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

Liam Ekblad[1,2] and Jonathan Herman[1,1]

[1]University of California, Davis
[2]University of California

November 30, 2022

**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 model 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.

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

**Liam Ekblad**[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 Ekblad, `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 model 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

Increasingly, water resources models combine observed data and computational experiments to support the development of theory regarding system processes (Clark et al., 2015a, 2015b), particularly those for which existing theory may insufficiently explain available observations (Karpatne et al., 2017; Schlüter et al., 2019). One such process is human behavior, which 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; Schill et al., 2019). 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 individuals and institutions across multiple sectors and scales (Mason et al., 2018; Monier et al., 2018; Muneepeerakul & Anderies, 2020). The increasing availability of multi-sectoral data describing these processes provides the opportunity to complement theory by inferring plausible models from data (Brunton et al., 2016; Montáns et al., 2019).

Many subfields of water resources have focused on the challenge of modeling human behavior, including: dynamical systems models, as in socio-hydrology (Sivapalan et al., 2012) and social-ecological systems (Berkes & Folke, 1998); hydro-economic models (Harou et al., 2009); and agent-based modeling (An, 2012). Each offers differing perspectives on which system components should be treated as exogenous, controlled, or self-organized, and which behaviors can be adequately described by data versus theory (Anderies, 2015). However, all share the goal of accurately describing observed dynamics of the system while managing the complexity of the spatial and temporal representation (Baumberger et al., 2017; Höge et al., 2018). These approaches are not necessarily exclusive, and can be connected through a common experimental framing—for example, Müller and Levy (2019) review how economic theory can be coupled with data-driven sociohydrologic modeling to support and develop theories of human influence in water systems. Similarly, agent-based modeling studies have integrated data-driven and theory-driven approaches to investigate system processes (Gunaratne & Garibay, 2017; Schlüter et al., 2019; Vu et al., 2019). By extricating the processes driving emergent and interdependent behaviors in coupled systems, data-driven models can be used beyond the integration of observations to advance theory.

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. Finally, 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 in water systems, while also identifying key challenges related to structure and complexity.

Model accuracy alone does not engender trust (Baumberger et al., 2017), particularly in the case of “black-box” models (Shen, 2018), though accuracy is often the primary metric by which model structure is validated (Eker et al., 2018). By starting from fixed model structures, many data-driven methods bypass the question of structural uncertainty (Walker et al., 2003). This complicates any eventual reconciliation with available theory or process knowledge to support interpretation and validation (Lipton, 2018; Knüsel et al., 2019; P. J. Schmidt et al., 2020). By contrast, data-driven methods for system identification are capable of searching both model structures and parameters to find candidate representations (Ljung, 2017). Methods have been demonstrated for systems in which the target relationships are well-known, such as the double pendulum (M. Schmidt & Lipson, 2009) and the Navier-Stokes equations (Rudy et al., 2017). In hydrology, data-driven system identification methods have been used to infer rainfall-runoff transfer functions (Klotz et al., 2017) and to automate the identification of rainfall-runoff model structures using global optimization (Spieler et al., 2020).

Generating model structures through data-driven system identification allows for the testing of multiple model structures and parameterizations as competing hypotheses (Beven, 2019), similar to how conceptual and theory-driven model components have been compared to reduce structural uncertainty (Clark et al., 2015a, 2015b; Nearing & Gupta, 2015; Knoben et al., 2020). Several specific challenges arise in the way candidate models are evaluated. First, data-driven system identification typically results in a tradeoff between model performance and complexity (Hogue et al., 2006; Bastidas et al., 2006; Pande et al., 2009). Second, additional criteria may be required for model evaluation, such as interpretability and agreement with available theory (Khatami et al., 2019; Knüsel et al., 2019). Opportunities remain for data-driven methods to identify model structures of water resources system components for which theory is still being developed, such as varied human influences. There remains a need for a general approach capable of generating and evaluating models of human interactions within water systems, with the simultaneous goals of accuracy and interpretability across a broad spectrum of possible representations (Schill et al., 2019).

This work contributes an approach to model generation and evaluation for the general challenge of deriving process representation and understanding from observed data in water resources systems. We focus on the particular challenge of modeling human behavior, an influential system process which poses significant uncertainty in hydrologic systems (Konar et al., 2019; Schill et al., 2019; Herman et al., 2020). By generating many candidate models as competing hypotheses and simultaneously evaluating models for performance and complexity, we operationalize a preference for parsimonious model structures in combinatorial search spaces. The structures resulting from search in broadly defined model spaces are consolidated through systematic decomposition and diagnostic assessment of plausible model sets to determine driving structure. The approach is demonstrated for a case study of agricultural land use decisions in California, a complex spatially distributed process through which humans exert substantial influence on the system. This approach provides a foundation for future studies of model structural uncertainty, reconciliation with theory, and integrated systems modeling, particularly regarding the role of these challenges in planning and management for coupled human-water systems under uncertainty.

## 2 Methodological Background

We extend data-driven system identification approaches to generate and evaluate plausible model structures describing human behavior in water resources systems (Figure 1). The experimental steps presented here share similarities with the problem of constructing emulators (surrogates) of environmental systems models (Castelletti et al., 2012; Kleijnen, 2015), though with the additional goal of generating models that support the development of candidate theory regarding system processes. This requires an evaluation phase in which the structures of generated models are examined directly. By searching over the space of model structures for a given problem definition, the uncertainty associated with selecting any given model can be visualized as a function of complexity and accuracy on held-out data.

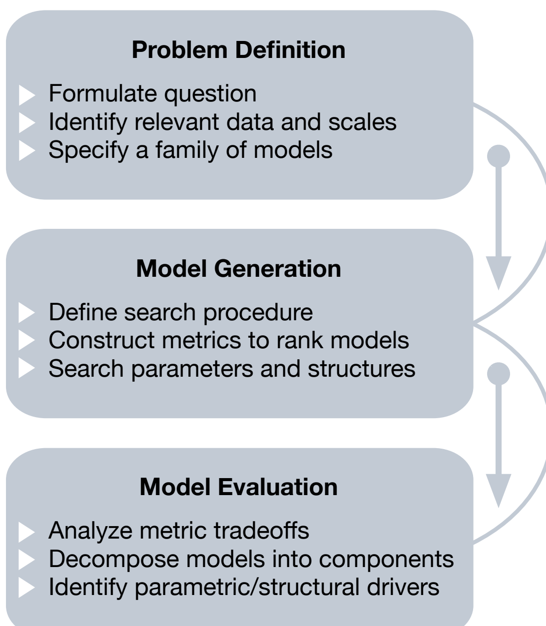


**Figure 1.** Flowchart of methodological steps involved in generating model structure from data.

### 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 generalized multivariate regression problem is an instance of supervised learning. However, it could alternatively be cast as a multivariate control problem, where rather than learning a model, a policy is learned based on rewards received by an agent after taking actions in an environment (Barto & Dietterich, 2004). The relationships between environmental observations and human decisions can also be framed in terms of causal inference, such as through instrumental variable and fixed effect experiments (Müller & Levy, 2019).

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 (Breiman, 2001; Huang et al., 2006, e.g.,) 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 (Quade et al., 2016). 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

The model generation process involves inferring models from data. Within water resources, model inference has largely focused on parameter estimation for given model structures. This is a broad category, including deterministic data-driven models with trainable weights such as neural networks (Hsu et al., 1995, e.g.,), physically-based model structures with probabilistic search procedures such as Markov Chain Monte Carlo (MCMC) (Vrugt & Beven, 2018, e.g.,), and general procedures for examining parametric uncertainty in conceptual linear and nonlinear model structures such as Generalized Likelihood Uncertainty Estimation (Beven & Binley, 1992). For example, Vrugt and Beven (2018) demonstrate the evolution and training of set-length Markov chains for different experiments, using differential evolution to explore a broadly-defined parameter space and maintaining a population of models in place of explicit structural search. These approaches are thus a combination of theory-driven structure and data-driven parameterization, which enables analysis of complexity and equifinality among parameter sets (discussed in Section 2.3).

Model structure can also be generated through a number of data-driven search methods that explicitly add or subtract elements—referred to as construction and pruning methods, respectively—which originate in the fields of machine learning and evolutionary algorithms. Construction methods include decision trees, which successively add linear decision rules to accurately classify samples (Quinlan, 1986), and genetic programming, the use of a genetic algorithm to build and search over graphical model structures composed of simple mathematical elements and inputs (Koza, 1992, 1995), among others. Broadly, global optimization 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 simultaneously. Though the target processes may be simple, basic implementations of these methods do not explicitly minimize model complexity. With increasing interest in model interpretability in machine learning (Lipton, 2018), pruning methods for the discovery of sparsely-connected sub-networks have been introduced that reproduce or improve performance of fully-connected neural networks after they have been trained (Frankle & Carbin, 2018). In contrast, multiple objectives can be used with construction methods to evaluate model structures simultaneously for error performance and structural complexity during optimization, codifying a preference during search for simpler models that perform equally well.

Genetic programming is particularly useful for its ability to conduct global multi-objective search over model structures of arbitrary complexity, i.e., symbolic regression (Quade et al., 2016). Symbolic regression 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 (M. D. Schmidt & Lipson, 2007), can be seeded with relations of interest during optimization (M. D. Schmidt & Lipson, 2009; Chadalawada et al., 2020, e.g.,), 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 grows combinatorially with the number of primitives (Vanneschi et al., 2010). 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 the examination of performance metrics and 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). Since the potential for a model to overfit to training data increases with complexity, the foremost 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). Generalization to unseen data is also required to appropriately accommodate non-stationarity in data, a necessity when seeking to describe dynamic human behavior over long time periods (Höge et al., 2018). 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 (Z. 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; M. D. Schmidt & Lipson, 2008; M. Schmidt & Lipson, 2009).

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 oversimplification 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), as well as in agent-based models (Williams et al., 2020). However, with the exception of a recent example from the social sciences (Vu et al., 2019), the equifinality problem is rarely approached in integrated studies by global search over model structures that considers both performance and complexity during training.

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

This approach is demonstrated on an application of agricultural land use change, one of the primary ways in which human decisions influence water resources systems, in addition to reservoir operations and urban consumption. Land use change represents a

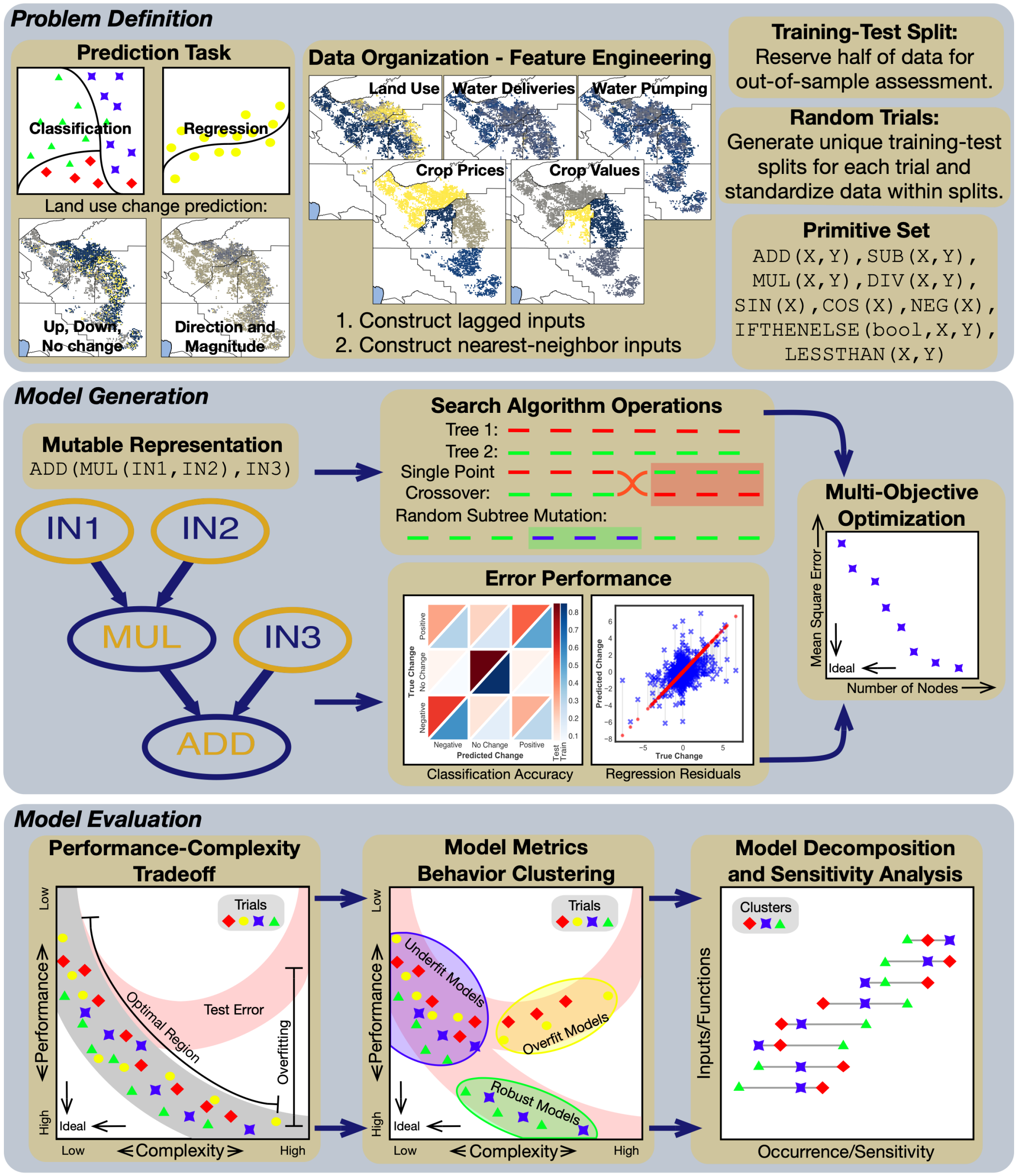


**Figure 2.** Schematic of experimental setup and workflow

complex test case in spatially distributed, heterogeneous decision-making (Groeneveld et al., 2017). Additionally, models of land use change depend on heterogeneous sources of information, such as water availability and socioeconomic factors (Nelson & Burchfield, 2017; Jasechko & Perrone, 2020; Malek & Verburg, 2020). This problem has been approached from multiple perspectives, including theoretical models based on economics and psychology (Schlüter et al., 2017), as well as statistical models (B. Sun & Robinson, 2018), which together suggest no clear agreement on process representation (Verburg et al., 2019). Economic models of land use change have been developed at the global scale (Prestele et al., 2017; Stehfest et al., 2019) and also at the regional scale (Howitt et al., 2012, e.g.,), and the integration of local and regional results into global models is currently being explored (Melsen et al., 2018; Schlüter et al., 2019; Malek & Verburg, 2020). In both cases, parameters are calibrated against historical observations. However, it is also acknowledged that land use decisions, like other water resources decisions, do not always follow the principle of full rationality (Groeneveld et al., 2017; Schlüter et al., 2017). By contrast, agent-based rules have also been developed for regional land use models, often ad hoc using expert judgment (Thober et al., 2017) informed by empirical studies (Robinson et al., 2007). There remains an opportunity to automate this process via model generation techniques, as has been explored elsewhere in the social sciences (Gunaratne & Garibay, 2017; Vu et al., 2019, e.g.,). While land use change presents a challenging test case, the methods proposed here also generalize to other aspects of human behavior in water resources systems, contingent on the availability of scale-appropriate datasets.

#### *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).

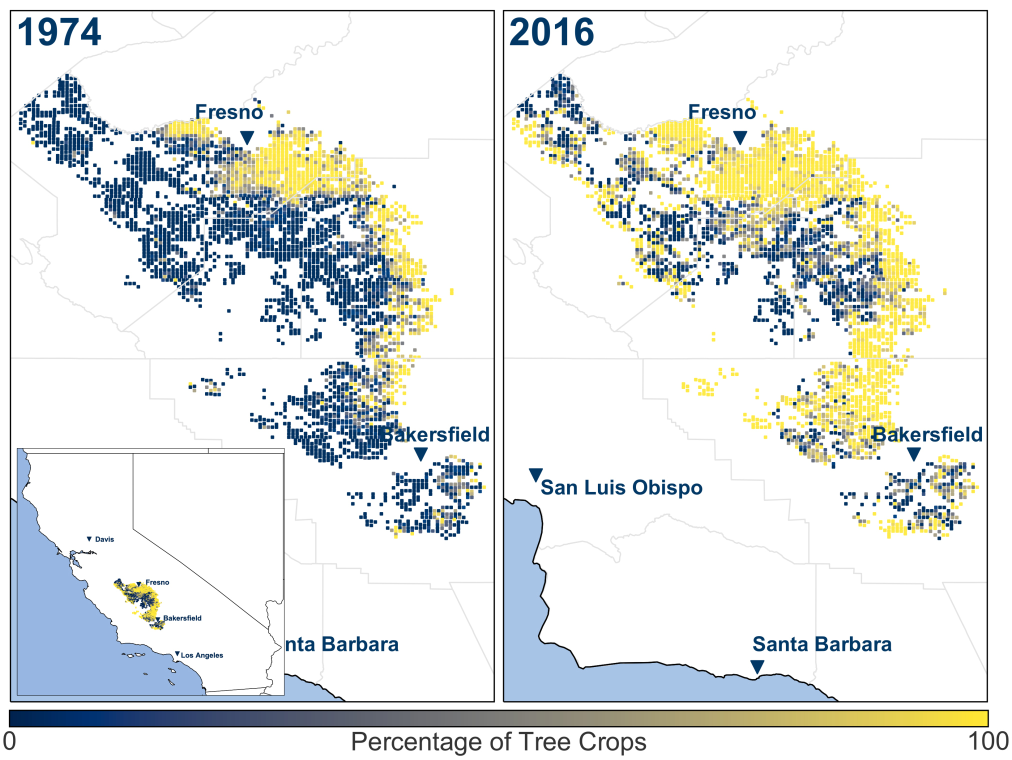


**Figure 3.** Historical change in crop type in the Tulare Basin, California from 1974 to 2016. Each grid cell is 1 $mi^2$, and tree crops are defined as in Mall and Herman (2019). The grey lines indicate county boundaries within which crop prices are reported annually.

#### 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 $(a_t, a_{t-1}, \ldots)$ and non-tree crops, the lagged change of tree-crops $(a'_{t-1}, a'_{t-2}, \ldots)$ 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.

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

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 (more information about these feature variables is given in Section 3.1.3):

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

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, and the target data are partitioned into tree and non-tree crops. Water availability data were taken from the C2VSim-IWFM groundwater model representing pumping and delivery estimates for the categories of urban, agricultural, rice crop, and refuge pumping and deliveries, further details for which are described in Kourakos et al. (2019). Lastly, county-level crop prices were taken from the California County Agricultural Commissioner reports, beginning in 1980 across Tulare, Fresno, Kings, and Kern counties (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.

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,

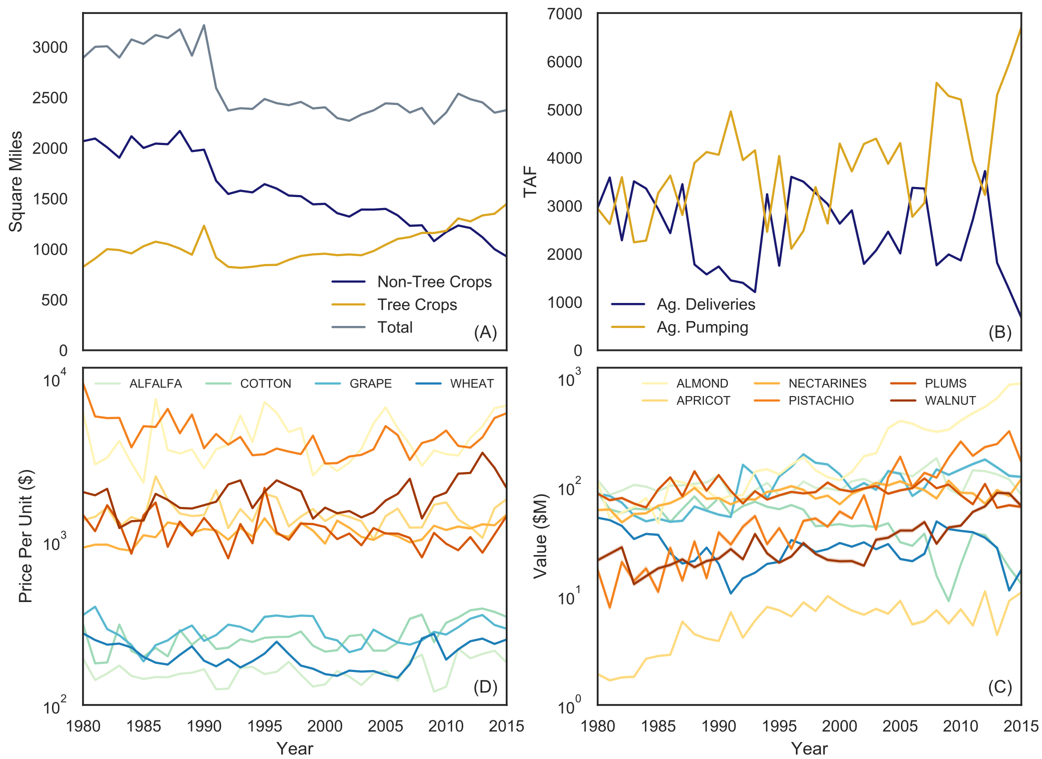


**Figure 4.** Historical trends in heterogeneous feature data. (A) Tree crop acreage, non-tree crop acreage, and total acreage planted; (B) Yearly total agricultural water deliveries and pumping; (C-D) Inflation-adjusted prices and total crop values for a selection of crops.

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 in order to maintain the interpretation of feature variables within the eventual model structures generated by this approach. 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. This spatiotemporal construction process also adds redundancy into the feature set, and we rely on the model search (Section 3.2.2) to navigate this redundancy to identify the most informative features while retaining interpretability.

#### *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.

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. Con-

## 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 $(a, b)$.

stants 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 (M. D. 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 (M. D. Schmidt & Lipson, 2009). 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 $F : \mathbb{R}^n \rightarrow \{Negative, No\ Change, Positive\}$:

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

Though the three classes are relatively balanced in this experiment $\{Negative \sim 25\%, No\ Change \sim 47\%, Positive \sim 28\%\}$, this simple accuracy metric might promote models that perform well on only a subset of classes. This can be a problem particularly when classes are not equally represented in the training set (Provost & Fawcett, 2001). Multi-class metrics such as the macro/micro-averaged F1-measure (Lipton et al., 2014) and receiver operating characteristic (ROC) curve (Fawcett, 2006) can account for class imbalance by weighting measures based on individual class accuracies. However, we find that for this problem, alternate metrics do not significantly change the rank order of models within each class (see Supplemental Material). In regard to improving regression metrics, the water resources field has thoroughly considered how error metrics for natural process models can incorporate available process knowledge (Gupta et al., 2009; Khatami et al., 2019; Lamontagne et al., 2020, e.g.,). These approaches are also relevant in scenarios lacking process knowledge but with known statistical relationships in the error signals.

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 (Vanneschi et al., 2010), 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 a customized genetic programming algorithm, an evolutionary approach that encodes mathematical expressions in a tree structure to support symbolic regression. Modular components of the algorithm were drawn from the package Distributed Evolutionary Algorithms in Python, or DEAP (De Rainville et al., 2012). As depicted in the Model Generation panel of Figure 2, mutation and crossover operators act on ordered representations of models, where each tree is flattened into an ordered list of elements, to generate new structures from promising candidates and explore the model space during optimization. The mutation operator adds a randomly initialized sub-tree of depth 1-2, representing a random addition into the model element list. Single-point crossover randomly selects a location along paired model element lists and exchanges the elements beyond this location to generate a new model, an example of which is depicted in Figure 5. 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 models are discarded before evaluation.

During training, the performance and complexity objectives are both minimized. 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. The algorithm follows a $\mu$+$\lambda$ evolution strategy, which allows parents to persist in the population. At each generation, a number of offspring $\mu$ are generated from $\lambda$ parents in the population by ap-

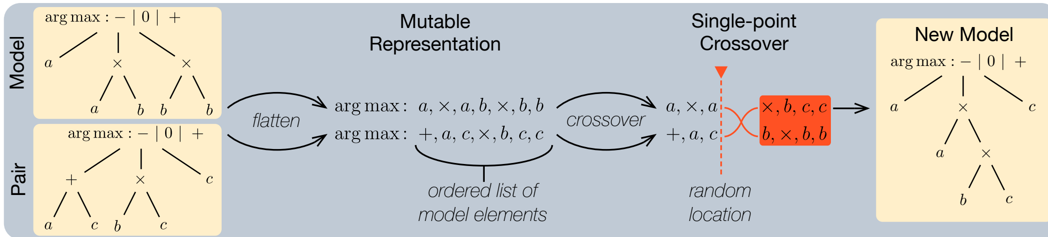


**Figure 5.** Detailed view of crossover operations, expanded from Figure 2. Two models from the population are used to generate a new model by splitting and recombining the ordered list representations at a random location, a process repeated throughout the search. Mutation operates similarly by adding a random sub-tree at a random location in a single model.

plying mutation and crossover with a given probability. The population is updated by applying deterministic crowding selection analogous to NSGA-II (Deb et al., 2000) to the collection of individuals $\mu+\lambda$, selecting $\lambda$ individuals to be used as parents in the next generation. The use of deterministic crowding for selection is intended to promote diversity within populations by spacing out models along the Pareto front. This ensures that no single model dominates in all objectives and is therefore used to generate all new individuals in the next generation. Separately from the population, an archive of Pareto-approximate model structures is maintained and updated through strict non-dominated sorting of the archive and population together in each generation, with no crowding distance selection applied. This archive represents the best approximation of the Pareto front at each iteration of the optimization (including the final result), and allays degradation known to occur in populations when using deterministic crowding for selection.

Experiments were run using the UC Davis College of Engineering HPC1 Cluster with 96 processors, employing DEAP package support for distributed computing. Each population of models contains 96 individuals, and each tree is initialized randomly with depth 1-3. Trials run for a maximum of 20,000 generations with a stagnation convergence criterion of 2,500 generations, which will stop the algorithm if performance improvements are not detected during this time. Performance improvements can be found throughout the optimization, but can become exceedingly small as models start to overfit. As in many high-dimensional sampling problems, it is not possible to prove that the global optimum has been reached. Though the algorithm is likely to comprehensively sample low-complexity models, the size of the primitive set (number of inputs, constants, and functions) dictates that the sampling coverage of possible models decreases at least factorially with additional model primitives (Knuth, 2011). Combinatorial expansion reflects the curse of dimensionality, and complicates the search for medium- and high-complexity models, though more efficient algorithms are an active area of research (Hadka & Reed, 2013; Vrugt & Beven, 2018; Conti et al., 2018, e.g.,). This complexity increases the likelihood of optimization trials getting stuck in local minima as trees grow, and emphasizes the importance of appropriately defining the model space during problem definition. To account for this stochasticity in optimization, 21 randomized trials are performed, which includes the initialization of the train-test split. The code to reproduce this study can be found at DOI: 10.5281/zenodo.3887360.

This algorithm configuration may generate spurious structure and/or redundant features within the same model. The Supplemental Material includes more details about the feature variables and their correlation. However, the algorithm performs variable selection to some extent when feature variables that lead to improved objective performance are introduced through mutation or crossover, suggesting an informative relationship. Even with correlated features, we expect that over the course of many iterations of the mutation operator, and multiple random seeds, the most informative features will occur most often in the resulting sets of models. This stochasticity in model structural identification reinforces the need for multiple trials, ensemble averaging across optimization trials during model evaluation, and summary statistics describing high-complexity regions of the model space, as any one model structure by itself may be subject to feature redundancy.

### 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 using Sobol sensitivity analysis with the Python package SALib (Herman & Usher, 2017). The goal of this sensitivity analysis is to determine whether the different clusters of model behavior are influenced by different feature variables, for example if certain features appear primarily in overfit models. To perform this step, 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. For example, if a model has 5 unique inputs, the model would be tested with 5000 samples for each unique input to appropriately characterize pairwise and total-order sensitivities in the Sobol method.

## 4 Results

### 4.1 Model Performance-Complexity Tradeoff

Figure 6 shows the tradeoff between model performance and complexity across the Pareto set of candidate model structures for both (a) regression and (b) classification experiments. Each point represents the performance (MSE) on the test data, while the gold background shading shows the distribution of performance for the same set of models on the training data. Figure 6 highlights four different regions: Parsimonious, Overfit, Equifinal, and Dominated model clusters. These designations are subjective, but separate the models for discussion according to their primary evaluation characteristics. During each trial, initial structure building occurs in the Parsimonious cluster in both Figure 6a and 6b. The Overfit clusters in Figure 6 are highlighted as the regions where models begin to rely on spurious structure discovered later in the trial. The Equifinal cluster in Figure 6a represents a region where multiple model structures exist at roughly the same level of performance. The Dominated cluster in Figure 6b represents models that are both relatively complex and do not generalize well to unseen data.

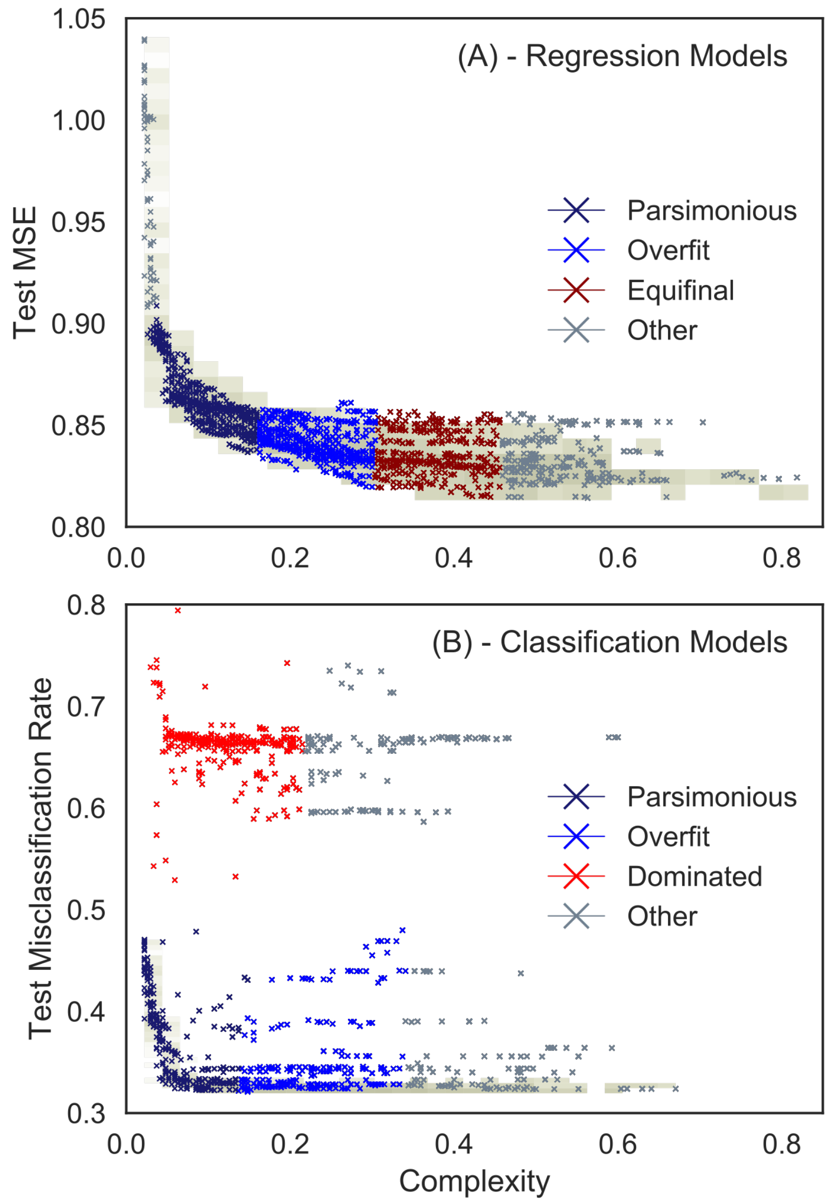


**Figure 6.** Tradeoff between performance (test error) and complexity for model structures across (A) all regression trials and (B) all classification trials. Light gold shading indicates the distribution of the same models evaluated on the training data. Models are clustered according to their behavior in this three-dimensional space (training error, test error, and complexity).

These results indicate several points. First, regression trials in Figure 6a 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. The progress of the optimization trials is determined by the model structures developed in the Parsimonious clusters; insufficient exploration may explain why significant overfitting occurs in Figure 6b. Equifinal model structures are observed in both cases, as many models with increasing complexity demonstrate similar performance.

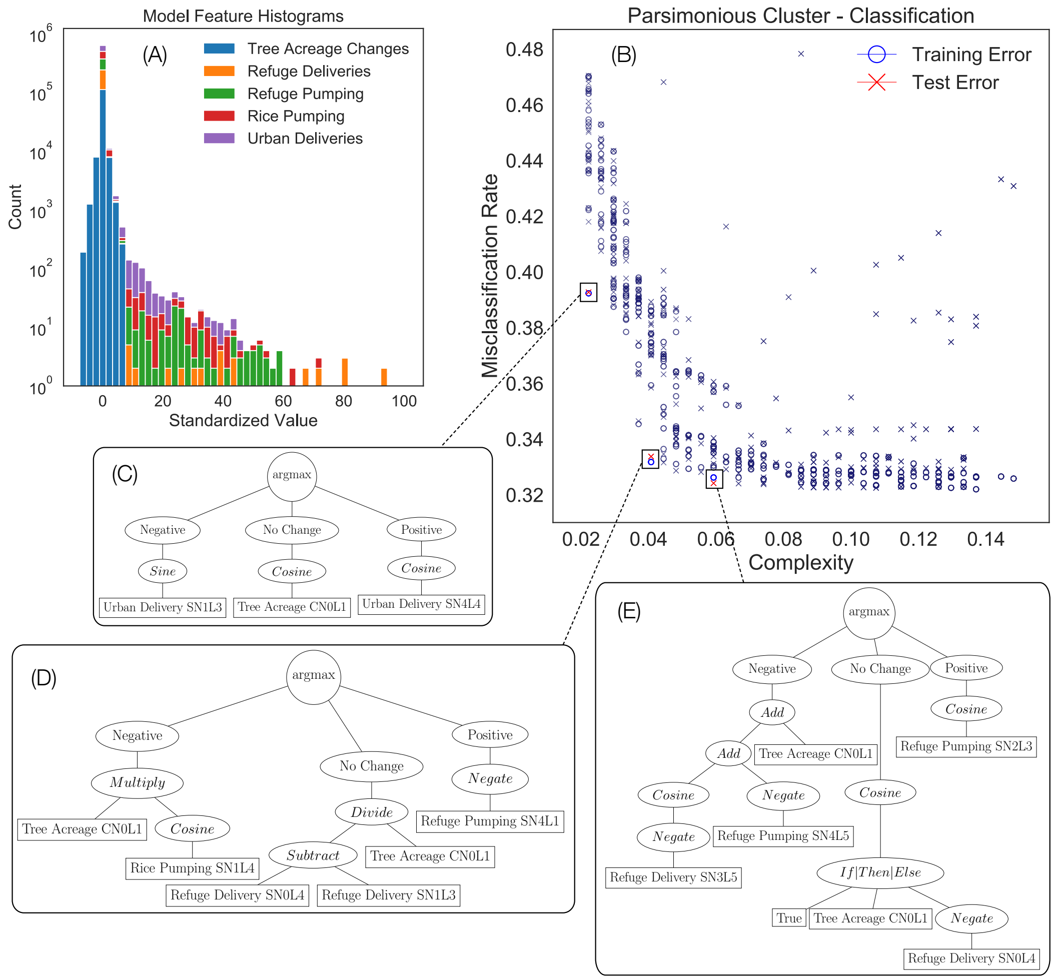


**Figure 7.** (a) Histograms of standardized feature data ($\frac{x-\mu}{\sigma}$) represented in the models; (b) Training and test error for models in the Parsimonious cluster; (c-e) selection of models from the Parsimonious cluster. Feature constructions are annotated as {**S**tate/**C**hange | **N**eighbor **0–5** | **L**ag **1-5**}. In (a), some of these distributions are asymmetric even after standardization; skewed features such as refuge deliveries and pumping occur more infrequently than the relatively balanced tree acreage changes. In (c-e), the $\arg\max$ operator returns the class $\{Negative, No\ Change, Positive\}$ of maximum value for a given sample.

The classification results in Figure 6b show model structures with a variety of macroscopic behavior that can be investigated further. We proceed with the classification results to determine the drivers of model behavior, and also to examine the structure of three models selected from the Parsimonious cluster that perform well on both training and test data in Figure 7. These three classification models depend on a variety of feature variables and structural elements. Figure 7a displays a histogram of standardized feature data represented in the models to understand any patterns shared among the distributions of feature variables selected by the algorithm for these three model structures. While the models occasionally rely on sparse, skewed feature distributions such as non-agricultural water use, they mainly rely on tree acreage changes. Specifically, all three models use the acreage change in the previous timestep (lag-1) and same location, indicating that decision-making agents are informed by past decisions. Additionally, the tree acreage change feature tends to occur closer to the output of each model structure (Figure 7c-e), and as a result is less modified than other features by the sequence of arithmetic operations in each model.

### 4.2 Feature Occurrence and Sensitivity

Large differences among models regarding the selection of other feature variables indicate that some of these structural components may be spurious. The distribution of features chosen by the algorithm might be a result of their different spatiotemporal resolutions. For example, the lack of consensus on the use of economic data could be due to its coarser resolution in space and limited coverage in time, or the inability of the search method to find informative features beyond the lag-1 tree acreage change. To investigate this further, we aim to identify the structural drivers separating robust models in the Parsimonious and Overfit Clusters from models that do not generalize well (i.e., the Dominated cluster). First, we start by analyzing the occurrence of features and function primitives among models in each cluster, displayed in Figure 8.

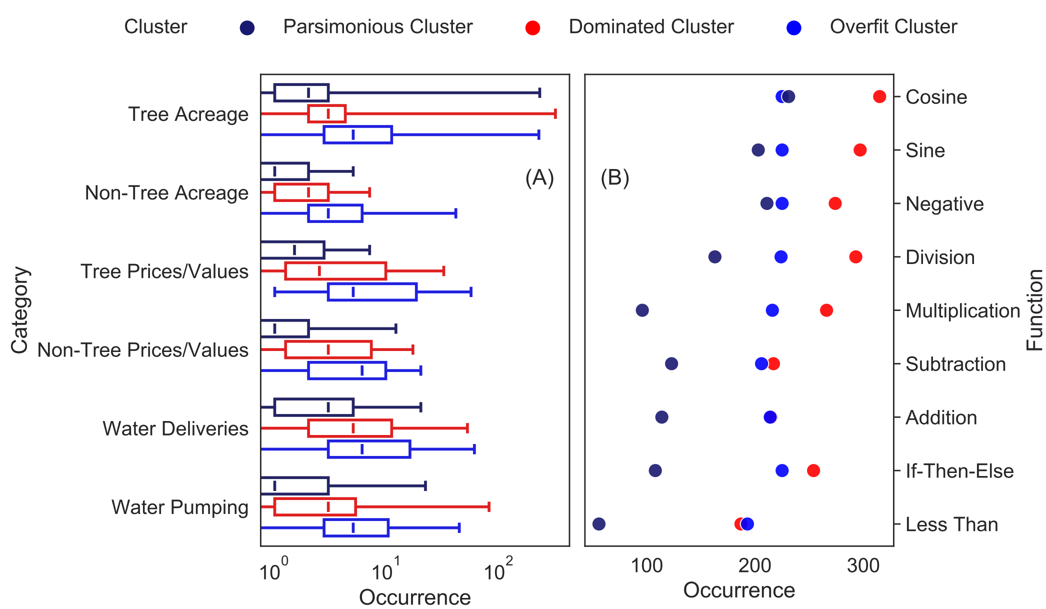


**Figure 8.** Occurrence of feature variables and function primitives among classification models. (A) Distribution of occurrence by feature category; (B) occurrence of functions. The former is a distribution because each feature category contains multiple feature variables, while the functions are not grouped.

Figure 8a shows the distribution of feature occurrence counts in each model cluster, where the features are grouped into categories (y-axis). The boxplots and ranges suggest several key points. All model clusters show a dependence on the group of inputs related to tree acreage data (all lagged and neighboring states and values for tree crops). The lag-1 tree acreage change in the same location (categorized under Tree Acreage) appear in every model across all clusters, indicated by the range of the whiskers at the top of Figure 8a. The Overfit cluster contains more instances of features from each category as compared to the Dominated and Parsimonious clusters, suggesting a higher level of feature complexity overall. Lastly, the largest differences in feature usage between the Parsimonious cluster and the Overfit cluster is in the tree and non-tree prices/values and water pumping feature categories.

Figure 8b shows the occurrence count of function primitives for models in each cluster; the primitives are not categorized into groups, so the values are a single count rather than a distribution. The Overfit cluster exhibits a more even distribution of function occurrence across primitives than the Parsimonious and Dominated clusters, suggesting an increase in the diversity of function primitives relative to the Parsimonious cluster. Both the Overfit cluster and Dominated cluster learn a dependence on the two conditional primitives. Finally, models in the Dominated cluster contain more instances of nearly every function type, particularly deviating from the Overfit and Parsimonious clusters for single-input functions, suggesting a higher level of functional complexity and feature transformations than either the Overfit or Parsimonious clusters.

Figure 8a-b together indicate that robustness to test data may be extended for models in the Parsimonious cluster by increasing reliance on feature complexity versus functional complexity. This contrast may also explain why additional complexity in two- and three-input functions for combining features is warranted over single-input functions that merely transform individual features. However, feature occurrence alone does not explain which features drive model output. Model responses to feature variable changes are quantified using Sobol sensitivity analysis. Results for total sensitivity indices are presented in Figure 9 as empirical cumulative distributions. The sensitivities are presented for two categories of feature variables, tree acreage and non-tree acreage, across the three clusters of classification models.

Figure 9 shows that over 60% of the tree acreage features (including lagged and neighboring feature occurrences) in models from the Overfit cluster have a total-order sensitivity index near zero, meaning that these features have a negligible effect on the class prediction. Both the Overfit and Dominated models show lower sensitivity to both categories of features relative to the Parsimonious cluster, indicating that the best-performing models are driven by a wider range of features. In the case of tree acreage inputs, over 70% of features in the Overfit cluster show small sensitivities ($S_T < 0.2$) 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 models are high ($S_T > 0.8$), illustrating a high reliance on fewer feature variables, which may reduce the ability of these models to generalize out-of-sample. Conversely, both the Overfit and Dominated models do not show the same high sensitivities to non-tree acreage data that appear in the Parsimonious models.

This result confirms the conclusion from Figure 8 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 crop switching requires respecializing and alternate scheduling, it is perhaps unexpected that over 60% of non-tree crop features had negligibly small impacts on the class prediction. Similarly, there were very few features with sensitivity indices greater than 0.6 among the Overfit or Dominated models. Sensitivity testing was only applied to features selected during the generation of each model, so the distributions of sensitivity indices are not affected by the frequency with which a feature is included in each model cluster.

Finally, the average total-order sensitivity indices within each feature category and variable construction are displayed across model clusters in Figure 10. Parsimonious mod-

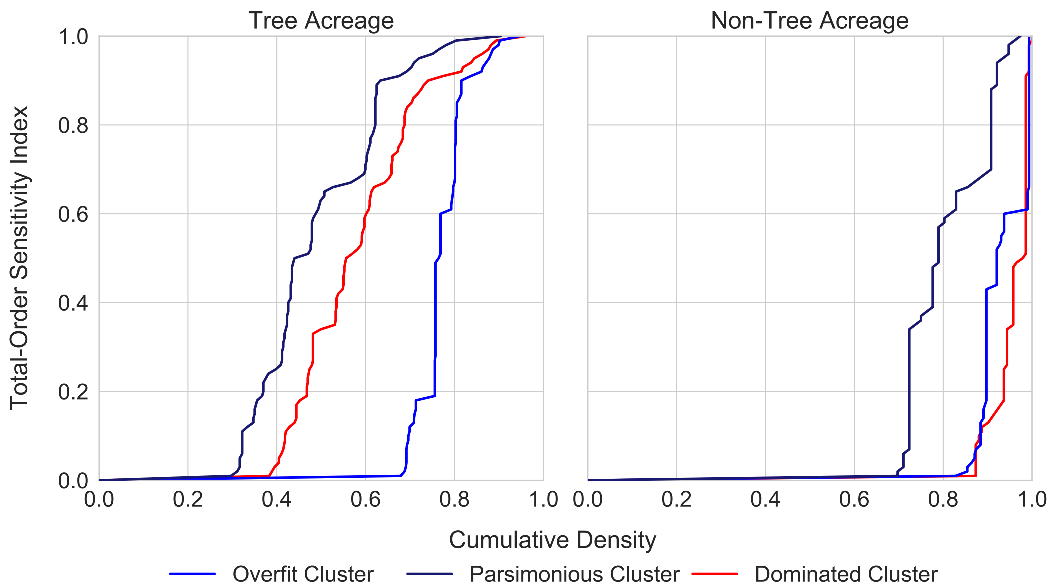


**Figure 9.** Empirical cumulative distribution of total-order sensitivity indices for two categories of feature variables: tree acreage and non-tree acreage, separated by model cluster (color). Only the feature variables appearing in each model were included in the sensitivity analysis.

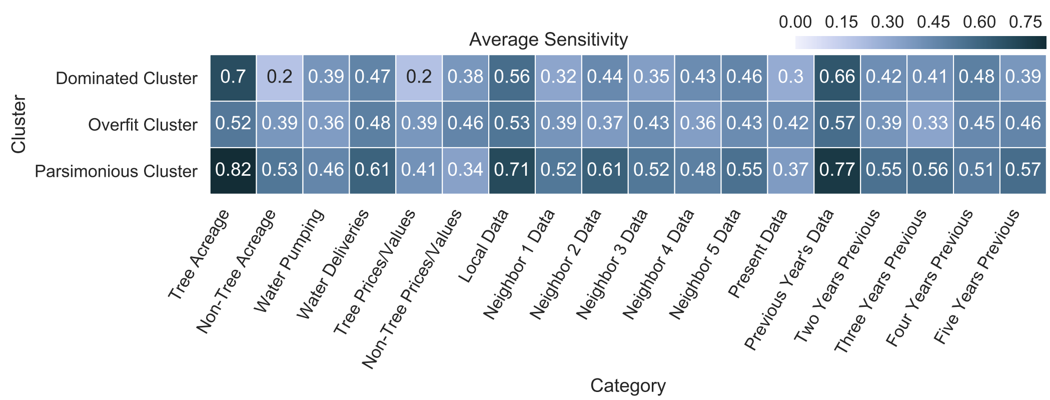


**Figure 10.** Average total-order sensitivity indices of feature variables across input categories for each cluster of model structures. In the feature grouping labels, "data" refers to the combination of state, change, temporal lags, and spatial neighbors for each type of feature.

els demonstrate elevated sensitivities across many of the feature categories and constructions. Since Parsimonious models are less complex than either Overfit or Dominated models, that their predictions are highly affected by a wide range of input variables makes intuitive sense, as Parsimonious models explain a similar fraction of total variance with fewer features. Parsimonious models are also particularly sensitive to tree acreage features from the previous year, as shown in prior figures. Though we might expect high bias in these Parsimonious models, bias seems to be minimized fairly quickly by selective inclusion of feature variables.

Models in the Dominated cluster share high average sensitivity for only some feature categories but not others, as do models from the Overfit cluster to a lesser degree. Models from the Overfit cluster exhibit relatively equal sensitivities across all feature categories as compared to the range of sensitivities represented in the Parsimonious and Dominated clusters. This, combined with the function occurrence result from Figure 8b, suggests that the Overfit models avoid becoming overly sensitive to individual features from any one category by over-engineering the function structure, which likely leads to their improved generalization ability over the Dominated models. This result demonstrates how averaging sensitivity to certain categories of features within model clusters can reveal the extent to which models should be sensitive to feature data given target properties, such as model robustness to unseen data. However, averaging across the set of models may obscure the sensitivities of individual models, the distribution of which is better shown in Figure 9.

## 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; Herman et al., 2020). This must include representations (Schill et al., 2019) and feedbacks (Calvin & Bond-Lamberty, 2018) that may not be implemented in existing model structures. This study proposes methods to automate the exploration of model structure along the canonical tradeoff between performance and complexity to describe human behavior. In this illustrative case study focused on agricultural land use and water demand, enumerating the range of model performance with increasing model complexity by drawing structures from a general, unconstrained space 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) by considering model structures as competing hypotheses (Beven, 2019), which could be compared alongside theory-based models.

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 feature data and primitives (Bongard & Lipson, 2007; M. Schmidt & Lipson, 2009), though informing and bounding search through process understanding and structural priors (Knüsel et al., 2019), constrained problem framings (Dobson et al., 2019; Müller & Levy, 2019, e.g.,), and structured generation schemes (Chadalawada et al., 2020; Spieler et al., 2020, e.g.,), and using advanced interpretation tools post-search (Worland et al., 2019; Quinn et al., 2019, e.g.,) could uncover more specific emergent phenomena in the data and resulting models. However, 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 and evaluating model structure can support the design of integrated system models such as agent-based or hydro-economic models.

This case study was encumbered by two primary sources of difficulty: (1) algorithmic search in combined parametric-structural model spaces, and (2) heterogeneous feature data across multiple temporal and spatial scales. First, the search space of candidate model structures grows combinatorially with the number of features and primitives, 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 6b), a saturation often interpreted as convergence, could be driven by a structural boundary beyond which improvements could not easily be found. Since search effectiveness is partially determined by the size of the model space, available theory regarding target or related processes can be used to plausibly constrain model generation, reinforcing the need for process knowledge alongside data in data-driven analysis (Karpatne et al., 2017; Knüsel et al., 2019). Additionally, 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 this limit is difficult to identify *a priori*. Hybrid methods, such as evolutionary strategies to approximate a gradient, are promising for tractable search in combined model-parameter spaces (Conti et al., 2018; Miikkulainen et al., 2019), as well as approaches that asynchronously tune parameters and structure (Frankle & Carbin, 2018). However, even when appropriately complex models can be identified, their often black-box nature does not guarantee interpretability. The results presented here indicate how increasing equifinality as a function of complexity can inhibit interpretability. Diminishing returns to model accuracy as complexity increases highlight the importance of parsimony as a key model evaluation and selection mechanism. More strategic analysis can be done to interpret the underlying logic behind model predictions, such as explaining the importance of features and structure in neural networks (Montavon et al., 2018; Worland et al., 2019, e.g.,), and using sensitivity analysis to explicate structural dependence in space and time (Quinn et al., 2019, e.g.,).

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, a task which may be better served by developing an ensemble of functions across the spatial region. Additionally, while the feature engineering applied to the data helps discern the importance of correlations in space and time, it also obfuscates the resulting model structures by increasing the interdependence among features. This could be resolved in future work with dimension reduction techniques (Giuliani & Herman, 2018; Cominola et al., 2019), potentially at the cost of feature interpretability. 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. Many-class classification could provide a middle ground between these two tasks, as well as the incorporation of metrics that more realistically reflect model accuracy across classes, such as weighting by class prevalence (Provost & Fawcett, 2001; Lipton et al., 2014) or adding process-informed definitions of model error as objectives (Gupta et al., 2009; Lamontagne et al., 2020). Using 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 priors are assumed on the model search space beyond the function primitives and feature data, including some 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. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0304380011003802` (Modeling Human Decisions) doi: https://doi.org/10.1016/j.ecolmodel.2011.07.010

Anderies, J. M. (2015). Understanding the dynamics of sustainable social-ecological systems: Human behavior, institutions, and regulatory feedback networks. *Bulletin of Mathematical Biology*, *77*(2), 259–280. Retrieved from `https://doi.org/10.1007/s11538-014-0030-z` doi: 10.1007/s11538-014-0030-z

Barto, A. G., & Dietterich, T. G. (2004). Reinforcement learning and its relationship to supervised learning. *Handbook of learning and approximate dynamic programming*, *10*, 9780470544785.

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

Berkes, F., & Folke, C. (Eds.). (1998). *Linking social and ecological systems: Management practices and social mechanisms for building resilience.* Cambridge University Press.

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., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279-298. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.3360060305` doi: 10.1002/hyp.3360060305

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, jun). 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

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., Agrawal, S., Pratap, A., & Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In (pp. 849–858). Springer.

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.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. Retrieved from http://www.sciencedirect.com/science/article/pii/S016786550500303X doi: https://doi.org/10.1016/j.patrec.2005.10.010

Frankle, J., & Carbin, M. (2018). *The lottery ticket hypothesis: Finding sparse, trainable neural networks.*

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

Groeneveld, J., Müller, B., Buchmann, C. M., Dressler, G., Guo, C., Hase, N., ... Schwarz, N. (2017). Theoretical foundations of human decision-making in agent-based land use models –a review. *Environmental Modelling & Software*, *87*, 39–48. Retrieved from http://www.sciencedirect.com/

science/article/pii/S1364815216308684 doi: https://doi.org/10.1016/j.envsoft.2016.10.008

Gunaratne, C., & Garibay, I. (2017). Alternate social theory discovery using genetic programming: Towards better understanding the artificial anasazi. In *Proceedings of the genetic and evolutionary computation conference* (p. 115–122). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3071178.3071332 doi: 10.1145/3071178.3071332

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022169409004843 doi: https://doi.org/10.1016/j.jhydrol.2009.08.003

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

Hadka, D., & Reed, P. (2013). Borg: An auto-adaptive many-objective evolutionary computing framework. *Evolutionary Computation*, *21*(2), 231-259. Retrieved from https://doi.org/10.1162/EVCO_a_00075 (PMID: 22385134) doi: 10.1162/EVCO_a_00075

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

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., 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

Howitt, R. E., Medellín-Azuara, J., MacEwan, D., & Lund, J. R. (2012). Calibrating disaggregate economic models of agricultural production and water management. *Environmental Modelling & Software*, *38*, 244–258. Retrieved from http://www.sciencedirect.com/science/article/pii/S136481521200196X doi: https://doi.org/10.1016/j.envsoft.2012.06.013

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

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.

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

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ... Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318-2331.

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

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resources Research*, *56*(9), e2019WR025975. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR025975 (e2019WR025975 10.1029/2019WR025975) doi: 10.1029/2019WR025975

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

Knuth, D. E. (2011). *The art of computer programming: Combinatorial algorithms, part 1* (1st ed.). Addison-Wesley Professional.

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).

Lamontagne, J. R., Barber, C. A., & Vogel, R. M. (2020). Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resources Research*, *56*(9), e2020WR027101. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR027101 (e2020WR027101 2020WR027101) doi: 10.1029/2020WR027101

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

Lipton, Z. C., Elkan, C., & Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize F1 measure. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *Machine learning and knowledge discovery in databases* (pp. 225–239). Berlin, Heidelberg: Springer Berlin Heidelberg.

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

Malek, Ž., & Verburg, P. H. (2020). Mapping global patterns of land use decision-making. *Global Environmental Change*, *65*, 102170. Retrieved from http://www.sciencedirect.com/science/article/pii/S0959378020307536 doi: https://doi.org/10.1016/j.gloenvcha.2020.102170

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

Melsen, L. A., Vos, J., & Boelens, R. (2018). What is the role of the model in socio-hydrology? Discussion of "Prediction in a socio-hydrological world". *Hydrological Sciences Journal*, *63*(9), 1435-1443. Retrieved from https://doi.org/10.1080/02626667.2018.1499025 doi: 10.1080/02626667.2018.1499025

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.

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., & Levy, M. C. (2019). Complementary vantage points: Integrating hydrology and economics for sociohydrologic knowledge generation. *Water Resources Research*, *55*(4), 2549-2571. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR024786 doi: 10.1029/2019WR024786

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

Nelson, K. S., & Burchfield, E. K. (2017). Effects of the structure of water rights on agricultural production during drought: A spatiotemporal analysis of California's Central Valley. *Water Resources Research*, *53*(10), 8293-8309. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017WR020666 doi: 10.1002/2017WR020666

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

Prestele, R., Arneth, A., Bondeau, A., de Noblet-Ducoudré, N., Pugh, T. A. M., Sitch, S., ... Verburg, P. H. (2017). Current challenges of implementing anthropogenic land-use and land-cover change in models contributing to climate change assessments. *Earth System Dynamics*, *8*(2), 369–386. Retrieved from https://esd.copernicus.org/articles/8/369/2017/ doi: 10.5194/esd-8-369-2017

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*(3), 203–231. Retrieved from https://doi.org/10.1023/A:1007601015854 doi: 10.1023/A:1007601015854

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

Quade, M., Abel, M., Shafi, K., Niven, R. K., & Noack, B. R. (2016, Jul). Prediction of dynamical systems by symbolic regression. *Phys. Rev. E*, *94*, 012214. Retrieved from https://link.aps.org/doi/10.1103/PhysRevE.94.012214 doi: 10.1103/PhysRevE.94.012214

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. Retrieved from https://doi.org/10.1007/BF00116251 doi: 10.1007/BF00116251

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

Robinson, D. T., Brown, D. G., Parker, D. C., Schreinemachers, P., Janssen, M. A., Huigen, M., ... Barnaud, C. (2007). Comparison of empirical methods for building agent-based models in land use science. *Journal of Land Use Science*, *2*(1), 31-55. Retrieved from https://doi.org/10.1080/17474230701201349 doi: 10.1080/17474230701201349

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

Schlüter, M., Baeza, A., Dressler, G., Frank, K., Groeneveld, J., Jager, W., . . . Wijermans, N. (2017). A framework for mapping and comparing behavioural theories in models of social-ecological systems. *Ecological Economics*, *131*, 21–35. Retrieved from http://www.sciencedirect.com/science/article/pii/S0921800915306133 doi: https://doi.org/10.1016/j.ecolecon.2016.08.008

Schlüter, M., Orach, K., Lindkvist, E., Martin, R., Wijermans, N., Bodin, Ö., & Boonstra, W. J. (2019). Toward a methodology for explaining and theorizing about social-ecological phenomena. *Current Opinion in Environmental Sustainability*, *39*, 44–53. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877343519300491 doi: https://doi.org/10.1016/j.cosust.2019.06.011

Schmidt, M., & Lipson, H. (2009). 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. D., & 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. D., & 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. D., & Lipson, H. (2009). 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

Schmidt, P. J., Emelko, M. B., & Thompson, M. E. (2020). Recognizing structural nonidentifiability: When experiments do not provide information about important parameters and misleading models can still have great fit. *Risk Analysis*, *40*(2), 352-369. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13386 doi: 10.1111/risa.13386

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., 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

Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., & Schütze, N. (2020). Automatic model structure identification for conceptual hydrologic models. *Water Resources Research*, *56*(9), e2019WR027009. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR027009 (e2019WR027009 10.1029/2019WR027009) doi: 10.1029/2019WR027009

Stehfest, E., van Zeist, W.-J., Valin, H., Havlik, P., Popp, A., Kyle, P., . . . Wiebe, K. (2019). Key determinants of global land-use projections. *Nature Communications*, *10*(1), 2166. Retrieved from https://doi.org/10.1038/s41467-019-09945-w doi: 10.1038/s41467-019-09945-w

Sun, B., & Robinson, D. T. (2018). Comparison of statistical approaches for modelling land-use change. *Land*, *7*(4). Retrieved from https://www.mdpi.com/2073-445X/7/4/144 doi: 10.3390/land7040144

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

Thober, J., Müller, B., Groeneveld, J., & Grimm, V. (2017). Agent-based modelling of social-ecological systems: Achievements, challenges, and a way forward. *Journal of Artificial Societies and Social Simulation*, *20*(2), 8. Retrieved from http://jasss.soc.surrey.ac.uk/20/2/8.html doi: 10.18564/jasss.3423

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., Alexander, P., Evans, T., Magliocca, N. R., Malek, Z., Rounsevell, M. D., & van Vliet, J. (2019). Beyond land cover change: towards a new generation of land use models. *Current Opinion in Environmental Sustainability*, *38*, 77–85. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877343518301362 doi: https://doi.org/10.1016/j.cosust.2019.05.002

Vrugt, J. A., & Beven, K. J. (2018). Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM(LOA) algorithm. *Journal of Hydrology*, *559*, 954–971. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022169418301021 doi: https://doi.org/10.1016/j.jhydrol.2018.02.026

Vu, T. M., Probst, C., Epstein, J. M., Brennan, A., Strong, M., & Purshouse, R. C. (2019). Toward inverse generative social science using multi-objective genetic programming. In *Proceedings of the genetic and evolutionary computation conference* (p. 1356–1363). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3321707.3321840 doi: 10.1145/3321707.3321840

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. *Integrated 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

Williams, T. G., Guikema, S. D., Brown, D. G., & Agrawal, A. (2020). Assessing model equifinality for robust policy analysis in complex socio-environmental systems. *Environmental Modelling & Software*, 104831. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364815220308884 doi: https://doi.org/10.1016/j.envsoft.2020.104831

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

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

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

Liam Ekblad[1], Jonathan D. Herman[1]

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

## Contents of this file

| Land Feature Data | | Water Feature Data | Economic Feature Data | |
|---|---|---|---|---|
| Tree Crops | Non-Tree Crops | Non-Ponded Crop Deliveries | Alfalfa | Almond |
| | | Non-Ponded Crop Pumping | Apricot | Beeswax |
| | | Rice Crop Deliveries | Cotton | Grape |
| | | Rice Crop Pumping | Honey | Milk |
| | | Urban Deliveries | Nectarine | Pistachio |
| | | Urban Pumping | Plum | Walnut |
| | | Refuge Deliveries | Wheat | |
| | | Refuge Pumping | | |
| | | Total Pumping | | |

**Table S1.** Feature data used to generate models during the experiment.


**Figure S1.** Nonlinear correlations for a subset of the features represented in the table above.

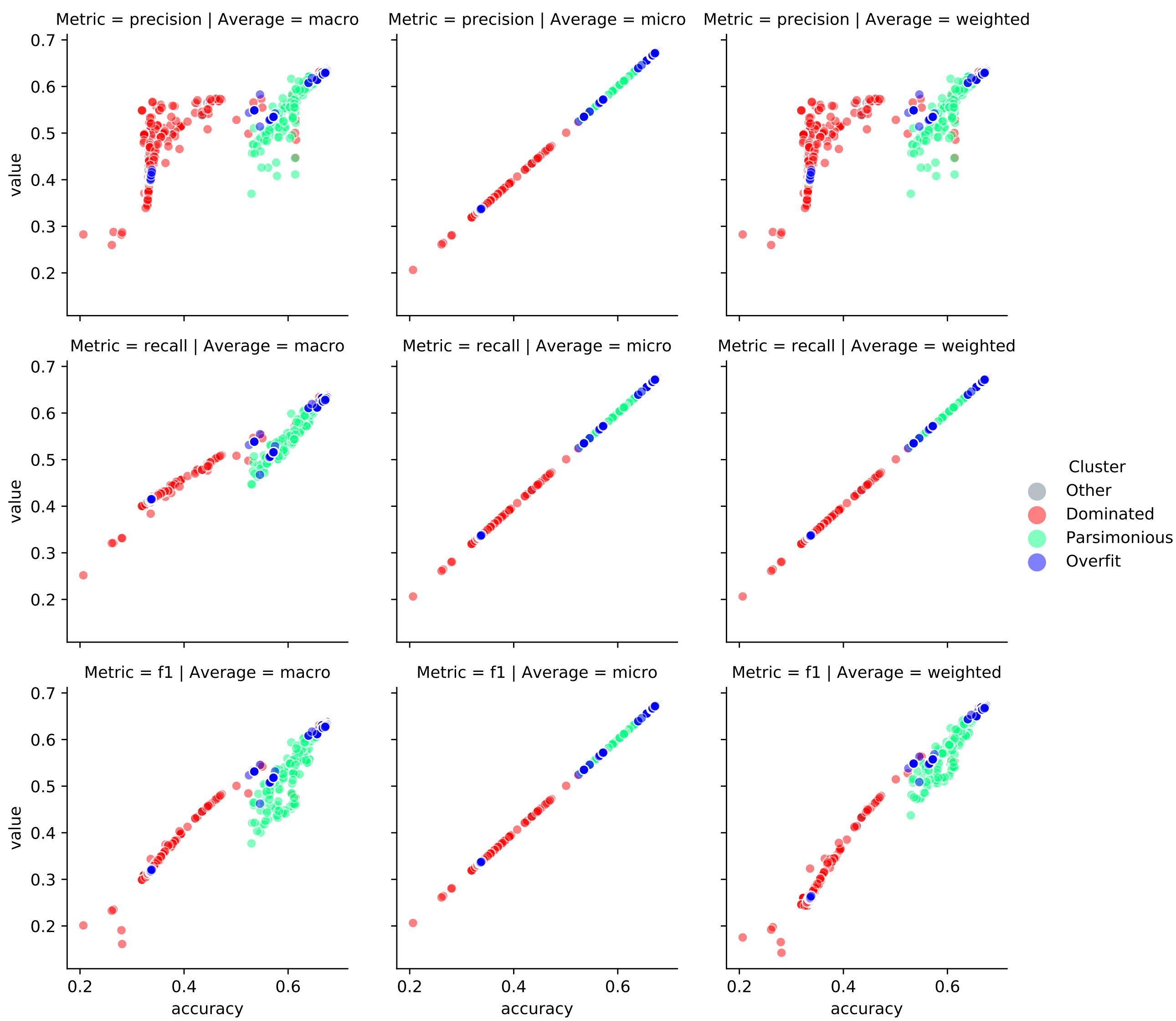


**Figure S2.** Visualization of classification model performance metrics in relation to the simple accuracy metric used in the paper.

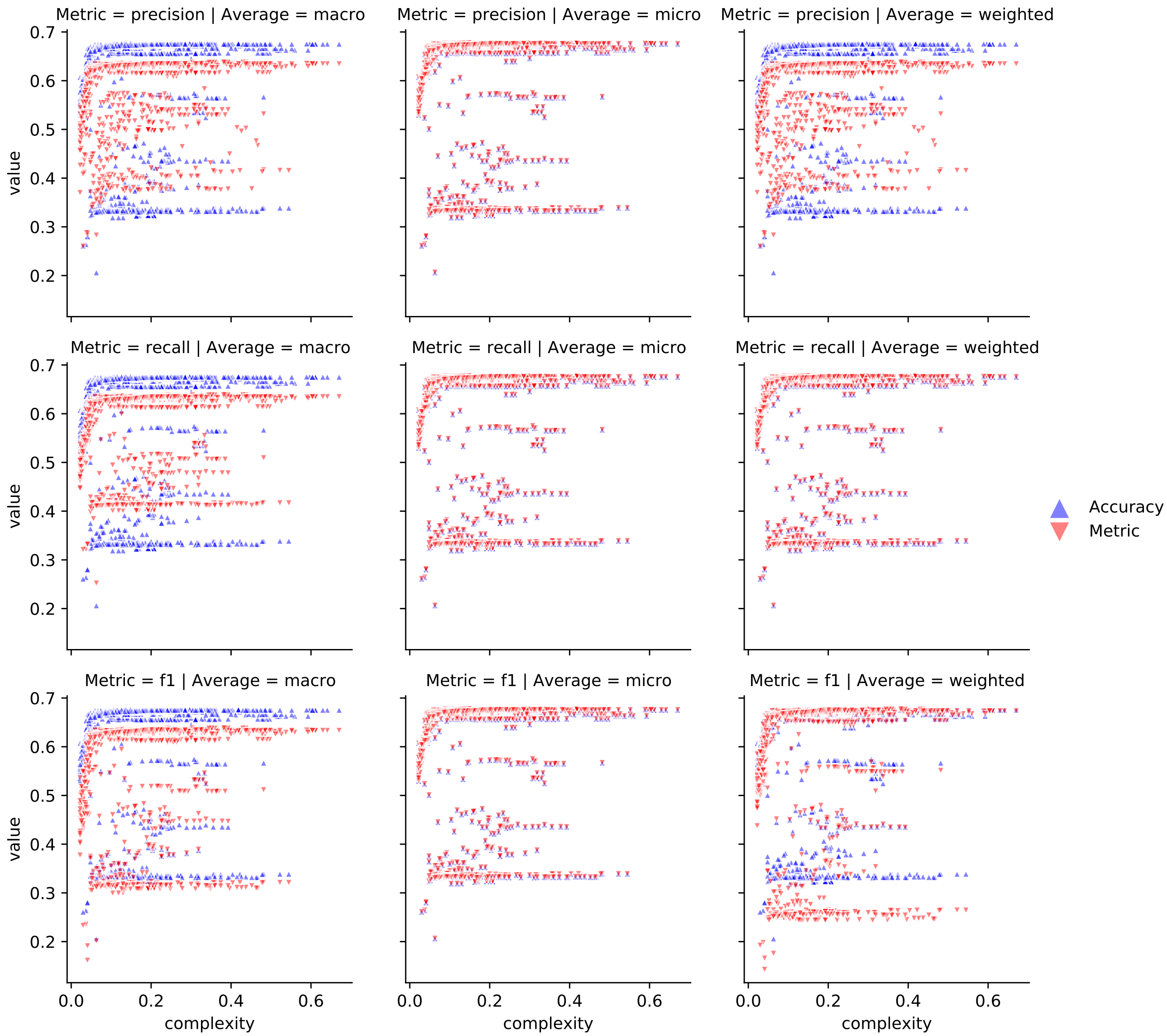


**Figure S3.** Visualization of the effect of classification model performance metric selection on the resultant performance-complexity tradeoff.