PLOS ONE

RESEARCH ARTICLE

Mapping access to domestic water supplies from incomplete data in developing countries: An illustrative assessment for Kenya

Weiyu Yu[1]*, Nicola A. Wardrop[1], Robert E. S. Bain[2], Victor Alegana[1,3], Laura J. Graham[1], Jim A. Wright[1]

1 University of Southampton, Southampton, Hampshire, United Kingdom, **2** Division of Data, Research and Policy, United Nations Children's Fund (UNICEF), New York, New York, United States of America, **3** Population Health Unit, Kenya Medical Research Institute—Wellcome Trust Research Programme, Nairobi, Kenya

* W.Yu@soton.ac.uk

Citation: Yu W, Wardrop NA, Bain RES, Alegana V, Graham LJ, Wright JA (2019) Mapping access to domestic water supplies from incomplete data in developing countries: An illustrative assessment for Kenya. PLoS ONE 14(5): e0216923. https://doi.org/10.1371/journal.pone.0216923

Editor: Laura Scherer, Leiden University, NETHERLANDS

Received: November 14, 2018

Accepted: May 1, 2019

Published: May 17, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Water point mapping databases, generated through surveys of water sources such as wells and boreholes, are now available in many low and middle income countries, but often suffer from incomplete coverage. To address the partial coverage in such databases and gain insights into spatial patterns of water resource use, this study investigated the use of a maximum entropy (MaxEnt) approach to predict the geospatial distribution of drinking-water sources, using two types of unimproved sources in Kenya as illustration. Geographic locations of unprotected dug wells and surface water sources derived from the Water Point Data Exchange (WPDx) database were used as inputs to the MaxEnt model alongside geological/hydrogeological and socio-economic covariates. Predictive performance of the MaxEnt models was high (all > 0.9) based on Area Under the Receiver Operator Curve (AUC), and the predicted spatial distribution of water point was broadly consistent with household use of these unimproved drinking-water sources reported in household survey and census data. In developing countries where geospatial datasets concerning drinking-water sources often have necessarily limited resolution or incomplete spatial coverage, the modelled surface can provide an initial indication of the geography of unimproved drinking-water sources to target unserved populations and assess water source vulnerability to contamination and hazards.

1. Introduction

Access to water is a basic human necessity and plays a key role in well-being and sustainable development [1]. Following the recognition of the human right to drinking-water [2], the United Nations' Sustainable Development Goals (SDGs) included a target that aims to achieve universal and equitable access to safe and affordable drinking-water by 2030 (Goal 6 Target 6.1) [3]. Realisation of the ambitious SDG target requires robust progress monitoring to underpin development planning for service improvements and to identify disadvantaged groups, so that those with the greatest needs can be prioritised accordingly. Geographic

location is often an important factor that influences access to basic housing infrastructures including water supply [4]. International guidance on monitoring of inequalities in drinking-water services therefore recommends data are disaggregated by geographic location and rurality to reveal and address sub-national disparities [5]. As such, this expanded need for detailed locational information where water is accessed for household use is likely to place greater demands on existing data sources.

Currently, there are few data sets describing the spatial distribution of access to drinking-water sources. Conventional monitoring of drinking-water services at international and national levels relies heavily on household surveys and censuses [6]. Even though several previous studies have created spatially disaggregated map layers using such data [7,8], these data sets necessarily have limited spatial resolution because of data protection needs (i.e. scrambling of household survey GPS cluster coordinates, aggregation of census data, etc.) and often focus on water sources for drinking only. Other relevant geospatial data sources include water point mapping databases, which contain geo-referenced water source locations alongside associated attributes. Water point data are not affected by data protection concerns, so have greater spatial precision than household surveys and censuses. Often, such databases capture water points used for many purposes, not solely for drinking by households. However, very often they are project-specific, focus on particular geographic areas and water source types, and are collected and stored by multiple agencies. Because of geographic gaps in project coverage and agency responsibilities, there thus remain few nationwide, complete and consistent water point datasets [9]. To gain a more complete picture of domestic water use from water point mapping, one therefore has to 'gap-fill' the data first.

One potential way to address concerning water points is to estimate their overall geographical distribution by predicting the probability of water point presence or relative site suitability across the landscape of interest. This problem is analogous to mapping habitat suitability from incomplete species occurrences in ecological studies. Various spatial predictive modelling techniques have been developed to address this issue, including a variety of algorithms based on bioclimatic envelopes [10,11], Gower distance [12], Mahalanobis distance [13,14], statistical regression [15–17], and machine learning [18–22]. These methods have been used to examine impacts of climate changes [23–26], invasive species [27–30], conservation assessment [31–33], and species richness [34]. In addition, they have also been used to track disease vectors [35–40], assess landslide susceptibility [41–43], map soil phosphorus [44], and wildfire risk [45–47]. Furthermore, spatial predictive modelling techniques have been combined with water point data for groundwater potential delineation for groundwater resource assessment [48–55] in recent years. However, to our knowledge, spatial predictive modelling has not yet been used for predicting the spatial distribution of infrastructure or services such as domestic water supplies by incorporating socio-economic as well as biophysical covariates.

Based on the type of observational data used, spatial predictive models can be classified as presence-only, presence-absence, and presence-background (or presence-pseudo-absence) methods. Since water point data only record observed locations of water access points, whilst the absence of a water source at a given location cannot be inferred from such records, they may be considered as presence data. In this study, we therefore employ a machine learning approach developed for spatial ecology [22], namely maximum entropy (MaxEnt) modelling, since it only requires 'presence' data from water point mapping enables use of categorical variables as predictors, has good predictive performance [56], user-friendly software [57,58], and is suitable for integration into a reproducible workflow [59,60]. As a novel spatial predictive modelling method, MaxEnt is more realistic than simple measures such as distance models [12,14], but more straightforward to implement than more complex methods, such as likelihood analysis [61] and hierarchical species distribution models [62]. However, despite its

advantages, because MaxEnt relies on presence-only data, it lacks information on the proportion of occupied sites (i.e. prevalence) and the logistic transformation of raw outputs used by MaxEnt only represents a relative ranking rather than a true probability [63]. In this study, we adopted a MaxEnt model merely as an illustrative example of how a wider suite of techniques can make predictions of the potential geographical distribution of unimproved domestic water sources from physical and socio-economic characteristics. We used unimproved water point data to examine the feasibility of introducing this method into the water sector, assuming that the locations of observed water points reflect suitable conditions for siting such water sources. Kenya was selected as a case study, given household use of unimproved domestic water sources there [64] and availability of suitable data. The main objectives of this study are to (1) examine the potential applicability of MaxEnt modelling for predicting the geographical occupancy (or relative site suitability) of access to domestic water supplies using water point data; and (2) analyse the importance of predictive covariates that potentially explain the spatial distribution of drinking-water sources.

2. Methods

2.1 Study area

Kenya is a lower-middle income country [65] in Sub-Saharan Africa (SSA), which did not achieve the MDG drinking-water target by 2015. However, substantial progress in access to improved drinking-water sources was demonstrated during the MDG period [66]. According to the most recent Demographic and Health Survey (DHS) [67], among Kenya's estimated population of 43 million in 2014, 31.6% were using unimproved drinking-water sources, including 7.3% using unprotected dug wells, 4.4% unprotected springs, 1.5% tanker trucks or carts with drum, and 18.4% surface water. The majority (85.7%) of the urban population had access to improved drinking-water sources, whilst nearly half (41.5%) of the rural population were using unimproved drinking-water sources.

2.2 Target water sources and locations

For illustration purposes, we focused specifically on two types of unimproved water sources, namely points representing unprotected dug wells ($n_u = 523$) and surface water (i.e. where households directly draw untreated water from rivers, streams, ponds, and lakes, $n_s = 212$) at a spatial resolution of 1 km (see Fig 1). This was selected considering both the availability and quality of relevant predictive covariates as well as water point data. We acquired water point data from the WPDx (http://www.waterpointdata.org/) database on 10th April 2018. This included user-uploaded water point inventories from diverse data sources, and therefore increases the variety of sample data. We restricted our analysis to household domestic sources, excluding institutional sources and water points potentially used for non-domestic purposes such as irrigation and watering livestock. For surface water, we further excluded water points where an ambiguous water source type (e.g. spring, dam, pan etc.) or a water lifting or extraction mechanism was recorded, as this may indicate a higher service level than direct consumption of surface water, as defined in the water ladder used in international monitoring [68]. To increase temporal consistency with covariate map layers, water points reportedly installed after 2009 were excluded in our analysis, assuming that sources lacking installation dates were installed before 2009. There was insufficient information recorded to exclude non-functional sources, so water sources with reported functionality and service continuity issues were retained for analysis. Water points located outside the study area or within large inland water bodies were also excluded.

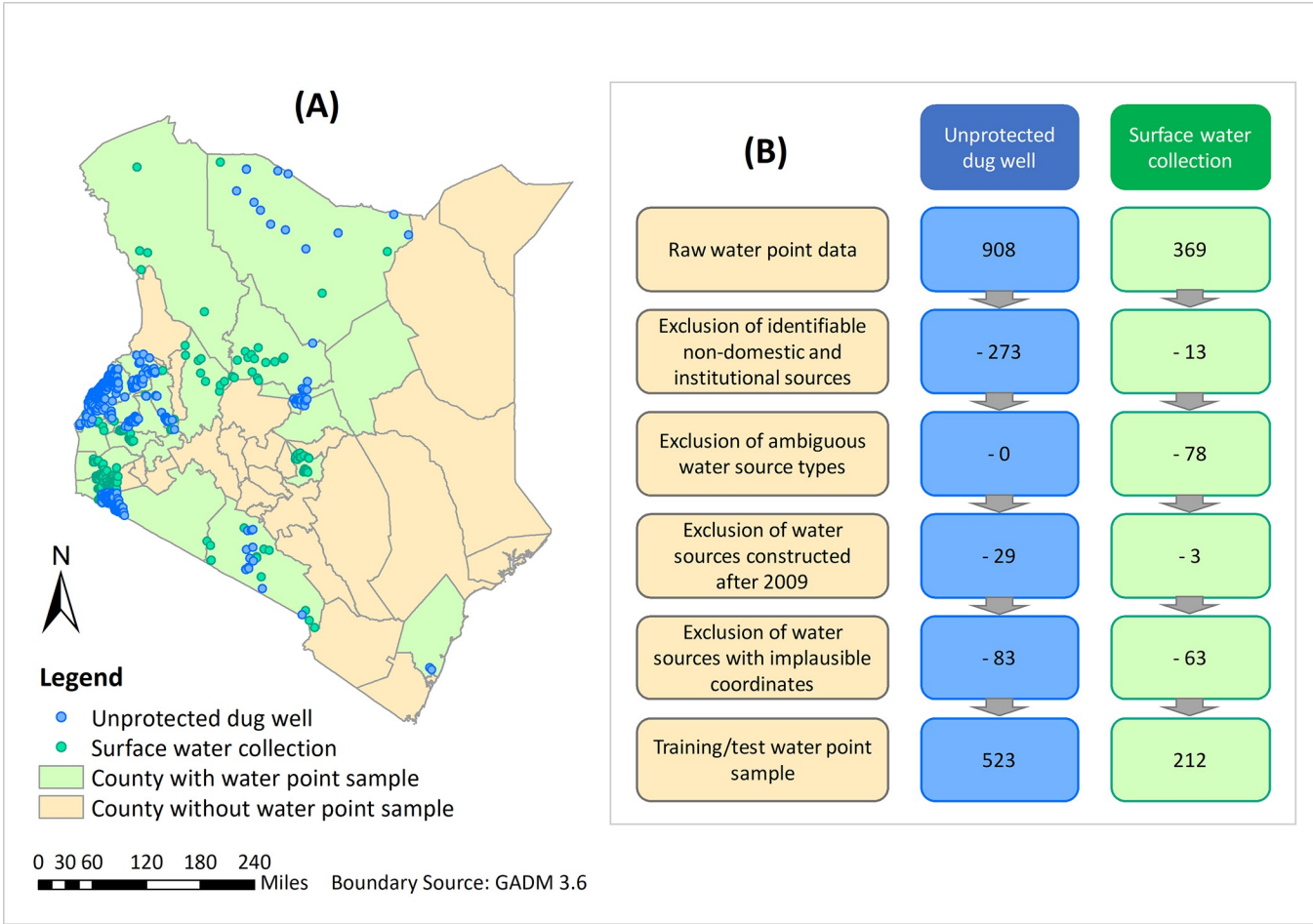

Fig 1. Unprotected dug wells and surface water points included for the MaxEnt modelling in this study. (A) Map showing the distribution of included unprotected dug wells and surface water points; (B) numbers of water points excluded during data pre-processing based on the exclusion criteria.

https://doi.org/10.1371/journal.pone.0216923.g001

Table 1. Conceptual framework of potential factors influencing the distribution of unprotected dug wells and surface water.

Type of factor	Perspective	Unprotected dug well	Surface water
Environmental factors	Original source of water	Potential presence of available groundwater according to hydrogeological and geological conditions (e.g. groundwater productivity; depth to groundwater; etc.)	Presence of available surface water (e.g. rivers, streams, lakes, ponds, etc.)
Technological factors	Technological preferences and avoidance	Feasibility of manual digging; ease of access for construction; avoidance of contamination hazards	N/A
Socio-economic factors	Demand for water	Presence of people without alternative water sources, or with alternative water sources that are not reliable (e.g. of poor quality, poor accessibility, with supply interruptions, etc.)	Presence of people without alternative water sources, or with alternative water sources that are not reliable (e.g. of poor quality, poor accessibility, with supply interruptions, etc.)
Socio-economic factors	Consumer accessibility	Ease of access for water collection (e.g. shorter walking-distance or water fetching times; no restriction due to land ownership; etc.)	Ease of access for water collection (e.g. less walking-distance; less time to spend; no restriction due to land ownership; etc.)
Socio-economic factors	Local preferences	Affordable water source; good perceived water quality; avoidance of places of perceived local importance	Affordable water source; good perceived water quality; potential avoidance of interference with places of high local importance

https://doi.org/10.1371/journal.pone.0216923.t001

2.3 Predictive covariates

We identified predictive covariates that may affect the distribution of unprotected dug wells and surface water based upon our conceptual framework (**Table 1**): surface water or groundwater availability, feasibility of shallow well construction, and compatibility of water source type with local socio-economic conditions. For all predictive covariates, we selected data sources from as close to the year 2009 as possible. When a covariate layer lacked temporal meta-data, we assumed that the state of that covariate did not change significantly over time.

For environmental and technological factors, we selected covariates characterising the hydrogeological and geological environment which relate to groundwater and surface water availability and in turn affect feasibility of shallow well construction [48–52,54,55,69–72], including depth to groundwater, groundwater productivity, groundwater storage, drainage density, elevation, slope, topographic wetness index (TWI), proximity to inland water, land use, lithology, and soil texture. For unprotected dug wells, we created groundwater productivity and storage covariate layers using the Surficial Geology of Africa data developed by U.S. Geological Survey (USGS), Central Energy Resources Team as geological base map (nominally at 1:5,000,000), subsequently rasterising this layer at 1km spatial resolution. Detailed aquifer types, groundwater productivity and storage maps were defined for this layer with reference to 5 km resolution quantitative digital groundwater maps of Africa [73]. Polygons smaller than the 5 km grid resolution were characterised by visual comparison with the hydrogeological information published on the British Geological Survey (BGS) channel (http://earthwise.bgs.ac.uk).

For socio-economic factors, the main covariates employed in this study include Euclidean distance to 1km grid cells containing buildings, town centres, villages and roads in a consideration of accessibility and proximity to human settlements. For surface water, we also created a gridded layer of cost distance (i.e. walking time) to inland water based on slope and land use to reflect ease of access for water collection, as this is an important criteria for households when selecting water sources [74,75]. In addition, since healthcare facilities are often considered a strong predictor of population presence [76], we assumed that they may in turn correlate with constructed water sources and therefore included Euclidean distance to healthcare facilities in modelling unprotected dug wells. Furthermore, we searched for map layers reflecting planning restrictions on well development [77], using Euclidean distance to areas protected for conservation as one such covariate. Furthermore, given the many accounts of disparities in unimproved source use between urban and rural areas and between rich and poor [67,74,75], we also included poverty defined by a Multidimensional Poverty Index (MPI) [78] and urban/rural settlement areas in both models.

All covariate layers were prepared at a spatial resolution of 30 arc-seconds (approximately 1km at the equator) showing terrestrial areas only, excluding large water bodies and all covariate map layers used the same extent and resolution as this layer. In addition, we calculated the Pearson's correlation [79] between continuous covariates; Polychoric correlation [80] between ordinal categorical covariates; and Polyserial correlation [80] between continuous and ordinal categorical covariates for each model. This analysis was to identify and remove strongly correlated covariate pairs (correlation coefficient > 0.7 or < -0.7) to reduce collinearity, as recommended in a previous study [57], retaining the variable in each pair most obviously related to water point distribution. All data pre-processing was carried out using ArcGIS 10; whilst all correlation analyses were carried out using R 3.4.0, where Polychoric and Polyserial correlations were computed with the polycor package [81]. Details of the predictive covariates used and corresponding data sources can be found in S1 File.

2.4 Model implementation

Fig 2 depicts the processing flow for the MaxEnt modelling adopted in this study. Each model was built using a random sample of 70% of the included water points. The remaining 30% of water point presences were used to test model performance based on Monte Carlo cross-validation [82]. We applied two different strategies to control for spatial variation in water point mapping effort and resultant sampling bias, namely use of a bias file and a restricted background correction [83]. For the first method, we generated 10,000 background points by randomly selecting points within the entire study area, weighted by a bias file. A kernel density surface derived from locations of all obtained water points was used as the bias file, assuming that survey effect was concentrated around these known water points. For the second method, we restricted the selection of 10,000 background points to 100 km buffer areas around water point locations. For both unprotected dug wells and surface water, we repeated the sampling of training and test points, model fitting, model evaluation and prediction 50 times for each bias correction strategy, and then computed aggregated predictions and model performance metrics from all 50 model runs.

Model fitting and analysis were carried out using Maxent v3.3.3k [22]. We kept all non-correlated covariates identified by our framework in this illustrative study, since one of our aims is to examine the potential environmental and socio-economic drivers of domestic water access distribution. In addition, we selected all available functional transformations of the predictive covariates (i.e. 'linear', 'quadratic', 'product', 'threshold', 'hinge', and 'discrete') [57] to capture the potential non-linear relationships between the covariates and target water points, to allow for complex relationships between socio-ecological factors and locations of water points. To optimise MaxEnt model complexity, we selected the best model corresponding to

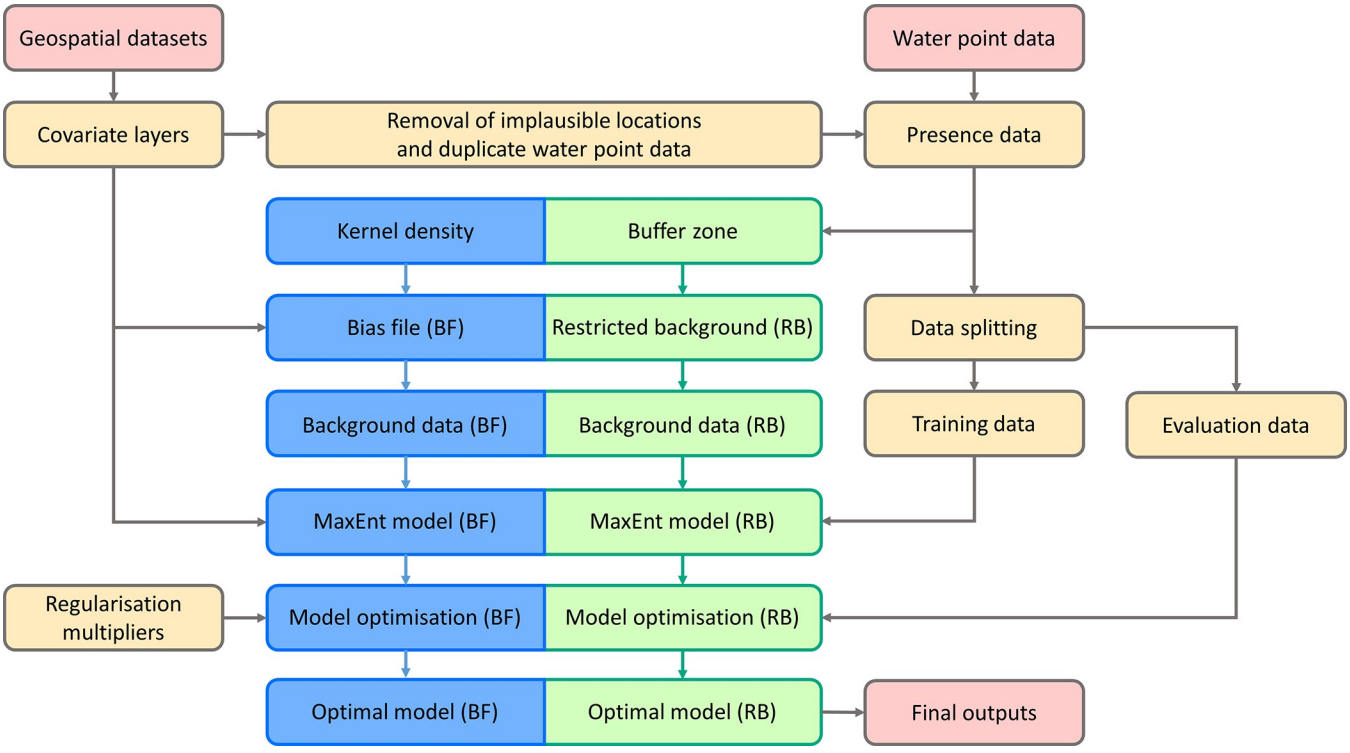

Fig 2. Flow chart of the MaxEnt modelling methodology.

https://doi.org/10.1371/journal.pone.0216923.g002

the regularisation multiplier with the largest AUC instead of using information criteria (AIC, AICc, BIC), due to the known issues with threshold and hinge features [84]. For the purposes of simple illustration and reduction of computational cost, we restricted our test of regularisation multipliers to integers from 1 to 10 based on the results of 50 replicated runs of each model, and selected the best model accordingly. We kept all other settings at their default value, and used logistic output for easier interpretation. However, as this logistic output arbitrarily assumes an occurrence prevalence of 0.5 [57], we interpreted this as the relative ranking of each cell across the landscape rather than true occurrence probabilities. The changes in regularised gains (percent contribution), changes in training AUC based on permuted data (permutation importance), and Jackknife test were used to evaluate the contribution of the covariates to the MaxEnt predictions.

2.5 Performance evaluation

Evaluation of model performance was carried out using Area Under the Receiver Operator Curve (AUC) [85] from the 30% testing sample for the 50 replicated runs. The AUC value is in the range of 0.5 to 1.0, where 1.0 reflects perfect discriminatory power, whilst 0.5 indicates that the prediction failed to capture any patterns and was no better than a random distribution. As the AUC value approaches 1.0, it indicates potentially useful discrimination by the model [58,86].

As an additional means of model evaluation, we calculated the density (persons per hectare) of population using unprotected dug wells or surface water as their main drinking-water source at the most disaggregated levels available via open data sources. For surface water, county level (post-2013 administrative level 1) data from the 2009 Kenya Population and Housing Census was acquired from the Kenya Open Data portal (http://opendata.go.ke/). For unprotected dug wells, publicly available census data did not distinguish water wells from boreholes and springs, so we acquired 2008–09 DHS data at regional level (pre-2013 administrative level 1) [87]. We used Spearman's Rho to examine the correlation across sub-national areas between water consumer density and (a) the density of 'raw' input water points; (b) the model output, namely the average predicted relative occupancy/suitability in populated areas. The populated areas were defined based on the Global Human Settlement (GHS) grid [88].

2.6 Ethics statement

This study involved use of aggregated nationally representative household survey data and population and housing census data from openly accessible data sources; it did not entail any work with records concerning individual human subjects nor collect any data. The research proposal (ID 18551) was submitted to the University of Southampton's Ethics and Research Governance Online (ERGO) system on 10th December 2015 and was reviewed and approved by the ethics committee on 15th January 2016.

3. Results

3.1 Model output and performance evaluation

In this illustrative study, 1 is found to be the optimal regularisation multiplier among the tested integers for both types of water point and bias correction methods. Our models display high predictive power according to the AUC values (all above 0.9), which suggests that the predictions successfully captured relationships between water points and relevant covariates. For unprotected dug wells (Fig 3A), a substantial area of western and central Kenya was predicted to have higher unprotected dug well occupancy, with isolated spots of relatively high

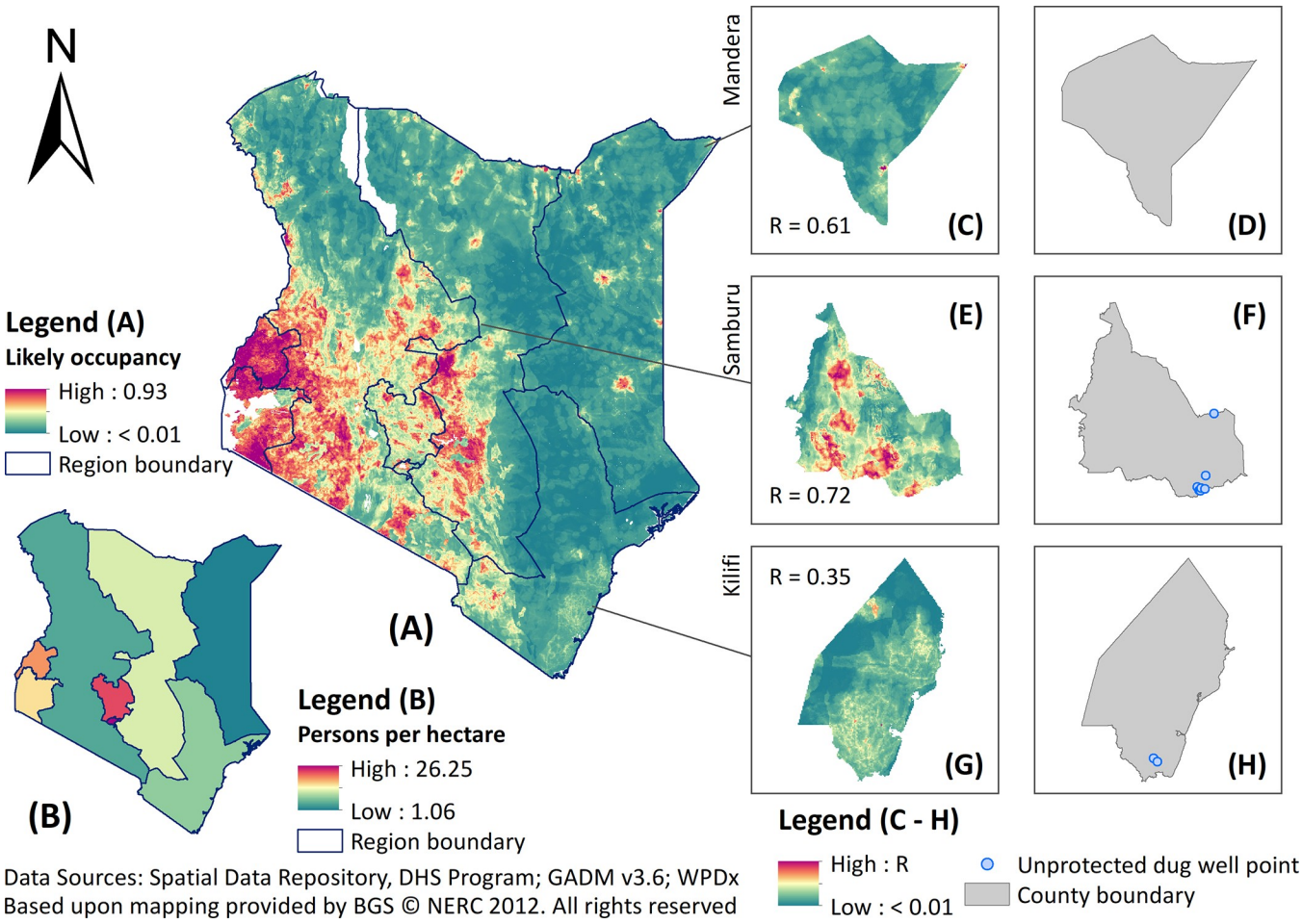

Fig 3. Predicted unprotected dug well occupancy across Kenya with inset maps highlighting Mandera, Samburu, and Kilifi counties. (A) Unprotected well occupancy predicted by MaxEnt at 1 km resolution; (B) DHS-based consumer density map of unprotected dug wells at the regional level; inset maps show MaxEnt prediction versus the initial water point coverage in Mandera (C, D), Samburu (E, F), and Kilifi (G, H) counties.

https://doi.org/10.1371/journal.pone.0216923.g003

occupancy in towns and urban centres in the north and northeast parts of the country, as well as the coastal regions in the south. For surface water (Fig 4A), the areas predicted to have high likely occupancy were also concentrated in western and central Kenya, but sparsely distributed in eastern Kenya. In arid regions lacking mapped water points, such as the insecurity-affected counties bordering Somalia, both models showed limited variation in predicted relative occupancy values relative to western and central parts of Kenya, which had more input water points. The final predictions for unprotected dug wells show patterns consistent with water consumer density (persons per hectare) derived from 2008–09 DHS (Fig 3B; $r_s = 0.714$), whilst those for surface water were consistent with 2009 census data (Fig 4B; $r_s = 0.722$). These correlations with water consumers reported in demographic data were greater than those calculated using the 'raw' input water points ($r_s = -0.073$ for unprotected dug wells and $r_s = 0.188$ for surface water). This suggests some success in gap-filling the initial, incomplete spatial coverage of water point mapping, as detailed in the inset maps in Figs 3 and 4 for example. S1 File provides further details about each prediction and associated uncertainty maps by source type and bias correction method.

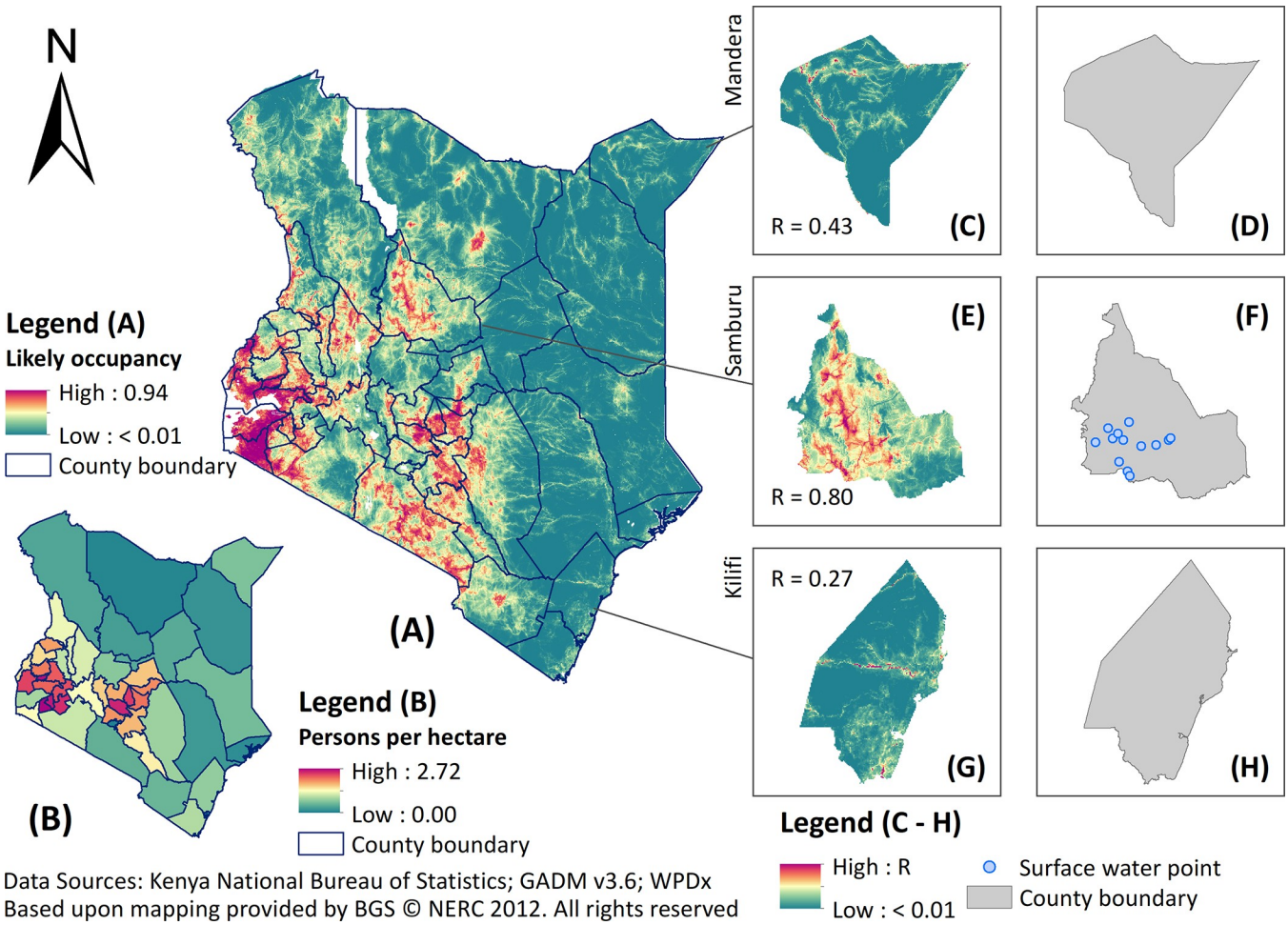

Fig 4. Predicted surface water occupancy across Kenya with inset maps highlighting Mandera, Samburu, and Kilifi counties. (A) Surface water occupancy predicted by MaxEnt at 1 km resolution; (B) census-based consumer density map of surface water at the county level; inset maps show MaxEnt prediction versus the initial water point coverage in Mandera (C, D), Samburu (E, F), and Kilifi (G, H) counties.

https://doi.org/10.1371/journal.pone.0216923.g004

3.2 Covariate contribution analysis

The relative contribution of the covariates varied, depending on bias correction method. For the unprotected dug wells model based on the restricted background correction, percent contribution analysis indicated that urban-rural divide and groundwater storage provided the most useful information, with elevation closely following (Fig 5, data in blue bars). However, elevation had the greatest contribution to the unprotected dug wells model when the bias file correction was applied, whilst urban-rural divide only had a moderate contribution. Groundwater storage and Euclidean distance to towns and urban centres were respectively the second and third most important covariates in the model (Fig 5, data in green bars). For surface water (Fig 6), percent contribution analysis indicated annual rainfall was the most influential covariate in the model based on restricted background correction, whilst elevation had the greatest contribution to the model with bias file correction.

Different covariate contribution analysis methods also yielded different results (see S1 Table). In general, however, for unprotected dug wells, elevation, poverty, and urban-rural divide were found to have the most useful information, whilst other covariates such as groundwater storage, drainage density, Euclidean distance to buildings and Euclidean distance to

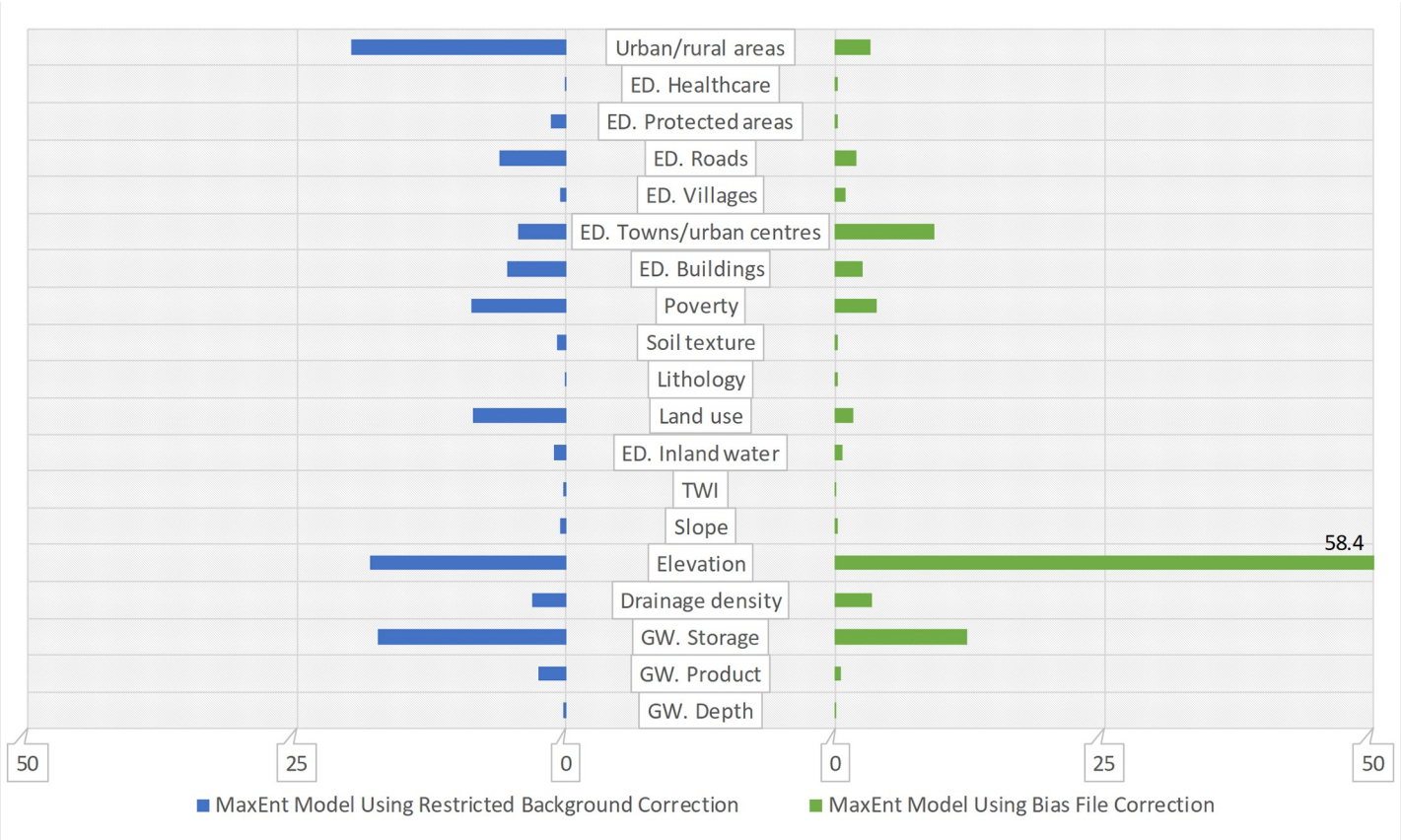

Fig 5. Covariate contribution to the MaxEnt model of unprotected dug wells for two bias correction methods. ED. denotes Euclidean distance; GW. denotes groundwater; the x-axis is the percent contribution of the predictive covariate based on changes in regularised gain.

https://doi.org/10.1371/journal.pone.0216923.g005

towns and urban centres were also very important to the models. For surface water, elevation, annual precipitation and drainage density were found to be the most important covariates, whilst Euclidean distance to buildings and cost distance to inland water also appeared relatively important. A full list of the response curves of predictive covariates can be found in S1 Fig.

4. Discussion

For the first time, our study has applied a spatial ecology technique, namely maximum entropy modelling, to map point locations of access to domestic water supplies. The output gridded map layer of relative likely occupancy adjusts the raw input water point mapping data to account for incomplete coverage, which forms the principle advantage of the technique. Incompleteness is a known issue with some water point mapping data sets: for example, Yu et al. [9] found that Tanzanian and Cambodian databases contained insufficient water points to account for the number of households reporting use of such water sources in census data, which therefore undermined their utility in national level monitoring of water access. The frequency of missing values in water point attribute fields (including water source type) is often high [89], further exacerbating incompleteness when enumerating water points of a particular type. By examining the relationship between covariates and the presence of water points, this spatial predictive modelling technique adjusts for such coverage gaps.

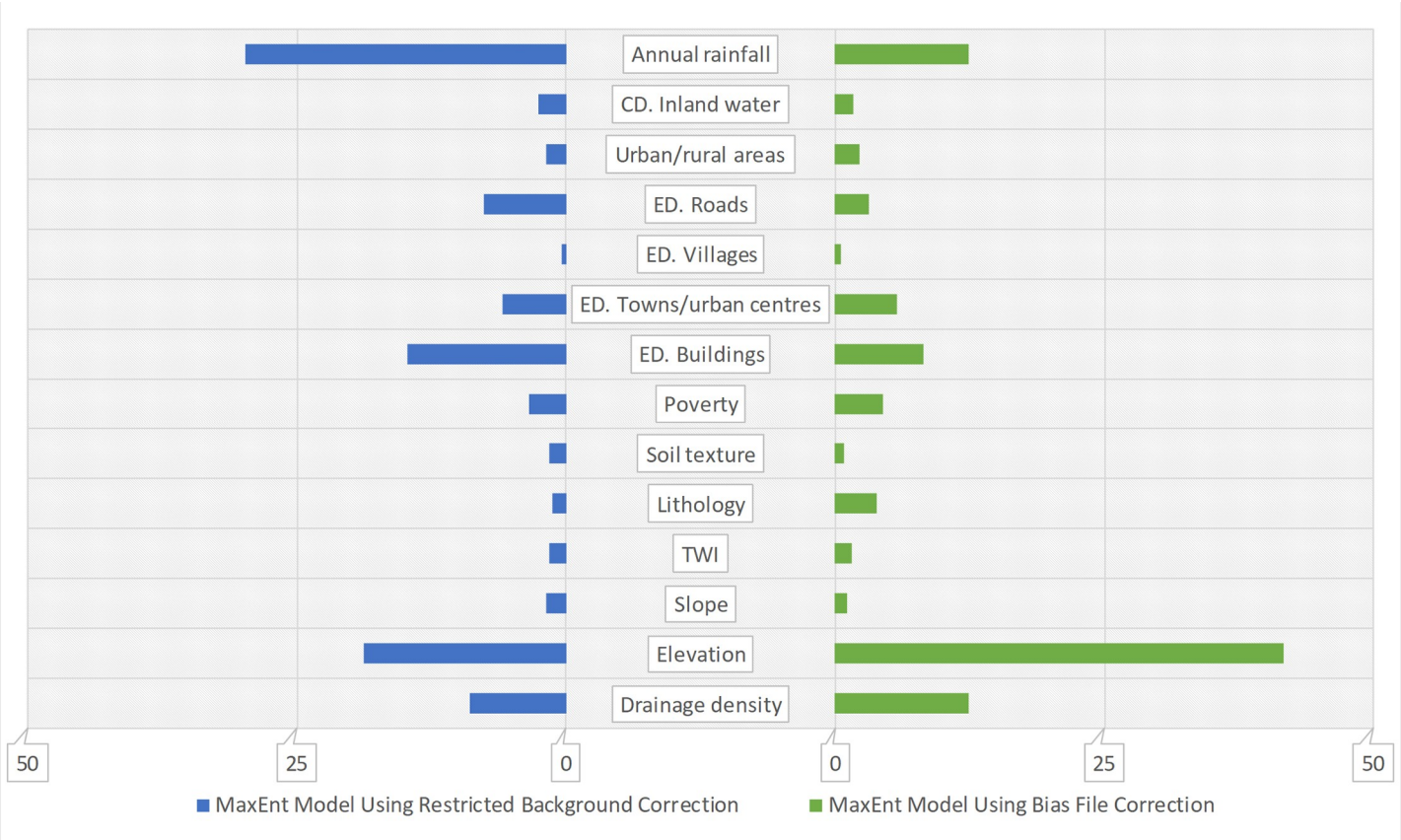

Fig 6. Covariate contribution to the MaxEnt model of surface water for two bias correction methods. CD. denotes cost distance. ED. denotes Euclidean distance; the x-axis is the percent contribution of the predictive covariate based on changes in regularised gain.

https://doi.org/10.1371/journal.pone.0216923.g006

Spatial predictive modelling techniques have been widely employed in different fields for depicting spatial distributions with incomplete locational observations. Although previous studies suggest algorithms based on bioclimatic envelopes [10,11], Gower distance [12] and Mahalanobis distance [13,14] may be unsuitable for categorical covariates, have poor predictive performance, and thus be unsuitable for this application, this illustrative assessment with a MaxEnt model indicates the potential for using other novel spatial predictive models in isolation or in ensemble to map access to domestic water supplies. Some novel methods such as random forests require presence-absence data, but these can still be used with water point data by treating cells unoccupied by water points as pseudo-absence data. Regardless of which method is used, however, the outputs need to be interpreted with caution because no absence data are available.

In this study, MaxEnt's logistic output was used instead of the raw output for easier interpretation, since the logistic output values typically correspond better with predicted relative site occupancy/suitability in comparison with the raw output. However, due to the arbitrarily assumed occurrence prevalence (0.5) in the calculation of the logistic output [57,90], the predicted likely occupancy values only indicate ranked relative site suitability, as they are theoretically the order-preserved transformations of the probability of water point occurrence. In addition, due to the use of background data rather than true absence data, a predicted likely occupancy value should not be considered as the true probability of occurrence.

The gridded output surfaces could be used in several ways. In the same way that ecologists have successfully used such model outputs to target follow-up surveys to find unmapped species occurrences [91], so the output map layers could potentially be used to target follow-up surveys of areas likely to contain unmapped water points. The output layer could also be compared to similar, gridded layers depicting household use of different water source types [8], which have been generated in several countries by applying geostatistical methods to household survey data. The WHO/UNICEF core question on drinking-water used in household surveys asks 'what is the main source of drinking-water for members of your household?' [92] and therefore captures only the main water source for drinking purpose. In contrast, water point mapping captures secondary, seasonal, and domestic water sources for purposes other than drinking. Gridded outputs derived from water points versus household surveys would thus potential capture different dimensions of domestic water resource use. Furthermore, because of the need to protect data concerning human subject, household survey GPS locations are randomly displaced [93], which restricts the resolution of derived gridded representations of household source use to 5km by 5km grid cells. In contrast, this data protection issue does not affect the spatial precision of water points, since they do not relate to human subjects. The MaxEnt output could also be combined with map layers depicting aquifer vulnerability [94] to contamination or geogenic contaminants [95], so as to identify areas where water users are exposed to such hazards, since relevant map layers often come in the gridded format.

Our case study predicted high relative occupancy of unprotected wells in Kenya's densely populated urban areas. This is consistent both with the high population densities and high unprotected well use reported in the Langas slum in Eldoret [96] and in Kisumu's slums [97]. Since many rivers and streams in semi-arid and arid Kenyan counties are seasonal rather than perennial, predicted direct domestic consumption of surface waters in these counties may also be seasonal.

A secondary potential benefit of the technique is the understanding gained of the association between covariate map layers and the spatial distribution of water points. Given ever-increasing demand on water resources [98], such insights into landscape-level characteristics associated with water resource exploitation could prove valuable. In this Kenyan case study, a high percentage contribution to surface water and unprotected well occurrence was attributed to elevation. A similarly strong correlation with elevation has been found for population density [99,100], so this apparent association between elevation and water points may reflect human settlement patterns. Globally, according to census data, unimproved source use, particularly direct consumption of surface waters, is known to be higher among rural populations and the poor [64]. This is reflected in the high covariate contributions for map layers depicting poverty in both sets of models and a high covariate contribution for rurality in the models based on restricted background correction. Unsurprisingly, rainfall and surface drainage density also had great covariate contributions to the surface water model.

In general, the response curves are consistent with our understanding of the hydrogeological and socio-economic determinants of unimproved water access distribution. However, some model covariates exhibit complex, random or somewhat implausible response curves. Some covariates included in the model for unprotected dug well and surface water collection (see S1 Fig) exhibit such response curves, likely due to potential sampling biases. However, such covariates had only small contributions to the final model, reducing concern over their impact on predictions. Given that we wished to illustrate application of a simplified model to water point mapping here, we did not attempt to reduce model complexity to resolve such issues, though further optimisation would be possible. The spatial ecology literature identifies several known limitations of MaxEnt modelling. The output is highly dependent on the spatial distribution of the input occurrences and choice of covariates. In our case study, the human

settlements layer that we used as a covariate does not differentiate slums from other urban areas, which affects predicted relative occurrence of unprotected wells in such areas. Our Kenyan case study used very coarse spatial resolution hydrogeological data, which may have reduced the apparent association between unprotected well occurrences and hydrogeology. In addition, some covariate layers (e.g. annual rainfall, inland water, etc.) dated from different periods to the water point survey dates. Such differences in temporal coverage could have reduced the strength of association between these layers and water point occurrences. Similarly, in our case study implementation, the input water points are heavily concentrated in only a few Kenyan counties. A particular concern where the input occurrence data are concentrated in a small area is out-of-sample prediction whereby the modelled output makes predictions for areas that experience combinations of environmental and socio-economic conditions not found in the training data [63]. In our case study implementation, for example, the training data include very few water points with elevations below 500 metres. More generally, the quality of the input water point data will affect the output gridded probability layer. An example of one such issue would be the misclassification of water source types arising from difficulty in differentiating protected versus unprotected wells or boreholes versus protected wells.

Aside from these issues affecting model calibration, there are further drawbacks of this approach to analysing water points. One difficulty is in evaluating the output relative occurrence grids, which we attempted via census and survey data in the Kenyan case study implementation. However, the coarse spatial resolution of household survey data in particular somewhat limits the usefulness of this evaluation. Apart from the WPDx there are other water point databases for Kenya [101,102], but differences between databases in the water source typologies used inhibit their usefulness for model validation. Thus, unprotected wells can be unambiguously identified in WPDx records, but not in the other data set. Points where surface water is used for domestic purposes could be corroborated via remotely sensed images of such surface water bodies, though not all surface water bodies will be used by households. In future, evaluation of model outputs by an expert panel may therefore be more effective than data-driven model validation. As with all data products generated through complex analysis protocols, a further difficulty is in communicating the nature of the product and the underlying data processing to a wider audience.

Although our case study implementation of the methodology in Kenya may be subject to these limitations, there are freely available, easy-to-use interfaces to both MaxEnt and other related environmental niche modelling techniques [103]. It would therefore be possible for other research groups with access to more appropriate covariate map layers or more representative, higher quality input water point data sets to refine the methodology and address any limitations in input data or covariate choices in our case study implementation. However, appropriate model development and parameterisation require a strong understanding of the technique's underlying principles [57]. Subject to sufficient expertise in the software's use, the approach could thus be customised to local conditions and data availability and, given the existence of a global water point data exchange platform [104], is potentially also scaleable. In our case study, we used a single 'presence-background' technique (MaxEnt) for illustration, but many ecologists use an ensemble approach, whereby multiple environmental niche modelling techniques are used in combination [105,106]. An ensemble approach to water point analysis would thus be a logical future methodological refinement. Although we have applied the method to just two types of water point here, in principle it could be applied to other types such as boreholes, kiosks, or rainwater harvesting systems. Mobile forms of water provision, such as water sold from tanker truckers or vendors using carts, are however inherently difficult to capture via water point mapping. In ecology, environmental niche modelling has proved more successful for endemic or specialist species occupying narrow environmental niches,

rather than more generalist species found in many environments [106,107]. This implies that the technique may perform better for water source types found only in a narrow range of socio-economic and environmental conditions, as opposed to source types that are installed under wide-ranging conditions.

5. Conclusion

This illustrative study takes a technique widely used in spatial ecology to analyse incomplete occurrence data and applies it to two types of water point in Kenya. The technique has potential to correct water point databases for incomplete survey coverage and provide insights into environmental and socio-economic characteristics associated with water points as landscape features. However, the spatial ecology literature also highlights some important limitations of the approach. These include the potential pitfalls of making predictions for environmental conditions not represented in training data and poorer model performance when predicting occurrences of generalist species (or here water points) found in a wide range of environments. Although we only applied the maximum entropy model in Kenya, the methodology could potentially be adapted to other predictive modelling algorithms, settings, types of water points and local data availability. It is also potentially scaleable given the existence of a global water point mapping data exchange, but its further uptake requires expert knowledge and strong understanding of the underlying principles of ecological niche understanding.

Supporting information

S1 File. A document containing additional information on covariate layers, data sources, and output prediction surfaces by water source type and bias correction method.
(PDF)

S1 Table. A Microsoft Excel file containing tables that describe covariate contributions by water source type and bias correction method.
(XLSX)

S1 Fig. A document containing graphs of response curves by predictive covariate, by water source type and bias correction method.
(PDF)

Acknowledgments

We gratefully thank Professor Felix Eigenbrod from the University of Southampton and Dr Ricard Giné Garriga from the Universitat Politècnica de Catalunya (Polytechnic University of Catalonia) for their comments on the manuscript.

Author Contributions

Conceptualization: Weiyu Yu, Nicola A. Wardrop, Robert E. S. Bain, Jim A. Wright.

Data curation: Weiyu Yu, Jim A. Wright.

Formal analysis: Weiyu Yu.

Methodology: Weiyu Yu, Nicola A. Wardrop, Robert E. S. Bain, Victor Alegana, Laura J. Graham, Jim A. Wright.

Resources: Weiyu Yu, Victor Alegana.

Supervision: Nicola A. Wardrop, Robert E. S. Bain, Victor Alegana, Laura J. Graham, Jim A. Wright.

Validation: Weiyu Yu, Jim A. Wright.

Visualization: Weiyu Yu, Victor Alegana.

Writing – original draft: Weiyu Yu, Jim A. Wright.

Writing – review & editing: Weiyu Yu, Nicola A. Wardrop, Robert E. S. Bain, Victor Alegana, Laura J. Graham, Jim A. Wright.

References

1. Bartram J, Cairncross S. Hygiene, sanitation, and water: Forgotten foundations of health. PLoS Med. 2010; 7: 1–9. https://doi.org/10.1371/journal.pmed.1000367 PMID: 21085694
2. United Nations General Assembly. The human right to water and sanitation (A/RES/64/292). 2010.
3. United Nations General Assembly. Transforming our world: The 2030 agenda for sustainable development (A/RES/70/1). 2015. https://doi.org/10.1007/s13398-014-0173-7.2
4. Paes de Barros R, Ferreira FHG, Vega JRM, Chanduvi JS. Measuring Inequality of Opportunities in Latin America and the Caribbean. Washington, DC: The World Bank; 2009.
5. WHO, UNICEF. Report of the task force on monitoring inequalities for the 2030 sustainable development agenda [Internet]. 2017. Available: https://washdata.org/report/jmp-taskforce-monitoring-inequalitiesmeeting-report
6. Bartram J, Brocklehurst C, Fisher MB, Luyendijk R, Hossain R, Wardlaw T, et al. Global monitoring of water supply and sanitation: History, methods and future challenges. Int J Environ Res Public Health. 2014; 11: 8137–8165. https://doi.org/10.3390/ijerph110808137 PMID: 25116635
7. Pullan RL, Freeman MC, Gething PW, Brooker SJ. Geographical inequalities in use of improved drinking water supply and sanitation across sub-Saharan Africa: Mapping and spatial analysis of cross-sectional survey data. PLoS Med. 2014; 11. https://doi.org/10.1371/journal.pmed.1001626 PMID: 24714528
8. Gething P, Tatem A, Bird T, Burgert-Brucker CR. Creating spatial interpolation surfaces with DHS Data, DHS Spatial Analysis Reports No. 11 [Internet]. Rockville, Maryland, USA; 2015. Available: http://dhsprogram.com/publications/publication-SAR11-Spatial-Analysis-Reports.cfm
9. Yu W, Wardrop NA, Bain RES, Wright JA. Integration of population census and water point mapping data—A case study of Cambodia, Liberia and Tanzania. Int J Hyg Environ Health. 2017; https://doi.org/10.1016/j.ijheh.2017.04.006 PMID: 28506523
10. Nix HA. Biogeographic analysis of Australian elapid snakes. In: Longmore R, editor. Atlas of Elapid Snakes of Australia Australian Flora and Fauna Series No 7. Canberra: Australian Government Publishing Service; 1986. pp. 4–15.
11. Busby JR. BIOCLIM—A Bioclimate Analysis and Prediction System. In: Margules CR, Austin MP, editors. Nature conservation: cost effective biological surveys and data analysis. 1991.
12. Carpenter G, Gillison AN, Winter J. Domain—a Flexible Modeling Procedure for Mapping Potential Distributions of Plants and Animals. Biodivers Conserv. 1993; 2: 667–680.
13. Mahalanobis PC. On the generalized distance in statistics. Proceedings of the National Institute of Science of India. 1936.
14. Clark JD, Dunn JE, Smith KG. A Multivariate Model of Female Black Bear Habitat Use for a Geographic Information System. J Wildl Manage. 1993; 57: 519–526.
15. McCullagh P, Nelder J a. Generalized Linear Models, Second Edition [Internet]. 2nd ed. London: Chapman and Hall; 1989. p. 532. https://doi.org/10.1007/978-1-4899-3242-6
16. Hastie TJ, Tibshirani RJ. Generalized Additive Models. London: Chapman and Hall/CRC; 1990.
17. Friedman JH. Multivariate Adaptive Regression Splines. Ann Stat. 1991; 19: 1–67. https://doi.org/10.1214/09-AOAS284
18. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. CRC; 1984.
19. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Ann Stat. 2000; 28: 337–407.
20. Breiman L. Random forest. Mach Learn. 2001; 45: 15–32.

21. Stockwell D, Peters D. The GARP modelling system: problems and solutions to automated spatial prediction. Int J Geogr Inf Sci. 1999; 13: 143–158. https://doi.org/10.1080/136588199241391

22. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Modell. 2006; 190: 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

23. Peterson AT, Ortega-Huerta MA, Bartley J, Sánchez-Cordero V, Soberón J, Buddemeier RH, et al. Future projections for Mexican faunas under global climate change scenarios. Nature. 2002; 416: 626–629. https://doi.org/10.1038/416626a PMID: 11948349

24. Thomas CD, Cameron A, Green RE, Bakkenes M, Beaumont LJ, Collingham YC, et al. Extinction risk from climate change. Nature. 2004; 427: 145–148. https://doi.org/10.1038/nature02121 PMID: 14712274

25. Hijmans RJ, Graham CH. The ability of climate envelope models to predict the effect of climate change on species distributions. Glob Chang Biol. 2006; 12: 2272–2281.

26. Fitzpatrick MC, Gove AD, Sanders NJ, Dunn RR. Climate change, plant migration, and range collapse in a global biodiversity hotspot: the Banksia (Proteaceae) of Western Australia. Glob Chang Biol. 2008; 14: 1337–1352.

27. Peterson AT, Robins CR. Using ecological-niche modeling to predict Barred Owl invasions with implications for Spotted Owl conservation. Conserv Biol. 2003; 17: 1161–1165.

28. Mohamed KI, Papes M, Williams R, Benz BW, Peterson AT. Global invasion potential of 10 parasitic witchweeds and related Orobanchaceae. Ambio. 2006; 35: 281–288. PMID: 17240760

29. Ficetola GF, Thuiller W, Miaud C. Prediction and validation of the potential global distribution of a problematic alien invasive species—the American bullfrog. Divers Distrib. 2007; 13: 476–485.

30. Ward DF. Modelling the potential geographic distribution of invasive ant species in New Zealand. Biol Invasions. 2007; 9: 723–735.

31. Anderson RP, Lew D, Peterson AT. Evaluating predictive models of species distributions: criteria for selecting optimal models. Ecol Modell. 2003; 162: 211–232.

32. Ortega-Huerta MA, Peterson AT. Modelling spatial patterns of biodiversity for conservation prioritization in North-eastern Mexico. Divers Distrib. 2004; 10: 39–54.

33. Peterson AT, Sánchez-Cordero V, Martínez-Meyer E, Navarro-Sigüenza AG. Tracking population extirpations via melding ecological niche modeling with land cover information. Ecol Modell. 2006; 195: 229–236.

34. Graham CH, Hijmans RJ. A comparison of methods for mapping species ranges and species richness. Glob Ecol Biogeogr. 2006; 15: 578–587.

35. Moffett A, Shackelford N, Sarkar S. Malaria in Africa: Vector species' niche models and relative risk maps. PLoS One. 2007;2. https://doi.org/10.1371/journal.pone.0000824 PMID: 17786196

36. Masuoka P, Klein TA, Kim H-C, Claborn DM, Achee N, Andre R, et al. Modeling the distribution of Culex tritaeniorhynchus to predict japanese encephalitis distribution in the Republic of Korea. Geospat Health. 2010; 5: 45–57. https://doi.org/10.4081/gh.2010.186 PMID: 21080320

37. Costa J, Peterson AT, Beard CB. Ecologic niche modeling and differentiation of populations of Triatoma brasiliensis neiva, 1911, the most important Chagas' disease vector in northeastern Brazil (Hemiptera, Reduviidae, Triatominae). Am J Trop Med Hyg. 2002; 67: 516–520. PMID: 12479554

38. Beard CB, Pye G, Steurer FJ, Rodriguez R, Campman R, Peterson AT, et al. Chagas disease in a domestic transmisson cycle in Southern Texas, USA. Emerg Infect Dis. 2003; 9: 103–105. https://doi.org/10.3201/eid0901.020217 PMID: 12533289

39. Peterson AT, Shaw J. Lutzomyia vectors for cutaneous leishmaniasis in Southern Brazil: Ecological niche models, predicted geographic distributions, and climate change effects. Int J Parasitol. 2003; 33: 919–931. https://doi.org/10.1016/S0020-7519(03)00094-8 PMID: 12906876

40. Levine RS, Peterson AT, Yorita KL, Carroll D, Damon IK, Reynolds MG. Ecological Niche and Geographic Distribution of Human Monkeypox in Africa. PLoS One. 2007; 2: 1–7. https://doi.org/10.1371/journal.pone.0000176 PMID: 17268575

41. Stumpf A, Kerle N. Remote Sensing of Environment Object-oriented mapping of landslides using Random Forests. Remote Sens Environ. Elsevier Inc.; 2011; 115: 2564–2577. https://doi.org/10.1016/j.rse.2011.05.013

42. Vorpahl P, Elsenbeer H, Märker M, Schröder B. How can statistical models help to determine driving factors of landslides? Ecol Modell. Elsevier B.V.; 2012; 239: 27–39. https://doi.org/10.1016/j.ecolmodel.2011.12.007

43. Park N-W. Using maximum entropy modeling for landslide susceptibility mapping with multiple geoenvironmental data sets. Environ Earth Sci. 2015; 73: 937–949. https://doi.org/10.1007/s12665-014-3442-z

44. Pearce A, Johns J, Hansen N. MaxEnt and Soil Phosphorus Predictions in a Mixed-use Montane Watershed. ASA, CSSA and SSSA International Annual Meetings (2016). Phoenix, AZ, USA; 2016. Available: https://scisoc.confex.com/crops/2016am/webprogram/Paper101228.html

45. Arnold JD, Brewer SC, Dennison PE. Modeling climate-fire connections within the great basin and upper colorado river basin, western united states. Fire Ecol. 2014; 10: 64–75.

46. Devisscher T, Anderson LO, Aragão LEOC, Galván L, Malhi Y. Increased wildfire risk driven by climate and development interactions in the Bolivian Chiquitania, Southern Amazonia. PLoS One. 2016; 11: 1–29. https://doi.org/10.1371/journal.pone.0161323 PMID: 27632528

47. Renard Q, Pélissier R, Ramesh BR, Kodandapani N. Environmental susceptibility model for predicting forest fire occurrence in the Western Ghats of India. Int J Wildl Fire. 2012; 21: 368–379.

48. Falah F, Nejad SG, Rahmati O, Daneshfar M, Zeinivand H. Applicability of generalized additive model in groundwater potential modelling and comparison its performance by bivariate statistical methods. Geocarto Int. Taylor & Francis; 2017; 32: 1069–1089. https://doi.org/10.1080/10106049.2016.1188166

49. Nejad SG, Falah F, Daneshfar M, Haghizadeh A, Rahmati O. Delineation of groundwater potential zones using remote sensing and GIS-based data-driven models. Geocarto Int. Taylor & Francis; 2017; 32: 167–187. https://doi.org/10.1080/10106049.2015.1132481

50. Miraki S, Zanganeh SH, Chapi K, Singh VP, Shirzadi A, Shahabi H, et al. Mapping Groundwater Potential Using a Novel Hybrid Intelligence Approach. Water Resour Manag. Water Resources Management; 2019; 33: 281–302.

51. Naghibi SA, Pourghasemi HR, Dixon B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monit Assess. 2016; 188: 44. https://doi.org/10.1007/s10661-015-5049-6 PMID: 26687087

52. Naghibi SA, Ahmadi K, Daneshi A. Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. Water Resour Manag. Water Resources Management; 2017; 31: 2761–2775. https://doi.org/10.1007/s11269-017-1660-3

53. Rahmati O, Amir S, Shahabi H, Tien D, Pradhan B, Azareh A, et al. Groundwater spring potential modelling: Comprising the capability and robustness of three different modeling approaches. J Hydrol. 2018; 565: 248–261. https://doi.org/10.1016/j.jhydrol.2018.08.027

54. Rahmati O, Kornejady A, Samadi M, Donato A, Melesse AM. Development of an automated GIS tool for reproducing the HAND terrain model. Environ Model Softw. Elsevier Ltd; 2018; 102: 1–12. https://doi.org/10.1016/j.envsoft.2018.01.004

55. Rahmati O, Pourghasemi HR, Melesse AM. Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. Catena. Elsevier B.V.; 2016; 137: 360–372. https://doi.org/10.1016/j.catena.2015.10.010

56. Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography (Cop). 2006; 29: 129–151. https://doi.org/10.1111/j.2006.0906–7590.04596.x

57. Merow C, Smith MJ, Silander JA. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography (Cop). 2013; 36: 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

58. Phillips SJ, Dudík M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography (Cop). 2008; 31: 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

59. Golding N, August TA, Lucas TCD, Gavaghan DJ, van Loon EE, McInerny G. The ZOON R package for reproducible and shareable species distribution modelling. Methods Ecol Evol. 2018; 9: 260–268. https://doi.org/10.1111/2041-210X.12858

60. Thuiller W, Georges D, Engler R, Breiner F. biomod2: Ensemble Platform for Species Distribution Modeling [Internet]. 2016. Available: https://cran.r-project.org/package=biomod2

61. Royle JA, Chandler RB, Yackulic C, Nichols JD. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods Ecol Evol. 2012; 3: 545–554. https://doi.org/10.1111/j.2041-210X.2011.00182.x

62. Hefley TJ, Hooten MB. Hierarchical Species Distribution Models. Curr Landsc Ecol Reports. Current Landscape Ecology Reports; 2016; 1: 87–97. https://doi.org/10.1007/s40823-016-0008-7

63. Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. Divers Distrib. 2011; 17: 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

64. WHO, UNICEF. Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines. Geneva; 2017.

65. Fantom N, Serajuddin U. The World Bank's classification of countries by income, Policy Research Working Paper 7528 [Internet]. 2016. https://doi.org/10.1596/1813-9450-7528

66. WHO, UNICEF. Progress on water and sanitation: 2015 update and MDG assessment. 2015.

67. Kenya National Bureau of Statistics, Ministry of Health, National AIDS Control Council, Kenya Medical Research Institute, National Council for Population and Development, The DHS Program ICF International. Kenya 2014 Demographic and Health Survey final report [Internet]. 2015. Available: https://dhsprogram.com/pubs/pdf/fr308/fr308.pdf

68. WHO, UNICEF. Progress on drinking water and sanitation: special focus on sanitation [Internet]. Geneva & New York; 2008. Available: http://www.wssinfo.org

69. Carter R, Chilton J, Danert K, Olschewski A. Siting of drilled water wells—A guide for project managers. St Gallen, Switzerland: Rural Water Supply Network (RWSN); 2014.

70. Rahmati O, Nazari Samani A, Mahdavi M, Pourghasemi HR, Zeinivand H. Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. Arab J Geosci. 2015; 8: 7059–7071. https://doi.org/10.1007/s12517-014-1668-4

71. Ramu MB, Vinay M. Identification of ground water potential zones using GIS and Remote Sensing Techniques: A case study of Mysore taluk -Karnataka. Int J Geomatics Geosci. 2014; 5: 393–403.

72. Aneesh R, Deka PC. Groundwater Potential Recharge Zonation of Bengaluru Urban District—A GIS based Analytic Hierarchy Process (AHP) Technique Approach. Int Adv Res J Sci Eng Technol. 2015; 2: 129–136.

73. MacDonald AM, Bonsor HC, Dochartaigh BÉÓ, Taylor RG. Quantitative maps of groundwater resources in Africa. Environ Res Lett. 2012; 7: 1–7. https://doi.org/10.1088/1748-9326/7/2/024009

74. Wijk-Sijbesma C van. Participation of women in water supply and sanitation: roles and realities. (Technical Paper Series no.22). The Hague, The Netherlands; 1985.

75. Wijk-Sijbesma C van. Gender in water resources management, water supply and sanitation: roles and realities revisited. The Hague, The Netherlands: IRC International Water and Sanitation Centre; 1998.

76. Stevens FR, Gaughan AE, Linard C, Tatem AJ. Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data. PLoS One. 2015; 10: 1–22. https://doi.org/10.1371/journal.pone.0107042 PMID: 25689585

77. UNICEF. Towards better programming: A Water Handbook. 1999. Report No.: 2.

78. Alkire S, Foster J, Seth S, Santos ME, Roche JM, Ballon P. Multidimensional Poverty Measurement and Analysis. Oxford University Press; 2015.

79. Snedecor GW, Cochran WG. Statistical Methods. 6th ed. Ames, Iowa: The Iowa State University Press; 1968.

80. Drasgow F. Polychoric and polyserial correlations. In: Kotz S, Johnson N, editors. The Encyclopedia of Statistics. Wiley; 1986. pp. 68–74.

81. Fox J. Package> 'polycor' [Internet]. 2016. Available: https://cran.r-project.org/web/packages/polycor/index.html

82. Dubitzky W, Granzow M, Berrar DP. Fundamentals of Data Mining in Genomics and Proteomics. 1st ed. Dubitzky W, Granzow M, Berrar DP, editors. Springer US; 2007. https://doi.org/10.1007/978-0-387-47509-7

83. Fourcade Y, Engler JO, Rödder D, Secondi J. Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. PLoS One. 2014; 9: e97122. https://doi.org/10.1371/journal.pone.0097122 PMID: 24818607

84. Warren DL, Wright AN, Seifert SN, Shaffer HB. Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. Divers Distrib. 2014; 20: 334–343. https://doi.org/10.1111/ddi.12160

85. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44: 837–845. https://doi.org/10.2307/2531595 PMID: 3203132

86. Elith J. Quantitative Methods for Modeling Species Habitat: Comparative Performance and an Application to Australian Plants. In: Ferson S, Burgman M, editors. Quantitative Methods for Conservation Biology. 1st ed. Springer New York; 2000. pp. 39–58. https://doi.org/10.1007/0-387-22648-6_4

87. ICF International Funded by the United States Agency for International Development (USAID). Spatial Data Repository, The Demographic and Health Surveys Program [Internet]. Funded by the United States Agency for International Development (USAID); [cited 27 Sep 2017]. Available: http://spatialdata.dhsprogram.com/

88. Pesaresi M, Freire S. GHS Settlement grid following the REGIO model 2014 in application to GHSL Landsat and CIESIN GPW v4-multitemporal (1975-1990-2000-2015) [Dataset]. European Commission, Joint Research Centre (JRC); 2016.

89. Verplanke J, Georgiadou Y. Wicked Water Points: The Quest for an Error Free National Water Point Database. ISPRS Int J Geo-Information. 2017; 6: 244. https://doi.org/10.3390/ijgi6080244

90. Yackulic CB, Chandler R, Zipkin EF, Royle JA, Nichols JD, Campbell Grant EH, et al. Presence-only modelling using MAXENT: When can we trust the inferences? Methods Ecol Evol. 2013; 4: 236–243. https://doi.org/10.1111/2041-210x.12004

91. Rhoden CM, Peterman WE, Taylor CA. Maxent-directed field surveys identify new populations of narrowly endemic habitat specialists. PeerJ. 2017; 5: e3632. https://doi.org/10.7717/peerj.3632 PMID: 28785520

92. WHO, UNICEF. Core questions on drinking-water and sanitation for household surveys. 2006.

93. Burgert CR, Colston J, Roy T, Zachary B. Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys, DHS Spatial Analysis Reports No. 7 [Internet]. Calverton, Maryland, USA; 2013. DHS Spatial Analysis Reports No. 7

94. Ouedraogo I, Defourny P, Vanclooster M. Mapping the groundwater vulnerability for pollution at the pan African scale. Sci Total Environ. Elsevier B.V.; 2016; 544: 939–953. https://doi.org/10.1016/j.scitotenv.2015.11.135 PMID: 26771208

95. Amini M, Abbaspour KC, Berg M, Winkel L, Hug SJ, Hoehn E, et al. Statistical modeling of global geogenic arsenic contamination in groundwater. Environ Sci Technol. 2008; 42: 3669–3675. PMID: 18546706

96. Kimani-Murage EW, Ngindu AM. Quality of water the slum dwellers use: The case of a Kenyan slum. J Urban Heal. 2007; 84: 829–838. https://doi.org/10.1007/s11524-007-9199-x PMID: 17551841

97. Odiere MR, Opisa S, Odhiambo G, Jura WGZO, Ayisi JM, Karanja DMS, et al. Geographical distribution of schistosomiasis and soil-transmitted helminths among school children in informal settlements in Kisumu City, Western Kenya. Parasitology. 2011; 138: 1569–1577. https://doi.org/10.1017/S003118201100059X PMID: 21679486

98. Mekonnen MM, Hoekstra AY. Four Billion People Experience Water Scarcity. Sci Adv. 2016; 2: 1–7. https://doi.org/10.1126/sciadv.1500323 PMID: 26933676

99. Schumacher J V., Redmond RL, Hart MM, Jensen ME. Mapping patterns of human use and potential resource conflicts on public lands. Environ Monit Assess. 2000; 64: 127–137.

100. Cohen JE, Small C. Hypsographic demography: The distribution of human population by altitude. Proc Natl Acad Sci U S A. 1998; 95: 14009–14014. https://doi.org/10.1073/pnas.95.24.14009 PMID: 9826643

101. International Livestock Research Institute. Water points in nothern Kenya by the German Technical Cooperation (GTZ) [Internet]. [cited 22 May 2018]. Available: http://192.156.137.110/gis/search.asp?id=401

102. International Livestock Research Institute. Distribution of water points, Almanac Characterisation Tool (ACT) database [Internet]. [cited 22 May 2018]. Available: http://192.156.137.110/gis/search.asp?id=312

103. University of Notre Dame. SPACES—Spatial Portal for Analysis of Climatic Effects on Species [Internet]. 2010 [cited 4 Oct 2018]. Available: http://spaces.crc.nd.edu/

104. WPDx. Water Point Data Exchange (WPDx) is the global platform for sharing water point data [Internet]. 2015 [cited 25 Jan 2016]. Available: https://www.waterpointdata.org/

105. Araújo MB, New M. Ensemble forecasting of species distributions. Trends Ecol Evol. 2007; 22: 42–7. https://doi.org/10.1016/j.tree.2006.09.010 PMID: 17011070

106. Grenouillet G, Buisson L, Casajus N, Lek S. Ensemble modelling of species distribution: The effects of geographical and environmental ranges. Ecography (Cop). 2011; 34: 9–17. https://doi.org/10.1111/j.1600-0587.2010.06152.x

107. Guisan A, Zimmermann NE, Elith J, Graham CH, Phillips S, Peterson AT. What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? Ecol Monogr. 2007; 77: 615–630.