



Accidental infrastructure for groundwater monitoring in Africa



Farah E. Colchester^a, Heloise G. Marais^a, Patrick Thomson^b, Robert Hope^b,
David A. Clifton^{a,*}

^a Department of Engineering Science, University of Oxford, UK

^b School of Geography and the Environment, University of Oxford, UK

ARTICLE INFO

Article history:

Received 21 October 2016

Received in revised form

16 January 2017

Accepted 27 January 2017

ABSTRACT

A data deficit in shallow groundwater monitoring in Africa exists despite one million handpumps being used by 200 million people every day. Recent advances with “smart handpumps” have provided accelerometry data sent automatically by SMS from transmitters inserted in handles to estimate hourly water usage. Exploiting the high-frequency “noise” in handpump accelerometry data, we model high-rate wave forms using robust machine learning techniques sensitive to the subtle interaction between pumping action and groundwater depth. We compare three methods for representing accelerometry data (wavelets, splines, Gaussian processes) with two systems for estimating groundwater depth (support vector regression, Gaussian process regression), and apply three systems to evaluate the results (held-out periods, held-out recordings, balanced datasets). Results indicate that the method using splines and support vector regression provides the lowest overall errors. We discuss further testing and the potential of using Africa's accidental infrastructure to harmonise groundwater monitoring systems with rural water-security goals.

© 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Timely and cost-effective groundwater monitoring is a global challenge in both industrialised and developing countries (Gorelick and Zheng, 2015; Giordano, 2009; Shah, 2010; Llamas and Martínez-Santos, 2005; Nelson, 2012; Foster and Garduño, 2013). Increasing urgency for policy action is driven by global groundwater depletion rates doubling between 1960–2000 and 2000–2009 from 56 km³ per year to 113 km³ per year (Doell et al., 2014). However, global groundwater data vary in extent and quality (Giordano, 2009; Wada et al., 2010; Mulligan et al., 2014). Africa is the most data-poor region with limited records (<0.001%) of global shallow groundwater records (Fan et al., 2013). Though high-yielding groundwater sites (<5 l s⁻¹) are limited and unevenly-distributed, groundwater is a strategic resource for Africa's growth and development, with groundwater storage estimated to be over 100 times greater than annual renewable freshwater sources (MacDonald et al., 2012). Africa's systematic data deficit in

shallow groundwater monitoring is juxtaposed by rapid and often competing demands from domestic, industrial, and agricultural sectors with regulatory and enforcement systems either weak or absent. An unpredictable future climate will place new pressures on managing and allocating groundwater, thereby increasing the need for high-quality, low-cost shallow groundwater data in a distributed monitoring system.

Africa's shallow groundwater systems (<80 m depth) supply domestic water for around 200 million rural Africans lifted by one million handpumps distributed across rural areas (Foster and Garduño, 2013). Handpumps emerged as a low-cost, durable technology in the 1980s to supply drinking water to rural communities (Hope, 2015). Shallow groundwater accessed by handpumps operating throughout the year provides generally good quality water to buffer dry periods. With ongoing challenges of repairing broken handpumps in remote rural areas, a transmitter was designed, tested, and successfully deployed in handpump handles to automatically send data on pump usage via the GSM network. Volumetric abstraction is calculated from accelerometry data generated by the movement of the pump handle (Thomson et al., 2012). Since 2012, these data have provided information on hourly pump use, and have allowed local mechanics to be alerted when failure events occur, thus reducing the down-time following such events from over a month to several days (University of

* Corresponding author. Institute of Biomedical Engineering, Department of Engineering Science, Old Road Campus Research Building, University of Oxford, Headington, Oxford, OX3 7DQ, UK.

E-mail address: david.clifton@eng.ox.ac.uk (D.A. Clifton).

Oxford/RFL, 2015). Methods using pressure sensors and water detection have also been developed to monitor pump usage remotely, similarly aimed at reducing pump downtimes (Nagel et al., 2015).

Further analysis of the accelerometry data revealed that elements of the high-frequency components appeared to correspond to groundwater depth. This study provides proof-of-concept analysis of novel methods based on machine learning to predict aquifer depth from the high-frequency signal. The implications of the findings present a potentially scalable approach to address Africa's groundwater monitoring deficit by harnessing handpumps as accidental infrastructure (Frischmann, 2012) to improve groundwater resource management; regulate and monitor irrigation, mining, or other commercial groundwater users; and provide early-warning systems for vulnerable populations dependent on shallow groundwater resources.

In this article, we model the high-rate waveforms from the accelerometry data using robust machine learning techniques that are sensitive to the subtle interaction between the dynamics of the handpump and the depth of the aquifer beneath the pump. We compare the ability of various candidate machine-learning models for the purposes of estimating aquifer depth.

2. Materials and methods

2.1. Study site and data description

The work described considers two datasets of accelerometry recordings, collected from two different models of handpumps: the Afridev and the India MK II. The first set of recordings, referred to as the “Oxford” dataset, was collected from an India MK II handpump installed at the University of Oxford, UK, between April and November, 2014. The second set of recordings, referred to as the “Kenya” dataset was collected from 11 Afridev handpumps installed in Kwale County, located between Mombasa and Tanzania's northern border, over a two-week period in April, 2014.

Each dataset consists of recordings taken at the pump location. To obtain recordings during our experiments, a consumer-grade accelerometer was mounted to the handle of each pump, and connected to a nearby laptop via a Bluetooth data connection. Each recording comprises a single person pumping for 20 s–120 s. The resulting accelerometry measurements in three orthogonal (“triaxial”) dimensions are recorded at 96 Hz. The signal recorded by the accelerometer is proportional to the force applied to the handle during the pumping motion. As the angle of the handle changes, the axis along which the acceleration is sensed changes. The lateral movement of the handle results in the presence of additional acceleration components; however, the accelerometer is mounted close to the fulcrum of the motion, and the angular velocity of the handle is low, and so these additional components are small compared to the effect of the applied force (Thomson et al., 2012).

Depth measurement at the Oxford site was performed using a manual “dipper”, which is lowered into the borehole and which sounds on contact with water. Measurements were made before each recording to the nearest 1 cm, a level of precision deemed to be appropriate with respect to the measurement error (as shown later). The depths for the Kenya dataset are estimates based on the known depth of the pumps rods. The volume of water abstracted in Oxford was very low, being tens of litres using a pump that is capable of pumping over a thousand litres per hour. Combined with the properties of the shallow aquifer in Oxford, this level of pumping would have no impact on water level, meaning that aquifer level could be viewed as being constant over the period of each recording. This is addressed further in section 5.

Of the three measurement dimensions recorded by the device, we use the dimension perpendicular to the pump handle, associated with the waveform of the largest amplitude for our analysis. Intervals of 5 s of accelerometry data collected using this method are shown in Fig. 1.

These time-series accelerometry signals show the fundamental pumping motion, similar to the motion of the handle. The increasing parts of each waveform in Fig. 1 correspond to the handle being pushed downwards to lift water and the decreasing parts to the handle lifting to reset the pump. We refer to each of these cycles as being a *period*.

The example waveforms in the figure also show the noise present in the data, which is mostly caused by the mechanical vibration in the pump due to the motion of the handle. The figure shows that this noise is of larger amplitude on the increasing part each period than on the decreasing part; this effect is anticipated, because the increasing parts of each period correspond to mechanical loading of the handpump, while the weight of the water is being lifted, as described above. The examples in the figure have different levels of noise; it may be seen that the India MK II pump at Oxford (shown in the lowermost plot in the figure) has the noise with the highest amplitude – the handle rubs against the body of the pump which causes substantial vibration levels when water is being lifted. Infrequent use of this pump (because it is a prototype in the university setting) means that the pump has not yet reached a dynamic equilibrium by being “worn in”.

We separate the time-series accelerometry data at the troughs (after smoothing using a low-pass filter) to divide the recordings into individual periods. Thus, each recording for each pump yields a series of periods of accelerometry data. The latter are generally between 0.8 s and 1.2 s in length, and hence contain approximately 80 and 120 data points (sampled at 96 Hz).

2.2. Representing each period of accelerometry data

The next step of our analysis aims to reduce the high-rate (96 Hz) time-series accelerometry data contained within each period into quantities that capture their dynamical characteristics in a parsimonious manner, suitable for modelling. We consider two main characteristics for each period: the shape (representing the pumping movement) and the high-frequency vibrations in the handle during the movement.

We chose to investigate three methods for representing each period. For each method, we summarise each period of waveform data from each recording using (i) a *feature vector* representing the shape of that period and (ii) a feature vector representing the vibration levels observed during that period. These feature vectors are sets of scalar variables (defined below), and labelled **s** and **v**, for shape and vibration, respectively.

2.2.1. Representation method I: wavelets

The wavelet transform (Torrence and Compo, 1998) provides information about the magnitude of different frequency components present in a time-series, and how these change over time. The wavelet transform should reveal both the underlying shape of the waveform corresponding to the gross pumping motion (which corresponds to relatively low frequencies in the signal), in addition to components describing the vibration (which correspond to relatively high frequencies in the signal).

Fig. 2 shows the wavelet transform applied to an example waveform of accelerometry data. Fig. 2a shows the original 96 Hz waveform Fig. 2b shows its wavelet transform. The figure shows a time vs. frequency plot for this wavelet transform, which is a heatmap corresponding to the strength of frequency content of the signal, through all points in time. (Higher frequencies correspond to

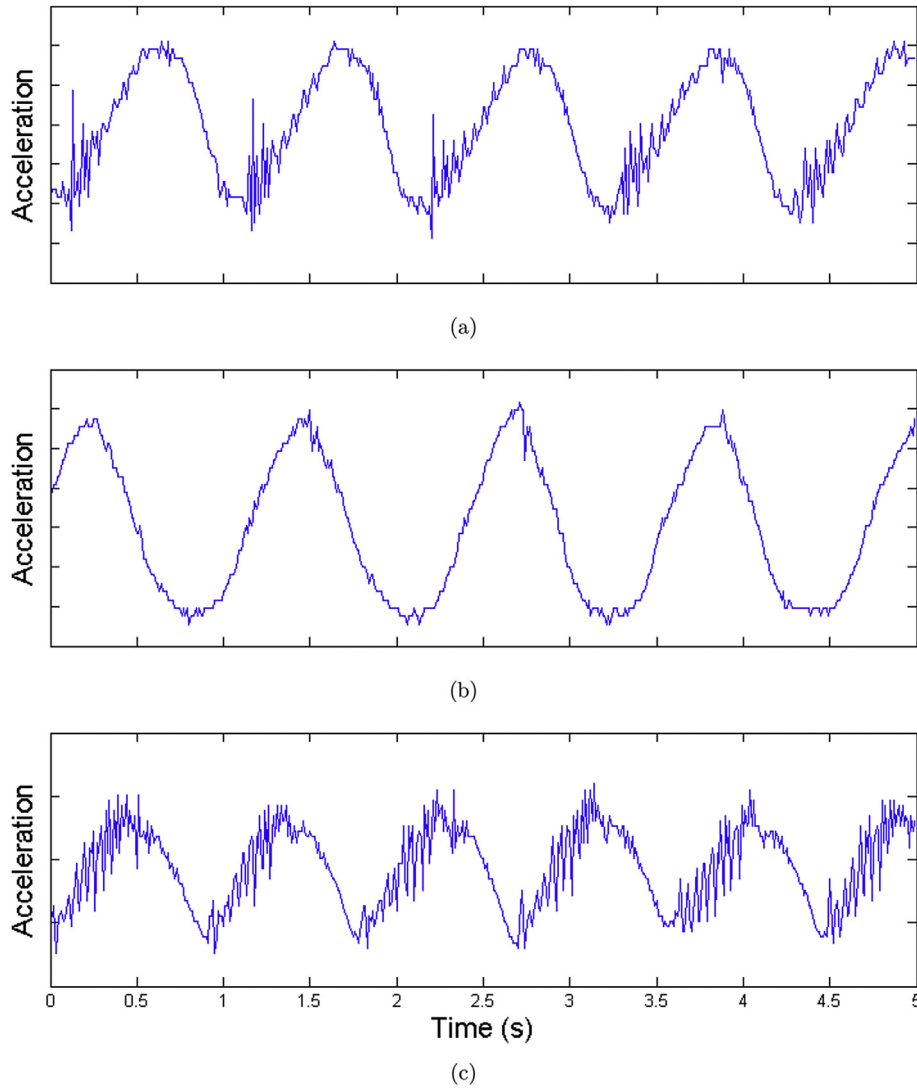


Fig. 1. Intervals of 5 s, showing accelerometry data from (a) 7 m and (b) 40 m boreholes in Kenya, and (c) the Oxford pump. Vertical axes are in arbitrary units, proportional to force exerted on the accelerometry sensor.

larger indices on the vertical axis in the figure.)

It can be seen that the periodic motion of the original waveform corresponds to high amplitudes in the wavelet transform (regions shown in red). The large amplitude of these frequencies in the wavelet transform corresponds to the fact that the dominant signal in the original waveform is due to the lower-frequency movements of the handle caused by the pumping motion.

To represent each period in the waveform, we use 12 equally-spaced frequencies between 4 Hz and 12 Hz, at p time intervals to create the feature vector, \mathbf{s} , defining the shape of the waveform in each period. Thus, each of the periods shown in the figure is represented by a feature vector \mathbf{s} containing $12p$ values.

We then apply a high-pass filter to the interval, to remove the “shape”, leaving the vibrations (Fig. 2c). We subsequently apply the wavelet transform to the result (Fig. 2d). We use 40 frequencies between 6 Hz and 36 Hz at p time intervals to create the feature vector, \mathbf{v} for each period, representing the vibrations in that period. Thus, each of the five periods shown in the figure is also represented by a feature vector \mathbf{v} that contains $40p$ values. The value of p is set during model training, described later.

2.2.2. Representation method II: splines

Smoothing splines (de Boor, 2001) are a solution to the general problem of finding a curve of best fit, $f(t)$, through noisy data $x(t)$, where t are the times associated with each measurement $x(t)$. A smoothing spline compromises between having small errors $\epsilon_i = |f_i - x_i|$ for the i values of the data, and having a slowly-changing curve $f(t)$ that does not “overfit” the data $x(t)$ by varying too rapidly.

We sample the fitted spline $f(t)$ at p intervals to create the feature vector \mathbf{s} representing the shape of a period. To obtain a feature vector \mathbf{v} representing the vibration for the same period, we take the mean of ϵ_i between each of the p samples, such that the feature vector \mathbf{v} for a period contains $p - 1$ values.

2.2.3. Representation method III: Gaussian processes

Gaussian processes (Rasmussen, 2006) (GPs) are a method for classification and regression which can be used for modelling time-series, with some mean function $m(t)$ and covariance function $K(t, t')$, so that the data have distribution $x(t) \sim N(m(t), \sigma)$ for some σ , and where the covariance between two points is given by

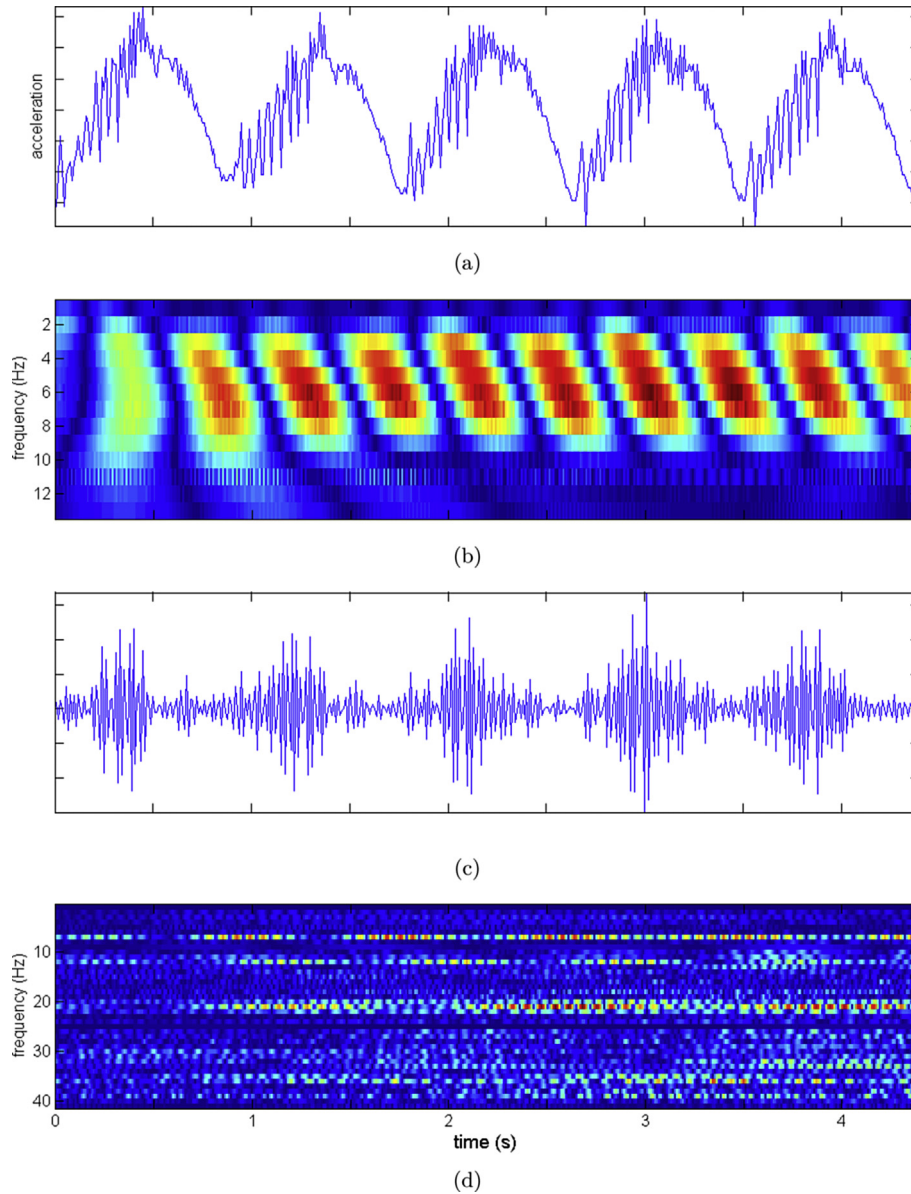


Fig. 2. (a) Accelerometry recording; (b) components of low frequency from the wavelet transform, as described in the main text; (c) Accelerometry recording after high-pass filtering; (d) components of high frequency from the wavelet transform applied to the filtered data. Wavelet transform heatmaps indicate the index of each frequency component on the vertical axis.

$$\text{cov}(x(t), x(t')) = K(t, t').$$

We expect two points closer together in time to have higher correlation, and hence higher covariance, and so for K we take the (commonly-used) squared exponential function (Rasmussen, 2006): as desired, this assigns covariance close to 1 for points that are very close in time, decreasing to covariance close to 0 for points that are far apart in time.

A standard Gaussian process assumes that the variance of the time-series is constant, which is clearly not the case for our data – the noise varies throughout a period, as described earlier. Gaussian processes for heteroscedastic data (i.e., in which the noise level is allowed to vary) have been developed (Goldberg et al., 1997; Lázaro-Gredilla and Titsias, 2011). These assume $x(t) \sim N(m(t), \sigma(t))$, where we note that $\sigma(t)$ is now a quantity that depends on time. This quantity must be positive, and so we take $\sigma(t) = \exp(g(t))$ and model $g(t)$ as a second Gaussian process. The latter assumes the distribution $g(t) \sim N(0, \sigma_g)$ and that there is a covariance function $k_g(t, t')$. We assume that noise values are

independent of one another, and so k_g is the white-noise covariance function (Rasmussen, 2006), that is, all off-diagonal elements of $k_g = 0$. Fig. 3 represents an HGP applied to accelerometry data. Fig. 3b shows $m(t)$ and the shaded area is of width $1.96\sigma(t)$ to represent the 95% confidence interval. Fig. 3c shows $g(t)$.

Gaussian processes are dependent on hyperparameters, the values of which are chosen to fit the data without overfitting (i.e., to minimise the negative log marginal likelihood, as is standard practice (Rasmussen, 2006)).

To create the shape feature vector \mathbf{s} for each period, we sample the Gaussian process $f(t)$ at p intervals. Similarly, to create the vibration feature vector \mathbf{v} for each period, we sample the Gaussian process $g(t)$ at p intervals.

2.3. Estimating aquifer levels from feature vectors

Having obtained feature vectors \mathbf{s}, \mathbf{v} for each period, using three candidate representation methods described previously, we now

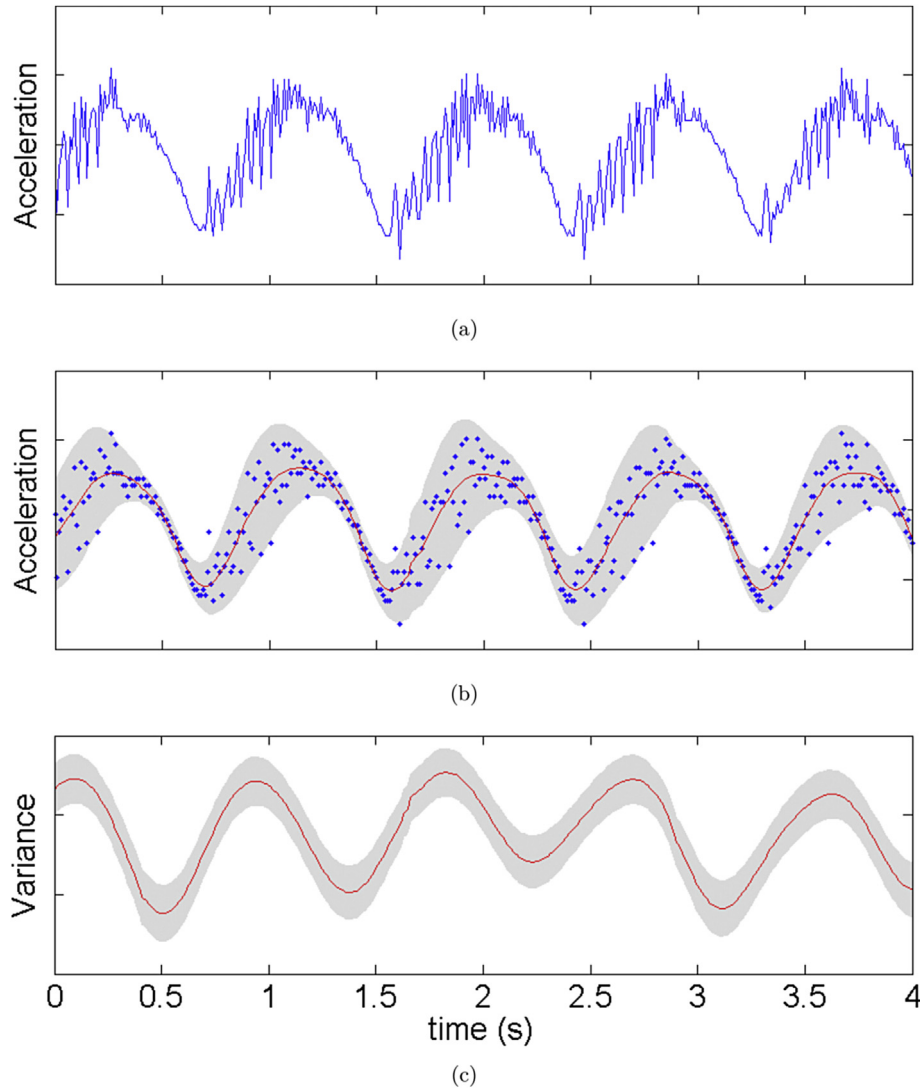


Fig. 3. Heteroscedastic GP representation of each period; (a) original accelerometry data; (b) latent function for the f GP (red) with 95% confidence interval (grey) and raw accelerometry (blue); (c) latent function for the g GP (red) with respective 95% confidence interval (grey). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

describe methods for subsequent estimation of aquifer depth. For each period i , we have feature vectors \mathbf{s}_i and \mathbf{v}_i , along with a depth measurement y_i . We aim to be able to produce models that produce estimates y_i^* which are as close as possible to actual aquifer depth y_i , given the feature vectors \mathbf{s}_i and \mathbf{v}_i . We developed two systems for depth estimation, based on (i) support vector regression and (ii) Gaussian processes, both of which are machine learning methods for modelling complex data.

2.3.1. Depth estimation system I: support vector regression, SVR

Support vector regression (SVR) (Hsu et al., 2003) aims to map feature vectors onto some target quantity y by transforming the original data from its original format (with each feature vector containing p scalar values) to some new format; in mathematical notation, the data from the p -dimensional input space are projected into some q -dimensional space ($q > p$). This transformation is provided by a kernel function, which has a small number of parameters that can be learned from pairs of input “training” data \mathbf{x} and target output values y . The parameters of the model are set so that the system makes predictions y^* given an input feature vector

\mathbf{x} , where there is minimal error between the predictions y^* and the targets y .

For the work described by this paper, we used a (commonly-chosen) Gaussian kernel function (Hsu et al., 2003); the resulting SVR subsequently has two parameters to set: one associated with the Gaussian kernel, and one pertaining to the fitting of the SVR to the data - this is the standard setting for the SVR. The training set comprises a set of feature vectors, each of which is associated with its depth measurement y . The feature vectors provided to the method are those for shape \mathbf{s} and vibrations \mathbf{v} described earlier, concatenated into a single, larger feature vector for use by the depth estimation system, $\mathbf{x} = \{\mathbf{s}, \mathbf{v}\}$.

2.3.2. Depth estimation system II: Gaussian process regression, GPR

The Gaussian process regression (GPR) framework is readily extended to map p -dimensional feature vectors onto target values y , using the same algorithms as described previously. In this case, the GPR model is constructed using “training” data comprising p -dimensional feature vectors along with their associated targets y - proceeding as that described for SVR.

2.3.3. Depth estimation system III: random forest regression, RFR

We include random forests in our analysis as a comparator method. Decision trees (Rokach and Maimon, 2014) are a classification and regression technique that uses a binary tree where each node represents a binary decision based on the value of one feature, and where each leaf is a target value. A datapoint x hence passes down branches of the tree depended on its feature values, until it reaches a target value. A random forest (Kam Ho, 1995) is an amalgamation of n decision trees, where n overlapping subsamples of data are taken and a decision tree is trained for each. Each datapoint x is then regressed using each decision tree i to give predicted value y_i^* , and the final prediction y^* is a weighted average of y_i^* , $i = 1, \dots, n$.

2.4. Implementation

All data analysis was carried out in Matlab. Wavelet analysis was performed using the Matlab signal processing toolbox. Heteroscedastic Gaussian processes were developed using the VHGP toolbox (Miguel and Michalis,). Smoothing splines were implemented using the Matlab curve-fitting toolbox. Gaussian process regression was implemented using the GPML toolbox (Rasmussen and Nickisch, 2010). Support vector regression used the libSVM toolbox (Chang and Lin, 2011). Random forest regression was based on the Matlab statistics and machine learning toolbox.

3. Training-testing methodology, and results

3.1. Training and testing

We evaluate our methods via schemes based on hold-out validation: a method for evaluating the performance of machine learning algorithms by randomly “holding out” a proportion of the available data as a *test* set, and using the remainder of the data as the *training* set. The optimal length of the feature vector p and the values of the hyperparameters for each algorithm are chosen using four-fold cross-validation. That is, the training set is divided into four quarters. For the combination or parameter values in question, the algorithm is trained on three quarters of the training data, and evaluated on the fourth in turn. The average of its performance in the four experiments is taken to give the overall performance associated with each set of values of the hyperparameters. We can thereby find the “optimal” value of the hyperparameters using grid search. This approach to evaluation avoids evaluating the performance of the model using the same data as were used to train it; this avoids “over-fitting” to the training data, and gives a more robust estimate of how the model would perform when applied to previously-unseen data.

3.2. Evaluation scheme I: held-out periods

In the first evaluation scheme, denoted “held-out periods”, we hold out 25% of individual periods, selected at random from all those available, and use the remainder as a training set.

3.3. Evaluation scheme II: held-out recordings

In the second evaluation scheme, denoted “held-out recordings”, we hold out 30% of entire recordings. The latter is a more difficult test because the models are being compared by evaluating their performance using previously-unseen entire recordings of data. In the former, a single recording might contribute some periods to the training set and some periods to the test set, making for a potentially easier challenge.

3.4. Evaluation scheme III: balanced datasets

We observe that a large proportion of our data happen to be concentrated around the average aquifer depth. To determine whether or not this biases our results, we use an additional third evaluation scheme, in which the training set is *balanced*. In this scheme, we divide all recordings into three quantiles according to aquifer depth, and take at random an equal number of periods from each of the resulting “shallow”, “intermediate”, and “deep” quantiles to form the training and test sets.

3.5. Combining methods

We have, in our description so far, defined:

- three methods of representing periods of accelerometry data using feature vectors (I: wavelets, II: splines, III: GP)
- two systems for estimating aquifer depth using feature vectors (I: SVR, II: GPR)
- three schemes for evaluating the proposed techniques (I: held-out periods, II: held-out recordings, III: balanced datasets)

All $3 \times 2 \times 3 = 18$ experiments could be performed; for clarity of description in this proof-of-principle study, we will limit the results described here to the following combinations. This selection of is based on those methods for creating feature vectors that perform well with the corresponding systems for depth estimation. For example, we combine the Bayesian probabilistic methods HGP and GPR, and the non-probabilistic methods of splines and random forests:

1. Wavelets for creating feature vectors, followed by depth estimation using GPR
2. Splines for creating feature vectors, followed by depth estimation using SVR
3. GPs for creating feature vectors, followed by depth estimation using GPR
4. Splines for creating feature vectors, followed by depth estimation using RFR

Each of these combinations will be investigated with each of the three evaluation schemes (held-out periods, held-out recordings, balanced datasets).

4. Results

Table 1 shows the median absolute errors (and interquartile range of these errors) of aquifer depth predictions y^* , for each of the three modelling approaches described above, using the “Oxford” dataset. For each of the three modelling approaches (Wavelets-GPR, Splines-SVR, and GP-GPR), we show results obtained using our three evaluation schemes. We also provide the performance of splines-RFR as a comparison.

Table 1

Errors in aquifer estimation, shown as median (IQR) in cm, for the three modelling approaches described in the text (by row), according to each of the three schemes for evaluation (by column).

	Held-out periods	Held-out recordings	Balanced
Wavelets-GPR	3.6 (1.7, 8.9)	5.9 (2.4, 15.8)	3.9 (1.2, 8.1)
Splines-SVR	2.5 (1.0, 6.4)	2.9 (1.1, 11.6)	2.1 (0.9, 6.4)
GP-GPR	3.8 (2.07, 9.65)	5.8 (2.5, 13.4)	2.4 (1.1, 5.1)
Splines-RFR	2.93 (1.14, 8.11)	5.23 (2.0, 11.48)	5.75 (2.5, 12.54)

Bold shows the best-performing case.

The results in Table 1 show that, as expected, performance for all models decreases when moving from “held-out periods” to “held-out recordings” - this is evident both in the median error for each of the three modelling approaches, and in the increased interquartile range of those errors. As described earlier, the latter corresponds to a more difficult level of evaluation, which might be described as being the most critical test of the methods.

The results in the table also demonstrate that balancing of the training set results in an improvement in performance for the Splines-SVR and GP-GVR approaches, but not for Wavelets-GPR or Splines-RFR. In most applications, balancing the training set provides a better model as removing the imbalance in the data allows the model to learn more effectively any underlying relationships

between input feature vectors and output estimates.

In all evaluation schemes, the Splines-SVR method provides the lowest overall errors and lowest associated interquartile ranges on the distributions of those errors. The SVR is very effective at learning non-linear relationships between input and output quantities in many machine learning applications. While the GPR-based methods can offer additional flexibility (such as fully probabilistic output, and the ability to cope with partially-missing input data), this often comes at the cost of lower overall accuracy when compared to less-flexible approaches that aim to maximise prediction accuracy (such as the SVR).

Examining these results in more detail, Fig. 4 shows the spread of individual depth estimates for each recording, along with their

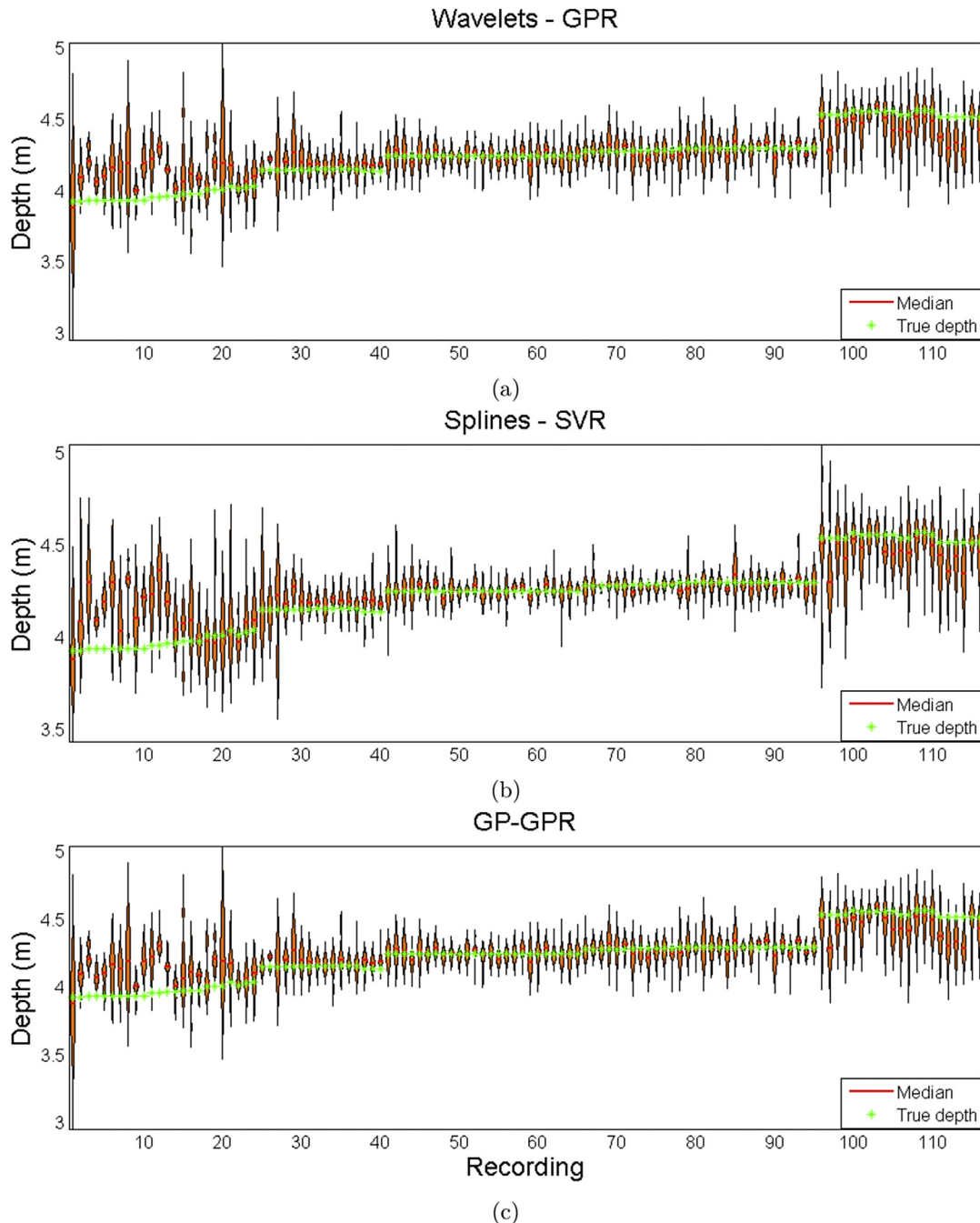


Fig. 4. Violin plots showing results on a randomly-selected test set for all three of our methods. Along the x-axis is each recording, and the “violin” shows the spread of individual depth estimates for the periods of that recording.

median, with the depth measurements target shown for comparison. It may be seen from the figure that the depth of the aquifer from this Oxford-based handpump varied over approximately 50 cm during the period of data acquisition.

Accuracy can be seen to depend on the true value of the aquifer depth. At medium depths (recordings 25–95 in the figure), the estimates for each recording are centred around the true depth with a small spread on either side. For those data from increased depths (recordings 96–116), there is a larger spread in predictions for each individual period of these recordings. Moreover, there is a bias in the predictions, such that many more predictions are below the true depth. For those data from the shallowest depths (recordings 1–24), this problem is amplified, but less consistent.

This pattern in the distribution of errors appears to be present in all three modelling approaches, but it may be seen that the Splines-SVR approach has smaller overall error around the true depth - this latter effect corresponds to the smaller error values shown in Table 1.

Fig. 5 shows the depth estimates for one modelling approach (GP-GPR) to the Kenya dataset which had associated median errors of 1.9 cm (IQR 4.1, 6.7). The figure shows that depth predictions are reasonable for data from shallowest and deepest aquifer depths. However, as with the Oxford dataset, there is a bias in the predictions such that the shallowest depths (recordings 1–37) are overestimated and the deepest ones (recordings 59–85) are underestimated. The model also fails to discriminate well between those with medium depths (38–58), possibly because we have very little data for each point of aquifer depth in that range.

4.1. Hyperparameters and sensitivity analysis

We list here (and in Table 2) the values of the hyperparameters used in each depth-prediction model. The covariance function used in our Gaussian processes is

$$\text{cov}(t, t^*) = \sigma_f^2 \exp\left(-\frac{(t - t')^2}{2l^2}\right) + \sigma_n \delta_{t=t^*}$$

Support vector regressors use the kernel function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

and soft margin parameter, C , using the standard nomenclature (Hsu et al., 2003).

We present a sensitivity analysis for Splines-SVR, by way of example, to demonstrate the variability in performance around the “optimal” value obtained from the training process described earlier. Table 3 shows the error rate in the case of the optimum hyperparameters being multiplied by a factor of $\{0.8, 0.9, 1.0, 1.1, 1.2\}$, using the held-out periods evaluation

Table 2

The length of the feature vector, p , and the values of the hyperparameters used in the depth-prediction models in each case. In the case of Gaussian process regression, the values given are for σ_f, l, σ_n respectively. For support vector regression, the values given are for γ, C respectively.

	p	Held-out periods	Held-out recordings
Wavelets-GPR	64	2.96, 1.59, -2.6	2.95, 1.59, -2.78
Splines-SVR	16	1, 50	1, 32
GP-GPR	100	2.5, 1.66, -2.4	2.41, 1.73–2.53

method. The table shows that, as expected, performance varies around the “optimal” value of the hyperparameters, but that the grid-search method is sufficient for recovering those values that yield highest performance.

5. Study limitations

A well or borehole will have draw-down and recovery as water is pumped from it, and it then settles to its quiescent, static, level. The pumping that is generating the data is itself affecting the level that it is measuring. The extent of this draw-down will be a function of two key variables: the rate of water abstraction and the hydraulic conductivity of the aquifer. These two variables must be taken into account to ensure an accurate measurement. Understanding of the underlying geology and borehole dynamics is therefore important.

Many handpumps are only used during daytime so that even where there is significant draw-down, the water level can recover overnight. In this case, a depth measurement made at the start of the day will be closest to the static level, and so taking a measurement at an appropriate time is critical. In addition, the system that is generating the depth estimate is also estimating the volume of water being pumped from the well (Thomson et al., 2012). Given knowledge of the hydraulic conductivity of the aquifer, this abstraction rate can be integrated with the estimate to reduce its error.

Factoring draw-down into the model would enable some disaggregation of locally-induced water level change and change in

Table 3

Table showing the median error on the test set when optimal value of hyperparameters C and γ are multiplied by the given factors, in the case of splines/SVR/ held-out periods.

		γ				
		0.8	0.9	1	1.1	1.2
C	0.8	2.59	2.72	2.76	2.69	2.54
	0.9	2.7	2.61	2.5	2.45	2.44
	1	2.59	2.5	2.42	2.43	2.43
	1.1	2.53	2.57	2.56	2.47	2.48
	1.2	2.48	2.44	2.44	2.43	2.44

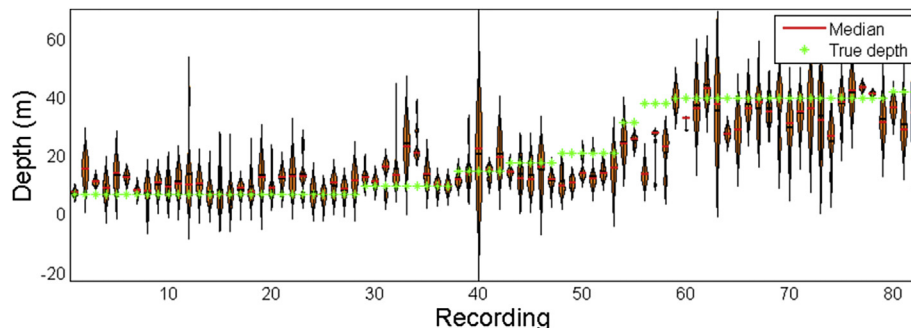


Fig. 5. Violin plots showing results of the GP-GPR model on the Kenya dataset.

gross aquifer level. In the case of a well at risk of going dry, understanding how local draw-down and wider aquifer depth interact would provide immediately-useful operational information. This could allow us to distinguish between a short-term imbalance between supply and demand (due to seasonal factors which could be managed through reducing pump usage over the period of risk), vs. whether abstraction levels are unsustainable in the longer term.

The same understanding of the surrounding aquifer will inform how multiple depth calculations from a number of pumps can be combined to provide a better understanding of the aquifer. As discussed, the water level beneath individual pumps will have reasons to be different to each other and fluctuate differently (e.g., different levels of abstraction or proximity to a river). Though not the same, these will be correlated between pumps and thus could be modelled. In contrast, most of the factors causing errors (e.g., different pump users, pump condition) will be uncorrelated. Thus combining calculations from multiple pumps could significantly reduce the error of the estimates.

6. Discussion and further work

Machine Learning is being increasingly used in environmental monitoring (Hill and Minsker, 2010; Naghibi et al., 2016). This proof-of-concept work demonstrates that in both a controlled and operational context, machine learning can use high-frequency vibration data from rural handpumps to generate useful estimates of groundwater depth. These estimates can be generated from the usual day-to-day pumping of rural water users. With a combination of on-board and centrally-managed data processing, the network of handpumps across rural Africa has the potential to be transformed into a distributed monitoring system which can provide timely information on the magnitude and direction of change in shallow groundwater levels as required by government, enterprise, communities and other stakeholders.

Daily or weekly data will provide objective and automated information to help ensure that water supplies for communities, schools, and clinics remain secure, and provide early warning of emerging problems. Monthly or seasonal data will help government and regulators manage the resource by identifying sustainable levels of abstraction that can (i) promote growth from agriculture, mining and other commercial demands while (ii) protecting the resource to fulfil the needs of domestic drinking water, livestock watering, and small-scale irrigation. Finally, spatially-distributed, longitudinal data will assist monitoring and managing climate variability through improved understanding of relationships between rainfall and recharge patterns. Further work on the same data will address questions specific to the handpumps as well as the aquifer, such as condition monitoring for pre-emptive maintenance.

Frischmann et al (Frischmann, 2012). argue that all infrastructure is accidental to some degree, with benefits being realised that are beyond those designed. Innovation can generate new ways for existing infrastructure to generate value, such as the example described by Overeem et al. (2013). Similarly, the accidental infrastructure of community handpumps, along with advances in remote sensing (Tapley et al., 2004), has the potential to complement Africas existing, but sparse, monitoring infrastructure. This will reduce the continents groundwater data-deficit, thereby helping sustainably manage Africa's abundant (but not yet fully-understood) groundwater resources, to promote growth and development.

Author information

Notes

The authors declare no competing financial interest.

Data availability

This work is linked to an ongoing NERC/ESRC/DFID funded project (UPGro Consortium Grant: *Gro for GooD* NE/M008894/1). All relevant data for this project will be deposited at the National Geoscience Data Centre at the end of the project in 2019. However, for the benefit of readers and reviewers we have published data directly relevant to this article submission on the Oxford Research Archive: <http://dx.doi.org/10.5287/bodleian:nbKjNaMj>.

Acknowledgements

This paper is an output from the “Groundwater Risk Management for Growth and Development” project (NE/M008894/1) funded by NERC/ESRC/DFID's UPGro programme; the “New Mobile Citizens and Waterpoint Sustainability in Rural Africa” (ES/JO18120/1) funded by ESRC/DFID; Oxford University's John Fell Fund; and, the “Smart Water Systems” project (R5737) under DFID's New and Emerging Technologies programme. FEC was supported by the EPSRC Doctoral Training Centre at the Life Sciences Interface, Oxford. HGM was supported by UNICEF. DAC was supported by the Royal Academy of Engineering, Balliol College, Oxford, and the EPSRC via a “Challenge Award”. PT was supported by NERC, ESRC, DFID and UNICEF.

References

- Chang, Chih-Chung, Lin, Chih-Jen, 2011. LIBSVM: a library for support vector machines. *ACM Trans. Intelligent Syst. Technol.* 2 (27), 1–27, 27, Software available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- de Boor, Carl, 2001. *A Practical Guide to Splines*. Springer.
- Doell, Petra, Mueller Schmied, Hannes, Schuh, Carina, Portmann, Felix T., Eicker, Annette, 2014. Global-scale assessment of groundwater depletion and related groundwater abstractions: combining hydrological modeling with information from well observations and grace satellites. *Water Resour. Res.* 50 (7), 5698–5720.
- Fan, Ying, Li, H., Miguez-Macho, Gonzalo, 2013. Global patterns of groundwater table depth. *Science* 339 (6122), 940–943.
- Foster, Stephen, Garduño, Héctor, 2013. Groundwater-resource governance: are governments and stakeholders responding to the challenge? *Hydrogeology J.* 1–4.
- Frischmann, Brett M., 2012. *Infrastructure: the Social Value of Shared Resources*. Oxford University Press.
- Giordano, Mark, 2009. Global groundwater? Issues and solutions. *Annu. Rev. Environ. Resour.* 34, 153–178.
- Goldberg, Paul W., Williams, Christopher K.I., Bishop, Christopher M., January 1997. Regression with input-dependent noise: a Gaussian process treatment. In: *Advances in Neural Information Processing Systems*, vol. 10. MIT Press, pp. 493–499.
- Gorelick, Steven M., Zheng, Chunmiao, 2015. Global change and the groundwater management challenge. *Water Resour. Res.* 51 (5), 3031–3051.
- Hill, David J., Minsker, Barbara S., September 2010. Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. *Environ. Model. Softw.* 25 (9), 1014–1022.
- Hope, Rob, 2015. Is community water management the community's choice? Implications for water and development policy in Africa. *Water Policy* 17 (4), 664–678.
- Hsu, Chih-Wei, Chang, Chih-Chung, Lin, Chih-Jen, 2003. *A practical Guide to Support Vector Classification*.
- Kam Ho, Tin, 1995. Random decision forests. In: *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1. IEEE, pp. 278–282.
- Lázaro-Gredilla, Miguel, Titsias, Michalis, 2011. *Variational Heteroscedastic Gaussian Process Regression*.
- Llamas, Manuel Ramón, Martínez-Santos, Pedro, 2005. Intensive groundwater use: silent revolution and potential source of social conflicts. *J. Water Resour. Plan. Manag.* 131 (5), 337–341.
- MacDonald, Alan M., Bonsor, Helen C., Dochartaigh, Brighid É.Ó., Taylor, Richard G., 2012. Quantitative maps of groundwater resources in Africa. *Environ. Res. Lett.* 7(2) (024009).
- Miguel Lázaro-Gredilla and Michalis Titsias. VHGP toolbox. <http://www.tsc.uc3m.es/~miguel/downloads.php>. [Accessed 2016].
- Mulligan, Kevin B., Brown, Casey, Yang, Yi-Chen E., Ahlfeld, David P., 2014. Assessing groundwater policy with coupled economic-groundwater hydrologic modeling. *Water Resour. Res.* 50 (3), 2257–2275.
- Nagel, Corey, Beach, Jack, Iribagiza, Chantal, Thomas, Evan A., 2015. *Evaluating*

- cellular instrumentation on rural handpumps to improve service delivery a longitudinal study in rural Rwanda. *Environ. Sci. Technol.* 49 (24), 14292–14300.
- Naghibi, Seyed A., Pourghasemi, Hamid R., Dixon, Barnali, January 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ. Monit. Assess.* 188 (1), 1–27.
- Nelson, Rebecca L., 2012. Assessing local planning to control groundwater depletion: California as a microcosm of global issues. *Water Resour. Res.* 48 (1).
- Overeem, Aart, Leijnse, Hidde, Uijlenhoet, Remko, 2013. Country-wide rainfall maps from cellular communication networks. *Proc. Natl. Acad. Sci.* 110 (8), 2741–2745.
- Rasmussen, Carl Edward, 2006. *Gaussian Processes for Machine Learning*. MIT Press.
- Rasmussen, Carl Edward, Nickisch, Hannes, December 2010. Gaussian processes for machine learning (gpml) toolbox. *J. Mach. Learn. Res.* 11, 3011–3015.
- Rokach, Lior, Maimon, Oded, 2014. *Data mining with decision trees: theory and applications*. World Sci. 1–20.
- Shah, Tushaar, 2010. *Taming the Anarchy: Groundwater Governance in South Asia*. Routledge.
- Tapley, Byron D., Bettadpur, Srinivas, Ries, John C., Thompson, Paul F., Watkins, Michael M., 2004. GRACE measurements of mass variability in the earth system. *Science* 305 (5683), 503–505.
- Thomson, Patrick, Hope, Rob, Foster, Tim, 2012. GSM-enabled remote monitoring of rural handpumps: a proof-of-concept study. *J. Hydroinformatics* 14 (4), 829–839.
- Torrence, Christopher, Compo, Gilbert P., 1998. A practical guide to wavelet analysis. *Bull. Am. Meteorological Soc.* 79 (1), 61–78.
- University of Oxford/RFL, 2015. *Financial Sustainability for Rural Water Services – Evidence from Kyuso, Kenya*. water programme, Working paper 2. Smith School of Enterprise and the Environment, Oxford University, UK. Available at: <http://www.smithschool.ox.ac.uk/research-programmes/water.php>.
- Wada, Yoshihide, van Beek, Ludovicus P.H., van Kempen, Cheryl M., Reckman, Josef W.T. M., Vasak, Slavek, Bierkens, Marc F.P., 2010. Global depletion of groundwater resources. *Geophys. Res. Lett.* 37 (20).