# Toward data-driven generation and evaluation of model structure for integrated representations of human behavior in water resources systems

**Liam Lynch**[1], **Jonathan D. Herman**[1]

[1]Department of Civil and Environmental Engineering, University of California, Davis, CA, USA

**Key Points:**

- Automated generation of model structure from data to describe human behavior in water systems.
- Systematic model evaluation along performance-complexity tradeoff by clustering models with similar behavior.
- Diagnostic assessment of model generalization skill using global sensitivity analysis of features.

Corresponding author: Liam Lynch, `wdlynch@ucdavis.edu`

## Abstract

Simulations of human behavior in water resources systems are challenged by uncertainty in model structure and parameters. The increasing availability of observations describing these systems provides the opportunity to infer a set of plausible model structures using data-driven approaches. This study develops a three-phase approach to the inference of model structures and parameterizations from data: problem definition, model generation, and model evaluation, illustrated on a case study of land use decisions in the Tulare Basin, California. We encode the generalized decision problem as an arbitrary mapping from a high-dimensional data space to the action of interest and use multi-objective genetic programming to search over a family of functions that perform this mapping for both regression and classification tasks. To facilitate the discovery of models that are both realistic and interpretable, the algorithm selects model structures based on multi-objective optimization of (1) their performance on a training set and (2) complexity, measured by the number of variables, constants, and operations composing the model. After training, optimal model structures are further evaluated according to their ability to generalize to held-out test data and clustered based on their performance, complexity, and generalization properties. Finally, we diagnose the causes of good and bad generalization by performing sensitivity analysis across model inputs and within clusters. This study serves as a template to inform and automate the problem-dependent task of constructing robust data-driven model structures to describe human behavior in water resources systems.


## 1 Introduction

Human behavior represents a significant source of uncertainty in simulation models of water resources systems (Konar et al., 2019), as humans interact with and depend on water systems in numerous ways (Lund, 2015). Examples include urban and agricultural water demand (Chini et al., 2017; Marston & Konar, 2017), population displacement (Müller et al., 2016), and the nonstationary behavior of decision-makers and regulatory institutions across multiple sectors and scales (Mason et al., 2018; Monier et al., 2018; Muneepeerakul & Anderies, 2020). Many different modeling approaches have been adopted for this problem, including: dynamical systems models, as in socio-hydrology (Sivapalan et al., 2012); hydro-economic models (Harou et al., 2009); and agent-based modeling (An, 2012). Each approach employs a structurally distinct perspective to link human decisions to the state of the hydrologic system (Schill et al., 2019). These approaches are not necessarily exclusive, and can be connected through a common experimental framing. Across all, the goal is to accurately describe observed dynamics of the system while managing the complexity of the spatial and temporal representation (Baumberger et al., 2017; Höge et al., 2018).

The increasing availability of multi-sectoral data describing water resources systems provides the opportunity to learn a set of plausible model structures using data-driven approaches (Brunton et al., 2016; Montáns et al., 2019). Data-driven methods are particularly useful for handling heterogeneous or unstructured data, and where existing theory may insufficiently explain available observations. In the latter case, however, care must be taken in the interpretation and application of the resulting models (Knüsel et al., 2019). There is growing interest in applying data-driven methods to calibrate parameters of integrated human-water models, such as smart-meter data (Cominola et al., 2019), water demand modeling (Oyebode et al., 2019), groundwater irrigation decisions (Hu et al., 2017), and water reservoir operations (Giuliani & Herman, 2018). While even simple data-driven models can sometimes outperform theory-driven models (Haughton et al., 2016), performance alone does not engender trust; model interpretability in the context of available theory is also needed to support both design and evaluation, though this may be limited in some systems (Baumberger et al., 2017; Lipton, 2018).

Several recent studies highlight the value and range of applications for data-driven approaches in water resources. For example, Giuliani et al. (2016) generate adaptive behavioral rules from historical climate and land use data by coordinating reservoir decisions with downstream cropping decisions from an economic model. Similarly, Quinn et al. (2018) employ policy emulation methods for coupled reservoir and irrigation decisions to reduce the computational cost of exploring a range of future hydroclimate scenarios. Worland et al. (2019) combine heterogeneous attributes of stream gauge networks to reconstruct observed flow duration curves under human influence with high accuracy using multi-output neural networks. Zaniolo et al. (2018) use data-driven variable selection across hydroclimate indicators and observed state variables to automatically design Pareto-optimal drought indices (i.e., constructing a function) to balance tradeoffs between complexity and performance. These studies have underscored the significant potential for data-driven methods to advance model design, while also identifying key challenges related to structure and complexity.

While data-driven approaches are adept at identifying parameters of a given model structure, there has generally been less focus on the identification of the structure itself, which is often not well-known (Blöschl et al., 2019). Model structural uncertainty arises from a lack of knowledge regarding the system, its behavior, and interactions between components (Walker et al., 2003). This is broadly the domain of data-driven system identification methods, which search both model structures and parameters to find candidate representations. System identification originated in the automatic control field to discover the components of interpretable mathematical models solely through data (black-box models), or by combining data with prior mechanistic knowledge about the system (gray-box models) (Ljung, 2017). Methods have been tested for systems in which the target relationships are known, such as the double pendulum (Schmidt & Lipson, 2009a) and the Navier-Stokes equations (Rudy et al., 2017). In hydrology, methods related to system identification have been applied for the general exploration of structural uncertainty in process-based modeling (Clark et al., 2015a, 2015b). Hydrologic studies have also considered data-driven approaches to system identification, such as the discovery of neural network structures for rainfall-runoff modeling (Hsu et al., 1995), the comparison of multiple regression methods for streamflow prediction (Wu et al., 2009), and the learning of transfer functions with symbolic regression (Klotz et al., 2017). Opportunities remain to leverage these developments for the identification of descriptive model structures of dynamic human behavior in water resources.

Several specific challenges arise in the process, as have been observed in hydrologic modeling where the question of structural uncertainty has been more widely studied (Young, 1998; Clark et al., 2008; Fenicia et al., 2011, e.g.,). First, data-driven system identification can result in many candidate models with varying levels of performance and complexity (Hogue et al., 2006; Bastidas et al., 2006; Pande et al., 2009). Second, additional criteria may be required for model evaluation (Beven & Freer, 2001; Höge et al., 2018; Eker et al., 2018), such as interpretability in the case of black-box models. For example, the introduction of deep learning methods into water resources has resulted in non-parsimonious models that often perform inexplicably well on unseen data (Shen, 2018). Conversely, data-driven system identification also allows for the testing of multiple model structures and parameterizations as competing hypotheses (Beven, 2019), often through search methods capable of adding complexity as needed. There remains a need to explore these challenges in the context of models of human behavior, where the goals of interpretability and parsimony apply simultaneously with the need for a broad spectrum of possible representations (Schill et al., 2019).

This study investigates the generation and evaluation of model structures for representations of human behavior for water resources systems. We propose a data-driven system identification approach to explore many candidate models as competing hypotheses. This approach operationalizes a preference for parsimonious model structures in combinatorial search spaces, along with the decomposition and diagnostic assessment of plausible model sets to determine driving structure. General modeling objectives are quantified at different phases of the experiment: generality, performance, complexity, and the importance of features and structural elements. This approach provides a foundation for future studies of model structural uncertainty and integrated systems modeling, particularly regarding the role of these issues in decision support for coupled human-water systems.

## 2 Methodological Background

We extend prior developments in environmental systems modeling to investigate structural uncertainty in models of human behavior through data-driven experimentation (Figure 1). This framing automates the identification and evaluation of plausible model structures within a general problem definition, quantifying a number of modeling objectives in the process. The phases presented here share similarities with the problem of constructing emulators (surrogates) of environmental systems models (Castelletti et al., 2012; Kleijnen, 2015). While system identification also seeks to generate models that accurately reproduce observed data, system identification has the additional goal of generating models that can support new understanding of the system.

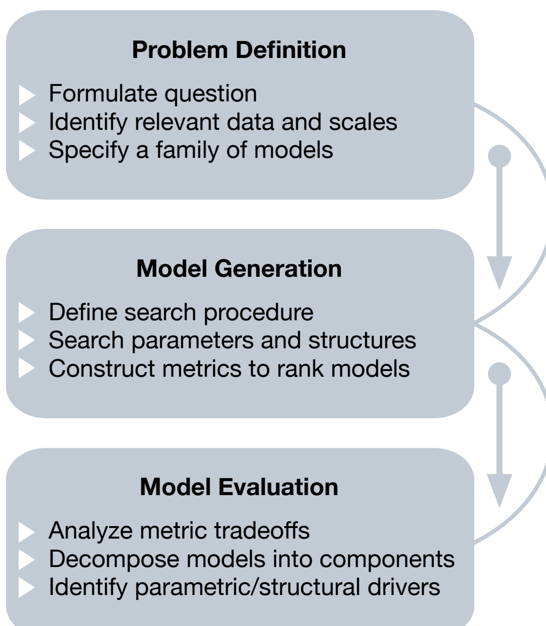


**Figure 1.** Flowchart of methodological steps involved in generating model structure from data.

### 2.1 Problem Definition

Problem definition for data-driven modeling includes the formulation of a question about the system, the collection and organization of available data at relevant spatial and temporal scales, and the specification of a family of models to answer the question. A data-driven system identification approach to problem definition can avoid human-intuited priors in the form of model structure and feature engineering, in favor of discovering useful constructions of both the data and the model simultaneously (Knüsel et al., 2019). First, the heterogeneous feature types common to integrated settings and observed human behavior can be considered across spatio-temporal scales. Feature engineering is then performed by transforming the observations, typically along with some form of dimension reduction such as eigenvalue decomposition (Giuliani & Herman, 2018). Variables at incongruent spatial and temporal scales and categorical variables can also be incorporated, for example through encoding schemes (Cerda et al., 2018).

In formulating the question, the model $\phi$ must be identified to map predictor variables $X$ (input samples) to the response variable $y$ in a multivariate regression problem: $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^1$. For modeling dynamical systems, the problem might involve learning the next state or derivative of a state variable in time given the current and previous states. The goal is to automatically reverse-engineer structure in $\phi$ that enables novel insights of the system (Bongard & Lipson, 2007). Discovering the optimal $\phi$ without pre-specifying the form of the function invokes exploration over both the structure and parameterization of $\phi$. This multivariate regression problem, an instance of supervised learning, can also be used in a control context, or to learn the behavior of agents in an environment given stochastic, noisy rewards (Barto & Dietterich, 2012).

There are a number of model families from which functions could be drawn to perform this mapping, such as linear additive models or neural networks. Functions can be most generally encoded as trees or graphs, either of which can be used to represent a universal approximator (e.g. Breiman, 2001; Huang et al., 2006) of highly complex, nonlinear human behavior. A common approach for the automatic construction of models of arbitrary mathematical structure and complexity is to combine objects from a primitive set of basic functions. As an instance of a process influencing the natural system, human behavior is integrated in model computation graphs, the network representing model operations and numerical fluxes (Gupta & Nearing, 2014; Khatami et al., 2019), by defining representational nodes and specifying links. Taken together, nodes and links in a model's graph form a natural measure of model integration (Claussen et al., 2002).

### 2.2 Model Generation

Model generation requires a search procedure over model parameters and structure, with performance represented by one or more metrics such as accuracy and complexity. Relatively few studies in the water resources field have considered an optimization over model structures, and most of these focus on normative rather than descriptive modeling. Many of these studies come from applications of data-driven methods to direct policy search (Rosenstein & Barto, 2001; Giuliani et al., 2014). For example, Herman and Giuliani (2018) test operating rule structures via the optimization of binary trees using genetic programming. In general, heuristic methods such as evolutionary algorithms have proven useful for this task (Reed et al., 2013), given the potentially non-convex or discontinuous objective surface that results from optimizing both structure and parameters.

The two primary tools for generation of model structures are neuro-evolution, the evolution of neural network topologies (Stanley & Miikkulainen, 2002), and symbolic regression via genetic programming, the evolution of nonlinear regression models composed of symbolic mathematic elements from a primitive set (Koza, 1992, 1995). Regarding neuroevolution, Stanley and Miikkulainen (2002) introduced a method for parsimonious neural network generation by initializing small random networks and adding connections with random nodes and weights when performance improved. The space of possible network configurations is intractably large for most applications, making the method relatively slow to converge. Deep neural networks generated using evolution strategies (e.g. Lehman et al., 2018; Miikkulainen et al., 2019) for reinforcement learning (e.g. Conti et al., 2018) have generated comparable results to deep Q-networks (e.g. Mnih et al., 2015) and other fixed networks trained through backpropagation, but are not completely gradient-free.

Gradient-free genetic algorithms have been used for faster training of deep neural network weights (e.g. Such et al., 2017), but not successfully for the discovery of structure, as originally intended in Stanley and Miikkulainen (2002). The selection of search method will dictate the success of finding appropriate models to describe human behavior in a high-dimensional search space.

Symbolic regression similarly uses linear and nonlinear operators as base functions, and can, for example, learn to compose nested functions and automate the process of feature engineering. Symbolic trees can also incorporate noise (Schmidt & Lipson, 2007), can be seeded with relations of interest during optimization (e.g. Schmidt & Lipson, 2009b; Chadalawada et al., 2020), and can be strongly-typed to incorporate and handle heterogeneous data types or function outputs (Montana, 1995). Model evaluations of symbolic regression trees are generally faster than traditional feed-forward neural networks because each model evolves a sparse input representation based only on the inputs that improve performance. These factors make symbolic trees suited for iterative and exploratory model generation when using a gradient-free optimization method. The primitive set of structures for building symbolic trees determines the size of the search space, which often grows combinatorially with the number of primitives (Vanneschi et al., 2010). The selection of search method should consider the breadth of the resulting space of possible model structures. In applications where the target functions are not known, as in the modeling of complex and highly nonlinear human behavior, the space of possible model structures can be broadened to include a large number of possible functional relationships.

### 2.3 Model Evaluation

Model evaluation consists of performance metrics, component-level behavior, and the identification of parametric and structural drivers. This section reviews different approaches and perspectives regarding model evaluation for data-driven system identification, recognizing that the implementation of this phase is problem-dependent, and that integrated systems models including human behavior may be difficult to validate against theoretical or conceptual results depending on their scale.

The minimization of one or more error metrics between the model and data defines its proximity to the "true" model (Haussler & Warmuth, 1993; Kearns et al., 1994; Valiant, 2013). The different methodological and philosophical details of model evaluation in these settings are reviewed by Höge et al. (2018). Accordingly, the most prominent issue regarding model evaluation is the test error, the indicator of a model's ability to generalize to unseen data by balancing model bias and variance (Friedman, 1997; Pande et al., 2009; Höge et al., 2018). Generalization to unseen data is required to appropriately accommodate non-stationarity in data, a necessity when seeking to describe dynamic human behavior over long time periods. Finally, standard error metrics can be supplemented by additional criteria such as the information content learned from a model (Nearing & Gupta, 2015; Nearing et al., 2020), or when functional relationships are known, the evaluation of structural error through tradeoffs between predictive and functional performance (Ruddell et al., 2019).

For data-driven model structures describing human behavior, several extensions arise that deserve consideration during the model evaluation phase. The first is model complexity, recognizing that additional components or parameters do not necessarily result in the ability to represent increasingly complex system behavior (Sun et al., 2016). Instead, the goal is to find a parsimonious model, or the simplest model that still describes the data accurately. This has been identified as a challenge for heuristic approaches to data-driven system identification (Bongard & Lipson, 2007; Schmidt & Lipson, 2008, 2009a; Schmid, 2010). The second extension is model equifinality, or lack of uniqueness, which occurs when many model structures produce comparable predictions even after being tuned, trained, constrained, or optimized (Beven, 1993). This can suggest possible redundancy or over-simplification in the model, meaning that the parsimonious model may not have been found or the collected data is not diverse enough to fully represent the underlying process. For data-driven system identification this is especially challenging given the large space of possible model structures and conflicting performance metrics (Curry & Dagli, 2014). The concept of equifinality has been widely explored in hydrology and water resources (Khatami et al., 2019), but has been less emphasized in studies of human behavior modeling with competing structures and is more likely when less prior structural information is provided.

Finally, when model generation results in a large number of plausible model structures, a range of diagnostic tools can be applied to further assess the common structures and parameters driving model behavior. For example, Pruyt and Islam (2015) use clustering to partition exploratory model parameterizations based on their behavior as transfer functions mapping input to output. In the absence of well-characterized uncertainty, sensitivity analysis can diagnose model prediction behavior and provide a metric by which to justify the inclusion of parameters (Pianosi et al., 2016; Gupta & Razavi, 2018; Wagener & Pianosi, 2019). Dobson et al. (2019) design a scenario resampling strategy to show the importance of contextual uncertainty in the performance of operational rules of water systems. These and similar approaches assist with the evaluation of models of human behavior in the abstract, through which key structural elements can be identified post-optimization.

# 3 Experiment

Figure 2 outlines the computational steps for the three experimental phases: problem definition, model generation, and model evaluation. The Problem Definition phase includes the definition of prediction tasks, feature engineering, and the specification of function primitives. The Model Generation phase includes the selection of an encoding representation and search procedure, the definition of metrics to use for evaluating models during search, and the search over candidate model structures in a multi-objective space. The Model Evaluation phase for these experiments focuses on the collection and analysis of many plausible model sets across many random trials. Clustering and sensitivity analysis techniques are employed to determine driving structure and features in different regions of the performance space.

## 3.1 Problem Definition

### *3.1.1 Case Study*

This approach is applied to the problem of understanding dynamic agricultural land use patterns in the Tulare Basin region of California. In this case study, we use data-driven system identification to discover a mathematical function to predict the year-to-year change in tree crop acreage for all continuously planted square-mile sections of land in the Tulare Basin from 1974 to 2016. This is a human response variable that is of particular interest for water resources management because of a strong historical trend towards tree crops (Figure 3) that has exacerbated groundwater overdraft, especially in times of drought (Jasechko & Perrone, 2020).

### *3.1.2 Problem Definition*

The state of the system $x_t$ is defined as an $n$-tuple drawn from $\mathbb{R}^n$ that includes the current and previous state of tree crops $(a_t, a_{t-1}, \dots)$ and non-tree crops, the lagged change of tree-crops $(a'_{t-1}, a'_{t-2}, \dots)$ and non-tree crops since the current change is being predicted, and other current and lagged information such as the current crop price, agricultural pumping, and surface water deliveries.

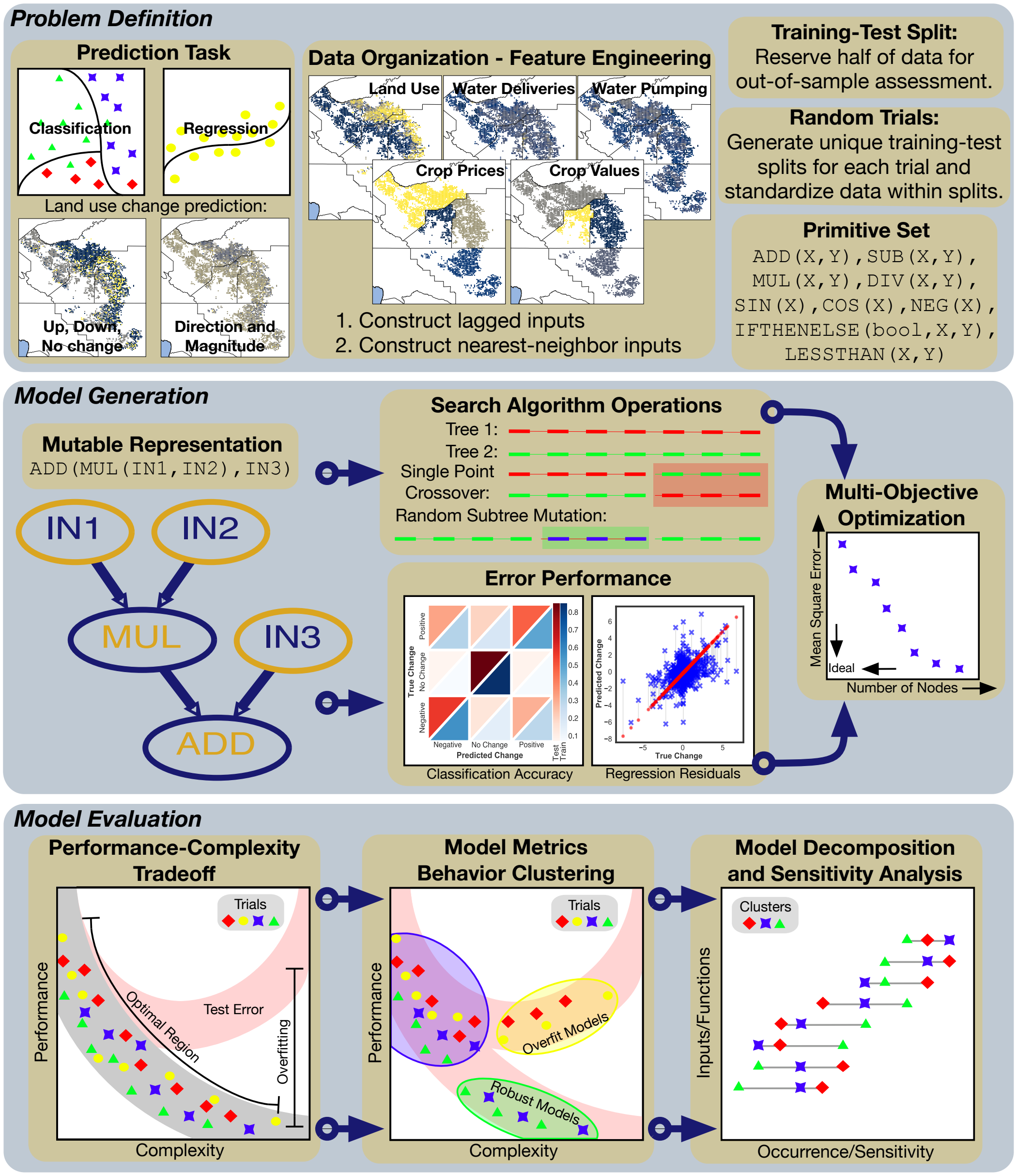


**Figure 2.** Schematic of experimental setup and workflow

$$x_t \coloneqq \left(a_t, b_t, c_t, \ldots, a_{t-1}, a'_{t-1}, b_{t-1}, c_{t-1}, \ldots\right) \tag{1}$$

where $a_t = a_{t-1} + a'_{t-1}$. Given the state of the system $x_t$ representing all current and previous information at a given spatial index, in learning the dynamics of the system we aim to predict the annual change in acreage at the same spatial index, $a'_t$, as a function of previous changes, current and previous states, and other features:

$$D_{x_t} \coloneqq \frac{\Delta x_t}{\Delta t} = F\left(x_t\right) \tag{2}$$

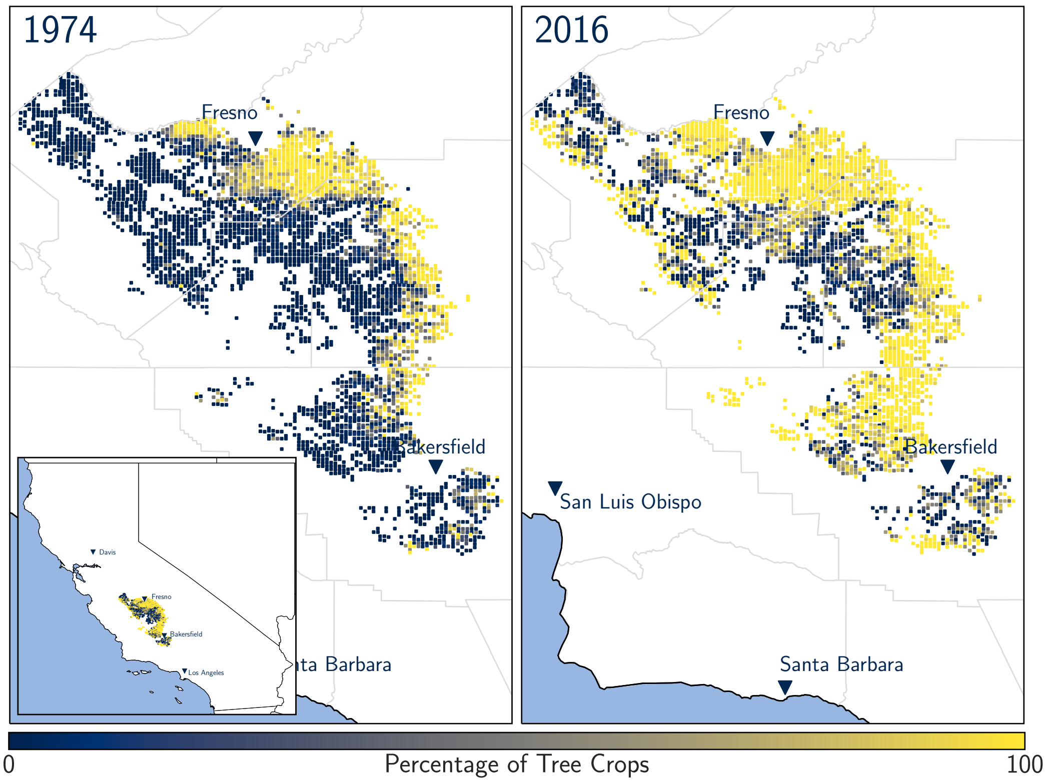


**Figure 3.** Historical change in crop type in the Tulare Basin, California from 1974 to 2016. Each grid cell is 1 mi$^2$, and tree crops are defined as in Mall and Herman (2019).

The notation $D_{x_t}$ is used to refer to the difference in tree crops $a'_t$ that would advance the tree crop state forward in time, $a_{t+1} = a_t + a'_t$. $x_t$ includes lagged responses such as $D_{x_{t-1}}$, the response of the previous state at the same index. The problem of learning model structure is therefore to determine the function $F$ that maps a given set of features to the annual change in state. $x_t$ includes potentially high-dimensional information describing the current state and any number of previous states (Lusch et al., 2018). When the dynamics of $F$ are unknown, general function forms are initialized randomly and trained to approximate system dynamics by learning from observed or measured data.

We explore two different prediction tasks related to this problem, regression and classification. In the regression formulation, models predict the magnitude and direction of the annual change in tree crop acreage. In the classification problem, models predict the direction of change only - positive, negative, or no change - as displayed under Prediction Task in Figure 2. Regression is generally considered a more difficult problem as functions must predict a continuous value, whereas this classification task requires predicting the most likely of three classes.

#### *3.1.3 Feature Engineering*

Feature data describing land use, water availability, and economics were organized into samples to train and test candidate model structures. Land use data was taken from the California Pesticide Use Reports, available digitally beginning in 1974 and extracted by Mall and Herman (2019). Annual crop type data are taken from 1974-2016 at the square-mile scale for over 3000 grid cells in the Tulare Basin. Water availability data were taken from the C2VSim-IWFM groundwater model output representing pumping and delivery estimates (Kourakos et al., 2019). Lastly, county-level crop prices were taken from the California County Agricultural Commissioner reports, digitized beginning in 1980 across Tulare, Fresno, Kings, and Kern counties, the four counties represented in the study area (USDA National Agricultural Statistics Service - California Field Office, 2019). Crop prices were adjusted for inflation using the producer price index for agriculture, based on the year 2016, published by the U.S. Bureau of Labor Statistics (U.S. Bureau of Labor Statistics, 2019). A summary of trends for this heterogeneous data set is presented in Figure 4.

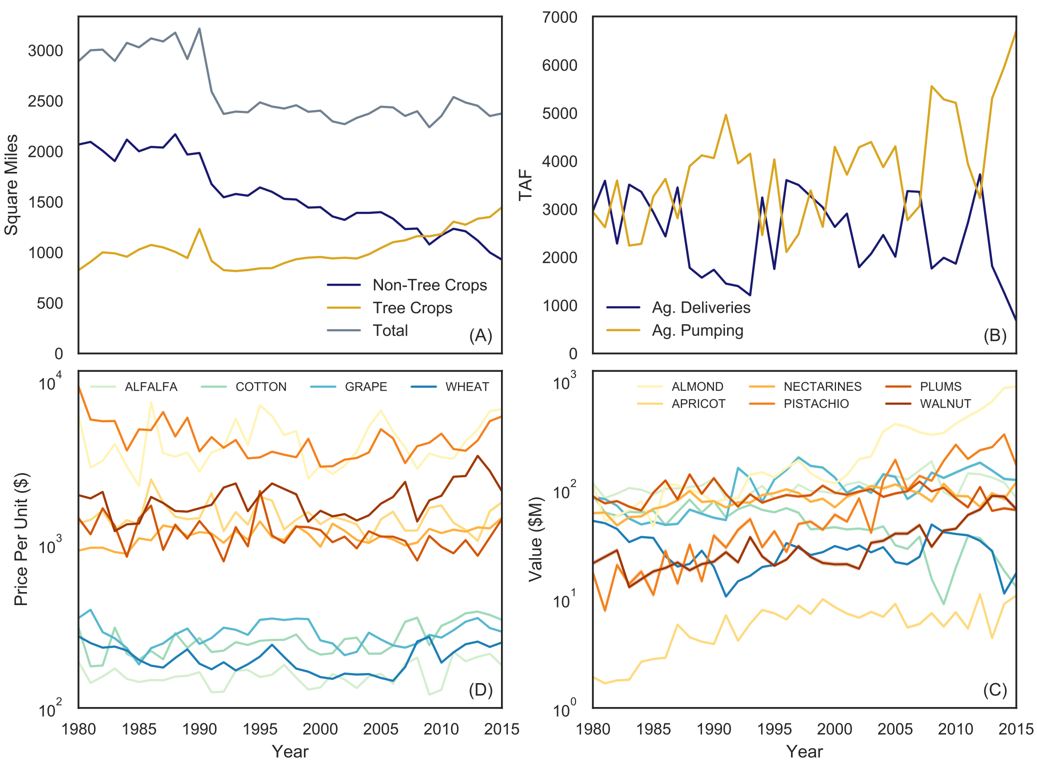


**Figure 4.** Historical trends in heterogeneous feature data. (A) Tree crop acreage, non-tree crop acreage, and total acreage planted; (B) Yearly total agricultural water deliveries and pumping; (C-D) Inflation-adjusted prices and total crop values for a selection of crops.

Additional features were included to account for the space-time dependence of the problem. Samples were organized such that each grid-cell sample was tagged with its data, the previous six years of data, and the same data from each of 5 neighboring grid cells in space. Since economic information is only available from 1980 onward and spatially distributed at the county scale, this space-time extension was only implemented for land and water data. Absolute data, such as the year and location, were excluded from the set of features to avoid overfitting. The resulting dimensions of the data were on the order of 500 predictor variables and 130,000 samples. No explicit dimension reduction steps were implemented, primarily to maintain the interpretation of feature variables and their eventual use within model structures. Samples were split into 50% training and 50% test, and both the features and response variable were standardized to $\mathcal{N}(0,1)$. Other than the bias introduced by constructing variables representing temporal lags and spatial neighborhoods, no empirical or theoretical priors were provided to inform the search.

#### *3.1.4 Model Structural Elements*

In addition to the feature variables, the primitive set of functions composing the feasible model structures must also be specified. The primitive set includes the mathematical relationships detailed in Table 1:

## Functions

| | |
|---|---|
| [float] = add([float],[float]) | [float] = sin([float]) |
| [float] = subtract([float],[float]) | [float] = cos([float]) |
| [float] = multiply([float],[float]) | [float] = negative([float]) |
| [float] = divide([float],[float]) | [bool] = less_than([float],[float]) |
| [float] = if_then_else([bool],[float],[float]) | |

## Constants

| | |
|---|---|
| (1,[bool]) | (RandInt(0,100)/10.,[float]) |
| (0,[bool]) | (RandInt(0,100)/1.,[float]) |

**Table 1.** The primitive set functions and constants, as defined for both regression and classification experiments. The space of feasible models is constrained by strong typing. The function RandInt$(a, b)$ generates a uniform random integer on $(a, b)$.

To include relational and logical operators in addition to mathematical operators in the primitive set, the functions are strongly typed, meaning that intermediate variables must match data types for the input and output of each component function. Constants are also defined as either boolean or floating point values as indicated in Table 1 and appear as terminal nodes in an expression, as do the model inputs (features). Constants are drawn from a distribution, though the resulting model is deterministic after the constants have been generated. However, the distributions themselves can be included in the primitive set, allowing the automatic construction of stochastic models (Schmidt & Lipson, 2007). In addition, search over the model space can be biased by providing a specific set of operators, inputs, or constants as seeds (Schmidt & Lipson, 2009b). By defining the primitive set and input space in this way, we ensure that search over the model space covers a broad general space of models, including linear and higher-order combinations of inputs and discontinuous functions.

### 3.2 Model Generation

#### *3.2.1 Search Objectives*

For the regression problem, the performance objective used to train model structures is the mean squared error (MSE), a commonly-used error metric that emphasizes larger residuals. A baseline performance value for MSE on the response variable—standardized to $\mathcal{N}(0, 1)$—is 1.0, which results from using the average prediction (zero) for every sample. For a given regressor $F : \mathbb{R}^n \to \mathbb{R}^1$:

$$MSE_{train} \coloneqq ave_{x_t \in X_{train}} (\hat{D}_{x_t} - D_{x_t})^2 \tag{3}$$

In the classification experiment, the multi-class output is addressed via ensemble learning, a common method in genetic programming studies (Espejo et al., 2010). The performance objective is the percent of misclassified samples. This is equivalent to $1-$ *Accuracy*, where accuracy is the percentage of classes predicted correctly. A baseline performance for misclassification percentage for this application is approximately 0.54, which results from predicting the most common class (no change) for every sample. The misclassification percentage can be calculated using the Hamming loss, $l(\hat{y}, y)$, which takes the value 1 for predictions that do not match the response and 0 otherwise. For a given classifier $F : \mathbb{R}^n \rightarrow \{Negative, No\ Change, Positive\}$:

$$MCP_{train} \coloneqq ave_{x_t \in X_{train}} l(\hat{D}_{x_t}, D_{x_t}) \tag{4}$$

A second objective, model complexity, is formulated and optimized concurrently with the performance objectives above using multi-objective optimization. The complexity metric is taken to be the representation length, a commonly used surrogate for computational or algorithmic complexity of a model, which in this case is the number of elements (nodes) in the ordered list representing the model. The complexity value is normalized by the maximum depth of recursive function calls in Python (90) to roughly match the scale and precision of the performance objectives.

#### *3.2.2 Search Algorithm*

The search over candidate model structures and parameterizations employs genetic programming, an evolutionary approach that encodes mathematical expressions in a tree structure to support symbolic regression. Mutation and crossover operators act on list representations of the models to generate new structures from promising candidates. In this study, the mutation operator adds a randomly initialized sub-tree of depth 1-2, and single-point crossover randomly selects a location along two separate model element lists. Mutation explores the model space by introducing new model structures, and crossover exploits the attributes of current models by testing new combinations of existing model structures. The mutation and crossover operations can result in invalid models according to the strong typing criteria, where intermediate data types among tree operations do not match; these are discarded before evaluation.

During training, the performance and complexity objectives were both minimized, and deterministic crowding was used for model selection (Deb et al., 2002). This has two implications: (1) the minimum complexity (maximum interpretability) model is preferred among two models with the same performance, (2) if the space of possible models is searched exhaustively, the resulting tradeoffs between models should be the minimum complexity model for a given level of performance. An archive of Pareto-approximate model structures is maintained and updated through non-dominated sorting of the archive and population together among the two objectives. The use of deterministic crowding is intended to promote diversity within populations by spacing models out along the Pareto front. Diversity is important to promote within populations for a number of reasons, but primarily to ensure that no single model dominates in all objectives and is used to generate all new individuals in the next generation.

Experiments were run with the Distributed Evolutionary Algorithms in Python package, or DEAP (De Rainville et al., 2012), using the UC Davis College of Engineering HPC1 Cluster with 96 processors. DEAP supports distributed computing, a number of evolutionary strategies, symbolic regression via genetic programming, and multi-objective optimization. Each population of models is made up of 96 individuals, and each tree is initialized randomly with depth 1-3. Trials run for a maximum of 20,000 generations, with a stagnation criterion of 2,500 generations. 21 iterations of the training-test split were performed. The code to reproduce this study can be found at DOI: 10.5281/zenodo.3887360.

### 3.3 Model Evaluation

Following the model training, candidate structures are evaluated in three ways: tradeoffs between performance objectives, model behavior in the metric space, and decomposition and sensitivity of the underlying structure and features. The approach to model evaluation taken during this phase depends on modeling decisions during problem definition and model generation. In these experiments, the feature data and primitive set together define a combinatorially large space of possible models, creating substantial uncertainty that must be acknowledged in the analysis that follows.

#### *3.3.1 Performance-Complexity Tradeoff*

After evaluating performance on the test set, models are placed in a three-dimensional performance-complexity tradeoff, as illustrated under Model Evaluation in Figure 2. Along the Pareto front, training error within a given trial will strictly decrease as complexity increases. However, as complexity of the model increases, test error can diverge from training error if the model overfits. If error performance changes relatively little across a broad range of model structures, this is an indicator of equifinality. To investigate this outcome further, candidate models can be clustered into groups with similar behavior. Specifically, k-means clustering is used to separate models according to training error, test error, and complexity.

#### *3.3.2 Model Decomposition and Sensitivity Analysis*

The collection of Pareto-optimal sets of models constitutes a new high-dimensional data set of structured model components and their associated performance metrics. Among many network analysis tools for structural and dynamic analysis of graphical models, model decomposition is a very simple initial step. The driving structural properties of each model—number of metrics, attributes, inputs, functions, and constants—are linked to their behavior cluster as described above. Each model is also tested for its sensitivity to individual features and their interactions through global sensitivity analysis. Along with the assessment of model responses to observed conditions in the training and test data, each model is re-evaluated with 1000 samples scaled by the cardinality of its unique feature set to ensure sufficient coverage of the sample space. Sobol sensitivity analysis is performed using the Python package SALib (Herman & Usher, 2017).

## 4 Results

Figure 5 shows the tradeoff between performance and complexity across the Pareto set of candidate model structures for both regression and classification experiments. Each point represents the performance of a model on the test data, while the gold shading shows the distribution of performance for the same models on the training data. Figure 5 highlights four different regions: Parsimonious, Overfit, Equifinal, and Dominated model clusters. Initial structure building during each trial occurs in the Parsimonious cluster in both Figure 5a and 5b. The Overfit clusters in Figure 5 are highlighted as the regions where models begin to rely on spurious structure discovered at any point during the trial and maintain a level of robustness on test data. The Equifinal cluster in Figure 5a represents a region where multiple model structures exist at roughly the same level of performance. The Dominated cluster in Figure 5b represents models that are both equifinal and do not generalize well to unseen data.

These results indicate several points. First, regression trials in Figure 5a exhibit better robustness to test data, with most models remaining within the region of the training error displayed in the gold background. Classification experiments show diminishing returns to increasing complexity much faster than regression experiments. Optimization trials are locked into a specific model structure by the development in the Parsimo-

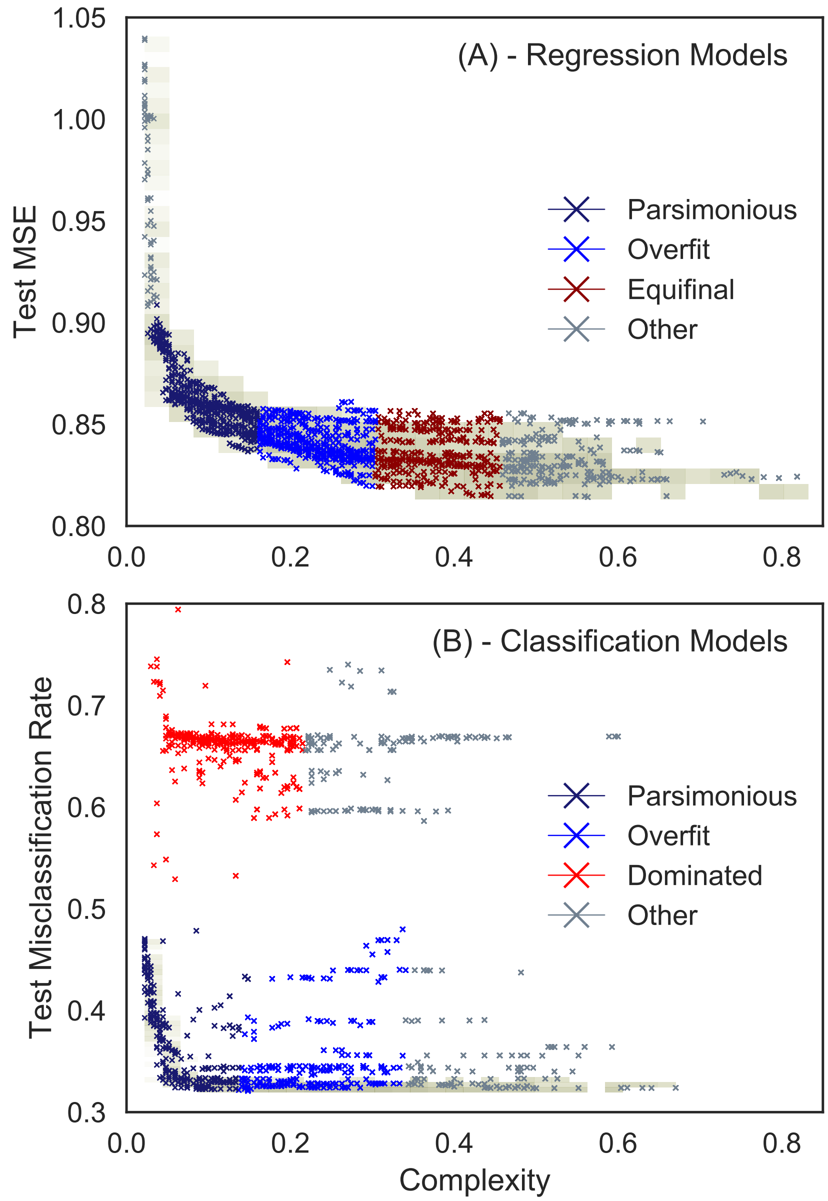


**Figure 5.** Tradeoff between performance (test error) and complexity for model structures across (A) all regression trials and (B) all classification trials. Light gold shading indicates the distribution of the same models evaluated on the training data. Models are clustered according to their behavior in this three-dimensional space (training error, test error, and complexity).

nious clusters; if this structure is developed before enough exploration has happened, it may explain why significant overfitting occurs in Figure 5b. Equifinal model structures are observed in both cases, as many models with increasing complexity demonstrate similar performance.

Figure 5b shows model structures with a variety of macroscopic behavior that can be investigated further. We proceed with the classification results to determine the drivers of behavior in the three highlighted clusters in Figure 5b. The Parsimonious cluster represents the initial set of low-complexity models prior to their divergence into either the

Overfit cluster or the Dominated cluster, so we examine the structure of three models from the Parsimonious cluster that perform well on both training and test data in Figure 6.

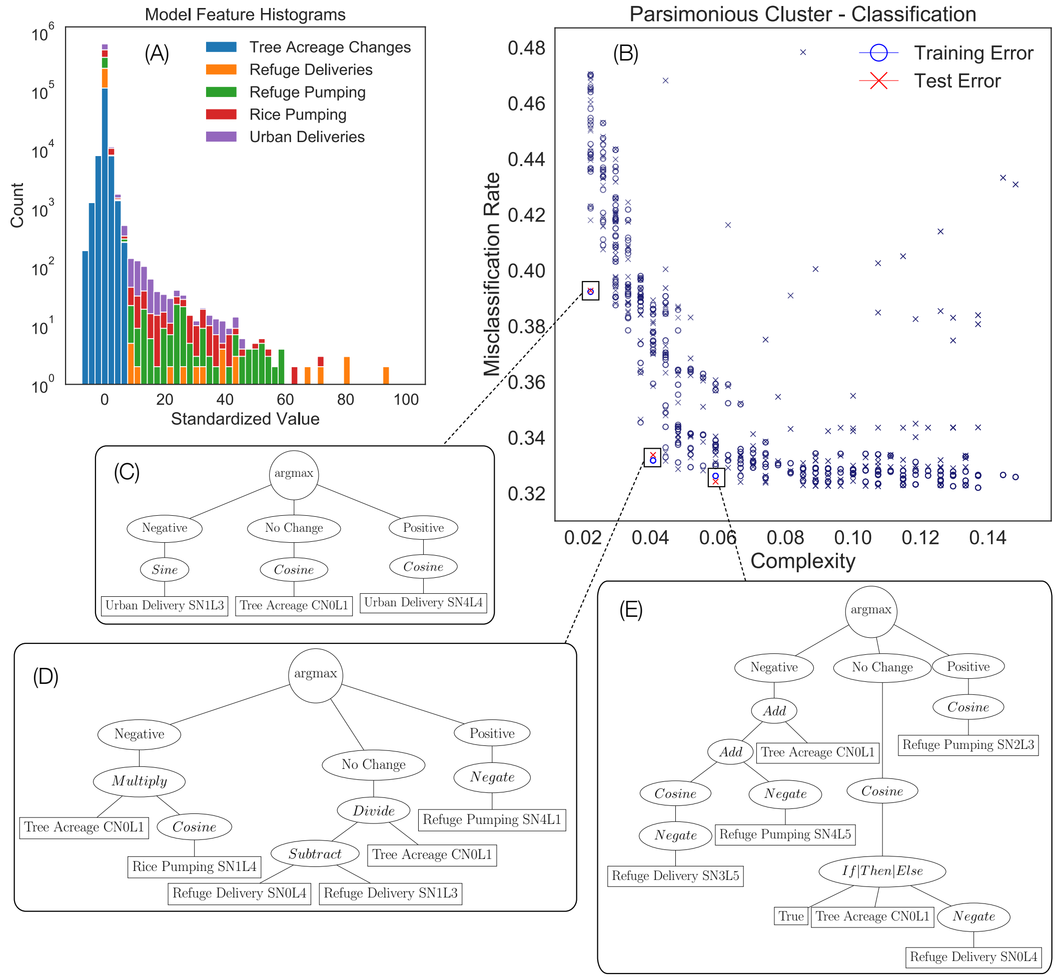


**Figure 6.** Parsimonious cluster training and test error (B), a selection of robust classifiers (C-E), and histograms of feature data represented in the models (A). Feature constructions are annotated as {**S**tate/**C**hange | **N**eighbor **0**–**5** | **L**ag **1-5**}.

The three classifiers shown in Figure 6 depend on a variety of feature data and constructions. A single construction of the tree acreage change—the previous change in the same location—was used by all three classifiers, whereas many different constructions of sparsely distributed (mostly zero) and asymmetric water data were used among the three models. In two of the models, this construction of tree acreage change occurs multiple times. Additionally, the tree acreage change feature tends to occur closer to the output of models, and is less engineered than the water data as a result. In inspecting individual models, the lag-1 tree acreage change is often used directly when appearing near the output of models, whereas additional complexity is often used to engineer other feature data as nonlinear scaling of the lag-1 tree acreage change.

Across all model structures, there is a clear dependence on the lag-1 tree acreage change, indicating that decision-making agents are informed by past decisions. Lack of consensus regarding other feature constructions indicate that these structural connec-

482 tions may be spurious. The distribution of features used in models must be interpreted
483 in the context of spatio-temporal resolutions. For example, the lack of consensus on the
484 use of economic data could be due to its coarser resolution in space and limited cover-
485 age in time, or the inability of the search method to find advantageous structure beyond
486 the lag-1 tree acreage change. In this case, we aim to identify the structural drivers sep-
487 arating robust models in the Parsimonious and Overfit Clusters from models that do not
488 generalize well in the Dominated cluster. First, we start by decomposing the models in
489 each cluster into their components to assess the structural differences in the occurrence
490 of feature variables and function primitives in each cluster, displayed in Figure 7.

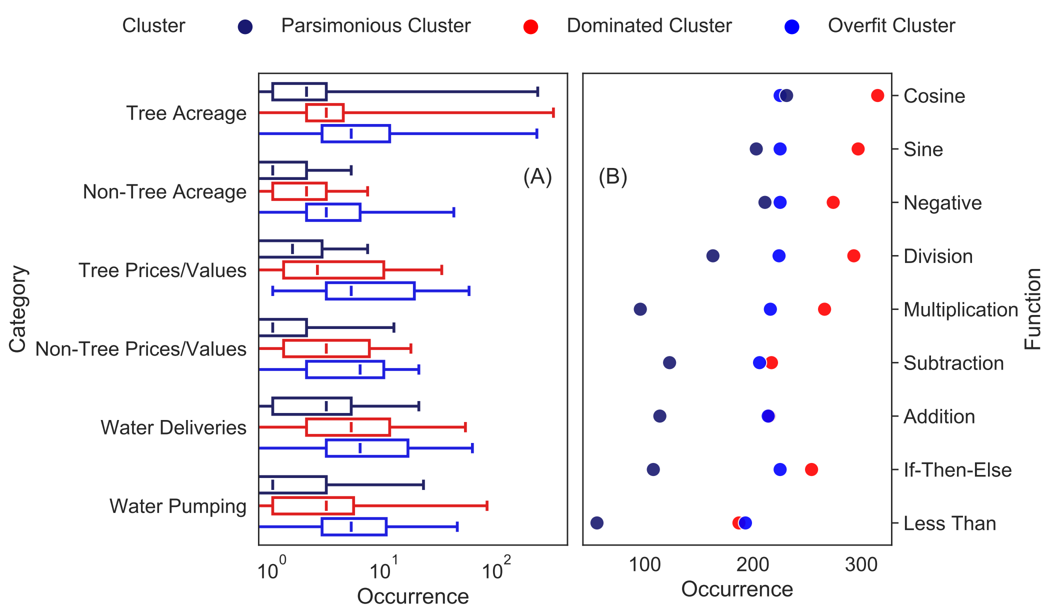


**Figure 7.** Total feature occurrence distributions within each category of inputs by model cluster (A) and function occurrence (B) across all classification models found during search. In (A), each feature category holds a number of feature constructions of input data from that category, leading to a distribution of total occurrences across all models in a given cluster within each category of feature data.

491 For feature variables (Figure 7a), all clusters show a dependence on the group of
492 inputs related to tree acreage data (all lagged and neighboring states and values for tree
493 crops). The lag-1 tree acreage change in the same location (categorized under Tree Acreage)
494 appear in every model across all clusters, indicated by the range of the whiskers at the
495 top of Figure 7a. The Overfit cluster contains more instances of inputs from each cat-
496 egory as compared to the Dominated and Parsimonious clusters. Almost the opposite
497 is true for function occurrence, where the Dominated cluster learns greater function de-
498 pendence than the Overfit cluster from the Parsimonious cluster for almost all primitives.
499 The Overfit cluster exhibits a more even distribution of function occurrence across prim-
500 itives than the Parsimonious and Dominated clusters, suggesting an increase in the di-
501 versity of function primitives relative to the Parsimonious cluster. Both the Overfit clus-
502 ter and Dominated cluster learn a dependence on the two conditional primitives.

503 The occurrence of the features does not by itself describe the response of the model
504 output to the values of the features, which is the goal of the sensitivity analysis step. Re-
505 sults for total sensitivity indices are presented in Figure 8 as non-exceedance curves for
506 two categories, tree acreage and non-tree acreage.

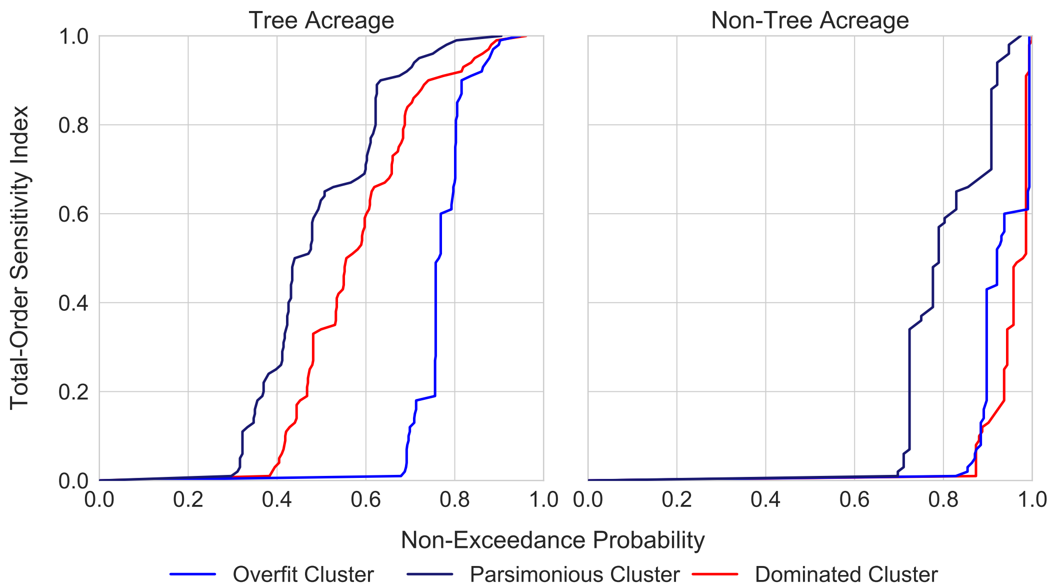


**Figure 8.** Empirical distribution of total-order sensitivity indices for two categories of feature variables: tree acreage and non-tree acreage, separated by metric space clustering (color).

Both the Overfit and Dominated cluster models show decreased sensitivity to both features relative to the Parsimonious cluster, indicating that the original occurrences of inputs become less influential as training proceeds. In the case of tree acreage inputs, over 70% of sensitivity indices to features in the Overfit cluster were small ($S_T < 0.2$) compared to less than 50% for the model structures from the Dominated and Parsimonious clusters. However, at least 20% of tree acreage inputs to both the Overfit and Dominated cluster models are high ($S_T > 0.8$), illustrating the existence of a spectrum of sensitivities to tree acreage data across the set of models. The transition between learning small and large sensitivities to tree acreage data in the Overfit cluster models appears to be a unique structural driver of the Overfit cluster's behavior in the metric space for this problem. On the other hand, both the Overfit cluster models and Dominated cluster models do not show the same high sensitivities to non-tree acreage data that appear in the Parsimonious cluster models.

This result confirms the conclusion from Figure 7 that previous tree acreage states and changes are a main driver for this problem. The results also indicate a partition in the information important to the decision problem; since changing from tree crops and non-tree crops requires respecializing and alternate scheduling, it runs counter to intuition to note that over 80% of non-tree crop input occurrences had negligibly small impacts on the decision to change towards tree crops, and there were very few input occurrences among the Overfit or Dominated model clusters with sensitivity indices greater than 0.6.

Finally, the average total-order sensitivity indices within each category of feature variables are displayed across clusters in Figure 9.

Models from the Overfit cluster exhibit relatively equal sensitivities across all feature categories as compared to models from the Parsimonious or Dominated clusters. Figure 9 also reveals that models from the Overfit cluster learn to be more sensitive on av-

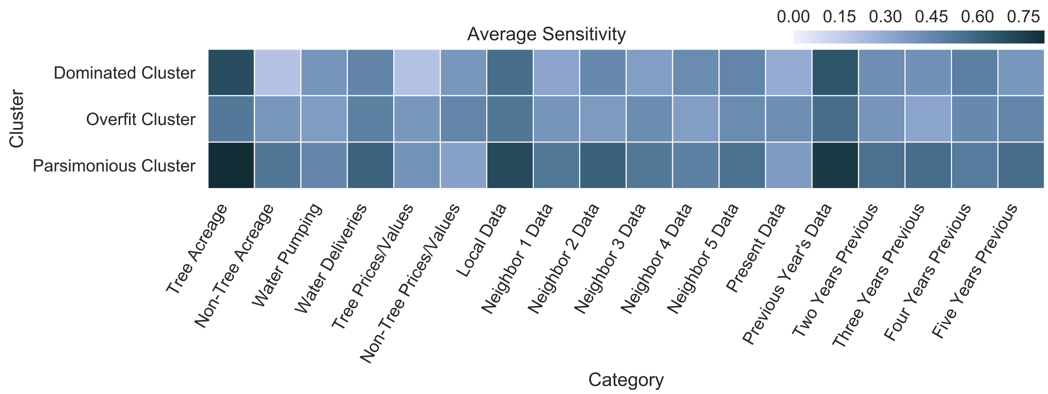


**Figure 9.** Average total-order sensitivity indices of feature variables across input categories for each cluster of model structures. In the feature grouping labels, "data" refers to the combination of state, change, temporal lags, and spatial neighbors for each type of feature.

erage to the prices of tree crops and non-tree acreage data than models from the Dominated cluster. In general, the behavior that allows models in the Overfit cluster to generalize to new data is the incorporation of more occurrences of all categories of feature variables, but without becoming too sensitive to individual occurrences from any one category by over-engineering with function structure. However, it is noted that averaging across the set of models may obscure the sensitivities of individual models, the distribution of which is better shown in Figure 8.

## 5 Discussion

There is a distinct need for integrated systems models when descriptions of the physical system are incomplete without consideration of the human component (Konar et al., 2019; Schill et al., 2019; Herman et al., 2020). This must include feedbacks that may not be represented by combinations of existing model structures (Calvin & Bond-Lamberty, 2018). This study proposes methods to automate the exploration of model structure to describe human behavior along the canonical tradeoff between performance and complexity. In this illustrative case study focused on agricultural land use and water demand, no priors or constraints were placed on the space of possible model structures. However, enumerating the range of optimal performance with increasing complexity provides context for any prior-informed solutions that might arise in the same context. The relative performance demonstrated here thus forms a basis for the analysis of model structural uncertainty (Walker et al., 2003) through casting of models as hypotheses (Beven, 2019). These outcomes follow from the quantification of a number of general model evaluation goals summarized in Figure 10.

Generating candidate model structures includes automatic feature selection and requires no prior knowledge of the system's mechanics, constraints, or information requirements beyond the basic provision of data and primitives (Bongard & Lipson, 2007; Schmidt & Lipson, 2009a; Knüsel et al., 2019). Though more concise problem framings (e.g. Dobson et al., 2019), generation schemes (e.g. Chadalawada et al., 2020), or post-search analysis tools (e.g. Worland et al., 2019) could uncover more specific emergent phenomena in the data and resulting models, framing model structural experimentation according to this generic framework enables a baseline contextualization of the complex integrated systems problem. In this way, a data-driven approach to generating model structure could support the design of agent-based or hydro-economic models.

| Phase | Model Goal | Approach | Encoding |
|---|---|---|---|
| Definition | Generality | Function Space | Primitives |
| | | Data Space | Features |
| Generation | Accuracy | Performance | Average Error |
| | | | Misclassification |
| | Interpretability | Complexity | # Nodes |
| Evaluation | Generalization | Overfitting | Test Error |
| | Behavior | Clustering | k-means |
| | Importance | Feature | Occurrence |
| | | | Sensitivity |
| | | Structure | Occurrence |

**Figure 10.** Model evaluation phases, general model evaluation goals, the approaches used to address each goal, and how the assessment was encoded in each approach.

Describing the human decision in this case study was encumbered by two primary sources of difficulty: (1) the difficulty of search in combined parametric-structural model spaces, and (2) the difficulty when incorporating noisy or incomplete heterogeneous feature data appropriate to the temporal and spatial scale of the problem. First, the search space of candidate model structures grows combinatorially, making it extremely unlikely to identify unique optimal solutions. In this study, the sudden failure to improve in performance past a given level of complexity in the classification experiment (Figure 5b), a saturation often interpreted as convergence, could be driven by a structural boundary beyond which improvements could not easily be found. Studies have argued for an upper limit on the description length of a model (Vanneschi et al., 2010) as done in Chadalawada et al. (2020), though it is difficult to know without doing an unconstrained search what the upper limit should be. Hybrid methods, such as evolutionary strategies to approximate a gradient, are promising for tractable search in vast model spaces (Conti et al., 2018; Miikkulainen et al., 2019). Even when model complexity is considered, black-box models do not guarantee interpretability, and the results presented here indicate that more strategic analysis can be done to interpret how models are making predictions, such as explaining the importance of features and structure in neural networks (e.g. Montavon et al., 2018; Worland et al., 2019), and using sensitivity analysis to explicate structural dependence in space and time (e.g. Quinn et al., 2019).

Second, the performance-complexity tradeoff of candidate model structures is tied to the choice of feature variables at the appropriate scale, and observed with the necessary accuracy, to generate acceptable test performance (Höge et al., 2018). This is also the case when the relations that would model such data do not exist or are not included in the primitive set (Kearns et al., 1994). This study incorporates land use and economic data across multiple decades and at a relatively fine spatial resolution to derive a single decision model, which may be better served by developing multiple functions across the spatial region. Additionally, while the feature engineering applied to the data helps discern the importance of certain autocorrelated structure, it is also obfuscatory, as the representation of an agent's decision-making context using neighborhoods could be improved upon to further explore spatial dependence while avoiding unnecessary correlations within samples. The feature data itself may not provide the right signal to adequately model the underlying process in this setting, due to noise in measurement or observation error, or the choice of inadequate features. However, examining multiple problem formulations allows the comparison of relative performance, as in the regression and classification experiments in this study; while classification is the easier problem, it shows higher potential for overfitting and may be underrepresenting the complexity in the data. Making use of heterogeneous data to identify the model structure of integrated systems is not simple or straightforward, but the explanation of decisions made by complex behavioral agents based on multiple sources of information is enabled by the methodological template presented here.

# 6 Conclusion

This study develops an approach to the inference of model structures and parameterizations from data describing human behavior in water resources systems. Three phases are considered: problem definition, model generation, and model evaluation, demonstrated on a case study of land use decisions in the Tulare Basin, California. No prior model structure is assumed, beyond the feature engineering to build a high-dimensional dataset reflecting land use, water use, and crop prices. Results indicate a tradeoff between model performance and complexity, with substantial equifinality in model structures that require additional diagnostic analysis. To this end, model structures are clustered according to similar behavior, and driving structural features are diagnosed by considering function importance and input sensitivity. Specific challenges arise due to identifying spatially distributed decisions from heterogeneous, multi-sectoral data, generally preventing the identification of a single "best" model from the performance-complexity tradeoff. This provides a basis for analyzing structural uncertainty under broadly-defined problem contexts, and a possible path forward for the generation of model components from observed data to support integrated representations of human actors in water systems.

## Acknowledgments

This work was supported by National Science Foundation grant CNH-1716130. All conclusions are those of the authors. We would also like to thank Dr. Giorgos Kourakos and Dr. Helen Dahlke for providing model results from C2VSim IWFM, and Natalie Mall for compiling of the land use data set analyzed here. Lastly, we would like to thank Dr. Julianne Quinn for helpful comments on a preliminary draft. Data and code are available at DOI:10.5281/zenodo.3887360.

## References

An, L. (2012). Modeling human decisions in coupled human and natural systems: Review of agent-based models. *Ecological Modelling*, *229*, 25–36. doi: 10.1016/j.ecolmodel.2011.07.010

Barto, A. G., & Dietterich, T. G. (2012). Reinforcement learning and its relationship to supervised learning. In *Handbook of learning and approximate dynamic programming* (p. 45-63). John Wiley & Sons, Ltd. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470544785.ch2` doi: 10.1002/9780470544785.ch2

Bastidas, L. A., Hogue, T. S., Sorooshian, S., Gupta, H. V., & Shuttleworth, W. J. (2006). Parameter sensitivity analysis for different complexity land surface models using multicriteria methods. *Journal of Geophysical Research: Atmospheres*, *111*(D20). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006377` doi: 10.1029/2005JD006377

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017). Building confidence in climate model projections: an analysis of inferences from fit. *WIREs Climate Change*, *8*(3), e454. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.454` doi: 10.1002/wcc.454

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, *16*(1), 41 - 51. Retrieved from http://www.sciencedirect.com/science/article/pii/030917089390028E (Research Perspectives in Hydrology) doi: https://doi.org/10.1016/0309-1708(93)90028-E

Beven, K. (2019). Towards a methodology for testing models as hypotheses in the inexact sciences. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *475*(2224), 20180862. Retrieved from https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2018.0862 doi: 10.1098/rspa.2018.0862

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. *Journal of Hydrology*, *249*(1), 11 - 29. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022169401004218 doi: https://doi.org/10.1016/S0022-1694(01)00421-8

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., ... Zhang, Y. (2019). Twenty-three unsolved problems in hydrology (UPH) – a community perspective. *Hydrological Sciences Journal*, *64*(10), 1141-1158. Retrieved from https://doi.org/10.1080/02626667.2019.1620507 doi: 10.1080/02626667.2019.1620507

Bongard, J., & Lipson, H. (2007). Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *104*(24), 9943–9948. Retrieved from https://www.pnas.org/content/104/24/9943 doi: 10.1073/pnas.0609476104

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. Retrieved from https://doi.org/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *113*(15), 3932–3937. Retrieved from https://www.pnas.org/content/113/15/3932 doi: 10.1073/pnas.1517384113

Calvin, K., & Bond-Lamberty, B. (2018, June). Integrated human-earth system modeling–state of the science and future directions. *Environmental Research Letters*, *13*(6), 063006. Retrieved from https://iopscience.iop.org/article/10.1088/1748-9326/aac642 doi: 10.1088/1748-9326/aac642

Castelletti, A., Galelli, S., Restelli, M., & Soncini-Sessa, R. (2012). Data-driven dynamic emulation modelling for the optimal management of environmental systems. *Environmental Modelling & Software*, *34*, 30 - 43. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364815211002015 (Emulation techniques for the reduction and sensitivity analysis of complex environmental models) doi: https://doi.org/10.1016/j.envsoft.2011.09.003

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*(8), 1477–1494. Retrieved from https://doi.org/10.1007/s10994-018-5724-2 doi: 10.1007/s10994-018-5724-2

Chadalawada, J., Herath, H. M. V. V., & Babovic, V. (2020). Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction. *Water Resources Research*, *56*(4), e2019WR026933. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR026933 (e2019WR026933 10.1029/2019WR026933) doi: 10.1029/2019WR026933

Chini, C. M., Konar, M., & Stillwell, A. S. (2017). Direct and indirect urban water footprints of the united states. *Water Resources Research*, *53*(1), 316-327. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/

10.1002/2016WR019473 doi: 10.1002/2016WR019473

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., … Rasmussen, R. M. (2015a). A unified approach for process-based hydrologic modeling: 1. modeling concept. *Water Resources Research*, *51*(4), 2498-2514. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR017198 doi: 10.1002/2015WR017198

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., … Marks, D. G. (2015b). A unified approach for process-based hydrologic modeling: 2. model implementation and case studies. *Water Resources Research*, *51*(4), 2515-2542. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR017200 doi: 10.1002/2015WR017200

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., … Hay, L. E. (2008). Framework for understanding structural errors (fuse): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*(12). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007WR006735 doi: 10.1029/2007WR006735

Claussen, M., Mysak, L., Weaver, A., Crucifix, M., Fichefet, T., Loutre, M. F., … Wang, Z. (2002). Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. *Climate Dynamics*, *18*(7), 579–586. Retrieved from https://doi.org/10.1007/s00382-001-0200-1 doi: 10.1007/s00382-001-0200-1

Cominola, A., Nguyen, K., Giuliani, M., Stewart, R. A., Maier, H. R., & Castelletti, A. (2019). Data mining to uncover heterogeneous water use behaviors from smart meter data. *Water Resources Research*, *55*(11), 9315-9333. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR024897 doi: 10.1029/2019WR024897

Conti, E., Madhavan, V., Such, F. P., Lehman, J., Stanley, K., & Clune, J. (2018). Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. In *Advances in neural information processing systems* (pp. 5027–5038).

Curry, D. M., & Dagli, C. H. (2014). Computational complexity measures for many-objective optimization problems. *Procedia Computer Science*, *36*, 185 - 191. Retrieved from http://www.sciencedirect.com/science/article/pii/S187705091401326X (Complex Adaptive Systems Philadelphia, PA November 3-5, 2014) doi: https://doi.org/10.1016/j.procs.2014.09.077

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182-197.

De Rainville, F.-M., Fortin, F.-A., Gardner, M.-A., Parizeau, M., & Gagné, C. (2012). Deap: A python framework for evolutionary algorithms. In *Proceedings of the 14th annual conference companion on genetic and evolutionary computation* (p. 85–92). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2330784.2330799 doi: 10.1145/2330784.2330799

Dobson, B., Wagener, T., & Pianosi, F. (2019). How important are model structural and contextual uncertainties when estimating the optimized performance of water resource systems? *Water Resources Research*, *55*(3), 2170-2193. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024249 doi: 10.1029/2018WR024249

Eker, S., Rovenskaya, E., Obersteiner, M., & Langan, S. (2018). Practice and perspectives in the validation of resource management models. *Nature Communications*, *9*(1), 5359. Retrieved from https://doi.org/10.1038/s41467-018-07811-9 doi: 10.1038/s41467-018-07811-9

Espejo, P. G., Ventura, S., & Herrera, F. (2010). A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(2), 121-144.

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. motivation and theoretical development. *Water Resources Research*, *47*(11). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010WR010174 doi: 10.1029/2010WR010174

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*(1), 55–77. Retrieved from https://doi.org/10.1023/A:1009778005914 doi: 10.1023/A:1009778005914

Giuliani, M., & Herman, J. D. (2018). Modeling the behavior of water reservoir operators via eigenbehavior analysis. *Advances in Water Resources*, *122*, 228 - 237. Retrieved from http://www.sciencedirect.com/science/article/pii/S0309170817311193 doi: https://doi.org/10.1016/j.advwatres.2018.10.021

Giuliani, M., Li, Y., Castelletti, A., & Gandolfi, C. (2016). A coupled human-natural systems analysis of irrigated agriculture under changing climate. *Water Resources Research*, *52*(9), 6928-6947. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016WR019363 doi: 10.1002/2016WR019363

Giuliani, M., Mason, E., Castelletti, A., Pianosi, F., & Soncini-Sessa, R. (2014). Universal approximators for direct policy search in multi-purpose water reservoir management: A comparative analysis. *IFAC Proceedings Volumes*, *47*(3), 6234 - 6239. Retrieved from http://www.sciencedirect.com/science/article/pii/S1474667016425899 (19th IFAC World Congress) doi: https://doi.org/10.3182/20140824-6-ZA-1003.01962

Gupta, H. V., & Nearing, G. S. (2014). Debates - the future of hydrological sciences: A (common) path forward? using models and data to learn: A systems theoretic perspective on the future of hydrological science. *Water Resources Research*, *50*(6), 5351-5359. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013WR015096 doi: 10.1002/2013WR015096

Gupta, H. V., & Razavi, S. (2018). Revisiting the basis of sensitivity analysis for dynamical earth system models. *Water Resources Research*, *54*(11), 8692-8717. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022668 doi: 10.1029/2018WR022668

Harou, J. J., Pulido-velazquez, M., Rosenberg, D. E., Medellín-azuara, J., Lund, J. R., & Howitt, R. E. (2009). Hydro-economic models : Concepts , design , applications , and future prospects. *Journal of Hydrology*, *375*(3-4), 627–643. Retrieved from http://dx.doi.org/10.1016/j.jhydrol.2009.06.037 doi: 10.1016/j.jhydrol.2009.06.037

Haughton, N., Abramowitz, G., Pitman, A. J., Or, D., Best, M. J., Johnson, H. R., ... Vuichard, N. (2016). The plumbing of land surface models: Is poor performance a result of methodology or data quality? *Journal of Hydrometeorology*, *17*(6), 1705-1723. Retrieved from https://doi.org/10.1175/JHM-D-15-0171.1 doi: 10.1175/JHM-D-15-0171.1

Haussler, D., & Warmuth, M. (1993). The probably approximately correct (pac) and other learning models. In A. L. Meyrowitz & S. Chipman (Eds.), *Foundations of knowledge acquisition: Machine learning* (pp. 291–312). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-585-27366-2_9 doi: 10.1007/978-0-585-27366-2_9

Herman, J., & Giuliani, M. (2018). Policy tree optimization for threshold-based water resources management over multiple timescales. *Environmental Modelling & Software*, *99*, 39-51. Retrieved from https://www.scopus.com/inward/

record.uri?eid=2-s2.0-85031504985&doi=10.1016%2fj.envsoft.2017.09.016&partnerID=40&md5=b9f887f1108ad06c668352e2fc9e70f4 (cited By 12) doi: 10.1016/j.envsoft.2017.09.016

Herman, J., Quinn, J., Steinschneider, S., Giuliani, M., & Fletcher, S. (2020). Climate adaptation as a control problem: Review and perspectives on dynamic water resources planning under uncertainty. *Water Resources Research*, *56*(2). Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081030619&doi=10.1029%2f2019WR025502&partnerID=40&md5=402e907ca2b2d2e9a1275d3cc0c4e0e6 (cited By 1) doi: 10.1029/2019WR025502

Herman, J., & Usher, W. (2017). Salib: An open-source python library for sensitivity analysis. *Journal of Open Source Software*, *2*(9), 97. Retrieved from https://doi.org/10.21105/joss.00097 doi: 10.21105/joss.00097

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, *54*(3), 1688-1715. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017WR021902 doi: 10.1002/2017WR021902

Hogue, T. S., Bastidas, L. A., Gupta, H. V., & Sorooshian, S. (2006). Evaluating model performance and parameter behavior for varying levels of land surface model complexity. *Water Resources Research*, *42*(8). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005WR004440 doi: 10.1029/2005WR004440

Hsu, K.-L., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, *31*(10), 2517-2530. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/95WR01955 doi: 10.1029/95WR01955

Hu, Y., Quinn, C. J., Cai, X., & Garfinkle, N. W. (2017). Combining human and machine intelligence to derive agents' behavioral rules for groundwater irrigation. *Advances in Water Resources*, *109*, 29–40. Retrieved from https://doi.org/10.1016/j.advwatres.2017.08.009 doi: 10.1016/j.advwatres.2017.08.009

Huang, G.-B., Chen, L., Siew, C. K., et al. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks*, *17*(4), 879–892.

Jafino, B. A., Haasnoot, M., & Kwakkel, J. H. (2019, Feb.). What are the merits of endogenising land-use change dynamics into model-based climate adaptation planning? *Socio-Environmental Systems Modelling*, *1*, 16126. Retrieved from https://sesmo.org/article/view/16126 doi: 10.18174/sesmo.2019a16126

Jasechko, S., & Perrone, D. (2020). California's central valley groundwater wells run dry during recent drought. *Earth's Future*, *8*(4), e2019EF001339. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019EF001339 (e2019EF001339 2019EF001339) doi: 10.1029/2019EF001339

Kearns, M. J., Schapire, R. E., & Sellie, L. M. (1994). Toward efficient agnostic learning. *Machine Learning*, *17*(2), 115–141. Retrieved from https://doi.org/10.1007/BF00993468 doi: 10.1007/BF00993468

Khatami, S., Peel, M. C., Peterson, T. J., & Western, A. W. (2019). Equifinality and flux mapping: A new approach to model evaluation and process representation under uncertainty. *Water Resources Research*, *55*(11), 8922-8941. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR023750 doi: 10.1029/2018WR023750

Kleijnen, J. P. C. (2015). Response surface methodology. In M. C. Fu (Ed.), *Handbook of simulation optimization* (pp. 81–104). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/978-1-4939-1384-8_4 doi: 10.1007/978-1-4939-1384-8_4

Klotz, D., Herrnegger, M., & Schulz, K. (2017). Symbolic regression for the estimation of transfer functions of hydrological models. *Water Resources Research*, *53*(11), 9402-9423. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017WR021253 doi: 10.1002/2017WR021253

Knüsel, B., Zumwald, M., Baumberger, C., Hirsch Hadorn, G., Fischer, E. M., Bresch, D. N., & Knutti, R. (2019). Applying big data beyond small problems in climate research. *Nature Climate Change*, *9*(3), 196–202. Retrieved from https://doi.org/10.1038/s41558-019-0404-1 doi: 10.1038/s41558-019-0404-1

Konar, M., Garcia, M., Sanderson, M. R., Yu, D. J., & Sivapalan, M. (2019). Expanding the scope and foundation of sociohydrology as the science of coupled human-water systems. *Water Resources Research*, *55*(2), 874-887. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024088 doi: 10.1029/2018WR024088

Kourakos, G., Dahlke, H. E., & Harter, T. (2019). Increasing groundwater availability and seasonal base flow through agricultural managed aquifer recharge in an irrigated basin. *Water Resources Research*, *55*(9), 7464-7492. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024019 doi: 10.1029/2018WR024019

Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection.* Cambridge, MA, USA: MIT Press.

Koza, J. R. (1995). Survey of genetic algorithms and genetic programming. In *Wescon conference record* (pp. 589–594).

Lehman, J., Chen, J., Clune, J., & Stanley, K. O. (2018). Es is more than just a traditional finite-difference approximator. In *Proceedings of the genetic and evolutionary computation conference* (p. 450–457). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3205455.3205474 doi: 10.1145/3205455.3205474

Lipton, Z. C. (2018, June). The mythos of model interpretability. *Queue*, *16*(3), 31–57. Retrieved from https://doi.org/10.1145/3236386.3241340 doi: 10.1145/3236386.3241340

Ljung, L. (2017). System identification. In *Wiley encyclopedia of electrical and electronics engineering* (p. 1-19). American Cancer Society. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/047134608X.W1046.pub2 doi: 10.1002/047134608X.W1046.pub2

Lund, J. R. (2015). Integrating social and physical sciences in water management. *Water Resources Research*, *51*(8), 5905-5918. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR017125 doi: 10.1002/2015WR017125

Lusch, B., Kutz, J. N., & Brunton, S. L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, *9*(1), 4950. Retrieved from https://doi.org/10.1038/s41467-018-07210-0 doi: 10.1038/s41467-018-07210-0

Mall, N. K., & Herman, J. D. (2019, oct). Water shortage risks from perennial crop expansion in California's Central Valley. *Environmental Research Letters*, *14*(10), 104014. Retrieved from https://doi.org/10.1088%2F1748-9326%2Fab4035 doi: 10.1088/1748-9326/ab4035

Marston, L., & Konar, M. (2017). Drought impacts to water footprints and virtual water transfers of the Central Valley of California. *Water Resources Research*, *53*(7), 5756-5773. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016WR020251 doi: 10.1002/2016WR020251

Mason, E., Giuliani, M., Castelletti, A., & Amigoni, F. (2018). Identifying and modeling dynamic preference evolution in multipurpose water resources systems. *Water Resources Research*, *54*(4), 3162-3175. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017WR021431 doi: 10.1002/2017WR021431

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., ... others (2019). Evolving deep neural networks. In *Artificial intelligence in the age of neural networks and brain computing* (pp. 293–312). Elsevier.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Monier, E., Paltsev, S., Sokolov, A., Chen, Y. H. H., Gao, X., Ejaz, Q., ... Haigh, M. (2018). Toward a consistent modeling framework to assess multi-sectoral climate impacts. *Nature Communications*, *9*(1), 660. Retrieved from `https://doi.org/10.1038/s41467-018-02984-9` doi: 10.1038/s41467-018-02984-9

Montana, D. J. (1995). Strongly typed genetic programming. *Evolutionary Computation*, *3*(2), 199-230. Retrieved from `https://doi.org/10.1162/evco.1995.3.2.199` doi: 10.1162/evco.1995.3.2.199

Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., & Kutz, J. N. (2019). Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique*, *347*(11), 845 - 855. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1631072119301809` (Data-Based Engineering Science and Technology) doi: https://doi.org/10.1016/j.crme.2019.11.009

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1 - 15. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1051200417302385` doi: https://doi.org/10.1016/j.dsp.2017.10.011

Müller, M. F., Yoon, J., Gorelick, S. M., Avisse, N., & Tilmant, A. (2016). Impact of the Syrian refugee crisis on land use and transboundary freshwater resources. *Proceedings of the National Academy of Sciences*, *113*(52), 14932–14937. Retrieved from `https://www.pnas.org/content/113/52/14932` doi: 10.1073/pnas.1614342113

Muneepeerakul, R., & Anderies, J. M. (2020). The emergence and resilience of self-organized governance in coupled infrastructure systems. *Proceedings of the National Academy of Sciences*, *117*(9), 4617–4622. Retrieved from `https://www.pnas.org/content/117/9/4617` doi: 10.1073/pnas.1916169117

Nearing, G. S., & Gupta, H. V. (2015). The quantity and quality of information in hydrologic models. *Water Resources Research*, *51*(1), 524-538. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014WR015895` doi: 10.1002/2014WR015895

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020). Does information theory provide a new paradigm for earth science? Hypothesis testing. *Water Resources Research*, *56*(2), e2019WR024918. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR024918` (e2019WR024918 2019WR024918) doi: 10.1029/2019WR024918

Oyebode, O., Babatunde, D. E., Monyei, C. G., & Babatunde, O. M. (2019). Water demand modelling using evolutionary computation techniques: integrating water equity and justice for realization of the sustainable development goals. *Heliyon*, *5*(11), e02796. Retrieved from `http://www.sciencedirect.com/science/article/pii/S2405844019364564` doi: https://doi.org/10.1016/j.heliyon.2019.e02796

Pande, S., McKee, M., & Bastidas, L. A. (2009). Complexity-based robust hydrologic prediction. *Water Resources Research*, *45*(10). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008WR007524` doi: 10.1029/2008WR007524

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, *79*, 214 - 232. Retrieved from `http://www.sciencedirect.com/science/article/`

pii/S1364815216300287 doi: https://doi.org/10.1016/j.envsoft.2016.02.008

Pruyt, E., & Islam, T. (2015). On generating and exploring the behavior space of complex models. *System Dynamics Review*, *31*(4), 220-249. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84955562469&doi=10.1002%2fsdr.1544&partnerID=40&md5=0eab6d448ea9b3acdb34d7b1553d2ebb (cited By 9) doi: 10.1002/sdr.1544

Quinn, J. D., Reed, P. M., Giuliani, M., & Castelletti, A. (2019). What is controlling our control rules? opening the black box of multireservoir operating policies using time-varying sensitivity analysis. *Water Resources Research*, *55*(7), 5962-5984. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024177 doi: 10.1029/2018WR024177

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*, *54*(7), 4638-4662. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022743 doi: 10.1029/2018WR022743

Reed, P., Hadka, D., Herman, J., Kasprzyk, J., & Kollat, J. (2013). Evolutionary multiobjective optimization in water resources: The past, present, and future. *Advances in Water Resources*, *51*, 438 - 456. Retrieved from http://www.sciencedirect.com/science/article/pii/S0309170812000073 (35th Year Anniversary Issue) doi: https://doi.org/10.1016/j.advwatres.2012.01.005

Rosenstein, M. T., & Barto, A. G. (2001). Robot weightlifting by direct policy search. In *Proceedings of the 17th international joint conference on artificial intelligence - volume 2* (p. 839–844). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ruddell, B. L., Drewry, D. T., & Nearing, G. S. (2019). Information theory for model diagnostics: Structural error is indicated by trade-off between functional and predictive performance. *Water Resources Research*, *55*(8), 6534-6554. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR023692 doi: 10.1029/2018WR023692

Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2017). Data-driven discovery of partial differential equations. *Science Advances*, *3*(4). Retrieved from https://advances.sciencemag.org/content/3/4/e1602614 doi: 10.1126/sciadv.1602614

Schill, C., Anderies, J. M., Lindahl, T., Folke, C., Polasky, S., Cárdenas, J. C., ... Schlüter, M. (2019). A more dynamic understanding of human behaviour for the anthropocene. *Nature Sustainability*, *2*(12), 1075–1082. Retrieved from https://doi.org/10.1038/s41893-019-0419-7 doi: 10.1038/s41893-019-0419-7

Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, *656*, 5–28.

Schmidt, M., & Lipson, H. (2007). Learning noise. In *Proceedings of the 9th annual conference on genetic and evolutionary computation* (p. 1680–1685). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1276958.1277289 doi: 10.1145/1276958.1277289

Schmidt, M., & Lipson, H. (2008, 07). Data-Mining Dynamical Systems: Automated Symbolic System Identification for Exploratory Analysis. In (Vol. Volume 2: Automotive Systems; Bioengineering and Biomedical Technology; Computational Mechanics; Controls; Dynamical Systems, p. 643-649). Retrieved from https://doi.org/10.1115/ESDA2008-59309 doi: 10.1115/ESDA2008-59309

Schmidt, M., & Lipson, H. (2009a). Distilling free-form natural laws from experimental data. *Science*, *324*(5923), 81–85. Retrieved from https://science.sciencemag.org/content/324/5923/81 doi: 10.1126/science.1165893

Schmidt, M., & Lipson, H. (2009b). Incorporating expert knowledge in evolutionary search: A study of seeding methods. In *Proceedings of the 11th annual conference on genetic and evolutionary computation* (p. 1091–1098). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1569901.1570048 doi: 10.1145/1569901.1570048

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558-8593. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022643 doi: 10.1029/2018WR022643

Sivapalan, M., & Blöschl, G. (2015). Time scale interactions and the coevolution of humans and water. *Water Resources Research*, *51*(9), 6988-7022. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR017896 doi: 10.1002/2015WR017896

Sivapalan, M., Savenije, H. H. G., & Blöschl, G. (2012). Socio-hydrology: A new science of people and water. *Hydrological Processes*, *26*(8), 1270-1276. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.8426 doi: 10.1002/hyp.8426

Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, *10*(2), 99-127. Retrieved from https://doi.org/10.1162/106365602320169811 doi: 10.1162/106365602320169811

Such, F. P., Madhavan, V., Conti, E., Lehman, J., Stanley, K. O., & Clune, J. (2017). Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *arXiv preprint arXiv:1712.06567*.

Sun, Z., Lorscheid, I., Millington, J. D., Lauf, S., Magliocca, N. R., Groeneveld, J., … Buchmann, C. M. (2016). Simple or complicated agent-based models? A complicated issue. *Environmental Modelling & Software*, *86*, 56 - 67. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364815216306041 doi: https://doi.org/10.1016/j.envsoft.2016.09.006

U.S. Bureau of Labor Statistics. (2019). *Producer Price Indexes - PPI databases.* (data retrieved from https://www.bls.gov/ppi/data.htm)

USDA National Agricultural Statistics Service - California Field Office. (2019). *County Ag Commissioners' Data Listing.* (data retrieved from https://www.nass.usda.gov/Statistics_by_State/California/Publications/AgComm/index.php)

Valiant, L. (2013). *Probably approximately correct: Nature's algorithms for learning and prospering in a complex world.* USA: Basic Books, Inc.

Vanneschi, L., Castelli, M., & Silva, S. (2010). Measuring bloat, overfitting and functional complexity in genetic programming. In *Proceedings of the 12th annual conference on genetic and evolutionary computation* (p. 877–884). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1830483.1830643 doi: 10.1145/1830483.1830643

Verburg, P. H., Dearing, J. A., Dyke, J. G., van der Leeuw, S., Seitzinger, S., Steffen, W., & Syvitski, J. (2016). Methods and approaches to modelling the anthropocene. *Global Environmental Change*, *39*, 328 - 340. Retrieved from http://www.sciencedirect.com/science/article/pii/S0959378015300285 doi: https://doi.org/10.1016/j.gloenvcha.2015.08.007

Wagener, T., & Pianosi, F. (2019). What has global sensitivity analysis ever done for us? A systematic review to support scientific advancement and to inform policy-making in earth system modelling. *Earth-Science Reviews*, *194*, 1 - 18. Retrieved from http://www.sciencedirect.com/science/article/pii/S0012825218300990 doi: https://doi.org/10.1016/j.earscirev.2019.04.006

Walker, W., Harremoës, P., Rotmans, J., van der Sluijs, J., van Asselt, M., Janssen, P., & von Krauss, M. K. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Inte-*

*grated Assessment*, *4*(1), 5-17. Retrieved from https://doi.org/10.1076/iaij.4.1.5.16466 doi: 10.1076/iaij.4.1.5.16466

Worland, S. C., Steinschneider, S., Asquith, W., Knight, R., & Wieczorek, M. (2019). Prediction and inference of flow duration curves using multioutput neural networks. *Water Resources Research*, *55*(8), 6850-6868. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024463 doi: 10.1029/2018WR024463

Wu, C. L., Chau, K. W., & Li, Y. S. (2009). Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resources Research*, *45*(8). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007WR006737 doi: 10.1029/2007WR006737

Young, P. (1998). Data-based mechanistic modelling of environmental, ecological, economic and engineering systems. *Environmental Modelling & Software*, *13*(2), 105 - 122. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364815298000115 doi: https://doi.org/10.1016/S1364-8152(98)00011-5

Zaniolo, M., Giuliani, M., Castelletti, A. F., & Pulido-Velazquez, M. (2018). Automatic design of basin-specific drought indexes for highly regulated water systems. *Hydrology and Earth System Sciences*, *22*(4), 2409–2424. Retrieved from https://www.hydrol-earth-syst-sci.net/22/2409/2018/ doi: 10.5194/hess-22-2409-2018