obtained from the SCORE datasets. The SCORE project conducted a randomized clinical trial from 2011 to 2015, during which the 295 study villages were randomly assigned to six arms and received annual mass drug administration (MDA) with praziquantel (Figure S1). In Kenya, the infection prevalence ranged from 8.33% to 100%, with a mean of 61.51% and a median of 59.79%. In Tanzania, the minimum prevalence was 4.6%, the average prevalence 54.55%, and the median prevalence 55.32%, all of which were smaller than in Kenya, except for the maximum prevalence of Tanzania, which was the same as that of Kenya. However, the average intensity (49.54 epg (eggs per gram)) and median intensity (89.19 epg) in Kenya were lower than in Tanzania, which were 63 epg and 130.36 epg, respectively. Furthermore, the variation in infection intensity in Kenya was much smaller compared to Tanzania (Figure S2), with the intensity range in the former (3.68-454.86 epg) being much narrower than in the latter (0.96-1138.16 epg). Generally, larger variations can make hotspots more unpredictable, resulting in lower accuracy in predictions (Table S3).

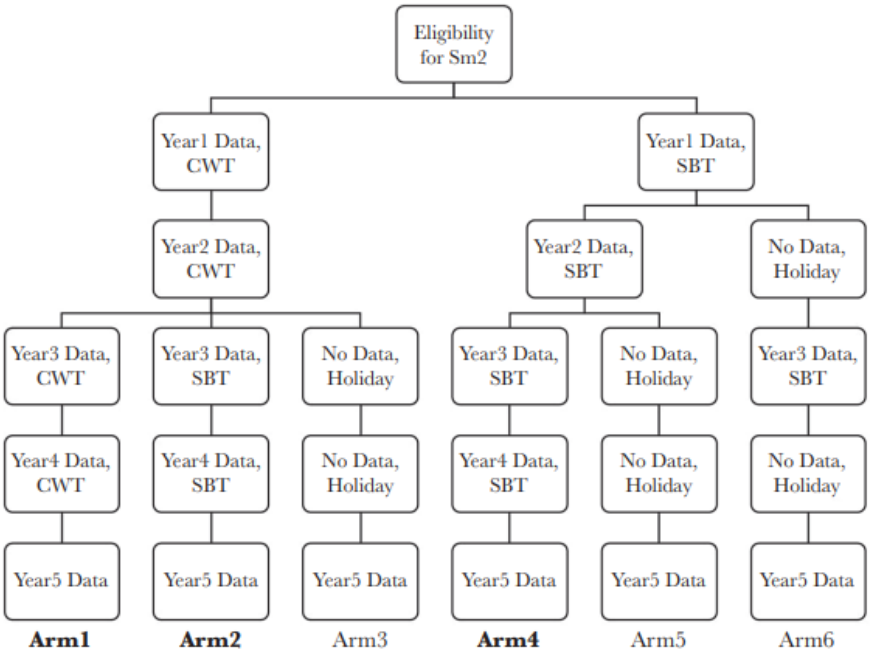

Fig S1: Study arms and timeline for the studies of gaining control of schistosomiasis from SCORE (Shen et al., 2020).

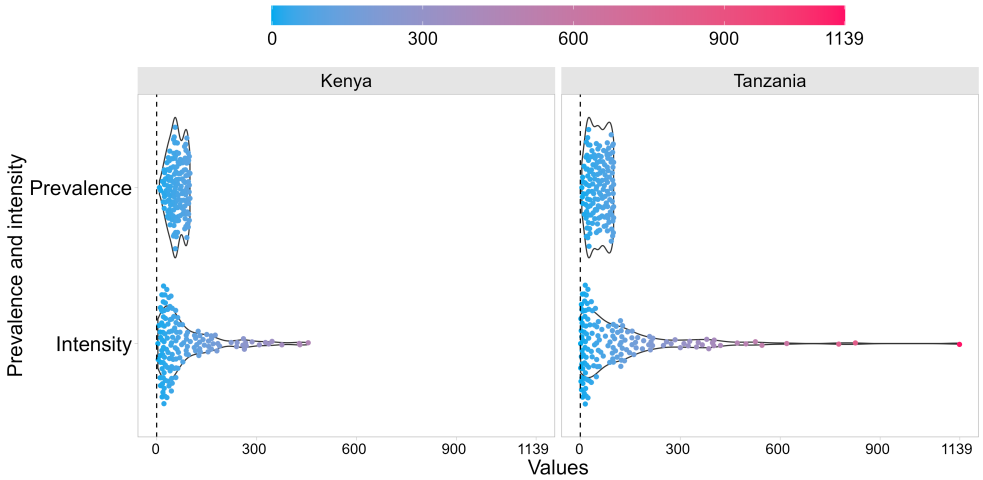

Fig S2: Distribution of infection prevalence and intensity in Kenya and Tanzania.

S2. Persistent hotspots

Table S1 Based on two persistent hotspot (PHS) definitions, the proportion of hotspots (1) and non-hotspots (0) in Kenya and Tanzania.

PHS	PHS definition I			PHS definition II		
	Kenya	Tanzania	Combined-countries	Kenya	Tanzania	Combined-countries
0	95 (65%)	42 (28%)	137 (46%)	106 (72%)	56 (38%)	162 (55%)
1	52 (35%)	106 (72%)	99 (50%)	41 (28%)	92 (62%)	133 (45%)
Total	147	148	198	147	148	295

S3. Categorizing datasets and spatially weighted data fusion

For categorizing datasets, prevalence and intensity were used as baseline disease inputs in prediction models to predict the binary outcome of the PHS status for S. *mansoni*, referred to as the baseline infection data. Furthermore, another predictor, namely the prevalence of infections $\geq$ 200 epg (denoted by prevalence.200), was also incorporated into the baseline infection data (Shen et al., 2020). In addition, 41 other predictors were employed to improve the accuracy of the models in hotspot predictions. To ensure the interpretability of the hotspot prediction results and gain deeper insight into the factors that potentially drive hotspot formation, a knowledge-driven category formation method was used to classify the datasets. This approach resulted in a categorization of predictors into six distinct categories: *infection data around villages*, *environment*, *agriculture*, *geography*, *biology*, and *society*.

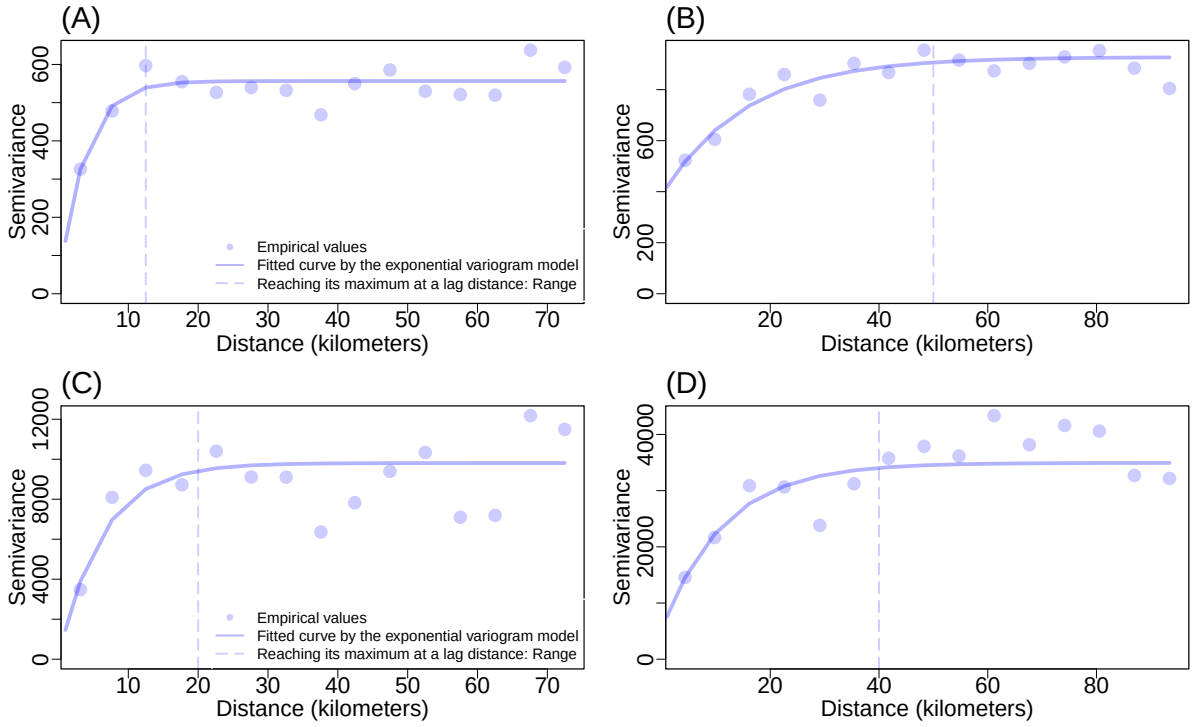

Fig S3: Empirical variograms and fitted curves by exponential variogram models. (A) Prevalence in Kenya, (B) Infection prevalence in Tanzania, (C) Infection intensity in Kenya, and (D) Intensity in Tanzania.

In spatially weighted data fusion for predictor construction, this study not only integrated spatial information from neighboring areas within a specified distance (threshold) but also simultaneously filtered out redundant noise beyond that threshold. This can be achieved using the spTIDW method (Chen et al., 2023, 2024), and the threshold can be determined by examining a spatial correlation range based on empirical variograms (Figure S3). Using a 20 km threshold in spTIDW, three predictors (prevalence, intensity, and prevalence.200) of the baseline infection data were spatially weighted to build predictors of *infection data around villages*. To capture similar patterns of S. *mansoni* in different villages, the K-Means clustering method with K=2 was used for prevalence and intensity in the baseline year, resulting in a binary variable. Then, this variable was assigned to the category of *infection data around villages* as a predictor. Using a threshold of 50 km, spTIDW was applied to construct other spatially weighted predictors from the remaining categories. In contrast to the annual scale of the SCORE datasets, the weighted predictors from other categories were typically on a smaller scale, such as monthly. To align these with infection data on a yearly scale, the weighted predictors were calculated as mean, maximum, minimum, or sum over time. See Table S2 of **SI** for more details of these predictors.

Table S2 List of predictors used in this study for developing prediction models for early identification of schistosomiasis hotspots. These predictors were collected from SCORE (Schistosomiasis Consortium for the Operational Research and Evaluation) (Dan et al., 2022), ERA5 (European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5) (ECMWF, 2024), GFAS (Global Fire Assimilation System) (GFAS, 2024), CGLS (Copernicus Global Land Service) (CGLS, 2020), GLAD (Global Land Analysis & Discovery) (Potapov et al., 2022; GLAD, 2022), and SDAC (Socioeconomic Data and Applications Center) (SDAC, 2022).

Categories	**Predictors**	**Label**	**Included** [1]	**Form**	**Resolution**	**Date**	**Source**
Baseline infection data	Prevalence	Prevalence	Yes	Points	-	2011	SCORE
	Intensity	Intensity	Yes	Points	-	2011	SCORE
	Prevalence of infections with epg $\geq$ 200	Prevalence.200	Yes	Points	-	2011	SCORE
Infection data around villages	Spatially weighted prevalence	sPrevalence	No	Points	-	2011	SCORE
	Spatially weighted Prevalence of infections with epg $\geq$ 200	sPrevalence.200	No	Points	-	2011	SCORE
	Spatially weighted intensity	sIntensity	No	Points	-	2011	SCORE
	Clustering variable[2]	Cluster.Pre.Int	No	Points	-	2011	SCORE
Geography	Longitude	Longitude	No	Points	-	-	SCORE
	Latitude	Latitude	No	Points	-	-	SCORE
	Elevation	Elevation	Yes	Grids	20km	-	ERA5
Environment	Minimum of 2m dew point temperature	min.2m.Dew.Temp	Yes	Grids	10km	2011	ERA5
	Maximum of 2m dew point temperature	max.2m.Dew.Temp	No	Grids	10km	2011	ERA5
	Minimum precipitation	min.Precipitation	No	Grids	10km	2011	ERA5
	Maximum precipitation	max.Precipitation	Yes	Grids	10km	2011	ERA5
	Minimum surface pressure	min.S.Pressure	No	Grids	10km	2011	ERA5
	Maximum surface pressure	max.S.Pressure	No	Grids	10km	2011	ERA5
	Fine particulate matters[3]	PM25	No	Grids	10km	2011	GFAS
	Maximum temperature of the soil in layer 1	max.Temp.Soil	No	Grids	27km	2010	ERA5
	Maximum horizontal irradiation	max.H.Irradiation	No	Grids	10km	2011	ERA5
	Maximum normal irradiation	max.N.Irradiation	No	Grids	10km	2011	ERA5
	Maximum diffuse irradiation	max.D.Irradiation	No	Grids	10km	2011	ERA5
	Cumulative global horizontal irradiation	sum.G.Irradiation	No	Grids	10km	2011	ERA5
	Maximum surface shortwave irradiation	max.S.Irradiation	No	Grids	10km	2011	ERA5
	Maximum evaporation from the top of canopy	max.Evaporation	No	Grids	10km	2011	ERA5
Agriculture	Permanent water cover fraction	Permanent.Water	Yes	Grids	0.1km	2015	CGLS
	Seasonal water cover fraction	Seasonal.Water	Yes	Grids	0.1km	2015	CGLS
	Cropland	Cropland	No	Grids	3km	2011	GLAD
	Maximum of soil moisture one-month return period	max.Period.Moisture	No	Grids	10km	2011	SDAC
	Minimum of soil moisture one-month return period	min.Period.Moisture	No	Grids	10km	2011	SDAC
	Maximum scientific units of soil moisture	max.Units.Moisture	No	Grids	10km	2011	SDAC
	Minimum scientific units of soil moisture	min.Units.Moisture	No	Grids	10km	2011	SDAC
	Maximum anomalies of soil moisture	max.Anomalies	No	Grids	10km	2011	SDAC
	Minimum anomalies of soil moisture	min.Anomalies	No	Grids	10km	2011	SDAC
Biology	Forest cover fraction	Forest.Cover	No	Grids	0.1km	2015	CGLS
	The number of mammals	Mammals	No	Grids	1km	2020	SDAC
Society	Population density	Population	Yes	Grids	1km	2010	SDAC
	Human modification of terrestrial lands	mod.Lands	No	Grids	1km	2010	SDAC
	Deprivation index	Deprivation.Index	No	Grids	5.5km	2010-2020	SDAC
	Child dependency ratio	Child.Ratio	No	Grids	5.5km	2010-2020	SDAC
	Building cover fraction	Building.Fraction	No	Grids	5.5km	2015	CGLS
	Building component	Building.Component	No	Grids	5.5km	2010-2020	SDAC
	Nighttime lights	Lights	No	Grids	5.5km	2010-2020	SDAC

[1] Whether one predictor has been considered to predict hotspots in prior studies;
[2] This predictor is binary, generated via the K-means method with K = 2;
[3] Fine particulate matters with aerodynamic diameters less than 2.5 micrometers.

S4. Comparisons

Table S3 Detailed accuracy of each models for predictor configurations C1-C8 under the setting of PHS definition I. Italics are used to indicate the results from the previous six models (GBM, RF, Tree, Logit, LASSO, and LGT) using the baseline infection data only (Shen et al., 2020). The highest accuracy for predicting hotspots in each scenario and each category is highlighted in bold, while the lowest accuracy in the gray color.

Predictor configurations	Scenario	GBM	RF	Tree	Logit	LASSO	LGT	SVM	Ensemble	LogitGPs	sparseSVM	dynaTrees	regLogit	Probit	DNN
Baseline infection data (C1)	Kenya	*0.687*	*0.634*	*0.677*	*0.594*	*0.587*	*0.585*	0.749	0.633	0.745	0.749	0.753	**0.754**	0.690	0.681
Infection around (C2)	Kenya	0.708	0.563	0.662	0.642	0.636	0.659	**0.756**	0.633	0.754	**0.756**	0.754	0.749	0.722	0.634
Environment (C3)	Kenya	0.760	0.712	0.717	0.683	0.692	0.699	0.732	0.715	**0.755**	0.743	0.739	0.744	0.709	0.724
Agriculture (C4)	Kenya	0.744	0.672	0.683	0.676	0.673	0.652	0.727	0.644	0.741	0.738	**0.749**	0.737	0.678	0.723
Geography (C5)	Kenya	0.719	0.660	0.696	0.681	0.685	0.664	**0.741**	0.667	0.730	0.734	0.742	0.745	0.692	0.702
Biology (C6)	Kenya	0.740	0.687	0.688	0.683	0.677	0.672	0.738	0.687	0.737	0.734	**0.745**	0.741	0.721	0.696
Society (C7)	Kenya	0.756	0.686	0.678	0.722	0.712	0.703	0.725	0.694	**0.742**	0.734	0.740	0.740	0.703	0.722
All (C8)	Kenya	0.754	0.709	0.678	0.713	0.718	0.684	0.716	0.700	0.734	0.699	**0.742**	0.729	0.739	0.688
Baseline infection data (C1)	Tanzania	*0.624*	*0.701*	*0.595*	*0.699*	*0.699*	*0.701*	**0.711**	0.697	0.688	**0.711**	**0.711**	0.711	0.664	0.693
Infection around (C2)	Tanzania	0.598	0.704	0.593	0.682	0.679	0.671	**0.711**	0.701	0.708	0.709	**0.711**	0.703	0.683	0.690
Environment (C3)	Tanzania	0.686	0.704	0.646	0.732	0.732	0.684	0.726	0.699	0.711	0.728	**0.737**	0.732	0.708	0.736
Agriculture (C4)	Tanzania	0.685	0.711	0.637	0.722	0.718	0.700	0.714	0.709	0.717	0.716	**0.740**	0.717	0.729	0.724
Geography (C5)	Tanzania	0.673	0.704	0.648	0.720	**0.727**	0.714	0.703	0.702	0.697	0.716	0.724	0.709	0.680	0.715
Biology (C6)	Tanzania	0.611	0.705	0.593	0.691	0.691	0.688	0.707	0.700	0.703	0.705	**0.709**	0.708	0.690	0.701
Society (C7)	Tanzania	0.657	0.705	0.640	0.699	0.693	0.690	0.709	0.696	0.710	0.696	0.711	**0.715**	0.674	0.708
All (C8)	Tanzania	0.650	0.729	0.653	**0.744**	0.737	0.601	0.664	0.724	0.711	0.706	0.739	0.732	0.727	**0.744**
Baseline infection data (C1)	Combined	*0.605*	*0.551*	*0.588*	*0.567*	*0.570*	*0.571*	0.582	0.560	**0.608**	0.584	0.585	0.566	0.587	0.531
Infection around (C2)	Combined	0.656	0.619	0.583	0.676	0.676	0.676	0.696	0.670	**0.681**	0.665	0.610	0.665	0.618	0.655
Environment (C3)	Combined	0.725	0.702	0.675	0.720	0.716	0.687	0.719	0.697	0.734	0.717	0.717	**0.736**	0.716	0.718
Agriculture (C4)	Combined	0.704	0.691	0.662	0.705	0.703	0.682	**0.715**	0.689	0.712	0.704	0.713	0.706	0.700	0.699
Geography (C5)	Combined	0.694	0.671	0.646	0.703	0.705	0.703	**0.729**	0.684	0.722	0.724	0.720	0.726	0.700	0.724
Biology (C6)	Combined	0.664	0.673	0.628	0.707	0.708	0.707	0.722	0.706	0.722	0.713	0.711	**0.725**	0.706	0.715
Society (C7)	Combined	0.720	0.691	0.645	0.707	0.708	0.706	**0.720**	0.717	0.703	0.707	0.707	0.714	0.680	0.708
All (C8)	Combined	0.725	0.717	0.658	0.713	0.716	0.694	0.699	0.715	0.723	0.716	0.720	**0.727**	0.718	0.714
Baseline infection data (C1)	Between I	0.511	0.532	0.492	0.590	**0.604**	0.605	0.442	0.567	0.435	0.368	0.404	0.469	0.478	0.522
Infection around (C2)	Between I	0.517	0.604	0.553	0.624	**0.643**	0.576	0.399	0.572	0.423	0.399	0.400	0.433	0.447	0.586
Environment (C3)	Between I	0.341	**0.541**	0.460	0.296	0.297	0.279	0.281	0.479	0.280	0.311	0.370	0.286	0.344	0.300
Agriculture (C4)	Between I	0.360	**0.517**	0.440	0.292	0.291	0.314	0.340	0.435	0.289	0.342	0.387	0.355	0.445	0.346
Geography (C5)	Between I	0.512	0.652	0.531	0.703	0.703	0.705	0.637	0.681	0.612	0.523	0.428	0.682	0.526	**0.712**
Biology (C6)	Between I	0.511	0.613	0.507	0.708	0.708	**0.709**	0.529	0.649	0.504	0.444	0.422	0.655	0.483	0.671
Society (C7)	Between I	0.386	0.560	0.498	0.670	0.654	0.685	0.645	0.636	0.556	0.550	0.348	0.677	0.367	**0.693**
All (C8)	Between I	0.316	**0.578**	0.471	0.352	0.321	0.352	0.335	0.493	0.283	0.378	0.376	0.348	0.432	0.529
Baseline infection data (C1)	Between II	**0.505**	0.374	0.494	0.368	0.368	0.368	0.370	0.361	0.388	0.370	0.370	0.365	0.466	0.367
Infection around (C2)	Between II	**0.470**	0.371	0.426	0.373	0.373	0.379	0.370	0.382	0.370	0.370	0.370	0.377	0.390	0.373
Environment (C3)	Between II	**0.448**	0.376	0.455	0.375	0.374	0.385	0.375	0.379	0.370	0.411	0.375	0.374	0.420	0.387
Agriculture (C4)	Between II	0.441	0.386	**0.459**	0.388	0.392	0.387	0.421	0.383	0.370	0.371	0.378	0.388	0.406	0.377
Geography (C5)	Between II	0.423	0.371	0.465	0.482	0.481	0.479	0.382	0.420	0.424	0.381	0.370	0.483	0.449	**0.509**
Biology (C6)	Between II	0.469	0.377	0.439	0.484	0.484	0.481	0.369	0.382	0.378	0.370	0.370	0.456	0.408	**0.526**
Society (C7)	Between II	0.404	0.370	0.445	0.395	0.413	**0.474**	0.371	0.418	0.370	0.370	0.370	0.468	0.419	0.370
All (C8)	Between II	0.420	0.376	0.483	0.395	0.410	0.551	0.557	0.380	0.370	0.413	0.376	**0.582**	0.382	0.376

[1] Only using the baseline infection data.

Table S4 Relative improvements (IRs, %) for each scenario obtained using the proposed spatially weighted data fusion methods with different predictor configurations (C2-C8), compared to the method with configuration C1. IRs were obtained from the worst of the 14 models for each method, based on the lowest prediction accuracy for each scenario and predictor configuration.

Combining baseline infection data	Within-country		Combined-countries	Between-countries		Average [1]
with additional predictors	Kenya	Tanzania	Kenya and Tanzania	*Between I*	*Between II*	
Infection around	-3.76	-0.34	9.98	8.42	2.49	3.36
Environment	16.75	8.57	20.34	-24.18	2.49	4.79
Agriculture	10.09	7.06	24.86	-21.47	2.49	4.61
Geography	12.82	8.91	21.85	16.30	2.49	12.47
Biology	14.87	-0.34	4.52	14.67	2.22	7.19
Society	15.90	7.56	21.28	-5.43	2.49	8.36
All	15.90	1.01	24.29	-23.10	2.49	4.12

[1] Average across CV scenarios from columns 2 to 6.

Table S5 Hotspot prediction accuracy on test sets obtained using models developed using synthetic sampling training sets, compared to those developed using imbalanced training sets.

Cross-validation	Imbalanced	Additional	**Accuracy**													
scenario	training sets [1]	predictors	GBM	RF	Tree	Logit	LASSO	LGT	SVM	Ensemble	LogitGPs	sparseSVM	dynaTrees	reglogit	Probit	CNN
Between I	Before	Disease around	0.517	0.604	0.553	0.624	0.643	0.576	0.399	0.572	0.423	0.399	0.400	0.433	0.447	0.586
	After	Disease around	0.353	0.716	0.634	0.722	0.722	0.722	0.587	0.574	0.722	0.588	0.399	0.549	0.664	0.722
	Before	Geography	0.512	0.652	0.531	0.703	0.703	0.705	0.637	0.681	0.612	0.523	0.428	0.682	0.526	0.712
	After	Geography	0.722	0.722	0.721	0.722	0.722	0.722	0.599	0.718	0.722	0.579	0.722	0.542	0.722	0.722
	Before	Environment	0.341	0.541	0.460	0.296	0.297	0.279	0.281	0.479	0.280	0.311	0.370	0.286	0.344	0.300
	After	Environment	0.724	0.723	0.722	0.722	0.722	0.722	0.712	0.720	0.722	0.657	0.722	0.615	0.722	0.722
	Before	Agriculture	0.360	0.517	0.440	0.292	0.291	0.314	0.340	0.435	0.289	0.342	0.387	0.355	0.445	0.346
	After	Agriculture	0.659	0.707	0.689	0.715	0.715	0.715	0.683	0.712	0.722	0.636	0.721	0.570	0.722	0.712
	Before	Biology	0.511	0.613	0.507	0.708	0.708	0.709	0.529	0.649	0.504	0.444	0.422	0.655	0.483	0.671
	After	Biology	0.720	0.720	0.720	0.719	0.719	0.719	0.605	0.722	0.716	0.570	0.720	0.547	0.721	0.718
	Before	Society	0.386	0.560	0.498	0.670	0.654	0.685	0.645	0.636	0.556	0.550	0.348	0.677	0.367	0.693
	After	Society	0.707	0.717	0.712	0.704	0.704	0.704	0.675	0.721	0.691	0.614	0.720	0.568	0.721	0.705
	Before	All	0.316	0.578	0.471	0.352	0.321	0.352	0.335	0.493	0.283	0.378	0.376	0.348	0.432	0.529
	After	All	0.426	0.718	0.699	0.711	0.711	0.712	0.713	0.605	0.701	0.666	0.677	0.700	0.722	0.719
Between II	Before	Disease around	0.470	0.371	0.426	0.373	0.373	0.379	0.370	0.382	0.370	0.370	0.370	0.377	0.390	0.373
	After	Disease around	0.630	0.630	0.630	0.528	0.528	0.531	0.579	0.636	0.630	0.560	0.630	0.542	0.630	0.558
	Before	Geography	0.423	0.371	0.465	0.482	0.481	0.479	0.382	0.420	0.424	0.381	0.370	0.483	0.449	0.509
	After	Geography	0.628	0.625	0.603	0.476	0.476	0.477	0.549	0.627	0.618	0.537	0.621	0.530	0.616	0.509
	Before	Environment	0.448	0.376	0.455	0.375	0.374	0.385	0.375	0.379	0.370	0.411	0.375	0.374	0.420	0.387
	After	Environment	0.630	0.629	0.619	0.586	0.586	0.586	0.623	0.630	0.630	0.604	0.629	0.576	0.625	0.601
	Before	Agriculture	0.441	0.386	0.459	0.388	0.392	0.387	0.421	0.383	0.370	0.371	0.378	0.388	0.406	0.377
	After	Agriculture	0.630	0.630	0.624	0.552	0.552	0.546	0.606	0.628	0.630	0.592	0.628	0.548	0.628	0.590
	Before	Biology	0.469	0.377	0.439	0.484	0.484	0.481	0.369	0.382	0.378	0.370	0.370	0.456	0.408	0.526
	After	Biology	0.630	0.630	0.630	0.489	0.489	0.489	0.552	0.628	0.630	0.544	0.630	0.532	0.630	0.524
	Before	Society	0.404	0.370	0.445	0.395	0.413	0.474	0.371	0.418	0.370	0.370	0.370	0.468	0.419	0.370
	After	Society	0.630	0.629	0.626	0.572	0.572	0.572	0.602	0.635	0.630	0.576	0.630	0.545	0.626	0.589
	Before	All	0.420	0.376	0.483	0.395	0.410	0.551	0.557	0.380	0.370	0.413	0.376	0.582	0.382	0.376
	After	All	0.632	0.630	0.627	0.625	0.626	0.624	0.628	0.630	0.627	0.628	0.630	0.626	0.629	0.624

[1] Before: unprocessed training sets; After: pre-processed training sets.

Table S6 Relative improvements in hotspot prediction accuracy on test sets obtained using models developed on synthetic sampling training sets, compared to those developed on imbalanced training sets.

Persistent hotspots	Scenario	Additional predictors[1]	**Relative improvements (%)**													
			GBM	RF	Tree	Logit	LASSO	LGT	SVM	Ensemble	LogitGPs	sparseSVM	dynaTrees	reglogit	Probit	CNN
Definition I	Between I	Disease around	-31.72	18.54	14.65	15.71	12.29	25.35	47.12	0.35	70.69	47.37	-0.25	26.79	48.55	23.21
		Environment	112.32	33.64	56.96	143.92	143.10	158.78	153.38	50.31	157.86	111.25	95.14	115.03	109.88	140.67
		Agriculture	83.06	36.75	56.59	144.86	145.70	127.71	100.88	63.68	149.83	85.96	86.30	60.56	62.25	105.78
		Geography	41.02	10.74	35.78	2.70	2.70	2.41	-5.97	5.43	17.97	10.71	68.69	-20.53	37.26	1.40
		Biology	40.90	17.46	42.01	1.55	1.55	1.41	14.37	11.25	42.06	28.38	70.62	-16.49	49.28	7.00
		Society	83.16	28.04	42.97	5.07	7.65	2.77	4.65	13.36	24.28	11.64	106.90	-16.10	96.46	1.73
		All	34.81	24.22	48.41	101.99	121.50	102.27	112.84	22.72	147.70	76.19	80.05	101.15	67.13	35.92
	Between II	Disease around	34.04	69.81	47.89	41.55	41.55	40.11	56.49	66.49	70.27	51.35	70.27	43.77	61.54	49.60
		Environment	40.63	67.29	36.04	56.27	56.68	52.21	66.13	66.23	70.27	46.96	67.73	54.01	48.81	55.30
		Agriculture	42.86	63.21	35.95	42.27	40.82	41.09	43.94	63.97	70.27	59.57	66.14	41.24	54.68	56.50
		Geography	48.46	68.46	29.68	-1.24	-1.04	-0.42	43.72	49.29	45.75	40.94	67.84	9.73	37.19	0.00
		Biology	34.33	67.11	43.51	1.03	1.03	1.66	49.59	64.40	66.67	47.03	70.27	16.67	54.41	-0.38
		Society	55.94	70.00	40.67	44.81	38.50	20.68	62.26	51.91	70.27	55.68	70.27	16.45	49.40	59.19
		All	50.48	67.55	29.81	58.23	52.68	13.25	12.75	65.79	69.46	52.06	67.55	7.56	64.66	65.96

[1] Combining baseline infection data with additional predictors from different categories.

S5. Importance of predictor categories and predictors

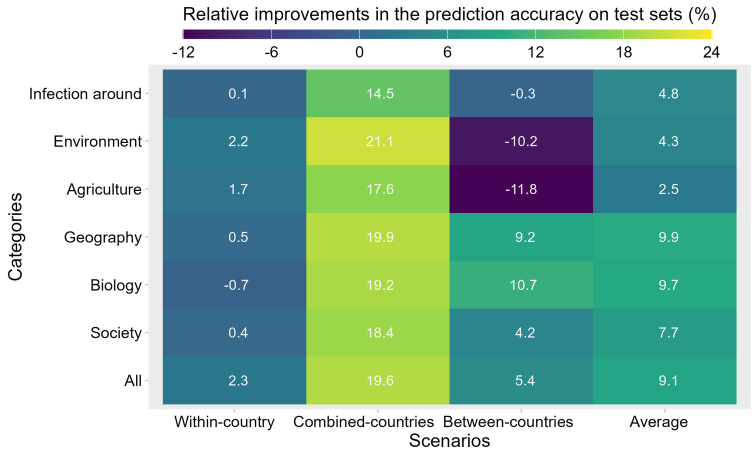

Fig S4: Relative improvements (RIs) in prediction accuracy on test sets for the three scenarios from the proposed spatially weighted data fusion method using different predictor categories, compared to the approach using only baseline infection data.

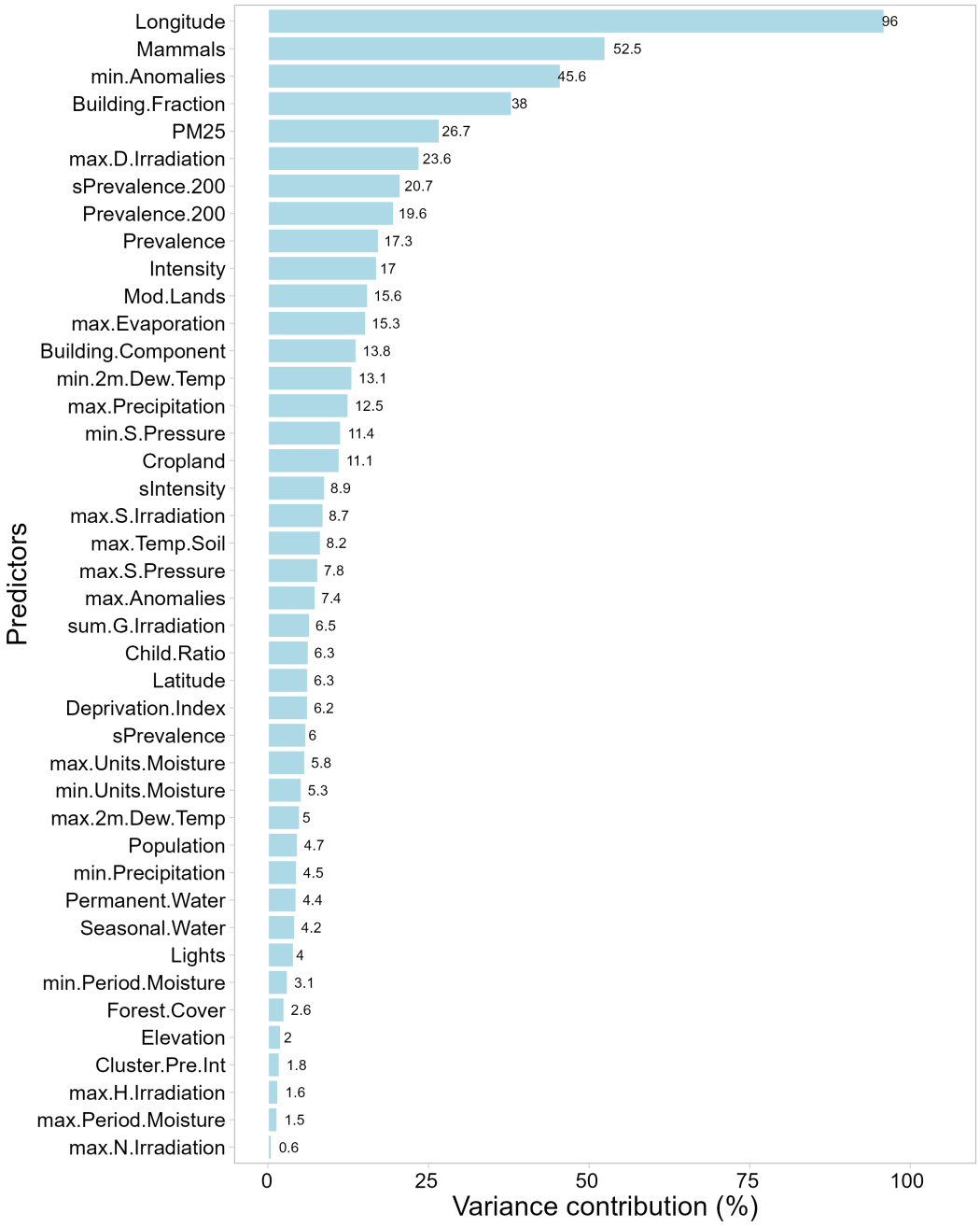

Fig S5: The importance of predictors was assessed by examining the relative effects of each predictor, with the effect quantified as the percentage of variance contribution based on multiple Gradient Boosting Machine (GBM) with different hyperparameters. The assessment was repeated 50 times using different training sets, and ten GBM models were run for each evaluation.

S6. Scalability of the proposed method

Table S7 Detailed accuracy of each models for predictor configurations C1-C8 under the setting of PHS definition II. The highest accuracy for predicting hotspots in each scenario and each category is highlighted in bold, while the lowest accuracy in the gray color.

Categories	Scenario	GBM	RF	Tree	Logit	LASSO	LGT	SVM	Ensemble	LogitGPs	sparseSVM	dynaTrees	regLogit	Probit	DNN
Baseline infection data (C1)	Kenya	0.789	0.791	0.760	0.779	0.774	0.776	0.816	0.772	0.817	0.816	0.820	**0.825**	0.796	0.757
Infection around (C2)	Kenya	0.796	0.771	0.741	0.723	0.723	0.748	**0.823**	0.757	0.817	0.821	0.821	0.820	0.805	0.736
Environment (C3)	Kenya	0.807	0.794	0.772	0.787	0.781	0.737	0.793	0.795	0.805	0.801	0.800	0.798	0.779	**0.810**
Agriculture (C4)	Kenya	0.798	0.787	0.768	0.783	0.781	0.754	0.786	0.760	0.795	0.806	**0.819**	0.808	0.769	0.788
Geography (C5)	Kenya	0.805	0.786	0.777	0.773	0.765	0.744	0.802	0.753	0.803	0.800	0.812	**0.821**	0.792	0.797
Biology (C6)	Kenya	0.804	0.770	0.772	0.749	0.752	0.753	0.796	0.767	0.804	0.798	0.813	**0.814**	0.778	0.768
Society (C7)	Kenya	**0.823**	0.776	0.768	0.772	0.767	0.748	0.792	0.768	0.807	0.800	0.815	0.808	0.783	0.773
All (C8)	Kenya	0.796	0.797	0.758	0.784	0.777	0.720	0.760	0.791	0.738	0.788	0.814	**0.815**	0.804	0.767
Baseline infection data (C1)	Tanzania	0.565	0.614	0.548	**0.620**	**0.620**	0.619	0.617	0.609	0.582	0.582	0.604	0.592	0.589	0.613
Infection around (C2)	Tanzania	0.564	0.611	0.560	0.598	0.596	0.591	0.596	0.603	**0.621**	0.583	0.616	0.602	0.588	0.608
Environment (C3)	Tanzania	0.661	0.636	0.632	0.674	0.678	0.680	0.676	0.647	**0.681**	0.672	0.668	0.671	0.663	0.672
Agriculture (C4)	Tanzania	0.645	0.688	0.613	0.663	0.662	0.669	0.655	0.682	**0.696**	0.677	0.678	0.657	0.668	0.694
Geography (C5)	Tanzania	0.644	0.622	0.621	0.650	0.649	0.646	0.661	0.620	0.661	**0.672**	0.645	0.660	0.628	0.654
Biology (C6)	Tanzania	0.593	0.617	0.572	0.603	0.604	0.604	0.600	0.608	**0.632**	0.586	0.615	0.588	0.586	0.610
Society (C7)	Tanzania	0.608	0.632	0.580	**0.648**	0.652	0.647	0.627	0.641	0.644	0.605	0.629	0.644	**0.648**	0.638
All (C8)	Tanzania	0.628	0.702	0.624	**0.704**	0.698	0.593	0.655	0.674	0.650	0.688	0.694	0.694	0.673	0.699
Baseline infection data (C1)	Combined	0.648	0.548	0.594	0.578	0.593	0.594	0.666	0.592	**0.685**	0.651	0.657	0.656	0.651	0.560
Infection around (C2)	Combined	0.669	0.607	0.604	0.642	0.644	0.644	0.659	0.639	0.672	0.664	0.655	**0.678**	0.656	0.611
Environment (C3)	Combined	**0.746**	0.728	0.699	0.723	0.721	0.715	0.731	0.739	0.744	0.724	0.719	0.745	0.737	0.715
Agriculture (C4)	Combined	0.732	**0.739**	0.673	0.696	0.698	0.698	0.703	0.733	0.725	0.686	0.713	0.697	0.707	0.666
Geography (C5)	Combined	0.721	0.711	0.676	0.698	0.694	0.696	0.728	0.711	**0.734**	0.697	0.706	0.691	0.713	0.695
Biology (C6)	Combined	0.689	0.669	0.651	0.693	0.694	**0.697**	0.691	0.680	0.696	0.686	0.682	0.686	0.684	0.689
Society (C7)	Combined	0.716	0.694	0.660	0.709	0.713	**0.717**	0.704	0.704	0.700	0.684	0.683	0.691	0.693	0.682
All (C8)	Combined	0.744	0.750	0.696	0.742	0.736	0.715	0.731	0.752	0.764	0.729	0.726	**0.765**	0.727	0.730
Baseline infection data (C1)	Between I	0.528	0.522	0.514	0.584	**0.588**	**0.588**	0.516	0.557	0.500	0.454	0.489	0.534	0.510	0.576
Infection around (C2)	Between I	0.436	0.616	0.590	**0.629**	**0.629**	**0.629**	0.524	0.571	**0.629**	0.559	0.492	0.524	0.618	**0.629**
Environment (C3)	Between I	**0.630**	**0.630**	**0.630**	0.629	0.629	0.629	0.624	0.629	0.629	0.589	0.629	0.571	0.629	0.629
Agriculture (C4)	Between I	0.569	0.610	0.604	0.626	0.626	0.625	0.593	0.606	**0.629**	0.567	0.625	0.538	**0.629**	0.622
Geography (C5)	Between I	0.620	**0.629**	**0.629**	**0.629**	**0.629**	**0.629**	0.547	0.624	**0.629**	0.531	**0.629**	0.515	**0.629**	**0.629**
Biology (C6)	Between I	0.628	0.628	0.628	0.626	0.626	0.626	0.536	**0.631**	0.624	0.529	0.627	0.525	0.628	0.626
Society (C7)	Between I	0.618	0.628	**0.629**	0.624	0.624	0.624	0.570	0.626	0.623	0.555	0.627	0.533	**0.629**	0.620
All (C8)	Between I	0.500	**0.630**	0.620	0.624	0.624	0.622	0.625	0.573	0.620	0.582	0.604	0.612	0.629	0.628
Baseline infection data (C1)	Between II	**0.566**	0.309	0.526	0.320	0.326	0.320	0.326	0.310	0.522	0.298	0.370	0.329	0.497	0.300
Infection around (C2)	Between II	**0.702**	**0.702**	**0.702**	0.549	0.549	0.561	0.640	0.697	**0.702**	0.611	**0.702**	0.570	**0.702**	0.616
Environment (C3)	Between II	**0.702**	0.701	0.684	0.631	0.631	0.631	0.687	**0.702**	**0.702**	0.667	0.694	0.628	0.698	0.685
Agriculture (C4)	Between II	**0.702**	**0.702**	0.693	0.568	0.568	0.568	0.666	**0.702**	**0.702**	0.633	0.700	0.575	0.699	0.646
Geography (C5)	Between II	**0.701**	0.700	0.652	0.486	0.486	0.486	0.609	0.699	0.689	0.578	0.687	0.555	0.684	0.516
Biology (C6)	Between II	0.702	0.702	**0.703**	0.494	0.494	0.494	0.612	**0.703**	0.702	0.592	0.702	0.556	0.702	0.555
Society (C7)	Between II	**0.702**	**0.702**	**0.702**	0.573	0.573	0.567	0.657	0.697	**0.702**	0.619	**0.702**	0.567	**0.702**	0.631
All (C8)	Between II	**0.702**	**0.702**	0.697	0.699	0.699	0.699	0.699	**0.702**	0.700	0.700	**0.702**	0.698	0.700	0.697

S7. Highlighted models

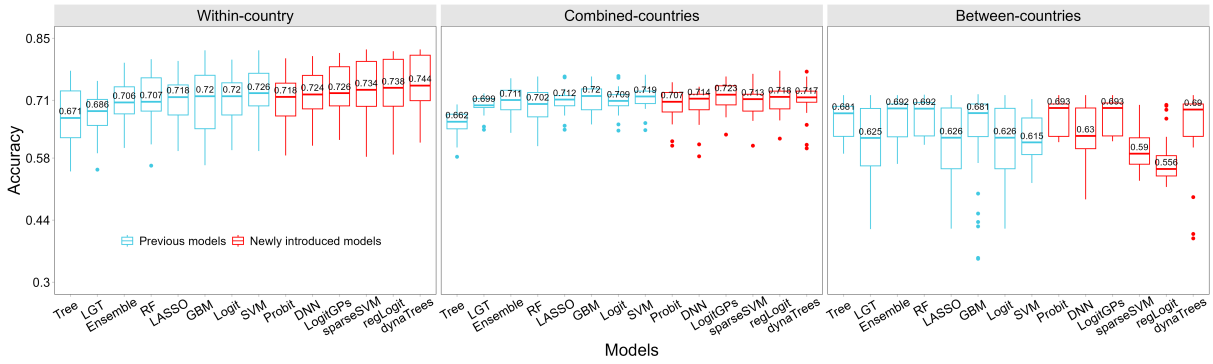

Fig S6: Overall performance of each model in each of the three scenarios for predicting PHSs for S. *mansoni*, where the horizontal lines in the boxplots represent the median of accuracy.

S8. Spatially weighted data fusion methods vs. previous non-baseline methods

Table S8 Detailed accuracy of each models for predictor configurations under the setting of PHS definition I, where the previous non-baseline method used the infection data in years 1 and 3 (Shen et al., 2020), while the proposed data fusion method used the infection data only from year 1. The highest accuracy for predicting hotspots in each scenario and each category is highlighted in bold, while the lowest accuracy in the gray color.

Categories	Scenario	GBM	RF	Tree	Logit	LASSO	LGT	SVM	Ensemble	LogitGPs	sparseSVM	dynaTrees	regLogit	Probit	DNN
Infection data (1-3)[1]	Kenya	0.823	0.777	0.752	0.765	0.772	0.728	0.769	0.777	0.808	**0.829**	0.815	0.799	0.798	0.781
Baseline infection data (C1)	Kenya	0.757	0.774	0.733	0.744	0.735	0.741	0.819	0.751	0.812	0.825	**0.822**	0.815	0.765	0.782
Infection around (C2)	Kenya	0.749	0.755	0.723	0.713	0.701	0.671	**0.814**	0.714	0.810	0.813	0.811	0.811	0.779	0.759
Environment (C3)	Kenya	0.811	0.803	0.753	0.749	0.741	0.740	0.760	0.787	0.797	0.803	**0.825**	0.807	0.749	0.775
Agriculture (C4)	Kenya	0.768	0.749	0.740	0.756	0.739	0.661	0.779	0.740	0.787	**0.825**	0.818	0.799	0.723	0.765
Geography (C5)	Kenya	0.757	0.728	0.733	0.763	0.747	0.689	0.785	0.723	0.795	0.809	**0.821**	0.817	0.749	0.775
Biology (C6)	Kenya	0.782	0.777	0.760	0.748	0.745	0.735	0.812	0.740	0.806	0.811	**0.819**	0.808	0.791	0.767
Society (C7)	Kenya	0.763	0.753	0.727	0.719	0.717	0.701	0.775	0.721	0.783	0.798	**0.823**	0.785	0.736	0.768
All (C8)	Kenya	0.792	0.779	0.726	0.802	0.799	0.689	0.760	0.771	0.713	0.763	**0.812**	0.804	0.782	0.751
Infection data (1-3)	Tanzania	0.633	0.688	0.629	0.654	0.642	0.616	0.657	0.684	**0.701**	0.695	0.697	0.673	0.673	0.669
Baseline infection data (C1)	Tanzania	0.603	0.697	0.568	0.673	0.671	0.673	0.700	0.675	0.668	0.697	**0.700**	**0.700**	0.615	0.680
Infection around (C2)	Tanzania	0.586	0.677	0.565	0.664	0.658	0.628	**0.700**	0.672	0.695	0.693	0.695	0.683	0.646	0.665
Environment (C3)	Tanzania	0.731	0.740	0.685	0.750	0.745	0.707	0.736	0.728	0.701	0.741	**0.779**	0.760	0.715	0.742
Agriculture (C4)	Tanzania	0.703	0.753	0.660	0.737	0.730	0.703	0.717	0.741	0.729	0.727	**0.767**	0.729	0.728	0.741
Geography (C5)	Tanzania	0.720	0.697	0.652	0.754	0.758	0.724	0.737	0.718	0.724	0.732	**0.773**	0.691	0.688	0.723
Biology (C6)	Tanzania	0.594	0.681	0.550	0.666	0.656	0.649	0.697	0.679	0.680	0.689	**0.700**	0.695	0.618	0.667
Society (C7)	Tanzania	0.605	0.686	0.624	0.665	0.672	0.641	0.683	0.684	0.692	0.669	**0.700**	0.689	0.636	0.667
All (C8)	Tanzania	0.664	0.763	0.663	0.760	0.753	0.554	0.692	0.758	0.700	0.742	**0.783**	0.753	0.723	0.777
Infection data (1-3)	Combined	**0.687**	0.667	0.637	0.656	0.668	0.682	0.676	0.672	0.683	0.677	0.671	0.685	0.658	0.621
Baseline infection data (C1)	Combined	0.630	0.549	0.597	0.552	0.552	0.552	0.614	0.553	**0.637**	0.612	0.614	0.607	0.609	0.495
Infection around (C2)	Combined	0.664	0.627	0.613	**0.656**	0.653	0.651	0.643	0.637	0.633	0.608	0.602	0.624	0.608	0.584
Environment (C3)	Combined	0.764	0.755	0.699	0.761	0.762	0.712	0.768	0.760	0.763	0.720	**0.775**	0.771	0.745	0.741
Agriculture (C4)	Combined	0.740	0.745	0.676	0.717	0.717	0.696	0.713	0.739	0.741	0.733	**0.762**	0.718	0.728	0.705
Geography (C5)	Combined	0.736	0.725	0.681	0.727	0.725	0.726	0.749	0.723	0.746	0.715	**0.764**	0.723	0.739	0.738
Biology (C6)	Combined	0.672	0.686	0.650	0.723	0.726	0.725	0.741	0.720	**0.743**	0.730	0.738	0.731	0.728	0.733
Society (C7)	Combined	0.716	0.705	0.679	0.713	0.712	0.714	0.704	0.711	0.700	0.675	**0.731**	0.685	0.676	0.715
All (C8)	Combined	0.740	0.764	0.701	0.765	0.764	0.699	0.758	0.752	0.749	0.770	0.758	**0.777**	0.751	0.738
Infection data (1-3)	Between I	0.413	**0.513**	0.479	0.455	0.451	0.480	0.409	0.493	0.369	0.377	0.391	0.453	0.449	0.460
Baseline infection data (C1)	Between I	**0.489**	0.466	0.483	0.471	0.471	0.479	0.415	0.488	0.433	0.409	0.410	0.429	0.453	0.448
Infection around (C2)	Between I	0.355	0.649	0.605	**0.693**	**0.693**	**0.693**	0.564	0.567	0.693	0.566	0.409	0.546	0.671	**0.693**
Environment (C3)	Between I	0.690	0.688	**0.696**	0.693	0.693	0.693	0.693	0.693	0.693	0.653	0.693	0.614	0.693	0.693
Agriculture (C4)	Between I	0.659	0.690	0.681	0.687	0.687	0.687	0.677	0.685	0.693	0.603	0.693	0.581	**0.692**	0.681
Geography (C5)	Between I	0.693	0.693	0.693	0.693	0.693	0.693	0.601	**0.694**	0.693	0.570	0.693	0.523	0.693	0.693
Biology (C6)	Between I	0.690	0.688	0.688	0.688	0.688	0.688	0.591	**0.696**	0.683	0.560	0.689	0.539	0.694	0.689
Society (C7)	Between I	0.673	0.690	0.681	0.683	0.683	0.683	0.651	0.691	0.655	0.591	**0.693**	0.587	**0.693**	0.683
All (C8)	Between I	0.455	0.688	0.657	0.685	0.685	0.681	0.687	0.617	0.672	0.638	0.677	0.669	**0.693**	0.689
Infection data (1-3)	Between II	**0.485**	0.330	**0.485**	0.381	0.381	0.405	0.351	0.346	0.295	0.302	0.317	0.399	0.410	0.349
Baseline infection data (C1)	Between II	**0.467**	0.313	0.459	0.285	0.284	0.283	0.295	0.284	0.334	0.295	0.295	0.281	0.419	0.295
Infection around (C2)	Between II	0.705	0.705	0.705	0.536	0.536	0.536	0.616	**0.718**	0.705	0.575	0.709	0.549	0.707	0.576
Environment (C3)	Between II	0.701	0.705	0.691	0.593	0.593	0.593	0.672	**0.708**	0.705	0.645	0.690	0.587	0.685	0.667
Agriculture (C4)	Between II	0.705	0.705	0.691	0.547	0.547	0.547	0.633	**0.706**	0.705	0.608	0.703	0.557	0.702	0.602
Geography (C5)	Between II	**0.705**	0.684	0.681	0.421	0.421	0.420	0.574	0.681	0.693	0.531	0.684	0.532	0.691	0.487
Biology (C6)	Between II	0.705	0.705	0.705	0.469	0.469	0.470	0.576	**0.712**	0.705	0.549	0.708	0.530	0.708	0.526
Society (C7)	Between II	0.703	0.704	0.688	0.616	0.616	0.616	0.615	0.701	**0.705**	0.597	0.697	0.555	0.701	0.681
All (C8)	Between II	0.693	**0.705**	0.678	0.690	0.699	0.689	0.704	0.700	**0.705**	0.691	**0.705**	0.689	0.698	0.701

[1] Using infection data in years 1 and 3.

Table S8 shows the accuracy of each model for the predictor configurations, where the models were developed based on four of the six arms. This is because two of the six arms lacked infection data in the third year (Figure S1), and infection data from the third year were necessary inputs for the previous non-baseline method required

(Shen et al., 2020). Compared to the method using only baseline infection data, the previous non-baseline method usually exhibited higher accuracy. However, based on the best model, the data fusion method improved the hotspot prediction results of the previous non-baseline method on test sets for the Tanzania, combined-countries, and between-countries scenarios.

REFERENCES

CGLS, 2020. Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2015: Globe. https://zenodo.org/records/3939038.

Chen, Y., Chang, X., Luo, F., Huang, H., 2023. Additive dynamic models for correcting numerical model outputs. Computational Statistics & Data Analysis , 107799.

Chen, Y., Chang, X., Zhang, B., Huang, H., 2024. Efficient and effective calibration of numerical model outputs using hierarchical dynamic models. The Annals of Applied Statistics 18, 1064–1089.

Dan, C., Nupur, K., Jennifer, C., 2022. Dataset: SCORE S. mansoni Cluster Randomized Trial. ClinEpiDB. https://clinepidb.org/ce/app/workspace/analyses/DS_d6a1141fbf/new. Accessed 1 February 2023.

ECMWF, 2024. The European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5). https://cds.climate.copernicus.eu/.

GFAS, 2024. Global Fire Assimilation System. https://ads.atmosphere.copernicus.eu/datasets/cams-global-fire-emissions-gfas?tab=overview.

GLAD, 2022. Global Land Analysis & Discovery: Global cropland expansion in the 21st century. https://glad.umd.edu/dataset/croplands.

Potapov, P., Turubanova, S., Hansen, M.C., Tyukavina, A., Zalles, V., Khan, A., Song, X.P., Pickens, A., Shen, Q., Cortez, J., 2022. Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. Nature Food 3, 19–28.

SDAC, 2022. The Water Security Indicator Model-Global Land Data Assimilation System (WSIM-GLDAS. https://sedac.ciesin.columbia.edu/data/set/water-wsim-gldas-v1/data-download.

Shen, Y., Sung, M.H., King, C.H., Binder, S., Kittur, N., Whalen, C.C., Colley, D.G., 2020. Modeling approaches to predicting persistent hotspots in score studies for gaining control of schistosomiasis mansoni in kenya and tanzania. The Journal of infectious diseases 221, 796–803.