Hydrogeology Journal
https://doi.org/10.1007/s10040-018-1900-5

PAPER

Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale

Issoufou Ouedraogo [1,2] · Pierre Defourny [1] · Marnik Vanclooster [1]

Received: 11 March 2018 / Accepted: 10 November 2018

Abstract
Groundwater management decisions require robust methods that allow accurate predictive modeling of pollutant occurrences. In this study, random forest regression (RFR) was used for modeling groundwater nitrate contamination at the African continent scale. When compared to more conventional techniques, key advantages of RFR include its nonparametric nature, its high predictive accuracy, and its capability to determine variable importance. The latter can be used to better understand the individual role and the combined effect of explanatory variables in a predictive model. In the absence of a systematic groundwater monitoring program at the African continent scale, the study used the groundwater nitrate contamination database for the continent obtained from a meta-analysis to test the modeling approach; 250 groundwater nitrate pollution studies from the African continent were compiled using the literature data. A geographic information system database of 13 spatial attributes was collected, related to land use, soil type, hydrogeology, topography, climatology, type of region, and nitrogen fertilizer application rate, and these were assigned as predictors. The RFR performance was evaluated in comparison to the multiple linear regression (MLR) methods. By using RFR, it was possible to establish which explanatory variables influence the occurrence of nitrate pollution in groundwater (population density, rainfall, recharge, etc.). Both the RFR and MLR techniques identified population density as the most important variable explaining reported nitrate contamination. However, RFR has a much higher predictive power ($R^2 = 0.97$) than a traditional linear regression model ($R^2 = 0.64$). RFR is therefore considered a very promising technique for large-scale modeling of groundwater nitrate pollution.

Keywords Groundwater modeling · Nitrate · Random forest · Geographic information system · Sub-Saharan Africa

Introduction

Groundwater constitutes one of the most valuable global natural resources, serving as a major source of water for communities, agriculture, and industrial purposes. Due to the importance of groundwater, many studies with a focus on groundwater issues can be found in the literature (Teng et al. 2018; Foster et al. 2018; Hao et al. 2018). Concerning Africa, groundwater is also a crucial natural resource supporting economic development, but it is subject to many pressures. Xu and Usher (2006) noted that degradation of groundwater represented the most serious water resource problem in Africa. A recent study by Lapworth et al. (2017) focused on urban groundwater quality in Sub-Saharan Africa. MacDonald et al. (2013) affirmed that the two main threats were overexploitation and contamination. Nitrate is a common chemical contaminant of groundwater, and the level of contamination is also increasing in many African aquifers (Spalding and Exner 1993; Puckett et al. 2011). Nitrate ingestion has been linked to methemoglobinemia, adverse reproductive outcomes, and specific cancers (Ward et al. 2005). In addition, nitrate is often a proxy for other possible pollutants of groundwater. Nitrate contamination is therefore very informative for overall groundwater quality. However, it is a space–time variable, and the level of contamination depends on many other space–time environmental and anthropogenic attributes. The

✉ Issoufou Ouedraogo
ouedraogo.issoufou03@gmail.com; ouediss6@yahoo.fr

[1] Earth and Life Institute, Université catholique de Louvain, Croix du Sud 2, Box 2, B-1348 Louvain-la-Neuve, Belgium

[2] Ecole Nationale Supérieur d'Ingénieurs de Fada (ENSI-F), Université de Fada N'Gourma, BP 46, Fada N'Gourma, Burkina Faso

Published online: 06 December 2018

regional modeling of nitrate contamination of groundwater therefore remains a technical and scientific challenge.

Statistical models are often deployed to explain the spatial distribution of observed nitrate concentration in terms of available environmental and anthropogenic attributes, or to discriminate sources of contamination (Nolan and Hitt 2006). Most statistical models used in such context use multiple linear regression (MLR), nonlinear regression, logistic regression, Bayesian approaches, artificial neural networks (ANN), or classification and regression trees in order to extract the variables that explain nitrate contamination (Rawlings et al. 1998; Burow et al. 2010; Mair and El-Kadi 2013; Gurdak and Qi 2012; Mattern et al. 2012). For example, Bauder et al. (1993) investigated the major controlling factors for nitrate contamination of groundwater in agricultural areas using land use, climate, soil characteristics, and cultivation types as explanatory variables. However, these different techniques exhibit a variety of problems, such as a lack of sensitivity toward outlier values of logistic regression, and the opacity of neural networks (Abrahart et al. 2008).

The use of modern data mining or machine learning approaches avoids many of these limitations. An emerging technique that utilizes ensembles of regressions is receiving particular interest in other fields of knowledge (Hansen and Salamon 1990; Steele 2000; Khalil et al. 2005; Sesnie et al. 2008). Ensemble learning algorithms use the same base algorithm to produce repeated multiple predictions, which are averaged in order to produce a unique model (Breiman 2001a; Friedl et al. 1999). They may reduce bias and improve prediction efficiency (Breiman 2001a; Cutler et al. 2007; Peters et al. 2007; Fernández-Delgado et al. 2014).

Random forest regression (RFR) is an example of such an ensemble regression method. RFR is nonparametric, and thus data do not need to come from a specific distribution (e.g. Gaussian) and can contain collinear variables (Cutler et al. 2007). Furthermore, RFR works well with very large numbers of predictors (Cutler et al. 2007). These methods can deal with model selection uncertainty, as predictions are based upon a consensus of many models and not simply a single model selected with some measure of goodness of fit. RFR is appropriate for illustrating the nonlinear effect of variables; it can handle complex interactions among variables and is not affected by multicollinearity (Breiman 2001b). Major applications of RFR are found in environmental and ecological modeling (e.g. Yost et al. 2008; Moreno et al. 2011; Oliveira et al. 2012; Oppel et al. 2012; Hoyos et al. 2015), eco-hydrological distribution modeling (e.g. Peters et al. 2007), landslide susceptibility mapping (Park 2014; Youssef et al. 2015), and remote sensing (e.g. Gislason et al. 2006; Pal 2005; Rodriguez-Galiano et al. 2012a, 2012b). In groundwater research, random forest (RF) methods have been applied to model nitrate and arsenic in aquifers of the southwestern United States (Anning et al. 2012), nitrate in an unconsolidated aquifer in southern Spain (Rodriguez-Galiano et al. 2014), nitrate in private wells in Iowa, USA (Wheeler et al. 2015), and nitrate in shallow and deep wells of the US Central Valley (Nolan et al. 2014). Mendes et al. (2016) applied random forest methodology to assess the vulnerability of groundwater to nitrate pollution for the Vega de Granada aquifer (Spain), while Norouz et al. (2016) used the random forest method to determine the Malekan Aquifer (Iran) vulnerability with several variables, including variables related to the DRASTIC method. More recently, Sahoo et al. (2017), Ransom et al. (2017), Sajedi-Hosseini et al. (2018), and Barzegar et al. (2018) used the machine learning algorithm in groundwater modeling. Naghibi et al. (2017) and Golkarian et al. (2018) used the random forest algorithm to study groundwater. A perceived disadvantage of other machine learning methods such as ANN is their "black-box" nature: without estimated coefficients, it is difficult to show significant relations between the response and predictor variables (Nolan et al. 2015). According to Rodriguez-Galiano et al. (2014), RFR is relatively robust to outliers and it can overcome the black-box limitations of ANN by assessing the relative importance of the explanatory variables and selecting the most important ones (features), hence reducing the dimensionality.

RFR is based on bootstrap aggregation of regression trees, and typically outperforms traditional models such as logistic regression (Breiman 2001a; Breiman et al. 1984). Moreover, the parameterization of RFR is very simple and it is computationally less demanding (Rodriguez-Galiano and Chica-Rivas 2012). RFR provides very good results compared to other machine learning techniques such as support vector machines (SVM) or ANN, or to other decision tree algorithms (Breiman 2001a; Liaw and Wiener 2002). Loosvelt et al. (2012) demonstrated that uncertainty estimates could be easily assessed when RFR was applied.

This study uses RFR to explain and predict groundwater nitrate contamination at the continental scale and compares the performance of RFR with MLR techniques. The study uses a nitrate contamination data set compiled through a meta-analysis (Ouedraogo and Vanclooster 2016a). The modeling of nitrate contamination at this scale can provide guidance for the planning and implementation of groundwater monitoring programs, in particular the design of transboundary groundwater management strategies adjusted to the conditions in different regions of Africa.

Materials and methods

Study area

Groundwater is an important water resource for the African continent. It is the source of drinking water for 75% of the continent's population (UNECA et al. 2000). This proportion is higher in some arid and semi-arid countries. For example, in the case of Libya, it is as high as 95% (Margat 2010). Groundwater occurrence depends primarily on geology, geomorphology/weathering, and rainfall (both current and historic). The interplay of these factors gives rise to complex hydrogeological environments with countless variations in the quantity, quality, ease of access, and renewability of groundwater resources. Across the African continent (including island countries), there are four major aquifer types: Precambrian basement rocks (covering approximately 34% of the land surface), consolidated sedimentary rocks (37%), unconsolidated sediments (25%), and volcanic rocks (4%) (MacDonald et al. 2009).

The groundwater nitrate contamination data set for the African continent

In the major part of Africa, there is very little or no systematic monitoring of groundwater. In the absence of monitoring data, the study used compiled nitrate pollution data at the African continental scale from a meta-analysis (Ouedraogo and Vanclooster 2016a). A total of 250 groundwater pollution studies in the literature from available books and the internet Web of Science (Scopus™, ScienceDirect®, Google™, and Google Scholar™) were collected. The literature data were filtered using the following criteria: (i) the publication should explicitly report on nitrate concentrations in groundwater; (ii) the publication should be published after 1999. Various aggregated measurements of nitrate concentration were derived from the different studies. Thus, it was possible to retain 206 points where the maximum concentration of nitrate was reported, 187 points where the minimum concentration was reported, and 94 studies on the mean concentration of nitrate. Of the 94 data sets for which mean values were reported, 12 field sites had nitrate concentration smaller than 1 mg/L. Therefore, only 82 values for the mean nitrate concentration were retained in this study. Figure 1 gives an example of the distribution of mean nitrate concentrations collected and overlaid on a groundwater pollution risk map (Ouedraogo et al. 2016). The figure shows that nitrate concentration in groundwater is irregularly distributed in Africa. However, this shows a coherent distribution according to a groundwater vulnerability map, i.e. where the groundwater is subjected to pollution risk (Ouedraogo et al. 2016). In this paper, only the reported mean nitrate concentrations were studied. This choice was guided by the fact that this category of data is less sensitive to outliers and more robust than minimum or maximum concentration data. Groundwater nitrate concentration data are summarized in Table 1. In the present study, the modeled response variable is the natural log of sampled groundwater nitrate concentration. The average mean nitrate concentration is 54.85 mg/L, the standard deviation is 89.91 mg/L, and the median concentration is 27.58 mg/L. The log transform reduces the influence of very high nitrate values (up to 648 mg/L) on model predictions. The nitrate studies were separated into two groups: 80% of the data set for building the model (training data set) and 20% for validation of the model (testing data set).

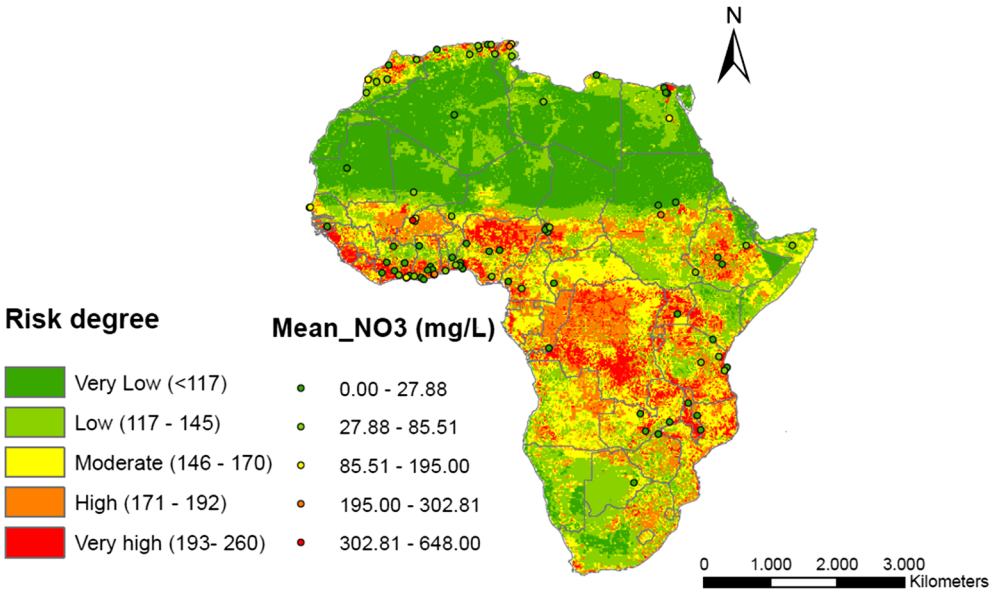

Fig. 1 Spatial distribution of mean nitrate concentration in groundwater in Africa (Ouedraogo et al. 2016)

Table 1 Summary statistics of original and ln-transformed nitrate data (from Ouedraogo and Vanclooster 2016a)

Variable	Mean NO_3^- concentration	Mean $\ln(NO_3^-)$ concentration
Minimum [mg/L or ln(mg/L)]	1.26	0.231
Maximum [mg/L or ln(mg/L)]	648	6.473
Mean [mg/L or ln(mg/L)]	54.85	3.169
CV (−)	8085.08	1.935
Standard deviation [mg/L or ln(mg/L)]	89.91	1.391
Median [mg/L or ln(mg/L)]	27.58	3.317
Variance [$(mg/L)^2$ or $\ln(mg/L)^2$)]	163.92	43.901
Kurtosis [mg/L or ln(mg/L)]	23.99	−0.167
Skewness [mg/L or ln(mg/L)]	4.31	−0.294
Number of observations (−)	82	82

CV coefficient of variation

The data set of explanatory factors

Data were collected from a total of 13 possible explanatory factors, extracted from several high-resolution databases covering physical and anthropogenic attributes. These attributes are related to characteristics including land use, soil type, hydrogeology, topography, and climatology. Table 2 presents the 13 explanatory factors, their spatial resolution, and their data sources. All explanatory factors were integrated into a geographical information system (GIS) and processed in

Table 2 Explanatory variables and nitrate concentrations used in the random forest model

Name	Type	Units or categories	Spatial resolution/ scale	Date	Data source(s)
Explanatory variables					
Land cover/land use	Categorical data	–	300 m	2014	Personal communication, UCL/ELIe-Geomatics (Belgium) (2014)
Population density	Continuous point data	people/km^2	2.5 km	2004	ESRI (1969) ArcGIS
Nitrogen application	Continuous point data	kg/ha	0.5° × 0.5°	2010	SEDAC (2010)
Climate class	Categorical data	–	0.5°	1997	Personal communication: P. Trambauer (UNESCO-IHE, Delft, Netherlands) (2015)
Type of region	Categorical data	–	0.5°	2014	Personal communication: P. Trambauer (UNESCO-IHE, Delft, Netherlands) (2015)
Rainfall class	Categorical data	mm/year	3.7 km	1986	UNEP (1986)
Depth to groundwater	Categorical data	m	0.5° × 0.5°	2011	British Geological Survey (BGS 2011)
Aquifer type	Categorical data	–	1:3750 000	2012	Personal communication: N. Moosdorf (Hamburg University) (2014)
Soil type	Categorical data	–	1 km × 1 km	2014	ISRIC (2014), World Soil Information:
Unsaturated zone (impact of vadose zone)	Categorical data	–	1:3750 000	2012	Personal communication: N. Moosdorf (Hamburg University) (2014)
Topography/slope	Continuous point data	%	90 m	2000	Personal communication: UCL/ELIe-Geomatics (Belgium) and CGIAR/CSI (SRTM data) (2014)
Recharge	Continuous point data	mm/year	5 km	2008	Personal communication: P. Döll and F. Portman (University of Frankfurt) (2014)
Hydraulic conductivity	Continuous point data	m/day	Average size of polygon ~100km^2	2014	Personal communication: T.P. Gleeson (McGill University) (2014)
Response variables					
Nitrate	Continuous point data	mg/L	–	2000 to 2015	Ouedraogo and Vanclooster (2016a)

ArcGIS™ 10.3 in a raster format of 15 km × 15 km spatial resolution. This resolution is the best compromise considering the resolution of the different available data sets and the large extent of the study area.

Climate

Climate conditions have an effect on the probability of nitrate occurrence in groundwater. The climate class data used is the "Derived Climate Classes for the African Continent" raster format with a resolution of 0.5° and grouped into five classes (Trambauer et al. 2014). The type of region data in raster format were obtained directly from P. Trambauer from UNESCO-IHE (Delft, the Netherlands) and grouped into six classes according to the UNEP classification (1997). The rainfall map was generated from the UNEP/FAO World and Africa GIS database.

Topography

Land surface slope indicates whether the runoff will remain on the surface to allow contamination percolation to the saturated zone. The slope was inferred from the 90-m Shuttle Radar Topography Mission (SRTM90) topographic map, using ArcGIS™10.3 Spatial Analyst software. Elevation values at 90-m resolution were aggregated at 15 km.

Soil type and land use

Soil texture influences nitrogen loss. Nitrate leaching is generally greater from poorly structured sandy soils than from clay soils, because of the slower water movement and the greater potential for denitrification to occur in the clay soils (Cameron et al. 2013). The soil texture was derived in this study from soil grids at a 1-km resolution produced by Hengl et al. (2014). There are nine classes of soil texture: sandy, sandy loam, loamy sand, loam, clay loam, sandy clay, sandy clay loam, silty clay loam, and clay.

Land cover and land use were produced from the high-resolution GlobCover data set (Defourny et al. 2014). To create a consistent measurement of land cover, the original 22 classes representing the African area were aggregated into six dominant classes (water bodies, bare area, grassland/shrubland, forest, urban, croplands).

Geology and hydrogeology

The water table and vadose zone thickness determine the nitrogen transfer time and dilution potential from the surface to the groundwater. The aquifer medium is a hydrogeological factor that describes the ability of pollutants to move within this aquifer according to its type. In this study, the depth to groundwater map was inferred from data presented by Bonsor and MacDonald (2011). The aquifer and vadose zone type were derived from the high-resolution global lithological database (GliM) of Hartmann and Moosdorf (2012). The aquifer type and unsaturated lithological zone were derived for each of the five hydrolithological and lithological categories as defined by Gleeson et al. (2014). These categories are unconsolidated sediments, siliciclastic sediments, carbonate rocks, crystalline rocks, and volcanic rocks.

Recharge to the groundwater as a portion of rainfall amount depends on rainfall characteristics, soil permeability, and topographic setting. Groundwater recharge is a function of many parameters, including soil type, antecedent soil water content, land cover, and rainfall, among others (Anuraga et al. 2006; Sophocleous 2004). In this study, the recharge was inferred from the global-scale modeling of groundwater recharge presented by Döll and Fiedler (2008).

Hydraulic conductivity and transmissivity are important for the assessment of the aquifer's ability to transmit water, and to determine the rate of flow of a contaminant material within the groundwater. The hydraulic conductivity of aquifers was determined using the Global Hydrogeology MaPS (GHYMPS) of permeability and porosity data (Gleeson et al. 2014).

Nitrogen fertilizer application

The mean nitrogen fertilizer application map (kg/km^2) was generated from the Potter et al. (2010) data set. The values shown on this map represent an average application rate for all crops over 0.5° resolution grid cell. Following this study, the highest rate of nitrogen fertilizer application (i.e. 220 kg/ha) is found in Egypt's Nile Delta.

Demography

The population density represents the distribution of potentially causative agents, considering that nitrate contamination on the African continental scale is mainly human-caused. Data on population density at the African scale were obtained from the UNEP website. The population density for the year 2000, considered in this study, was produced by Nelson (2004).

Description of the models

The preliminary analysis of the mean nitrate data revealed a non-normal distribution. Different transformations were tested in order to obtain a normal distribution of the dependent variable, as required in the multiple linear regression (MLR) model. The original data set was randomly divided into calibration (80%) and validation (the remaining 20%) samples. An MLR and a random forest regression (RFR) model were identified and validated. All models were implemented using R statistical software (version 3.1.1; R Development Core

Team 2015). The goodness of fit was evaluated using the FITEVAL model (Ritter and Muñoz-Carpena 2013).

Multiple linear regression (MLR)

Regression techniques such as linear regression (Boy-Roura et al. 2013) or logistic regression (Nolan and Hitt 2006) have been widely applied in nitrate modeling. Typically, these models aim to use the fewest predictors to explain the greatest variability in the response variable (Graham 2003). Stepwise approaches are used to select the most relevant predictors in regression models, using different selection criteria such as the Akaike information criterion (AIC), Schwarz's Bayesian information criterion of the *F*-statistic, and others (Murtaugh 2009). In the case of this study, the AIC criterion was used to select the predictors. The rationale for this selection method is to combine the measure of fit with a penalty term based on the number of parameters used in the model. If more parameters (i.e., the number of trends or explanatory variables) are used, the model fit can be better, but the penalty for the extra parameters is higher as well. The smallest AIC indicates the most appropriate model.

Random forest regression (RFR)

An RFR model was identified on the same data set. RFR modeling is an ensemble machine learning method for classification and regression that operates by constructing a multitude of decision trees (Breiman 2001a). The philosophy behind ensemble learning techniques is based on the premise that its accuracy is higher than other machine learning algorithms because the combination of predictions performs more accurately than any single constituent model. The individual decision trees in RFR tend to learn highly irregular patterns, i.e. they overfit their training data sets. RFR is a means of averaging multiple decision trees, trained on different parts of the same training data set, with the goal of reducing the prediction variance (Hastie et al. 2008). RFR modeling is appropriate for modeling the nonlinear effect of variables. It can handle complex interactions among variables, and is not affected by multicollinearity (Breiman 2001b). RFR can assess the effects of all explanatory variables simultaneously, and automatically ranks the importance of these variables in descending order (Rodriguez-Galiano et al. 2014). Rodriguez-Galiano et al. 2014 argue that the generalization error converges as the number of trees increases; therefore the random forest (RF) does not overfit the data.

A detailed description of the mathematical formulation for the RFR model is found in Breiman (2001a) and Liaw and Wiener (2002). The algorithm for RFR consists in building a forest of uncorrelated trees. Each individual tree is grown using a randomized subset of predictor variables. The trees are grown to the largest extent possible without pruning, and they are aggregated by averaging them. Out-of-bag (OOB) samples are used to calculate variable importance and to get an unbiased estimate of the test set error, which is one of the advantages of RFR, because there is no need for cross-validation. The method behaves essentially as a "black box" since the individual trees cannot be examined separately (Prasad et al. 2006) and it does not calculate regression coefficients or confidence intervals (Cutler et al. 2007). Nevertheless, it allows the computation of variable importance measures that can be compared to other regression techniques (Grömping 2009). To run the analysis of the decision trees, regression methods are carried out from the data set (for details, e.g. flowchart of RF, see Rodriguez-Galiano et al. 2014). In this study, the "randomforest" package of R version 3.1.1 (open-source statistical software; R Development Core Team 2015) was used for all modeling.

In addition, 13 explanatory variables were used (Table 2). In order to maintain a similar procedure between MLR and RFR, RFR was applied on the training and the validation data set. Even though independent validation samples are not required in RFR, they provide the opportunity to assess the generalization capability of this method (Cutler et al. 2007). To run the RFR model, it was necessary to define a priori two parameters: the number of variables or factors to be used in each tree-building process (mtry), and the number of trees to be built in the forest to run (ntree). Concerning the number of trees, Breiman (1996) demonstrated that by increasing the number of trees, the generalization error always converges; hence, overtraining is not a problem, and the number of trees can be fixed once the error has converged. Rodriguez-Galiano et al. (2014) demonstrated that the error was minimum and stable when considering 1000 trees. The parameter mtry was determined via the internal random forest function TuneRF, which recognizes the optimal number of factors (the default value of mtry is a total number of variables/3 for regression), and looks below and above this threshold for the value of the minimum OBB error rate. Breiman (2001a) and Liaw and Wiener (2002) stated that even a single variable or factor (mtry = 1) could generate good accuracy, while Grömping (2009) proved the need to include at least two variables/factors (i.e. mtry = 2, 3, 4,…, *m*) in order to avoid using the weaker regressors as splitters.

Variable importance in random forest

The RFR, used to predict nitrate concentration in groundwater, allows one to estimate the variable importance of environmental attributes in the explanatory model. The advantage of the RFR algorithm is that it allows one to explicitly measure variable importance with two metrics: mean decrease in the GINI index and mean decrease in accuracy (%IncMSE). [GINI is defined as "inequality" when used in describing a society's distribution of income, or a measure of "node

impurity" in tree-based classification.] The mean decrease in the GINI index is used to measure the quality of a split for each variable in a tree, while the mean decrease in accuracy, based on the mean squared error (MSE), measures the mean decrease in prediction accuracy. In others words, the former measures the node impurity of the explanatory factors, while the latter measures the contribution of the factor toward the overall fit. Each of these metrics measures the impact of the explanatory factor on the overall prediction (Breiman 2001a). However, the GINI index has been shown to have a bias (Strobl et al. 2007). Also, Genuer et al. (2010) determined that the percentage in explained mean squared error is a more reliable measure than the decrease in node impurity. Therefore, the percentage in explained mean squared error was used to assess variable importance.

Model validation

The quality of the statistical models was evaluated by means of the FITEVAL code (Ritter and Muñoz-Carpena 2013). This code uses the most general formulation of the coefficient of efficiency. For a complete model goodness-of-fit evaluation, the graphical results of FITEVAL contain the following elements (Ritter and Muñoz-Carpena 2013): (1) a plot of observed versus estimated values illustrating the match of the 1:1 line; (2) the evaluation of NSE (Nash-Sutcliffe coefficient of efficiency) and the root mean square error (RMSE) and their corresponding 95% interval; (3) the qualitative goodness-of-fit interpretation based on established classes (unsatisfactory, acceptable, good, very good); (4) a verification of the presence of bias or the possible presence of outliers; (5) the plot of the NSE cumulative probability function superimposed on the NSE class region [the NSE has several class regions defined by Ritter and Muñoz-Carpena (2013)]; and (6) a plot illustrating the evolution of observed and estimated values.

Results

Modeling results

Multiple linear regression (MLR)

The final MLR model built includes four variables: (i) depth to groundwater, (ii) recharge, (iii) aquifer type, and (iv) population density. Table 3 summarizes the results of this MLR model. The percentage of variance explained by the model is 64%, and the residual standard error is 0.95 [ln(mg/L)]. The sign of the parameter coefficient indicates the direction of the relationship between independent and dependent variables (Boy-Roura et al. 2013). The lower the p value, the more

Table 3 Optimal MLR model for explaining the ln-transformed mean nitrate concentration. See Table 2 for units

Variable	Estimate	SE	t value	Probability, Pr (>|t|)
(Intercept)	3.427e+00	7.306e-01	4.690	2.08e-05 ***
Depth [0–7]	1.384e+00	5.003e-01	2.766	0.00789 **
Depth [7–25]	7.322e-01	4.603e-01	1.591	0.11788
Depth [25–50]	1.408e+00	5.645e-01	2.493	0.01594 **
Depth [50–100]	1.000e+00	5.185e-01	1.929	0.05928*
Depth [100–250]	1.332e+00	8.694e-01	1.532	0.13175
Recharge [0–45]	−6.094e-01	6.903e-01	−0.883	0.38154
Recharge [45–123]	−1.580e+00	6.775e-01	−2.333	0.02365 **
Recharge [123–224]	−1.334e+00	6.638e-01	−2.010	0.04974 **
Recharge [224–355]	−9.021e-01	6.535e-01	−1.380	0.17350
Aquifer media [Crystalline rocks]	−1.116e+00	4.010e-01	−2.783	0.00753 **
Aquifer media [Siliciclastic sedimentary rocks]	−1.196e-01	4.770e-01	−0.251	0.80306
Aquifer media [Unconsolidated sediments rocks]	−8.010e-01	3.954e-01	−2.026	0.04802 **
Aquifer media [Volcanic rocks]	−4.044e-01	6.658e-01	−0.607	0.54631
Population density (people/km^2)	5.982e-04	8.536e-05	7.008	5.29e-09 ***

Residual standard error: 0.95 on 51 degrees of freedom

Multiple R-squared: 0.64

F-statistic: 6.75 on 14 and 51 degrees of freedom, p value = 1.688e-07 < 0.001

*** significant at $p < 0.001$; ** significant at $p < 0.05$ and * significant at $p < 0.1$

SE standard error

significant the model parameter. Only explanatory variables with p values ≤0.1 were retained.

The MLR results show that the population density has a strong positive relationship with ln-transformed mean nitrate concentration. As the p value is <0.0001, this variable strongly affects the nitrate occurrence in groundwater at the African continental scale.

The second variable included in the model is the depth to groundwater. The three classes [0–7, 25–50, and 50–100 m below ground level (b.g.l)] of this variable are all statistically significant. Among these three classes it was observed that the 0–7 m class has the strongest statistical significance and a positive coefficient, indicating a large contamination for shallow groundwaters. Regarding the largest groundwater depth class (100–250 m b.g.l), this class is not statistically significant (p value >0.05). Therefore, it is affirmed that the shallow groundwaters in Africa are more vulnerable than deep aquifers to nitrate pollution.

The third variable retained in the model is recharge. The two classes of recharge rate (45–123 and 123–224 mm/year) are statistically significant and correspond in general to semi-arid and dry sub-humid climate. The high nitrate in groundwater associated with these climate conditions can be explained by the intensive agricultural development in the regions that strongly relies on irrigation.

The last variable included in the final model is the aquifer media. It was observed that two categories of aquifer media were significant in the model: crystalline rocks and unconsolidated sediment rocks. Indeed, the analysis of their negative regression coefficient estimates shows that the likelihood of nitrate contamination decreases with the presence of unconsolidated sediments and crystalline rocks. The other categories of aquifer type, namely siliciclastic sedimentary rocks and volcanic rocks, are found to be statistically insignificant in the model. However, the variable of aquifer type is an important parameter for assessing groundwater vulnerability and provides information about the hydrogeological setting.

A plot of predicted versus observed log nitrate (ln NO_3) concentration for the training data, shown in Fig. 2a, indicates that the MLR model shows an acceptable fit for the observed data.

Estimating independent variable importance

The variable importance plot is a critical output of the random forest algorithm. Figure 3 shows the ranking of the relative importance of environmental attributes with regard to groundwater nitrate occurrence. Higher values of percentage increase in mean squared error (MSE) indicate higher importance. Population density is the most important predictor of nitrate concentration. This can be considered a relevant finding, because population has a direct effect on nitrate pollution in groundwater. If this variable were omitted, the quality of the model could be drastically reduced. Rainfall is also found to a relevant predictor of nitrate contamination, followed by recharge and aquifer type. The most important intermediate variables are nitrogen fertilizer application, climate classes, and hydraulic conductivity, while land use, depth to water, slope, soil media, and type of region are in the bottom of the ranking, and thus have the smallest influence on the quality of the random forest (RF) model.

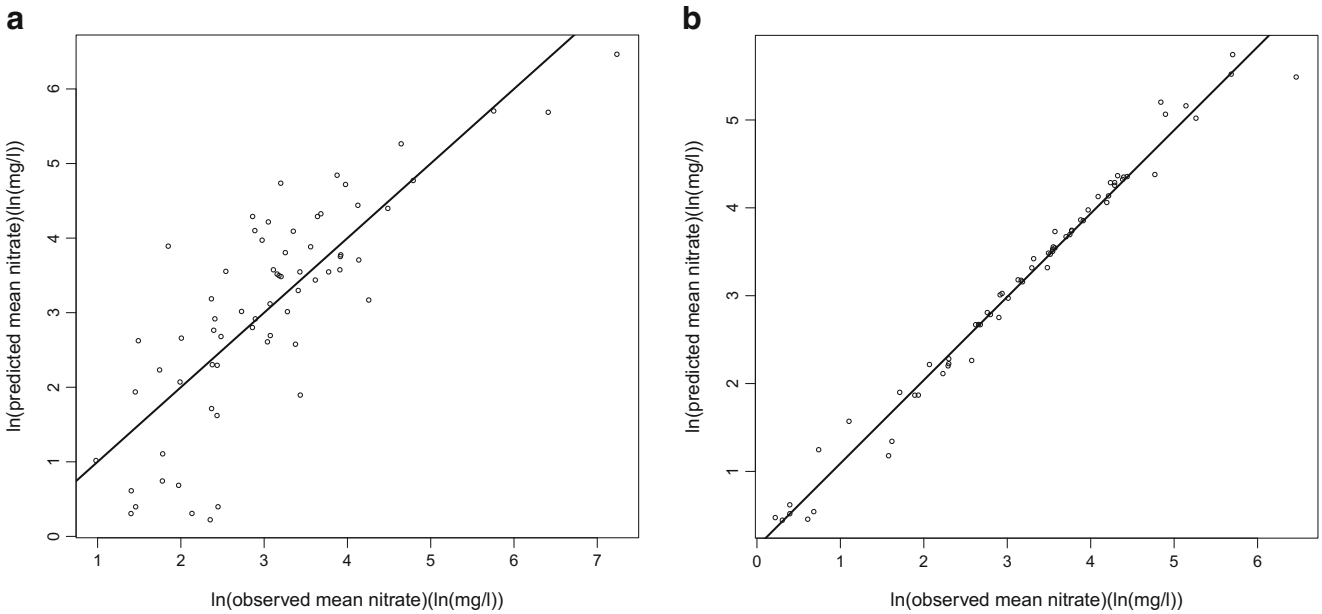

Fig. 2 Comparison of observed and predicted log (ln) nitrate concentration on training data set: (**a**) linear regression and (**b**) random forest regression

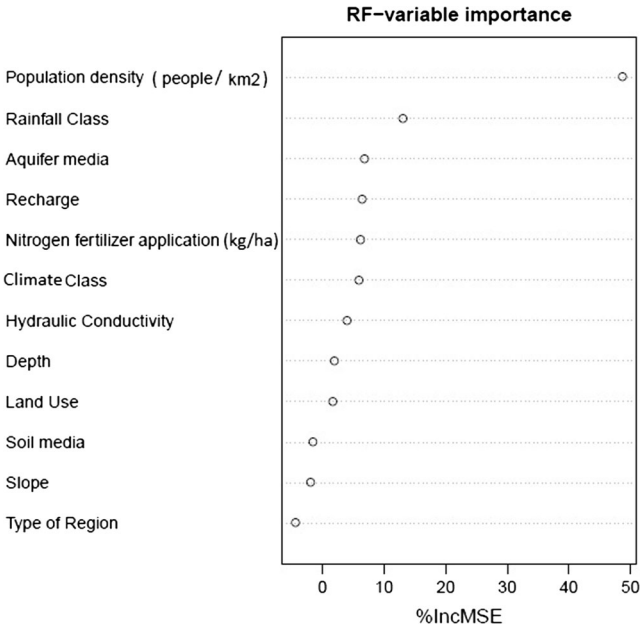

Fig. 3 Variable importance according to the percentage increase in mean squared error (%IncMSE)

Application of random forest regression

The mtry parameter is the number of predictors used at each split. The tuneRF function suggests an optimal value for mtry = 4. This value is equivalent to the default value for mtry. The ntree was specified as follows: ntree = 1000. The final model was built with the first four variables shown in Table 4. The percentage of variance explained with this model is 97.9% [mean of squared residuals = 0.0407 ln(mg/L)]. In this model, the fourth most important variable is population density, rainfall class, recharge, and aquifer type. Figure 2b shows a plot of the predicted versus the observed ln-transformed nitrate concentration values for the training data, based on only 80% of observations. The RFR shows a better fit. Figure 4b shows the predicted versus the observed values for the test data, based on 20% of observations.

Table 4 Variables included in the final RFR model, in descending order of importance based on percentage of mean decrease in accuracy (% IncMSE)

Variable	% IncMSE
Population density (people/km^2)	50.3
Rainfall class	10.2
Aquifer media	5.3
Recharge	5.2

Evaluation of model performance

The results of the MLR and RFR models explain 64 and 97%, respectively, of ln-transformed mean nitrate concentration. To validate the predictive ability of these two models, FITEVAL was used for evaluating first the training and then the performance of the validation. Figures 5 and 6 illustrate the results of the FITEVAL evaluation for the two models. Figure 5a suggests that MLR yields unsatisfactory to acceptable results (NSE = 0.554 [0.129–0.743]). In this figure, scattered data follow the 1:1 line, but calculated values deviate from the observations positively and negatively along the prediction range. Figure 5b suggests that RFR exhibits very good (NSE = 0.998 [0.996–0.999]) prediction. This model demonstrates very good performance, since it clearly shows that computed values are very similar to the observations.

As regards the testing data set, the results of the MLR model (Fig. 6a) vary from an unsatisfactory to acceptable prediction (NSE = 0.289 [−0.462 to 0.7]). The scattered data follow the 1:1 line, but the plot scale of observed versus predicted values is reduced substantially. On the basis of the qualitative goodness of-fit, the model performance is considered unsatisfactory. Figure 6b illustrates the results of the RFR model. The fit shows very good prediction of groundwater nitrate pollution (NSE = 0.981 [0.959–0.991]). The model performance is significant even though there is a probability of obtaining an NSE < 0.65 (28.1%). The performance is considered good on the basis of the qualitative goodness of fit.

Discussion

Nitrate contamination of groundwater at the African continental scale is spatially variable. The results of this study show that the contamination by nitrate is influenced by both physical and anthropogenic factors. Previous studies suggest that nonlinear relationships exist between nitrate concentration and the independent variables (Rodriguez-Galiano et al. 2014; Kihumba et al. 2015; Wheeler et al. 2015). In these studies, the superiority of nonlinear techniques for predicting groundwater nitrate concentrations as compared to linear multiple regression models was demonstrated. A nonparametric nonlinear statistical model is therefore considered suitable for modeling observed nitrate concentrations at the continental scale.

The comparison between the results shows that RFR has much higher predictive ability than MLR, although MLR shows an acceptable but weak relationship between the dependent variable and the predictors. The RFR exhibited better performance than MLR in both the training and validation data set, explaining 97.9 and 98% of the variation in ln-transformed nitrate concentrations, respectively (see Figs. 2b and 5b). By comparison, the MLR explained only 64% of the

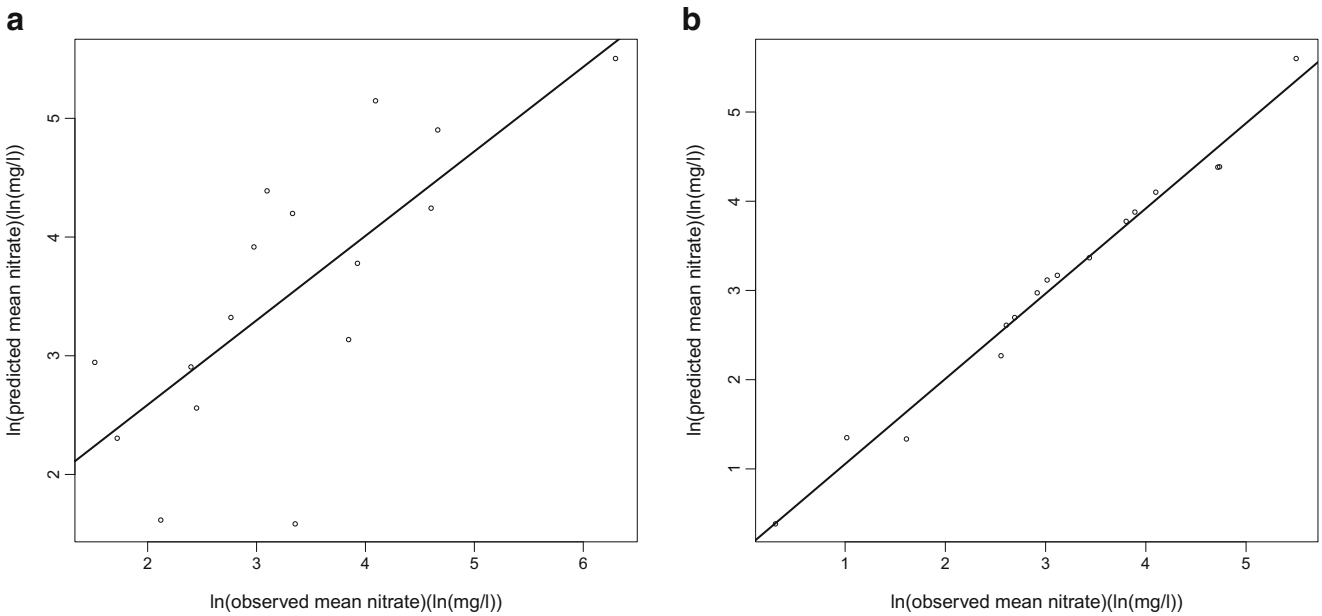

Fig. 4 Comparison of validation results on the tested data set: (**a**) linear model and (**b**) random forest regression

variation in ln-transformed nitrate concentration for the training data set. The better performance of RFR is due to its ability to handle nonlinear relationships between the nitrate pollution and explanatory factors.

The prediction errors for the final random forest model and linear model are listed in Table 5 for both the training and testing data sets.

The most important variable identified in both models was population density. This is most likely related to the lack of sanitation in major parts of the African continent and the development of small-scale farming systems in peri-urban environments. The strong influence of population density on groundwater nitrate contamination is consistent with the findings of a previous UNEP study (UNEP/DEWA 2014).

Rainfall is identified in the RFR as the second most important variable, whereas this variable is not among the four variables identified as important in the final MLR model. This may indicate that a nonlinear association exists between rainfall and nitrate concentration, and as such, it was better identified by the RFR procedure. According to Pearson (2015), rainfall is indeed a major explanatory factor for nitrate occurrence in groundwater environments. Following Kulabako et al. (2007), rainfall is the primary climatological control factor that aids in washing of contaminants down to the shallow groundwater. Kulabako et al. (2007) add that the increased nitrate in the shallow groundwater after rain suggests that the system receives high loads of organic nitrogen through leaching (from pit latrines, drains, solid waste dumps, animal waste dumps, etc.). Several studies have found that increasing precipitation may positively affect groundwater nitrate concentrations (Davis and Sylvester-Bradley 1995; Rankinen et al. 2007). Conversely, other studies suggest that higher average precipitation could promote the uptake of nitrogen by crops (Schweigert et al. 2004; Sieling and Kage 2006) or support the dilution of nitrate-containing substances (Hofreither and Pardeller, 1996, cited in Wick et al. 2012), and hence reduce potential nitrate leaching. These opposing effects suggest that the coefficient of precipitation could have either a negative or a positive sign. In the MLR analysis in this study, the sign is unknown, because the precipitation parameter is not explicitly in the final model, which indicates that the linear model is not well adapted for the interpretation of this data set of nitrate at the African continent scale.

Recharge and aquifer type come in third and fourth in the RFR model, and they occupy the third and second positions, respectively, in the MLR model. The influence of groundwater recharge rate on nitrate contamination is consistent with studies such as Saffigna and Keeney (1997) and Hanson (2002). In addition, according to UNEP/DEWA (2014), groundwater microbial and chemical water quality is influenced by recharge from multiple sources. The groundwater recharge rate is interlinked with many other environmental variables, including soil type, aquifer type, antecedent soil water content, land use/land cover type, and rainfall (Sophocleous 2004; Anuraga et al. 2006). A recent study in the shallow unconfined aquifer of the Piemonte plain (northern Italy) reported that dilution could be regarded as the main cause of nitrate attenuation in groundwater (Debernardi et al. 2007). Furthermore, Andrade and Stigter (2009) and Nolan and Hitt (2006) found that dilution by surface-water irrigation was an important attenuation process. The negative sign on the recharge factor in the MLR method is consistent with these studies, but caution should be taken in interpreting the sign

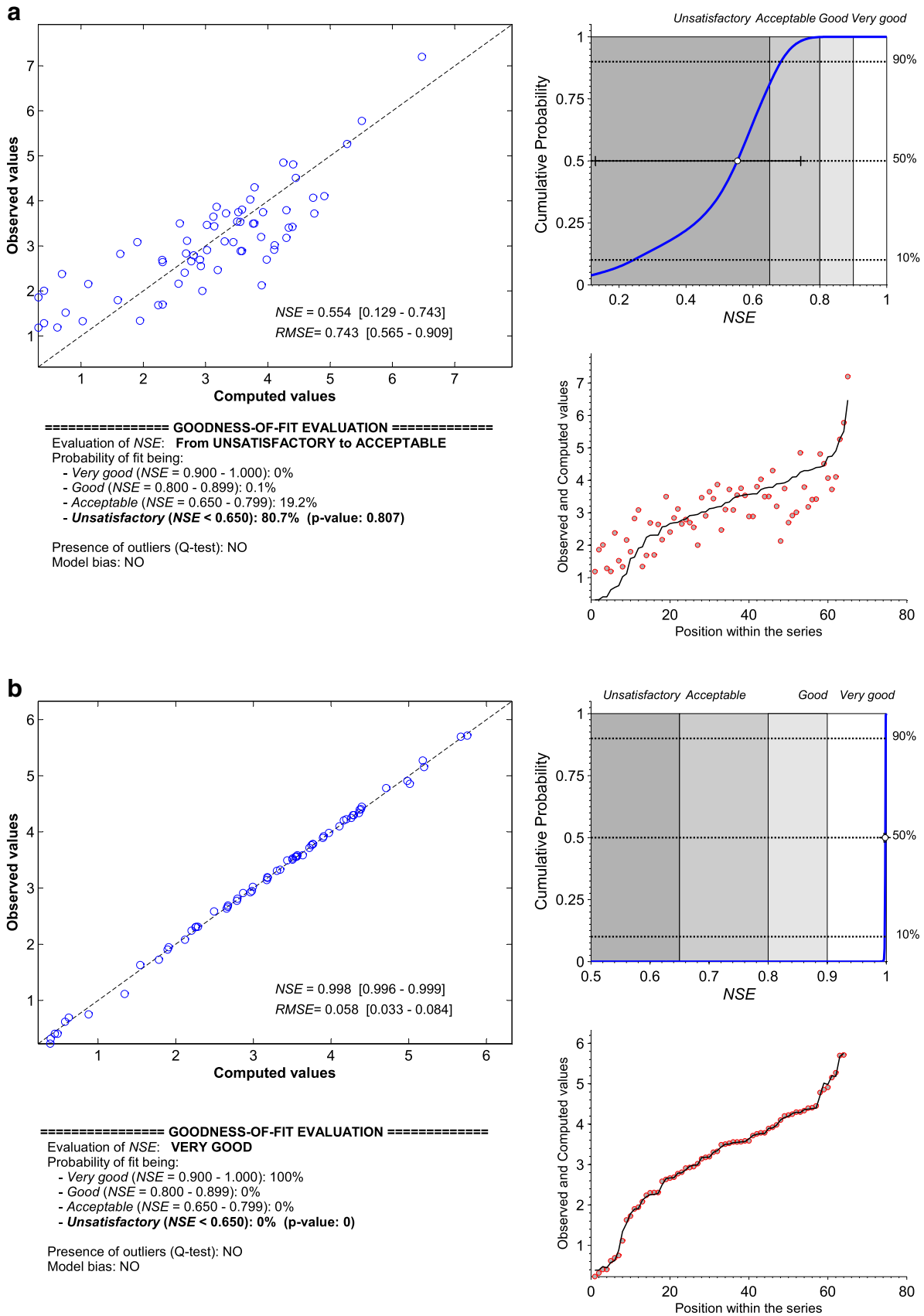

Fig. 5 Goodness-of-fit evaluation for the training data set in the (**a**) MLR and (**b**) RFR. NSE = Nash-Sutcliffe coefficient of efficiency

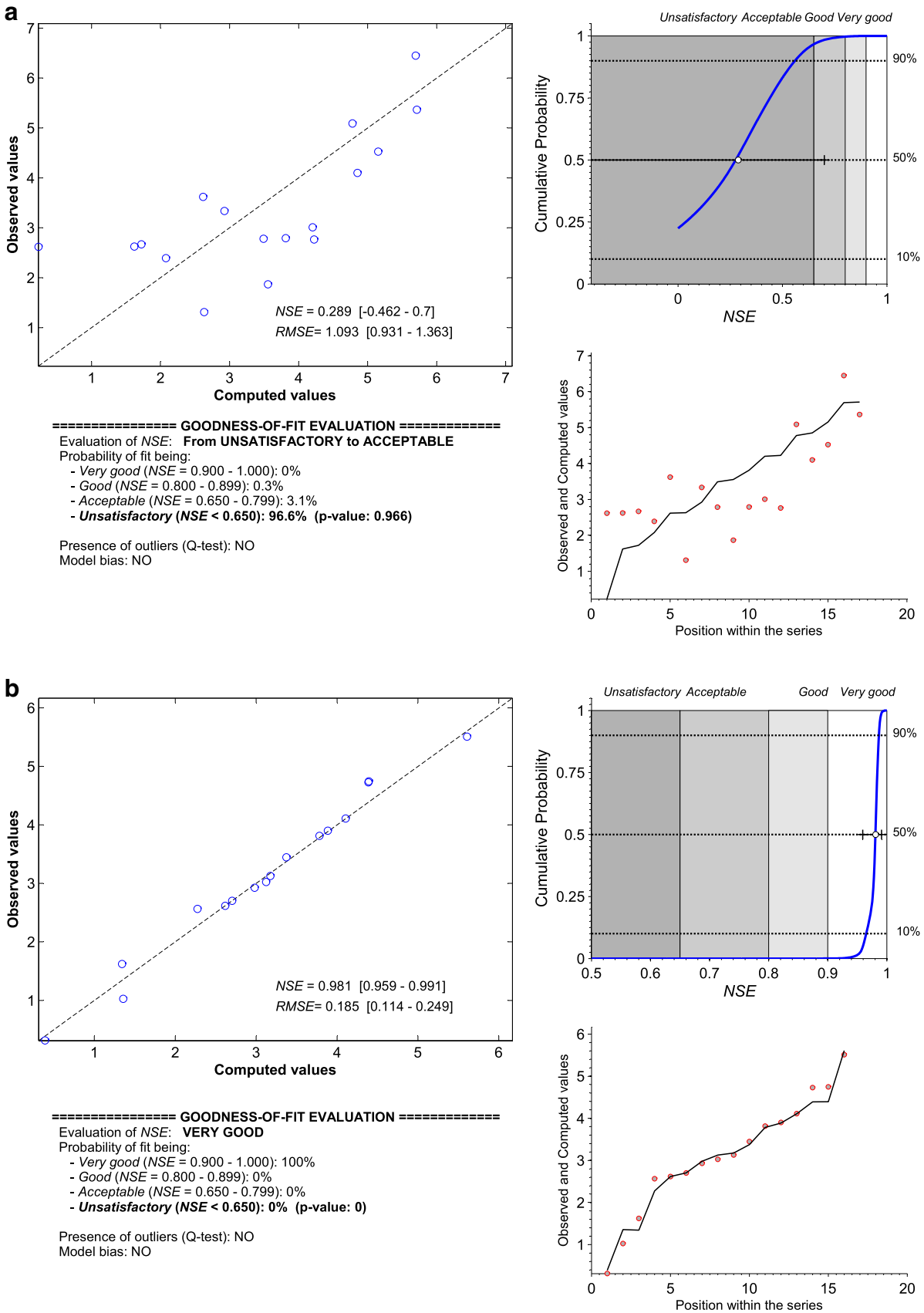

Fig. 6 Goodness-of-fit evaluation for the validation data set in the (**a**) MLR and (**b**) RFR

Table 5 Predictive performance of random forest and multiple linear regression

Model	Training data		Testing data	
	MSE	% variation explained	MSE	% variation explained
Random forest	0.04	0.97	0.01	0.98
Linear regression	0.95	0.64	0.75	0.54

MSE mean squared error

in the MLR result for this factor, because MLR does not deal with the nonlinearity of possible relationships. The influence of aquifer type on nitrate contamination was demonstrated by Boy-Roura et al. (2013), while Stigter et al. (2008) showed that nitrate concentration correlated with aquifer media and land use.

Nitrogen fertilizer is the fifth most important variable in the RFR model. Previous studies have already observed an association between groundwater nitrate pollution and nitrogen fertilizer loading/manure application (Greene et al. 2005; Nolan et al. 2002). Nolan et al. (2014) found that nitrogen fertilizer was the most important in their RFR model, while Boy-Roura et al. (2013) observed that net nitrogen load (kg/ha) provided better performance for their MLR. They concluded that nitrate concentration in groundwater increased with increasing net loading.

Climate classes are found in sixth position in the RFR. This is consistent with previous studies showing the importance of climatic conditions in groundwater degradation (Wick et al. 2012; Ramasamy et al. 2003). For instance, Fram and Belitz (2011) used aridity index data to develop a logistic regression model for predicting the probability of detecting perchlorate at significant concentrations in California and the southwestern United States. They demonstrated that the predicted probability of perchlorate occurrence is a function of climate, expressed in terms of the aridity index. This index incorporates precipitation and evapotranspiration.

As regards the groundwater depth factor, the absence of this important factor among the top six variables in the RFR model could be due to the use of interval depth as a proxy. This surprising result is in contrast to various other studies showing that the nitrate concentration in groundwater decreases with increasing sampling depth (Tesoriero and Voss 1997; Nolan and Hitt 2006; Nolan et al. 2014; Wheeler et al. 2015; Ouedraogo and Vanclooster 2016b). Groundwater age is a fundamental characteristic of groundwater that is affected by diverse geologic processes (Gassiat et al. 2013). Geological age is defined as the time that has passed since the water entered the groundwater system (Alley et al. 2002; Kazemi et al. 2006). The distribution of groundwater age depends on many factors, including permeability, recharge rate, aquifer geometry, and topography (Gassiat et al. 2013). Deeper groundwater is typically older and may predate periods of intensive fertilizer application (1950 to present), and there is greater opportunity for denitrification because groundwater requires more time to travel to deeper aquifers. According to Dubrovsky et al. (2010), the deeper the well, the more likely that the sampled groundwater represents a mixture of different ages and land uses. Furthermore, Luo et al. (2003) observed that there is a significant downward movement of nitrate, and hence nitrate concentration at a certain depth will decrease with time. Age is among the most important variables controlling groundwater nitrate concentration, but it remains one that is difficult to estimate (Nolan et al. 2015). Numerical groundwater flow models may allow for estimation of the groundwater travel time. Such models have demonstrated that travel time increases with increasing distance up-gradient of a well (Masterson et al. 2002). According to Dubrovsky et al. (2010), a 10-year travel time from recharge to a well is reasonable for areas with well-drained soils and flat topography. Similar studies of large recharge times for deep sandy aquifers were also identified by Mattern and Vanclooster (2009). As a corollary to age and travel time, denitrification is recognized for its role in significantly reducing nitrate in soil and groundwater (Stevenson and Cole 1999; Thayalakumaran et al. 2004; Aljazzar 2010).

Other variables, such as hydraulic conductivity, land cover/land use, slope, soil media, and type of region, were not included in either final model. This is in contrast to studies where these factors were considered as being important for explaining groundwater degradation (Gemitzi et al. 2009; Wick et al. 2012; Liu et al. 2013; Jung et al. 2015; Kihumba et al. 2015).

Validating predictions for the entire continent of Africa is certainly not trivial. Figure 1 clearly demonstrates that efforts to study the groundwater nitrate contamination problem are not distributed equally among the African countries. For many smaller countries, such as Liberia, Guinea, Equatorial Guinea, and Sierra Leone, but also for large land areas such as the Democratic Republic of the Congo, Chad, Namibia, Angola, and South Sudan, there are few available studies on the nitrate pollution problem. Many countries in Africa face considerable difficulties with systematic groundwater monitoring. The major constraint in the validation of the modeling study, therefore, lies in the unavailability of a homogeneous data set on groundwater nitrate contamination. Results from this analysis should therefore be interpreted with caution. While the available data inferred from a meta-analysis provide a useful preliminary assessment of the nitrate contamination issue in

Africa at a continental scale, there are clear limitations. First, the data are from different sources, and the methods used to collect and produce the results of each study were not the same. Second, bias may have been introduced due to the setup of the different reported studies. Certain studies addressed groundwater nitrate contamination as support for a drinking water supply project and others for an irrigation project, while still others addressed the issue within a mining context. The different setups of these studies may therefore introduce a possible bias. The data used to calibrate and validate the models thus should not be treated the same as nitrate measurements that are collected in standardized groundwater monitoring programs. Further studies based on less biased data sets should be performed to demonstrate the robustness of the developed models.

It may be expected, therefore, that a revised robust model can be identified if new homogeneous data sets become available. It may also be expected that such new data sets will be produced and improved continuously over time. As time progresses, better and more homogeneous data sets will become available. It is thus suggested that the statistical models that were identified in this study would be integrated within a time-related dynamic data assimilation framework. The RFR model structure that outperforms in this study, as demonstrated by the objective goodness-of-fit analysis, would be an excellent modeling structure for integration into such a framework.

Despite all these limitations, the results in this study have already provided valuable insight into the potential causes of nitrate occurrence in groundwater across the African continent that may be considered in groundwater management and protection programs. The analysis confirmed the importance of population density as a controlling factor, in addition to a set of other physical environmental attributes, particularly those related to aquifer properties (rock materials) and climatic conditions. Groundwater management and protection programs should therefore first focus on the densely populated areas and consider the remediation of anthropogenic pollution sources such as leaking water sanitation systems, poorly controlled livestock systems, and urban agriculture.

Conclusion

In this study, the potential for using random forest regression (RFR) techniques to model nitrate in groundwater at the African continental scale was explored. To this end, the performance of the RFR was compared with that of a multiple linear regression (MLR) model.

The results of the study illustrate that the RFR outperforms MLR in modeling nitrate concentration at the African continental scale. The good performance of the RFR is attributed to its nonparametric nature, i.e., it does not need to follow a normal distribution. Moreover, its robustness against outlier values is greater than that of other methods, as each tree in the RFR is generated from different data subsets. A perceived disadvantage of this machine learning method is its "black-box" nature. The RFR does not allow for directly estimating coefficients related to explanatory variables. However, the RFR allowed ranking of the variables according to their relative contribution to the model using a nonparametric approach.

The validation of the MLR and RFR models based on an objective goodness of fit confirmed the good performance of the RFR as compared to the MLR. The RFR is, therefore, a promising technique for modeling groundwater degradation because of its ability to provide meaningful analysis of nonlinear and complex relationships such as the ones found in hydrogeological studies.

However, model performance could be influenced by the sample size and bias in nitrate concentration values collected. Furthermore, groundwater nitrate pollution is not only a spatial but also a temporal variable process. Nitrate is not a conservative tracer, since its concentration can be affected by complex biogeochemical processes, in particular redox chemistry. This temporal dimension has not yet been included in the current study. The relative lack and bias of groundwater quality data have been pointed out as a major limitation for the systematic investigation of nitrate in groundwater at the continental scale. This prompts the need for consolidating and further developing groundwater monitoring programs at the continental level. Nevertheless, despite the data scarcity and bias issues, some overall conclusions could be drawn related to important groundwater pollution sources. Results have demonstrated that groundwater nitrate contamination is strongly linked to population density.

The current study is a novel application of machine learning techniques for groundwater nitrate contamination modeling at the African continent scale. Further studies are needed to reduce the uncertainty and to incorporate homogenous data to test the RFR model and increase the accuracy of this model. The analysis presented here represents an important step toward developing tools that will help to accurately predict the distribution of groundwater nitrate contamination within the context of climate change. Such conclusions could prompt the promotion of targeted local investigations by national or international authorities. The work also yields important baseline information for monitoring progress in the implementation of the United Nations Sustainable Development Goals (UN SDGs) for water.

Acknowledgments This work was funded by the IDB (Islamic Development Bank) under its Ph.D. Merit Scholarship Program (MSP). GIS shape files for generating generic attributes were obtained from different sources throughout the world and also online. In this regard, special thanks go to T. Gleeson, P. Döll, N. Moosdoorf, and P. Trambauer. I would like to thank all colleagues, particularly Mr. V. Antharam, for their valuable discussions on the random forest method. We also thank Dr. Lixiang Lin and two reviewers for their constructive comments on the initial version of the paper.

References

Abrahart RJ et al (2008) Practical hydroinformatics. computational intelligence and technological developments in water applications. Open Model Integration in Flood Forecasting 68

Aljazzar TH (2010) Adjustment of DRASTIC vulnerability index to assess groundwater vulnerability for nitrate pollution using the advection-diffusion cell. Von der Fakultät für Georessourcen und Materialtechnik der Rheinisch-Westfälischen Technischen Hochschule Aachen Ph.D. thesis, 146 pp

Alley WM, Healy RW, LaBaugh JW, Reilly TE (2002) Flow and storage in groundwater systems. Science 296(5575):1985–1990

Andrade AIASS, Stigter TY (2009) Multi-method assessment of nitrate and pesticide contamination in shallow alluvial groundwater as a function of hydrogeological setting and land use. Agric Water Manag 96(12):1751–1765

Anning DW, Paul AP, McKinney TS, Huntington JM, Bexfield LM, Thiros SA (2012) Predicted nitrate and arsenic concentrations in basin-fill aquifers of the southwestern United States. US Geological Survey Scientific Investigations Report 2012–5065

Anuraga TS, Ruiz L, Kumar MSM, Sekhar M, Leijnse A (2006) Estimating groundwater recharge using land use and soil data: a case study in South India. Agric Water Manag 84(1–2):65–76

Barzegar et al (2018) Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Sci Total Environ 621(2018):697–712. https://doi.org/10.1016/j.scitotenv.2017.11.185

Bauder J, Sinclair KN, Lund RE (1993) Physiographic and land use characteristics associated with nitrate-nitrogen in Montana groundwater. J Environ Qual 22(2):255–262. https://doi.org/10.2134/jeq1993.00472425002200020004x

BGS (2011) Depth to groundwater map. https://www.bgs.ac.uk/downloads/browse.cfm?sec=9&cat=38. Accessed 19 April 2014

Bonsor HC, MacDonald AM (2011) An initial estimate of depth to groundwater across Africa. British Geological Survey Open Report OR/11/067: 26pp

Boy-Roura M, Nolan BT, Menció A, Mas-Pla J (2013) Regression model for aquifer vulnerability assessment of nitrate pollution in the Osona region (NE Spain). J Hydrol 505:150–162

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001a) Random forests. Mach Learn 45:5–32

Breiman L (2001b) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16(3):199–231

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Boca Raton

Burow KR, Nolan BT, Rupert MG, Dubrovsky NM (2010) Nitrate in groundwater of the United States, 1991−2003. Environ Sci Technol 44(13):4988–4997

Cameron KC, Di HJ, Moir JL (2013) Nitrogen losses from the soil/plant system: a review. Ann Appl Biol 162(2):145–173

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson JC, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792. https://doi.org/10.1890/07-0539.1

Davis DB, Sylvester-Bradley R (1995) The contribution of fertiliser nitrogen to leachable nitrogen in the UK: a review. J Sci Food Agric 68:399–406. https://doi.org/10.1002/jsfa.2740680402

Debernardi L, De-Luca DA, Lasahna M (2007) Correlation between nitrate concentration in groundwater and parameters affecting aquifer intrinsic vulnerability. Environ Geol 55:539–558

Defourny P, Kirches G, Brockmann C, Boettcher M, Peters M, Bontemps S, et al (2014) Land cover CCI product user guide version 2. 2014

Döll P, Fiedler K (2008) Global-scale modeling of groundwater recharge. Hydrol Earth Syst Sci 12:863–885. https://doi.org/10.5194/hess-12-863-2008,2008

Dubrovsky NM, Burow KR, Clark GM, Gronberg JM, Hamilton PA, Hitt KJ, Mueller DK, Munn MD, Nolan BT, Puckett LJ, Rupert MG, Short TM, Spahr NE, Sprague LA, Wilber WG (2010) The quality of our nation's waters—nutrients in the nation's streams and groundwater, 1992–2004. US Geological Survey Circular 1350, 174 pp

ESRI (1969) ArcGIS, www.arcgis.com/home. Accessed 23 June 2015

Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15(1):3133–3181

Foster S, Pulido-Bosch A, Vallejos Á, Molina L, Llop A, MacDonald AM (2018) Impact of irrigated agriculture on groundwater-recharge salinity: a major sustainability concern in semi-arid regions. Hydrogeol J. https://doi.org/10.1007/s10040-018-1830-2

Fram MS, Belitz K (2011) Probability of detecting perchlorate under natural conditions in deep groundwater in California and the southwestern United States. Environ Sci Technol 45(4):1271–1277

Friedl MA, Brodley CE, Strahler AH (1999) Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE Trans Geosci Remote Sens 37(2 II):969–977

Gassiat C, Gleeson T, Luijendijk E (2013) The location of old groundwater in hydrogeologic basins and layered aquifer systems. Geophys Res Lett 40(12):3042–3047. https://doi.org/10.1002/grl.50599

Gemitzi A, Petalas C, Pisinaras V, Tsihrintzis VA (2009) Spatial prediction of nitrate pollution in groundwaters using neural networks and GIS: an application to south Rhodope aquifer (Thrace, Greece). Hydrol Process 23(3):372–383. https://doi.org/10.1002/hyp.7143

Genuer R, Poggi JM, Christine TM (2010) Variable selection using random forests. Pattern Recogn Lett 31(14):2225–2236

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recogn Lett 27(4):294–300

Gleeson T, Moosdorf N, Hartmann J, van Beek LPH (2014) A glimpse beneath earth's surface: global HYdrogeology MaPS (GLHYMPS) of permeability and porosity. Geophys Res Lett 41(11):3891–3898. https://doi.org/10.1002/2014GL059856

Golkarian A, Naghibi SA, Kalantar B, Pradhan B (2018) Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. Environ Monit Assess 190(3):149. https://doi.org/10.1007/s10661-018-6507-8

Greene EA, LaMotte AE, Cullinan KA (2005) Ground-water vulnerability to nitrate contamination at multiple thresholds in the Mid-Atlantic region using spatial probability models. US Geological Survey Scientific Investigations Report 2004–5118, p 24

Graham MH (2003) Confronting multicollinearity in ecological multiple regression. Ecology 84(11) pp. 2809–2815. https://www.jstor.org/stable/3449952. Accessed 3 Feb 2016

Grömping U (2009) Variable importance assessment in regression: linear regression versus random forest. Am Stat 63(4):308–319. https://doi.org/10.1198/tast.2009.08199

Gurdak JJ, Qi SL (2012) Vulnerability of recently recharged groundwater in principle aquifers of the United States to nitrate contamination. Environ Sci Technol 46(11):6004–6012

Hansen LK, Salamon P (1990) Neural network ensembles. IEEE Trans Pattern Anal Mach Intell (10):993–1001

Hanson CR (2002) Nitrate concentrations in Canterbury ground water – a review of existing data. Report no. R02/17. Environment Canterbury Technical Report, 87 pp

Hao A, Zhang Y, Zhang E, Li Z, Yu J, Wang H, Yang J, Wang Y (2018) Review: groundwater resources and related environmental issues in China. Hydrogeol J. https://doi.org/10.1007/s10040-018-1787-1

Hartmann J, Moosdorf N (2012) The new global lithological map database GLiM: a representation of rock properties at the earth surface. Geochem Geophys Geosyst 13:Q12004. https://doi.org/10.1029/2012GC004370

Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning, 2nd edn. Springer

Hengl T, Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, Ribeiro E, Samuel-Rosa A, Kempen B, Leenaars JGB, Walsh MG, Gonzalez MR (2014) Soil-Grids1km – global soil information based on automated mapping. PLoS One 9:e105992. https://doi.org/10.1371/journal.pone.0105992

Hoyos ICP, Krakauer N, Khanbilvardi R (2015) Random forest for identification and characterization of groundwater dependent ecosystems. WIT Trans Ecol Environ 196:89–100

ISRIC (2014) SoilGrids – Global gridded soil information. (https://www.isric.org/explore/soilgrids, Accessed 19 July 2014). [Reference to paper: Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, et al. (2014) SoilGrids1km — global soil information based on automated mapping. PLoS ONE 9(8):e105992. https://doi.org/10.1371/journal.pone.0105992]

Jung Y-Y, Dong-Chan K, Won-Bae P, Kyoochul H (2015) Evaluation of multiple regression models using spatial variables to predict nitrate concentrations in volcanic aquifers. Hydrol Process 30(5):663–675

Kazemi G, Lehr J, Perrochet P (2006) Groundwater age. Wiley-Interscience, Hoboken, New Jersey. 325pp

Khalil A, Almasri MN, McKee M, Kaluarachchi JJ (2005) Applicability of statistical learning algorithms in groundwater quality modeling. Water Resour Res 41(5)

Kihumba AM, Longo JN, Vanclooster M (2015) Modelling nitrate pollution pressure using a multivariate statistical approach: the case of Kinshasa groundwater body. Democratic Republic of Congo. Hydrogeol J: 1–13. https://doi.org/10.1007/s10040-015-1337-z

Kulabako N, 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(1):180–199. https://doi.org/10.1016/j.scitotenv.2007.03.035

Lapworth DJ, Nkhuwa DCW, Okotto-Okotto J, Pedley S, Stuart ME, Tijani MN, Wright J (2017) Urban groundwater quality in sub-Saharan Africa: current status and implications for water security and public health. Hydrogeol J 25(4):1093–1116. https://doi.org/10.1007/s10040-016-1516-6

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22

Liu CW, Wang Y-B, Jang C-S (2013) Probability-based nitrate contamination map of groundwater in Kinmen. Environ Monit Assess 185(12):10147–10156

Loosvelt L, Petersb J, Skriverc H, Lievensa H, Van Coillied FMB, De Baetsb B, Verhoesta NEC (2012) Random forests as a tool for estimating uncertainty at pixel-level in SAR image classification. Int J Appl Earth Obs Geoinf 19:173–184

Luo Y, Qiao X, Song J, Christie P, Wong M (2003) Use of a multi-layer column device for study on leachability of nitrate in sludge-amended soils. Chemosphere 52:1483–1488

MacDonald AM, Calow RC, MacDonald DM, Darling WG, Dochartaigh BÉÓ (2009) What impact will climate change have on rural groundwater supplies in Africa. Hydrol Sci J 64(690–703). 18pp

MacDonald AM, Taylor RG, Bonsor HC (2013) Groundwater in Africa – is there sufficient water to support the intensification of agriculture from "Land Grabs"? Hand book of land and water grabs in Africa, 9pp

Mair A, El-Kadi AI (2013) Logistic regression modeling to assess groundwater vulnerability to contamination in Hawaii, USA. J Contam Hydrol 153:1–23

Margat J (2010) Ressources et utilisation des eaux souterraines en Afrique. Managing Shared Aquifer Resources in Africa, Third International Conférence Tripoli 25–27 may 2008. International Hydrological Programme, Division of Water Sciences, IHP-VII Series on groundwater No.1, UNESCO, p 26–34

Masterson, JP, Hess KM, Walter DA, LeBlanc DR (2002) Simulated changes in the sources of ground water for public-supply wells, ponds, streams, and coastal areas on Western Cape Cod, Massachusetts. US Geological Survey Water Resources Investigations Report 02–4143

Mattern S, Vanclooster M (2009) Estimating travel time of recharge water through the unsaturated zone using transfer function model. Environ Fluid Mech. https://doi.org/10.1007/s10652-009-9148-1

Mattern S, Raouafi W, Bogaert P, Fasbender D, Vanclooster M (2012) Bayesian data fusion (BDF) of monitoring data with a statistical groundwater contamination model to map groundwater quality at the regional scale. J Water Resour Prot 4(11):929–943

Mendes MP, Rodriguez-Galiano V, Luque-Espinar JA, Ribeiro L, Chica-Olmo M (2016) Applying random forest to assess the vulnerability of groundwater to pollution by nitrates. geoENV 2016. The 11th International Conference onGeostatistics for Environmental Applications. Lisbon, Portugal. geoENV2016BookofAbstractsMPM

Moreno R, Zamora R, Molina JR, Vasquez A, Herrera MÁ (2011) Predictive modeling of microhabitats for endemic birds in south Chilean temperate forests using maximum entropy (Maxent). Eco Inform 6(6):364–370

Murtaugh PA (2009) Performance of several variable-selection methods applied to real ecological data. Ecol Lett 12(10):1061–1068

Naghibi SA, Ahmadi K, Daneshi A (2017) Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour Manag 31(9):2761–2775. https://doi.org/10.1007/s11269-017-1660-3

Nelson A (2004) Population Density for Africa in 2000, 4th edn. Retrieved 1/27/2011 from UNEP/GRID Sioux Falls. https://databasin.org/datasets/4d59b959e8b040688037d2fe83a3f369. Accessed 19 April 2015

Nolan BT, Hitt KJ (2006) Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. Environ Sci Technol 40(24):7834–7840. https://doi.org/10.1021/es060911u

Nolan BT, Hitt KJ, Ruddy BC (2002) Probability of nitrate contamination of recently recharged groundwaters in the conterminous United States. Environ Sci Technol 36(10):2138–2145. https://doi.org/10.1021/es0113854

Nolan BT, Fienen MN, Lorenz DL (2015) A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. J Hydrol 531:902–911. https://doi.org/10.1016/j.jhydrol.2015.10.025

Nolan BT, Gronberg JM, Faunt CC, Eberts SM, Belitz K (2014) Modeling nitrate at domestic and public-supply well depths in the Central Valley, California. Environ Sci Technol 48(10):5643–5651. https://doi.org/10.1021/es405452q.

Norouz H, Negar AM, Attaallah N (2016) Determining vulnerable areas of Malekan Plain aquifer for nitrate, using random forest method. Journal of Environmental Studies, vol 41, no 4 (76), pp 923–942. http://www.sid.ir/En/Journal/ViewPaper.aspx?ID=550917. Accessed online 2 August 2018

Oliveira S, Oehler F, San-Miguel-Ayanz J, Camia A, Pereira JMC (2012) Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. For Ecol Manag 275:117–129

Oppel S, Meirinho A, Ramírez I, Gardner B, O'Connell AF, Miller PI, Louzao, M (2012) Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biol Conserv 156:94–104. https://doi.org/10.1016/j.biocon.2011.11.013

Ouedraogo I, Vanclooster M (2016a). A meta-analysis and statistical modelling of nitrates in groundwater at the African scale. In: Hydrology and Earth System Sciences 20(6):2353–2381

Ouedraogo I, Vanclooster M (2016b) Shallow groundwater poses pollution problem for Africa. SciDev.Net, 4 pp, http://hdl.handle.net/2078.1/169630

Ouedraogo I, Defourny P, Vanclooster M (2016) Mapping the groundwater vulnerability for pollution at the pan-African scale. In: Science of the Total Environment, 544:939–953. https://doi.org/10.1016/j.scitotenv.2015.11.135

Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217–222

Park N-W (2014) Using maximum entropy modeling for landslide susceptibility mapping with multiple geoenvironmental data sets. Environ Earth Sci 73(3):937–949

Pearson S (2015) Identifying Groundwater Vulnerability from Nitrate Contamination: Comparison of the DRASTIC model and Environment Canterbury's method. Degree of Master of Applied Science (Environmental Management). Lincoln University. 58 pp

Peters J, Baets BD, Verhoest NEC, Samson R, Degroeve S, Becker PD, Huybrechts W (2007) Random forests as a tool for ecohydrological distribution modelling. Ecol Model 207(2–4):304–318

Potter P, Ramankutty N, Bennett EM, Donner SD (2010) Characterizing the spatial patterns of global fertilizer application and manure production. Earth Interact 14:1–22. https://doi.org/10.1175/2009EI288.1

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2):181–199. https://doi.org/10.1007/s10021-005-0054-1

Puckett LJ, Tesoriero AJ, Dubrovsky NM (2011) Nitrogen contamination of surficial aquifers–a growing legacy. Environ Sci Technol 45(3):839–844. https://doi.org/10.1021/es1038358

R Development Core Team (2015) A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.r-project.org/. Last accessed 6 March 2015)

Ramasamy N, Krishnan P, Bernard JC, Ritter WF(2003) Modeling Nitrate Concentration in Ground Water Using Regression and Neural Networks. Department of Food and Resource Economics. College of Agriculture and Natural Resources. University of Delaware(ORES SP03–01). 10pp

Rankinen K, Salo T, Granlund K, Rita H (2007) Simulated nitrogen leaching, nitrogen mass field balances and their correlation on four farms in South-Western Finland during the period 2000–2005. Agric Food Sci 16:387–406

Ransom et al (2017). A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. https://doi.org/10.1016/j.scitotenv.2017.05.192

Rawlings JO, Pantula SG, Dickey DA (1998) Applied regression analysis, a research tool. Springer, Berlin. 658p

Ritter A, Muñoz-Carpena R (2013) Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J Hydrol 480:33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004

Rodriguez-Galiano VF, Chica-Rivas M (2012) Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and digital terrain models. Int J Digital Earth 7(6):492–509

Rodriguez-Galiano VF, Chica-Olmo M, Abarca-Hernandez F, Atkinson PM, Jeganathan C (2012a) Random forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sens Environ 121:93–107

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012b) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 67:93–104

Rodriguez-Galiano V, Mendes MP, Garcia-Soldado MJ, Chica-Olmo M, Ribeiro L (2014) Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (southern Spain). Sci Total Environ 476-477:189–206. https://doi.org/10.1016/j.scitotenv.2014.01.001

Saffigna PG, Keeney DR (1997) Nitrate and chloride in groundwater under irrigated agriculture in Central Wisconsin. Groundwater 15(2):170–177

Sahoo S, Russo TA, Elliott J, Foster I (2017) Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. Water Resour Res 53:3878–3895. https://doi.org/10.1002/2016WR019933

Sajedi-Hosseini F, Malekian A, Choubin B, Rahmati O, Cipullo S, Coulon F, Pradhan B (2018) A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci Total Environ 644(2018):954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054

Schweigert P, Pinter N, van der Ploeg R (2004) Regression analyses of weather effects on the annual concentrations of nitrate in soil and groundwater. J Plant Nutr Soil Sci 167(3):309–318

Sesnie SE, Gessler PE, Finegan B, Thessler S (2008) Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. Remote Sens Environ 112(5):2145–2159

Sieling K, Kage H (2006) N balance as an indicator of N leaching in an oilseed rape – winter wheat – winter barley rotation. Agric Ecosyst Environ 115:261–269

Sophocleous M (2004) Groundwater recharge. In: Silveira L, Wohnlich S, Usunoff EL (eds), Groundwater. Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford, UK. http://www.eolss.net. Accessed 9 September 2015

Spalding RF, Exner ME (1993) Occurrence of nitrate in groundwater- a review. J Environ Qual 22:392–402. https://doi.org/10.2134/jeq1993.00472425002200030002x

Steele BM (2000) Combining multiple classifiers: an application using spatial and remotely sensed information for land cover type mapping. Remote Sens Environ 74(3):545–556

Stevenson FJ, Cole MA (1999) Cycles of soil carbon, nitrogen, phosphorus, sulfur, micronutrients, 2nd edn. Wiley, Hoboken

Stigter TY, Ribeiro L, Dill AMMC (2008) Building factorial regression models to explain and predict nitrate concentrations in groundwater under agricultural land. J Hydrol 357(1–2):42–56

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources, and a solution. BMC Bioinf 8:25. https://doi.org/10.1186/1471-2105-8-25

Teng Y, Hu B, Zheng J, Wang J, Zhai Y, Zhu C (2018) Water quality responses to the interaction between surface water and groundwater along the Songhua River, NE China. Hydrogeol J. https://doi.org/10.1007/s10040-018-1738-x

Tesoriero AJ, Voss FD (1997) Predicting the probability of elevated nitrate concentrations in the Puget Sound-Basin, implications for aquifer susceptibility and vulnerability. Ground Water 35(6):1029–1039

Thayalakumaran T, Charlesworth PB, Bristow K, van Bemmelen RJ, & Jaffres J (2004) Nitrate and ferrous iron concentrations in the lower Burdekin aquifers: assessing denitrification potential. In B. Singh (Ed), SuperSoil 2004 Conference 3rd Australian New Zealand Soils Conference (pp. 1-9). Sydney: The Regional Institute Ltd. https://researchoutput.csu.edu.au/en/publications/nitrate-and-ferrous-iron-concentrations-in-the-lower-burdekin-aqu, https://www.researchgate.net/publication/228513222_Nitrate_and_ferrous_iron_concentrations_in_the_lower_Burdekin_aquifers_assessing_denitrification_potenti. Accessed 17 Feb 2016

Trambauer P, Dutra E, Maskey S, Werner M, Pappenberger F, van Beek LPH, Uhlenbrook S (2014) Comparison of different evaporation estimates over the African continent. Hydrol Earth Syst Sci 18(1):193–212

UNECA, AU, AfDB (2000) The Africa Water Vision 2025: Equitable and Sustainable Use of Water for Socioeconomic Development. http://www.afdb.org/fileadmin/uploads/afdb/Documents/Generic-Documents/african%20water%20vision%202025%20to%20be%20sent%20to%20wwf5.pdf. Accessed 11 February 2016

UNEP (1986) Final Report: UNEP/FAO World and Africa GIS Data Base; December 1984. http://www.grid.unep.ch/data/summary.php?dataid=GNV38&category=atmosphere&dataurl=http://www.

grid.unep.ch/data/download/gnv038.zip&browsen=http://www.grid.unep.ch/data/download/gnv038.gif. Accessed 17 June 2015

UNEP/DEWA (2014) Sanitation and Groundwater Protection – a UNEP Perspective. http://www.bgr.bund.de/EN/Themen/Wasser/Veranstaltungen/symp_sanitat-gwprotect/present_mmayi_pdf.pdf?__blob=publicationFile&v=2. Accessed 14 August 2014

Ward MH, deKok TM, Levallois P, Brender J, Gulis G, Nolan BT, VanDerslice J (2005) Workgroup report: drinking-water nitrate and health—recent findings and research needs. Environ Health Perspect 113(11):1607–1614. https://doi.org/10.1289/ehp.8043

Wheeler DC, Nolan BT, Flory AR, DellaValle CT, Ward MH (2015) Modeling groundwater nitrate concentrations in private wells in Iowa. Sci Total Environ 536:481–488. https://doi.org/10.1016/j.scitotenv.2015.07.080

Wick K, Heumesser C, Schmid E (2012) Groundwater nitrate contamination: factors and indicators. J Environ Manag 111:178–186

Xu Y, Usher B (2006) Groundwater pollution in Africa. Taylor & Francis/Balkema, the Netherlands, 353 pp

Yost AC et al (2008) Predictive modeling and mapping sage grouse (Centrocercus urophasianus) nesting habitat using maximum entropy and a long-term dataset from southern Oregon. Eco Inform 3(6):375–386

Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM (2015) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides 13(5):839–856