Predictive modelling of water losses using random forests on weather covariates

J. M. Jenkins, M. Kowalski and E. F. S. Alvarenga

ABSTRACT

Conventionally, multiple linear regression methods have been used to predict water losses (a proportion of which is real losses) some weeks or months ahead, based upon various weather parameters. This paper describes the development of an alternative method to predict water losses using random forests and compares model performance with linear regression using a case study approach from one water utility. It suggests that a random forest approach can significantly improve the ability to predict water losses based on readily available covariates. Further validation work on holdout data is recommended to ensure the model is not over-fitted to the learning set.

Key words | leakage, non-revenue water, random forest, water losses, weather

J. M. Jenkins (corresponding author)
M. Kowalski
E. F. S. Alvarenga
WRc plc,
Frankland Road, Blagrove,
Swindon SN5 8YF,
UK
E-mail: *james.jenkins@wrcplc.co.uk*

INTRODUCTION

As evidence grows for the increase in frequency of rainfall and temperature extremes (as cited in e.g. Water UK 2016), it becomes ever more important to understand the relationship between observed weather conditions and their impact on losses from water supply networks. Water losses is the generic term for the imbalance in recorded flow between water supply system input and authorised consumption, as defined by IWA. Water losses comprise both real losses (referred to here as 'leakage') and apparent losses (e.g. from water meter measurement uncertainty). This relationship can be combined with short-term weather forecasts to provide early warning of real losses (also referred to here as leakage) for operational decision-making. Longer term weather-based predictions can be fed into the decision-making process to ensure that leakage targets are attainable. This is particularly important as regulatory and public scrutiny surrounding water losses continues to rise.

UKWIR (2006, 2013) linked temperature and soil moisture deficit (SMD) to leakage due to stresses from soil movement and pipe material expansion. These factors have cumulative, long-term effects on leakage and increase the probability of catastrophic failure in vulnerable pipelines.

Previous studies (Kowalski 2006; UKWIR 2013) used multiple linear regression techniques to predict leakage based on weather parameters. The major benefit of the use of this model class is that outputs are readily interpretable for inference. However, as parametric models, when their assumptions are not appropriate, the predictive power of these linear models is reduced.

Maidment & Miaou (1986) questioned the linearity of the relationship between weather and consumer demand and, in the first decade of this century, machine learning techniques have been employed successfully in demand forecasting (Adamowski *et al.* 2012). Machine learning models, which include random forests and other non-parametric methods, have been shown to perform better than parametric models when the relationship between the predictors and response is non-linear (Herrera *et al.* 2010).

Because these methods have proven useful in demand forecasting, this paper developed a methodology to estimate water losses based on weather parameters derived from

doi: 10.2166/ws.2018.044

observed or forecast meteorological data. A case study approach using data from one UK water utility is adopted to investigate how random forests (Breiman 2001) provide improved performance over linear regression.

DATA

A water supply network was subdivided into a number of catchments and further into supply zones. Daily data from the gridded meteorological weather dataset provided by the Centre for Ecology and Hydrology (Robinson *et al.* 2015a, 2015b) were gathered for the centroid of each supply zone for the latest 6 years available (2005 to 2011). Effort to reduce water losses by water utilities (so-called active leakage control, or ALC, effort) can significantly reduce leakage. Hence we requested data from the water utility on both water losses (m^3/h) for each supply zone, and active leakage control (ALC) effort (person-hours). As the volume of ALC employed was only available at a catchment scale, modelling of water losses was undertaken at this spatial resolution. Data on water losses were available at a monthly timestep.

Exploratory analysis

Water losses were estimated according to Equation (1). From the minimum night flow (MNF), provided at a monthly timestep, the estimated or logged night use of both household and non-household properties is subtracted.

$$\text{Water losses} = N - U_c - n_p U_p \quad (1)$$

where:

N is the MNF;
U_c is the logged commercial night use;
n_p is the number of properties; and
U_p is the estimated night use per property.

The following daily meteorological parameters were obtained for the period 2005–2011: minimum and maximum temperature (°C), precipitation ($kgm^{-2}s^{-1}$), and potential evapotranspiration (mm/day). The rolling 30-day mean of water losses was modelled against the summarised weather for the same period: the mean value of water losses on day t was based on the weather on day t, and the 29 previous days.

SMD was calculated from the precipitation and potential evapotranspiration as described by Schulte *et al.* (2005). Evapotranspiration was calculated from potential evapotranspiration (Schulte *et al.* 2005). SMD was included due to prior research (UKWIR 2006, 2013) showing a relationship between SMD and leakage.

Temperature can physically affect leakage in a number of different ways, some of which are best captured by a small number of derived variables, e.g.:

- *Number of air frost days* provides an indication of the number of days in which the ground may have undergone a freeze-thaw cycle.
- A *run length* variable provides information on extended periods of freezing.
- *Diurnal temperature difference* highlights larger temperature swings that would enhance ground movement in prone soil types.

These variables were selected in consultation with leakage experts based on data that were, and would continue to be, readily available for ongoing model use.

From the exploratory analysis phase, a set of candidate predictor variables was devised:

- $AF_{i,t}$: count of air frost days in the 30-day window ending on day t, in catchment i.
- $RunLen_{i,t}$: longest contiguous period (days) of air frost days in the 30-day window ending on day t, in catchment i.
- $Swing_{i,t}$: mean daily temperature difference (C°) in the 30-day window ending on day t, in catchment i.
- $SMD_{i,t}$: mean SMD (mm) in the 30-day window ending on day t, in catchment i.
- $ALC_{i,t}$: mean ALC effort (hours) in the 30-day window ending on day t, in catchment i.

Plotting these key variables as a time series (Figure 1) allowed visual inspection of potential relationships between weather information and water losses (WL) before formal modelling commenced.

Figure 1 shows correlation between the count of air frost days and run length; this is to be expected as they are closely linked. Both the $Swing_{i,t}$ and $SMD_{i,t}$ are negatively

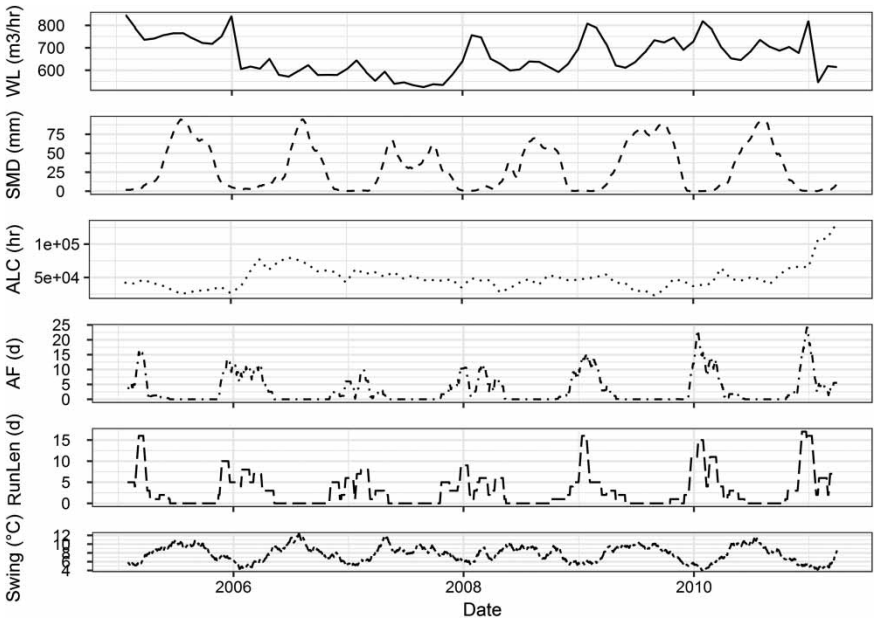

Figure 1 | Time series variation of all model variables in catchment 1 showing the relative variations in predictors and the associated variation in the response variable of water losses (WL).

correlated to $AF_{i,t}$. This is also intuitive, since there is typically more rain and less evapotranspiration in winter than in summer and so it follows that SMD peaks in late summer when there are also fewer air frost days.

Looking at the shape of the water losses response, there are some clear peaks in 2008, 2009 and 2010 which coincide with the winter peaks in air frost and run length. Because of the smooth nature of ALC, it is more difficult to make out any relationship between water losses and ALC, but it can be seen that step changes in water losses at the start of 2006 and 2011 are coincident with step changes in ALC.

EXPERIMENTAL DESIGN

In order to compare performance of a machine learning method over conventional linear regression, a benchmark model was devised against which performance could be measured. The benchmark used ordinary least squares multiple linear regression and the machine learning model was created using random forests.

The models were trained using the R (R Core Team 2013) caret package (Kuhn 2008). random forest (RF) models in caret are trained through the use of the R randomForest (Liaw & Wiener 2002) package. These allow for the computation of both RMSE and R-squared. Lowest RMSE was used for model comparison.

Linear model

In multiple linear regression, the dependent variable, Y (in this case water losses) is modelled as a linear combination of independent variables $X_1, X_2, \ldots, X_p = \boldsymbol{X}$, according to Equation (2).

$$Y = \beta_0 + \beta^T \boldsymbol{X} + \epsilon \tag{2}$$

where $\beta^T = \beta_1, \beta_2, \ldots, \beta_p$ are the model coefficients and ϵ represents the noise in the measurement of Y. It is possible to include interactions between predictor variables $\boldsymbol{X}$; thus:

$$Y = \beta_0 + \beta^T \boldsymbol{X} + \gamma \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}} + \epsilon \tag{3}$$

but they must be explicitly included by the operator. The inclusion of all possible interaction terms would cause an exponential increase in the number of possible predictors and can swiftly become intractable.

Linear regression is a global model: a single model is assumed to hold over the entire observed locus (and

beyond if extrapolation is employed). If the relationship between the response and predictors is truly linear, then this assumption is satisfied.

Random forests

By contrast, random forests (Breiman 2001) are nonparametric models consisting of an ensemble of classification or regression trees (Breiman *et al.* 1984). When the output required is quantitative, regression trees are used.

Regression trees (Breiman *et al.* 1984) work by partitioning the training data into smaller subspaces by setting a threshold on one of the predictors and splitting the training set cases into those above the threshold and those below. At each split, the predictor variable and threshold that provide the best prediction (as measured by mean absolute error (MAE)) are chosen.

The resulting partitions have a simpler relationship to the response variable. Thus, a regression tree consists of a set of rules for partitioning the data and a simple model that can be applied to the data in each of those parts. In the Breiman *et al.* (1984) implementation of regression trees, the model for each node is a constant estimate. A detailed explanation of the regression trees method is provided in Hastie *et al.* (2001).

Trees are grown to a certain size dependent on a stopping criterion. Usually, this is either a maximum number of nodes in the tree or a minimum size of terminal nodes. If, as was the case for this paper, a minimum size of terminal node is used, all nodes to which more than the minimum number of training samples are allocated will be further subdivided.

Regression trees tend to overfit data because they have very high variance and low bias (Hastie *et al.* 2001). The expected error on unseen data is given by Equation (4). The high variance and low bias allows for the approximate function, $\hat{f}(x)$, to fit noise in the data rather than the underlying relationship.

$$\mathrm{E}\left[\left(y-\hat{f}(x)\right)^2\right] = \mathrm{Bias}\left[\hat{f}(x)\right]^2 + \mathrm{Var}\left[\hat{f}(x)\right] + \sigma^2 \quad (4)$$

where:

y is the unseen observation;

$\hat{f}(x)$ is the model function that approximates the true function, $f(x)$; and

σ is the noise in $f(x)$.

This risk can be mitigated by pruning the tree (removing nodes that do not provide significant improvement) which can be done through the cost-complexity pruning method (Hastie *et al.* 2001). However, because the expectation of the average of an ensemble of trees is the same as the expectation of any one of them, an ensemble has the same bias as an individual tree while maintaining a reduced variance. For this reason, methods that average the prediction of an ensemble of trees, like random forests, do not require pruning.

Random forests build an ensemble of trees under the following algorithm:

- If the number of cases in the training set is N, sample N cases at random – but with replacement – from the original data. This sample is the training set for growing the tree.
- If there are M predictor variables, a number $m < M$ is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the growth of the forest.

Predictions are made by the forest by first creating a prediction for each tree in the forest and then taking the average prediction across all trees.

Random forests offer several theoretical advantages over ordinary linear regression. Because of the use of regression trees, the algorithm is capable of automatically detecting and modelling interactions between variables, removing the requirement on the operator to specify interactions. This becomes even more important when the number of predictors is very high and it is not practical for the operator to specify all possible interaction terms.

Random forests do not assume any particular relationship between the response and the predictors. In the case that the relationship is highly non-linear, random forests would be expected to perform significantly better than ordinary linear regression.

One key limitation of random forests is that they do not allow for extrapolation beyond the locus of the training set.

If predictions were required for more extreme weather than observed in the training data, then alternative methods would be required. However, for the future utility of this research, it was helpful that the observation period included multiple periods of severe winter weather relative to the climatological average for the catchments studied.

Repeated k-fold cross-validation

To train and test the result of both linear and RF models, we applied a technique known as repeated k-fold cross validation. Cross-validation refers to partitioning the original sample into training (for fitting the model) and validation (for assessing the model's predictive power).

K-fold cross-validation works by first randomly subdividing the training set into a number, k, of equally sized partitions or *folds*. Then, leaving out one of the folds as validation data, a model is trained on k-1 folds. This model, which has no prior knowledge of the validation data, is then used to create predictions on the validation data. By comparing the predictions to the known values, an estimate of the model's performance on unknown data is obtained.

This process results in some loss of information because $1/k$ of the potential training sets are not used. Cross-validation works by then repeating the above steps until each fold has been left out once and each fold has been used to train k-1 models. The performance of a model is the average performance over all k iterations.

In repeated cross-validation (Kohavi 1995), k-fold cross-validation is performed a number, n, times with different random folds each time. Because the variance of the model increases with less data, the magnitude of k cannot become too large as it would cause the size of the folds to become too small. Repeated cross-validation allows for more folds to be attempted without reducing their size and so decreasing the variance of the performance metric. This study made use of 10-fold cross-validation repeated three times.

Performance metric

The measure chosen to assert the predictive power of the model variants was root mean squared error (RMSE). Its units are the same as the response variable: m^3/h in our case. It serves to aggregate the prediction errors across all test cases into a single measure of predictive performance. RMSE was chosen in preference to MAE, as the former combine information on the magnitude of the average error and the variance of the errors, making it sensitive to outliers. For RMSE, a value of zero would indicate a perfect fit but it can take any positive value. RMSE formulation is given by Equation (5).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2} \tag{5}$$

where:

N is the number of test cases;
Y_i, is the true value of the i^{th} test case; and
$\hat{Y}_i$ is the predicted value of the i^{th} test.

Additionally, R-squared was provided due to its prevalence in statistical modelling and its ease of understanding. R-squared can be understood as the proportion of the variance in the test set that can be explained by the model and, as such, is bounded between 0 and 1.

Model construction

Models were created for each modelling technique by examining the possible predictor variables that could be included. Expert elicitation was used to remove permutations that were not likely to outperform models of equivalent complexity.

Ten candidate RF models were created constituting ten different permutations of the candidate predictors. All models included a fixed 'catchment' term to allow for the fact that individual catchments may differ for reasons which are not characterised by the ALC effort and weather predictors. The individual models were ordered by increasing complexity:

- RFv1: catchment, ALC;
- RFv2: catchment, SMD;
- RFv3: catchment, AF;
- RFv4: catchment, ALC, SMD;
- RFv5: catchment, ALC, AF;
- RFv6: catchment, SMD, AF;
- RFv7: catchment, ALC, SMD, AF;

- RFv8: catchment, ALC, SMD, AF, Swing;
- RFv9: catchment, ALC, SMD, AF, RunLen; and
- RFv10: catchment, ALC, SMD, AF, Swing, Runlen.

A larger number of linear models were included due to the need to test a number of candidate interaction terms (not explicitly required in random forests):

- LinearV1: catchment, ALC;
- LinearV2: catchment, SMD;
- LinearV3: catchment, AF;
- LinearV4: catchment, ALC, SMD;
- LinearV5: catchment, ALC, AF;
- LinearV6: catchment, SMD, AF;
- LinearV7: catchment, ALC, SMD, AF;
- LinearV8: catchment, ALC, SMD, AF, Swing;
- LinearV9: catchment, ALC, SMD, AF, RunLen;
- LinearV10: catchment, ALC, SMD, AF, Swing, RunLen;
- LinearV11: catchment, ALC, SMD, AF, Swing, RunLen, interaction between SMD and AF;
- LinearV12: catchment, ALC, SMD, AF, Swing, RunLen, interaction between SMD and Swing;
- LinearV13: catchment, ALC, SMD, AF, Swing, RunLen, interaction between SMD and RunLen;
- LinearV14: catchment, ALC, SMD, AF, Swing, RunLen, interaction between SMD and AF, interaction between SMD and Swing;
- LinearV15: catchment, ALC, SMD, AF, Swing, RunLen, interaction between SMD and AF, interaction between SMD and RunLen;
- LinearV16: catchment, ALC, SMD, AF, Swing, RunLen, interaction between SMD and Swing, interaction between SMD and RunLen;
- LinearV17: catchment, ALC, SMD, AF, Swing, RunLen, interaction between SMD and Swing, interaction between SMD and RunLen, interaction between SMD and AF.

RESULTS AND DISCUSSION

Model comparison

Figure 2 shows the performance metrics for each model investigated. The two performance metrics are shown on separate scales and the models are ordered by increasing performance.

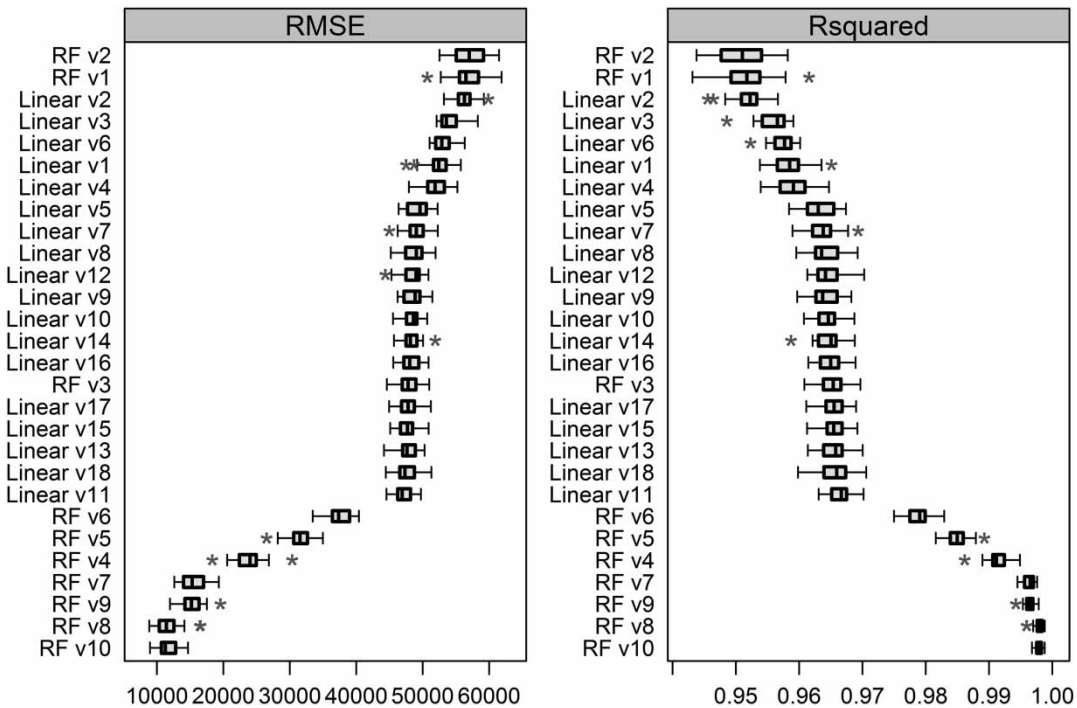

Figure 2 | Comparison of models using RMSE and R-squared as performance metrics.

Figure 2 shows that very simple random forests, with few predictor variables, can have similar or worse performance than corresponding linear models. RFv1 and RFv2 each have the same covariates as LinearV1 and LinearV2, respectively, and exhibit worse performance.

Note that, because RMSE is on the same scale as water losses, it has large values that, nevertheless, are small relative to the total catchment value.

As models increase in complexity and have more predictor variables on which to perform splits, random forests outperform linear models by as much as 35 m^3/h (left hand side of Figure 2). We can also see a steep increase in R-squared performance, boosting the variance explained from around 0.95 to values higher than 0.99 (right hand side of Figure 2).

The best RF model, RF v10, is based on all candidate variables. The best linear model, Linear v17, makes use of all the candidate variables and the three additional interaction terms.

Prediction using best model variants

From Figure 2, we can see that the best RF model outperforms the best linear model considerably on both performance metrics. RF v10 has an RMSE of 11.7 m^3/h, whilst Linear v17 has an RMSE of 47.6 m^3/h.

Figure 3 shows the predictive performance of the best models (linear and RF) by overlaying predicted and observed values.

The best linear model (left hand side of Figure 3) does a reasonable job of approximating the observed losses. In particular, it can be seen that the peaks in the predicted water losses are concurrent with observed values. However, it struggles to predict major step changes in water losses: note that the trough in 2007 observed in catchment 1 is not reflected in the prediction. More generally, the predicted peaks and troughs in catchment 1 have a smaller magnitude than those observed. The linear model for catchment 2 appears to perform better, but this is probably due to the much lower

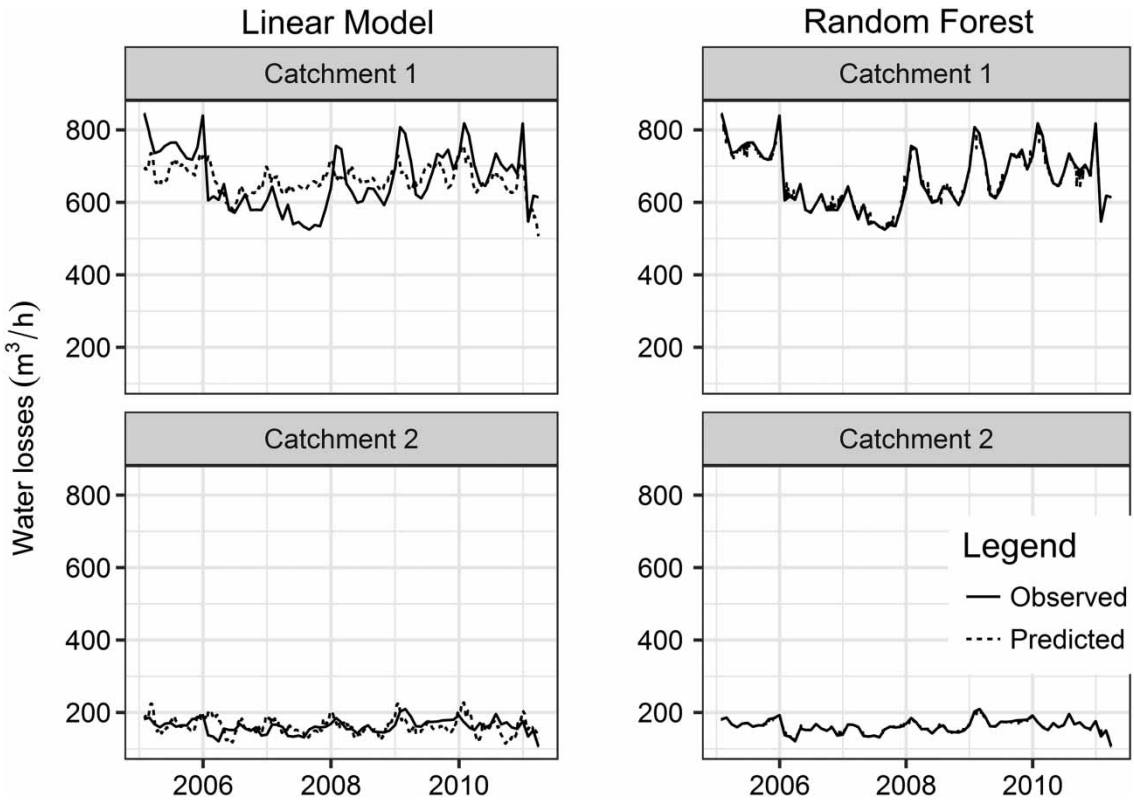

Figure 3 | Comparison of water losses predicted by linear and random forest models and the observed monthly water losses for two catchments.

magnitude and variation in water losses observed in catchment 2, reducing the absolute magnitude of the differences.

From the best RF model (right hand side of Figure 3) water losses predictions are indistinguishable from the observed values for almost all of the period. The model shows some variance around the observed values, as the model is not perfect. Visual comparison of both sides of Figure 3 demonstrates the advantages of using RF models over linear regression.

CONCLUSIONS

This study has compared the performance of traditional linear regression models and random forests for the prediction of water losses. Through experimental comparison of the two techniques using data from one UK water utility, random forests were identified as the more successful technique.

Linear models are sufficient for determining temporal trends in water losses but under-represent the variance that has been historically observed. This results in under-predicting the peaks and over-predicting the troughs. The use of these models to aid in the setting or evaluation of leakage targets would not bear the current level of public and regulatory scrutiny. Random forests provide a more robust alternative with a significantly higher predictive power.

The cross-validation performance of the models provides an estimate of the performance of the model on new data but the true performance is unknown. It is possible that the models presented in this report provide better predictions on the training dataset than they would on other data. This study could be developed further by validating the models on more recent data between 2012 and 2017 to test its performance on truly unseen data. This would provide additional confidence in future predictions using random forests.

REFERENCES

Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B. & Sliusarieva, A. 2012 Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research* **48** (1).

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984 *Classification and Regression Trees*. Wadsworth & Brooks, Monterey, CA.

Hastie, T., Tibshirani, R. & Friedman, J. 2001 Regression Trees. In: *The Elements of Statistical Learning* (Hastie, T., Friedman, J. H. and Tishbirani, R., eds). Springer New York Inc., New York, pp. 307–310.

Herrera, M., Torgo, L., Izquierdo, J. & Pérez-García, R. 2010 Predictive models for forecasting hourly urban water demand. *Journal of Hydrology* **387** (1), 141–150.

Kohavi, R. 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection. *In Ijcai* **14** (2), 1137–1145.

Kowalski, M. 2006 *The Role of Meteorological Processes in Leakage From Buried Water Supply Pipes*. MSc Dissertation, University of Reading.

Kuhn, M. 2008 Caret package. *Journal of Statistical Software* **28** (5), 1–26.

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

Maidment, D. R. & Miaou, S. P. 1986 Daily water use in nine cities. *Water Resources Research* **22** (6), 845–851.

R Core Team 2013 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Robinson, E. L., Blyth, E., Clark, D. B., Finch, J. & Rudd, A. C. 2015a Climate hydrology and ecology research support system meteorology dataset for Great Britain (1961–2012) [CHESS-met]. NERC Environmental Information Data Centre. https://doi.org/10.5285/80887755-1426-4dab-a4a6-250919d5020c.

Robinson, E. L., Blyth, E., Clark, D. B., Finch, J. & Rudd, A. C. 2015b Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961–2012) [CHESS-PE]. NERC Environmental Information Data Centre. https://doi.org/10.5285/d329f4d6-95ba-4134-b77a-a377e0755653.

Schulte, R. P. O., Diamond, J., Finkele, K., Holden, N. M. & Brereton, A. J. 2005 Predicting the soil moisture conditions of Irish grasslands. *Irish Journal of Agricultural and Food Research* **44** (1), 95–110.

UK Water Industry Research 2006 *Managing Seasonal Variations in Leakage*. Report number 07/WM/08/35.

UK Water Industry Research 2013 *Effect of Weather on Leakage and Bursts*. Report number 13/WM/08/50.

Water UK 2016 Water resources long term planning framework (2015-2065).

First received 15 August 2017; accepted in revised form 6 February 2018. Available online 20 February 2018