UTRECHT UNIVERSITY
Earth Science

Honglei Zhang 6752608

SURFACE WATER SIMULATION WITH THE PYCATCH MODEL AND VERIFICATION WITH REMOTE SENSING DATA FOR MALARIA PREDICTION

MSC Research Thesis

1st Supervisor
Prof. dr. Derek Karssenberg
2st Supervisor
dr. Edwin Sutanudjaja

2021 Year

Statement of originality of the MSc thesis

I declare that:

1. this is an original report, which is entirely my own work,

2. where I have made use of the ideas of other writers, I have acknowledged the source in all instances,

3. where I have used any diagram or visuals I have acknowledged the source in all instances,

4. this report has not and will not be submitted elsewhere for academic assessment in any other academic course.

Student data:

Name: Honglei Zhang

Registration number: 6752608

Date: 9 June, 2021

Signature: Honglei Zhang

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisors Prof.Dr.Derek Karssenberg and Dr.Edwin Sutanudjaja for all their help in providing such an interesting project, setting up the fleets for running PyCatch and guiding where I should head. From them I learned not only professional hydrology knowledge and modeling skills, but also the qualities and perspectives required in my future academic field.

I'm grateful to all institutes and organizations for offering open-access data and tools, such as Copernicus, ECMWF, and ISRIC, laying the foundation of my master thesis project. Thanks to Dr. Oliver Schmitz for the authority of using fleets.

I would also like to mention all my friends accompanying me in this hard situation, extending their helping hand and brightening every single day. They are Pauline Colucci, Jiacheng Cai, Jingkun Dai, Tianyun Zong, Yihong Li, and Yuanhang Luo. Especially, Zhaopeng Liu always inspires me with his talented academic thoughts and stands by me supportively. His tips of being organized in a presentation really helps.

Many thanks to my parents for their unconditional love and support in finance, academic and mental health. Though all our communication is through video calls, their happy faces and positive attitude towards difficulties never fail to cheer me up. I also owe thanks to the staff of Chinese Embassy in the Netherlands for handing out free masks and personal hygiene products. They made me feel safe even in the corona situation.

Abstract

Malaria affects more than 200 million people in the world each year. It is increasingly restricted to tropical and subtropical areas, especially in Africa, resulting in lower dependence on air temperature, so more attention is paid to modeling malaria with surface water processes in recent years. The PyCatch model enables dynamically simulating the processes of interception, evapotranspiration, surface storage, infiltration, subsurface flow, and overland flow. To examine the PyCatch capability in simulating malaria-relevant surface water processes, we applied the PyCatch model to a small catchment and evaluated the model performance in discharge and soil moisture simulation. The results show the consistency between PyCatch hydrograph and other global hydrological models, and the PyCatch skillfulness in discharge simulation. The soil moisture simulated by PyCatch is stable and has evident seasonal and interannual variation patterns, despite some differences against the remote sensing data. Through linear regression, PyCatch is proved to be competent for malaria burden prediction.

KEYWORDS: PyCatch, model, malaria, soil moisture

Table of Contents

1. Introduction

Malaria is one of the first identified infectious diseases and caused 405000 deaths worldwide in 2018, among which 94% occurred in Africa [1]. Children under five years old are most vulnerable. In 2018, 93% of the 228 million cases worldwide occurred in Africa. Understanding the climatic suitability and distribution of malaria is vital for making intervention measures and preventing future malaria outbreaks. However, since malaria transmission is highly dependent on various climatic, environmental, hydrological, and anthropogenic factors, malaria distribution and incidence prediction remain challenging[2].

Malaria transmission nature is complicated because of the complicity of interaction between the parasite, the vector, and the host. Malaria is caused by the Plasmodium parasite, including the species P. vivax, P. malaria, P. falciparum, P. ovale, and P. knowlesi [3, 4]. Among them, P. falciparum is a fatal pathogen. Female anopheles search for a blood meal to develop eggs. After biting an infectious human, they are likely to get infected. Then the sporogonic processes of Plasmodium start in the mosquito midgut, and these mosquitoes get infectious. In the next gonotrophic cycle of mosquitoes, mosquitoes inject saliva with around 10 to 100 sporozoites, a spore-like stage of Plasmodium, into the human's blood system [5]. Then the person gets the disease with a possibility. After that, they develop into merozoites, break out of the liver, and invade blood cells to grow and divide further. Anopheles gambiae is considered the most efficient malaria transmission vector, especially in sub-Saharan Africa [6, 7]. The lifecycle of subadult mosquitoes also affects malaria transmission. Female mosquitoes lay eggs on the water surface, and most eggs hatch and become larvae within 2-3 days, depending on temperature. At last, the larvae become pupae that will finally turn into mosquitoes.

The climatic suitability mainly includes suitable temperature and sufficient water availability [8]. The air temperature mainly affects the adult mosquito's survival and gonotrophic cycle duration. In contrast, the water temperature significantly impacts the Plasmodium parasite's sporogonic duration and the larvae survival rate[9]. Generally, higher temperature accelerates the gonotrophic cycle and sporogonic cycle, but a too high temperature adds to the mortality of larvae and adult mosquitoes. The most suitable temperature for vector reproduction and malaria transmission is unclear because of the differences between laboratory data and the actual situation in the field and the complication of micro-habitats caused by small water bodies and vegetation shading. However, associated with implementing mitigation measures and improving control tools, the malaria spatial distribution is increasingly restricted to the tropical and subtropical areas [10]. In consequence, the malaria incidence is decreasingly sensitive to temperature. Instead, malaria is inextricably related to temporal and spatial water distribution [11].

Both long-lasting and temporary water availabilities contribute to malaria transmission by generating suitable habitats and attracting more hosts. Female vector mosquitoes lay eggs in water, and the immature development stages rely on the aquatic environment. Consequently, arid regions or drought periods are usually free of malaria because of the lack of Anopheles habitats. Besides, hydro-processes also affect malaria transmission by determining vegetation type and cover. Vegetation acts as resting places and refuges for mosquitoes [12], and it attracts host animals that mosquitoes can feed on. Therefore, the abundant precipitation intensity and duration facilitate vegetation growth and then indirectly boost malaria prevalence. Besides, water accessibility is one of the prime conditions for regions being populated, and the gathering of both vectors and human hosts results in high malaria incidence. Instead of large lakes and rivers, P. falciparum mainly lays eggs in temporary water pits and small pools, even in hoof prints and wheel ruts. Whether these temporary tiny water bodies contribute to malaria incidence depends on whether they last longer than mosquito gonotrophic cycle. The lifespan and shrink rate of these water bodies are strongly influenced by hydrological processes, such as evaporation, infiltration, and runoff, which vary over time and space. Thus, hydrological processes, especially surface water processes, need to be considered to study malaria distribution and incidence. Also, cropland irrigation and corresponding dams prolong the wet period, thus aggravating malaria transmission [2].

After the late 1990s and before the 2010s, many models have been built to map the malaria distribution. They have a coarse resolution and a large scale, either a global or continental scale. As temperature and precipitation are major determiners for malaria vector development, some climatic models are used for malaria prevalence forecasts. Martens [13] includes a General Circulation Model to explore the climate change impact on malaria epidemic potential, and the climate change is represented by rainfall and temperature variation. However, the mosquito survival probability used in the model is arbitrary. Similarly, global coupled ocean-atmosphere climate models were used for interannual climate variability simulation, through which anomalous malaria incidence can be predicted [14].

Many of these models use the regression method, especially logistic regression, and emphasize the temperature impacts on malaria incidence. The relationship between environmental conditions and malaria vector population was studied. Craig [15] adapted temperature and the rainy season length into a fuzzy logistic model of malaria transmission diagnosis, and this is one of the first numerical models for mapping malaria risk. Hoshen [16] modeled both immature and mature vector dynamics and the infected host population, within which a daily stepped 10-day accumulated rainfall was used. Besides, distance to water and normalized difference vegetation index (NDVI) was also utilized in a logistic model to map malaria [17]. Omumbo [18] firstly used satellite data to generate monthly cold cloud duration (CCD) images as a proxy of rainfall, and a supervised classification method was implemented to classify the malaria burden. Based on former studies, Rogers used NDVI and CCD to map the distribution and abundance of four vector species. Then the maps were combined with the entomological inoculation rate to evaluate the malaria burden [19]. Moreover, the calibration was done according to observed precipitation [14]. In these models, vector dependence on water is highly simplified, and only vegetation index, rainfall, or its proxy, rather than detailed hydrological processes, were included. It is proved that malaria transmission is much more sensitive to the rainfall threshold than the thermal response [8].

Recently, some researches focus more on surface water hydrological processes for malaria modeling. There is excellent progress in improving the spatio-temporal resolution and modeling the malaria incidence and transmission at a village scale. New data sets, new technical methods like GIS and remote sensing images are incorporated for malaria or vector modeling. Shaman [20] ran a hydrological model with meteorological inputs to get the dynamic surface humidity and correlate spatial and temporal varying surface humidity with Anopheles abundance using a logistic regression model. Though only simple hydrological components were included, the model paid close attention to water accessibility. Bomblies [11] built a model called HYDREMATS, incorporating precipitation, soil moisture profiles, and runoff to simulate hourly surface water body locations and soil moisture fields, and then input them into the entomology model to get malaria prevalence. This agent-based model gained great attention and was extended to include the immune process. HYDREMATS was used to assess the impact of environmental management on malaria control and build potential early warning systems [21–23]. However, model adaption and calibration are needed because of the unique hydrological processes in different villages [24]. Another hydrological model - VECTRI formulates both permanent water bodies and temporary pond dynamics determined by surface water runoff, infiltration, evaporation, and pond overflow [25]. To be used in regions without enough rain-gauge data, the VECTRI model was revised to incorporate satellite data, and the performance was compared with the HYDREMATS model [11, 25, 26].

Though more attention was paid to modeling malaria with hydrology at a fine scale, their application to the whole world is complicated. As HYDREMATS incorporates a land surface scheme with six soil layers, Bombies uses time domain reflectometry (TDR) soil moisture profiles to model vertical hydrological processes, making it costly and labor-intensive to apply HYDREMATS globally [11]. The hydrological part of the VECTRI model is simple enough but quite empirical. Thus Asare [26] provides a method for VECTRI parametrization, but the calibration in each grid cell remains a challenge. Besides, VECTRI neglects the vegetation effects and regards the infiltration and evaporation constant in space and time. For easier application at a global scale, model inputs should be available globally; the hydrological outputs should be consistent with observed data or other models; the model should be robust enough so that the hydrological processes will be stable and trustable. Lana-Renault and Karssenberg [27] built a new component-based hydrological catchment model-PyCatch that simulates detailed interception processes, evapotranspiration, surface storage, infiltration, and surface overland flow. However, the model's potential to be used for global malaria modeling remains unclear.

The main objective of this research is to evaluate the performance of the PyCatch model regarding its capability to predict the surface water occurrence and serve as a predictor of Malaria incidence rate, at a global scale. The main question is how the model performs in simulating malaria-relevant hydrological processes. To realize the main objective, we divide the main question into five sub-questions and provide possible methods. Firstly, what global datasets are available at the spatio-temporal resolution required by PyCatch, and what is the quality of these data? Secondly, how well does the model perform in predicting the river discharge? Thirdly, what is the PyCatch capability of predicting soil moisture and surface water occurrence? Fourthly, what is the relationship between soil moisture and malaria? At last, how capable would PyCatch be to predict surface water globally, and how feasible would it be given the current data availability and hardware availability?

In this research, we applied PyCatch to the Okole catchment in Uganda. We combined the reanalysis data, remote sensing data, model outputs and literature data as PyCatch inputs, all of which are globally available. Then we compared the hydrograph simulated by PyCatch with four other global hydrological models and score the PyCatch skillfulness in simulating discharge. To answer the third question, we applied sensitivity analysis to test the model stability, analyzed the PyCatch ability in reflecting soil moisture variation, and then calculated the surface water occurrence to compare it with remote sensing data. Regarding the malaria modeling, we applied a linear regression between the soil moisture and the parasite rate. Lastly, We timed the model running on our study area and calculated the time consumption if PyCatch is applied globally.

2. Data and Methods

We create a system to examine the PyCatch capability regarding to simulate malaria-relevant hydrological processes (figure 1) and apply it to the Okole catchment. The PyCatch model structure and relevant input data is introduced first, followed by the model evaluation in aspects of discharge and soil moisture simulation. We give a brief introduction of malaria modeling method at the end of this chapter.

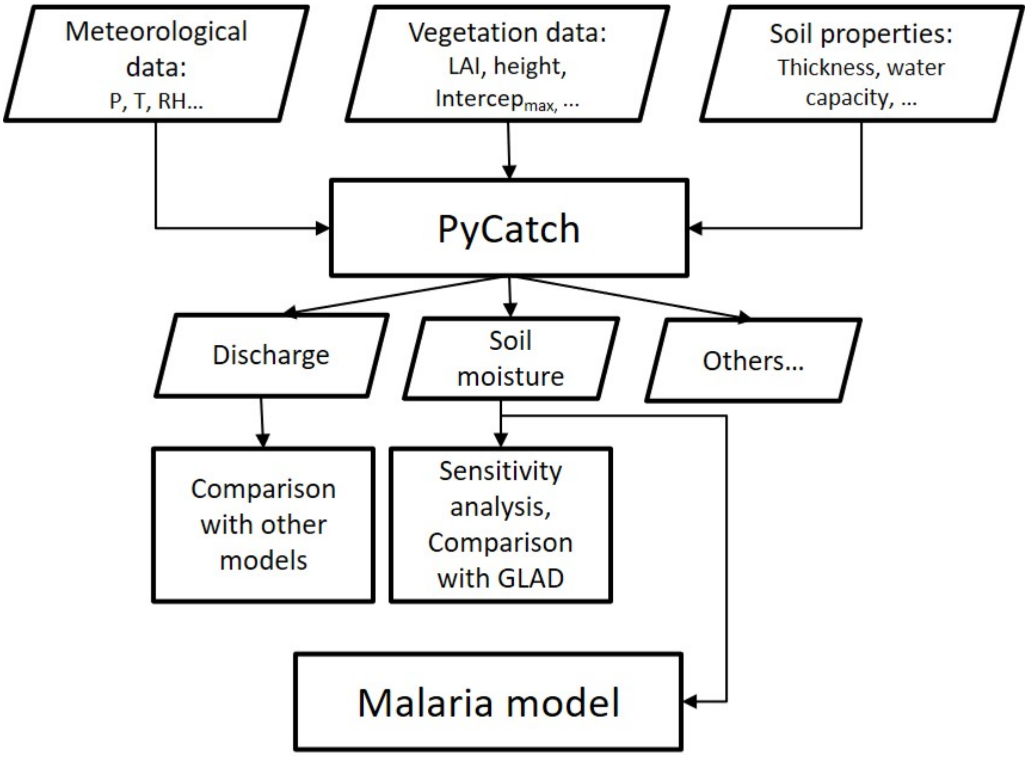


Figure 1. *A model evaluation system in terms of simulating malaria-relevant surface water processes.*

2.1 The PyCatch model

The PyCatch model is constructed by Lana-Renault and Karssenberg in 2013, using the easy-to-learn environmental programming language PCRaster Python, a powerful tool for dynamic model construction, requiring low programming skills [27]. Users can cut down the programming time and technical details, improve code reusability and avoid coding errors compared with using system programming languages [28]. Particularly, the dynamic framework interpolated in PCRaster allows users to establish interactive temporal models to simulate dynamical process easily [29].

PyCatch simplifies complicated hydrological processes, with components of interception processes, evapotranspiration, infiltration, surface storage, subsurface flow, overland flow and their interactions(see figure 2). In space, PyCatch is distributed, dividing the study area into square cells interacting with neighbors, enabling both vertical and lateral flow simulation. The spatial resolution is 100m, which is moderate for surface water simulation that aims at studying mosquito richness. The spatially heterogeneous hydrological components determined by vegetation coverage, topography, and meteorology within the catchment can be simulated with the PyCatch model at satisfactory resolution.

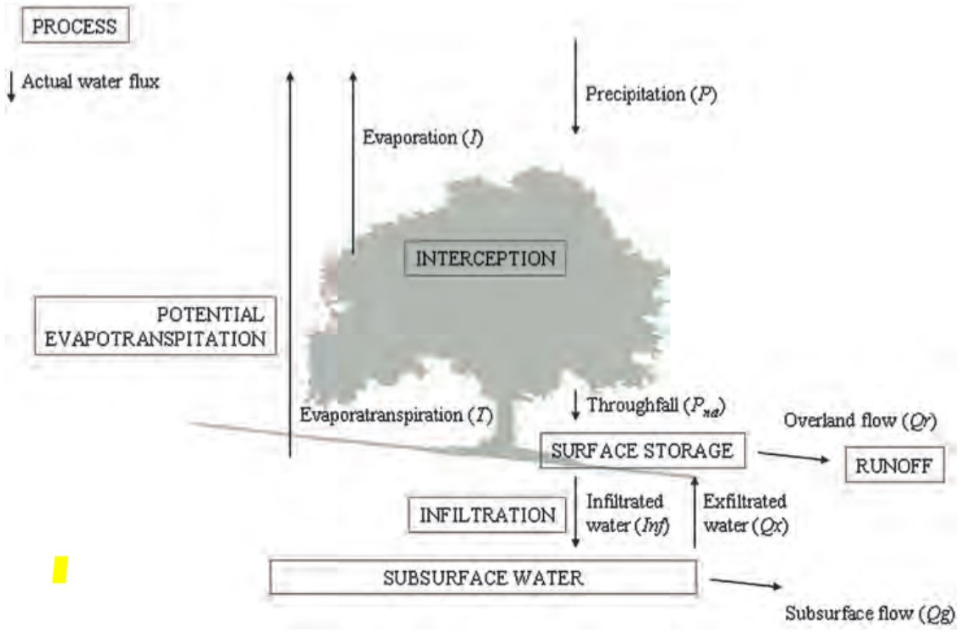


Figure 2. *Schematic overview of the hydrological model PyCatch [27].*

In the time domain, PyCatch is a dynamic model that uses a transition function to update state variables at time t from previous time step $t - 1$, enabling to reconstruct the environmental complexity with relatively simple transition functions (see figure 3). The temporal resolution is one hour, so it allows to simulate detailed hydrological processes. As surface water processes, combined with temperature, determine the gonotrophic cycle duration (approximate one week), the high time resolution helps to understand mosquito vector amount dynamics. PyCatch is comprised of three main process-based distributed hydrological submodels, interception and evapotranspiration model, soil water budget and subsurface flow model, and overland flow model, where all components interacts with each other.

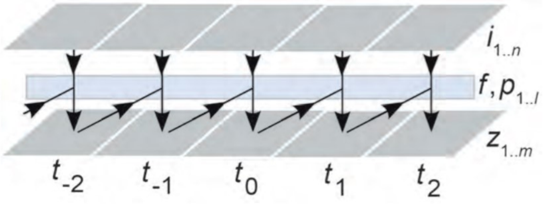


Figure 3. *A dynamic model in two dimensional space; $z_{1..m}$ are state variables, f the transition functions, $i_{1..n}$ are inputs, and $p_{1..l}$, the parameters of f [27].*

2.1.1 Interception and evapotranspiration

Precipitation directly reaches the soil, with a part intercepted by vegetation. The amount of interception part of rainfall depends on the total precipitation and canopy gaps, which is further determined by the leaf area index (LAI [-]). The interception part (P_{int} [m h^{-1}]) and the canopy gap fraction (f_{gap} [-]):

$$P_{\text{int}} = (1 - f_{\text{gap}}) * P \tag{2.1}$$

$$f_{\text{gap}} = e^{-kLAI} \tag{2.2}$$

Where:

k [-]: light extinction coefficient

With intercepted water in the canopy, the leaf surface is saturated, so energy-limit evaporation occurs with no transpiration. The evapotranspiration is calculated with the Penman-Monteith equation. The potential evapotranspiration (E_p [m h^{-1}]) is influenced by the radiation, the latent heat, and meteorology variables:

$$E_p = \frac{1}{\lambda}\left(\frac{R_n \cdot \delta + \rho_a \cdot c_p \cdot v/r_a}{\delta + \gamma(1 + r_c/r_a)}\right) \tag{2.3}$$

Where:

λ [J kg^{-1}]: the latent heat coefficient
R_n [W m^{-2}]: net radiation
δ [Pa K^{-1}]: the gradient of the vapour pressure-temperature curve
ρ_a [kg m^{-3}]: mean air density
c_p [J kg^{-1} K^{-1}]: specific heat of the air at constant pressure
v [Pa]: the deficit of air vapour pressure
γ [Pa K^{-1}]: the psychrometric constant
r_c [h m^{-1}]: canopy resistances
r_a [h m^{-1}]: aerodynamic resistances

Soil potential evapotranspiration (ET [m h^{-1}]) and soil actual evapotranspiration (T [m h^{-1}]):

$$ET = E_p \cdot (E_0 - I)/E_0 \tag{2.4}$$

$$T = min(ET, W) \tag{2.5}$$

Where:
W [m h^{-1}]: available water amount in the soil

When there is no water in the canopy, transpiration depends on potential evapotranspiration and vegetation stomatal conductance determined by soil water content.

2.1.2 soil water budget and subsurface flow

The potential infiltration (Inf_p [m h^{-1}]) and actual infiltration (Inf [m h^{-1}]) are respectively:

$$Inf_p = K_{inf} \cdot \left(\frac{\phi \cdot \triangle\theta}{F} + 1 \right) \tag{2.6}$$

$$Inf = min\left(K_{inf}, Qs_t\right) \tag{2.7}$$

Where:
K_{inf} [m h^{-1}]: saturated conductivity
ϕ [m]: wetting front capillary pressure head
$\triangle\theta$ [-]: available pore space
F [m]: commutative infiltration
Qs_t [m h^{-1}]: available water for infiltration

Soil water storage at time step t depends on the water storage in the last time step, soil evapotranspiration, infiltration, subsurface flow from the upstream neighboring cells and to the downstream cells:

$$G_t = G_{t-1} - T + Inf + Qg_i - Qg_0 \tag{2.8}$$

where:

Qg_i[m h^{-1}]: the subsurface inflow

Qg_0[m h^{-1}]: the subsurface outflow

G[m]: the storage of soil water

The subsurface flow to the downstream area is calculated with Darcy's law:

$$Qg = K_{sat} \cdot L \cdot \alpha \cdot G \tag{2.9}$$

where:

K_{sat}[m h^{-1}]: the saturated conductivity

L[m]:the cell length

α[-]: the slope to the downstream cell

2.1.3 Overland flow

Overland flow is calculated with the water balance on the earth surface [m h^{-1}]. It depends on not only the state of net precipitation, infiltration, and exfiltrated water of the current time step but also the overland flow of the last time step:

$$Qr = P_{net} + Qx + Qs_{t-1} - Inf \tag{2.10}$$

Where:

Qx[m h^{-1}]: exfiltrated water

2.1.4 Study area

The Okole catchment (32.68°E,2.23°N) in the northern plateau of Uganda is selected as the study area, as shown in figure 4 circled with the red line. Figure 5 shows the digital elevation model (DEM) and the land cover classifications of the Okole catchment. A tropical savanna climate characterizes the Okole river catchment, where the average rainfall is around 1200 mm-1400 mm with a bimodal precipitation pattern. The two dry seasons are from June to October and December to February next year. The average temperature is between 23°C and 27°C, close to the most suitable malaria transmission temperature 23°C-28°C, based on the work of Mordecai [30]. The Okole river starts from the north of Lira city, flows for more than 90 km to the west, and joins the Victoria Nile.

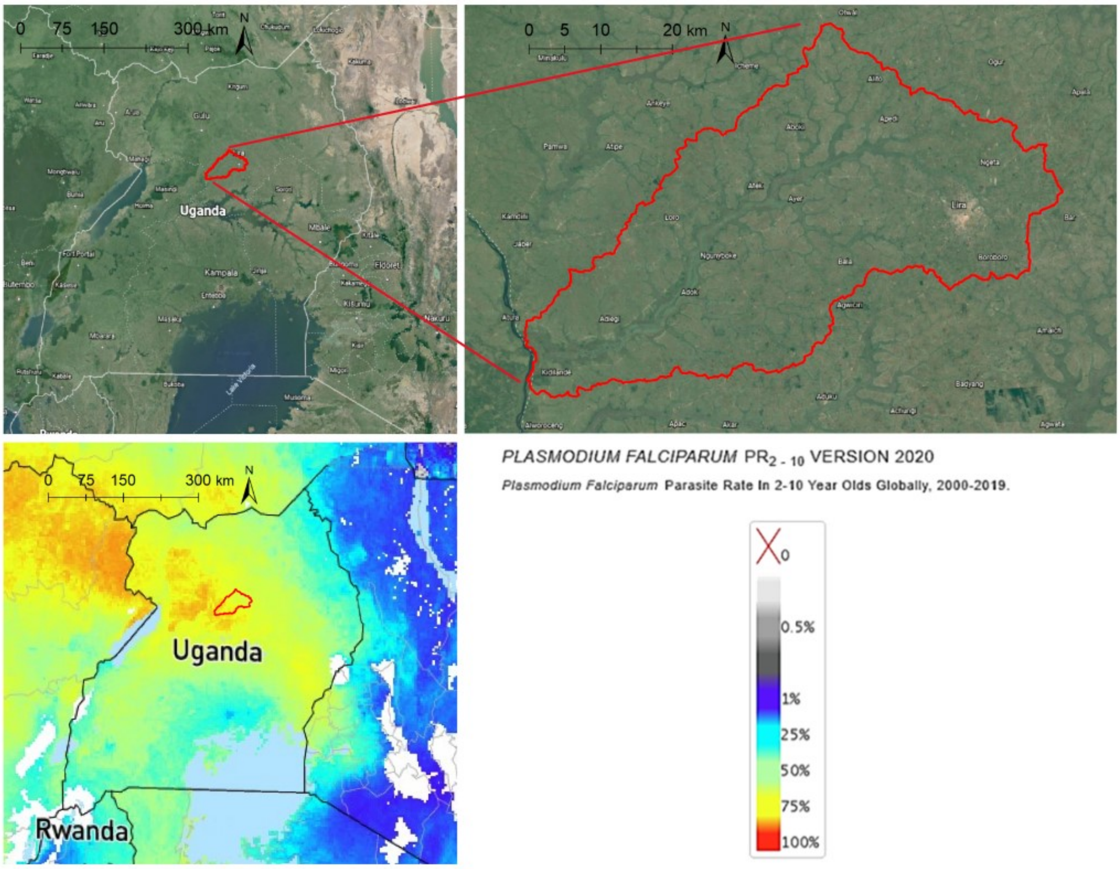

0 75 150 300 km
N
Uganda
Kampala
0 5 10 20 km
N
Lira
0 75 150 300 km
N
Uganda
Rwanda
PLASMODIUM FALCIPARUM PR$_{2-10}$ VERSION 2020
Plasmodium Falciparum Parasite Rate In 2-10 Year Olds Globally, 2000-2019.
0
0.5%
1%
25%
50%
75%
100%

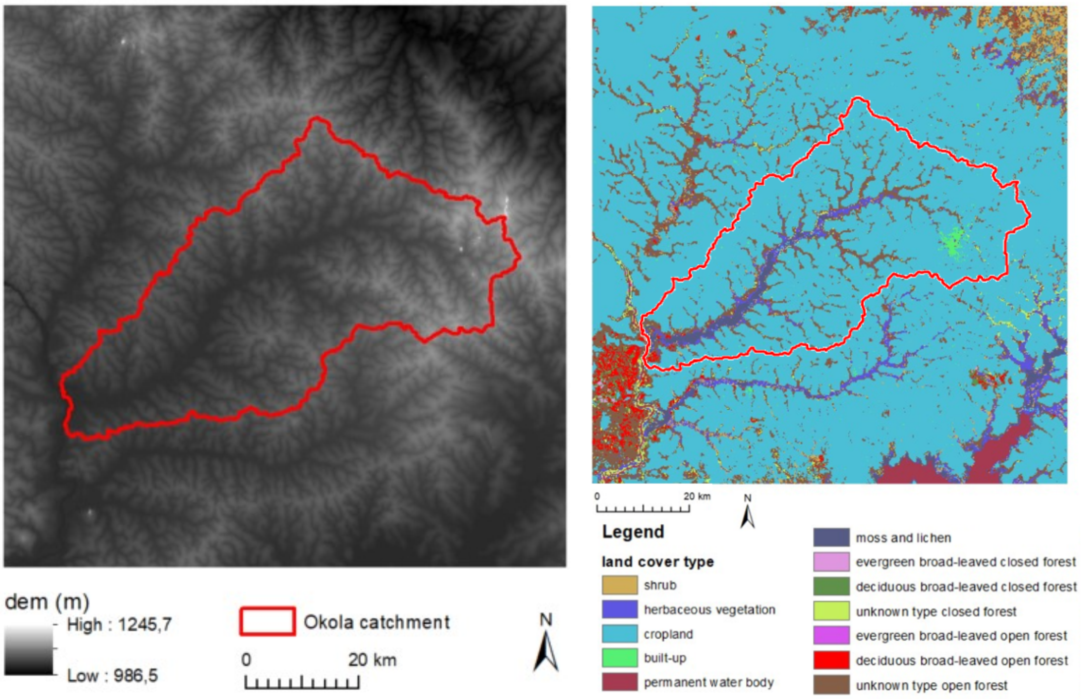
Figure 4. *Study area: the Okole catchment in the north of Uganda (top); Plasmodium falciparum incidence (bottom).*

The catchment is mainly comprised of cropland, taking up more than 70%. Herbaceous vegetation, herbaceous wetland, and shrub are distributed along the streams. Spotted deciduous broad-leaved open forest randomly appears near the rivers. For the malaria modeling part, described in sections 2.4, we expand the study area to the whole extent shown in figure 5 to cover more malaria data. The expanded study area is around 100 km*100 km surrounding the Okole catchment, so there are 1000*1000 cells.

2.2 Data source

The inputs of the model include catchment characteristics, meteorological data, soil variables, and vegetation variables. As for meteorological data, ERA5-Land hourly data from 1981 to the present [33] is used for temperature, precipitation, wind speed, radiation, and relative humidity that is calculated with temperature and dew point temperature [34]. Table 2 shows an overview of all model inputs and their data source.

ERA5-Land data is a reanalysis data, with a resolution of 0.1*0.1 degrees, approximately equivalent to 9km at the equator. The temporal resolution is one hour, the same as the



(a) (b)

Figure 5. Digital Elevation Model [31] (left) and land cover type classification [32] (right) of the study area

model resolution. Meteorological data are spatially averaged and inputted into PyCatch as spatially lumped and temporally dynamic variables. Though the ERA5-Land precipitation has a enhanced spatial resolution [35], it is still much coarser than the observed data. Therefore, we calibrate the precipitation according to the annual total rainfall amount of the Okole catchment, and the hourly precipitation after calibration is:

$$PC_t = P_t * (5\overline{P_{year}}/\sum_{t=0}^{n} P_t) \tag{2.11}$$

where:

P_t: the hourly precipitation at time step t

$\overline{P_{year}}$: long-term average annual precipitation

n: the total time step of the 5 year study period

We calculate the relative humidity as a ratio of the actual air vapor pressure to the saturated vapor pressure.

$$e_s = 6.107 * exp\frac{Ta \cdot 17.4}{239 + Ta} \tag{2.12}$$

$$VPA = 6.107 \cdot exp\frac{Td \cdot 17.4}{239 + Td} \tag{2.13}$$

where:

e_s: the saturated vapour pressure

Ta: the air temperature

VPA: the actual vapour pressure in the air

Td: the dew point temperature in °C

Then, the relative humidity is:

$$RH(\%) = 100 \cdot \frac{VPA}{e_s} \tag{2.14}$$

The African soil property maps at 250m resolution produced by ISRIC-World Soil Information are used for the soil inputs [36]. We use the derived available soil water capacity with FC = pF 2.5 as the field capacity and the water capacity with FC = pF 4.2 as the wilting point. The saturated water content is used as the porosity. All above data include seven maps at depths of 0cm, 5cm, 15cm, 30cm, 60cm, and 100cm, which are averaged according to soil thickness. Also, the 250m data needs to be resampled into the model resolution 100m. Considering the Okole catchment is on a plateau, we set the regolith thickness to be 10m, uniform within the study area.

For the vegetation variables, LAI, vegetation albedo, vegetation height, and vegetation stomatal conductance are needed. The 16-day MODIS-MCD15A2Hv006 data at 500m resolution [37] is resampled to the model resolution, and the average LAI of 2005-2009 is calculated as PyCatch LAI input. The albedo input is also from the ERA5-Land hourly data using a similar processing method as LAI. The vegetation height is computed by combining the land cover classification data shown in figure5 [32] and the vegetation height of each land cover type (see table 1) [38]. As the built-up area and permanent water bodies are beyond the PyCatch simulation, the vegetation height in these areas is set to be a minimal value, approximately equal to 0. The land cover classification data is also used to get the forest area. The vegetation stomatal conductance for the forested area and the non-forested area is set to be 0.0032 and 0.0064, respectively [27]. Different from meteorological data, vegetation condition data, as well as soil property data, are processed as spatially distributed and temporally static inputs.

Table 1. *Vegetation height of different land covers*

Land cover classification	**Vegetation height(m)**
Shrub	1.0
Herbaceous vegetation	0.3
Cropland	1.0
Built-up	0.000001
Permanent water bodies	0.000001
Herbaceous wetland	0.3
Evergreen broad-leaved open or closed forest	6.8
Deciduous broad-leaved open or closed forest	10.0
unknown type closed forest	8.4

Table 2. *Model inputs and relevant ERA5-Land data*

Model inputs	**Data source**
Meteorological inputs: lumped and dynamic	
Incoming shortwave radiation [W m^{-2}]	ERA5-land: surface radiation downwards [J m^{-2}]
Air temperature [°C]	ERA5-land: 2m temperature [K]
Air relative humidity [-]	ERA5-land: 2m temperature [K] and dew point temperature [K]
Wind velocity [m s^{-1}]	ERA5-land: 10m u-component of wind and 10m v-component of wind [m s^{-1}]
Vegetation parameters: distributed and static	
LAI [-]	MODIS 16-day LAI data [-]
Albedo [-]	ERA5-land: forecast albedo [-]
Vegetation height	Copernicus Global Land Service land cover classification data and former literature [38]
Vegetation stomatal conductance	Literature data [27]
Soil parameters: distributed and static	
Soil water content at which root water uptake by the plant declines [-]	Literature data [27]
Soil water content at saturation [-]	ISRIC 250m soil maps: saturation soil water content
Soil water content at wilting point [-]	ISRIC 250m soil maps: derived available soil water capacity (volumetric fraction) with FC = pF 4.2
Soil water content at field capacity [-]	ISRIC 250m soil maps: derived available soil water capacity (volumetric fraction) with FC = pF 2.5
Others	
DEM	Multi-Error-Removed Improved-Terrain DEM

Continues...

Table 2 – *Continues...*

Model inputs	**Data source**
Flow direction map, catchment and stream	Derived from DEM
Forest area and Non-forest area	Derived from land cover classification map
Latitude, longitude	WGS-84

2.3 PyCatch evaluation methods

To examine the PyCatch capability in simulating surface water hydrological processes, 1) we compare the PyCatch hydrograph with other models and score the PyCatch skillfulness in discharge simulation; 2) we apply a sensitivity analysis of soil moisture to parameter changes and validate PyCatch soil moisture against remote sensing data.

2.3.1 Discharge comparison with other models

To evaluate the PyCatch performance, we compare the discharge from PyCatch with four global hydrological models. The discharge data is easy to access and can be used for model calibration. It is informative for the catchment hydrology behavior since it reflects how a catchment responds to the rainfall under specific topographic, soil, and vegetation characteristics. However, discharge is not as widely used as soil moisture in malaria modeling, so the PyCatch seasonal and interannual soil moisture data is also analyzed. The multi-year average water possibility is calculated with PyCatch hourly soil moisture data, and the result is compared with remote sensing data. To evaluate the model stability, we also calculate soil moisture sensitivity to changes in precipitation, interception, and regolith thickness.

The time step of the PyCatch model is 1 hour. To compare PyCatch with four other models, HTESSEL [39], LISFLOOD [40], WaterGAP3 [41], and PCR-GLOBWB, we calculate the monthly average PyCatch discharge from January 2005 to December 2009. The HTESSEL model divides the land surface into tiles of bare soil, low vegetation, high vegetation, foliage covered by water, shaded snow and exposed snow and describes the heat and water exchange between the atmosphere and the land surface. The climatic data are from ERA-40 [42] and Balsamo [39] corrects the precipitation with GLOBAL Precipitation Climatology Project data. WaterGAP3 contains a rainfall-runoff model, five water-used models and a water quality model. The Observed Atmospheric Climate Data Sets [43] are used as climate drivers of WaterGAP3. LISFLOOD aims to simulate the long-term water balance and the flood events of the mid or large catchments, based on the rainfall-runoff process and the reservoir rules. The meteorology data are from the EFAS-Meteo dataset [44, 45]. PCR-GLOBWB model simulates the water exchanges between soil, atmosphere and groundwater, as well as the human water demands and consumption [46]. The ERA40 and ERA-Interim data [47] are used as forcing inputs, and the precipitation data is calibrated according to the observed data.

The Pearson correlation method is used to validate how well the PyCatch discharge is related to four other models. The average discharge of HTESSEL, LISFLOOD, WaterGAP, PCR-GLOBWB model is calculated as the approximate observation discharge. We assess the PyCatch model against the approximate discharge with a regression method and two efficient scores, Nash-Sutcliffe efficiency score (NSE) [48] and Kling-Gupta skill score (KEG^*). The NSE equation is:

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_{sim}(t) - Q_{obs}(t))^2}{\sum_{t=1}^{T}(Q_{obs}(t) - \overline{Q_{obs}})^2} \tag{2.15}$$

(2.16)

where:

Q_{sim}: the simulated discharge

Q_{obs}: the observed discharge

$\overline{Q_{obs}}$: the mean observed discharge

When NSE<0, the model prediction is worse than the mean value of the observed data; when NSE equals 0, the model has the same explainability as the observed mean; when NSE>0, the model has a better explanatory power. Kling-Gupta efficiency score (KEG) and KEG^* score are widely used for evaluating hydrological model performance [48–50]. Similar to NSE, they reveal the model performance in aspects of correlation with observations, but they are also capable to evaluate the bias. When the skill score is larger than 0.5, the model is skillful. The equations for calculating KEG and $KEG^*(Q)$ are:

$$KGE = 1 - \sqrt{(\rho - 1)^2 + (\frac{\mu_{sim}}{\mu_{obs}} - 1)^2 + (\frac{\sigma_{sim}}{\sigma_{obs}} - 1)^2} \tag{2.17}$$

$$KGE^*(Q) = \frac{KGE(Q_{sim}) - KGE(\overline{Q_{obs}})}{1 - KGE(\overline{Q_{obs}})} \tag{2.18}$$

where:

ρ: the linear correlation between observations and simulations
μ_{obs}: the observation mean
μ_{sim}: the simulation mean
σ_{sim}: the standard deviation in simulations
σ_{obv}: the standard deviation in observations

In lack of observation data, we use the monthly approximate observation discharge of Okole catchment from 2005 to 2009 to calculate μ_{obs} and σ_{obv}.

Comparing PyCatch discharge with other models validates the accuracy of the PyCatch model. Nevertheless, discharge is not widely used in modeling malaria. Vectors require stable surface water for reproduction, but the discharge data does not provide temporary water stability information. Instead, soil moisture, NDVI, surface water store, pool size, and pool persistence are more frequently used as hydrological inputs. These variables are all somewhat related to soil moisture which PyCatch simulates. The monthly soil moisture of PyCatch is calculated, and the 5 years' seasonal variation is analyzed. The soil moisture is related to precipitation to see how much variation in soil moisture can be explained by precipitation variation in the PyCatch model.

2.3.2 Soil moisture stability

At first, we apply a sensitivity analysis to the PyCatch model. Precipitation, vegetation interception, and soil thickness are chosen as representative for climatic, vegetation, and soil properties, respectively, for sensitivity analysis. Thesis variables are set to be 80%, 95%, 100%, 105%, and 120% of original levels, the corresponding fluctuation percentage are -20%, -5%, 0%, +5% and +20%. The soil moisture changing percentage will be calculated:

$$smcp_{i,j} = (sm_{i,j} - sm_{i,0})/sm_{i,0} \tag{2.19}$$

$$\tag{2.20}$$

where:
i: the variable changed, i = precipitation, vegetation interception or soil thickness
j: changing percentage of a specific variable, $j = -20\%, -5\%, 0\%, +5\%$ or $+20\%$
$smcp_{i,j}$: the soil moisture changing percentage of variable i changing $j\%$
sm: soil moisture

The monthly mean water percent data from Global Surface Water Dynamics 1999-2018 [51] (GLAD) provides information for the discrimination of water from land, and the surface water percent is calculated according to the frequency of water observed. The monthly mean water percent, the time step number of surface water present divided by the total time step number, is used for validating the PyCatch model. For PyCatch outputs of every time step, each pixel's maximum soil moisture is regarded equal to the porosity fraction, so the pixel with soil moisture reaches the maximum value is saturated, thus classified as water area. The monthly water possibility is calculated as the number of saturated time steps divided by the total time step number in a month. Then we calculate the average monthly water possibility from 2005 to 2009 and compare it with the GLAD surface water percent. The PyCatch water possibility time series and GLAD monthly water possibility time series of Okole catchment are plotted in the same figure.

2.4 Malaria modeling methods

In this section, the malaria burden is represented with the P. falciparum parasite rate (pf_pr) [52], referred to as parasite rate in the following parts. As the Okole catchment covers only four pieces of parasite rate data, we expand the study area to the 100 km*100 km area to cover 24 more data. Still, the parasite rate data were collected in different time from 1985 to 2010, during which the malaria epidemic area was shrinking. Therefore, we use all parasite rate data in Africa to calculate the parasite rate decreasing trend (equation 2.21) in time and calibrate the 28 pieces of data to the level of the year 2009 (equation 2.22) so that the interannual variation trend is excluded.

$$pf\_pr_{oy} = k \cdot y + c \tag{2.21}$$

$$pf\_pr_{o2009} = (2009 - y) \cdot k + pf\_pr_{oy} \tag{2.22}$$

where:

pf_pr_{oy}: the observed P. falciparum parasite rate in the year y

k: the regression coefficient

c: the regression interception

We take the average soil moisture (sm_a) of all time steps, and get the soil moisture value of all parasite data from the soil moisture map. Subsequently, we apply a linear regression between parasite rate data and corresponding average soil moisture. The parasite rate changes from the year 2005 and 2009 are calculated with the yearly-averaged soil moisture spatial data accordingly.

3. Results

3.1 Hydrograph comparison

We compared the PyCatch discharge with four other models. Figure 6 shows the monthly discharge of the Okole catchment from January 2005 to December 2009 simulated by PyCatch and four comparison models, HTESSEL, LISFLOOD, WaterGAP3, and PCR-GLOBWB. In spite of some gaps, the five hydrographs are consistent with each other, with two peak discharges in May and October and the lowest discharge in January. Compared with the HTESSEL model, PyCatch gets a minor peak discharge and a much larger base flow. WaterGAP3 hydrograph coincides nicely with PyCatch only after September 2007, with a smooth discharge variation before October 2006 and a sudden peak after that. PCR-GLOBWB and LISFLOOD hydrographs strongly agree with the PyCatch, except for the year 2007, which is also indicated in table 3. In 2007, the PyCatch discharge behaves comparably to other years, but PCR-GLOBWB and HTESSEL yield an abnormal autumn peak; for WaterGAP3, the autumn peak in 2006 merges into the spring peak in 2007, producing an abnormal large flood in February that is usually a dry season.

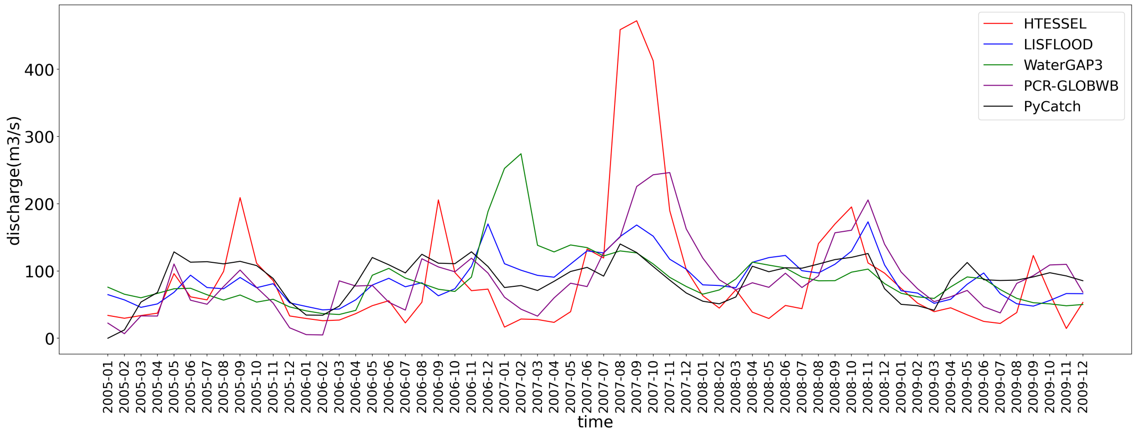


Figure 6. *Comparison of monthly-averaged discharge time series of PyCatch model with HTESSEL, LISFLOOD, WaterGAP3, PCR-GLOBWB model over the Okole catchment from 2005 to 2009.*

Furthermore, figure 7 illustrates the average Okole river discharge series of four comparison models from 2005 to 2009, where the bimodal pattern also appears. PyCatch discharge has a good consistency with the average discharge time series, but it is more evenly distributed among years.

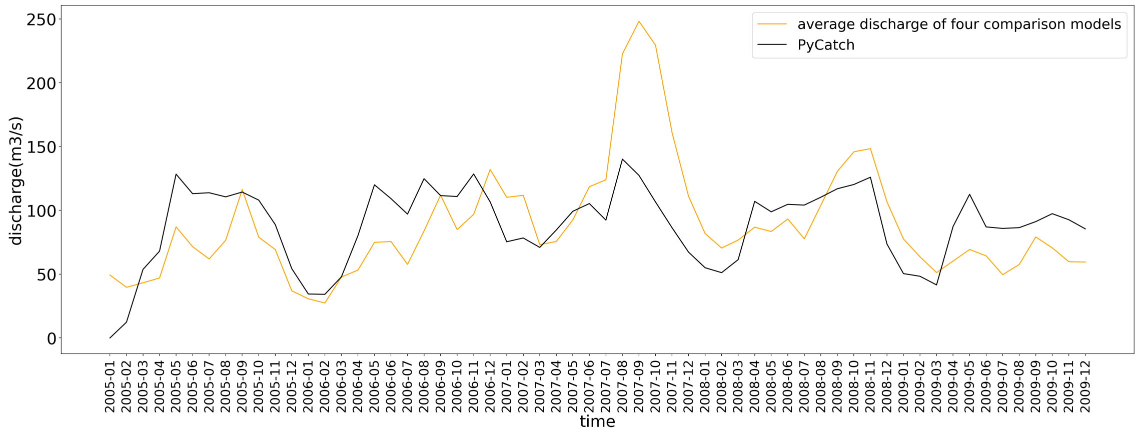


Figure 7. *Average monthly discharge of HTESSEL, LISFLOOD, WaterGAP3 and PCR-GLOBWB model of the Okole catchment from 2005 to 2009.*

Table 3. Pearson correlation between monthly Okola catchment discharge from PyCatch model and four comparison models during 2005-2009

model name	HTESSEL	LISFLOOD	WaterGAP3	PCR-GLOBWB	Average
Pearson correlation with PyCatch	0.450	0.544	0.230	0.500	0.551
P-value	3.10e-04	7.02e-05	0.077	4.78e-05	5.17e-06

Table 3 shows the Pearson correlation between 5-year monthly discharge results of PyCatch and four comparing models. The PyCatch discharge is most consistently correlated to the LISFLOOD discharge in four comparison models (largest Pearson correlation coefficient and smallest P-value). WaterGAP3 discharge is the only model result insignificantly correlated to the PyCatch discharge (P-value > 0.05). After taking the average of comparison models, the correlation between PyCatch discharge and comparison discharge is more significant. This is also reflected in the linear regression between the discharge of PyCatch and comparison models shown in figure 8, where the scatters are evenly located on both sides of the diagonal line representing PyCatch discharge equivalent to the average discharge of comparison models.

Besides hydrograph comparison, we also score the PyCatch skillness in simulating discharge with NSE and KGE*. The NSE score of PyCatch discharge is 0.29 (>0), indicating that the PyCatch model predictions are more reliable than the mean of the comparison discharge. The KGE* score of PyCatch discharge is 0.61 (>0.5), implying the PyCatch model is skillful. Removing the WaterGAP3 model (P-value>0.05, see table 3) from comparison models, KGE* score decreased by 0.06 and the NSE remains the same. Though the WaterGAP3 model is insignificantly correlated with PyCatch, involving WaterGAP3 in the PyCatch evaluation system improves the model efficiency score.

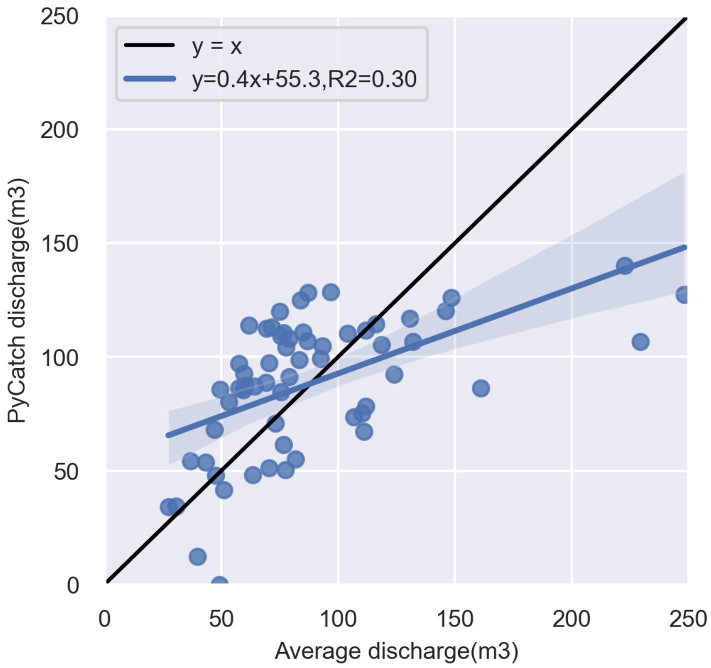


Figure 8. *Regression of monthly PyCatch discharge and average discharge of HTESSEL, LISFLOOD, WaterGAP3, PCR-GLOBLWB and PyCatch model*

3.2 Soil moisture evaluation

Soil moisture is a proxy for surface water. When the soil is saturated (soil moisture fraction >= porosity), water can not infiltrate anymore, and then temporary surface water starts to accumulate, acting as mosquito habitats. Similar to surface water, soil moisture also indicates vegetation conditions, which influence malaria vector and parasite host amounts. The PyCatch soil moisture is validated and analyzed in this section in terms of sensitivity to variable fluctuation, capability in reflecting seasonal and interannual soil moisture variation, and consistency with remote sensing data.

3.2.1 Sensitivity analysis

A sensitivity analysis is implemented for the precipitation, maximum vegetation interception, and regolith thickness, and the result is shown in figure 9. For example, the group of blue bars in the top right of figure 9 indicates that when the precipitation increase around 20% (y-axis), the soil moisture increase less than 2% (x-axis) than the soil moisture at the original precipitation level. The 12 bars from the bottom to the top in each group indicate soil moisture fluctuation in 12 months.

The PyCatch model is relatively robust as the soil moisture fluctuation percentage is less than 2.2%, with each input changing between -20% - +20%. The PyCatch soil moisture increases with the precipitation increasing, the regolith thickness decreasing, and the vegetation interception decreasing. If extra precipitation is added to the Okole catchment, more water infiltrates into the soil, contributing to the increase of the soil moisture. With a thicker regolith, the water travels deeper into the soil through the pores and cracks, and hence the volumetric soil water content decreases. If the vegetation canopy intercepts more precipitation in a storm, less rainfall finally reaches the soil, resulting in a dryer soil layer.

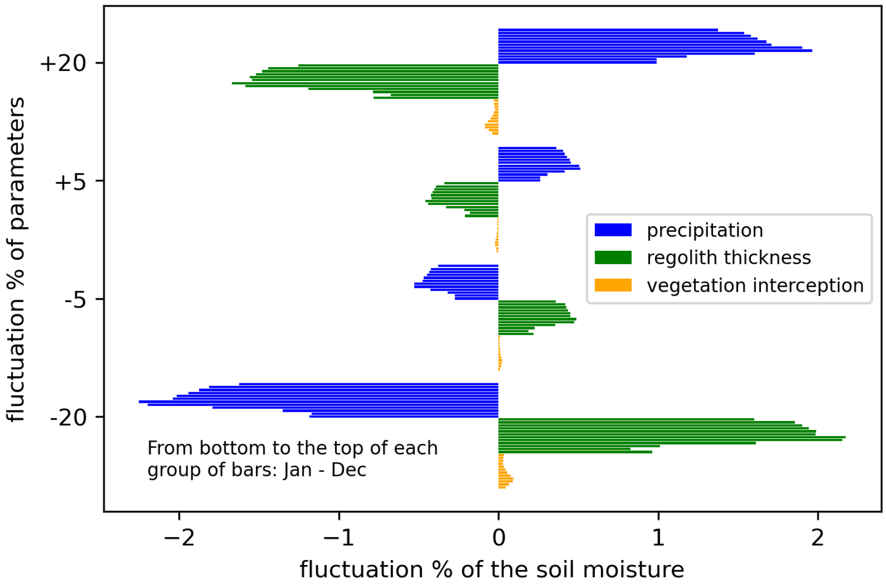


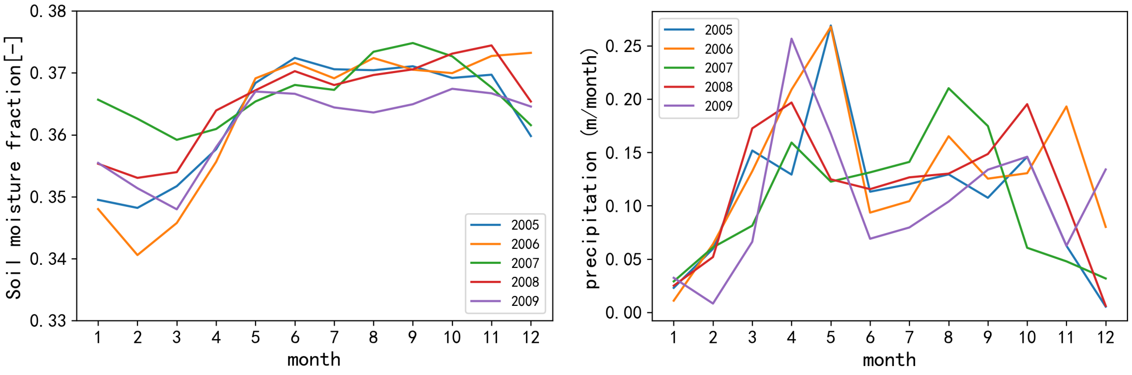
Figure 9. *The fluctuation percentage of the average soil moisture over the Okole catchment from 2005 to 2009 with relative precipitation, regolith thickness and vegetation interception fluctuation simulated by the PyCatch model.*

PyCatch soil moisture is least sensitive to the maximum vegetation interception that depends on the land cover fraction and the vegetation type. The species with a large leaf area, a funnel-like shape, and tough petioles tend to bear more water. As the annual precipitation of the Okole catchment reaches over 1300mm, the interception amount (P_{int}) is ignorable compared with the direct throughfall, so the maximum vegetation interception parameter impacts the output soil moisture marginally.

The PyCatch soil moisture sensitivity varies among different months. The model is more sensitive to precipitation and regolith thickness changes in May and June than in other months, and less sensitive in January and February. For maximum vegetation interception, the maximum sensitivity is in March and April.

3.2.2 Seasonal and interannual variation

The seasonal and interannual soil moisture and precipitation variations are plotted in figure 10. The seasonal soil moisture variation coincides with the annual precipitation pattern,



(a)

(b)

Figure 10. *Soil moisture seasonal variation averaged over the Okole catchment from 2005 to 2009 simulated by PyCatch (left) and corresponding monthly precipitation calculated from ERA5-land data (right)*

which is reasonable as the rainfall is the main driver of soil moisture change. The soil moisture is the lowest in January. The magnitude depends on the previous soil condition in the last December as the January precipitation is similarly low. The soil moisture keeps climbing until April or May, when the first rainfall peak occurs. Subsequently, the soil moisture decreases a little and then slowly rises until the second peak between September and November, which is also the time for the second rainfall peak. The soil moisture drops again after the second peak, and the dropping steepness is influenced by the peak time and the precipitation dropping slope.

PyCatch performs well in simulates interannual soil moisture variation. Changes in soil moisture lag behind the changes in precipitation, since the infiltration rate is limited and most precipitation transforms into runoff, so the soil moisture increase needs a couple of storms. The interannual soil moisture differences are much more significant in January - March, and November - December than in other months. On the contrary, the precipitation differences among years during the two rainy seasons are much larger than the previous and following months. Figure 10 shows the year 2009 is the driest year. According to NASA Earth Observation [53], in 2009, the precipitation from May to September in Africa was significantly lower than in normal years, resulting in a severe drought that caused punishing effects on the vegetation growth. The vegetation started recovering when October 2009 brought heavy rain. PyCatch successfully simulates the drought, though the high ERA5-Land precipitation input overall makes the simulated soil moisture overestimated.

Figure 11 depicts the relationship between the monthly-averaged soil moisture and precipitation of the PyCatch model during 2005-2009. The x-axis takes the logarithm of

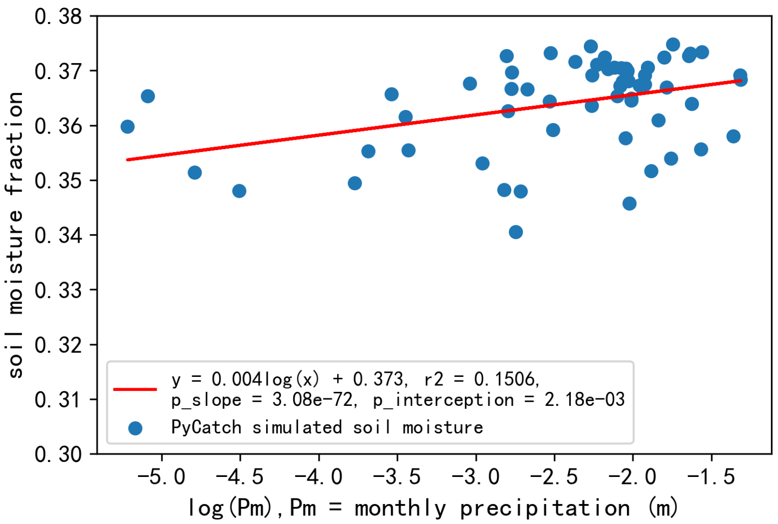


Figure 11. *The relationship between monthly precipitation and PyCatch soil moisture fraction of the Okole catchment during 2005-2009*

precipitation; the y-axis is the monthly soil moisture fraction. R^2 is around 0.15, which means 15% of the soil moisture seasonal variation is explained by the precipitation change. Though the explanation power is low, their logarithmic relationship is significant as p-values are small. This means precipitation is significantly related to soil moisture change, but there are also other important determinants, like topography, vegetation conditions, soil properties, and other climatic parameters.

3.2.3 Surface water occurrence validation with GLAD

There is a large gap between the average GLAD and PyCatch water possibilities over the study area. The GLAD water percentage and PyCatch water possibility are plotted in figure 12. Generally, PyCatch result (solid line) is much larger than GLAD throughout the year, with GLAD and PyCatch water possibility ranging from 0.04 to 0.25 and 0.159 to 0.34 respectively. The PyCatch rising limb has a smaller slope, and it hits the highest point sooner. After the rainy season, PyCatch water possibility keeps at a high level until November. Water possibility series calculated with 0.3 times and 0.1 times ERA5-Land precipitation are drown with dashdot and dashed line. With lower precipitation input, surface water occurrence frequency is smaller, and its variation range also decreases. When only 0.1 times precipitation is input into PyCatch, the water possibility peak moves to November because soil water needs more time to accumulate. Though the GLAD and PyCatch water possibility seems different, we cannot draw a conclusion that PyCatch has a poor performance in water occurrence simulation, because the quality and limitation of GLAD data in the Okole catchment needs to be discussed and tested.

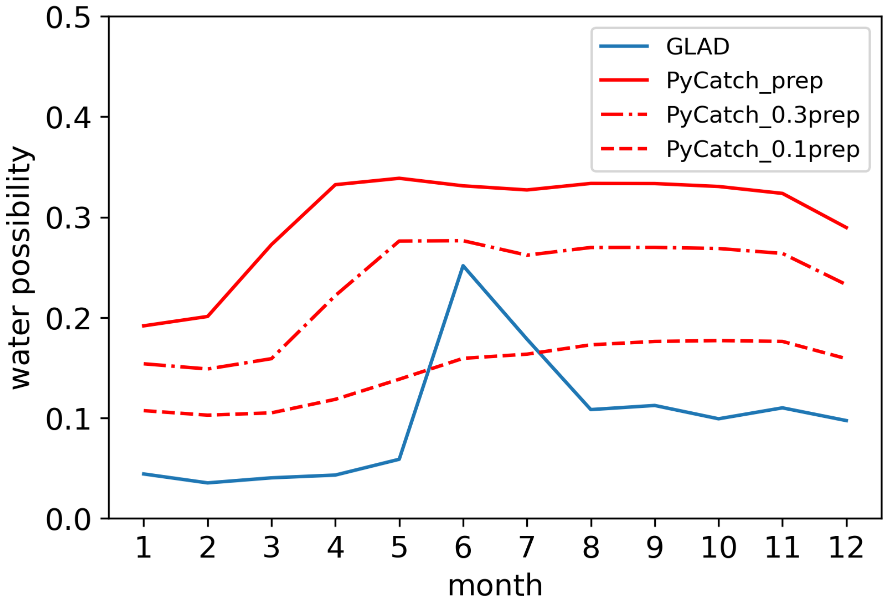
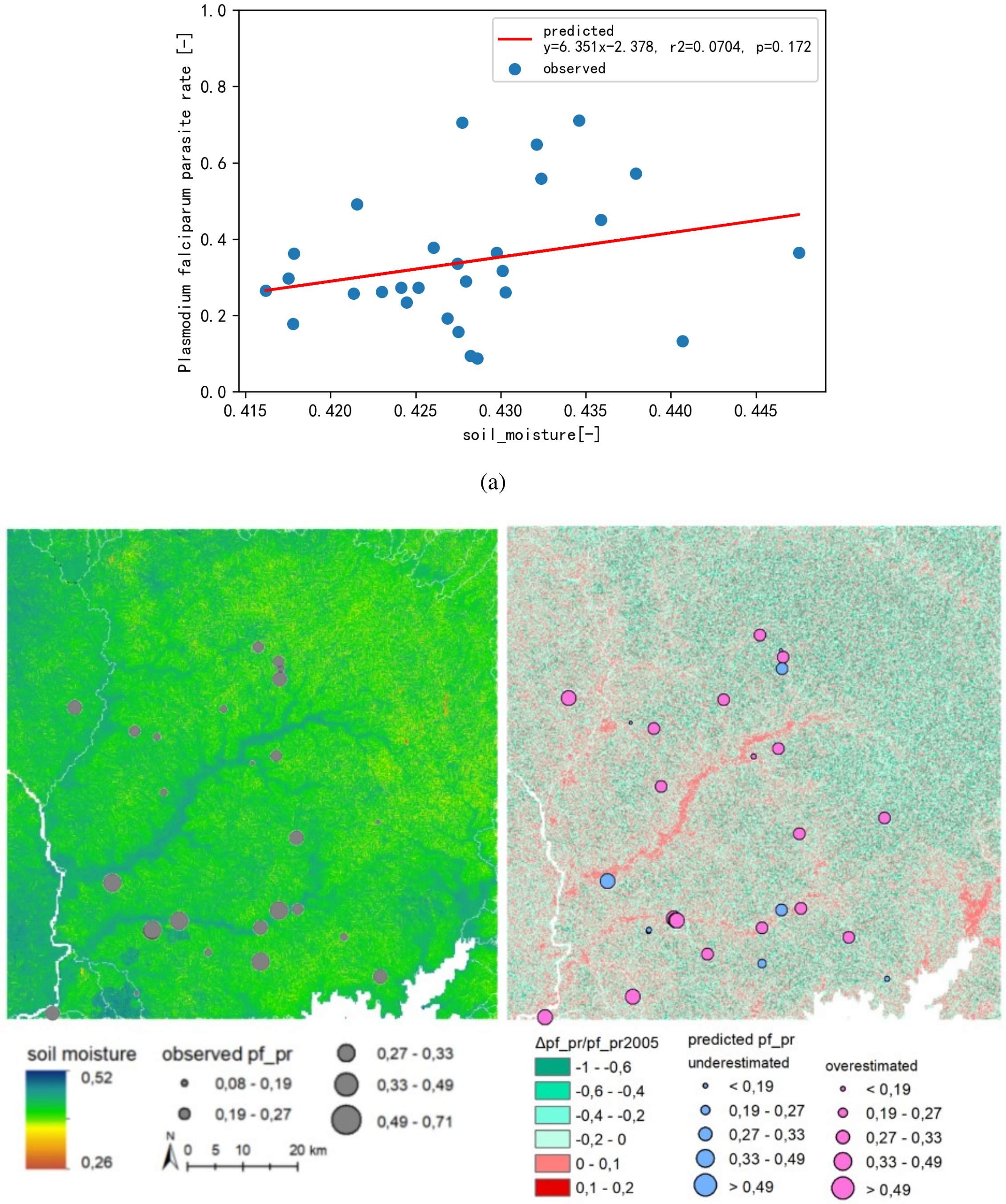


Figure 12. *PyCatch and GLAD monthly Water possibility averaged over the Okole catchment from 2005 to 2009: the solid, dashdot, and dashed line represent water possibility respectively calculated from the exact ERA5-Land precipitation, 0.3 times ERA5-Land precipitation and 0.1 times ERA5-Land precipitation.*

3.3 Malaria modeling

To predict malaria parasite rate, we do a linear regression between the P. falciparum parasite rate and the PyCatch soil moisture within the expanded study area, and the result is shown in figure 13(a). The increasing trend of the regression model is reasonable and realistic, since the soil moisture increment promotes the occurrence and retention of the surface water. The simple regression model shows that the soil_moisture is insignificantly related to parasite rate (P-value = 0.142, larger than 0.05) and has low explainability to the parasite rate change, indicating that the PyCatch soil moisture has potential to be used in more complicated malaria modeling.

Figure 13 presents the 5-year average soil moisture map ((b)left) generated by the linear regression and the resulting parasite rate variation map ((b)right) of the expanded study area from 2005 to 2009. The parasite rate of the riverine area somewhat increases, as well as the northeast of the Kwania lake. The area between rivers has a lighter malaria burden in 2009. The dots in the two map show the inconsistency between predicted malaria and the observed data. The gray dots in (b)left shows the observed parasite rate and the colored dots in (b)right shows the predicted parasite rate, with blue and prink representing the linear regression underestimating or overestimating the parasite rate. The linear regression model, predicts a higher malaria burden than observed in the Araca catchment situated in the south of the expanded study area.



(a)

(b)

Figure 13. *The regression between the parasite rate and soil moisture simulated with PyCatch (a); The long-term average soil moisture map (b(left)) and the predicted relative p.falciparum parasite rate change from 2005 to 2009*($(pf\_pr2009 - pf\_pr2005)/pf\_pr2005$)*. Gray dot size: the parasite rate from survey data; pink and blue dot size: the predicted parasite rate from the linear regression described in (a), with pink indicating overestimated positions and blue for underestimated positions compared with the survey data.*

4. Conclusion and Discussion

4.1 Discussion

In this research, we established a system to look over whether PyCatch is capable of being used in predicting malaria burden at a global scale in terms of data accessibility, discharge and soil moisture consistency with other models or remote sensing data, and the soil moisture potential in reflecting malaria parasite rate. We implemented this system in the Okole catchment from 2005 to 2009.

Data accessibility We blend the data from reanalysis, remote sensing, literature and observation to generate globally available PyCatch inputs, contributing to the feasibility of the global PyCatch application. ERA5-Land reanalysis data, combining model results, and the observed data across the world [35], are used as meteorological inputs. ERA5-Land enhances the spatial resolution to 9km from 30km of ERA5 and upgrades the temporal data continuity through a new initialization method. Though the resolution is still much coarser than PyCatch, it still serves as an ideal meteorological data source for the global hydrological model because of its accuracy and long-term global accessibility. Since ERA5-Land data exclude the precipitation assimilation process, we calibrate the ERA5-Land precipitation with observed annual rainfall. However, this calibration ignores the global water balance closure and heat circulation without recalculating other meteorological variables. This ignorance may cause severe mistakes on a global scale, but for simplicity, we correct the precipitation with this method as we focus on the earth surface hydrological processes of a small open catchment, where ERA5-Land data too much overestimate the precipitation. When applying PyCatch at a large scale, the forcing process needs to be implemented.

The soil property maps at 250m resolution from ISRIC are derived from the machine learning classification method using 150,000 soil profiles and remote sensing data [54]. In this soil database, the average model explained variance of all soil properties is around 61%. As field experiments for soil property measurement are challenging to conduct at a global scale, the open-access dataset solves the soil data acquisition issue in global hydrology modeling. However, the soil water content varies within days, resulting in its instability. The regolith thickness is set to be uniform within the Okole catchment, which actually varies in space with a smaller magnitude in mountainous areas. We can overcome this shortage by adopting the 1km gridded upland hill-slope regolith thickness [55]. In this research, we resampled the meteorological data and soil water content data at various resolutions into the model resolution, which can be further improved by applying a downscaling method according to the elevation and slope aspects.

For the vegetation input, MODIS provides widely used and reliable LAI data. The Copernicus Land Monitoring Service produced a 100m-resolution global land cover data, though the historical land cover types before 2015 are unavailable. Here we ignore the land cover type changes from 2005 till now of the Okele catchment and get the stomatal conductance and the vegetation height from literature data accordingly, yet the rationality of the assumption needs to be examined.

Hydrograph comparison PyCatch is skillful with regard to discharge simulation, and the hydrograph simulated with PyCatch coincides well with other models. Lack of observed discharge data, we compared the Okole river hydrograph simulated by PyCatch against the HTESSEL, WaterGAP3, PCR-GLOBWB, and LISFLOOD model. The hydrographs of the five models seem diverse to each other, but after taking the average value of four comparison models, the PyCatch model hydrograph has a good consistency with the average discharge, which is also reflected by the Pearson correlation. Assuming the average discharge of comparison models represents the actual stream runoff, the PyCatch performance regarding the hydrograph simulation is satisfactory. However, the regression result shows that the PyCatch discharge is irreplaceable by the average of comparison models according to the low variance explainability, and this is likely to be related to the high spatio-temporal resolution of PyCatch. The NSE and KGE score of the PyCatch demonstrates that the model is skillful and accurate in the overall performance of mean value, errors, and bias of the simulated discharge.

Nevertheless, the consistency among the five models is partly attributed to the similar forcing processes for atmosphere variables, soil properties, or other components. Forcing is either a part of these hydrological models or included in the model input dataset. If we could find the observed discharge data of the Okole catchment, the verifying result would be more convincing. Besides, the model comparison can be performed with a higher time resolution other than monthly as data assimilation included in the comparison models may correct the monthly discharge to the same level, but more differences might be observed with an hourly or daily step.

Soil moisture validation The PyCatch model is robust and performs well in simulating seasonal and interannual soil moisture variation, but the water occurrence frequency needs to be further tested. We proved the model is robust by a sensitivity analysis in surface water simulation, because the soil moisture fluctuates slightly even with a significant change in vegetation, meteorology, and soil property. However, the soil moisture is less sensitive to the precipitation than our expectation, possibly because the precipitation in the Okole catchment is so high (1361 mm per year) that other variables mainly restrict the soil moisture. We tried to amplify tenfold the precipitation input and apply the sensitivity analysis again. It turns out that the PyCatch soil moisture is even less sensitive to precipitation fluctuation.

PyCatch successfully simulates the bimodal seasonal soil moisture variation, consistent with the annual precipitation changes. Unlike precipitation, the soil moisture between June and October keeps at a high level despite a modest drop. Though precipitation decreases dramatically during this period, around 100mm rainfall per month is still a lot, resulting in high soil moisture levels. The autumn soil moisture peak is larger than the spring peak, which is opposite to the precipitation pattern. For one thing, the winter dry season is stronger and more precipitation is required in the following spring to produce the same soil moisture level as in autumn. For another, when the precipitation is adequate, the water infiltration rate stays unchanged with higher precipitation, so the soil moisture variation depends on successive high precipitation accumulation. The second soil precipitation peak lasts longer, contributing to the high soil moisture. PyCatch also reproduces the African drought in the 2009 summer (see figure 10), indicating the model capability relevant to reflecting interannual soil moisture variation.

However, figure 12 shows a large divergence between the GLAD water possibility and the PyCatch water frequency calculated from soil moisture. The comparison is based on the assumption that the 5-year average soil moisture reflects the long-term average soil conditions, but an abnormal drought or flood during the five years will significantly impact the validation results. Besides, the GLAD water percentage might be underestimated due to vegetation occlusion, and even in the Okole river, where the water possibility should be close to 1, the GLAD water percentage is still small. Additionally, the GLAD data is aimed to distinguish the permanent water area from the land area, so some clear observations that determine the cell as the opposite state of water or land in different scenes are filtered [51]. These observations indicate temporary water body information, essential for simulating malaria and can be simulated with PyCatch. A pure water cell is easily recognized from land cells according to the spectral information, but most cells with temporary surface water storage are mixed with vegetation, bare soil or other land cover types, so it is hard to monitor surface water occurrence with remote sensing data. For example, a large area of the Okole catchment is covered with cropland, and unless a flood totally submerges them, the shallow surface water at the bottom of the plants is hard to be caught by satellite sensors. Although GLAD is tested qualified for permanent water body classification globally, the large gaps between the PyCatch and GLAD data within the small catchment possibly originate from the unreliability of the GLAD data in temporary water indication. Therefore, the PyCatch performance in simulating surface water occurrence needs to be further examined with replacement data.

Malaria modeling We successfully relate the soil moisture with malaria burden, although the explained variance and the P-value indicate that the regression is weak. The parasite rate non-linearly responds to the soil moisture variation in space and time, beyond the linear regression capability. The ignorance of the complex impact of hydrological processes, climatic conditions, vegetation properties, anti-malaria measures, population density, and immunity conditions also leads to poor model performance. We also apply multiple linear regression to include LAI and soil porosity influences, but hardly get any improvements. Furthermore, only 28 pieces of malaria parasite data are available in the expanded study area, less than the lowest limitation of statistics requirement. The data collection time is inconsistent with the modeling period. Though we apply a time calibration to the involving parasite data, it only calibrates the long-period parasite variation trend, ignoring the seasonal parasite variation pattern. However, here we focus on the PyCatch model performance, and the malaria modeling can be improved by selecting a more extensive study area to cover more malaria data, incorporating more variables and applying non-linear models, such as the agent-based entomology model incorporated in the HYDREMATS model [11]. Generally, the correct trend of the regression between the soil moisture and parasite rate still serves as proof for the PyCatch potential in malaria modeling.

Model running efficiency Running the PyCatch model over 100 km*100 km square area for five years takes around 84.5h on my laptop with an 8GB memory. Hay [10] estimates that the global land area under malaria risk decreased from 77.6 million km^2 in 1900 to around 39.8 million km^2 in 2002. Though the malarious area probably has shrunk after the year 2002, monitoring the whole 39.8 million km^2 helps to prevent areas free of malaria after 2002 being popular with malaria again. If we simultaneously conduct 4 runs considering the memory and capacity limitation of my laptop, the total time consumption on applying the PyCatch model to the malarious area in 2002 with a 5 year study period will be:

$$39.8 * 10^6/10000/4 * 84.5 = 84077.5h$$

. With a supercomputer or a computer cluster with 400 nodes, the running time will be reduced to around 210 hours or 8.76 days.

4.2 Conclusion

Many regional malaria models focused on surface water hydrological processes have been built in recent years in water-limited malaria-epidemic regions according to the vector reproduction need. We apply a process-based dynamic hydrological model PyCatch to the Okole catchment in Uganda. PyCatch inputs are globally accessible, laying a solid foundation for global application. The PyCatch capability with regard to discharge simulating is verified satisfactory by hydrograph comparison with other models and model efficiency evaluation. The soil moisture evaluation shows PyCatch is stable and capable of simulating soil moisture variation, but the surface water occurrence calculated with PyCatch needs to be further studied. It is proved that PyCatch soil moisture can be used for malaria parasite rate modeling. In conclusion, PyCatch performs well in simulating malaria-relevant hydrological processes, and has potential to be incorporated into global malaria incidence modeling.

References

[1] W. H. Organization *et al.*, "World malaria report 2019. 2019," *Reference Source*,

[2] F. F. Ateba, I. Sagara, N. Sogoba, M. Touré, D. Konaté, S. I. Diawara, S. A. S. Diakité, A. Diarra, M. D. Coulibaly, M. Dolo, *et al.*, "Spatio-temporal dynamic of malaria incidence: A comparison of two ecological zones in mali," *International journal of environmental research and public health*, vol. 17, no. 13, p. 4698, 2020.

[3] C. Caminade, S. Kovats, J. Rocklov, A. M. Tompkins, A. P. Morse, F. J. Colón-González, H. Stenlund, P. Martens, and S. J. Lloyd, "Impact of climate change on global malaria distribution," *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3286–3291, 2014.

[4] S. Antinori, L. Galimberti, L. Milazzo, and M. Corbellino, "Biology of human malaria plasmodia including plasmodium knowlesi," *Mediterranean journal of hematology and infectious diseases*, vol. 4, no. 1, 2012.

[5] S. E. Eikenberry and A. B. Gumel, "Mathematical modeling of climate change and malaria transmission dynamics: A historical review," *Journal of mathematical biology*, vol. 77, no. 4, pp. 857–933, 2018.

[6] M. N. Bayoh, "Studies on the development and survival of anopheles gambiae sensu stricto at various temperatures and relative humidities," Ph.D. dissertation, Durham University, 2001.

[7] S. N. Arifin, G. J. Davis, and Y. Zhou, "Modeling space in an agent-based model of malaria: Comparison between non-spatial and spatial models," in *Proceedings of the 2011 Workshop on Agent-Directed Simulation*, 2011, pp. 92–99.

[8] M. Smith, T. Willis, L. Alfieri, W. James, M. Trigg, D. Yamazaki, A. Hardy, B. Bisselink, A. De Roo, M. Macklin, *et al.*, "Incorporating hydrology into climate suitability models changes projections of malaria transmission in africa," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[9] S. Hay, J. Omumbo, M. Craig, and R. Snow, "Earth observation, geographic information systems and plasmodium falciparum malaria in sub-saharan africa," *Advances in parasitology*, vol. 47, pp. 173–215, 2000.

[10] S. I. Hay, C. A. Guerra, A. J. Tatem, A. M. Noor, and R. W. Snow, "The global distribution and population at risk of malaria: Past, present, and future," *The Lancet infectious diseases*, vol. 4, no. 6, pp. 327–336, 2004.

[11] A. Bomblies, J.-B. Duchemin, and E. A. Eltahir, "Hydrology of malaria: Model development and application to a sahelian village," *Water Resources Research*, vol. 44, no. 12, 2008.

[12] E. E. Ricotta, S. A. Frese, C. Choobwe, T. A. Louis, and C. J. Shiff, "Evaluating local vegetation cover as a risk factor for malaria transmission: A new analytical approach using imagej," *Malaria journal*, vol. 13, no. 1, p. 94, 2014.

[13] W. Martens, L. W. Niessen, J. Rotmans, T. H. Jetten, and A. J. McMichael, "Potential impact of global climate change on malaria risk.," *Environmental health perspectives*, vol. 103, no. 5, pp. 458–464, 1995.

[14] M. C. Thomson, F. J. Doblas-Reyes, S. J. Mason, R. Hagedorn, S. J. Connor, T. Phindela, A. P. Morse, and T. N. Palmer, "Malaria early warnings based on seasonal climate forecasts from multi-model ensembles," *Nature*, vol. 439, no. 7076, pp. 576–579, 2006. DOI: `10.1038/nature04503`.

[15] M. H. Craig, R. Snow, and D. le Sueur, "A climate-based distribution model of malaria transmission in sub-saharan africa," *Parasitology today*, vol. 15, no. 3, pp. 105–111, 1999.

[16] M. B. Hoshen and A. P. Morse, "A weather-driven model of malaria transmission," *Malaria Journal*, vol. 3, 2004. DOI: `10.1186/1475-2875-3-32`.

[17] I. Kleinschmidt, M. Bagayoko, G. P. Y. Clarke, M. Craig, and D. Le Sueur, "A spatial statistical approach to malaria mapping," *International Journal of Epidemiology*, vol. 29, no. 2, pp. 355–361, 2000. DOI: `10.1093/ije/29.2.355`.

[18] J. A. Omumbo, S. I. Hay, S. J. Goetz, R. W. Snow, and D. J. Rogers, "Updating historical maps of malaria transmission intensity in East Africa using remote sensing," *Photogrammetric Engineering and Remote Sensing*, vol. 68, no. 2, pp. 161–166, 2002.

[19] D. J. Rogers, S. E. Randolph, R. W. Snow, and S. I. Hay, "Satellite imagery in the study and forecast of malaria," *Nature*, vol. 415, no. 6872, pp. 710–715, 2002. DOI: `10.1038/415710a`.

[20] J. Shaman, M. Stieglitz, C. Stark, S. Le Blancq, and M. Cane, "Using a dynamic hydrology model to predict mosquito abundances in flood and swamp water," *Emerging infectious diseases*, vol. 8, no. 1, p. 8, 2002.

[21] R. L. Gianotti, A. Bomblies, and E. A. Eltahir, "Hydrologic modeling to screen potential environmental management methods for malaria vector control in niger," *Water resources research*, vol. 45, no. 8, 2009.

[22] T. K. Yamana and E. A. Eltahir, "Early warnings of the potential for malaria transmission in rural africa using the hydrology, entomology and malaria transmission simulator (hydremats)," *Malaria journal*, vol. 9, no. 1, pp. 1–10, 2010.

[23] T. K. Yamana, A. Bomblies, I. M. Laminou, J.-B. Duchemin, and E. A. Eltahir, "Linking environmental variability to village-scale malaria transmission using a simple immunity model," *Parasites & vectors*, vol. 6, no. 1, p. 226, 2013.

[24] A. Bomblies and E. A. Eltahir, "Assessment of the impact of climate shifts on malaria transmission in the sahel," *EcoHealth*, vol. 6, no. 3, pp. 426–437, 2009.

[25] A. M. Tompkins and V. Ermert, "A regional-scale, high resolution dynamical malaria model that accounts for population density, climate and surface hydrology," *Malaria journal*, vol. 12, no. 1, p. 65, 2013.

[26] E. O. Asare, A. M. Tompkins, and A. Bomblies, "A regional model for malaria vector developmental habitats evaluated using explicit, pond-resolving surface hydrology simulations," *PLoS One*, vol. 11, no. 3, e0150626, 2016.

[27] N. Lana-Renault and D. Karssenberg, "Pycatch: Component based hydrological catchment modelling," *Cuadernos de Investigación Geográfica*, vol. 39, no. 2, pp. 315–333, 2013.

[28] D. Karssenberg, "The value of environmental modelling languages for building distributed hydrological models," *Hydrological Processes*, vol. 16, no. 14, pp. 2751–2766, 2002.

[29] D. Karssenberg, K. de Jong, and J. Van Der Kwast, "Modelling landscape dynamics with python," *International Journal of Geographical Information Science*, vol. 21, no. 5, pp. 483–495, 2007.

[30] E. A. Mordecai, K. P. Paaijmans, L. R. Johnson, C. Balzer, T. Ben-Horin, E. de Moor, A. McNally, S. Pawar, S. J. Ryan, T. C. Smith, *et al.*, "Optimal temperature for malaria transmission is dramatically lower than previously predicted," *Ecology letters*, vol. 16, no. 1, pp. 22–30, 2013.

[31] D. Yamazaki, D. Ikeshima, R. Tawatari, T. Yamaguchi, F. O'Loughlin, J. C. Neal, C. C. Sampson, S. Kanae, and P. D. Bates, "A high-accuracy map of global terrain elevations," *Geophysical Research Letters*, vol. 44, no. 11, pp. 5844–5853, 2017.

[32] M. Buchhorn, M. Lesiv, N.-E. Tsendbazar, M. Herold, L. Bertels, and B. Smets, "Copernicus global land cover layers—collection 2," *Remote Sensing*, vol. 12, no. 6, p. 1044, 2020.

[33] J. Muñoz Sabater, “Era5-land hourly data from 1981 to present,” *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, 2019.

[34] J. Ephrath, J. Goudriaan, and A. Marani, “Modelling diurnal patterns of air temperature, radiation wind speed and relative humidity by equations from daily characteristics,” *Agricultural systems*, vol. 51, no. 4, pp. 377–393, 1996.

[35] J. Muñoz-Sabater, E. Dutra, A. Agustı-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, *et al.*, “Era5-land: A state-of-the-art global reanalysis dataset for land applications,” *Earth System Science Data Discussions*, pp. 1–50, 2021.

[36] T. Hengl, G. B. Heuvelink, B. Kempen, J. G. Leenaars, M. G. Walsh, K. D. Shepherd, A. Sila, R. A. MacMillan, J. Mendes de Jesus, L. Tamene, *et al.*, “Mapping soil properties of africa at 250 m resolution: Random forests significantly improve current predictions,” *PloS one*, vol. 10, no. 6, e0125814, 2015.

[37] R. Myneni, Y. Knyazikhin, and T. Park, *Mcd15a2h modis/terra+ aqua leaf area index/fpar 8-day l4 global 500m sin grid v006. nasa eosdis l. process. daac 2015.*

[38] L. Van Beek, “Forcing pcr-globwb with cru data,” *Report Department of Physical Geography, Utrecht University, Utrecht, the Netherlands*, 2008.

[39] G. Balsamo, A. Beljaars, K. Scipal, P. Viterbo, B. van den Hurk, M. Hirschi, and A. K. Betts, “A revised hydrology for the ecmwf model: Verification from field site to terrestrial water storage and impact in the integrated forecast system,” *Journal of hydrometeorology*, vol. 10, no. 3, pp. 623–643, 2009.

[40] J. Van Der Knijff, J. Younis, and A. De Roo, “Lisflood: A gis-based distributed model for river basin scale water balance and flood simulation,” *International Journal of Geographical Information Science*, vol. 24, no. 2, pp. 189–212, 2010.

[41] K. Verzano, “Climate change impacts on flood related hydrological processes: Further development and application of a global scale hydrological model,” Ph.D. dissertation, University of Kassel Kassel, 2009.

[42] S. M. Uppala, P. Kållberg, A. Simmons, U. Andrae, V. D. C. Bechtold, M. Fiorino, J. Gibson, J. Haseler, A. Hernandez, G. Kelly, *et al.*, “The era-40 re-analysis,” *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 131, no. 612, pp. 2961–3012, 2005.

[43] N. Rayner, D. Parker, E. Horton, *et al.*, “Hadley centre for climate prediction and research, met office, bracknell, uk global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century,” *J. Geophys. Res*, vol. 108, no. D14, p. 4407, 2003.

[44] V. Thiemig, P. Salamon, G. N. Gomes, J. O. Skøien, M. Ziese, A. Rauthe-Schöch, K. Rehfeldt, D. Pichon, and C. Schweim, "Emo-5: Copernicus pan-european high-resolution meteorological data set for large-scale hydrological modelling," in *EGU General Assembly Conference Abstracts*, 2020, p. 21 551.

[45] V. Ntegeka, P. Salamon, G. Gomes, H. Sint, V. Lorini, M. Zambrano-Bigiarini, and J. Thielen, *EFAS-Meteo A European daily high-resolution gridded meteorological data set for 1990-2011*. Jan. 2013, ISBN: 978-92-79-35006-1. DOI: 10.2788/51262.

[46] E. H. Sutanudjaja, R. Van Beek, N. Wanders, Y. Wada, J. H. Bosmans, N. Drost, R. J. Van Der Ent, I. E. De Graaf, J. M. Hoch, K. De Jong, *et al.*, "Pcr-globwb 2: A 5 arcmin global hydrological and water resources model," *Geoscientific Model Development*, vol. 11, no. 6, pp. 2429–2453, 2018.

[47] D. P. Dee, S. M. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, d. P. Bauer, *et al.*, "The era-interim reanalysis: Configuration and performance of the data assimilation system," *Quarterly Journal of the royal meteorological society*, vol. 137, no. 656, pp. 553–597, 2011.

[48] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part i—a discussion of principles," *Journal of hydrology*, vol. 10, no. 3, pp. 282–290, 1970.

[49] W. J. Knoben, J. E. Freer, and R. A. Woods, "Inherent benchmark or not? comparing nash–sutcliffe and kling–gupta efficiency scores," *Hydrology and Earth System Sciences*, vol. 23, no. 10, pp. 4323–4331, 2019.

[50] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling," *Journal of hydrology*, vol. 377, no. 1-2, pp. 80–91, 2009.

[51] A. H. Pickens, M. C. Hansen, M. Hancher, S. V. Stehman, A. Tyukavina, P. Potapov, B. Marroquin, and Z. Sherani, "Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full landsat time-series," *Remote Sensing of Environment*, vol. 243, p. 111 792, 2020.

[52] D. J. Weiss, A. Nelson, H. S. Gibson, W. Temperley, S. Peedell, A. Lieber, M. Hancher, E. Poyart, S. Belchior, N. Fullman, B. Mappin, U. Dalrymple, J. Rozier, T. C. Lucas, R. E. Howes, L. S. Tusting, S. Y. Kang, E. Cameron, D. Bisanzio, K. E. Battle, S. Bhatt, and P. W. Gething, "A global map of travel time to cities to assess inequalities in accessibility in 2015," *Nature*, vol. 553, pp. 333–336, 7688 2018, ISSN: 14764687. DOI: 10.1038/nature25181.

[53] *Drought in africa*. [Online]. Available: https://earthobservatory.nasa.gov/images/39363/drought-in-africa.

[54] T. Hengl, J. Mendes de Jesus, G. B. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, *et al.*, "Soilgrids250m: Global gridded soil information based on machine learning," *PLoS one*, vol. 12, no. 2, e0169748, 2017.

[55] J. PELLETIER, P. BROXTON, P. HAZENBERG, X. ZENG, P. TROCH, G. NIU, Z. WILLIAMS, M. BRUNKE, and D. GOCHIS, "Global 1-km gridded thickness of soil, regolith, and sedimentary deposit layers," en, 2016. DOI: 10.3334/ORNLDAAC/1304. [Online]. Available: http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1304.