



## **Applying spatial regression to evaluate risk factors for microbiological contamination of urban groundwater sources in Juba, South Sudan**


Downloaded from: <https://research.chalmers.se>, 2025-06-18 03:30 UTC

Citation for the original published paper (version of record):

Engstrom, E., Mortberg, U., Karlstrom, A. et al (2017). Applying spatial regression to evaluate risk factors for microbiological contamination of urban groundwater sources in Juba, South Sudan. *Hydrogeology Journal*, 25(4): 1077-1091.  
<http://dx.doi.org/10.1007/s10040-016-1504-x>

N.B. When citing this work, cite the original published paper.

# Applying spatial regression to evaluate risk factors for microbiological contamination of urban groundwater sources in Juba, South Sudan

Emma Engström<sup>1,2</sup>  · Ulla Mörtberg<sup>1</sup> · Anders Karlström<sup>2</sup> · Mikael Mangold<sup>3,4</sup>

Received: 4 May 2016 / Accepted: 20 November 2016 / Published online: 13 December 2016  
© The Author(s) 2016. This article is published with open access at Springerlink.com

**Abstract** This study developed methodology for statistically assessing groundwater contamination mechanisms. It focused on microbial water pollution in low-income regions. Risk factors for faecal contamination of groundwater-fed drinking-water sources were evaluated in a case study in Juba, South Sudan. The study was based on counts of thermotolerant coliforms in water samples from 129 sources, collected by the humanitarian aid organisation Médecins Sans Frontières in 2010. The factors included hydrogeological settings, land use and socio-economic characteristics. The results showed that the residuals of a conventional probit regression model had a significant positive spatial autocorrelation (Moran's  $I = 3.05$ ,  $I$ -stat = 9.28); therefore, a spatial model was developed that had better goodness-of-fit to the observations. The most significant factor in this model ( $p$ -value 0.005) was the distance from a water source to the nearest Tukul area, an area with informal settlements that lack sanitation services. It is thus recommended that future remediation and monitoring efforts in the city be concentrated in such low-income regions. The spatial model differed from the conventional approach: in contrast with the latter case, lowland topography was not

significant at the 5% level, as the  $p$ -value was 0.074 in the spatial model and 0.040 in the traditional model. This study showed that statistical risk-factor assessments of groundwater contamination need to consider spatial interactions when the water sources are located close to each other. Future studies might further investigate the cut-off distance that reflects spatial autocorrelation. Particularly, these results advise research on urban groundwater quality.

**Keywords** Sub-Saharan Africa · Health · Microbial processes · Statistical modeling · Urban groundwater

## Introduction

Human health is at risk when microbes are present in groundwater-fed sources of drinking water. Borchardt et al. (2003) reported that diarrhoea in children in Wisconsin (USA) was correlated with drinking from a household well contaminated with faecal enterococci. Beller et al. (1997) traced an outbreak of gastroenteritis in Alaska (USA) to water consumption from a contaminated well. The disease burden of water-related infectious diseases is the most severe in developing countries (Batterman et al. 2009). In 2010, diarrheal disease caused an estimated 0.8 million deaths in children under the age of 5 years, with approximately half of these occurring in Africa (Liu et al. 2012). Sorensen et al. (2015) detected DNA from the pathogens *Vibrio cholerae* and *Salmonella enterica* (cause of typhoid fever) in 41 and 16% of the analysed samples, respectively, in groundwater in the city of Kabwe, Zambia. In developing countries, groundwater often provides the most important sources of drinking water (Pedley and Howard 1997). In Sub-Saharan Africa, where most of the world's poorest countries are located, understanding of the mechanisms that cause faecal contamination of

Published in the special issue “Hydrogeology and Human Health”

✉ Emma Engström  
emmaeng@kth.se

<sup>1</sup> Department of Sustainable Development, Environmental Science and Engineering (SEED), KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden

<sup>2</sup> Department of Transport Science, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden

<sup>3</sup> Watsan, MSF-OCB, Hai Malakal, Juba, South Sudan

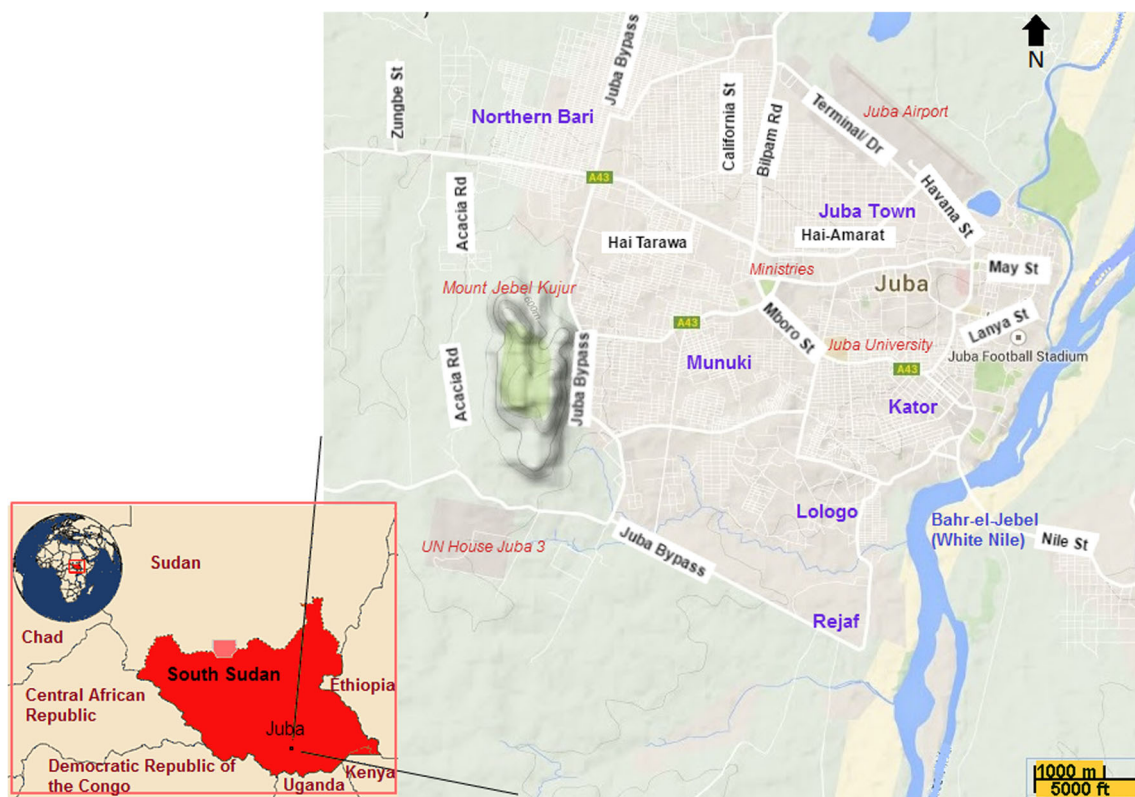
<sup>4</sup> Department of Civil and Environmental Engineering, Chalmers University of Technology, 412 96 Göteborg, Sweden

groundwater sources is still very limited (Kanyerere et al. 2012; Nyenje et al. 2013). It is thus imperative to improve guidelines and practices related to water and sanitation, particularly in Sub-Saharan Africa. For regions that lack water-quality data, the highest priority is to monitor the performance of improved (protected) sources (Abramson et al. 2013). As much as 86% of the population in low-income countries has access to such improved water sources (WHO/UNICEF 2012), which are typically derived from groundwater.

The current study focused on protected groundwater sources used for drinking water in Juba, the capital of South Sudan (Fig. 1). The analysis was based on the incidence of thermotolerant coliforms (TTCs) in samples from boreholes and hand-dug wells, collected in 2010 by the humanitarian aid organisation Médecins Sans Frontières-Belgique (MSF-B). The initial results of these investigations were presented by Engström et al. (2015a): 66% of the investigated sources, including 95 boreholes, were microbiologically contaminated at least once. The local topography and the accumulated long-term antecedent rainfall (5-day and monthly) were statistically associated with contamination events, in contrast with the wellhead drainage efficiency, the distance to the closest latrine, and level of short-term rainfall (Engström et al. 2015a). These findings indicated that the contributing groundwater had been contaminated. Hynds et al. (2014) made the

distinction that there are three processes by which well water contamination can occur: generalized aquifer contamination, localized source-specific contamination due to rapid and/or shallow groundwater pathways, or direct ingress at the well-head. Of these, the study by Engström et al. (2015a) addressed the latter two and the results indicated that direct ingress at the wellhead was not the main process. However, the importance of regional, as opposed to local, factors for groundwater contamination was not investigated. The significance of long-term precipitation suggested that contamination could have been caused by generalized aquifer contamination.

Statistical models provide means to identify risk factors for groundwater contamination—for example, they can indicate the likely route of contaminant entry, inform future well siting and improve the screening of wells (Hynds et al. 2014). They can also help specify where future monitoring efforts are most needed and the results based on a particular site can be used to guide field investigations in other areas with similar hydrogeology and land use (Mair and El-Kadi 2013). Regression-based models are particularly useful in operational contexts (de Brauwere et al. 2014). Their use is common in the literature in studies on risk factors for microbial groundwater contamination, which have focused on: coliform bacteria in rural wells in Iowa, USA (Glanville et al. 1997), the link between *Cryptosporidium* and onsite wastewater systems and private



**Fig. 1** Location map (inset) of the study area based on WHO data (2014b), showing an approximation of actual country borders. Map of Juba from Google (2014), with data on urban subdivisions (purple text)

and landmarks (red text) based on WHO (2014b) and Japan International Cooperation Agency (2009a; b) data

wells in New Mexico, USA (Tollestrup et al. 2014), *Escherichia coli* (*E. coli*) in 211 wells in the Republic of Ireland (Hynds et al. 2014), *E. coli* in groundwater sources in northern, rural Malawi (Kanyerere et al. 2012), coliform bacteria in shallow wells in Ibadan, Nigeria (Oguntoke et al. 2013), TTCs and faecal streptococci in shallow groundwater in Kampala, Uganda (Howard et al. 2003), enterococci and TTCs in shallow groundwater sources in Lichinga, Mozambique (Godfrey et al. 2006), and faecal coliform and faecal streptococci in rural areas in Burkina Faso (Guillemin et al. 1991).

Typically, the data used to develop regression models are assumed to be statistically independent, with residuals between observations and model estimates that are independent and identically distributed (iid). However, spatial data have a tendency to be autocorrelated, which implies that the residuals vary systematically over space (LeSage 2000; Mörtberg and Karlström 2005). If spatial effects are ignored, the estimates of the coefficients and the inferences based on such models might be inaccurate. An important characteristic in the current study was that the sources were located relatively close to each other, which might result in spatial interactions between data points, particularly in the event of regional aquifer contamination. Recently, spatial statistics has received increased attention, with applications in geology, economics and epidemiology (Pinkse and Slade 1998). However, to the authors' knowledge, spatial regression has not been used in research on risk factors for groundwater contamination.

The objectives of the current case study of Juba were to improve understanding of the factors that cause microbiological contamination of protected groundwater sources in areas with tropical climates, low incomes and high population densities and to advance hydrogeological research using statistical modelling as a tool to evaluate mechanisms of urban groundwater pollution. The study investigated the hypothesis that regression models of aquifer pollution should consider spatial autocorrelation when the sources are located near to each other. The risk factor analysis included land use, socio-economic factors and hydrogeological settings.

## Methods

### Case study area

The investigated groundwater sources were located in Juba, South Sudan, north of the equator in Sub-Saharan Africa. The country is afflicted with conflicts. Juba is segregated with locally dense, transient and low-income populations, as portrayed by the United States Agency for International Development (USAID 2005). Sudan's Peace Agreement was signed in 2005, after which Juba experienced unprecedented population growth, inducing the expansion and proliferation

of informal settlements (McMichael 2016). The area has a tropical climate, with a wet season that normally lasts from April through November and a dry season during the rest of the year. In the wet season, the monthly precipitation typically varies from 100 to 200 mm (Fig. 2). The rainfall events have short durations, lasting approximately 2 h, as described by the Japan International Cooperation Agency (JICA 2009a). In the dry season there is little rainfall, and the mean annual precipitation is approximately 1,000 mm. There is little variation in temperature across the wet season and temperatures do not vary largely between the wet and the dry seasons: the minimum monthly average temperature is approximately 20–25 °C, and the maximum monthly average temperature is 30–40 °C throughout the year; nevertheless, the maximum average monthly temperatures are somewhat lower in the wet season, based on 2006 data (JICA 2009a; Fig. 2).

### Hydrogeology

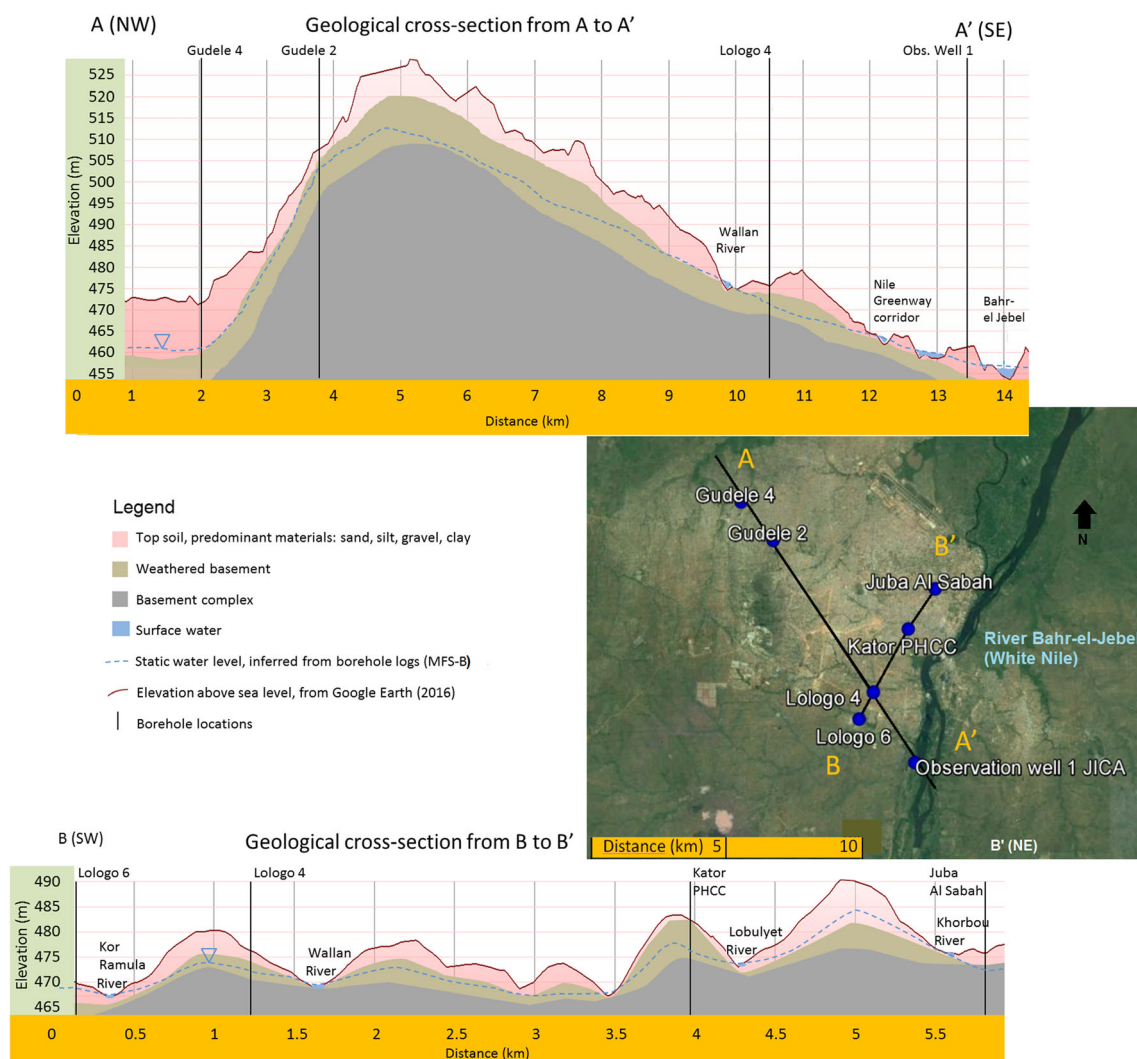
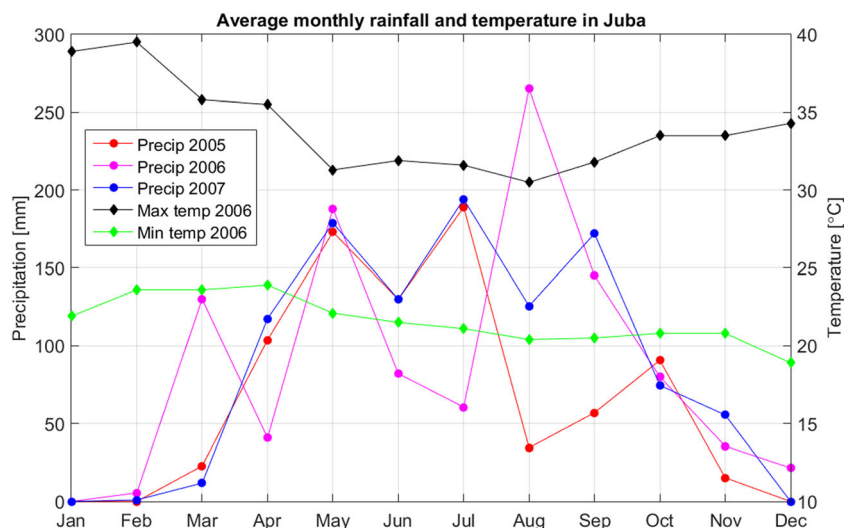
Juba is on an alluvial plain that slopes from the Mount Jebel Kujur foothill (550 m a.s.l.) in the south–southwest towards the Bahr-el Jebel (the White Nile) in the north–northeast (450 m a.s.l.), with an average gradient of 0.5% (Fig. 3). During the wet season, flooding affects approximately half of the city and five seasonal streams, or wadis, appear and flow eastward towards the Nile: Luri, Khorbou, Lobulyet, Wallan and Kor Ramula—from north to south (JICA 2009a). At this time, stagnant water typically covers the area along the river, which has alluvial surfaces, wadi fills and swamp deposits (JICA 2009a). The city is underlain by undifferentiated basement complex, and aquifers are typically located in fracture zones in the weathered basement rock (Panagos et al. 2011; Sudan Ministry of Energy and Mines Geological and Mineral Resources Department 1981; Vail 1989). Superficial groundwater also occurs in sand and gravel layers of alluvial deposits and, at greater depths, local perennial or ephemeral aquifers can take place in thin, saturated layers (JICA 2009a). Groundwater is recharged by rainfall and flooding of river water. As indicated by the sloping topography and the northward flow of the Bahr-el-Jebel, the direction of groundwater flow is expected to be from the south-west towards the north-east. Figure 3 shows a hydrogeological conceptual model of the area with two cross-sections.

### Variables and data sources

The spatial risk factor analysis included site-specific information and regional data, reflecting hydrogeological factors, land use, and socio-economics. The Appendix lists the variables used, their measurement units, and the corresponding reference.



**Fig. 2** Average monthly precipitation and temperature data from JICA (2009a), citing the Sudan Meteorological Authority



**Fig. 3** Hydrogeological conceptual model of Juba, inferred from JICA (2009a) and borehole logs (MSF-B, unpublished data, 2013). Map of Juba and ground elevation data from Google Earth 7.1.7. 04°48.61596'

N, 031°35.08757'E (2016). Lologo, Kator and Gudele are urban subdivisions, PHCC is short for Primary Health Care Corporation facility, and Al Sabah is a children's hospital

### Sample collection and analysis

The water quality data were collected by MSF-B during the wet season, from 6 April to 29 October 2010, with the purpose of identifying boreholes that could potentially spread cholera during outbreak events. Most of the sources were tested on two different dates, with approximately 3 months between sampling. Microbiological contamination was defined as  $>0$  CFU/100 ml, in agreement with the WHO (2011) guidelines for drinking-water quality. To assess faecal contamination, water samples were analysed for TTCs using an Oxfam-DelAgua kit (Oxfam-DelAgua 2009). TTCs are considered acceptable indicators of faecal pollution (WHO 2011), because their populations are dominated by *E. coli* in most environments. The effect of this assumption was previously discussed in Engström et al. (2015a), which further contains a more detailed account of the water sampling procedure and the microbiological analyses.

### Hydrogeological variables

The following hydrogeological characteristics were studied: marshlands, the Bahr-el-Jebel river and its tributaries, elevation above sea level, the local topography, and the static water level. The elevation and catchment areas were extracted using topographical data with  $30 \times 30$  m resolution, based on the ASTER Global Digital Elevation Model (NASA Jet Propulsion Laboratory (JPL) 2011). The local topography was based on an on-site assessment by MSF-B at the time of the water sampling. This factor was represented as a Boolean indicator, set to 1 if a water source was located in a lowland area and 0 otherwise. Its importance was investigated using cross-tabulation, which tests the null-hypothesis that a table is independent in each dimension. The static water level was based on data obtained by MSF-B. Independently of the microbiological examination, groundwater sources were examined in 2008, 2009 or 2010 in MSF-B campaigns of boreholes drilling and rehabilitation in cooperation with the government of Southern Sudan, the Ministry of Cooperatives and Rural Development, and the Directorate of Rural Water (MSF-B, unpublished data, 2013). At these evaluations, the static water level was recorded. The static water levels obtained from the rehabilitation and the drilling protocols from 33 sites were used to estimate the depth-to-groundwater elsewhere in Juba (Fig. 3). The groundwater level was calculated by subtracting the static water level from the ground surface elevation, obtained from the ASTER Global Digital Elevation Model. An inverse distance-weighting algorithm was then applied. The resulting raster was subsequently used to extract the static water levels at the borehole locations that were sampled for coliform bacteria.

### Land use and socio-economic data

Land cover information was defined via reports by USAID (2005) and JICA (2009a; b). Based on maps in those studies, four land cover categories were identified: bush, open ground or grassland, commercial and market areas, and roads or houses. Furthermore, socio-economic data were included using four land class categories, defined by USAID (2005) as follows: informal Tukul areas, which are low-income areas with squatter housing (532 inhabitants per ha); class 3–4 areas, with a transient, low-income housing mix of permanent and temporary materials (266 inhabitants per ha); class 2 areas, with middle-class cottage homes of simple construction, some with sanitation (200 inhabitants per ha); and Class 1 areas, with permanent structures and colonial-style homes with access to formal sanitation (128 inhabitants per ha). Additionally, the on-site hygiene level was accounted for in the regression. It had been categorized into three levels by MSF-B at the time of water sampling, as presented previously (Engström et al. 2015a). There were 129 water sources accounted for in the current study and 147 locations were evaluated in Engström et al. (2015a); however, spatial data could not be obtained for all sources.

### GIS data generation

The spatial features were geographical information system (GIS)-derived using image processing operations on maps. Features were accounted for in variables reflecting shares of circular areas centered on each water source. Different radii were considered to investigate the effect of lateral contaminant transport (30, 100 and 500 m). In some cases, the feature was lacking in the smaller buffers and these radii were omitted from the statistical analysis. The regression also included variables reflecting the Euclidean distance from each water source to the nearest feature.

### Statistical analyses

The statistical associations between contamination and the hydrogeological and anthropological risk factors were investigated. These tests were based on the two-sided Wilcoxon rank-sum test (or Mann-Whitney U-test). This identified the most important risk factors, which were subsequently considered in the multivariable models. The variables with individual significance of  $p < 0.10$  were assessed in these models, in agreement with Mair and El-Kadi (2013) and Hynds et al. (2014). A probabilistic (probit) regression model was developed to estimate the probability of contamination related to these predictors. It included only the factors for which the relationship corresponded with prior theories. The occurrence, defined as the presence/absence of TTCs in 100-ml samples, was considered rather than concentrations, in accordance with

Hynds et al. (2014), motivating a binary model with unquantifiable variability within the system.

### Conventional probabilistic regression

In a probit regression model, the inverse standard normal distribution of the probability is described as a linear combination of the most significant explanatory variables. The conventional probit model assumes that the error terms are iid with constant variance. The probability of contamination of sample  $i$ ,  $p_i$  was thus estimated according to (LeSage and Pace 2009; LeSage 2000):

$$p_i = P(Y_i = 1) = \Phi(\mathbf{x}_i' \boldsymbol{\beta}), \text{ and } 1-p_i = P(Y_i = 0) \\ = 1-\Phi(\mathbf{x}_i' \boldsymbol{\beta}) \quad (1)$$

for  $i = 1, \dots, n$ , where  $Y_i$  is a random variable representing contamination,  $\Phi$  is the cumulative distribution function of the standard normal distribution,  $\mathbf{x}_i$  is a vector of independent explanatory variables for sample  $i$  (assumed to be deterministic),  $\boldsymbol{\beta}$  is a vector of parameters to be estimated, and  $n$  is the number of observations. Thus:

$$\Phi^{-1}(p_i) = \mathbf{x}_i' \boldsymbol{\beta} \quad (2)$$

for  $i = 1, \dots, n$ . An equivalent model, with a latent variable  $Y_i^*$  can be formulated as:

$$Y_i^* = \mathbf{x}_i' \boldsymbol{\beta} + e_i \quad (3)$$

for  $i = 1, \dots, n$ , where the error terms  $e_i$  are iid, and  $N(0, \sigma)$  and  $Y_i$  indicates if  $Y_i^*$  is positive:  $Y_i = 1$  if  $Y_i^* > 0$  and  $Y_i = 0$  otherwise. Thus:

$$p_i(Y_i = 1 | x_i) = p_i(Y_i^* > 0) = p_i(\mathbf{x}_i' \boldsymbol{\beta} + e_i > 0) \\ = \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma) \quad (4)$$

for  $i = 1, \dots, n$ .

### Model evaluation and selection

The estimated  $\boldsymbol{\beta}$  values maximized the natural log of the likelihood function  $L(\boldsymbol{\beta})$  considering each selection of explanatory variables, according to:

$$\log L(\mathbf{y} | \boldsymbol{\beta}) = \sum_{i=1}^n y_i \log(p_i) + (1-y_i) \log(1-p_i) \quad (5)$$

where  $y_i$  is the observed binary dependent variable. In this study, this represents microbiological contamination of

sample  $i$ . The Akaike information criterion (AIC) Akaike (1973) was used to compare the models:

$$\text{AIC} = -2\log L_{\max} + 2k \quad (6)$$

where  $L_{\max}$  is the maximized value of the likelihood function and  $k$  is the number of covariates in the model. This criterion reflects the information lost when a particular model is used to represent the observations. The AIC decreases with higher goodness-of-fit, and increases with the number of model parameters. All combinations of the variables with individual significance of  $p < 0.10$  were evaluated and the one that resulted in the lowest AIC was selected. This model was thus the optimal estimate in terms of both the selection of explanatory variables and the values of the corresponding coefficients  $\boldsymbol{\beta}$ . To maintain as much information as possible, the whole data set was used for model development, in agreement with Mair and El-Kadi (2013) and Howard et al. (2003). The rationale was that there were relatively few data points and the purpose of the study was to make structural interpretations of the results to infer the key mechanisms that affect contamination, rather than prediction. A cut-off value for the probability of predicted contamination  $p_i$  was specified at 0.5 by convention.

### Testing for spatial autocorrelation

The effect of spatial autocorrelation is particularly important in binary-outcome models such as the probit model. In the presence of spatial interdependence, the standard maximum likelihood estimator of these models is miss-specified if the interdependence is ignored, because the spatial structure affects the error terms (Fleming 2004). Moran's  $I$  (MI) test statistic (Moran 1950) is the most popular test for spatial autocorrelation (Kelejian and Prucha 2001). Kelejian and Prucha (2001) generalized Moran's  $I$  to limited dependent variable models, allowing for heteroscedasticity in the error term, which results in the following probit specification (Amaral et al. 2012):

$$\text{MI} = \frac{(\mathbf{e}' \mathbf{W} \mathbf{e})^2}{\text{trace}(\mathbf{W} \boldsymbol{\Sigma} \mathbf{W} \boldsymbol{\Sigma} + \mathbf{W}' \boldsymbol{\Sigma} \mathbf{W} \boldsymbol{\Sigma})} \quad (7)$$

resulting in the  $I$ -statistic:

$$I^2 = \frac{(\mathbf{e}' \mathbf{W} \mathbf{e})^2}{\text{trace}(\mathbf{W} \boldsymbol{\Sigma} \mathbf{W} \boldsymbol{\Sigma} + \mathbf{W}' \boldsymbol{\Sigma} \mathbf{W} \boldsymbol{\Sigma})} \quad (8)$$

where  $\mathbf{W}$  is the weight matrix, with entries  $w_{ij}$  that specify whether the locations  $i$  and  $j$  are neighbours and  $\boldsymbol{\Sigma}$  is a diagonal matrix that contain the variances of the individual residual terms, i.e., between the observed values,  $y_i$ , and the predicted values,  $\hat{\Phi}_i = \Phi(\mathbf{x}_i' \hat{\boldsymbol{\beta}})$ , according to  $\sigma_i^2 = \hat{\Phi}_i (1 - \hat{\Phi}_i)$ , with  $\hat{\boldsymbol{\beta}}$  as the Maximum-Likelihood

estimated parameters. The variance is not constant because  $\hat{\Phi}_i$  changes with  $\mathbf{x}_i$ . The weight matrix defines the spatial structure and should be specified based on theory (Mörtberg and Karlström 2005). In the current study, it reflected an estimate of the maximum lateral microbial travel distance in the aquifers in Juba. Two different weight matrices,  $\mathbf{W}$ , were considered: one that accounted for lateral distance only, and one that additionally considered the direction of groundwater flow. In the former case, a source was defined as a neighbour if it was located within a radius of 300 m of the reference source; in the latter, a source was defined as a neighbour if it was located within 300 m and up-gradient or level with the reference but not lower than 2 m below it. By convention, the weight matrix was normalized, summing each row to unity and setting the diagonal to zero. Moran's  $I$  ranges from  $-1$  to  $1$  and a high value indicates a high positive autocorrelation, whereas a value close to zero indicates spatial independence. For zero spatial autocorrelation, Moran's  $I$  is  $N(0,1)$  (Kelejian and Prucha 2001). Amaral et al. (2012) compared three test statistics proposed to reflect spatial error autocorrelation in probit models and found that Kelejian and Prucha's (2001) generalized Moran's  $I$  statistic performed the best.

### Spatial probit regression

To develop the spatial probit model, a spatial autocorrelation parameter,  $\rho$ , was included in addition to the explanatory variables selected in the conventional probit models. The spatial error model is based on a spatial autoregressive error term, according to:

$$e_i = \rho \sum_{j=1}^n w_{ij} e_j + \mu_i \quad (9)$$

for  $i = 1, \dots, n$ , where the error terms are normal and iid  $\mu_i \sim N(0, 1)$ , and  $\rho$  reflects the spatial autocorrelation:  $\rho = 0$  for independent error terms and a positive value indicates positive autocorrelation. In the spatial probit, the probabilities,  $p_i$ , are not independent and a multidimensional integral needs to be calculated, reflecting the number of observations (LeSage and Pace 2009; LeSage 2000). Thus:

$$\begin{aligned} p_i(Y_i = 1 \mid \mathbf{x}_i) &= p_i(Y_i^* > 0) = p_i(\mathbf{x}_i' \boldsymbol{\beta} + e_i > 0) \\ &= p_i(e_i > -\mathbf{x}_i' \boldsymbol{\beta}) = p_i(e_i / \sigma_i < \mathbf{x}_i' \boldsymbol{\beta} / \sigma_i) \\ &= \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma_i) \end{aligned} \quad (10)$$

for  $i = 1, \dots, n$ . The individual error terms  $\sigma_i$  are heteroscedastic and the vector  $\boldsymbol{\sigma}$  follows a multivariate normal distribution with zero mean and variance-covariance matrix  $[(\mathbf{I} - \rho \mathbf{W})'(\mathbf{I} - \rho \mathbf{W})^{-1}]$  (Amaral et al. 2012). The recursive

importance sampling algorithm was applied to calculate the  $n$ -dimensional integral in the likelihood function and thus estimate the parameters in the spatial probit model. This method uses random draws of truncated normal distributions (Beron and Vijverberg 2004). This simulator is one of the most efficient techniques for estimating the likelihood function (Pace and LeSage 2011). Other alternative methods include Gibbs sampling (LeSage 2000), the generalized method of moments (Pinkse and Slade 1998), and the expectation-maximization algorithm (McMillen 1992). To assess the relevance of a spatial probit model, confidence intervals (95%) and  $p$ -levels were evaluated for the spatial parameter,  $\rho$ .

## Results and discussion

### Statistical evaluation

#### Individual analyses

The exploratory analyses showed that contamination was correlated with several factors, as in prior hypotheses (Table 1). The most important factor was related to the near proximity of Tukul areas. A Tukul is a circular dwelling made of mud with a roof with thatching such as straw and leaves; Tukul areas are low-income zones that are informally occupied by people from rural regions, as described by USAID (2005). Three related variables were significant at the 10% level: the distance to the nearest Tukul area ( $p = 0.002$ ), the share of a 500 m radius buffer ( $p = 0.009$ ), and the share of a 100 m radius buffer ( $p = 0.055$ ). Note that the different  $z$ -value signs reflect the situation that a *larger* share of Tukul areas in the surrounding area was statistically associated with *more* contamination, whereas a *larger* distance to the nearest Tukul area was associated with *less* contamination. The results further indicated that the proximity of rivers or wadis might constitute a risk factor ( $p = 0.023$ ), in accordance with prior hypothesis. Studying a peri-urban area in Malawi, Palamuleni (2002) found that the surface water was highly polluted and suggested that this might be attributed to the disposal of raw sewage and run-off from townships, the washing of diapers in the rivers, and workers using the river as a disposal system. The proximity of land class 3–4 areas, which have limited or no sanitation, was also associated with contamination ( $p = 0.059$ ). In the current study, the sign of the correlation indicated that the proximity of open ground or grassland was correlated with a lower risk of contamination ( $p = 0.070$ ). This agreed with prior hypothesis because open ground indicates fewer residences, which might indicate less sources of human and animal waste. In all but one instance, the evaluated factors followed the predefined hypotheses in terms of the signs of the test statistics (Table 1); however, the results indicated that a *long* distance to nearest marshland implied a *higher* risk for contamination ( $p = 0.096$ ). One explanation could be that an inverse relationship



**Table 1** Results from the bivariate risk factor analyses: variables with  $p$ -values < 0.10

Variable	Test statistic	$p$ -value
Distance to the nearest Tukul area	−3.07 ( $z$ -value)	0.002
Tukul areas, share of a 500 m radius buffer	2.63 ( $z$ -value)	0.009
Distance to the nearest river or wadi	−2.27 ( $z$ -value)	0.023
Lowland topography (Boolean for lowland or flatland/highland)	4.44 ( $\chi^2$ )	0.035
Tukul areas, share of a 100-m-radius buffer	1.92 ( $z$ -value)	0.055
Class 3–4 residence area, share of a 100-m-radius buffer	1.89 ( $z$ -value)	0.059
Open ground or grassland, share of a 500 m radius buffer	−1.81 ( $z$ -value)	0.070
Distance to the nearest marshland	1.67 ( $z$ -value)	0.096

could be seen between the location of Tukul areas and the marshlands, which were mainly found along the Nile, whereas the former were located closer to the Mt. Jebel Kujur (Fig. 1); in Juba, people rarely settle in the marshlands, which are prone to flooding. Nevertheless, all of the factors with  $p$ -values < 0.1 were included in the multivariable regression, because the final results were derived based on AIC. This criterion penalizes additional parameters and gives low scores to models with factors that do not add significantly to the variance in the responses.

#### Multivariable regression

In the conventional regression analysis, the model with the lowest AIC (model 1A) included three explanatory variables: the distance to the nearest Tukul area, the local topography, and the share of class 3–4 residence areas within a 100 m radius. The proximity of rivers or wadis was not significant in the multivariable regression, which was likely related to covariation with other factors. Further, these results indicated that the distance to open ground, grasslands and marshlands was not relatively important. Considering the Tukul areas, they showed that the three variables identified in the individual analyses were correlated (Table 1). In model 1A, the proximity of class 3–4 residence areas was not significant at the 5% level; therefore, another model was developed (model 2A), which was the model with the lowest AIC that included only significant explanatory variables at the 5% level. It accounted for the distance to the nearest Tukul area and the local topography. Explanations of the different models can be found in Table 2.

The residuals of the conventional probit models were spatially autocorrelated. For model 1A, Moran's  $I$  was 1.90 ( $I$ -stat 3.61) if a source was defined as a neighbour located nearby, and Moran's  $I$  was 2.88 ( $I$ -stat 8.29) if a source was defined as neighbor only if it was found both nearby and upstream. Considering model 2A, the corresponding values were 2.08 ( $I$ -stat 4.31) for neighbours located nearby, and 3.05 ( $I$ -stat 9.28), for neighbours located nearby and upstream. These results indicated that spatial autocorrelation was stronger for the narrower definition of a neighbour, which excluded sources that were located downstream of a reference source. This was anticipated, considering the direction of

groundwater flow. These results showed that subject knowledge is important to appropriately define the weight matrix when applying a spatial model.

These findings thus indicated the presence of spatial autocorrelation in the residuals of a conventional approach, so spatial versions of models 1A and 2A were developed. Based on the  $MI$  values, two different definitions of contiguity were considered in these spatial models: in the first, a water source was specified as a neighbour if it was located nearby another source (models 1B and 2B); in the second, a source was considered as neighbour if it was located both nearby and upstream (models 1C and 2C in Table 2). Table 3 lists the parameter estimates in the different multivariable probit regression models, their standard deviations and  $t$ -statistics, as well as the log-likelihood of making the observations given the model parameters and the corresponding AIC. Considering the AIC, the latter models (models 1C and 2C) consistently performed better than former ones (models 1B and 2B). This was anticipated, considering the direction of groundwater flow and the Moran's  $I$  values; the subsequent analysis therefore concentrates on models 1C and 2C. In model 1C, the spatial interaction parameter,  $\rho$ , was estimated at 0.48 (standard deviation, SD, 0.16), which was significantly above zero ( $p$ -value 0.004). In the case of model 2C,  $\rho$  was estimated at 0.50 (SD 0.15 and  $p$ -value 0.001). Introducing a spatial parameter improved the models: the lowest AIC obtained using a traditional approach was 158.31 (model 1A), whereas the spatial model with the highest goodness-of-fit was model 2C, with an AIC of 153.20. Furthermore, there was an important difference between the inferences drawn from the traditional and spatial models. The latter emphasized the relative importance of the presence of Tukul areas, whereas it reduced the importance of the local topography; this factor was no longer significant at the 5% level ( $p$ -value 0.074 vs. 0.040 for the conventional approach). Figure 4 depicts the location of such Tukul areas as well as the investigated water sources and their TTC contamination levels.

#### Contamination mechanisms and hydrogeology

The best model, the one with the lowest AIC, thus incorporated two explanatory variables: the distance to the nearest Tukul

**Table 2** Description of the different multivariable probit regression models developed

Model	Explanation
1A	Conventional model with the lowest AIC, considering all combinations of explanatory variables with individual significance of $p < 0.10$ . The included factors were: lowland topography [Boolean]; share of class 3–4 residences [%]; distance to the nearest Tukul area [m]
1B	Spatial model: the included factors were the same as for the corresponding conventional model (1A): lowland topography [Boolean]; share of class 3–4 residences [%]; distance to the nearest Tukul area [m]; in addition to a parameter for spatial interactions [-]. In this model a water source was considered a neighbour to another source if it was located near to it ( $< 300$ m)
1C	Spatial model: the included factors were the same as for the corresponding conventional model (1A): lowland topography [Boolean]; share of class 3–4 residences [%]; distance to the nearest Tukul area [m]; in addition to the parameter for spatial interactions [-]. In this model a water source was considered a neighbour to another source if it was located near to it ( $< 300$ m) and upstream of it
2A	Conventional model with the lowest AIC, considering all combinations of explanatory variables with individual significance of $p < 0.05$ . The included factors were: lowland topography [Boolean]; distance to the nearest Tukul area [m]
2B	Spatial model: the included factors were the same as for the corresponding conventional model (2A)—lowland topography [Boolean]; distance to the nearest Tukul area [m]; in addition to the parameter for spatial interactions [-]. In this model, a water source was considered a neighbour to another source if it was located near to it ( $< 300$ m)
2C	Spatial model: the included factors were the same as for the corresponding conventional model (2A)—lowland topography [Boolean]; distance to the nearest Tukul area [m]; in addition to the parameter for spatial interactions [-]. In this model a water source was considered a neighbour to another source if it was located near to it ( $< 300$ m) and upstream of it

area ( $\beta_3$ ), and the local topography ( $\beta_1$ ) (model 2C). The siting of the Tukul areas, as specified by USAID (2005), was clearly approximate, seeing that the zones were circular (Fig. 4); nevertheless, considering the negative sign of  $\beta_3$ , these results reasonably indicated that if a water source was located at a far distance (measured in m) from all of the Tukul areas, then the susceptibility to contamination was substantially reduced. For water sources located in the Tukul areas the effect of the corresponding variable coherently disappears from the equation.

The statistical significance of a factor could either be linked to the presence of contaminant sources or to transport pathways; of these, it was likely that the effect of the near presence of Tukul areas was primarily related to the

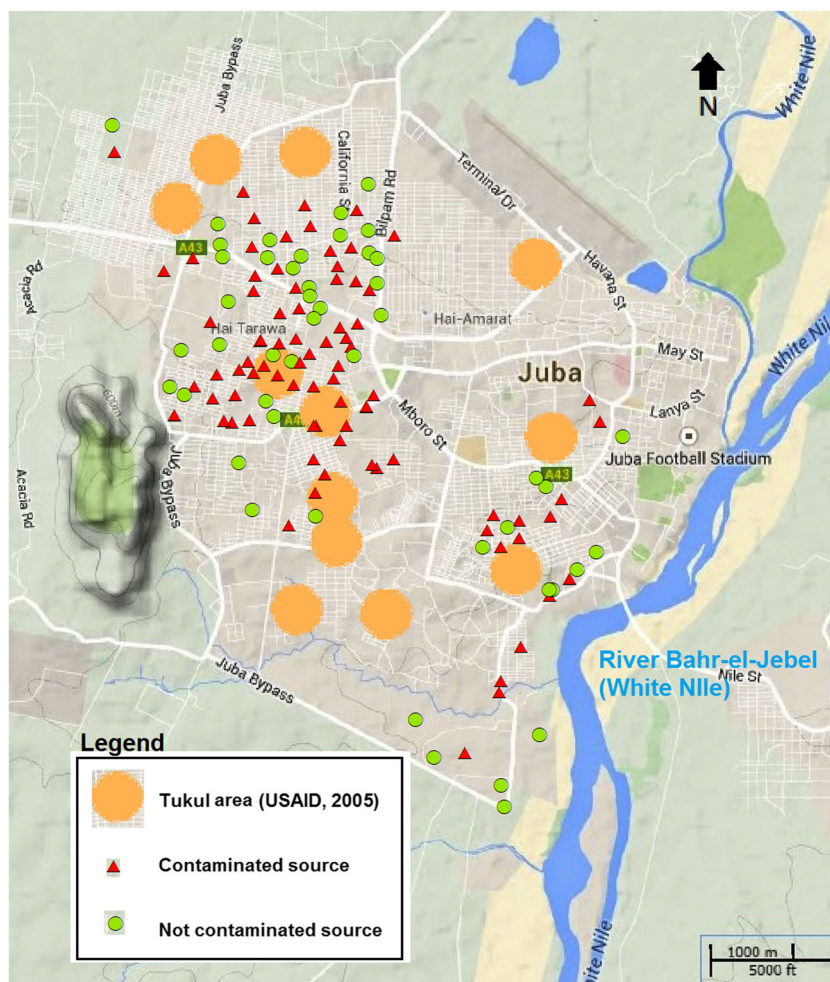
former. These areas typically have dense populations that reside in squatter housing and lack access to formal sanitation systems and the surrounding land is often used for rotational crops and subsistence farming (USAID 2005). In the proximity of Tukul areas, this suggests the high relative prevalence of animal and human waste, which provides sources of faecal coliforms. The Southern Sudan Commission for Census Statistics and Evaluation (2006) reported that 64% of the household population in the country used open-air spaces to dispose of human wastes. This part of the population is more likely to reside in informal Tukul areas than in the other zones where a larger share of the residents has access to sanitation.

**Table 3** The two best conventional models (models 1A and 2A) and the corresponding spatial models (models 1B, 1C, 2B, 2C) for explaining TTC contamination of water sources in Juba

Model type	Conventional	Spatial		Conventional	Spatial	
Variable	Model 1A	Model 1B	Model 1C	Model 2A	Model 2B	Model 2C
$\beta_0$ constant (SD)	0.62 (0.21)	0.69 (0.26)	0.75 (0.28)	0.70 (0.20)	0.78 (0.26)	0.85 (0.28)
<i>t</i> -stat/ <i>p</i> -value	2.98/0.003	2.62/0.010	2.68/0.008	3.47/0.001	2.97/0.004	3.09/0.002
$\beta_1$ lowland topography (SD)	0.63 (0.29)	0.57 (0.31)	0.63 (0.33)	0.60 (0.29)	0.53 (0.31)	0.59 (0.33)
<i>t</i> -stat/ <i>p</i> -value	2.16/0.033	1.84/0.068	1.91/0.058	2.08/0.040	1.71/0.090	1.80/0.074
$\beta_2$ class 3–4 residences within 100 m (SD)	0.94 (0.56)	0.94 (0.66)	0.87 (0.64)	—	—	—
<i>t</i> -stat/ <i>p</i> -value	1.68/0.096	1.44/0.152	1.36/0.176	—	—	—
$\beta_3$ distance to Tukul area (SD)	−0.0008 (0.0003)	−0.0009 (0.0003)	−0.0010 (0.0004)	−0.0009 (0.0003)	−0.0010 (0.0004)	−0.0011 (0.0004)
<i>t</i> -stat/ <i>p</i> -value	−3.08/0.003	−2.71/0.008	−2.77/0.006	−3.19/0.002	−2.77/0.006	−2.89/0.005
$\rho$ spatial interactions (SD)	—	0.29 (0.16)	0.48 (0.16)	—	0.33 (0.15)	0.50 (0.15)
<i>t</i> -stat/ <i>p</i> -value	—	1.87/0.064	2.94/0.004	—	2.14/0.034	3.25/0.001
Log-likelihood	−75.16	−73.55	−71.60	−76.72	−74.66	−72.60
AIC	158.31	157.10	153.21	159.45	157.32	153.20

*Italic font* indicates that the estimator is significant at the 5% level

**Fig. 4** Uncontaminated and contaminated water sources (Juba map from Google Maps 2014) and the locations of informal Tukul areas, from USAID (2005)



To identify a region of impact of each feature, different buffer zones were considered in the GIS analyses. In the case of Tukul areas, the most significant factor in the regression reflected the Euclidean distance (Table 3); additionally, the shares of Tukul areas within 500 m radii circular areas around each source were more significant than those within 100 m radii areas (Table 1). This suggested that the characteristics of an area further than 100 m from a borehole might influence its level of contamination. This result thus indicated generalized aquifer contamination, a contamination mechanism articulated by Hynds et al. (2014). Consistently, Batterman et al. (2009) found that the spreading of water-related infectious diseases is related to both ecologic and socio-economic processes, and that distal causes should be accounted for to enable sustainable interventions.

Seeing the positive sign of  $\beta_1$ , the results moreover indicated that lowland areas were more prone to contamination than highlands or flatland (Table 3). In the regression model, this factor was represented as a dummy variable, disappearing from the equation for water sources located in highlands or flatlands, and the coefficient would supplement the intercept for water sources located in lowland areas, such as valleys.

Assuming the presence of coliforms on the ground, this could be related to ponding in such areas, considering that Engström et al. (2015a) reported that the level of accumulated long-term precipitation was associated with contamination. The hydrogeology in Juba might allow for groundwater pollution. Basement complex aquifers generally imply large variations in groundwater velocities and vulnerability to contamination (Morris et al. 2003). Geological profiles from drilling protocols (MSF-B, unpublished data, 2013) specified that the top soil in Juba contained alluvial sediments with sand, loam, clay and weathered rock, which was underlain by rock of various degrees of weathering, and that the distance to the rock had large local variations. Lineaments in fractured rock do not provide substantial natural protective layers to reduce contamination (Kanyerere et al. 2012). Particularly, laterite zones near the surface can be quite transmissive and unconfined aquifers can enable contaminant transport from the ground towards the water table in a matter of days or weeks, with low attenuation potential and high to extreme pollution vulnerability (Morris et al. 2003).

The water samples considered in this study were collected by MSF-B to monitor the evolution of potential cholera

outbreaks and identify high-risk water sources. The sampling focused on areas previously affected by cholera, Kator and Munuki, where all of the water sources were tested. In developing countries, cholera is typically transmitted through water, and infected people could transmit the disease to other individuals via faecal contamination of water (Sack et al. 2004). It is thus reasonable to expect that boreholes contaminated with faecal indicators, such as TTCs, are more likely than clean ones to transmit cholera. *Vibrio cholerae* and TTCs have important similarities: they are gram-negative, facultatively anaerobic, and have similar size and shape (Cabral 2010), indicating that the strains would be transported in the same manner underground; however, this link requires further research. Nevertheless, the results could support future efforts that aim to reduce diarrheal disease. Cholera outbreaks have taken place in the South Sudan region every year from 2006 to 2009 and in 2014 (WHO 2014a). It remains a public health threat in Sub-Saharan Africa. According to Mengel et al. (2014), Sub-Saharan Africa accounted for 86% of reported cases of cholera and 99% of deaths due to cholera worldwide in 2011 (excluding the Haitian epidemic).

#### *Spatial regression*

This is the first study to use spatial regression models to assess risk factors for groundwater contamination, to the authors' knowledge. Hence, there was no previous literature to refer to when specifying the spatial model. The weight matrix should reflect the distance within which the response data are correlated. In theory, the groundwater in Juba might originate from the whole upstream Nile river basin, which would imply a vast zone of impact for each borehole and the possibility of spatial correlation among boreholes located very far from each other. However, the zone of impact would be limited by the fact that faecal coliforms typically die after 20 days in the field at 20–30 °C temperatures, based on Westcot (1997); nevertheless, it is not obvious how this would translate to distance, as discussed more thoroughly by Engström et al. (2015b). Notably, aquifers in weathered basement complexes often have anisotropic properties related to the orientation of the fractures, and pumping from boreholes could induce constricted and elongated zones of contribution (Tearfund 2007). In the current study, as an approximate approach, the presence of a water source within a fixed 300 m distance from a reference was defined as a neighbour and sources further away were not, which allowed for relatively lengthy transport. Shorter transport distances might also have been relevant. Hynds et al. (2012) estimated that the approximate zone of impact of septic tanks extended up to 110 m up-gradient of the wellhead, if high 120-h prior precipitation rates were considered. Conversely, in a review, Pang (2009) reported that the maximum observed *E. coli* transport distance was as great as 920 m, for sewage polluted groundwater in gravel aquifers in

Burnham, New Zealand, at velocities as high as 56–153 m/day (Sinton 1980). Future studies might investigate the cut-off distance for spatial autocorrelation as related to microbial transport in different hydrogeological environments.

The results in this study indicated that a spatial model might be more adequate than one that assumes all data are independent in space. The findings thus contribute to research on risk factors for urban (or peri-urban) groundwater contamination because sources that provide water in such areas are likely to be densely located. This is especially notable because groundwater provides an important component of the water supply system in 12 of the world's 23 megacities (>10 million inhabitants) (Hirata et al. 2006); in particular, groundwater is an essential water source in peripheral, poorer parts of many cities, which often do not receive piped water or formal sanitation services (Hirata et al. 2006).

#### **Limitations**

The regression resulted in 67% correct predictions using the model with the lowest AIC (model 2C). This was relatively low, indicating that the investigated features did not account for the whole variance in the response variable, which might be an effect of the low resolution of the maps. Other factors than those considered in the current study may have also influenced the water quality.

#### *Data resolution*

It is reasonable to expect that microbial contamination of groundwater sources is particularly prevalent in urban areas in developing countries; however, such environments are often relatively disorganized, imposing constraints on access to detailed spatial and temporal data. Batterman et al. (2009) stated that understanding of water-related infectious diseases in developing countries is often limited by knowledge and data gaps and that related analyses are often based on multiple and sparse data sets. The current study also faced some related restrictions. The Comprehensive Peace Agreement was signed between fighting parties in Sudan in 2005, ending decades of civil war. Few records of geological and hydrogeological surveys in Juba were centralized before 2005. The decades of conflict resulted in many internal refugees and very limited resources for systematic monitoring of environmental and socio-economic factors. Therefore, the analysis relied on reports by USAID (2005) and JICA (2009a; b) for spatial information. The resolution in these data varied. Furthermore, the report by USAID (2005) was developed 5 years prior to the sampling in the current study and spatial features could have changed during this time, which means that the exact location of features could not be determined. Instead, inferences need to be based on broad trends in the data.



### *Missing spatial risk factors*

In the regression, it would be preferable to account for the hydrogeological settings in the vicinity of each water source, such as the bedrock and the subsoil characteristics. Fine-resolution spatial data on the location of fracture zones or lithology could not be found, as the accessible hydrological and geological maps were on a country scale. Groundwater levels in Juba had to be estimated based on interpolation of the registered static water level from a limited number of sources. The static water level reported in these protocols varied from 2 m to more than 20 m below ground, with large variations. It was anticipated that the local water table level would be associated with contamination. For example, Kulabako et al. (2007) reported that the level of faecal contaminants increased in areas in Kampala with a higher water table. However, in the current study, the static water level was not significant; nevertheless, the results do not exclude the possibility that local and/or ephemeral aquifers influenced contamination, considering that local variations might not be correctly estimated based on the 33 locations used for estimation of the static water level elsewhere in Juba. Further, the static water level was measured at times other than the microbiological sampling dates and there could be seasonal variations. Future studies would thus preferably account for ephemeral and local aquifers.

Additionally, the results indicated that the proximity of houses or roads was not associated with borehole contamination; however, the map representing their locations did not thoroughly reflect the informal infrastructure in Juba, such as walkways and individual clay huts, which might be important. If possible, future studies should account for such data. Other potential risk factors include the number of users of each source and the locations of small-scale animal farming facilities or cultivated areas where manure might be used for fertilizer. Further, the distance to small ponds near each water source would preferably be included. Studying ponds in rural Bangladesh, Knappett et al. (2011) reported that the water in the majority of the ponds contained unsafe levels of faecal contamination, which was mainly attributed to the proximity of unsanitary latrines (visible effluent or open pits).

### *Time-variant risk factors*

The current study focused on spatial factors, although temporal factors are also likely to be important. Results from Engström et al. (2015a) indicated that both the level of on-site hygiene and contamination of groundwater sources varied considerably with time. The latter was transient in 43% of the investigated sources, and the level of on-site hygiene was a significant factor for contamination in pairwise comparisons of the sources with varying contamination at different times (Engström et al. 2015a). Water sampling was consistently conducted in the wet season in the current study; nevertheless,

there are weather variations in this period that might have impact on the susceptibility of wells to contamination. Engström et al. (2015a) found that accumulated long-term antecedent rainfall was associated with contamination events but temperature was not. It is therefore recommended that future studies in similar areas account for time-variant factors that might influence groundwater quality, particularly precipitation, in addition to spatial factors.

### **Summary and conclusions**

This study investigated potential risk factors influencing bacterial contamination of urban groundwater sources. The evaluated variables reflected site-specific information as well as regional land use, hydrogeological setting and socio-economic characteristic data in Juba, South Sudan. A conventional multivariable regression model was developed. This approach resulted in residuals that had significant, positive spatial autocorrelation. Therefore, a spatial model was estimated in which the parameter that reflected spatial interactions was significant ( $p$ -value 0.001) and estimated at 0.50 (SD 0.15). This model accounted for the proximity of areas with informal settlements, Tukul areas, as well as the local topography (lowland/no lowland indicator variable). The results indicated that the groundwater below these zones was contaminated. Tukul areas lack formal sanitation systems, rearing animals is common and the surrounding land is often used for subsistence farming, which might explain the increased risk for contamination in their vicinity. The results suggested that generalized aquifer contamination occurred. It is recommended that future remediation efforts and monitoring schemes in cities similar to Juba—in terms of climate, hydrogeology and socio-economic characteristics—focus on such low income and informal settlement areas.

This study contributed to methodological development in the subject area. The results showed that statistical studies of groundwater quality should consider the effects of spatial interactions when the investigated sources are located near to each other. Introducing a spatial term could have important effects on the other parameters in the model. In the current study, the spatial model indicated that the local topography was not significant at the 5% level, in contrast with inferences based on the conventional model. However, when applying spatial regression, it should be emphasized that subject knowledge is important to define the weight matrix that reflects spatial interactions. In this study, the spatial parameter was more significant when the direction of groundwater flow was considered in defining the weight matrix. In the field of groundwater quality, research based on statistical models can inform decision making by identifying priority land-use types and prioritizing remediation efforts. In cities, groundwater quality data are unlikely to be independent in space because

the water sources are often located near to each other. Future research should address the mechanisms for urban groundwater contamination; when using statistical models to do so, spatial effects should be accounted for. This is important considering that groundwater provides a large component of the water supply system in a majority of the world's megacities.

**Acknowledgements** The ASTER Global Digital Elevation Model (GDEM V2) data was retrieved from the online Data Pool, courtesy of the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, [https://lpdaac.usgs.gov/data\\_access/data\\_pool](https://lpdaac.usgs.gov/data_access/data_pool). The authors would like to acknowledge Dr. Berit Balfors and Dr. Roger Thunvik for their comments and suggestions.

## Appendix

**Table 4** Risk factors considered in the bivariate analyses

Variable	Description [unit]	Data source
<b>Land use</b>		
Bush	Euclidean distance to the closest feature from each source [m]; share of the area within 30, 100 and 500 m radii buffers from each source [%]	JICA (2009a; b)
Open ground or grassland	Euclidean distance to the closest feature from each source [m]; share of the area within 30, 100 and 500 m radii buffers from each source [%]	JICA (2009a; b)
Roads or houses	Euclidean distance to the closest feature from each source [m]; share of the area within 30, 100 and 500 m radii buffers from each source [%]	JICA (2009a; b)
Commercial areas	Euclidean distance to the closest feature from each source [m]; share of the area within a 500 m radius buffer from each source [%]	JICA (2009a; b) and USAID (2005)
<b>Land class characteristics</b>		
Class 1	Euclidean distance to the closest feature from each source [m]	USAID (2005)
Class 2	Euclidean distance to the closest feature from each source [m]	USAID (2005)
Class 3–4	Euclidean distance to the closest feature from each source [m]; share of the area within 100 and 500 m radii buffers from each source [%]	USAID (2005)
Informal Tukul areas	Euclidean distance to the closest feature from each source [m]; share of the area within 100 and 500 m radii buffers from each source [%]	USAID (2005)
<b>Hydrogeological settings</b>		
Marshland, including the Nile Greenway corridor	Euclidean distance to the closest feature from each source [m]; share of the area within a 500 m radius buffer from each source [%]	JICA (2009a; b) and USAID (2005)
Rivers: the White Nile and its tributaries	Euclidean distance to the closest feature from each source [m]; share of the area within 30, 100 and 500 m radii buffers from each source [%]	JICA (2009a; b)
Static water level	Estimated level in metres below ground at the water source [m]	MSF-B (unpublished data 2013)
Elevation	Metres above sea level at the source, 30 × 30 m spatial resolution [m]	ASTER Global Digital Elevation Model—GDEM V2; NASA Jet Propulsion Laboratory (JPL) 2011
<b>Source-site specific features</b>		
On-site hygiene	Acceptable/acceptable with cleaning of site/major cleaning of site needed (ordinal)	MSF-B on-site investigations
Local topography	1 if lowland/0 if flatland or highland (Boolean)	MSF-B on-site investigations

Buffer zones of 30, 100 and 500 m were considered for most factors; however, in some cases these zones did not contain any features and were omitted from the analysis. The numeric data were obtained using image processing operations

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Abramson A, Benami M, Weisbrod N (2013) Adapting enzyme-based microbial water quality analysis to remote areas in low-income countries. *Environ Sci Technol* 47:10494–10501. doi:[10.1021/es402175n](https://doi.org/10.1021/es402175n)
- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Paper presented at the 2nd International Symposium on Information Theory, Akademiai Kiado, Budapest, pp 267–281
- Amaral PV, Anselin L, Arribas-Bel D (2012) Testing for spatial error dependence in probit models. *Lett Spat Resour Sci* 6:91–101. doi:[10.1007/s12076-012-0089-9](https://doi.org/10.1007/s12076-012-0089-9)
- Batterman S, Eisenberg J, Hardin R, Kruk M, Lemos M, Michalak A, Mukherjee B, Renne E, Stein H, Watkins C, Wilson M (2009) Sustainable control of water-related infectious diseases: a review and proposal for interdisciplinary health-based systems research. *Environ Health Perspect* 117:1023–1032
- Beller M, Ellis A, Lee SH, Drebot MA, Jenkerson SA, Funk E, Sobsey MD, Simmons OD, Monroe SS, Ando T, Noel J, Petric M, Middaugh JP, Spika JS (1997) Outbreak of viral gastroenteritis due to a contaminated well: international consequences. *J Am Med Assoc* 278:563–568
- Beron KJ, Vijverberg WPM (2004) Probit in a spatial context: a Monte Carlo analysis. In: Anselin L, Florax RJGM, Rey SJ (eds) *Advances in spatial econometrics: methodology, tools and applications*. Springer, Heidelberg, Germany, pp 169–195
- Borchardt MA, Chyou PH, DeVries EO, Belongia EA (2003) Septic system density and infectious diarrhea in a defined population of children. *Environ Health Perspect* 111:742–748
- Cabral JPS (2010) Water microbiology: bacterial pathogens and water. *Int J Environ Res Public Health* 7:3657–3703. doi:[10.3390/ijerph7103657](https://doi.org/10.3390/ijerph7103657)
- de Brauwere A, Ouattara NK, Servais P (2014) Modeling fecal indicator bacteria concentrations in natural surface waters: a review. *Crit Rev Environ Sci Technol* 44:2380–2453
- Engström E, Balfors B, Mörtberg U, Thunvik R, Gaily T, Mangold M (2015a) Prevalence of microbiological contaminants in groundwater sources and risk factor assessment in Juba, South Sudan. *Sci Total Environ* 515–516:181–187. doi:[10.1016/j.scitotenv.2015.02.023](https://doi.org/10.1016/j.scitotenv.2015.02.023)
- Engström E, Thunvik R, Kulabako R, Balfors B (2015b) Water transport, retention and survival of *Escherichia coli* in unsaturated porous media: a comprehensive review of processes, models and factors. *Crit Rev Environ Sci Technol* 45:1–100. doi:[10.1080/10643389.2013.828363](https://doi.org/10.1080/10643389.2013.828363)
- Fleming MM (2004) Techniques for estimating spatially dependent discrete choice models. In: Anselin L, Florax RJGM, Rey SJ (eds) *Advances in spatial econometrics: methodology, tools and applications*. Springer, Heidelberg, Germany, pp 145–168
- Glanville TD, Baker JL, Newman JK (1997) Statistical analysis of rural well contamination and effects of well construction. *Trans ASAE* 40(2):363–370
- Godfrey S, Timo F, Smith M (2006) Microbiological risk assessment and management of shallow groundwater sources in Lichinga, Mozambique. *Water Environ J* 20:194–202. doi:[10.1111/j.1747-6593.2006.00040.x](https://doi.org/10.1111/j.1747-6593.2006.00040.x)
- Google Maps (2014) <https://www.google.se/maps/@4.8417987,31.5885441,13z/data=!5m1!1e4?hl=en>. Accessed October 23, 2014
- Google Earth (2016) 7.1.7. 04°48.61596'N, 031°35.08757'E. <http://www.google.com/earth/index.htm>. Accessed September 09, 2016
- Guillemin F, Henry P, Uwechue N, Monjour L (1991) Faecal contamination of rural water supply in the Sahelian area. *Water Res* 25:923–927
- Hirata R, Stimson J, Varnier C (2006) Urban hydrogeology in developing countries: a foreseeable crisis. Paper presented at the International Symposium on Groundwater Sustainability (ISGWAS) Alicante, Spain, January 2006
- Howard G, Pedley S, Barrett M, Nalubega M, Johal K (2003) Risk factors contributing to microbiological contamination of shallow groundwater in Kampala, Uganda. *Water Res* 37:3421–3429
- Hynds PD, Misstear BD, Gill L (2012) Development of a microbial contamination susceptibility model for private domestic groundwater sources. *Water Resour Res* 48. doi:[10.1029/2012/WR012492](https://doi.org/10.1029/2012/WR012492)
- Hynds P, Misstear BD, Gill LW, Murphy HM (2014) Groundwater source contamination mechanisms: physicochemical profile clustering, risk factor analysis and multivariate modelling. *J Contam Hydrol* 159:47–56. doi:[10.1016/j.jconhyd.2014.02.001](https://doi.org/10.1016/j.jconhyd.2014.02.001)
- Japan International Cooperation Agency (JICA) (2009a) Juba urban water supply and capacity development study in the Southern Sudan: final report. <http://libopac.jica.go.jp/top/index.do?method=change&langMode=ENG>. Accessed 1 April 2014
- Japan International Cooperation Agency (JICA) (2009b) Juba water supply and capacity development study in the Southern Sudan, Interim report 1, presentation, JICA, Tokyo
- Kanyerere T, Levy J, Xu Y, Saka J (2012) Assessment of microbial contamination of groundwater in upper Limphasa River catchment, located in a rural area of northern Malawi. *Water SA* 38:581–596
- Kelejian HH, Prucha IR (2001) On the asymptotic distribution of the Moran I test statistic with applications. *J Econ* 104:219–257. doi:[10.1016/S0304-4076\(01\)00064-1](https://doi.org/10.1016/S0304-4076(01)00064-1)
- Knappett PSK, Escamilla V, Layton A, McKay LD, Emch M, Williams DE, Huq R, Alam J, Farhana L, Mailloux BJ, Ferguson A, Sayler GS, Ahmed KM, van Geen A (2011) Impact of population and latrines on fecal contamination of ponds in rural Bangladesh. *Sci Total Environ* 409:3174–3182
- Kulabako NR, Nalubega M, Thunvik R (2007) Study of the impact of land use and hydrogeological settings on the shallow groundwater quality in a peri-urban area of Kampala, Uganda. *Sci Total Environ* 381:180–199
- LeSage JP (2000) Bayesian estimation of limited dependent variable spatial autoregressive models. *Geogr Anal* 32:19–35. doi:[10.1111/j.1538-4632.2000.tb00413.x](https://doi.org/10.1111/j.1538-4632.2000.tb00413.x)
- LeSage J, Pace RK (2009) *Introduction to spatial econometrics*. CRC, Boca Raton, FL
- Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, Rudan I, Campbell H, Cibulskis R, Li M, Mathers C, Black RE (2012) Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* 379:2151–2161. doi:[10.1016/S0140-6736\(12\)60560-1](https://doi.org/10.1016/S0140-6736(12)60560-1)
- Mair A, El-Kadi AI (2013) Logistic regression modeling to assess groundwater vulnerability to contamination in Hawaii, USA. *J Contam Hydrol* 153:1–23. doi:[10.1016/j.jconhyd.2013.07.004](https://doi.org/10.1016/j.jconhyd.2013.07.004)
- McMichael G (2016) Land conflict and informal settlements in Juba, South Sudan. *Urban Studies* 53(13):2721–2737
- McMillen DP (1992) Probit with spatial autocorrelation. *J Reg Sci* 32:335–348. doi:[10.1111/j.1467-9787.1992.tb00190.x](https://doi.org/10.1111/j.1467-9787.1992.tb00190.x)
- Mengel MA, Delrieu I, Heyerdahl L, Gessner BD (2014) Cholera outbreaks in Africa. *Curr Top Microbiol Immunol* 379:117–144. doi:[10.1007/82\\_2014\\_369](https://doi.org/10.1007/82_2014_369)
- Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika* 37:17–23. doi:[10.2307/2332142](https://doi.org/10.2307/2332142)
- Morris BL, Lawrence AR, Chilton PJ, Adams B, Calow RC, Klinck BA (2003) Groundwater and its susceptibility to degradation, a global

- assessment of the problem and options for management. Early Warning and Assessment Report Series, RS. 03-3. United Nations Environment Programme (UNEP), Nairobi, Kenya
- Mörtberg U, Karlström A (2005) Predicting forest grouse distribution taking account of spatial autocorrelation. *J Nat Conserv* 13:147–159. doi:10.1016/j.jnc.2005.02.008
- NASA Jet Propulsion Laboratory (JPL) (2011) Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model Version 2 (GDEM V2). NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, SD. <http://earthexplorer.usgs.gov/>. Accessed September 27, 2013
- Nyenje PM, Foppen JW, Kulabako R, Muwanga A, Uhlenbrook S (2013) Nutrient pollution in shallow aquifers underlying pit latrines and domestic solid waste dumps in urban slums. *Environ Manag* 122: 15–24
- Oguntoké O, Komolafe OA, Annegarn HJ (2013) Statistical analysis of shallow well characteristics as indicators of water quality in parts of Ibadan City, Nigeria. *J Water Sanit Hygiene Dev* 3:602–611. doi:10.2166/washdev.2013.066
- Oxfam-Delagua (2009) Oxfam-Delagua portable water testing kit user manual (version 4.2). [http://www.oxfam.org.uk/equipment/catalogue/resources-included-available/water-and-sanitation/water-treatment-and-testing/Delagua%20english\\_manual\\_2000-1.pdf/at\\_download/file](http://www.oxfam.org.uk/equipment/catalogue/resources-included-available/water-and-sanitation/water-treatment-and-testing/Delagua%20english_manual_2000-1.pdf/at_download/file). Accessed 28 November 2014
- Pace RK, LeSage JP (2011) Fast simulated maximum likelihood estimation of the spatial probit model capable of handling large samples. doi:10.2139/ssrn.1966039
- Palamuleni LG (2002) Effect of sanitation facilities, domestic solid waste disposal and hygiene practices on water quality in Malawi's urban poor areas: a case study of South Lunzu Township in the city of Blantyre. *Phys Chem Earth, parts A/B/C* 27:845–850. doi:10.1016/S1474-7065(02)00079-7
- Panagos P, Jones A, Bosco C, Senthil Kumar PS (2011) European digital archive on soil maps (EuDASM): preserving important soil data for public free access. *Int J Digit Earth* 4:434–443
- Pang L (2009) Microbial removal rates in subsurface media estimated from published studies of field experiments and large intact soil cores. *J Environ Qual* 38:1531–1559. doi:10.2134/jeq2008.0379
- Pedley S, Howard G (1997) The public health implications of microbiological contamination of groundwater. *Q J Eng Geol Hydrogeol* 30: 179–188. doi:10.1144/gsl.jgegh.1997.030.p2.10
- Pinkse J, Slade ME (1998) Contracting in space: an application of spatial statistics to discrete-choice models. *J Econ* 85:125–154. doi:10.1016/S0304-4076(97)00097-3
- Sack DA, Sack RB, Nair GB, Siddique AK (2004) Cholera. *Lancet* 363: 223–233. doi:10.1016/S0140-6736(03)15328-7
- Sinton LW (1980) Two antibiotic-resistant strains of *Escherichia coli* for tracing the movement of sewage in groundwater. *J Hydrol N Z* 19: 119–130
- Sorensen JPR, Lapworth DJ, Read DS, Nkhuwa DCW, Bell RA, Chibesa M, Chirwa M, Kabika J, Liemisa M, Pedley S (2015) Tracing enteric pathogen contamination in sub-Saharan African groundwater. *Sci Total Environ* 538:888–895. doi:10.1016/j.scitotenv.2015.08.119
- Southern Sudan Commission for Census Statistics and Evaluation (2006) Southern Sudan Household Health Survey. <http://www.bsf-south-sudan.org/sites/default/files/SHHS.pdf>. Accessed 10 October 2014
- Sudan Ministry of Energy and Mines Geological and Mineral Resources Department (1981) Geological map of the Sudan. [http://europa.eu/esdb\\_archive/eudasm/africa/images/maps/download/afr\\_sd2001\\_ge.jpg](http://europa.eu/esdb_archive/eudasm/africa/images/maps/download/afr_sd2001_ge.jpg). Accessed 4 November 2014
- Tearfund (2007) Darfur: water supply in a vulnerable environment—phase two of Tearfund's Darfur environment study. Summary report, USAID, Washington, DC; DFID, London; UNEP, Nairobi
- Tollestrup K, Frost FJ, Kunde TR, Yates MV, Jackson S (2014) Cryptosporidium infection, onsite wastewater systems and private wells in the arid Southwest. *J Water Health* 12:161–172. doi:10.2166/wh.2013.049
- USAID (United States Agency for International Development) (2005) Juba Assessment Town Planning and Administration Report September–October 2005 CA no. 623-A-00-05-00318, USAID, Washington, DC
- Vail JR (1989) Hydrological map of Sudan. South Sheet, series 2201. Ministry of Energy and Mines, Geological and Mineral Resources Department, Khartoum, Sudan
- Westcot DW (1997) Quality control of wastewater for irrigated crop production. FAO water report 10, Food and Agriculture Organization of the United Nations, Rome
- WHO (2011) Guidelines for drinking-water quality, 4th edn. WHO, Geneva
- WHO (2014a) Early warning and disease surveillance bulletin (IDP camps and communities) 11–17 August 2014. [http://www.who.int/hac/crises/ssd/south\\_sudan\\_ewarn\\_17august2014.pdf?ua=1](http://www.who.int/hac/crises/ssd/south_sudan_ewarn_17august2014.pdf?ua=1). Accessed September 2014
- WHO (2014b) South Sudan Country Profile. <http://www.who.int/countries/ssd/en/>. Accessed 30 September 2014
- WHO/UNICEF (2012) Progress on drinking water and sanitation Joint Monitoring Programme update 2012. UNICEF, New York and WHO, Geneva