Environmental Modelling and Software 131 (2020) 104781

Contents lists available at ScienceDirect

Environmental Modelling and Software

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

Detecting early warning signals of long-term water supply vulnerability using machine learning

Bethany Robinson, Jonathan S. Cohen, Jonathan D. Herman *

Department of Civil & Environmental Engineering, University of California, Davis, USA

A B S T R A C T

Adapting water resources systems to climate change requires identifying hydroclimatic signals that reliably indicate long-term transitions to vulnerable system states. While recent studies have classified the conditions under which vulnerability occurs (i.e., scenario discovery), there remains an opportunity to extend such methods into a dynamic planning context to design and assess early warning signals. This study contributes a machine learning approach to classifying the occurrence of long-term water supply vulnerability over lead times ranging from 0 to 20 years, using a case study of the northern California reservoir system. Results indicate that this approach predicts the occurrence of future vulnerabilities in validation significantly better than a random classifier, given a balanced set of training data. Accuracy decreases at longer lead times, and the most influential predictors include long-term monthly averages of reservoir storage. Dynamic early warning signals can be used to inform monitoring and detection of vulnerabilities under a changing climate.

1. Introduction

Climate change requires water supply planners to navigate significant uncertainty in future precipitation projections, which in many regions disagree on the magnitude and direction of change (Hallegatte, 2009). Much of this uncertainty centers on extreme drought and flood events, which are expected to become more frequent and severe in the coming decades (Trenberth et al., 2014; Polade et al., 2017). In general, this uncertainty prevents the optimal planning of adaptations and requires innovative approaches such as bottom-up vulnerability assessment, a key feature of robust planning frameworks (Lempert and Collins, 2007; Wilby and Dessai, 2010; Herman et al., 2015). Bottom-up methods focus on identifying the conditions under which vulnerability occurs, using either a wide range of plausible scenarios (e.g., Bryant and Lempert, 2010) or scenario narratives driven by local decision contexts (Rounsevell and Metzger, 2010; Carlsen et al., 2013). This classification problem, known as scenario discovery (Lempert et al., 2008), has benefitted from the application of a variety of statistical methods (Hadka et al., 2015; Kwakkel, 2015; Quinn et al., 2018). The main outcome of scenario discovery methods is a trained classifier capable of mapping uncertain scenario properties to a binary outcome—vulnerable or not—for each scenario aggregated over time.

However, a related question remains less explored: under climate uncertainty, can system vulnerabilities be detected in advance? This is highly relevant to dynamic planning approaches, where adaptations are taken over time in response to observed and projected information (Haasnoot et al., 2013; Zeff et al., 2016; Hui et al., 2018; Fletcher et al., 2019). It also relates to the challenge of anticipating tipping points in environmental systems (Scheffer et al., 2009, 2012). In the case that robust planning proves too costly (e.g., Borgomeo et al., 2018), dynamic planning may increase the effectiveness and appropriateness of adaptations, both preventing over-investments in unnecessary infrastructure and lessening the severity of vulnerabilities if detection methods are sufficiently accurate. Dynamic planning approaches have in common the need to design a policy mapping information to actions, which could include either observations or predictions of vulnerable states (Herman et al., 2020). While the use of observations to trigger adaptations has been explored in detail by bottom-up methods such as scenario discovery and Dynamic Adaptive Policy Pathways (Haasnoot et al., 2013), there remains significant opportunity to study the second case by predicting the occurrence of vulnerable states dynamically—in other words, by designing statistical early warning signals for adaptation.

The process of designing early warning signals first requires the selection of informative feature (predictor) variables. Due to the dynamic aspect of the problem, this includes both the type of variable as well as the timescale, aggregation window, and statistical transformation: for example, the 30-year moving average of annual reservoir inflow (Raso et al., 2019a). Feature selection can affect both efficiency and accuracy, and choosing informative features can leverage both human expertise and statistical techniques (Dietterich, 2002). This concept is analogous to input variable selection in the water resources field (Guyon and Elisseeff, 2003; Galelli et al., 2014), which has been widely used to support

* Corresponding author.
E-mail address: jdherman@ucdavis.edu (J.D. Herman).

https://doi.org/10.1016/j.envsoft.2020.104781
Received 30 March 2020; Received in revised form 17 June 2020; Accepted 17 June 2020
Available online 25 June 2020

reservoir policy search (e.g., Giuliani et al., 2015). However, the longer timescales involved in the climate adaptation problem create additional challenges: it may be more difficult to separate signal from noise (Hegerl and Zwiers, 2011; Hawkins and Sutton, 2012), and the decisions that the features are meant to inform are often irreversible (Raso et al., 2019b).

This challenge is closely related to the choice of signposts for adaptation policies, which are evaluated based on their relevance, credibility, and legitimacy (Haasnoot et al., 2018). Relevance refers to the predictive skill of a monitoring system in observing long-term trends amid short-term variability. Analysis of predictive skill has focused on both Type I errors, representing over-investment, as well as Type II errors, representing under-investment (Rosner et al., 2014; Stephens et al., 2018; Raso et al., 2019). The goals of credibility and legitimacy reflect the fact that monitoring systems must inform human decisions within a broader context of objectives, actions, and spatiotemporal scales among multiple actors and institutions (Hermans et al., 2013, 2017). The interpretability of early warning signals is therefore critical, and may be supported by the parsimony of signpost variables, as well as the completeness of the sources of uncertainty considered (Raso et al., 2019a). In total, these goals for a monitoring system may reflect a tradeoff between predictive skill and interpretability that is widely recognized in statistical modeling. For example, while several studies have considered adaptations triggered by linear threshold values (Hallegatte et al., 2012; Walker et al., 2013; Robinson and Herman, 2019), such signals could also be represented by more complex functions, potentially incorporating multiple variables on different timescales (Herman and Giuliani, 2018; Nayak et al., 2018). This study considers the potential for nonlinear multivariate classifiers to address this problem, recognizing that improvements in predictive skill will likely be met with a decrease in interpretability.

To address this challenge, this study frames the design and testing of early warning signals under climate change as a machine learning classification problem, using observations of human and hydrologic variables to predict the binary occurrence of future water supply vulnerability in a systems model. The goal is to detect vulnerability without knowledge of future forcing, which would be required in a forward simulation of the system. Specifically, this paper addresses the following research questions to investigate the utility of the proposed method:

1. Can machine learning classification techniques predict long-term vulnerability of a water resources system in advance, and are these predictions significantly better than random?
2. How is the accuracy of the prediction affected by the lead time and vulnerability threshold?
3. Can the interpretability of the classifiers be assessed and improved using feature importance, i.e., can the feature set be simplified while retaining accuracy to support real-world applications?

The proposed methodology is intended as a tool to support adaptive planning under a changing climate, specifically by using these early warning signals to inform monitoring and detection of vulnerabilities.

2. Methods

A flowchart of models and datasets used in this experiment is shown in Fig. 1. The experiment is demonstrated using a case study of the northern California reservoir system, a large network of storage and conveyance infrastructure designed primarily to move winter precipitation from north to south to support summer irrigation. The largest consumptive use is represented by the 7.9 million acres of irrigated farmland (generating $100 billion annually in agricultural production), followed by urban use by California's 39.7 million residents (Johnson and Cody, 2015). This system complexity yields many options for adaptation, but is complicated by the fact that it is managed by hundreds of distinct agencies, utilities, and districts, making coordination difficult

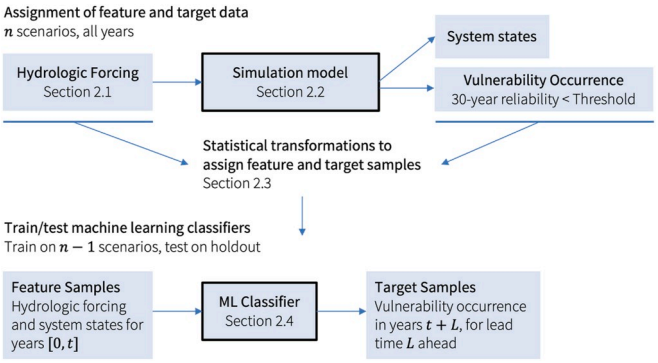

Fig. 1. Methods flowchart. In the first step, a reservoir simulation model is used to develop the feature and target data for all years throughout the century in all climate scenarios. In the second step, these samples are used to train and test machine learning classifiers for early warning of system vulnerability.

(Hanak and Lund, 2012). While the effects of climate change on average precipitation in California remain uncertain, projections agree on increases in the frequency of both wet and dry extremes, as well as the increased likelihood of these extremes occurring sequentially (Swain et al., 2018), placing more stress on the long-term planning and operation of water supply storage.

2.1. Data sources

Precipitation and temperature data are taken from an ensemble of downscaled CMIP5 projections publicly available from the U.S. Bureau of Reclamation (Brekke et al., 2014). This daily timestep dataset contains hydrologic projections (streamflow and snowpack) for multiple point and gridded locations created by routing precipitation and temperature through the Variable Infiltration Capacity (VIC) hydrologic model (Liang et al., 1994). Ensemble projections are available for 31 global climate models (GCMs) and four RCPs as shown in Supplemental Table S1. Fig. 2 shows the locations of streamflow (11), snowpack (4), and precipitation and temperature (3) gages used.

2.2. Simulation model

The Operation of Reservoirs in California (ORCA) model (Cohen et al., Accepted) is used to simulate the operation of the northern California reservoirs under all 97 CMIP5 climate scenarios on a daily timestep over the period 2000–2100. Given the input data shown in Fig. 2, the model simulates the operations of Shasta, Oroville, and Folsom reservoirs, which are located respectively on the Sacramento, Feather, and American Rivers. In addition to reservoir management, the model also simulates the operations of South-of-Delta exports (via pumping) from the Sacramento-San Joaquin Delta to meet urban and agricultural demands via the Central Valley Project (CVP) and State Water Project (SWP), while also meeting environmental flow and salinity requirements for the Delta. For the purposes of this study, the supply reliability of Delta exports is the key model output that determines system vulnerability. The model has been found to adequately reproduce historical operations of the system on a daily timestep, with Nash-Sutcliffe Efficiency above 0.9 for reservoir storage. Model code and documentation can be found at https://github.com/jscohen4/orca. More details about the origin, use, and locations of each type of data described above can be found in Supplemental Table S2.

2.3. Feature and target data

The classification problem is to predict system vulnerability at a certain lead time, given a set of feature variables. Vulnerability is determined based on the 30-year moving average of the supply

Fig. 2. Locations in northern California of the CMIP5 hydroclimatic projections used in this study, which include precipitation, temperature, streamflow, and snowpack (Brekke et al., 2014). Snowpack and precipitation values are spatially averaged over their respective basins (outlined in white), while streamflow and temperature data are taken at point locations. Water exports from the Sacramento-San Joaquin Delta (black arrow) are delivered throughout the southern half of the state for urban and agricultural uses. The supply reliability of these water deliveries are the focus of potential future vulnerability considered in this study.

reliability of water exports from the Delta, a metric that reflects a focus on identifying long-term trends rather than a single drought period. This metric is based on meeting the target demand for a certain fraction of monthly timesteps. Vulnerability occurs when the 30-year average supply reliability falls below a chosen threshold, resulting in a single binary target metric for classification. While this approach does not distinguish between different magnitudes of vulnerability, we adopt this binary classification following the standard for scenario discovery methods in the water resources field. This study tests a range of possible threshold values to understand the impact on classifier performance, recognizing that vulnerability definitions in practice are determined by decision makers. The threshold value plays a key role in determining the balance of positive and negative classifications in the training set.

The feature variables used to classify future vulnerabilities each year include time series of the hydrologic variables shown in Fig. 2, as well as several internal states of the simulation model. These variables are aggregated to annual and monthly values using either the mean, maximum, or sum. They are then translated into moving averages and standard deviations using timescales of 10, 20, 30, 40, and 50-year rolling windows. This differentiation by timescale and statistic is conducted to improve the ability of the machine learning methods to detect important trends in the data (Ahmed et al., 2010). Given that many of the moving windows share overlapping data, several of the features are expected to be correlated. A summary of feature variables is shown in Table 1.

2.4. Machine learning methods

Several classification methods are selected (Table 2) to learn from the annual and monthly data to classify water supply reliability as vulnerable (below a chosen threshold) or not vulnerable (above a chosen threshold) at lead times of 0, 1, 5, 10, and 20 years. These methods were chosen from the many classification methods available in the open source *scikit-learn* library (Pedregosa et al., 2011; Scikit-Learn, 2019) based on their widespread use and demonstrated effectiveness for nonlinear problems. By examining the performance of each classifier against the others and a random classifier, which serves as a baseline, application-specific insights can be drawn about the similarity of their performance and the likelihood of misclassifications. The random classifier guesses proportional to the ratio of possible outcomes with an accuracy equal to the square of the ratio of possible outcomes.

All methods in Table 2 are implemented using default parameter settings from the *scikit-learn* library. It is recognized that these parameter choices can significantly impact the performance, and that a meta-level analysis would be needed to determine the optimal parameter settings.

Each of the classification methods follows the same prediction structure:

Table 1
Summary of feature variables, including how they were aggregated, their timescales, their transformations, and the lead times applied to them. A total of 500 features are considered. Each feature variable encompasses 9603 observations (99 years of data x 97 scenarios).

Variables	Summary	Aggregations	Timescales	Transformations
Temperature (min, mean, max), Storages, X2 Salinity	Mean	Monthly and Annually	10, 20, 30, 40, 50 (-year windows)	Rolling mean, Rolling standard deviation
Streamflow, Precipitation, Demand, Pumping, Outflows, Inflows, Shortages	Sum	Monthly and Annually	10, 20, 30, 40, 50 (-year windows)	Rolling mean, Rolling standard deviation
Snowpack	Max	Monthly and Annually	10, 20, 30, 40, 50 (-year windows)	Rolling mean, Rolling standard deviation

Table 2
Summary of classification methods tested in this study.

Method	How classification is determined
K-Nearest Neighbors	Majority vote of the K nearest points in the training set
Logistic Regression	Fit a logistic function to binary data; round prediction to 0 or 1
SVM (3rd Degree)	Fit a decision boundary using cubic polynomial kernels for each point
Random Forest (Breiman, 2001)	Train ensemble of decision trees, use majority vote as the prediction
Multi-layer perceptron	Fit an arbitrary nonlinear decision boundary with a multi-layer neural network
AdaBoost (Freund and Schapire, 1996)	Ensemble of classifiers that trains new copies iteratively by increasing the weights of incorrectly classified points. Default estimator is decision tree.
Naïve Bayes	Maximum likelihood classification using Bayes' theorem assuming normality and independent features given the output class.
Random	Guesses proportional to the ratio of possible outcomes

$$F(\mathbf{X_t}) = \begin{cases} Vulnerable, & P_{t+lead} < threshold \\ Not\ Vulnerable, & P_{t+lead} > threshold \end{cases}$$

where F is the fitted function embedded in each of the classification methods, using the features $\mathbf{X}_t$to evaluate the function and make predictions. If the prediction, P, at the specified lead time is less than the threshold, the instance is classified as vulnerable; otherwise, the instance is classified as not vulnerable. The classification methods in Table 2 differ primarily in the family of functions used to represent F. More recently developed algorithms are denoted with citations in Table 2, while the other fundamental methods can be referenced in Hastie et al. (2009). Importantly, this classification approach is not meant to emulate the reservoir system model itself. It is predicting based on a combination of hydrologic and system observations whether it is likely to be on a trajectory toward a long-term vulnerable state.

2.5. Experimental design

Next, each method is fit to the training data (lagged predictors and binary targets) and re-evaluated against held-out validation data, using a repeated leave-one-out approach. From the ensemble of CMIP5 scenarios, one scenario is held out for testing while the classifier is trained on the remaining 96 scenarios. The process is then repeated for all scenarios. This design provides several benefits: the evaluation scenario is hidden from the classifier during training; the temporal structure of the data is preserved, which prevents biased training or testing with either too many values from the beginning or the end of the century; and it represents the realistic case in which the future hydrologic forcing is unknown (e.g. to evaluate in a systems model) but where recent observations can be used to make a statistical prediction. In the next section, all results will be reported in the validation stage, using the ensemble of leave-one-out experiments to estimate confidence intervals for the prediction accuracy.

In summary, this experiment tests seven classification methods, five lead times ranging from 0 to 20 years, vulnerability thresholds ranging from 0.60 to 0.86 (the full range in which both positive and negative classifications are possible), and three numbers of features (5, 10, and 500) as described in the following paragraph. In each of these cases the machine learning method classifies each prediction as either a true positive ($TP_i(t)$), false positive ($FP_i(t)$), false negative ($FN_i(t)$), or true negative ($TN_i(t)$). The primary metrics used to analyze classifier accuracy are the true positive and negative ratios, which are the fraction of possible positive/negative outcomes that are correctly predicted. The number of possible outcomes in each class is determined by counting the occurrences in the test set. The results from each of the combinations are evaluated against the baseline to test the null hypothesis that the accuracy of the machine learning classifiers is no better than random. If the null hypothesis is rejected ($p < 0.05$), then the accuracy of the machine learning classifiers is significantly better than random. We consider this the absolute minimum standard to evaluate the practical utility of the approach.

Finally, the original set of feature variables is too large for some of the classifiers to converge. The set is reduced using feature importance scores, determined based on the frequency of occurrence of each feature in a Random Forest of 10,000 trees (a default method from the *scikit-learn* library), with more occurrences corresponding to a higher importance. The 500 features with the highest importance scores are used in the training step. Before training, all feature and target data are scaled to unit variance. Additional cases are considered in which the feature set is reduced to 5 and 10 features, again based on the importance scores from the Random Forest method.

Fig. 3 shows conceptually how a classifier attempts to predict at every time step whether the long-term water supply reliability of the system will be vulnerable (below the threshold) or not vulnerable (above the threshold) at a given lead time, which in this example is 10 years. Time $t + L$ is the year the scenario becomes vulnerable, and only the five most informative features are shown. The gray dotted lines show information that is not available to the classification methods when they are making a prediction at time t years. The long-term vulnerability is always based on the 30-year trailing average water supply reliability, regardless of the lead time at which the prediction is made.

3. Results

3.1. Classifier accuracy as a function of threshold value

The ability to detect early warning signals of water supply vulnerability depends in part on how frequently these events occur in the training data, which is controlled by the threshold value, as well as the strength of the climate change signal relative to noise in each of the scenarios. Fig. 4 compares true positive and true negative ratios for vulnerability thresholds between 0.6 and 0.86 for a fixed lead time of five years. Each classification method contains the median, 10th, and 90th percentiles across the ensemble of leave-one-out experiments. The performance of a random classifier is included as a benchmark (black line), which always guesses proportional to the ratio of possible outcomes with an accuracy of the ratio squared. While all methods perform similarly, the Naïve Bayes Classifier shows a slight advantage, with only one median value below 0.8 for both true positives and true negatives.

Depending on the ratio of possible positives and negatives (dotted

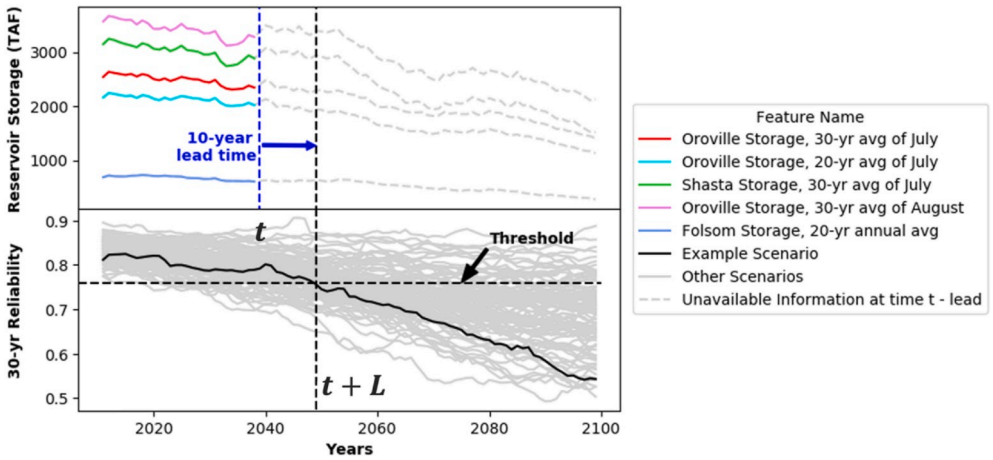

Fig. 3. Conceptual overview of the classification problem carried out each year. Time $t + L$ is the year the scenario becomes vulnerable, and the classification methods are attempting to predict that occurrence at a lead time of $L = 10$ years using the available feature information.

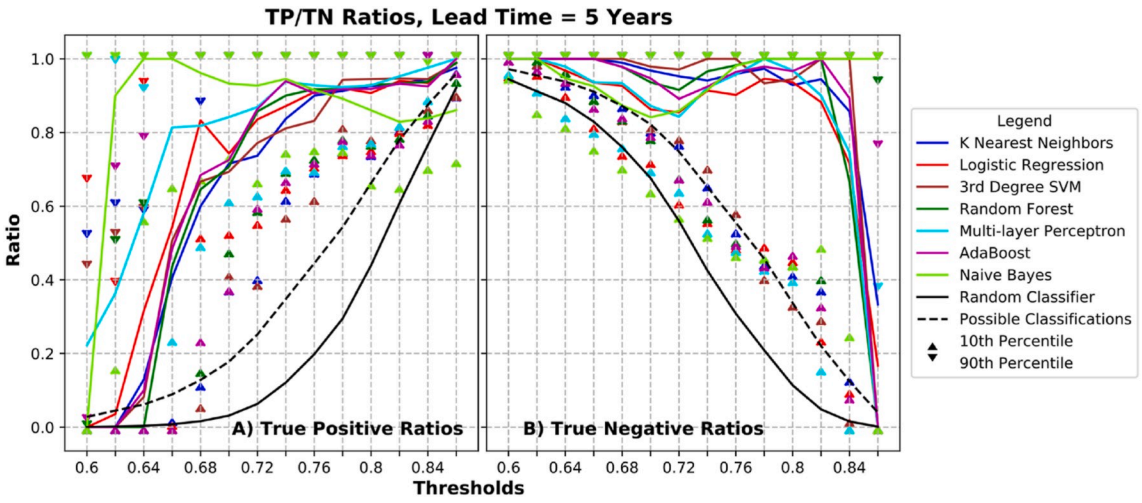

Fig. 4. (A) True positive and (B) true negative ratios for different vulnerability thresholds (0.6–0.86) for the classification methods for a 5-year lead time. The median of the validation ensemble is shown as a solid line, while the 10th and 90th percentiles are shown in colored triangles corresponding to each method. The random classifier benchmark is shown in black, and the ratios of possible positive and possible negative classifications are shown as dotted lines on each subplot.

line), imbalance in the training data may lead to low rates of true classifications, with the best performance near the center of the range of possible thresholds (0.4–0.6). This is reflected with the benchmark random classifier, showing that at unbalanced ratios (thresholds below 0.66 or greater than 0.84), some methods fail to perform better than random. Understanding the effects of the training set imbalance can inform whether true positive or true negative classifications are more likely to be accurate, a well-known challenge for machine learning methods. In water resources applications, the choice of the vulnerability threshold is left to the decision maker and cannot be selected arbitrarily to ensure an accurate classifier. However, a decision maker may choose a different classifier threshold to balance the tradeoff between false positives and false negatives to increase the effectiveness of the early warning signal in a real-world institutional context. This analysis underscores that if vulnerabilities are rare among the set of climate projections tested, then by definition it will be difficult to train a machine learning model to predict them, and that the estimates of prediction accuracy should be accompanied by confidence intervals to improve their interpretability for stakeholders.

In addition, while the median true positive and true negative ratios suggest substantial improvement over the random classifier benchmark, the 10th percentiles indicate the lower range of performance across the validation ensemble, and in particular sometimes fail to outperform the random classifier. The 90th percentile markers generally show an accuracy of 1.0 excepting some classifiers for thresholds below 0.70 (for true positive ratios) and above 0.84 (for true negative ratios). We return to the question of statistical significance in Section 3.3. Fig. 4 only shows the true positive and true negative ratios for a single lead time (5 years). Similar figures for other lead times can be found in Supplemental Fig. S1.

3.2. *Classifier accuracy as a function of lead time*

Accuracy was also evaluated across lead times for a fixed threshold of 0.76 (Fig. 5). In general, the median accuracy ratios decrease with lead time, with the exception of the true positive ratio for the AdaBoost classifier. The confidence intervals (triangles) suggest that all classifiers generally outperform the random classifier, except for the true negative ratios at a 20-year lead time, suggesting a lack of skill at the lower end of the validation ensemble. The spread of the confidence intervals

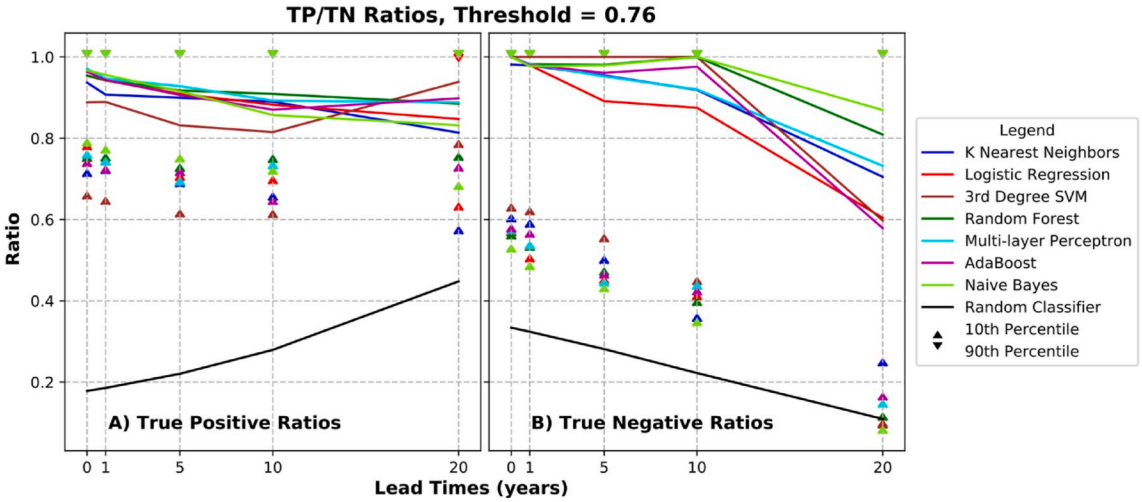

Fig. 5. The median (A) true positive and (B) true negative ratios compared to lead times for a threshold of 0.76 for all of the classification methods. The benchmark showing the accuracy of a random classifier is shown in black on each subplot. Colored markers show the 10th and 90th percentiles for each of the methods based on the leave-one-out ensemble testing.

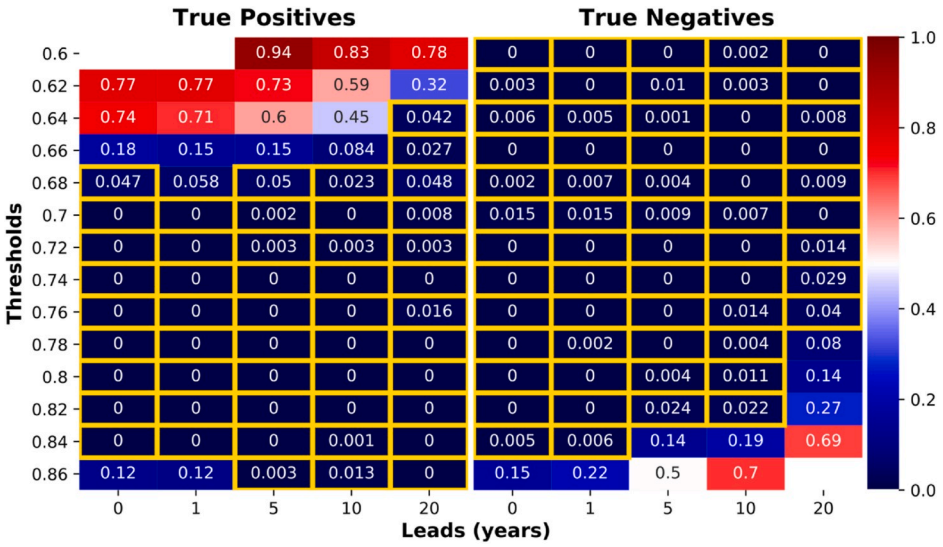

Fig. 6. Heatmap for the Random Forest classifier, showing (A) true positive and (B) true negative p-values corresponding to each lead time and each threshold value. All p-values less than or equal to 0.05 are highlighted in yellow, which indicate the conditions for which the Random Forest classifier outperforms the random classifier. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

increases with lead time, though this observation is most apparent for the true negative ratios. The 90th percentiles of all distributions fall at or near an accuracy of 1.0 across all lead times. Finally, the true positive ratio of the random classifier (black) increases with longer lead times due to the increase in possible positive outcomes further ahead in the century. The practical implication of this is to raise the standard of performance needed for the other algorithms to outperform the random benchmark at longer lead times.

3.3. Statistical significance

To determine whether to reject the null hypothesis that the accuracy of a machine learning classifier is equal to that of a random classifier, the p-values for all combinations of lead times and thresholds must be examined. By rejecting the null hypothesis, the alternative hypothesis (the accuracy of the machine learning classifier is better than a random classifier) will be accepted. The null hypothesis will be rejected for a particular combination of lead time and threshold value with $p \leq 0.05$.

Fig. 6 shows a heatmap of p-values for true positives and true negatives as a function of both lead time and threshold value for the Random Forest classifier. Similar heat maps showing results for the other classifiers can be found in Supplemental Figs. S2 through S7. The p-values are determined by the percentile of the leave-one-out distribution that falls below the accuracy of the random classifier, indicating the likelihood of the classifier performing worse than random. For many combinations of threshold and lead time, the classifier performs significantly better than random ($p \leq 0.05$). Fig. 6 also suggests that threshold values have a larger impact on accuracy than lead times, likely driven by the ratio of possible positive to possible negative classifications. Only a narrow range of threshold values (0.70–0.76) exists in which both the true positive and true negative classifications are significantly better than random for all lead times.

3.4. Feature importance

To reduce the number of features used to classify vulnerability, the importance of each feature must be determined. The top five features for each lead time are shown below (Table 3), ranked in the order of importance. The most important feature for all lead times is the 30-year moving average of Oroville reservoir storage, which appears as either the first or second ranked feature for each of the lead times. The 30-year average of Shasta Storage is the second most common feature in Table 3. However, for a 20-year lead time, the most important feature is the annual maximum air temperature at Folsom Dam, which likely reflects the longer-term temperature trends associated with climate change. The most common months in the important features are July, August, and June, which are the drier months in California and have the potential to carry important signals in a water system in which intra-annual water storage is vital. These influential features are also highly correlated with each other (see Supplemental Fig. S8).

Table 3
The five most important features for each lead time. Each feature is labeled first by the month, *M*, it corresponds to (*ANN* for annual, M01 for January, M02 for February, etc.), then by the rolling window (10, 20, 30, 40, or 50-year windows) used to calculate the metric (AVG for average, SD for standard deviation), and finally by the location and type of feature (e.g., Oroville Storage).

Rank	0-yr Lead	1-yr Lead	5-yr Lead	10-yr Lead	20-yr Lead
1	M07 30-yr AVG Oroville Storage	M07 30-yr AVG Oroville Storage	M07 30-yr AVG Oroville Storage	M07 20-yr AVG Oroville Storage	ANN 10-yr AVG Folsom T_{max}
2	M06 30-yr AVG Shasta Storage	M08 30-yr AVG Oroville Storage	M07 20-yr AVG Oroville Storage	M08 20-yr AVG Oroville Storage	M08 20-yr AVG Oroville Storage
3	M08 30-yr AVG Oroville Storage	M07 30-yr AVG Shasta Storage	M07 30-yr AVG Shasta Storage	M07 20-yr AVG Shasta Storage	ANN 10-yr AVG Folsom T_{avg}
4	M07 30-yr AVG Shasta Storage	M06 30-yr AVG Shasta Storage	M08 30-yr AVG Oroville Storage	M06 20-yr AVG Shasta Storage	M07 20-yr AVG Oroville Storage
5	M11 30-yr AVG Total Shortage	ANN 30-yr AVG Total Shortage	ANN 20-yr AVG Folsom Storage	M09 20-yr AVG Folsom Storage	ANN 20-yr AVG Folsom T_{max}

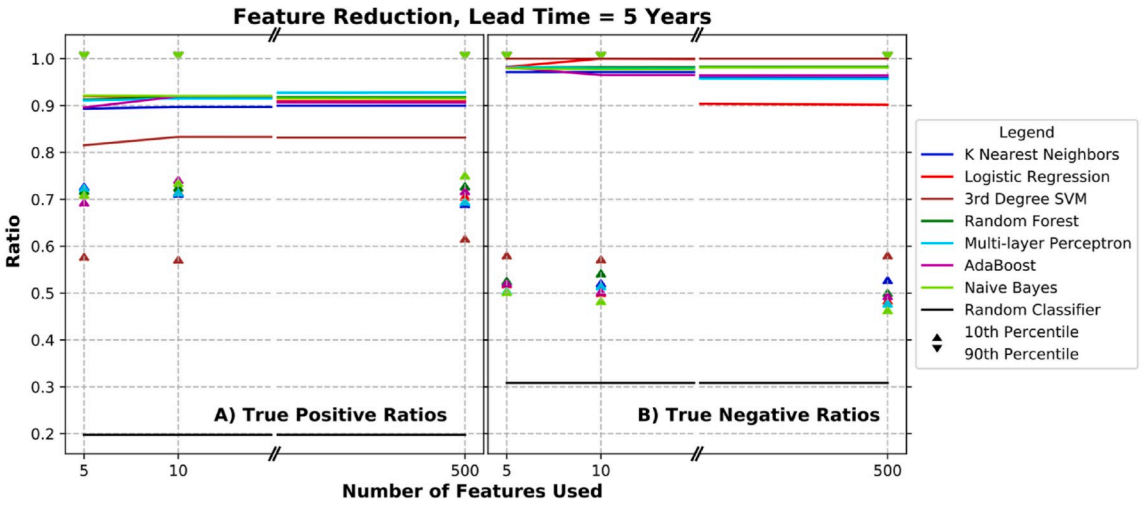

Fig. 7. Comparing the change in median (A) true positive and (B) true negative ratios to the number of features used for the classification methods for a 5-year lead time and a threshold of 0.76. The 10th and 90th percentiles are shown as triangles to better represent the distribution of the results, and a benchmark showing the accuracy of a random classifier is in black.

The finding that summer reservoir storage dominates the ability to detect early warning signals is not surprising, but also is not obvious. The target prediction is not water supply vulnerability in a given year, but rather a long-term trend in the 30-year average water supply reliability. This result suggests that the role of reservoir storage in integrating different aspects of the hydrologic cycle also make it a good indicator for future change, provided that the system operations remain the same as assumed in this study. This can also be interpreted in light of the storage-to-inflow ratios of each reservoir, which are approximately 0.80, 0.93, and 0.43 for Shasta, Oroville, and Folsom reservoirs, respectively.

Fig. 7 shows the effect of reducing the number of features used with the machine learning methods given a threshold of 0.76 and a 5-year lead time, where the number of features are prioritized according to their relative importance using the random forest method (Table 3). Across all methods, the true negative ratios are higher (above 0.95 from 5 to 10 features and above 0.9 for 500 features) than the true positive ratios (ranging between 0.8 and 0.95 for all features). In general, most ratios have a slight increase in performance from 5 to 10 features. Performance does not show significant changes between 10 and 500 features except for the Logistic Regression true negative ratio, which falls from about 1.0 to 0.9. The 10th percentiles range from 0.57 to 0.76 for the true positive ratios and range from 0.45 to 0.6 for the true negative ratios, both outperforming the random classifier. Overall, these results imply that the number of features can be reduced from 500 to 5 with only small reductions, if any, in the true positive and true negative ratios, due to the high correlation among features with overlapping rolling windows. In general, feature importance can be linked to the signal-to-noise ratio of each feature: variables that change more slowly, such as the storage of large reservoirs or the annual temperature, likely provide more reliable signals than observations with a more variable response to climate forcing. The reduced complexity of this problem will improve opportunities for practical application. Feature reduction figures for lead times of 0, 1, 10, and 20 years can be found in the supplemental material (Fig. S9).

3.5. Accuracy over time

The previous results consider the true positive and true negative ratios for different parameters, aggregated over the entire century. Fig. 8 shows the true positive and true negative ratios as they change over the century for classifiers trained on the full time period using a threshold of

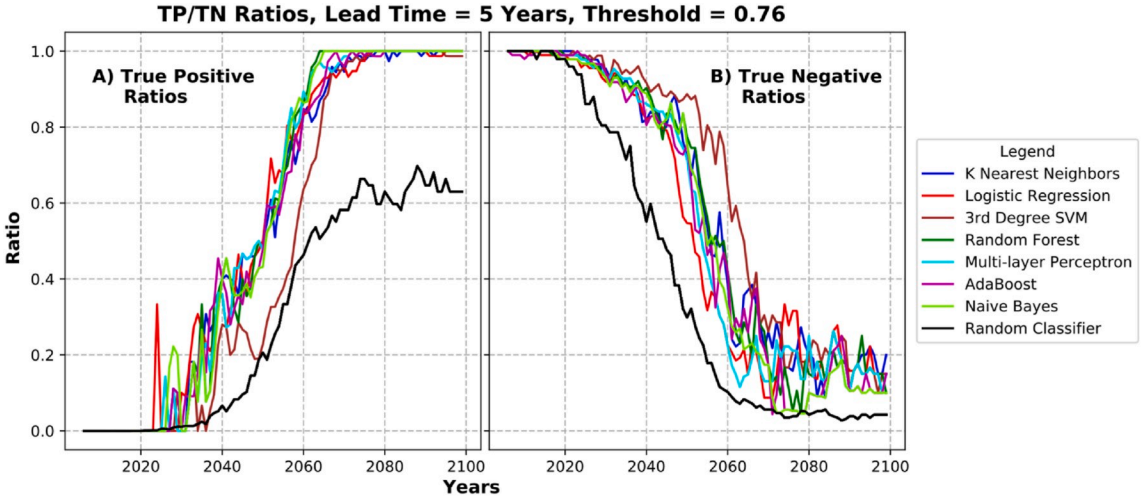

Fig. 8. The ratio of (A) true positives and (B) true negatives for each of the classification methods throughout the 21st century for a 5-year lead time and a threshold of 0.76. A benchmark showing the accuracy of a random classifier is in black.

0.76 and a 5-year lead time (see Supplemental Fig. S10 for 0, 1, 10, and 20 -year lead times). The random benchmark classifier for each year is also shown. In general, the true positive ratios become more accurate throughout the century, while the true negative ratios become less accurate. Importantly, the classifier is not being retrained over time, only applied to new data, which explains the decreasing true negative rate. A higher true positive ratio later in the century means that it is easier to correctly make a vulnerable (positive) classification later in the century, when many of the features show stronger climate change signals. Most of the methods perform better than the random classification benchmark, except for some true positive ratios before 2040 and some true negative ratios before 2020.

In summary, results suggest that the machine learning classifiers outperform the random classifier benchmark for most lead times under thresholds with a balanced ratio of training outcomes, for a reduced set of features, and for most years throughout the century. The classifications generally do not show statistically significant differences in performance (Figs. 4, 5 and 7), though this finding may not generalize to other applications. Additionally, only a few influential features (mostly reservoir storage variables and their transformations) are responsible for the predictions. This may indicate that the reservoir storage values are able to uniquely aggregate input information given that the other features are either influencing, or influenced by, reservoir storage.

4. Discussion

These experiments have analyzed the predictive skill of machine learning classifiers trained to detect future water supply vulnerabilities under climate change. The analysis has therefore considered several of the goals of a monitoring system (Haasnoot et al., 2018) by exploring how the reliability and observability of early warning signals change at different levels of timeliness and vulnerability. However, the remaining goals of credibility and legitimacy have not been quantified here, and it is recognized that the use of machine learning classifiers rather than linear thresholds will hinder the interpretability of this approach for stakeholders. The attempt to improve interpretability in this study depends on feature importance (Table 2) to prune the set of input variables, which may support the parsimony of a monitoring system (Raso et al., 2019b). Additionally, the logic of the dominant features is demonstrated within the context of the system: reservoir storage integrates hydrologic and demand dynamics over time, and therefore provides the most reliable signal of vulnerability. Interpretability may be further improved by monitoring a continuous variable from the classifier (such as the class probability) rather than the binary prediction, which might provide a more reliable signal of change similar to the p-value detection method proposed by Haasnoot et al. (2018). In general, explainability is a rapidly advancing area of machine learning (Doshi-Velez and Kim, 2017; Xie et al., 2020) that will likely yield developments to support environmental systems analysis in the coming years.

Classifier skill strongly depends on the extent to which the training data reflects the range of possible future scenarios. This is true of any machine learning problem, and arises in two key aspects of this study. First, unbalanced training data cause difficulty in classifying positive and negative outcomes. This may be amplified by relatively small sample sizes in the training set, with significant implications for water resources planning under climate extremes. The second challenge, more specific to this problem, is that of deep uncertainty in the climate scenarios. It is entirely possible that the future hydrology will depart significantly from the training data due to a combination of model uncertainty, emissions scenarios, and natural variability. This study employs a leave-one-out training and validation strategy to partially account for potential bias by testing whether the classifier can generalize to (1) other realizations of a similar uncertainty characterization, and (2) other GCM and RCP combinations with different uncertainty characterizations. However, good out-of-sample performance is perhaps less reassuring here than in typical machine learning problems relying on large datasets with well-characterized uncertainty. A more complex validation approach could consider alternate ensembles generated with different climate models, or expert judgment of bias in the training data. As in any study of deeply uncertain futures, the findings are contingent on the inherently subjective design of the training and validation experiments.

This study is only partially linked to a specific decision context: it aims to analyze the range of timescales and vulnerability thresholds over which reliable prediction of water supply vulnerability might be possible. A real-world decision context would also include the adaptations to be selected when detection occurs (a subject of ongoing work), as well as the necessary timescales for each. For example, water conservation and regulation may benefit from information on annual or sub-annual lead times, while infrastructure may require a decade or more. The findings are also specific to the range of uncertainty demonstrated in the water supply projections for this system, which are quite large—nearly 50% change in mean annual flows by the end of the century, arising from a combination of GCM and emissions uncertainty. Even so, we do not achieve a complete representation of all sources of uncertainty in the early warning system (Raso et al., 2019a), particularly the endogenous uncertainties arising from changes to system operations or water demand. Additionally, we do not attempt to evaluate how stakeholders learn from monitoring information in their decision making process (Hermans et al., 2013), or the extent to which stakeholders with different problem framings (Hermans et al., 2017; Quinn et al., 2018) may find the early warning signals convincing within their system of organizational decision-making (Haasnoot et al., 2018). Much interesting work remains at this intersection of statistical modeling and policymaking for "wicked" problems that by definition do not lend themselves to straightforward prediction (Rittel and Webber, 1973; Kwakkel et al., 2016).

5. Conclusions

This paper contributes a methodology for detecting early warning signals of water supply vulnerabilities under climate change using machine learning, demonstrated on a case study of the northern California reservoir system. Among the many goals of a monitoring system (relevance, credibility, and legitimacy) proposed by Haasnoot et al. (2018), this study has primarily focused on relevance, represented by the predictive skill of detecting future change. Results indicate that the classification methods generally outperform a benchmark random classifier, though the factor most strongly influencing this result is the balance of the training data determined by the vulnerability threshold. In addition, the overall classification accuracy decreases with larger lead times. To improve the interpretability and parsimony of the resulting classifiers (Raso et al., 2019b), the feature set can be reduced with minimal impact on accuracy due to high correlation between features at short lead times. The features most strongly influencing the predictions are long-term averages of summer reservoir storage, which demonstrates predictive power in the ability of storage to integrate different aspects of the hydrologic cycle.

With further work to analyze the credibility and legitimacy of this approach in a real-world decision context with significant human and institutional uncertainties (Hermans et al., 2013, 2017), this approach could be implemented as a tool to support water resources planning under climate uncertainty. An additional limitation is the assumption that the system infrastructure and operations remain unchanged throughout the century; the trained classifiers are expected to become less accurate over time as a result of endogenous adaptation, a topic of ongoing work. However, even when ensemble climate projections suggest substantial uncertainty in future hydrology, this approach can help to identify what signals should be monitored to inform adaptation. While this study has developed vulnerability classification methods in line with previous work on scenario discovery, future work will consider

regression methods to identify the magnitude of failure as well. Additional research will focus on integrating these dynamic vulnerability classifications with an adaptive infrastructure planning problem, where early warning signals can be used directly to trigger decisions. This analysis will provide insights into the benefits of predicting vulnerabilities along with the consequences of inaccurate classifications, including the costs of unnecessary adaptations and the regrets of foregoing beneficial ones.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements and Software Availability

This work was partially supported by the U.S. National Science Foundation grants CNH-1716130 and CBET-1803589. Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views or policies of the NSF. All code and data used in this study is available on GitHub (https://github.com/brobinson3/Early_Warning_Signals_ML_ORCA). We further acknowledge the World Climate Research Program's Working Group on Coupled Modeling and the climate modeling groups listed in the Supplement of this paper for producing and making available their model output.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2020.104781.

References

Ahmed, N.K., Atiya, A.F., El Gayar, N., El-Shishiny, H., 2010. An empirical comparison of machine learning models for time series forecasting. Econom. Rev. 29 (5), 594–621. https://doi.org/10.1080/07474938.2010.481556.

Borgomeo, E., Mortazavi-Naeini, M., Hall, J.W., Guillod, B.P., 2018. Risk, robustness and water resources planning under uncertainty. Earth's Future 6 (3), 468–487. https://doi.org/10.1002/2017EF000730.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brekke, L., Wood, A., Pruitt, T., 2014. Downscaled CMIP3 and CMIP5 Hydrology Climate Projections: Release of Hydrology Projections, Comparison with Preceding Information, and Summary of User Needs. US Bureau of Reclamation.

Bryant, B.P., Lempert, R.J., 2010. Thinking inside the box: a participatory, computer-assisted approach to scenario discovery. Technol. Forecast. Soc. Change 77, 34–49.

Carlsen, H., Dreborg, K.H., Wikman-Svahn, P., 2013. Tailor-made scenario planning for local adaptation to climate change. Mitig. Adapt. Strategies Glob. Change 18 (8), 1239–1255. https://doi.org/10.1007/s11027-012-9419-x.

Cohen, J., Zeff, H., & Herman, J. (Accepted). Adaptation of multi-objective reservoir operations to snowpack decline in the Western U.S. J. Water Resour. Plann. Manag.

Dietterich, T.G., 2002. Machine Learning for Sequential Data: A Review. Springer, Berlin, Heidelberg, pp. 15–30.

Doshi-Velez, F., Kim, B., 2017. Towards A rigorous science of interpretable machine learning. Retrieved from. http://arxiv.org/abs/1702.08608.

Fletcher, S.M., Lickley, M., Strzepek, K., 2019. Learning about climate change uncertainty enables flexible water infrastructure planning. Nat. Commun. 10 (1), 1782. https://doi.org/10.31223/OSF.IO/2TM7X.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning, pp. 148–156, 10.1.1.133.1040.

Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Software 62, 33–51. https://doi.org/10.1016/j.envsoft.2014.08.015.

Giuliani, M., Pianosi, F., Castelletti, A., 2015. Making the most of data: an information selection and assessment framework to improve water systems operations. Water Resour. Res. 51 (11), 9073–9093. https://doi.org/10.1002/2015WR017044.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (Mar), 1157–1182. Retrieved from. http://www.jmlr.org/papers/v3/guyon03a.html.

Haasnoot, M., Kwakkel, J.H., Walker, W.E., ter Maat, J., 2013. Dynamic adaptive policy pathways: a method for crafting robust decisions for a deeply uncertain world. Global Environ. Change 23 (2), 485–498. Retrieved from. http://www.sciencedirect.com/science/article/pii/S095937801200146X.

Haasnoot, M., van 't Klooster, S., van Alphen, J., 2018. Designing a monitoring system to detect signals to adapt to uncertain climate change. Global Environ. Change 52, 273–285. https://doi.org/10.1016/J.GLOENVCHA.2018.08.003.

Hadka, D., Herman, J.D., Reed, P.M., Keller, K., 2015. An open source framework for many-objective robust decision making. Environ. Model. Software 74, 114–129. https://doi.org/10.1016/j.envsoft.2015.07.014.

Hallegatte, S., 2009. Strategies to adapt to an uncertain climate change. Global Environ. Change 19 (2), 240–247. https://doi.org/10.1016/j.gloenvcha.2008.12.003.

Hallegatte, S., Shah, A., Lempert, R., Brown, C.M., Gill, S., 2012. Investment Decision Making under Deep Uncertainty: Application to Climate Change.

Hanak, E., Lund, J.R., 2012. Adapting California's water management to climate change. Climatic Change 111 (1), 17–44. https://doi.org/10.1007/s10584-011-0241-3.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning data mining, inference, and prediction. Retrieved from. https://books.google.com/books?hl=en&lr=&id=tVIjmNS3Ob8C&oi=fnd&pg=PR13&dq=hastie+elements&ots=ENJeR8K0U2&sig=v330svWtfeLRaE_2kixJn-uk2MQ.

Hawkins, E., Sutton, R., 2012. Time of emergence of climate signals. Geophys. Res. Lett. 39 (1) https://doi.org/10.1029/2011GL050087 n/a-n/a.

Hegerl, G., Zwiers, F., 2011. Use of models in detection and attribution of climate change. Wiley Interdis. Rev.: Clim. Change 2 (4), 570–591. https://doi.org/10.1002/wcc.121.

Herman, J.D., Reed, P.M., Zeff, H.B., Characklis, G.W., 2015. How should robustness Be defined for water systems planning under change? J. Water Resour. Plann. Manag. 141 (10), 4015012 https://doi.org/10.1061/(ASCE)WR.1943-5452.0000509.

Herman, Jonathan D., Giuliani, M., 2018. Policy tree optimization for threshold-based water resources management over multiple timescales. Environ. Model. Software 99, 39–51. https://doi.org/10.1016/j.envsoft.2017.09.016.

Herman, Jonathan D., Quinn, J.D., Steinschneider, S., Giuliani, M., Fletcher, S., 2020. Climate adaptation as a control problem: review and perspectives on dynamic water resources planning under uncertainty. Water Resour. Res. 56 (2) https://doi.org/10.1029/2019wr025502.

Hermans, L.M., Slinger, J.H., Cunningham, S.W., 2013. The use of monitoring information in policy-oriented learning: insights from two cases in coastal management. Environ. Sci. Pol. 29, 24–36. https://doi.org/10.1016/j.envsci.2013.02.001.

Hermans, L.M., Haasnoot, M., ter Maat, J., Kwakkel, J.H., 2017. Designing monitoring arrangements for collaborative learning about adaptation pathways. Environ. Sci. Pol. 69, 29–38. https://doi.org/10.1016/j.envsci.2016.12.005.

Hui, R., Herman, J.D., Lund, J.R., Madani, K., 2018. Adaptive water infrastructure planning for nonstationary hydrology. Adv. Water Resour. 118 https://doi.org/10.1016/j.advwatres.2018.05.009.

Johnson, R., Cody, B.A., 2015. California Agricultural Production and Irrigated Water Use Renée Johnson Specialist in Agricultural Policy Specialist in Natural Resources Policy.

Kwakkel, J.H., 2015. Exploratory modelling and analysis (EMA) workbench — exploratory modeling workbench. Retrieved from. http://simulation.tbm.tudelft.nl/ema-workbench/contents.html.

Kwakkel, Jan H., Walker, W.E., Haasnoot, M., 2016. Coping with the wickedness of public policy problems: approaches for decision making under deep uncertainty. J. Water Resour. Plann. Manag. 142 (3), 1816001 https://doi.org/10.1061/(ASCE)WR.1943-5452.0000626.

Lempert, R.J., Collins, M.T., 2007. Managing the risk of uncertain threshold responses: comparison of robust, optimum, and precautionary approaches. Risk Anal. 27 (4), 1009–1026. https://doi.org/10.1111/j.1539-6924.2007.00940.x.

Lempert, R.J., Bryant, B.P., Bankes, S.C., 2008. Comparing Algorithms for Scenario Discovery.

Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. 99 (D7), 14415. https://doi.org/10.1029/94JD00483.

Nayak, M.A., Herman, J.D., Steinschneider, S., 2018. Balancing Flood Risk and Water Supply in California : Policy Search Integrating Short-Term Forecast Ensembles with Conjunctive Use. American Geophysical Union. https://doi.org/10.1029/2018WR023177.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Polade, S.D., Gershunov, A., Cayan, D.R., Dettinger, M.D., Pierce, D.W., 2017. Precipitation in a warming world: assessing projected hydro-climate changes in California and other Mediterranean climate regions. Sci. Rep. 7 (1), 10783. https://doi.org/10.1038/s41598-017-11285-y.

Quinn, J.D., Reed, P.M., Giuliani, M., Castelletti, A., Oyler, J.W., Nicholas, R.E., 2018. Exploring How Changing Monsoonal Dynamics and Human Pressures Challenge Multireservoir Management for Flood Protection, Hydropower Production, and Agricultural Water Supply. Water Resources Research. https://doi.org/10.1029/2018WR022743.

Raso, L., Kwakkel, J., Timmermans, J., Panthou, G., 2019a. How to evaluate a monitoring system for adaptive policies: criteria for signposts selection and their model-based evaluation. Climatic Change 1–17. https://doi.org/10.1007/s10584-018-2355-3.

Raso, L., Barbier, B., Bader, J.C., 2019b. Modeling dynamics and adaptation at operational and structural scales for the ex-ante economic evaluation of large dams in an African context. Water Res. Econom. 26 https://doi.org/10.1016/j.wre.2018.08.001.

Rittel, H.W.J., Webber, M.M., 1973. Dilemmas in a general theory of planning. Pol. Sci. 4 (2), 155–169. https://doi.org/10.1007/BF01405730.

Robinson, B., Herman, J.D., 2019. A framework for testing dynamic classification of vulnerable scenarios in ensemble water supply projections. Climatic Change 152 (3–4), 431–448. https://doi.org/10.1007/s10584-018-2347-3.

Rosner, A., Vogel, R.M., Kirshen, P.H., 2014. A risk-based approach to flood management decisions in a nonstationary world. Water Resour. Res. 50 (3), 1928–1942. https://doi.org/10.1002/2013WR014561.

Rounsevell, M.D.A., Metzger, M.J., 2010. Developing qualitative scenario storylines for environmental change assessment. Wiley Interdis. Rev.: Clim. Change 1 (4), 606–619. https://doi.org/10.1002/wcc.63.

Scheffer, M., Carpenter, S.R., Lenton, T.M., Bascompte, J., Brock, W., Dakos, V., et al., 2012. Anticipating critical transitions. Science 338 (6105), 344–348. https://doi.org/10.1126/science.1225244.

Scheffer, Marten, Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, S.R., Dakos, V., et al., 2009. Early-warning signals for critical transitions. Nature 461 (7260), 53–59. https://doi.org/10.1038/nature08227.

Scikit-Learn, 2019. Supervised Learning (Version 0.21.3) Documentation.

Stephens, S.A., Bell, R.G., Lawrence, J., 2018. Developing signals to trigger adaptation to sea-level rise. Environ. Res. Lett. 13 (10), 104004 https://doi.org/10.1088/1748-9326/aadf96.

Swain, D.L., Langenbrunner, B., Neelin, J.D., Hall, A., 2018. Increasing precipitation volatility in twenty-first-century California. Nat. Clim. Change 8 (5), 427–433. https://doi.org/10.1038/s41558-018-0140-y.

Trenberth, K.E., Dai, A., Van Der Schrier, G., Jones, P.D., Barichivich, J., Briffa, K.R., Sheffield, J., 2014. Global warming and changes in drought. Nat. Clim. Change 4 (1), 17–22. https://doi.org/10.1038/nclimate2067.

Walker, W.E., Haasnoot, M., Kwakkel, J.H., 2013. Adapt or perish: a review of planning approaches for adaptation under deep uncertainty. Sustainability 5 (3), 955–979. https://doi.org/10.3390/su5030955.

Wilby, R.L., Dessai, S., 2010. Robust adaptation to climate change. Weather 65 (7), 180–185. https://doi.org/10.1002/wea.543.

Xie, N., Ras, G., van Gerven, M., Doran, D., 2020. Explainable deep learning: a field guide for the uninitiated. Retrieved from. http://arxiv.org/abs/2004.14545.

Zeff, H.B., Herman, J.D., Reed, P.M., Characklis, G.W., 2016. Cooperative drought adaptation: integrating infrastructure development, conservation, and water transfers into adaptive policy pathways. Water Resour. Res. 52 (9), 7327–7346. https://doi.org/10.1002/2016WR018771.