Environmental Modelling & Software 47 (2013) 207–217

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Recursive partitioning techniques for modeling irrigation behavior

Sanyogita Andriyas [a,b,*], Mac McKee [a,b]

[a] Department of Civil & Environmental Engineering, Utah State University, Logan, UT, USA
[b] Utah Water Research Laboratory, Department of Civil & Environmental Engineering, Utah State University, Logan, UT, USA

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form
14 May 2013
Accepted 21 May 2013
Available online 21 June 2013

Keywords:
Farmer
Irrigation behavior
Recursive partitioning
Trees
Decision

ABSTRACT

Accurate forecasts of short-term irrigation demands can provide information useful for canal operators to manage water diversions and deliveries more efficiently. This can be accomplished by analyzing the actions of the farmers who make water use decisions. Readily available data on biophysical conditions in farmers' fields and the irrigation delivery system during the growing season can be utilized to anticipate irrigation water orders in the absence of any predictive socio-economic information that could be used to provide clues into future irrigation decisions. Decision classification and the common factors, form a basis for division of farmers into groups, which can be then used to make predictions of future decisions to irrigate. In this paper, we have implemented three tree algorithms, viz. classification and regression trees (CART), random forest (RF), and conditional inference trees (Ctree), to analyze farmers' irrigation decisions. These tools were then used to forecast future decisions. During the training process, the models inferred connections between input variables and the decision output. These variables were a time series of the biophysical conditions during the days prior to irrigation. Data from the Canal B region of the Lower Sevier River Basin, near the town of Delta, Utah were used. The main crops in the region are alfalfa, barley and corn. While irrigation practices for alfalfa are dependent on the timing of cuts, for barley and corn the critical crop growth stages are often used as indicators of farmer decisions to irrigate. Though all the models performed well in forecasting farmer decisions to irrigate, the best prediction accuracies by crop type were: 99.3% for alfalfa using all the three models; 98.7% for barley, using the CART model; and 97.6% for corn, with Ctree approximately. Crop water use, which is the amount of water lost through evapotranspiration, was the prime factor across all the crops to prompt irrigation, which complies with the irrigation principles. The analyses showed that the tree algorithms used here are able to handle large as well as small data sets, they can achieve high classification accuracy, and they offer potential tools to forecast future farmer actions. This can be subsequently useful in making short-term demand forecasts.

1. Introduction

In this data-rich world, there is a lack of pertinent information about certain phenomena that are either hard to model or lack a complete physically based cause–effect description of the problem. This presents challenges in the use of conventional approaches such as deterministic models to predict future conditions. Such a problem exists in understanding and predicting a farmer's decision to irrigate. Substantial scientific theory and large quantities of data are available to analyze the irrigation problem and forecast short-term irrigation demand, but the problem of accurately anticipating short-term water demand of an individual irrigator still remains. This is due to a limited understanding of the irrigation practices that are followed by different farmers and how farmer preferences influence decisions about the timing of irrigation.

The site selected for this study is equipped with technologies to monitor reservoir releases and canal diversions, and it has dependable forecasts of evapotranspiration (ET). Some of the irrigated fields have real-time soil moisture measurements to study agricultural water use in the area. In spite of these developments, day-to-day irrigation demands are difficult to forecast. Information about such demands can be vital to help irrigation system operators achieve greater efficiency in water deliveries. In an on-demand irrigation delivery system, farmers make the basic water use decisions. Hence it is essential to consider their decision-making mechanisms in forecasting short-term irrigation demand.

Irrigation behavior has rarely been a topic of research. Each farmer has personal goals to achieve in a season, ranging anywhere

* Corresponding author. Utah Water Research Laboratory, Department of Civil & Environmental Engineering, Utah State University, Logan, UT, USA. Tel.: +1 435 797 2932; fax: +1 435 797 1185.
E-mail addresses: sandriyas@gmail.com, s.andriyas@aggiemail.usu.edu (S. Andriyas), mac.mckee@usu.edu (M. McKee).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.05.011

from profit maximization, to crop quality, to being environmentally conscious about saving water. A farmer whose primary profession is agriculture will make different choices from the one who considers agriculture as a secondary occupation. These characteristics make it more difficult to forecast behavior. The few studies that have dealt with irrigation behavior have been inconclusive in understanding the factors that contribute to decisions regarding if and when to irrigate. To find out the scope of studies done on farmer's behavior previously, we are presenting some of the notable ones in the field.

Becu et al. (2006) used a multi-agent system for a study of water sharing between two villages located at the upstream and downstream ends of a watershed. The objective was to evaluate various options to allot water to the villages and provide feasible solutions to different water users for dealing with water scarcity. The solutions were found by analyzing the impact of different land use and water management options on water deficit. Since it involved water use decisions, farmer's behavior was considered in terms of what crops are planted, when they are harvested, and how they are irrigated during the season. Farmers were grouped into various classes on the basis of different cropping patterns identified in the region. This study simulated irrigation decisions taken by the farmers on the basis of the crops they were growing. The paper primarily evaluated the use of a multi-agent system to support collective decision-making in a participatory modeling framework. The farmers initially had misunderstood the model as being the representation of the real world but the model tested scenarios and suggested possible solutions. The study showed that the players involved from the upstream village were concerned about the impact of water scarcity on both villages, while the ones downstream were only locally concerned. Bontemps and Couture (2002) developed a sequential decision model to study water use behavior under conditions when the farmers paid a negligible amount to obtain water and there was no charge for supplying it. The model required precisely calibrated crop-growth simulation models, irrigation practices and information about land use and climate for the region of interest. Data was obtained by integrating an agronomic model, and a solution-searching optimization methodology connected to an economic model. Irrigation demand functions were estimated using non-parametric methods. Three different functions were estimated keeping three types of representative weather conditions in mind. The method was applied to estimate crop demands in southwest France. Results for all types of weather conditions were found to be same (demand function curves had the same shape: decreasing and non-linear), suggesting that irrigation demand was inelastic for the small amount of water available, but if the total quantity of water was increased, the demand became more elastic. The results showed that the threshold price at which alteration of price-response seems to take place, depends on weather conditions. Le Bars et al. (2005) developed a multi-agent systems paradigm to simulate farmer-agents and their decisions over a number of years, under conditions where water supply was limited. The water manager controlled, the amount of water given to a farmer by using allocation rules that were based on the amount of water requested by farmers at the beginning of the season. The farmer-agents each owned a farm with several plots and could decide their own cropping plan. Weather variables were random. This agent-based model helped the negotiations between water suppliers, farmers, public servants and environmentalists by presenting the impacts of water allocating rules based on different criteria. In other words, rules can be tested and resulting consequences can be seen. It was found that for global corn profits, based on the information of whether the previous agent knows about the allocated water to the agent before them, the differences between farmers could be decreased. This decrease would also show a drastic effect on water use efficiency.

From the limited literature on farmer irrigation decision behavior, it is clear that few studies have been conducted to analyze decisions already made or to forecast future irrigation decision under simulated conditions. Models that could provide such forecasts could be potentially useful for improving irrigation system operations. The study reported here is a first attempt at analyzing farmers' decisions using "decision" trees. We use data about the biophysical conditions during the growing season to isolate information available to the farmer about differences on the days leading up to the time of irrigation. We also look into the possibility of using those differences to forecast farmer decision-making.

2. Theory

A wide range of machine learning techniques is available today to address modeling problems, where missing information is an issue. These show promise for the analysis of problems involving the forecasting of decision behavior under conditions where it is not possible to quantify all of the process-specific factors that affect the decision. Fig. 1 shows a tree structure. The nodes are the variables related to the process in form of root, intermediate or terminal nodes (which do not have any child nodes). As we descend in the tree the importance of the variable to the process decreases. The variable at the root node is the most important. The effect of all the variables leading to the terminal node is collective.

Trees are used to understand systems that have little *a priori* information about how and which variables are related. Classification trees have been used by Kastellec (2010) to analyze judicial decisions and laws. Random forests have been used to model ecology applications (Cutler et al., 2007). Das et al. (2009) used conditional inference trees to assess crash severity in road accidents and found the factors involved. These are some applications of trees to real problems.

Decision trees have been a powerful tool for classification and forecasting. The features and capabilities of the trees are described ahead (Hill and Lewicki, 2007). They can give insights into non-parametric, non-linear relationships between a large number of continuous and/or categorical predictor inputs, and output variables, which may be continuous or categorical. When the output variable is continuous it is a regression analysis and when categorical then a classification problem. They divide a heterogeneous group of features into small homogeneous groups with respect to the output variable. A binary tree formed by two child nodes split from each parent node is one such structure. The split is best when it separates the data into groups with two different predominant classes. "Levels" in a tree are referred to as the depth of the tree, and "size" is the number of nodes in the tree. The measure often used to estimate the split is known as "purity." The best split is defined as

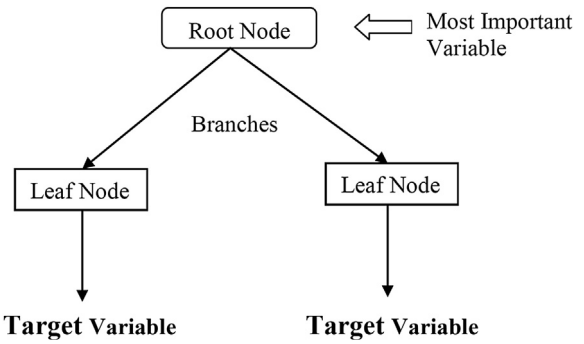

Fig. 1. A tree structure, showing root/parent nodes, branched into leaf nodes which can be intermediate nodes or terminal nodes and the variable of interest or the target variable.

the one which increases the purity of sub-groups by a considerable amount and creates nodes of similar size (not very small ones). Dense structures can often be simplified by pruning. The tree models make no prior assumptions about the data. No unit conversions are required. Raw data can be used as it is. The variables at the root of the tree are deemed the most important.

2.1. Classification and regression trees

Breiman et al. (1984) popularized the classification and regression tree, commonly known as C&RT or CART algorithm. Binary recursive partitioning forms the basis of CART analysis (Breiman et al., 1984). For the analysis, binary partitioning is repeated, hence the term recursive is used. Each parent node results in two child nodes and these nodes are further split into other child nodes. The data set itself is partitioned into sections to form homogeneous groups with similar features. As an example, we assume that the categorical variable at node *t* has two responses: 'Yes' and 'No'. There are basically four steps in CART analysis (Breiman et al., 1984):

i. Tree building : All the data are at first placed at the root node. During learning, the first variable in the sample is split at all the possible values in the data. For each split there are two resulting nodes, a 'Yes' and a 'No' response. All the cases with corresponding responses, 'Yes' and 'No', are classified accordingly. A node is assigned a class, even though it may or may not be split further into child nodes. After this a goodness-of-split criterion is applied to each split to assess the reduction in impurity (or heterogeneity). In CART, one of the measures of impurity is the Gini Index given as:

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t) \text{ if the misclassifications are equally costly,}$$

where, the $p(i|t)$ and $p(j|t)$ are the probabilities of category i and j, respectively, at the node t. Then the best split on the variable is the one for which reduction in impurity is the highest. The above is done for all the remaining variables. In the next step, CART ranks the "best" splits on each variable according to the Gini Index. The best split is the one which most reduces the Gini Index (Breiman et al., 1984). We repeat the above steps for all the non-terminal child nodes of the tree.

ii. Stopping criterion: At this point a large tree has been produced which over-fits the information contained within the learning data set. New splits are stopped when they result in very little or no improvement in the predictions. Resubstitution error rate of the classifier is one accuracy estimate used at this point. It is the proportion of cases misclassified on the same sample that was used during learning. It is given as (Hill and Lewicki, 2007):

$$R(d) = \frac{1}{N}\sum_{i=1}^{N} X(d(x_n) \neq j_n)$$

where, $d(x)$ is the classifier and X is the indicator function such that,

$X = 1$, if $X(d(x_n) \neq j_n)$ is true and

$X = 0$, if $X(d(x_n) \neq j_n)$ is false.

This indicator is usually biased and underestimates the true error rate.

iii. Tree pruning: CART reduces the splits using 10-fold cross-validation (for details refer to Hill and Lewicki (2007) and Breiman et al. (1984)), which results in the creation of a sequence of simpler tree structures by removing the unimportant nodes. For example, if 10 splits result in 90% accuracy during prediction, but 11 splits result in 91% accuracy, then 10 splits are preferred. This also gives more accurate ('honest') estimates of the (true) prediction error in comparison to the resubstitution error. For a sequence of trees, these estimates of error are plotted against tree size, and the size with the minimum error is selected.

iv. Optimal tree selection: The tree which does not over fit the information in the learning data set is selected from a sequence of pruned trees by evaluating the resulting cross-validation error. Breiman et al. (1984) suggested using the 1-SE rule where in the optimal tree is the smallest tree such that its estimated error rate is within one standard error of the minimum. The test sample estimate (error) is then evaluated as the mean squared error between the predicted and the observed data.

CART analysis also produces a variable importance table. This provides a list of all the explanatory variables used and not used in the tree-building process with a score linked to each variable. This score is based on the improvement each variable makes as a surrogate to the primary (the one which shows up in the tree structure) splitting variable. The variable with the highest sum of improvements is scored 100, and all other variables are scored lower, descending toward zero. This helps identify the variables whose significance is disguised by other variables in the tree building process.

2.2. Random forests

The random forest (RF) algorithm uses many decision trees to perform ensemble classification (Breiman, 2001). Random forest algorithms typically have good accuracy, do not over-fit, run efficiently on large data sets with complex interactions, and are robust (Strobl et al., 2008). In RFs, individual classification decisions from a large number of random classifiers (trees) are grouped. This is done when the tree "votes" (provides a classification) for that class. The classification with the largest number of votes across all the trees is chosen by the forest (which is comprised of N classification trees). This ensures more accurate predictions than a single tree classifier. Each tree in RF is built using bootstrapped samples, equivalent to approximately 63% of the observations from the original data set (Cutler et al., 2007), with replacements, leaving about one-third of the cases which are termed as *oob* (out-of-bag) data. Both inputs and variable selection are used randomly at each split in the tree. If there are N explanatory variables, then a number $n < N$ is selected such that n variables are randomly selected out of N and the best split on these n variables is used for node splitting. The number of variables n, remains constant during forest growth. The trees are not pruned. When a large number of classification trees have been grown, class membership of new data is predicted for the *oob* cases. Though the *oob* cases are from the original data set, they do not occur in a bootstrap sample. The predictions for these cases provide unbiased (cross-validated) estimates of the prediction accuracy of the model (Cutler et al., 2007). All the new cases are sent down the trees starting from the root. Every tree in the forest gives its classification for those cases at their terminal nodes. For example, if the classes are "Yes" and "No", then number of trees having "Yes" classification/votes are counted, and the percent of "Yes" votes of the total votes is the predicted probability. This gives a combined predicted classification and is referred in the literature as "majority

voting". Error rates are estimated using the *oob* predictions and are averaged over all the cases in the data set. Each tree in the forest can be associated with a misclassification rate for the *oob* cases. For variable importance, the values of '*n*' predictor variables in the *oob* data are randomly permuted and put down the tree to get new predictions. The measure of importance of the variable is the difference between the misclassification rate for the original and the permuted *oob* data, divided by the standard error (Cutler et al., 2007). Details can be found in Breiman (2001).

2.3. Conditional inference trees

Conditional inference tree (Ctree) models regress relationships between predictor variables and target variables by recursively partitioning data in a conditional inference framework (Hothorn et al., 2006). The trees are built using the following steps: The global null hypothesis of independence between the input and output variables is tested. The model terminates if the hypothesis is not rejected. In the case when the hypothesis is rejected, the algorithm selects that input variable which is strongly associated with the target variable using a *p*-value resulting from a test for partial null hypothesis. A binary split is performed on this input variable. Testing and splitting is repeated for all covariates, recursively. A certain stopping criterion based on hypothesis tests is adopted (e.g. $p < 0.05$). This usually avoids any over-fitting or biased variable selection (Hothorn et al., 2006). By using the Gini index, the chances of finding a good split increases if the variable is continuous or has numerous categories. CART is found to have a bias in variable selection for continuous variables. Conditional trees use a chi-square significance test for variable selection, as opposed to CART which selects the variable that maximizes an information measure like the Gini index. In spite of these advantages the model is still new and experimental.

3. Case study and methods

3.1. Study site and data available

The data used in this study are from the Canal B region of the Lower Sevier River Basin, near the town of Delta in south-central Utah This area covers approximately 20 square miles of irrigated farm land. Alfalfa, barley and corn are the main crops grown in the area. Irrigation consumes a large amount of water in this basin. Weather data for Delta was obtained from the following website: http://www.cemp.dri.edu/cgi-bin/cemp_stations.pl?stn=delu. Data estimated using Kimberly Penman Reference ET rules are available on this website.

Table 1 presents the variables used to build trees and predict the irrigation decision. Variables 1, 4, 5, 6, and 21 are weather variables. Data were also available for canal flow rate (Variable 13). Three types of soil are found at Delta: silty clay loam, silty clay, and loam. Farmer identification numbers convey information to the model that the data are from a different subject. Soil moisture contents were available from numerous soil moisture probes installed in the Canal B irrigation command area, and were corrected using a mass balance constraint on soil moisture. Variables 7–10, 18, and 20–25 are required for the soil moisture balance computation. A daily time series was created for market prices of alfalfa, corn, and barley

Table 1
Predictor variables, the represented factors as seen by the farmer, and the target variable used for trees analysis.

S. no.	Variable name	Represented factor	Continuous or categorical-(no. of classes)
1	AirTemp	Average air temperature	Continuous
2	GrowingDegDays	Growing degree days accumulated till a given day and reset on the day of irrigation	Continuous
3	GrowStageIrrigNeed	Sensitivity of growth stage to water stress as indicated by growing degree days	Categorical-(2)
4	WindSpeed	Wind speed	Continuous
5	RH	Relative humidity	Continuous
6	ET	Potential evapotranspiration (ET)	Continuous
7	ET_c	Crop evapotranspiration	Continuous
8	CropCoeff	Crop-specific coefficient	Continuous
9	SoilStressCoeff	Soil stress coefficient	Continuous
10	ET_a	Actual evapotranspiration	Continuous
11	$CumET_c$	Cumulative crop ET	Continuous
12	StressIrrigNeed	Consumptive use as indicated by $CumET_c$	Categorical-(2)
13	CanalFlow	Canal flow rates	Continuous
14	WaterSupplyIrrigNeed	If the farmer irrigated when his neighbors irrigate as indicated by CanalFlow	Categorical-(2)
15	JDay	Julian day in the season	Continuous
16	WeekEndORNOT	Saturday/sunday	Continuous
17	WkEndIrrigNeed	If the farmer irrigated on a weekend as indicated by WeekEndORNOT	Categorical-(2)
18	RootingDepth	Rooting depth of the plant	Continuous
19	CropIrrigNeed	Plant need indicator, deeper the root, frequent are the needs for water as indicated by RootingDepth	Categorical-(2)
20	SMCinit	Soil moisture content at the start of the day	Continuous
21	Rain	Precipitation amount	Continuous
22	AmountPercolation	Amount of irrigation water percolated	Continuous
23	IrrigationAmt	Estimated amount of irrigation from the soil moisture probes	Continuous
24	DepInit	Depletion at the start of the day	Continuous
25	DepEnd	Depletion at the end of the day	Continuous
26	SoilIrrigNeed	If the soil is dry or not (also indicated in plant condition) as indicated by SoilStressCoeff	Categorical-(2)
27	Year	Year, indicating a dry, moderate or wet year	Categorical-(4)
28	Yield	Yields estimated using ETa and ETc	Continuous
29	MarketPrice	Price of the crop	Continuous
30	ProfitORLoss	Profit or loss for the farmer	Continuous
31	EconIrrigNeed	Economic need to irrigate the crop as indicated by ProfitORLoss	Categorical-(2)
32	ID	Different farmers	Categorical-(39)
33	SoilType	Type of soil	Categorical-(2)
34	Irrigate	The irrigation decision	Categorical-(2), (0–1 for regression and Yes–No for classification)

using the monthly data available for the USDA website for Millard County, Utah. Approximate planting dates were established by initiating the soil moisture calculations from a random day such that soil moisture matched the day of first irrigation, which was known from the soil moisture probe data. We assumed that the initial depletion was zero and began the computations from field capacity.

Phenology coefficients (K_c) for all the crops were derived from Wright (1982) and FAO-56 (Allen et al., 1998) and were found to be quite representative. Since we were using the values of K_c for crop reference ET, we had to multiply the values with a factor of 1.2 to model a field crop, instead of grass reference ET.

All the other variables were either derived from the primary data or categorized to simplify their representation in the model. If both the data and the derived variable behave the same, then the derived ones can be removed.

3.2. Models, specifications and performance evaluation

All of the models used in this study were implemented using R-statistical software (R Development Core Team, 2013). For all the classification problems it is necessary that the target variable be categorical (e.g. "Yes" and "No"), which can be done using the "factor" function. For all the models, the data were randomly partitioned into training and testing sets. For all the data sets, one-fourth of the data were used for testing. The input to all these algorithms for our case was the decision to irrigate ("Yes") or not ("No"). During training, the model was tuned according to the irrigation decision. During testing, this target variable was forecasted. The outputs of all the models were the confusion matrix and the error rate. We used the accuracy rate (calculated as the difference between 100 and the error rate) and the confusion matrix to evaluate model performance.

To do a CART analysis, we used the "rpart" (Therneau et al., 2012) package in R. It is powerful and easy to use, and is based on the same algorithm as Breiman et al. (1984). During the model fitting process, the "rpart" function was applied to the training data with the dependent variable being the irrigation decision and method = "class". The "predict" function uses the fitted tree and predicts the classification for the test data set. The "table" function can be used to obtain a confusion matrix, and the accuracy rate can also be obtained. The accuracy rate is the sum of the diagonal elements of the confusion matrix, divided by the sum of all the elements. The "printcp" function can be used to print the complexity parameter (cp) table (Table 2 shows a sample output) for the fitted tree. This way we can find the optimal pruning of the tree based on 'cp'. It can be seen in Table 2 that nsplit (number of splits) denotes the size of the tree, and (nsplit + 1) is the number of nodes in a tree. Scaled errors are presented such that the error at the first node is 1. Using the 1-SE rule to find the best number of splits, the smallest "xerror" is added to the corresponding "xstd" as shown in the last column. The number of splits resulting in the smallest error is the "best split", e.g. in this case, the optimal number of splits is 1. For this case the pruned tree will have 2 nodes. If in case the sum of "xerror" and "xstd" are all equal, the "best split" will be the tree with the fewest number of splits other than zero, which would leave only the root node. To assess the performance of a predictive model, we have also used cross-validation. We cross-validated using bagging (Breiman, 1996), derived from "**b**ootstrap **agg**regat**ing**". Bagging is an ensemble method which decreases the variance of the original individual models by using a bootstrap of the training set to build every new model and then taking the average of the predictions from those models. For running 10-fold cross-validation with bagging we used the "adabag" (Alfaro et al., 2012) package in R. For cross-validation, two-thirds of the data were used for training and the remaining for testing. The function "bagging.cv" requires the target variable as input from the training data set and the number of iterations, "mfinal" variable (default = 100). The most important variable is at the root node.

We used the "randomForest" (Liaw and Wiener, 2012) package of R for model development. The "randomForest" function needs the following as parameters for tuning the forests

- "mtry" is the number of variables randomly sampled as possible choice at each split,
- "ntree" is the number of trees to be populated,
- "importance = TRUE" calculates the variable importance and can be retrieved using the "varImpPlot" function, which plots "MeanDecreaseAccuracy" and "MeanDecreaseGini". The plot presents the variables in the descending order of importance. We used "MeanDecreaseGini" to assess the important variables. This is related to the fact that the stopping criterion of splitting in a tree is when a new split does not reduce the Gini index any further (Breiman et al., 1984). This means that the variable is not important for tree building. "rfcv" with "cv.fold = 10", was used to evaluate the 10-fold cross-validated prediction performance of the models. This was done by sequentially reducing the number of predictors (as ranked in variable importance) using a nested cross-validation procedure. We also used out-of-bag error estimates also to evaluate the models.

Table 2
The cost–complexity parameter (cp), relative error, cross validation error (xerror) and cross validation standard deviation (xstd) for trees with nsplit from 0 to 8.

S. no.	CP	nsplit	rel error	xerror	xstd	Sum of xerror & xstd
1	0.709	0	1.000	1.069	0.050	1.119
2	0.030	**1**	0.291	0.291	0.035	**0.326**
3	0.020	5	0.172	0.345	0.038	0.382
4	0.015	7	0.133	0.330	0.037	0.367
5	0.010	8	0.118	0.310	0.036	0.346

The bold numbers indicate the number of splits resulting in the smallest error is the "best split", e.g. in this case, the optimal number of splits is 1.

For the analysis of conditional inference trees, the "party" (Hothorn et al., 2012) package in R library was used. "ctree_control()" sets the parameters for the tree. We ran with the default settings, which include mtry = 0, implying the variables were not selected randomly at each split. These settings were chosen to avoid different trees at each run. This feature is available in the "randomForest" package. The input to the "ctree" function was the irrigation decision and parameters were supplied through "ctree_control()". The "plot" command was used to produce the resulting trees. The mean "y" and the number of cases "n" ending up at the terminal nodes are also displayed. The predictions were obtained using the "predict" function and the accuracy rate was calculated.

4. Results and discussion

Fig. 2 presents a plot of some of the weather variables to illustrate some existing groups based on the irrigation decision. No obvious classes can be found in the plot.

If apparent groups existed, we would not use such algorithms. Usually in practical problems it is difficult to find classes to discern groups based on the target variable. We have to then seek the help of specialized techniques like recursive partitioning, in our case, for exploring the groups, if any.

Since the day of irrigation can be anywhere from 24 h to 3 days from the day of order (call-time), we decided to examine the importance of the variables presented in Table 1 for:

- all the days in the season (referred to as "All days" in the results)
- four days before the day of irrigation and the day itself (referred to as "4-days" in the results)
- the day of irrigation and a day before (referred to as "1-day" in the results).

The prediction accuracy for the tree analysis is presented in Table 3. The best results are displayed in bold. It can be seen that the predictions of the models, which were given the description of the whole season (i.e., the All-days model), performed better than the 1-day or 4-day models. Given that there is a possibility of some missing information, all three algorithms work exceptionally well to predict future decisions. All three algorithms had close to 99% accuracy for alfalfa irrigation decisions. CART had accuracy estimates of 98.7 and 96.9% for barley and corn. RFs predicted the decisions for barley and corn with an accuracy of 97.8 and 96.2%, respectively. Ctree had accuracy measures of 98.0 for barley and 97.6% for corn.

4.1. CART

A 10-fold cross validation was done on all the data sets and the cross-validation accuracy was reasonably close to the resubstitution accuracy, except for the corn 1-day model, as shown in Table 3. Cross-validation is performed to help determine if the classifier is being over fitted. For our case, the performance is promising. Fig. 3 shows trees built for the three crops in the study. Clearly there are different strategies for the three crops, starting with the same variable. The most important

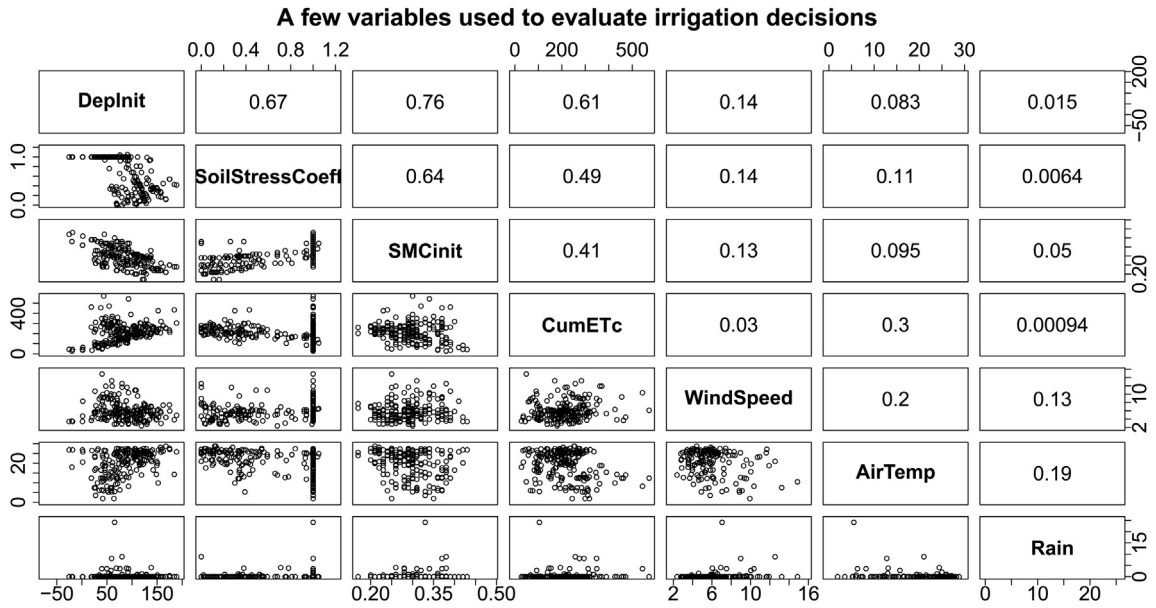

Fig. 2. Pairs plot of some weather variables used in the tree analysis, with the intention of finding groups of similar features.

variable for all three crops was cumulative crop ET, or the consumptive use.

According to many standard agricultural handbooks, crop growth stage and its sensitivity gauged by accumulated heat units (or growing degree days, "GDD"), and the consumptive use that this implies, is almost always considered by farmers when deciding about irrigation. StressIrrigNeed is the indicator for critical $CumET_c$, so it was expected that it would be next in importance in explaining irrigation decisions. At each terminal node, there is a group based on a variable or its level which makes it different from the others, according to CART interpretation. In other words, there is a class at every node where we see either a "Yes" or a "No" response. Since we are interested in the reasons behind irrigation decisions, we are only evaluating "Yes" responses.

Fig. 3(a) shows that the alfalfa growers in the first group were irrigating when their neighbors irrigated (Canal Flow $>$ 209.5cfs, which was a high flow rate), at medium soil moisture depletion ($>$65.16 mm), and when farmers other than the ones shown in Fig. 3(a) were irrigating according to these common rules. Since we wanted to avoid bias in selecting the training and testing sets, we used bootstrapping to sample the data sets. It also becomes important to note that these data sets were a mix of information from any of the four years (2007–2010). The second class evidently used the $CumET_c$ (StressIrrigNeed = Yes) measure to time the irrigations. $CumET_c$ is the crop ET accumulated between irrigations and is the same as depletion. As a general rule, alfalfa is primarily irrigated either before or after the cuts. All the resulting principles conform to recommended irrigation practices, since depletion would not be used generally as one of the indicators to trigger irrigation for alfalfa. The possibility that farmers are choosing to irrigate when their neighbors are irrigating is also refuted since the fields involved have different crop planting dates. This hints at different maturity and cut timings and might even point at a different crop quality.

Barley irrigation strategies were very straight forward, with the low depletion of 22 mm and consumptive use ($>$170 mm) being the only indicators CART could discriminate. This suggests that the farmers who were growing barley were not taking risks with respect to the irrigations because this would mean loss in yield to them, even though they could wait longer to irrigate. We did not have enough data to make strong conclusions but these may be probable reasons for the observed timing of irrigation.

For corn, CART presented a huge tree with many variables. It clearly showed three classes (i.e., three terminal nodes with "Yes" as the irrigate decision). For group one, the day in the growing season (JDays between 127 and 145 were crucial) and an irrigation amount more than 105 mm appeared to be the driving factors. The day in the season is indicative of a certain critical growth stage for corn. The irrigation amount may seem a strange choice for grouping farmers, but it implies that farmers who replenished the soil moisture to this level would irrigate similarly. This corresponds indirectly to consumptive use. For group two, in years 2007 and 2010, the farmers other than those shown irrigated while ET was constraining on the crop. This meant that temperatures could have been high for long periods when the farmers decided to irrigate. In the third class, the group irrigated when consumptive use was more than 122 mm and the predicted market price was higher than before. The consumptive use for corn is always a driving force for irrigation. If there is high moisture stress, the amount of carbohydrates available for kernel development in corn is inhibited, which can affect the yield. The implication is that corn growers were keeping the stress levels in control by irrigating at moderate levels of moisture depletion.

CART pruned trees performed the same as the full grown versions, but the advantage was the smaller number of variables for interpretation. We have opted for the full grown versions, however, since it gives us an in-depth analysis of the factors leading to irrigation. Though the pruned tree narrowed the choices of variables it

Table 3
Accuracy estimates on test data for CART, RF and Ctree models. Resub – resubstitution accuracy estimate, Xval – 10-fold cross-validation accuracy estimate. (a) 1-day, (b) 4-day, (c) All days models.

Crop	Alfalfa			Barley			Corn		
Model	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
CART – Resub	83.5	91.1	99.2	48.5	83.1	**98.7**	66.0	84.9	96.8
CART – Xval	83.3	94.5	99.4	53.4	81.2	98.0	46.0	82.1	97.1
RF	85.6	97.1	**99.3**	78.8	81.9	97.9	51.1	84.0	96.2
Ctree	80. 6	93.1	99.3	57.6	91.6	98.1	44.7	80.7	**97.6**

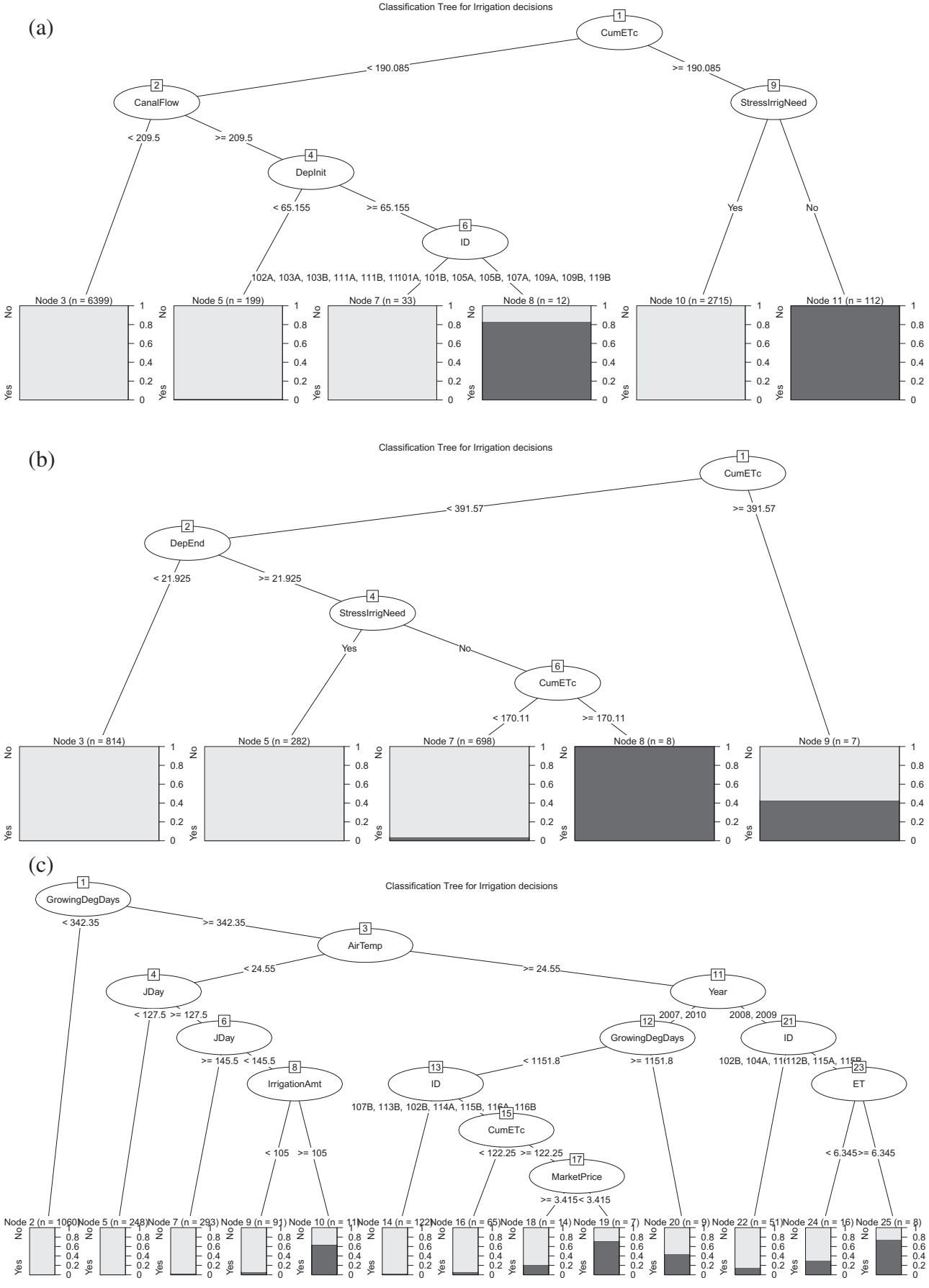

Fig. 3. CART structures for (a) alfalfa, (b) barley, and (c) corn irrigation decisions.

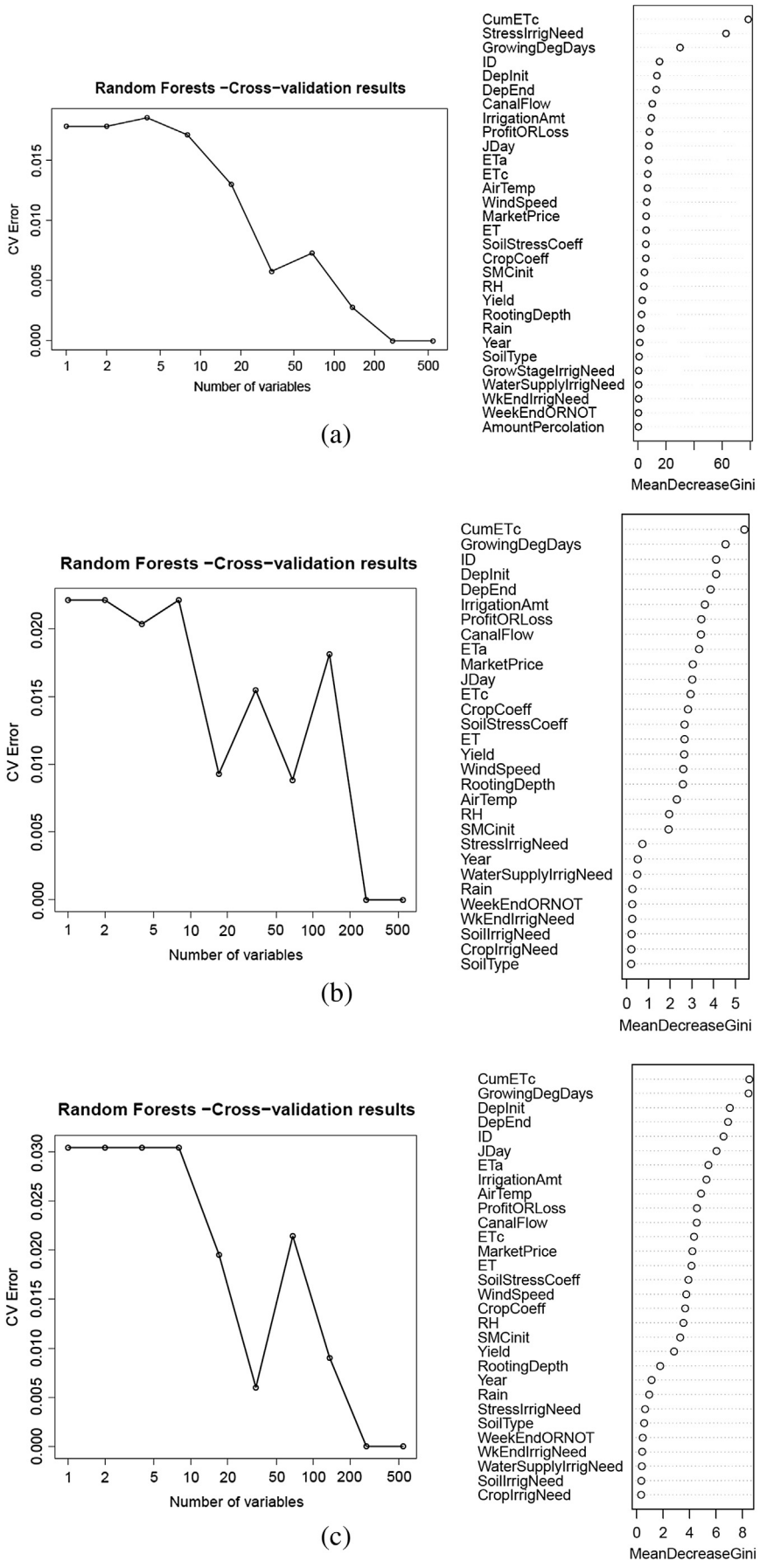

Fig. 4. Random forest 10-fold cross-validation performance and variable importance plots using Gini Index for (a) alfalfa, (b) barley, and (c) corn.

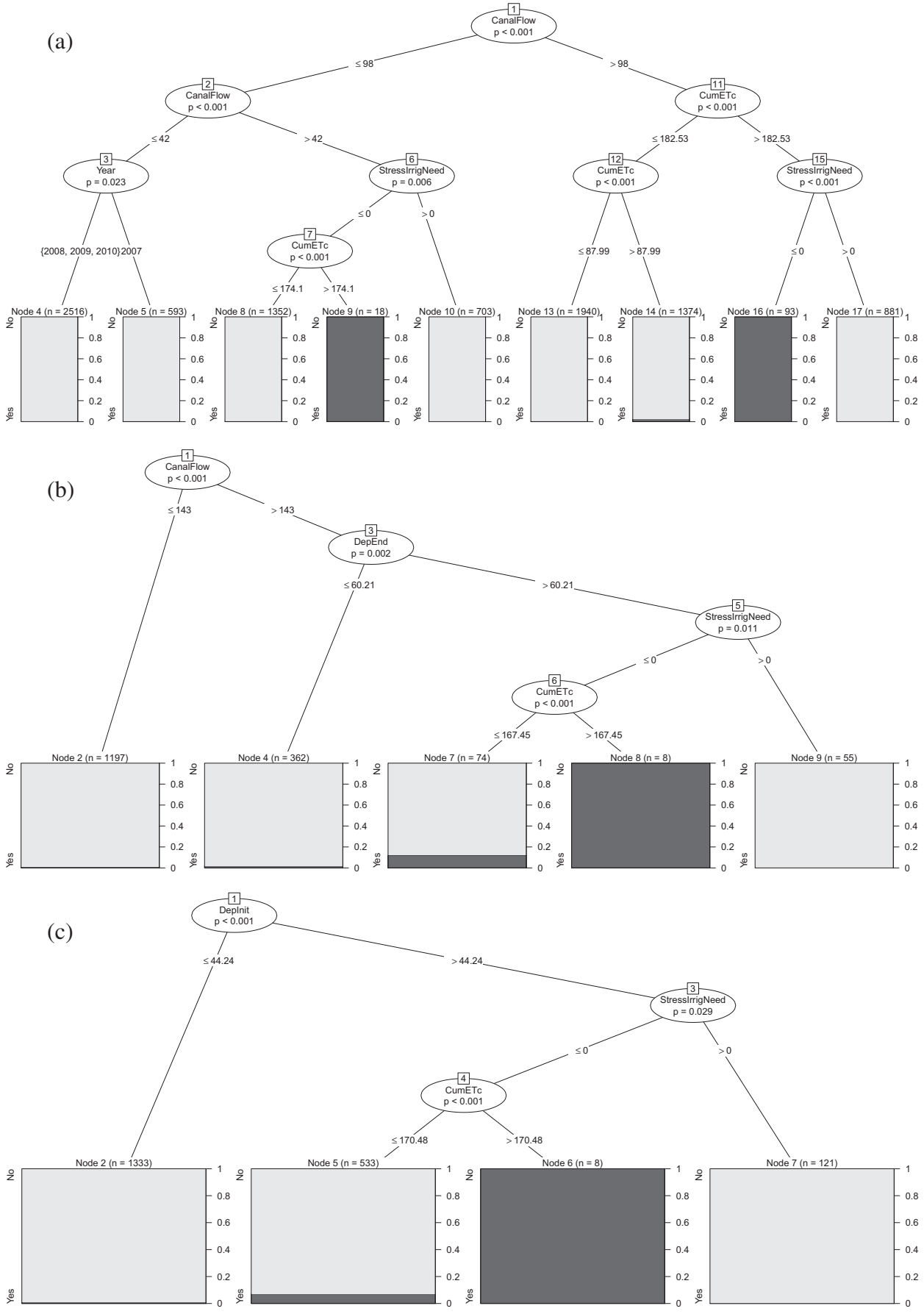

Fig. 5. Conditional inference trees for (a) alfalfa, (b) barley, and (c) corn irrigation decisions.

does not refute the fact that farmers consider multiple factors in the thought-process of scheduling irrigation.

4.2. Random forests

Random forests are a modern tool for classification. Since they have several trees, they have generally been shown to perform exceptionally well in grouping predictor variables according to target decisions. A 10-fold cross-validation (CV) of the All-days data is shown in Fig. 4. The CV error evaluates the effect of adding input variables to the classifier in the order of importance. For alfalfa, Fig. 4(a), the error remained the same at 1.8% for the addition of 1–5 variables, but started dropping sharply as the number of variables increased. For barley, Fig. 4(b), the error was unstable throughout. With variable addition from 1 to 5 it stayed around 2.2%, but with 20 variables it dropped to around 1%. For variable additions between 20 and 50 it rose to 1.5% and then dropped down again. For corn, Fig. 4(c), the error was inconsistent, with variables 1 to 10 maintaining a constant error of around 3%. Further addition of variables dropped the error to 0.5%, but it went up close to 2% between variables 50 to 100. The instability in the errors might be because alfalfa had large amounts of data for learning, while barley and corn did not. In spite of the lack of data, RFs performed better in predicting a decision to irrigate for the test cases across crops, as shown in Table 3. The data sets with limited information (1-day and 4-day models) were not found to perform well. The proportion of times that the predicted class is not the same as the observed class averaged over all cases is the oob error estimate. The out-of-bag error estimates were 0.59%, 1.89% and 2.66% for alfalfa, barley and corn crops, respectively. These indicated that the model built to forecast alfalfa irrigation practices was more reliable than the ones for barley and corn. The driving factors for irrigation in alfalfa, barley and corn as found in CART were confirmed in random forests and are distinct from the Gini Index. Consumptive use and growing degree days were found to be important factors. This means that temperature related factors were found to be important by RFs.

4.3. Conditional inference trees

Conditional inference trees also perform classification. Since we had continuous covariates, we attempted to analyze them using Ctrees. With the exception of a few cases where data were limited, they performed well (refer to Table 3). The tree structures are presented in Fig. 5. They gave insights into some factors which were ignored by other algorithms (Table 4). Ctree showed that alfalfa (Fig. 5(a)) and barley (Fig. 5(b)) growers might be irrigating with their neighbors (as measured by high canal flow levels), but we do not have any related information to confirm this suggestion. Additionally for alfalfa (Fig. 5(a)), year and $CumET_c$ helped farmers to decide the irrigation timing. "Year" variable denotes the year when the measurements were recorded. While $CumET_c$ due to high temperature is justified, the year factor would seem something strange. However, it is pertinent to alfalfa since it is a perennial crop and will be typically cultivated for a period of 3–5 years. The first year irrigation practices will be different for alfalfa since the crop will be germinating and growing, as opposed to the other years where it will emerge and the crop root will already be developed. For barley (Table 4), the farmers also used depletion levels, besides $CumET_c$, to decide irrigation timing. Corn planters (Fig. 5(c)) used consumptive use ($CumET_c$ and depletion) to make irrigation decisions.

Table 4
Important variables for irrigating different crops, according to various models.

Model	Alfalfa	Barley	Corn
CART	CumETc–Canal Flow	CumETc–Depletion	GDD–AirTemp–Day–Year
RF	CumETc–GDD	CumETc–GDD	CumETc–GDD
Ctree	CanalFlow–Year–CumETc	CanalFlow–CumETc–Depletion	Depletion–CumETc

5. Conclusions

Irrigation system managers would benefit from information about short-term irrigation demand. This study applied different types of classification trees to infer how farmers, the water users, make irrigation decisions. This information can be used to predict future actions and forecast short-term water demands, relying on readily measurable biophysical data alone as input. The results from this study show that biophysical conditions can be used as indicators of irrigation behavior, and have a potential to be used as predictors for future irrigation decisions.

The tree algorithms provide analysis of the factors leading to decisions and present a possible forecasting tool. RF, Ctree and CART are all classification algorithms. CART and Ctree present simplified trees, while RF has no means of representing the forest built by it. In terms of modeling different problems, it is important to tune the models and find the best-fit parameters to improve accuracy estimates. It was found that all the models had high classification accuracy to predict irrigation decisions when larger data sets (more information) were used. Smaller data sets supplied incomplete information to the models, resulting in poor classification rates.

All the three models picked logical factors which can possibly lead to an irrigation decision. From the point of view of crop to be studied, models performance cannot be compared for perennial or annual crops, since database for Alfalfa was bigger than that for the other two crops. Information for the whole growing season can increase the number of cases for training, resulting in better model performance. If we need to study the time critical for decision-making, random forests will be a good choice.

Table 4 summarizes the probable important factors exhibited in the tree structures and variable importance measures. The predictors which are most useful in forecasting irrigation decisions are consumptive use, growing degree days or cumulative temperatures, and irrigating when a neighbor irrigates. The variable Year is specific to a perennial crop like alfalfa. Since ET is dependent on temperature, temperature and canal diversion measurements can be used to forecast farmers' future actions. The other important aspect in getting accurate forecasts is the amount of information given to the model. Information for the full growing season should be provided, which means that the models will not be able to handle missing information for this problem. This feature is similar to a farmer managing his farm who monitors day-to-day crop and soil conditions and makes decision accordingly. If he skips a few days in observing these conditions, he will not be able to make appropriate decisions due to the gap in information. We conclude that the most important factor for irrigation behavior appears to be crop need, followed by farmers' observations of their neighbors' actions. These findings are promising and can be used to make estimates of short-term demand forecasts.

References

Alfaro, E., Gámez, M., García, N., 2012. Adabag: Applies AdaBoost.M1, AdaBoost-SAMME and Bagging. Retrieved from: http://cran.r-project.org/web/packages/adabag/index.html. R package version 3.1-52.

Allen, R.G., Pereira, L., Raes, D., Smith, M., 1998. Crop Evapotranspiration. FAO, Rome, Italy.

Becu, N., Sangkapitux, C., Neef, A., Kitchaicharoen, J., 2006. Participatory simulation sessions to support collective decision: the case of water allocation between a Thai and a Hmong village in northern Thailand. In: Paper Presented at the International Symposium Towards Sustainable Livelihoods and Ecosystems in Mountainous Regions, Chiang Mai, Thailand.

Bontemps, C., Couture, S., 2002. Irrigation water demand for the decision maker. Environment and Development Economics 7, 643–657.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140. URL: http://www.springerlink.com/index/10.1007/BF00058655.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

Das, A., Abdel-Aty, M., Pande, A., 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. Journal of Safety Research 40 (4), 317–327. http://dx.doi.org/10.1016/j.jsr.2009.05.003.

Hill, T., Lewicki, P., 2007. STATISTICS: Methods and Applications. Statsoft, Tulsa, OK.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., 2012. A Laboratory for Recursive Partytioning. Retrieved from: http://cran.r-project.org/web/packages/party/index.html. R package version 1.0-0.

Hothorn, T., Hornik, K., Zeilis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics 15 (3), 651–674.

Kastellec, J.P., 2010. The statistical analysis of judicial decisions and legal rules with classification trees. Journal of Empirical Legal Studies 7 (2), 202–230.

Le Bars, M., Attonaty, J.M., Pinson, S., Ferrand, N., 2005. An agent-based simulation testing the impact of water allocation on farmers' collective behaviors. Simulation 81 (3), 223–235. http://dx.doi.org/10.1177/0037549705053166.

Liaw, A., Wiener, M., 2012. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. Retrieved from: http://cran.r-project.org/web/packages/randomForest/index.html. R package version 4.6-6.

R Development Core Team, 2013. R: a Language and Environment for Statistical Computing. Foundation for Statistical Computing, Vienna, Austria. Retrieved from: http://www.R-project.org. R version 3.0.0.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9, 307.

Therneau, T.M., Atkinson, B., Ripley, B.D., 2012. rpart: Recursive Partitioning. Retrieved from: http://cran.r-project.org/web/packages/rpart/index.html. R package version 3.1-52.

Wright, J.L., 1982. New evapotranspiration crop coefficients. Journal of Irrigation and Drainage Engineering, ASCE 108 (IR2), 57–74.