Predicting malaria cases using remotely sensed environmental variables in Nkomazi, South Africa

Abiodun Morakinyo Adeola,[1,2] Joel Ondego Botai,[1,3,4] Jane Mukarugwiza Olwoch,[3,5] Hannes C.J. de W. Rautenbach,[1,2,6] Omolola Mayowa Adisa,[3] Christiaan de Jager,[2] Christina M. Botai,[1] Mabuza Aaron[7]

[1]South African Weather Service, Pretoria, South Africa; [2]UP Institute for Sustainable Malaria Control, School for Health Systems and Public Health, University of Pretoria, Pretoria, South Africa; [3]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Hatfield, South Africa; [4]School of Agricultural Earth and Environmental Sciences, University of KwaZulu-Natal, South Africa; [5]Southern African Science Service Centre for Climate Change and Adaptive Land Use (SASSCAL), Windhoek, Namibia; [6]Faculty of Natural Sciences, Akademia, Centurion, South Africa; [7]Malaria Control Programme, Mpumalanga Department of Health, Nelspruit, South Africa

Abstract

There has been a conspicuous increase in malaria cases since 2016/2017 over the three malaria-endemic provinces of South Africa. This increase has been linked to climatic and environmental factors. In the absence of adequate traditional environmental/climatic data covering ideal spatial and temporal extent for a reliable warning system, remotely sensed data are useful for the investigation of the relationship with, and the prediction of, malaria cases. Monthly environmental variables such as the normalised difference vegetation index (NDVI), the enhanced vegetation index (EVI), the normalised difference water index (NDWI), the land surface temperature for night (LSTN) and day (LSTD), and rainfall were derived and evaluated using seasonal autoregressive integrated moving average (SARIMA) models with different lag periods. Predictions were made for the last 56 months of the time series and were compared to the observed malaria cases from January 2013 to August 2017. All these factors were found to be statistically significant in predicting malaria transmission at a 2-months lag period except for LSTD which impact the number of malaria cases negatively. Rainfall showed the highest association at the two-month lag time (r=0.74; P<0.001), followed by EVI (r=0.69; P<0.001), NDVI (r=0.65; P<0.001), NDWI (r=0.63; P<0.001) and LSTN (r=0.60; P<0.001). SARIMA without environmental variables had an adjusted R^2 of 0.41, while SARIMA with total monthly rainfall, EVI, NDVI, NDWI and LSTN were able to explain about 65% of the variation in malaria cases. The prediction indicated a general increase in malaria cases, predicting about 711 against 648 observed malaria cases. The development of a predictive early warning system is imperative for effective malaria control, prevention of outbreaks and its subsequent elimination in the region.

Introduction

Malaria is a major endemic below 1,300 m above mean sea level in the north-eastern part of South Africa (SA), where it affects about 5 million of the population who has recently witnessed a surge in malaria morbidity and mortality (STATS SA,

2017). The endemic regions are in the provinces of Mpumalanga, Limpopo, and KwaZulu-Natal (South Africa National Department of Health, 2011). The disease is markedly seasonal, with varying intensity of transmission due to environmental and climatic factors such as rainfall, temperature, elevation, and humidity favouring the development of the vector and parasite (Teklehaimanot *et al*., 2004; Ngomane and de Jager, 2012; Midekisa *et al*., 2012). The transmission is highest during the wet summer months (September to May), and peak transmission occurs in January/February (Ngomane and de Jager, 2012). The *Plasmodium falciparum* parasite accounts for about 95% of the total malaria infections in SA and the mosquito *Anopheles arabiensis* is the major local vector (Govere *et al*., 2007).

More often than not, meteorological conditions, such as high rainfall and/or high temperature, are cited as the causing factors for malaria epidemics (Alemu *et al*., 2011; Ferrão *et al*., 2017). Hence, many efforts have been made to predict malaria epidemics by using climatic variables on the local (Craig *et al*., 2004; Teklehaimanot *et al*., 2004; Ceccato *et al*., 2007; Ferrão *et al*., 2017) and the regional (Craig *et al*., 1999; Gething *et al*., 2011; Weiss *et al*., 2014) scale. However, many of these studies report variations in the relationship between climatic factors and malaria occurrences from one geographic space to another. This suggests that one or more climate factors are more important than others for malaria. Rainfall has both direct and indirect relationship with its incidence by creating suitable breeding habitats for the vector. However, excessive rainfall can also have a negative impact on the mosquitoes by washing away the larvae (Teklehaimanot *et al*., 2004). Temperature impacts survival of both mosquito and parasite. Larval development takes about 47 days at 16°C, while parasite development ceases at temperatures of <15°C and the organism dies at temperatures >40°C (Gething *et al*., 2011; Weiss *et al*., 2014). Vegetation provides a resting habitat for adult *Anopheles* and also serves as an indicator of the availability of moisture (Machault *et al*., 2011; Sarfraz *et al*., 2014). Elevation is associated with the flight range of the vector as the proportion of moving mosquitoes declines exponentially with distance and height from the breeding habitat (Thomas *et al*., 2013). Nevertheless, as alluded to by (Ceccato *et al*., 2005; Adimi *et al*., 2010; Machault *et al*., 2011) the lack of adequate climatic data is a major constraint for the development of a reliable early warning system (EWS). With the recent surge in malaria morbidity and mortality in SA, there is no better time to tackle the urgent need for the development of an operational malaria EWS to predict when, where, and what magnitude of malaria epidemics might occur, with sufficient lead-time to target scarce resources for effective control measure. However, in contrast to the conventional data gathering methods (Jensen, 2007), remote Sensing (RS) offers advantages such as large area coverage and continuous the spatiotemporal representation of Earth's surface.

Environmental as well as climatic variables derived from Earth-observing satellites, for instance: land surface temperature (LST), the enhanced vegetation index (EVI), the normalized difference vegetation index (NDVI), the normalized difference water index (NDWI) and rainfall estimates have been used to identify mosquito breeding habitats (Mushinzimana *et al*., 2006; Julie *et al*., 2010), to predict outbreaks (Hay *et al*., 1998; Adimi *et al*., 2010) and for the development of malaria EWSs (Ceccato *et al*., 2005; Midekisa *et al*., 2012). These studies have led to a significant contribution to malaria control in various regions; however a large percentage of malaria studies in SA are based on the use of conventional climate data (Kleinschmidt *et al*., 2001; Craig *et al*., 2004; Silal *et al*., 2014) rather than RS despite its great potential. Although reported recently (Adeola *et al*., 2015; 2016; 2017b) no study has studied the direct relationship between remotely sensed environmental data and malaria in SA.

An in-depth understanding of the relationships between malaria cases and environmental factors over a period of time may offer a useful insight towards the effective control of the disease. The South African malaria elimination strategy aims to eliminate malaria within its borders by 2018 on the one hand, it is faced with surging epidemics on the other, the characterization of malaria cases with remotely derived environmental/climatic variables is imperative to provide adequate spatial orientation and predictive tool for targeted local malaria control efforts. This, in particular, should help to improve malaria control measures in Nkomazi municipality which records high malaria transmission in Mpumalanga Province (Figure 1). However, less attention has been paid to other malaria endemic areas of KwaZulu-Natal (Adeola *et al*., 2015) and this paper aims at extending the previous work done by the authors (Adeola *et al*., 2016) by i) determining the relationships between malaria cases and remotely sensed climatic and environmental variables; and ii) developing a flexible modeling framework for predicting malaria cases based on remotely sensed variables.

This study builds on the preliminary studies (Adeola *et al*., 2016; 2017b) in the quest of developing a robust and functional malaria early warning system using satellite derived environmental data due to its advantage over traditional climate data which are not available at sufficient spatial scale and are not adequate due to lengthy data gap over the study area

Materials and Methods

Study area

Nkomazi Municipality is located in the North-eastern part of SA, and lies between latitudes 25°19'0"S and 26°00'0"S and longitudes 31°15'0"E and 32°01'0"E. The municipality is bordered to the east by Mozambique and in the south by Swaziland and the Kruger National Park to the north. The area covers a total area of 3255, 67 km², representing 4.1% and 23% of the total land area of Mpumalanga Province and Ehlanzeni District, respectively (Figure 1). It has 54 towns/villages, mostly concentrated in the southern part of the municipality. The municipality had a total population of 277 864 in 1996 that has increased to 410 907 in 2016 (STATS SA, 2016). It enjoys sub-tropical weather conditions, with temperatures ranging between 2°C and 43°C, with an average of 22.6°C, and an annual average rainfall of 680 mm. Nkomazi Municipality varies in elevation from about 120 to about 1 250 m. The western part is densely vegetated with undulating hills and deeply incised valleys. The area is drained by two major rivers, namely the Komati River to the east, with its main tributary, the Lomati River to the west. The municipality is known for its richness in sugarcane, fruits and vegetable production.

Data collection

Both malaria and environmental data used in this study span a period of approximately 18 years (from 2000 to August 2017). Daily malaria data, including both passive and active data, were

acquired by the malaria control programme of the Department of Health, Mpumalanga. The passive data are those obtained from patients who present at the health facility and test positively for *Plasmodium,* while the active data are those collected through screening measures. These include people with non-specific symptoms such as fever, or those residing near or in the same homesteads with recently confirmed cases. There are 48 healthcare facilities in the area ranging from hospitals to clinics across the municipality. The records in the malaria information system contain information such as facility name, date of diagnosis, number of cases, deaths, age, gender, infection source and the facility coordinates. For this study, only localized notified malaria (56.38%) cases were utilized in the model.

Vegetation indices consisting of NDVI, EVI, and water index NDWI were extracted from a 16-day composite of the US National Aeronautics and Space Administration's (NASA) Moderate-resolution Imaging Spectroradiometer (MODIS) MOD13Q1 on board of both Terra and Aqua satellites system (NASA, 2017). The indices, providing regular spatial and temporal assessments of vegetation conditions, were computed from atmospherically corrected bi-directional surface reflectance masked for water, clouds, heavy aerosols, and cloud shadows (NASA, 2017).

The day-time (LSTD) and night-time (LSTN) values were derived from an 8-day composite of the MODIS MOD11A2 thermal sensor on board the NASA-Terra satellite system (NASA, 2017). The procedure for downloading and processing data is fully described by Busetto and Ranghetti (2016). Monthly rainfall estimates (January 2000-November 2015) were derived from the tropical rainfall measuring mission (TRMM)-3B43 and December 2015-August 2017 from the newer global precipitation measurement, a joint mission between the Japanese Aerospace Exploration Agency (JAXA) and NASA as the principal participants. The data was downloaded using the Network Common Data Form simple subset wizard in the NASA Goddard Earth Sciences Data and Information Services Center (TRMM, 2011).

Data analysis

All data processing, statistical analyses and modeling were performed in R (R Core Team, 2016). Malaria cases and environmental data were aggregated to monthly data. The malaria cases were used as the dependent variables with the environmental variables as the independent ones. Descriptive statistics using boxplot (Spitzer *et al*., 2014) was performed to explore the climatic variables and malaria case. Pearson's correlation was used to statisti-
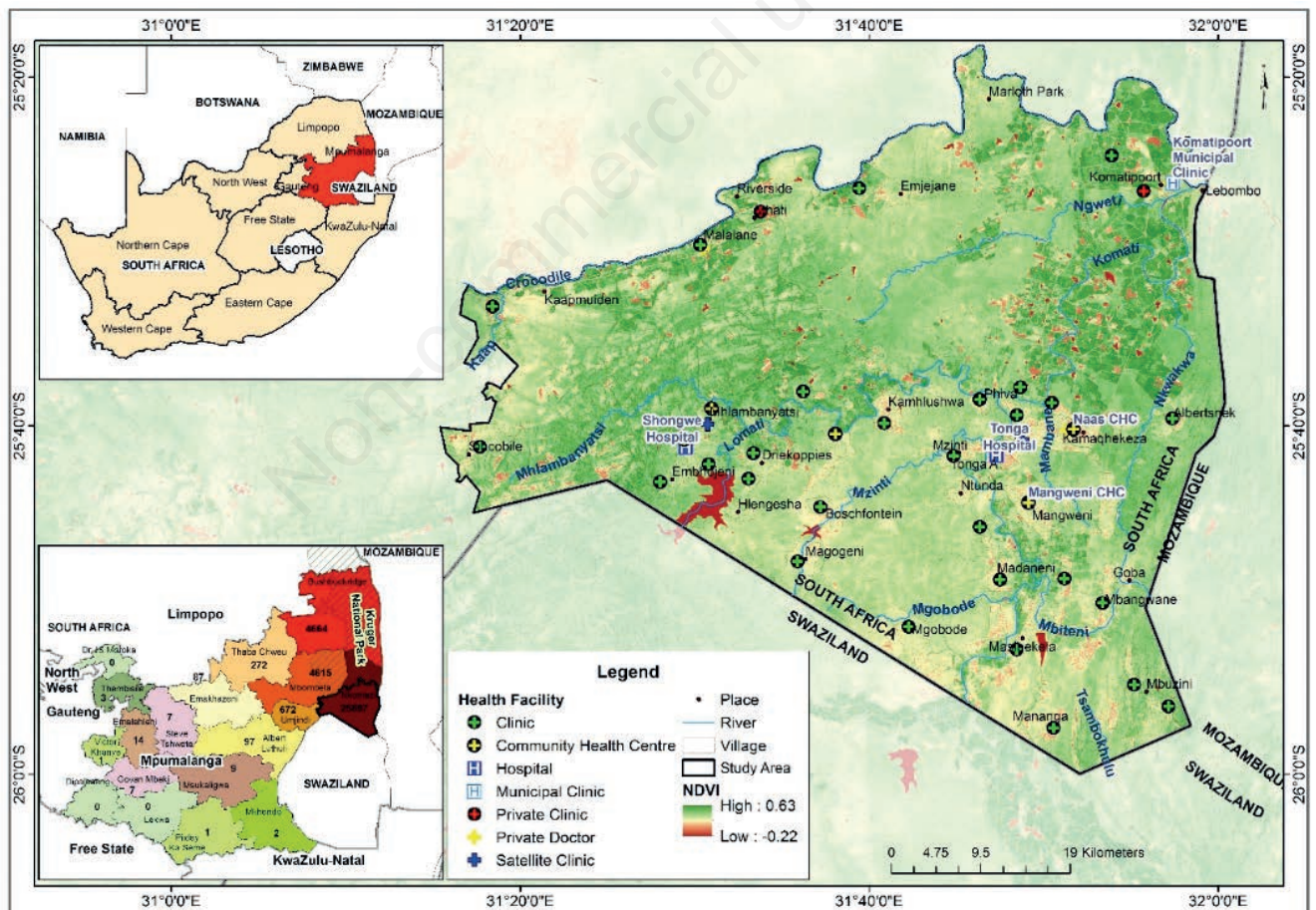

Figure 1. Maps of the study area in South Africa. Left top: South African provincial map showing Mpumalanga Province in red; left bottom: Mpumalanga Province showing the malaria case distribution from January 1997 to August 2017; right: Nkomazi Municipality showing the villages, health facilities and major rivers with the Normalized Difference Vegetation Index (NDVI) overlay.

cally determine the association between monthly malaria cases and environmental variables and at different lags at 0 to 3 months lagged periods. With the aim of determining the statistical associations of the environmental variables with malaria cases and predicting the number of malaria cases, a seasonal autoregressive integrated moving average (SARIMA) model without exogenous variables and with exogenous variables was employed as the baseline predictive model (Box and Jenkins, 2008).

The SARIMA model provides a robust set of tools for performing time series analysis, parameter estimation and forecasting. SARIMA is particularly appropriate in situations when the time series data exhibit seasonality periodic fluctuations. The SARIMA model structure is $SARIMA = (p,d,q)(P,D,Q)_S$, expressed as $\Phi(L^S)\phi(L)(1-L)^d(1-L^S)^D Y_t = \Theta(L^S)\theta(L)\varepsilon_t$, where L is the lag operator denoted by $L^k = Y_{t-k} / Y_t$, $\phi(L) = 1-\phi_1 L^1 -\phi_2 L^2 -\phi_3 L^3 -\ldots \phi_p L^p$ which is an Autoregressive (AR) polynomial function of order p with a vector of the coefficient $\phi' = (\phi_1, \phi_2, \phi_3,\ldots, \phi_p)$, $\theta(L) = 1 + \theta_1 L^1 + \theta_2 L^2 + \theta_3 L^3 +\ldots+ \theta_q L^q$. The latter is a Moving Average (MA) polynomial of order q with a vector of coefficients $\theta' = (\theta_1, \theta_2, \theta_3,\ldots, \theta_q)$, $\Phi(L^S) = 1 - \phi_{S,1} L^S - \phi_{S,2} L^{2S} - \phi_{S,3} L^{3S} -\ldots- \phi_{S,P} L^{PS}$, and $\Theta(L^S) = 1 + \theta_{S,1} L^S + \theta_{S,2} L^{2S} + \theta_{S,3} L^{3S} +\ldots+ \theta_{S,Q} L^{QS}$, which are seasonal polynomial functions of order P and Q, respectively, and which fulfill the invertibility and stationarity conditions. The variable d is the number of differencing passes needed to stationarise the series, while D is the number of seasonal differences, S is the seasonal period (*e.g.* in quarterly data s=4 and in monthly data s=12) and ε_t are error terms known as white noises since they are interpreted as an exogenous effect that the model is not able to explain. The white noise, in this case, maybe contributions of other environmental variables like altitude, the use of IRS, population movement and other factors which are not included in the model.

In order to induce constant variance and stationarity, the malaria case data were logarithmically transformed and differenced once. Augmented Dickey-Fuller (ADF) test, a formal statistical test for stationarity (Hyndman and Athanasopoulos, 2017), which was used to check for stationarity. The ADF tests if the change in Y can be explained by lagged value and a linear trend. A stationary time series is a time series without trend, having a constant mean and variance over time, a condition to be satisfied in using the Box-Jenkins approach (Box and Jenkins, 2008). The time series data were divided into training (80%) and testing (20%) data in order to allow for cross-validation. Cross-validation is primarily a way of assessing the predictive performance of a model against a set of data not used in estimation. The auto.arima function, which searches through combinations of order parameters and selects the order that best fits the model, was performed on the transformed malaria time series without environmental variables to identify the auto-regressive, moving average and differencing orders of the SARIMA model. The Autocorrelation Function (ACF) and Partial Autocorrelations Function (PACF) were examined. The model with the lowest Akaike Information Criterion (AIC) value, indicating the best-fit model, was selected (Burnham and Anderson, 2002). Consequently, using the identified best SARIMA model form, SARIMA with exogenous variables model (SARIMAX) was fitted with the transformed malaria and environmental variables at different lag periods. The SARIMAX model is mainly a linear regression model that uses SARIMA model type process with exogenous variables, $(p,d,q) (P,D,Q)_S (X)$, where X is the vector of external variables (Peter and Silvia, 2012). The lag periods were used to assess the associations between malaria cases and the environmental variables from 0 to 3 months using cross-

correlation analysis (a time lag was defined as the time duration between malaria incidence and environmental observation). The adequacy of each model was diagnosed by plotting the residuals of the ACF and PACF using Ljung_Box Q statistics (Ljung *et al.*, 1978). Out-of-sample predictions of the models (SARIMA with environmental variables; SARIMAX) for the last 56 months of the time series were made and compared with the observed malaria data (January 2013 to August 2017). The performance measures of prediction were assessed by computing the Root Mean Squared Error (RMSE), which gives an indication of how the predicted values differ from the observed values.

Results

During the study period, a total of 25,897 malaria cases were recorded in Nkomazi Municipality. This number of cases accounts for 32.3% of the total malaria cases in Mpumalanga Province (80,058 malaria cases - see inset map in Figure 1). Within this period, a total of 37,054 (46.3%) malaria cases were locally transmitted, 43,004 (53.6%) imported and 62 (0.08%) untraceable. Mozambique was the highest contributing source of imported malaria (41,550) accounting for 96.6% of the total figure. In Nkomazi Municipality, 24,684 (95.3%) of the cases were detected by passive surveillance and 1213 (4.7%) by active. Gender-wise, males accounted for 13,601 (52.5%) and females for 12,296 (47.5%) of the cases. Age was categorized into 3 major groups of 0-14, 15-64 and ≥65 representing the young, those working and the elderly, respectively. Malaria infection was more common in the economically active group - 16,459 (63.6%), while those belonging to the young amounted to 8,956 (34.6%) and those in the old-age group were only 482 (1.9%). Figure 2A indicates that there is monthly variation in the cases of malaria in Nkomazi, with incidence rising from September and dropping after May, while Figure 2B shows the annual locally notified malaria cases in Nkomazi Municipality from January 2000 to December 2016. There was a remarkable peak in 2000 (8,525) with an equally remarkable decline (1,128 cases) in 2006 soon afterwards. Peak malaria incidence was noticeable in March and few cases were recorded in June, July, and August, *i.e.* the winter months.

Shown in Figure 3 are monthly time series box plots of malaria cases and the environmental variables showing the minimum, first quartile, median, third quartile, and maximum values in Nkomazi for the period under investigation. At a monthly average rainfall of about 9 mm, July is the driest month. The greatest amount of rainfall is received in January, with an average of about 130 mm. Between the driest and wettest months, the difference in rainfall is 117 mm. January is the warmest month of the year at a day-time average temperature of 31.2°C and 20.0°C night-time average temperature. June is the coldest month of the year, with average temperatures of about 17.8°C. The variation in annual temperature is around 8.4°C. The highest mean NDVI (0.67), EVI (0.44) and NDWI (-0.058) values are recorded in March February and January, respectively. While the lowest mean NDVI (0.41), EVI (0.21) and NDWI (-0.14) values are recorded in August/September, August and August, respectively.

Over the period considered in the study, the mean maximum and minimum daytime temperature received was 41.7 and 21.3°C, respectively, with a standard deviation (SD) of 4.34. On the other hand, the mean maximum and minimum night-time temperatures recorded were 22.6 and 10.4°C, respectively (SD=3.08). The high-

est and lowest rainfall received were recorded in February 2000 amounting to 594.5 and 0.19 mm, respectively (SD=72.7). From the time series analysis, a peak; a little above the 5-year average was noticed in the total rainfall received from November 2016 till February 2017. A corresponding effect of the increased rainfall was noticed with regard to the other variables. The maximum, minimum averages of NDVI, EVI and NDWI values were 0.75; 0.34, (SD=0.11), 0.53; 0.17 (SD=0.09) and 0; -0.17 (SD=0.04), respectively.

The result of the Pearson's correlation indicated that the monthly environmental variables corrected with the number of the monthly malaria cases during the period under study (Appendix Figure S1). The correlation indicated that there was a statistically significant association between the monthly environmental variables time series and the malaria case time series.

The highest correlation between the number of malaria cases and an environmental variable was shown by rainfall (r=0.36; P<0.001), followed by NDVI and EVI (r=0.28; P<0.001), and NDWI (r=0.17; P=0.014). There was no statistically significant relationship between the temperatures and malaria case count (r=-0.094; P=0.179 and r=0.056; P=0.423 for LSTD and LSTN, respectively). Furthermore, as indicated in Table 1, none of the variables showed any strong positive or negative correlation with the number of malaria cases at 0-month lag period. Only at least 1-month time lag when the environmental variables time series lagged the malaria time series. A negative relationship was indicated between the number of malaria cases and LSTD. Although temperatures were not significantly associated with the number of malaria cases, LSTN was significant at a 2-month lag time with the number of malaria cases. Most significant associations were observed at the 2-month lag time. Rainfall showed the highest association at the two-month lag time (r=0.74; P<0.001), followed by EVI (r=0.69; P<0.001), NDVI (r=0.65; P<0.001), NDWI (r=0.63; P<0.001) and LSTN (r=0.60; P<0.001). When different combinations of the environmental variables with the number of malaria cases were performed, total monthly rainfall, monthly mean EVI, NDWI and LSTN, at a two-month lagged effect, were found to be the most significant climatic variables for malaria transmission in Nkomazi Municipality having an adjusted coefficient of determination (R^2=0.64; P<0.001).

Shown in Figure 4 are monthly mean aggregates of malaria cases against environmental variables. The number of malaria

Table 1. Correlation between malaria cases and environmental variables at lags of 0, 1, 2, and 3 months.

Climatic variables	Lag (month)	R	P-value
Rainfall	0	0.19	0.062
	1	0.32	<0.001
	2	0.74	<0.001
	3	0.56	<0.001
EVI	0	0.24	0.069
	1	0.42	<0.001
	2	0.69	<0.001
	3	0.56	<0.001
NDVI	0	0.22	0.083
	1	0.31	0.001
	2	0.65	<0.001
	3	0.56	<0.001
NDWI	0	0.31	0.111
	1	0.27	0.055
	2	0.63	<0.001
	3	0.56	<0.001
LSTD	0	-0.24	0.109
	1	-0.43	<0.001
	2	-0.51	<0.001
	3	-0.34	<0.001
LSTN	0	0.07	0.145
	1	0.12	0.115
	2	0.59	<0.001
	3	0.26	0.054

EVI, enhanced vegetation index; NDVI, normalized difference vegetation index; NDWI, normalised difference water index; LSTD, day-time land surface temperature; LSTN, night-time land surface temperature.
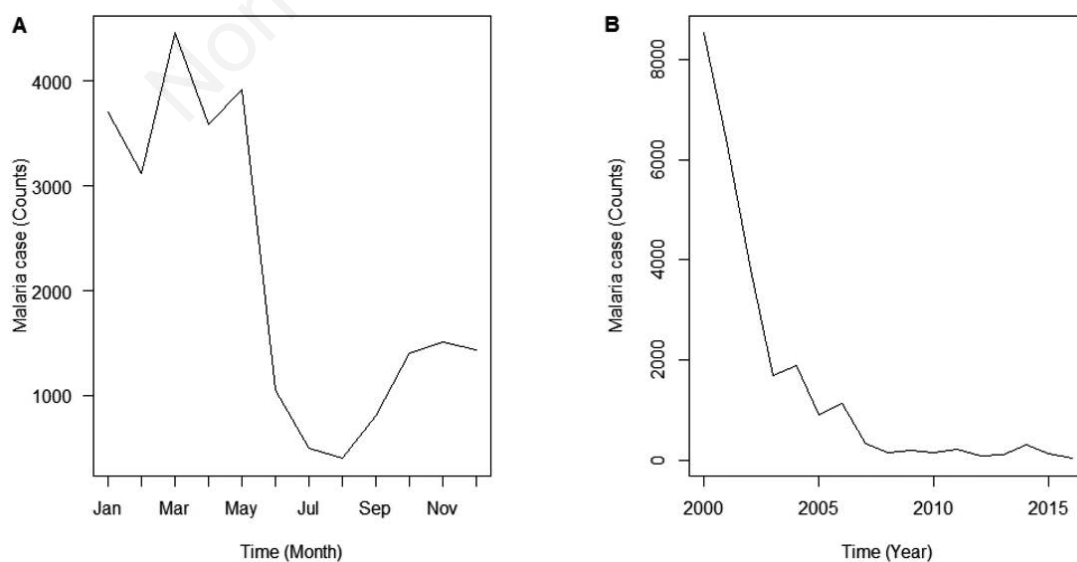

Figure 2. The temporal distribution of locally notified malaria in Nkomazi local Municipality, Mpumalanga January 2000-August 2017. (A) Monthly distribution; (B) annual distribution.

cases increased with increasing nighttime temperature (17-20°C) from January-May and decreased with increasing LSTD from June to November. NDVI values between 0.50-0.65 from January-May (r=0.78; P<0.001) and NDVI values of 0.42-0.45 from September-December (r=0.67; P<0.001) were found to be more significantly associated with the number of malaria cases. Similarly, EVI values between 0.28-0.45 from January-May (r=0.72; P<0.001) and EVI values of 0.22-0.26 from September-December (r=0.61; P<0.001) were found to be more significantly associated with malaria. NDWI values greater than -0.12 but less than -0.06 from January-May and values greater than -0.14 but less than -0.12 were significantly associated with the number of malaria cases. Monthly total rainfall amount of about 20-120 mm was seen to be a trigger for malaria from November-May having its peak in January and hence confirming the 2-months lag period.

The result from the ADF test indicated that the transformed time series of monthly malaria counts was stationary, *i.e.* the P-value was less than 0.05 meaning that there was no unit root. The auto.arima function suggested that the SARIMA $(2,1,1)$ $(2,0,1)_{12}$ model form is the best model fitting the training dataset by the lowest AIC values compared to other models. The AIC values for the SARIMA models fitted to the malaria case training data are shown in Table 2. The adequacy of the model was further assessed by diagnosing the residuals of the model. Appendix Figure S2 panel a indicates that the standardized residuals estimated from the model was an independent and identically distributed sequence with a mean of zero and a constant variance. Plotted in Appendix Figure

S2 panel b is the ACF of the residuals indicating that the autocorrelations were close to zero. This result implies that the residuals did not significantly deviate from a zero mean white noise (P<0.001). The plot of the P-values in Appendix Figure S2 panel d showed that the seasonal and non-seasonal autoregressive (AR) and moving average (MA) parameters of the model have P-values significant at the 5% level. Additionally, the ACF of the residuals shown in Appendix Figure S2 panel b indicates that these autocor-

Table 2. Akaike Information Criterion (AIC) values of log-transformed malaria cases in Nkomazi Municipality.

Model type	AIC values of SARIMA	AIC values of SARIMA plus environ. variables
$(2,1,1)(2,0,1)_{12}$	59.534	54.438
$(2,1,0)(2,0,0)_{12}$	74.234	71.597
$(2,1,1)(2,0,0)_{12}$	62.446	58.223
$(2,1,2)(2,0,0)_{12}$	62.974	58.256
$(2,1,1)(1,0,0)_{12}$	86.218	81.673
$(1,1,1)(2,0,0)_{12}$	58.754	56.377
$(3,1,2)(2,0,0)_{12}$	67.342	63.262
$(0,1,1)(2,0,0)_{12}$	62.151	58.186
$(1,1,0)(2,0,0)_{12}$	81.045	77.198

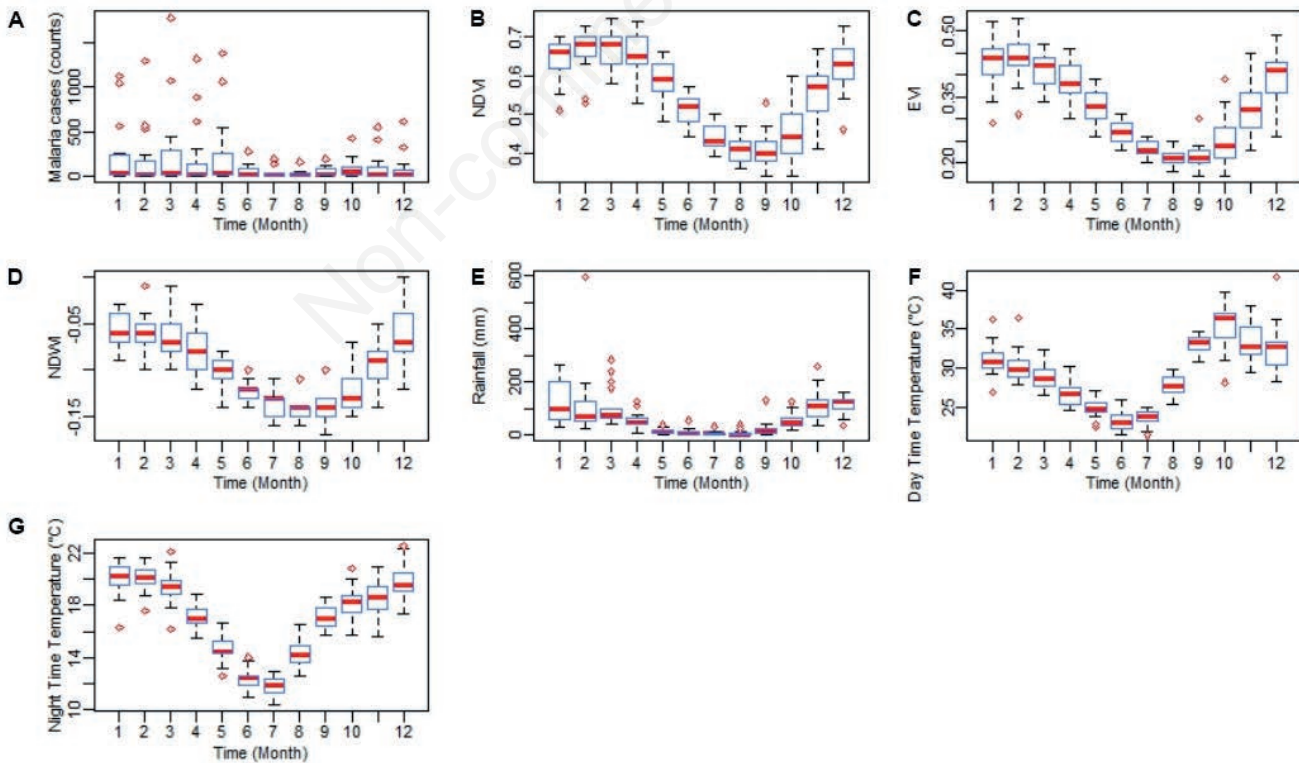SARIMA, Seasonal autoregressive integrated moving average.

Figure 3. Boxplots of malaria case distributions and environmental variables. (A) monthly number of malaria cases; (B) normalized difference vegetation index (NDVI); (C) enhanced vegetation index (EVI); (D) normalised difference water index (NDWI); (E) rainfall; (F) day-time land surface temperature; (G) night-time land surface temperature.

relations were also close to zero. Because of the high P-values, the null hypothesis of independence in the residual series cannot be rejected. It could, therefore, be concluded that the SARIMA model of the form $(2,1,1)$ $(2,0,1)_{12}$ is a good prediction model.

As shown in Table 2, the addition of the environmental variables improved the prediction performance of the SARIMA models from adjusted R^2 of from 0.41 to 0.65. Furthermore, the AIC (Table 2) and RMSE (Table 3) score reduced after the inclusion of the environmental variables. The decrease in the values of the RMSE and AIC indicated that the inclusion of the environmental variables improved the predicting accuracy of the models across all the villages.

Different combinations of SARIMA with environmental variables were tested (Table 3). SARIMA with total monthly rainfall, EVI, NDVI, NDWI and LSTN were able to explain about 65% of variation, while SARIMA with total monthly rainfall, EVI, NDWI and LSTN were able to explain about 64% of the variation. However, because EVI is more sensitive to differentiating vegetation canopy to the ground than NDVI (Matsushita *et al.*, 2007) and high collinearity between EVI and NDVI, only monthly total rainfall, EVI, NDWI and LSTN were left in the model to predict the number of malaria cases to be expected.

The number of malaria cases from January 2013 to August 2017, based on the SARIMA $(2,1,1)$ $(2,0,1)_{12}$ model including the significant predictors for malaria cases (rainfall, EVI, NDWI and LSTN) lagged at 2-month, was subjected to out-of-sample prediction and compared with the 20% observation testing data left out of the modeling procedure (January 2013 to August 2017) (Figure 5 and Table 4). The predicted number of malaria cases were relatively close to the number of cases observed, which indicates that

the model provides an acceptable fit to predict the number of malaria cases within the study area. The prediction indicated a general increase of malaria which is a deviation from the downward trends witnessed in the observed data after the major peak in the year 2000 although with few peaks in 2004, 2006, 2011 and 2014. The above-average numbers of observed malaria cases (major peaks) noticed form February 2015 through May 2015 and declining afterwards until the sudden rise in December 2016

Table 3. Forecasting performance measures of seasonal autoregressive integrated moving average (SARIMA) and SARIMA with exogenous variables models at 2-month lag period.

Forecast method	MAPE	RMSE	Adjusted R^2
SARIMA	27.621	19.541	0.413
SARIMA + rainfall	25.342	17.891	0.492
SARIMA + rainfall + EVI	23.453	16.342	0.527
SARIMA + rainfall + EVI + NDVI	23.112	15.876	0.532
SARIMA + rainfall + EVI + NDVI + NDWI	21.874	15.451	0.607
SARIMA + rainfall + EVI + NDVI + NDWI + LSTN	20.005	13.962	0.647
SARIMA + rainfall + EVI + NDVI + NDWI + LSTN + LSTD	25.211	17.561	0.535
SARIMA + rainfall + NDVI + NDWI + LSTN	20.449	14.215	0.624
SARIMA + rainfall + EVI + NDWI + LSTN	20.143	14.052	0.635

MAPE, Mean absolute percentage error; RMSE, Root mean squared error. EVI, enhanced vegetation index; NDVI, normalized difference vegetation index; NDWI, normalised difference water index; LSTN, night-time land surface temperature; LSTD, day-time land surface temperature.

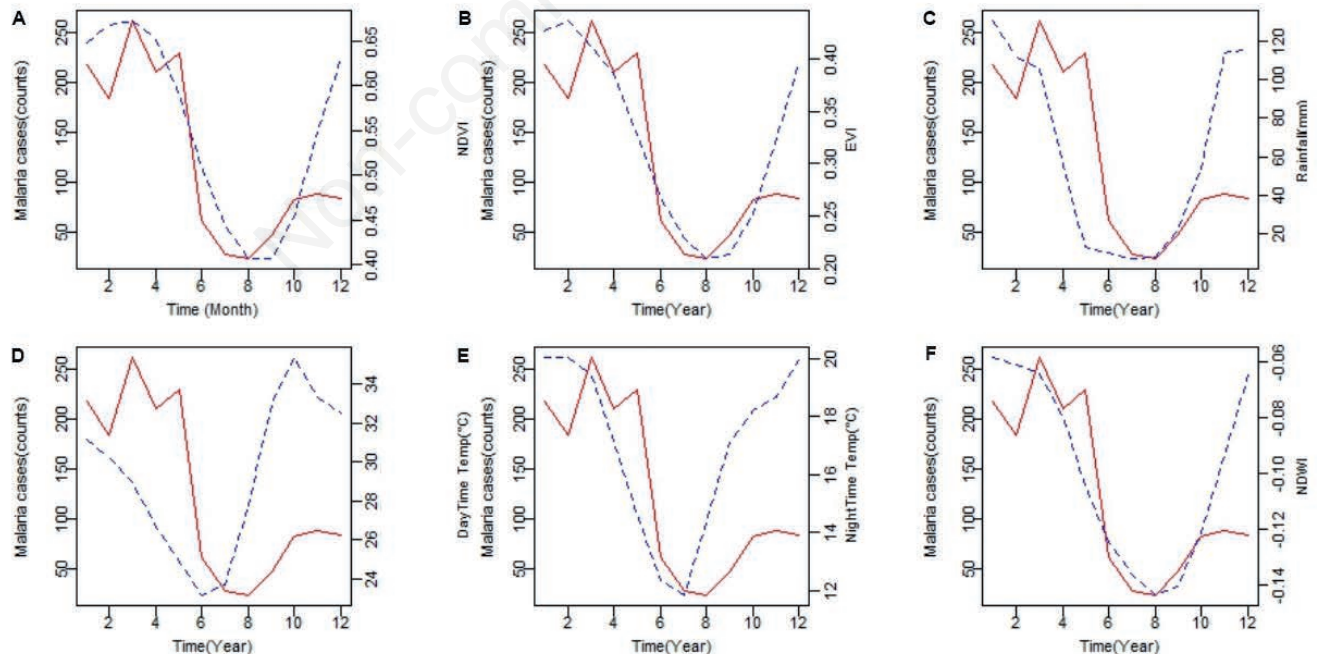**Figure 4. The plot of total monthly time series of malaria cases in relation to environmental variables. Red line, malaria case count; Blue line, the various environmental variables as follows: (A) normalized difference vegetation index (NDVI); (B) enhanced vegetation index (EVI); (C) rainfall; (D) day-time land surface temperature; (E) night-time land surface temperature; (F) normalised difference water index (NDWI).**

stretching into June 2017 was equally well captured by the model.

The model under-predicted the malaria cases in 2014, predicting 222 cases against 297 observed cases. In general, the SARIMA model over-predicted malaria cases in the study area. The model predicted a total number of 711 malaria cases as against 648 observed in the study area. As shown in Appendix Figure S3, the forecast for September to December of 2017 indicated that there would be an additional 64 cases of malaria in Nkomazi Municipality if all other conditions such as the implementation of malaria control measures remain unchanged. Out-of-sample predicted and observed malaria cases for the next 56 months are given in Table 4 and Table 5 gives the summary of the performance measures of the model's prediction accuracy.

Discussion

The spatial and temporal distribution of malaria transmission is largely determined by climatic and environmental variables (Weiss *et al*., 2014; Ferrão *et al*., 2017). These determinants range from the provision of suitable thresholds for the survival of both the parasites and vectors, to the provision suitable breeding habitats and to the availability of host (human or animals). For this study, the SARIMA model was developed and shown to be specifically useful for epidemiological studies that exhibit seasonal patterns. The SARIMA model in the $(2,1,1)(2,0,1)_{12}$ form that was developed in this study aims to provide a prediction tool to predict the expected number of malaria cases based on historically observed data with and without environmental variables. The latter (rainfall, NDVI, EVI, NDWI and LSTN at the 2-month lag time) were found to be significantly associated with the number of malaria cases in Nkomazi Municipality. These environmental variables were lagged by one to three months, considering the accumulation of rainfall which has an influence on the availability of soil moisture and indirectly on vegetation greenness and water availability in

ponds. This, in turn, influences the occurrence of the *Anopheles* vector which takes about two weeks to complete its life cycle, with an additional two weeks for the incubation of parasites in the human host.

The statistically significant relationship of rainfall with the number of malaria cases agrees with other studies where satellite-derived rainfall estimates were used to study its relationship with malaria transmission. For instance, Midekisa *et al*. (2012) have identified a lag of one to three months between satellite-derived rainfall and malaria cases in five study sites in Ethiopia. Similarly, Ceccato *et al*. (2007) in Eritrea have indicated that satellite-derived rainfall is statistically associated with malaria with a lead time of two to three months. In addition, Ngomane and de Jager (2012) used climatic data, including rainfall and temperature data from ground weather stations, to examine climatic associations with malaria cases in the Nkomazi Municipality. The study found that
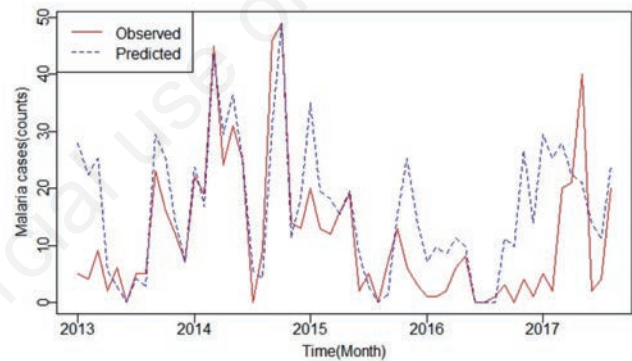

Figure 5. Plot of Actual (observed) and predicted (fit) malaria cases in Nkomazi Municipality January 2013-August 2017.

Table 4. Observed number of malaria cases and out-of-sample predicted values obtained in Nkomazi local Municipality January 2015-August 2017.

Year	Cases	Jan.	Feb.	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.	Total
2013	Observed	5	4	9	2	6	0	5	5	23	16	12	7	94
	Predicted	21	17	19	5	3	1	4	3	22	19	11	6	131
2014	Observed	22	19	45	24	31	25	0	9	46	49	14	13	297
	Predicted	18	13	32	22	27	19	5	4	23	36	9	14	222
2015	Observed	20	13	12	16	19	2	5	0	7	13	6	3	116
	Predicted	26	15	14	12	15	7	3	1	2	12	19	11	137
2016	Observed	1	1	2	6	8	0	0	1	3	0	4	1	27
	Predicted	20	11	7	9	8	1	1	1	9	8	6	8	89
2017	Observed	5	2	20	21	40	2	4	20	-	-	-	-	114
	Predicted	22	19	21	17	16	11	9	18	21	19	14	9	133+64

Based on SARIMA $(2,1,1)(2,0,1)12$ model with climatic variables as exogenous variable.

Table 5. Performance measures of the model for out-sample prediction.

Model	Prediction without environmental variables			Prediction with environmental variables		
SARIMA $(2,1,1)(2,0,1)_{12}$	MAPE	RMSE	Adjusted R^2	MAPE	RMSE	Adjusted R^2
	26.674	18.562	0.412	22.430	15.684	0.687

SARIMA, Seasonal autoregressive integrated moving average; MAPE, Mean absolute percentage error; RMSE, Root mean squared error.

rainfall is significantly associated with the number of malaria cases. EVI and NDVI are measures of vegetation greenness which customarily should respond positively to moisture availability (rainfall) and negatively to its absence. Hence, EVI and NDVI could serve as an indirect indicator of soil moisture, surface water and near-surface humidity. Previous studies, for instance Ceccato *et al.* (2007) have reported a 4-month lag association between malaria and NDVI. However, although NDVI in this study exhibits a statistically significant relationship with the number of malaria cases at the 1 to 3-month lag period, EVI seems to be more statistically associated with malaria at the same lag period. This could be as a result of the advantage the EVI has over the NDVI by being higher over areas of denser vegetation, thus allowing it to determine more changes and variation in a mature canopy (Matsushita *et al.*, 2007).

Although the use of NDWI to detect mosquitoes breeding habitats has previously been reported (McFeeters, 2013; Adeola *et al.*, 2017b), this water index has not been included in any model to determine its relationship with malaria infection. Indeed, this is a new result emanating from this study. The NDWI showed a statistically significant association with the number of malaria cases in the study area. The significant relationship of NDWI with malaria may be explained by the presence of open-water bodies (river and ponds) used for irrigation which amounts to about 18% of the total land area of the municipality (Adeola *et al.*, 2016). The NDWI is a well-established method used to detect surface water which could serve as potential breeding and resting sites for mosquitoes. Thus, NDWI may be an additional index for quantifying environmental factors in relation to malaria infection, and conceivably, other vector-borne diseases.

In this study, LSTD extracted from meteorological earth observation satellite indicated a negative relationship with the number of malaria cases at all tested lag times. However, LSTN at the 2-month lag period indicated a significant association with malaria, which corresponds with the results reported by Ngomane and de Jager (2012) who found temperature (based on climate data from a weather station) to be less of a predictor of malaria in Nkomazi Municipality. In contrast, Adeola *et al.* (2017a) found a positive significant association of minimum temperature with malaria cases locally in Mutale Municipality, a malaria-endemic area in Limpopo Province, SA. Obviously varying associations between malaria and climatic variables have been shown in the three malaria-endemic SA provinces, Limpopo, KwaZulu-Natal and Mpumalanga. For instance, Komen *et al.* (2015) found a strong positive correlation between temperature and malaria (*r*=0.5212), and a weak positive relationship with rainfall (*r*=0.2810) in Limpopo Province, while Craig *et al.* (2004) found that temperatures during the preceding summer and current spring were significantly associated with delta log of the number of malaria cases in KwaZulu-Natal.

In general, this study found that the spatial and temporal patterns of malaria cases in the Nkomazi Municipality is associated with satellite-derived environmental factors. The developed SARIMA model improved the adjusted R^2 from 0.41 to 0.64 after the inclusion of the environmental variables, *i.e.* the SARIMAX model with monthly total rainfall, EVI, NDWI and LSTN were able to explain about 64% of the variation. On the other hand, the SARIMA model explained only 41% of the variation. By including only rainfall into the SARIMA model, the adjusted R^2 increased from 0.413 to 0.492. It further increased to 0.527 by including EVI rising to 0.532 after adding NDVI and further to 0.607 by incorporat-

ing NDWI. The adjusted R^2 increased to 0.647 after the inclusion of LSTN but dropped to 0.535 when LSTD was included in the model indicating the negative effect of LSTD, which is in line with another previous study (Ngomane and de Jager, 2012). Although, SARIMA + rainfall + EVI + NDVI + NDWI + LSTN gave the highest adjusted R^2 (0.647), SARIMA + rainfall + EVI + NDWI + LSTN (adjusted R^2=0.635) was used in the final model because EVI is more sensitive in differentiating vegetation canopy from ground than NDVI (Huete *et al.*, 2002; Matsushita *et al.*, 2007) coupled with high collinearity between EVI and NDVI. The final selection was further confirmed by the adjusted R^2 of 0.624 for SARIMA + rainfall + NDVI + NDWI + LSTN as against adjusted R^2 of 0.635 for the selected model. Also, the decrease in MAPE and RMSE explicitly substantiates that the incorporation of the external variables improves the forecast accuracy. This implies that 64% of the total variation in the number of malaria cases can be explained by the linear relationship between the environmental variables and malaria. The other 36% of the total variation might be explained by other factors such as a high number of imported malaria cases and social factors like population mobility, housing type, sanitation, control measures, public health systems, *etc.* which were not considered in this present study. From the result, it can be deduced that malaria is expected to appear between 9-10 weeks following adequate rainfall of about 20-120 mm, at an average temperature between 17-20°C, NDVI of 0.40-0.65 or EVI of 0.28-0.45 and NDWI values of -0.12 to -0.06.s

The prediction indicated a general increase in malaria cases with a distinct seasonal pattern and significant peaks in February to April. This is as a result of about 2 to 3 months lag effect of environmental variables, particularly rainfall, which starts in September/October. Although, the model under-predicted malaria cases in 2014 which recorded a high malaria incidence, the model was able to depict monthly variation and distinct seasonal pattern. The under-prediction might be attributed to the influence of non-climatic external factors which could have triggered the increase. The over-prediction of the model could be due to the non-inclusion of the malaria control strategies which could have been reinforced after the notice of the sudden rise in the number of malaria cases.

Conclusions

As SA progresses in efforts to eliminate malaria in three endemic provinces by 2018, an effective and operational EWS as proposed by the World Health Organization becomes imperative. Despite the advantages offered by RS for epidemiology research, the use of remotely sensed environmental variables has not been fully been exploited for malaria research in SA. Hence, this study aimed at investigating the relationship between remotely sensed monthly environmental variables with monthly malaria case counts towards the development of a model capable of better prediction. Hence, a SARIMA model regressed with the external variables such monthly total rainfall, EVI, NDWI and LSTN is ideal for estimating the number of malaria cases. This study indicates that total rainfall is the most significant predictor of malaria cases in the study area. The lagged time of 1 to 3 months between the environmental variables and malaria can be used to develop a malaria EWS for municipalities commonly ravaged by high malaria transmission. The performance of the model has the potential for improvement by the inclusion of factors like population movement, migration, elevation, data on indoor residual spraying and

proximity to health. Further research will, therefore, be required in order to determine the impact of these factors and their possible inclusion into the model, which should be regularly updated to an operational malaria EWS.

References

Adeola AM, Botai OJ, Olwoch JM, Rautenbach CJ de W, Kalumba AM, Tsela PL, Adisa MO, Wasswa NF, Mmtoni P, Ssentongo A, 2015. Application of geographical information system and remote sensing in malaria research and control in South Africa: a review. South Afr J Infect Dis 30:114-21.

Adeola AM, Botai OJ, Olwoch JM, Rautenbach CJ de W, Adisa OM, Taiwo OJ, Kalumba AM, 2016. Environmental factors and population at risk of malaria in Nkomazi municipality, South Africa. Trop Med Int Health 21:675-86.

Adeola AM, Botai OJ, Rautenbach CJ de W, Adisa OM, Ncongwane KP, Botai CM, Adebayo Ojo TC, 2017a. Climatic variables and malaria morbidity in Mutale Local Municipality, South Africa: a 19-year data analysis. Int J Environ Res Public Health 14:E1360.

Adeola AM, Olwoch JM, Botai OJ, Rautenbach CJ de W, Kalumba AM, Tsele PL, Adisa OM, Nsubuga FW, 2017b. Landsat satellite derived environmental metric for mapping mosquitoes breeding habitats in the Nkomazi municipality, Mpumalanga Province, South Africa. S Afr Geogr J 99:14-28.

Adimi F, Soebiyanto RP, Safi N, Kiang R, 2010. Towards malaria risk prediction in Afghanistan using remote sensing. Malar J 9:125.

Alemu A, Gemeda A, Wondewossen T, Lemu G, 2011. Climatic variables and malaria transmission dynamics in Jimma town, South West Ethiopia. Parasit Vectors 4:30.

Box G, Jenkins G, 2008. Time series analysis: forecasting and control, 4th ed; John Wiley & Sons: New York, NY, USA.

Burnham KP, Anderson DR, 2002. Model selection and multimodel inference: A practical information-theoretic approach, 2nd ed. Springer: New York, NY, USA.

Busetto L, Ranghetti L, 2016. MODIStsp: An R package for automatic preprocessing of MODIS Land Products time series. Comput Geosci 97:40-8.

Ceccato P, Connor SJ, Jeanne I, Thomson MC, 2005. Application of geographical information systems and remote sensing technologies for assessing and monitoring malaria risk. Parassitologia 47:81-96.

Ceccato P, Ghebremeskel T, Jaiteh M, Graves PM, Levy M, Ghebreselassie S, Ogbamariam A, Barnston AG, Bell M, del Corral J, Connor SJ, Fesseha I, Brantly EP, Thomson MC, 2007. Malaria stratification, climate, and epidemic early warning in Eritrea. Am J Trop Med Hyg 77:61-8.

Craig MH, Kleinschmidt I, Nawn JB, Le Sueur D, Sharp BL, 2004. Exploring 30 years of malaria case data in KwaZulu-Natal, South Africa: Part I. The impact of climatic factors. Trop Med Int Health 9:1247-57.

Craig MH, Snow RW, Le Sueur D, 1999. A climate-based distribution model of malaria transmission in Sub-Saharan Africa. Parasitol Today 15:105-11.

Ferrão JL, Mendes JM, Painho M, 2017. Modelling the influence of climate on malaria occurrence in Chimoio Municipality, Mozambique. Parasit Vectors 10:260.

Gething P, Van Boeckel T, Smith D, Guerra C, Patil A, Snow R,

Hay S, 2011. Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. Parasit Vectors 4:92-108.

Govere JM, Durheim DN, Coetzee M, Hunt RH, La Grange JJ, 2007. Captures of mosquitoes of the *Anopheles gambiae* complex (*Diptera: Culicidae*) in the Lowveld Region of Mpumalanga Province, South Africa. Afr Entomol 8:91-9.

Hay SI, Snow RW, Rogers DJ, 1998. Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data. Trans R Soc Trop Med Hyg 92:12-20.

Huete A, Didan K, Miura T, Rodriguez EP, Gao X, Ferreira LG, 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ 83:195-213.

Hyndman RJ, Athanasopoulos G, 2017. Forecasting: principles and practice. Available from: https://www.otexts.org/book/fpp Accessed: 20 July 2017.

Jensen JR, 2007. Remote Sensing of the environment - an earth resource perspective. 2nd ed. Prentice Hall, Upper Saddle River, NJ.

Julie AC, Aniset K, Mulenga M, Clive S, Gregory EG, 2010. Identifying malaria vector breeding habitats with Remote Sensing data and terrain-based landscape indices in Zambia. Int J Health Geogr 9:58-69.

Kleinschmidt I, Sharp BL, Clarke GP, Curtis B, Fraser C, 2001. Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in Kwa-Zulu Natal, South Africa. Am J Epidemiol 153:1213-21.

Komen K, Olwoch J, Rautenbach H, Botai J, Adebayo A, 2015. Long-run relative importance of temperature as the main driver to malaria transmission in Limpopo Province, South Africa: a simple econometric approach. EcoHealth 12:131-43.

Ljung GM, Box GEP, 1978. On a measure of lack of fit in time series models. Biometrika 65:297-303.

Machault V, Vignolles C, Borchi F, Vounatsou P, Pages F, Briolant S, Lacaux J, Rogier C, 2011. The use of remotely sensed environmental data in the study of malaria. Geospat Health 5:151-68.

Matsushita B, Yang W, Chen J, Onda Y, Qiu GY, 2007. Sensitivity of the enhanced vegetation index (EVI) and normalized difference vegetation index (NDVI) to topographic effects: a case study in high-density cypress forest. Sensors 7:636-2651.

McFeeters SK, 2013. Using the normalized difference water index (NDWI) within a geographic information system to detect swimming pools for mosquito abatement: a practical approach. Remote Sensing 5:3544-61.

Midekisa A, Senay G, Henebry GM, Semuniguse P, Wimberly MC, 2012. Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. Malar J 11:165-74.

Mushinzimana E, Munga S, Minakawa N, Li L, Feng C, Bian L, Kitron U, Schmidt C, Beck L, Zhou G, Githeko A, Yan G, 2006. Landscape determinants and remote sensing of mosquito larval habitats in the high land of Kenya. Malar J 5:13-23.

NASA, 2017. NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) center, Sioux Falls, South Dakota. Available from: https://lpdaac.usgs.gov/node/844 accessed: 20 April 2017.

Ngomane L, de Jager C, 2012. Changes in malaria morbidity and mortality in Mpumalanga Province, South Africa (2001-2009): a retrospective study. Malar J 11:19-29.

Peter D, Silvia P, 2012. ARIMA vs ARIMAX which approach is better to analyze and forecast macroeconomic time series ? Proceedings of 30th International Conference Mathematical Methods in Economics. Karviná, Czech Republic. pp 136-140.

R Core Team, 2016. A language and environment for statistical computing; R foundation for statistical computing: Vienna, Austria. Available from: https://www.R-project.org/ Accessed: 26 October 2017.

Sarfraz MS, Tripathi NK, Faruque FS, Bajwa UI, Kitamoto A, Souris M, 2014. Mapping urban and peri-urban breeding habitats of *Aedes* mosquitoes using a fuzzy analytical hierarchical process based on climatic and physical parameters. Geospat Health 8:685-97.

Silal SP, Little F, Barnes KI, White LJ, 2014. Towards malaria elimination in Mpumalanga, South Africa: a population-level mathematical modelling approach. Malar J 13:297-310.

South Africa National Department of Health, 2011. National Malaria Report. Pretoria, South Africa.

Spitzer M, Wildenhain J, Rappsilber J, Tyers M, 2014. BoxPlotR: web tool for generation of box plots. Nature Methods 11:121-2.

STATS SA (Statistics South Africa), 2016. Community survey. South African statistics. Pretoria, South Africa. Available from: http://cs2016.statssa.gov.za/wp-content/uploads/2016/06/CS-2016-Provinces-at-a-glance.pdf

STATS SA (Statistics South Africa), 2017. Mid-year population estimates. South African statistics, Pretoria, South Africa. Available from: http://www.statssa.gov.za/publications/P0302/P03022017.pdf

Teklehaimanot HD, Lipsitch M, Teklehaimanot A, Schwartz J, 2004. Weather-based prediction of *Plasmodium falciparum* malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. Malar J 3:41.

Thomas CJ, Cross DE, Bøgh C, 2013. Landscape movements of *Anopheles gambiae* malaria vector mosquitoes in rural Gambia. PLoS One 8:e68679.

TRMM (Tropical Rainfall Measuring Mission), 2011. TRMM (TMPA/3B43) Rainfall estimate L3 1 month 0.25 degree x 0.25 degree V7, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC). Available from: doi: 10.5067/TRMM/TMPA/MONTH/7 Accessed: 11 April 2017.

Weiss DJ, Bhatt, Mappin B, Van Boeckel TP, Smith DL, Hay SI, Gething PW, 2014. Air temperature suitability for *Plasmodium falciparum* malaria transmission in Africa 2000-2012: a high-resolution spatiotemporal prediction. Malar J 13:171.