UC Davis
UC Davis Previously Published Works

Title
Machine Learning Predicts Reach��Scale Channel Types From Coarse��Scale Geospatial Data in a Large River Basin

Permalink
https://escholarship.org/uc/item/0qf3w3m6

Journal
Water Resources Research, 56(3)

ISSN
0043-1397 1944-7973

Authors
Guillon, Hervé
Byrne, Colin F
Lane, Belize A
et al.

Publication Date
2020-03-11

DOI
10.1029/2019WR026691

Peer reviewed

Machine Learning Predicts Reach-Scale Channel Types From Coarse-Scale Geospatial Data in a Large River Basin

Hervé Guillon,[1,*] Colin F. Byrne,[1] Belize A. Lane,[2] Samuel Sandoval Solis,[1] and Gregory B. Pasternack[1]

[1] *University of California Davis, Davis, CA, United States*
[2] *Utah State University, Logan, UT, United States*

ABSTRACT

Hydrologic and geomorphic classifications have gained traction in response to the increasing need for basin-wide water resources management. Regardless of the selected classification scheme, an open scientific challenge is how to extend information from limited field sites to classify tens of thousands to millions of channel reaches across a basin. To address this spatial scaling challenge, this study leverages machine learning to predict reach-scale geomorphic channel types using publicly available geospatial data. A bottom-up machine learning approach selects the most accurate and stable model among ∼20,000 combinations of 287 coarse geospatial predictors, preprocessing methods and algorithms in a three-tiered framework to: (i) define a tractable problem and reduce predictor noise; (ii) assess model performance in statistical learning; and (iii) assess model performance in prediction. This study also addresses key issues related to the design, interpretation and diagnosis of machine learning models in hydrologic sciences. In an application to the Sacramento River basin (California, USA), the developed framework selects a Random Forest model to predict 10 channel types previously determined from 290 field surveys over 108,943 two hundred-meter reaches. Performance in statistical learning is reasonable with a 61% median cross-validation accuracy – a sixfold increase over the 10% accuracy of the baseline random model, and the predictions coherently capture the large-scale geomorphic organization of the landscape. Interestingly, in the study area, the persistent roughness of the topography partially controls channel types and the variation in the entropy-based predictive performance is explained by imperfect training information and scale mismatch between labels and predictors.

I. INTRODUCTION

Classification is commonly used to characterize and interpret natural systems. In fluvial geomorphology, the observed geometry and behavior of a river characterize the distinct set of physical processes shaping a stream. This form-process paradigm [1] gave rise to a variety of reach-scale geomorphic classifications [e.g., 2, 3]. With increasing collaborations between scientists and stakeholders, basin-wide hierarchical classifications such as the River Styles framework have gained traction [e.g., 4–9]. Such classifications usually rely on field-based observations of valley confinement [10], channel geometry (e.g., width and sinuosity) and instream features (e.g., bar and pool). These mainly descriptive classifications have a fuzzy correspondence [11], suggesting that automation and standardization of reach-scale classification may not only be possible but beneficial [12, 13]. To that end, reproducible, physically interpretable river characterization within a GIS framework represents an important step-forward [e.g., 14–18].

Regardless of the classification scheme selected, an open scientific challenge is how to extend a field-based classification generated from a limited number of study sites to the regional scale over tens of thousands to millions of channel reaches to support regional water management efforts. Given the strong relationship between upstream catchment properties accessible through public geospatial data sets [e.g., 19] and streamflow response, extrapolating hydrologic classes to the stream network is well-established [e.g., 20]. By contrast, reach-scale geomorphic channel types depend more on local geology and confinement. Information related to these attributes is rarely available at sufficient resolution over entire regions to accurately extrapolate classes to a stream network. O'Brien *et al.* [7] relied on a rich geospatial data set including extensive remote sensing and field mapping to extrapolate channel types in a 2,050 km^2 catchment. Although it is possible to predict channel types using only information in available regional databases, reach-scale attributes in such databases often have significantly different values than those observed in the field [e.g., 21] so direct prediction of channel types without mindful consideration of data scaling problems will likely yield poor results. Machine learning (ML) offers an opportunity to perform such a scaling translation by exploring the potentially sophisticated linkages between coarse-scale geospatial information publicly available at the regional scale and field-based geomorphic surveys.

ML refers to models that improve performance during execution [e.g., 22]. ML models are data driven, scalable, and exhibit high-performance in classification tasks while identifying significant driving variables. Despite evidenced successes with ML in other fields, environmental sciences have been slower to adopt these statistical techniques and combine them

_____

* hguillon@ucdavis.edu

with process-based deterministic approaches. For example, recent reviews underline a limited use of the latest developments of ML in hydrologic sciences [23, 24], a slower adoption in geophysics [25], and emerging applications in Earth System Science [26]. However, with increasing volume of data, ease-of-access to computing resources and availability of ML toolboxes, this situation is rapidly evolving. In geomorphology in particular, novel ML applications include delineating landforms [27], predicting geomorphic disturbance [28] or dune erosion [29], mapping susceptibility to landslide and gully erosion [30–32], inferring ecohydrological parameters [33], analyzing model residuals [34], clustering river profiles [35], classifying and predicting sediment-discharge relationships [36, 37], and assessing stream diversity with large-scale top-down approaches [38, 39]. In the face of such rising popularity, the interpretability and assessment of uncertainty in ML models remain key issues [e.g., 26].

In this study, we develop a multitiered ML framework to extrapolate a geomorphic classification scheme to a regional stream network, assess model performance, and inform additional data collection. This framework is applied to the Sacramento River basin in California (USA) to predict the channel type for 108,943 two hundred-meter long stream intervals. The three tiers of the proposed ML framework (i) define a tractable problem from 290 field surveys and 287 predictors; (ii) assess the performance of $\sim$ 20,000 ML models in statistical learning; and (iii) assess the performance of $\sim$ 20,000 ML models in predictive modeling. In consequence, this study also adds to the growing body of literature describing approaches to better design, interpret and diagnose ML models in hydrological sciences [e.g., 40–43].

The paper is organized as follows: Section 2 presents the study area; section 3 describes the proposed three-tier ML framework used to predict the channel forms; and sections 4 and 5 present our results and discuss their implications.

II. CASE STUDY

Spanning 70,130 km^2, the Sacramento River basin in California (USA) exhibits diverse physiography and hydrology [44]. Flowing southward from its headwaters, the river crosses the Central Valley of California before reaching the Sacramento-San Joaquin river delta and Pacific Ocean through San Francisco Bay. The western and eastern portions of the basin drain the Coastal Range and the Sierra Nevada mountains, respectively. The northern portion transitions from Coastal Range to the southern end of the Cascade Range to the volcanic Modoc Plateau. The climate is Mediterranean with cool wet winters and warm dry summers. Nonetheless, varying precipitation intensity and seasonality yield differences in flow regimes across the basin and more generally across the state [20].

Byrne *et al.* [45] previously identified 10 dominant geomorphic channel types present in the Sacramento River basin from data acquired at 290 field sites (Table I). Channel types were statistically derived based on multivariate clustering and comparison to field observations to achieve a best fit to all field surveyed sites. Canals and ditches were excluded from Byrne et al's. (2019) analysis and, while some field sites exist in areas with human influence, channel types are not defined by anthropogenic alteration. The evidenced natural variability of rivers in the study basin [45, 46] makes it an ideal test bed for testing large-scale geomorphic classification needed for future integrative research [e.g., 47, 48].

III. A MULTITIERED ML FRAMEWORK

In the cognition process defining ML, model parameters are *internally* optimized against some performance metrics and the model *trains* and self-improves. *Hyper*parameters set prior to training define the architecture of the model and *tuning* is used to select hyperparameters leading to best performance. For a polynomial model, the order of the polynomial is the hyperparameter, its coefficients are parameters optimized by minimizing the sum of the squared residuals, and tuning identifies the order yielding the highest performance. Importantly, what separates ML from a curve fitting exercise is ML's ability to generalize patterns. Such generalization is achieved by *resampling* the initial data set into a training set and a test set such that the robustness of the learned pattern can be assessed against data unseen during training.

Learning tasks are usually separated into unsupervised and supervised ML approaches [e.g., 49]. Both approaches use input data, predictor variables (henceforth *predictors*) to extract or predict some information about a data set. Unsupervised learning identifies patterns in input data and uses these patterns to cluster observations. Supervised learning uses known information to approximate the relationship between input and output data which can then be used to predict output given new input. When the distribution of such output is continuous, the supervised learning task is a regression problem. When the output distribution is discrete, it corresponds to *labels*, observations of a set of *classes*, and the supervised learning task is a classification problem. For a regression problem, the learned relationship between output and input corresponds to a mathematical mapping. For a classification problem, what is learned pertains to *class boundaries*, divisions of the multidimensional predictor space separating subsets where observations of a given class are dominantly located.

This study aims to predict channel types for 108,943 two hundred-meter long stream intervals throughout the stream network. As 10 possible channel types have been previously identified [45], a supervised ML classification approach answers the following question: Given a set of predictors, which class (channel type) should be assigned to each stream reach? To achieve this study aim, we developed a three-tiered ML framework (Figure 1) to: (i) define a tractable problem and reduce predictor noise, (ii) assess the performance of models in statistical learning, and (iii) assess the per-

Channel type	Name	Number of observations	Prevalence
1	unconfined, boulder-bedrock, bed undulating	6	0.02
2	confined, boulder, high-gradient, step-pool/cascade	27	0.09
3	confined, boulder-bedrock, uniform	36	0.12
4	confined, boulder-bedrock, low-gradient step-pool	33	0.11
5	confined, gravel-cobble, uniform	43	0.15
6	partly-confined, low width-to-depth, gravel-cobble, riffle-pool	45	0.16
7	partly-confined, cobble-boulder, uniform	33	0.11
8	partly-confined, high width-to-depth, gravel-cobble, riffle-pool	24	0.08
9	unconfined, low width-to-depth, gravel	27	0.09
10	unconfined, gravel-cobble, riffle-pool	16	0.06

Table I: Channel Types Identified in the Sacramento Basin by Byrne *et al.* [45].

formance of models in predictive modeling.

A. Define a Tractable Problem and Reduce Predictor Noise

Our ML approach derives the relationship between classes corresponding to 10 channel types from Byrne *et al.* [45] and 287 predictors (Table II). Several categories of predictors with a documented influence on landscape and channel morphology were considered: (a) channel confinement [e.g., 10]; (b) stream network topology [e.g., 50, 51]; (c) statistical roughness or fractal dimension of the topography [52–59]; (d) contextual variables like geology, soils, land cover and climate [e.g., 60, 61]; and (e) nine Terrain Analysis Metrics (e.g. curvature, 2, 62, 63). Since the distribution of Terrain Analysis Metrics across the landscape differentiates stages of landscape maturity [64] and channel types [46], six Distribution Metrics were estimated: mean, median, minimum, maximum, standard-deviation and skewness. Most predictors were derived from three core data sets: (i) the 10-m National Elevation Dataset [65, NED]; (ii) the stream network from the National Hydrology Dataset [66, NHDPlusV2]; and (iii) the Stream-Catchment Dataset [StreamCat; 19] which aggregates previous data sources.

The essence of the problem addressed here is the prediction of channel types, derived from data at the 10^0 to 10^2 m scale, by predictors available at coarser spatial scales ranging from 10^2 to 10^5 m (Table II). In addition, defining the relevant scales to distinguish controlling physical processes is often difficult [e.g., 67]. To address this challenge, we considered 32 sets of five nested spatial scales with upper scales from ∼0.6 km to ∼82 km when calculating the fractal dimension following the box-counting method from Liucci and Melelli [58]. This can be seen as an alternative to using multiscale topographic decomposition [e.g., 68–71]. In addition, the Terrain Analysis Metrics and Distribution Metrics were estimated over two spatial coverages to represent hillslope and near-channel processes: a 512-m square tile centered at the midpoint of each stream-interval and along a 100-m wide near-channel buffer, respectively. These two spatial coverages, combined with nine Terrain Analysis Metrics (TAM) and six Distribution Metrics result in 108 Terrain Analysis Metrics-Distribution Metrics predictors (Table II).

Although ML approaches can incorporate a large number of predictors and labels, irrelevant or noisy data may deteriorate model performance and increase computational load. Because each channel type was expertly inspected in the unsupervised learning phase by Byrne *et al.* [45], the labels are treated as noiseless and we refrained from implementing techniques that evaluate and correct for uncertain labeling [e.g., 72–75]. To filter potential predictor noise, the complexity of the problem was assessed by calculating data complexity measures (DCM; 76). DCMs offer a model agnostic tool [76–78] to inform on the linearity of the task, the complexity of the class boundaries, and the underlying structure of the observations within the predictor space [e.g., 75]. DCMs were computed to evaluate the dimensionality of the predictor space, its complexity with neighborhood and network metrics, the discriminative power of predictors and the linearity of the task across four predictor spaces: the initial set of predictors and when removing the fractal dimension predictors, the stream network topology metrics and the contextual variables. While we provide a reference table for the DCMs with their name, acronym and use, the full definitions can be found in Lorena *et al.* [76] (Table III).

B. Assess Performance in Statistical Learning

1. Statistical Learning

While numerous ML models exist, the best model for a given task is often unknown at the start [89], leading to the common practice of training multiple models on the same task. Here we define a ML model as the combination of algorithm, preprocessing, predictors, oversampling and resampling. We describe these model components in the following.

Twelve different ML algorithms were trained, including prominent algorithms like support vector machine (SVM), random forest (RF) and artificial neural network (ANN) as well as partial least squares, multivariate adaptive regression splines, flexible discriminant analysis, *k*-nearest neighbors, classification and regression tree, bagged-trees, linear discriminant analysis, regularized linear discriminant analysis and naive bayes. Three key algorithms, SVM, RF and ANN, are described below. Implementation of these algorithms was performed using the R packages `caret` [90, 91] and `h2o` [92].
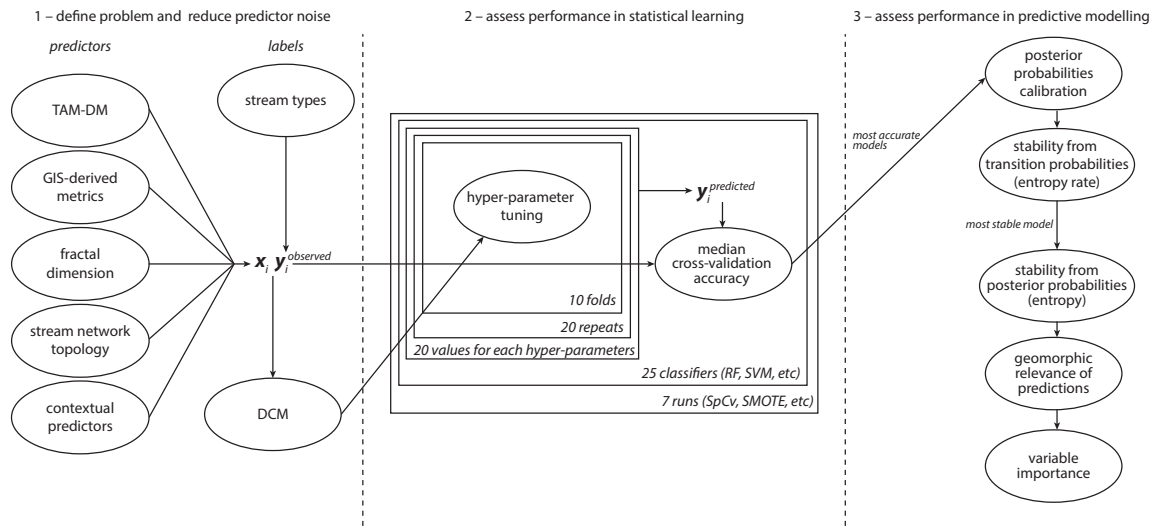
Figure 1: Schematic of the three-tiered machine-learning framework.

Predictors group	Predictor name	Spatial scale	Original data	Methodology
TAM-DM (108)	Elevation	512 m; 100-m buffer	Gesch *et al.* [65]	[79]
	Slope	512 m; 100-m buffer	Gesch *et al.* [65]	[79]
	Aspect	512 m; 100-m buffer	Gesch *et al.* [65]	[79]
	Roughness	512 m; 100-m buffer	Gesch *et al.* [65]	[79]
	Flow direction	512 m; 100-m buffer	Gesch *et al.* [65]	[79]
	Planform curvature	512 m; 100-m buffer	Gesch *et al.* [65]	[80]
	Profile curvature	512 m; 100-m buffer	Gesch *et al.* [65]	[80]
	Topographic position index	512 m; 100-m buffer	Gesch *et al.* [65]	[79]
	Terrain ruggedness index	512 m; 100-m buffer	Gesch *et al.* [65]	[79]
GIS-metrics (3)	Channel slope	200 m	Gesch *et al.* [65]	[81]
	Confinement	-	Gesch *et al.* [65]	[45]
	Sediment supply	-	[82]	Renard *et al.* [83]
Network topology (4)	Drainage area	-	McKay *et al.* [66]	[19]
	Strahler's stream order	-	McKay *et al.* [66]	[50]
	Local drainage density	-	McKay *et al.* [66]	[51]
Fractal dimension (32)	Hurst coefficients	640 m to 82 km	Gesch *et al.* [65]	[58]
Contextual predictors (140)	Lithology	>1 km	[84]	Hill *et al.* [19]
	Soil characteristics	1 km	[85]	Hill *et al.* [19]
	Land cover	30-m initial resolution	[86]	Hill *et al.* [19]
	1981-2010 climatologies	800-m initial resolution	[87]	Hill *et al.* [19]
	Indices of Catchment Integrity	-	[88]	Hill *et al.* [19]

TAM-DM : Terrain Analysis Metrics - Distribution Metrics

Table II: Predictors Used in the Machine Learning Framework. The 10-m National Elevation Data Set [65, NED] and the Stream-Catchment Data Set [StreamCat; 19] are publicly available on download platform from the United States Geological Survey and the United States Environmental Protection Agency, respectively. The stream network from the National Hydrology Data Set [66, NHDPlusV2] is publicly available on both platforms. TAM-DM: Terrain Analysis Metrics-Distribution Metrics

Linear SVM finds the linear boundary between two distinct classes by maximizing the margin between the class boundary and each class's closest point(s) [93]. Those points are the support vectors for the boundary. Nonlinear boundaries are obtained by a nonlinear kernel version of SVM, transforming predictor space so that the problem becomes linearly solvable. The most common kernel used to perform this so-called kernel trick is the radial basis function. SVMs solve multiclass problems by transforming them into a set of two-class problems for which multiple binarization strategies exist [e.g., 94].

RF is an ensemble of classification and regression trees built from random subsets of predictors [95]. At each split of each tree, a predictor is chosen as splitting variable based on an information selection pro-

cess (e.g. Gini coefficient) which ultimately provides a measure of variable importance. In addition, the ensemble decision process from uncorrelated (random) trees leads to great performance when the training data set is reduced, noisy or both [96].

ANN is formed by successive layers of connected neurons each characterized by a weight and an activation. The weight describes the strength of the connection of the neuron to neurons in the next layer. The activation results from the combination, through an activation function, of the inputs that a neuron receives from the previous layer. The first and last layers of such networks correspond to input and output while middle layers are termed hidden. A network with a large number of hidden layers is called a deep ANN with rapidly emerging applications in environ-

mental sciences [23, 25, 26].

Prior to statistical learning, a set of transformations was applied to predictors (Table IV). Such preprocessing includes estimating missing values with k-nearest neighbors imputation, removing predictors with near-zero variance and applying centering, scaling and Box-Cox transformations to collapse each predictors distribution to a normal distribution – an assumption behind numerous models [e.g., 97]. In addition, we also tested the influence of removing correlated predictors, applying principal component analysis, and applying independent component analysis.

To test the influence of predictors, oversampling, and resampling, seven different runs were performed (Table V). First, statistical learning was attempted both with and without contextual predictors. Second, the classes in this study suffer from the common ML challenge of unequal representation (Table I). In a base run, this data set imbalance was left untouched. The Synthetic Minority Oversampling Technique (SMOTE, 98) assigns predictors along the edges connecting the five-nearest neighbors from randomly selected observations and was used in SMOTE runs to address the imbalance of the training set. Third, because of the limited number of observations, resampling was performed by 20 repeats of 10-fold cross-validation, allowing all data to be used in both training and testing [99]. The data are randomly separated into 10 parts or folds and one fold is successively held out to assess model performance while the other nine folds are used for training. Repeated cross-validation addresses the potential bias introduced by the initial random selection of the folds. Spatial cross-validation ensures that resampling folds are spatially disjointed [100] and addresses the important issue of spatially correlated training data.

For each model (Table IV) and for each run (Table V), the set of best hyperparameters was tuned by a grid search across 20 different values per parameter resulting in the training of $\sim$ 20,000 models over 200 folds corresponding to four million realizations of the learning process. As the hyperparameter search space is larger for ANN, 20-hr discrete random search was performed before passing the resulting hyperparameters to the resampling scheme to improve computational efficiency. For all models, the one-standard-error rule was applied, selecting the simplest hyperparameters within one-standard-deviation of the most accurate hyperparameters [e.g., 95].

2. Model Performance Assessment

Model performance in statistical training is often reported as median cross-validation accuracy over resampling realizations. Accuracy corresponds to the ratio of accurately predicted classes to the total number of classes. The models exhibiting the highest median cross-validation accuracy were selected for calibration. A paired t test between the distribution of cross-validation accuracies from the best model and other models was performed to assess their similarity. These additional models are also reported.

ML models output posterior probabilities that often require calibration. Such calibration corrects the potential distortion of the posterior probabilities when compared to empirical probabilities and improves model performance [101–103]. Given the sigmoid shape of most distortions, Platt [104] proposed a sigmoid calibration to address this effect. Other useful approaches include Bayesian calibration and isotonic scaling [105] which both require binarizing the problem [106–112]. In this study, posterior calibration was performed using a multinomial regression, a straightforward extension of the binomial case corresponding to the logistic Platt's scaling [104]. The R package `glmnet` was used to fit a generalized linear model with an elastic net penalty and with a 10-fold cross-validation.

C. Assess Performance in Predictive Modeling

In probabilistic predictive modeling such as weather forecasting, prediction skill measures the difference between a predicted value and a reference forecast [113]. If a reference forecast is unavailable, as in the current study, one strategy is to estimate the entropy H associated with the predicted posterior probabilities p_i:

$$H = -\sum_i p_i \log p_i \tag{1}$$

Increasing entropy corresponds to increasing unpredictability [114] and prognosticates the prediction skill of a model [115, 116]. In addition, the entropy rate $H(\chi)$ defines (to the limit) how entropy varies in a Markovian sequence of prediction and how predictable the sequence is. Comparing the entropy associated with the outputs from different models is equivalent to comparing the log-likelihood of these models [117] and indicates which model provides the best information [40].

In this study, the predictive performance of the selected most accurate ML models was quantitatively assessed using entropy rate and stream interval entropy. First, at the network scale, the stability of the predictions was derived by computing the entropy rate from the transition probabilities between channel types. Interestingly, such an estimate of model performance uses only the prediction of the most probable channel type at each stream interval and addresses the issue that ML models may exhibit different shapes for their posterior probability distributions even after posterior calibration. While such a model-agnostic approach is relevant because of the hierarchical structure inherent to morphological units (e.g., pool, riffle and cascade) and channel types [e.g., 118, 119], similar reasoning can be applied for point pattern using spatial statistics (A).

Second, stream interval entropy was derived from the posterior probabilities of the model with the highest median-cross validation accuracy and the lowest entropy rate. This entropy measure represents the stability of the predictions at the stream interval scale and may be model-dependent. It is maximized when

Category	Acronym	Name	Information
Dimensionality	T2	Average number of points per dimension	Sparsity of the dataset
	T3	Average number of points per PCA dimension	Sparsity of the dataset after PCA
	T4	Ratio of the PCA dimension to the original dimension	Proportion of relevant predictors
Linearity	L1	Sum of the error distance to the SVM hyperplane	Linear separability
	L2	Error rate of the SVM classifier	Linear separability
	L3	Non-linearity of SVM classifier	Linearity of boundaries
Neighborhood	LSC	Local set average cardinality	Width of boundaries
	N1	Fraction of borderline points	Complexity of boundaries
	N2	Ratio of inter-/extra-class NN distance	Distance-based separability
	N3	Error rate of the NN classifier	Distance-based complexity
Network	ClsCoef	Clustering coefficient	Grouping tendency of observations
	Density	Density	Connectedness of observations
Overlapping	F1	Fisher's discriminant ratio	Predictor discriminative power
	F2	Volume of overlapping region	Predictor discriminative power
	F3	Maximal individual predictor efficiency	Predictor discriminative power
	F4	Collective predictor efficiency	Predictor discriminative power

PCA: principal component analysis

SVM: support vector machine

NN: nearest-neighbor

Table III: DCMs computed to characterize the problem complexity and to filter predictor noise. We follow definitions from Lorena *et al.* [76].

Model name	Centering, scaling Box-Cox transformations	k-NN imputation	PCA	ICA	Removing highly correlated predictors
PLS Corr	✓	✓	✗	✗	✓
PLS PCA Corr	✓	✓	✓	✗	✓
SVM Linear	✓	✓	✗	✗	✗
SVM Linear Corr	✓	✓	✗	✗	✓
SVM Linear PCA Corr	✓	✓	✓	✗	✓
SVM Radial	✓	✓	✗	✗	✗
SVM Radial Corr	✓	✓	✗	✗	✓
SVM Radial PCA Corr	✓	✓	✓	✗	✓
MARS	✓	✓	✗	✗	✗
FDA	✓	✓	✗	✗	✗
k-NN	✓	✓	✗	✗	✗
k-NN Corr	✓	✓	✗	✗	✓
k-NN PCA Corr	✓	✓	✓	✗	✓
CART	✓	✓	✗	✗	✗
CART Corr	✓	✓	✗	✗	✓
CART PCA Corr	✓	✓	✓	✗	✓
BaT	✓	✓	✗	✗	✗
BaT Corr	✓	✓	✗	✗	✓
RF	✓	✓	✗	✗	✗
LDA	✓	✓	✗	✗	✗
RLDA	✓	✓	✗	✗	✗
NB	✓	✓	✗	✗	✗
NB ICA	✓	✓	✗	✓	✗
ANN	✓	✓	✗	✗	✗

PCA: principal component analysis; ICA: independent component analysis;

PLS: partial least squares; SVM: support vector machine;

MARS: multivariate adaptive regression splines; FDA: flexible discriminant analysis;

k-NN: k-nearest neighbors; CART: classification and regression tree;

BaT: bagged-trees; RF: random forest; LDA: linear discriminant analysis;

RLDA: regularized linear discriminant analysis; NB: naive bayes; ANN: artificial neural network

Table IV: Types of Preprocessing.

Run	Predictors	Oversampling	Resampling
base-a	All	no	20 x 10-fold cross-validation
base-b	No contextual predictors	no	20 x 10-fold cross-validation
SpCv-b	No contextual predictors	no	20 x 10-fold Spatial Cross-Validation
SMOTE-a	all	SMOTE	20 x 10-fold cross-validation
SMOTE-b	No contextual predictors	SMOTE	20 x 10-fold cross-validation
SpCV-SMOTE-a	All	SMOTE	20 x 10-fold Spatial Cross-Validation
SpCV-SMOTE-b	No contextual predictors	SMOTE	20 x 10-fold Spatial Cross-Validation

Table V: Type of Runs Performed to Evaluate the Influence of Predictors, Oversampling and Resampling.

all channel types are equally probable for a given stream interval. Stream-interval entropy was used to assess the spatial structure of model uncertainty and its correlation with predictors.

The map of predicted channel types was investigated using expert-knowledge, with a focus on the general spatial organization of channel types across the study area as well as their geomorphic relevance. Aerial imagery was used to qualitatively confirm predictions at selected sample locations. Finally, the variable importance of predictors was investigated to support geomorphic interpretability of model predictions.

IV. RESULTS

Key results from the three-tier ML framework are presented below. In particular: (i) removing coarse-scale contextual predictors leads to a simpler classification problem, (ii) RF outperforms other models, (iii) prediction skill varies across the study area, and (iv) RF predictions capture the large-scale organization of the landscape.

A. Removing Coarse-scale Contextual Predictors Leads to a Simpler Classification Problem

DCMs characterize the classification problem (Figure 2a) and underline that its complexity decreases when removing contextual predictors and increases when removing topology or fractal dimension predictors (Figure 2b). In the following, the results from dimensionality, neighborhood, network, discriminative power and linearity DCMs are detailed.

Dimensionality DCMs underline a sparse predictor space (Figure 2a). As a consequence, removing predictors mechanically makes the problem simpler by increasing the number of observations per original (T2) or PCA dimensions (T3). Nonetheless, removing fractal dimension predictors increases the complexity of the relationship between predictors (T4, Figure 2b).

High neighborhood and network DCMs indicate a complex distribution of the classes observations in the predictor space (Figure 2a). The class boundaries appear narrow (LSC), complex (N1), with possibly overlapping classes (N1, N2) and highly disconnected observations with little clustering (Density, ClsCoef). Such a complex predictor space leads to a high error rate for the distance-based nearest-neighbor model (N3).

Despite the overall complexity of the predictor space, the discriminative power of some predictors makes the problem workable (Figure 2a). Similar to neighborhood and network DCMs, the mean discriminative power of predictors (F1) indicates a complex problem. However, the volume of overlapping region (F2) and the maximum individual predictor efficiency (F3) suggests that complete class separability is achievable. Importantly, the iterative use of high efficiency predictors to separate classes measures the

collective efficiency of predictors (F4) and leads to a tractable problem. This result contrasts with neighborhood DCMs which take into account the relationship of observations in the entire predictor space and implies predictor noise in the four predictor spaces considered. Furthermore, removing topology predictors leads to the most complex problem underlining that at least one of these predictors is prominent at separating classes (F3, F4, Figure 2b). The maximum individual predictor efficiency (F3) was computed for each pair of classes in the study area (Figure 3). Such an account of class separability underlines that the unconfined channel types 1, 9, and 10 are the most easily discriminated from other channel types (Table I). In contrast, the confined or partly-confined channel types 2–8 appear more complex to separate.

Similar to overlapping DCMs, linearity DCMs indicate a tractable problem. The error distance of observations to the SVM hyperplane (L1), the error rate of the SVM (L2) and its nonlinearity (L3) are all negligible. While these measures show that it is possible to find a linear hyperplane to separate all classes. Nonetheless, the underlying calculation of the SVM does not assess overfitting by cross-validation and is specific to the internal calculation of the DCMs. Importantly, this SVM model is distinct from SVM models included in the ML models benchmark during which ability to generalize is properly assessed by resampling.

B. Random Forest Outperforms Other Models

Seven different model runs were performed to assess the influence of predictors, oversampling, and resampling (Table VI). The removal of the contextual predictors leaves prediction accuracy unchanged. In contrast, balancing the channel types with SMOTE significantly improves prediction accuracy and using spatial cross-validation leads to decreased accuracy for all runs indicating some degree of spatial correlation in the data. In a multiclass problem with an expected significant spatial variability, ensuring that each fold contains examples of all channel types is complex. For 15 of the 20 spatial cross-validation repeats, 10 out of 10 folds contain examples of all channel types; one repeat has two folds with missing observations for one channel type, three repeats have four problematic folds and one repeat has six. Because addressing spatial correlation is more conservative than a minor loss in fold stratification, for the remainder of the paper, we only consider the run using spatial cross-validation, SMOTE, and without coarse contextual predictors (SpCV-SMOTE-b, Table VI).

From the statistical learning step of our three-tier framework, three models emerge as the most accurate for the Sacramento basin: L-SVM, R-SVM and RF (Figure 4). The 61% median cross-validation of these models represents a sixfold increase over the 10% accuracy of the baseline random model. L-SVM had the lowest computational cost (<1 minute), while R-SVM and RF were the most costly (>2 hours). Good performance of these models were expected from the
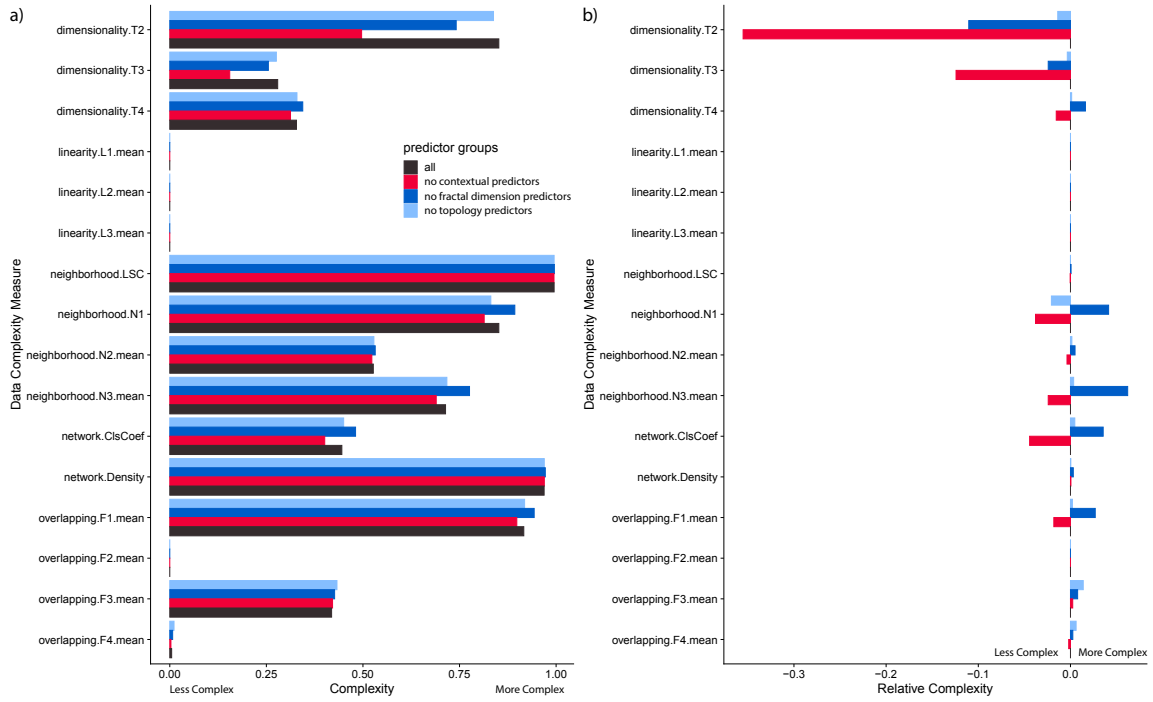
Figure 2: Summmary of data complexity measures computed following [76] for four sets of predictors with the initial set of predictors in black: (a) DCMs values; (b) relative complexity to the initial predictor sets. Increasing values indicate increasing complexity. Except for F1, mean values represent the average over one-versus-one binarization of the problem ($n = 45$ here). For F1, mean values represent the average over the number predictors.
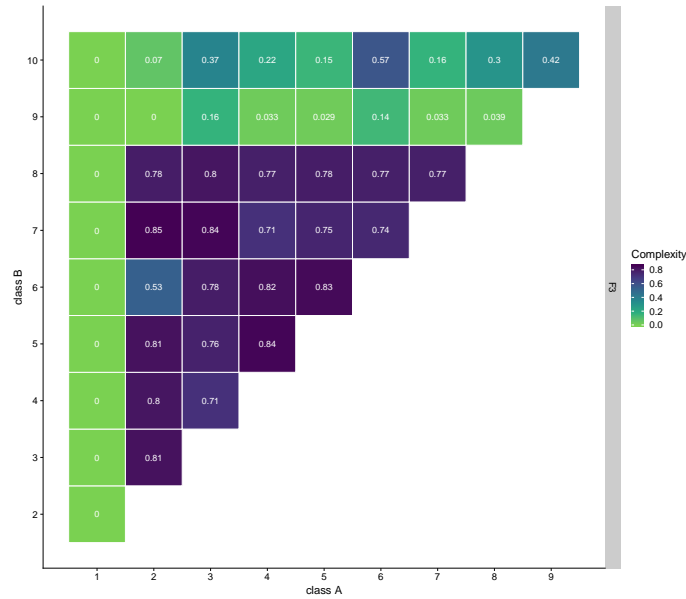

Figure 3: Maximum individual predictor efficiency (F3) for each pair of classes. For each pair of classes, values are between 0 and 1, with 0 indicating a simple problem and 1 a complex one.

DCMs analysis as they likely benefit from the collective efficiency of predictors and from the linearity of the problem (F4, L1, Figure 2a). Such evidenced linearity leads us to limit further analysis to the simpler L-SVM model.

Comparing the entropy rate associated with the predictions from RF and L-SVM identifies RF as the most accurate and stable model. The entropy rate is derived from the transition probabilities of the predic-

tions and is lower for RF ($H(\chi) = 1.35$) than L-SVM ($H(\chi) = 1.83$). While the transition probabilities are later described in detail to assess model predictive performance, their more organized representation provides compelling visual evidence of the higher stability of RF predictions (Figure 5). In addition, the best number of predictors tried at each split (RF hyperparameter) is 17; compared to the number of predictors, this value is low enough so that the trees of the forest

Run	Max median x-val accuracy	Best model(s)
base-a	0.39	MARS, FDA
base-b	0.38	MARS, FDA, RF
SpCv-b	0.34	MARS, FDA, CART, (RF[a])
SMOTE-a	0.70	R-SVM
SMOTE-b	0.70	R-SVM
SpCV-SMOTE-a	0.60	L-SVM, R-SVM
SpCV-SMOTE-b[b]	0.61	L-SVM, RF, R-SVM

[a]: RF achieved a high accuracy but was not considered statistically similar to MARS
[b]: Run selected

Table VI: Results From the Different Runs Performed (Table V). The maximum value for the median cross-validation is reported for each run as well as the best classifier(s). If more than one classifier is reported, the distributions of the cross-validation of these classifiers are indistinguishable from a statistical point of view.

are not significantly correlated, making the decision process of the ensemble model more stable [120]. In contrast, the best value of the cost parameter of the LSVM is $C = 0.25$, meaning that the accuracy of the L-SVM comes at the cost of some underfitting which translates into poorer predictions for data unseen in training as the class boundaries may not be well constrained (Figure 2).

C. Prediction Skill Differs Within Study Area

Stream-interval posterior entropy provides a quantitative assessment of uncertainty in model predictions (Figure 7). In the Sacramento basin, the predictions appear highly stable in the Central Valley, with more uncertainties (represented by higher entropy) in the mountainous area and a high level of instability in the Modoc Plateau. Such spatial differences in prediction skills are related to the varying complexity in separating channel types with unconfined channel types more easily separated from others (Figure 3). Nonetheless, the correlation of entropy with predictor variables was investigated to reveal the underlying structure of the uncertainty associated with the predictions (Figure 7). Only three predictor values were significantly correlated based Spearman's rank correlation ρ, elevation, slope and confinement distance with values of 0.30, 0.25 and -0.37, respectively.

Transition probabilities were computed for the Sacramento basin predictions (Figure 5). A chord diagram displays the probability of transitioning from one channel type to another by links between each sectors. These probabilities are undirected and are thus better understood as probabilities of being adjacent to another interval of a given channel type. Black arrows highlight the highest probability link. Self-transition, the transition from an interval of a given channel type to an interval of the same channel type, is represented by the bar in each sectors. In the Sacramento basin, the channel type with the lowest self-transition probability is the high gradient step-pool/cascade. This channel type transitions mainly to boulder-bedrock uniform streams and to the other confined channel types. In turn, these have links to partially confined channel types. The unconfined channel types transition mostly to themselves and to other unconfined channel types.

D. Random Forest Predictions Capture the Large-scale Organization of the Landscape

The geomorphic relevance of the RF model predictions was investigated in the study area. Overall, RF predictions of channel type exhibit the expected large-scale organization of the landscape (Figure 6): unconfined meandering streams occur in the Central Valley, boulder dominated step-pools occur in the mountainous areas. Additional qualitative investigations of the predictions were performed in combination with aerial imagery and indicated general good agreement with expectations. In the Sacramento basin, noisy predictions could be expected both from the pairwise DCMs analysis (Figures 2-3) and from the median cross-validation accuracy (61%).

The variable importance of the RF model identifies driving predictors and overcomes the "black box" nature of some ML approaches (Figure 8). Three predictor variables appear significantly more important than the others: valley confinement, drainage area and stream order. Apart from these three variables, channel slope and local drainage density, the most important predictors are fractal dimension predictors (Hurst coefficients), underlining their significance in the model performance. The importance of the fractal dimension and topology predictors is also suggested by the DCM analysis (Figures 2-3) and is supported by similar variable importance assessment for the ANN model.

V. DISCUSSION

A. Predicting Reach-scale Labels with Coarse-scale Predictors

Channel types are determined from relatively coarse predictors with reasonable performance in statistical learning (Figure 4), uncertainty (Figure 5 and 7), and geomorphic relevance (Figure 6). Five points appear critical in assessing this performance and will be useful to expand these methods to other areas of study and other bottom-up spatial predictions problems related to hydrologic regimes [20], surface–groundwater interactions [121], landsliding risk [30], or erosion processes [32]: (i) using DCM to understand the characteristics of the problem and to reduce predictor noise significantly improves statistical learning accuracy (Fig-
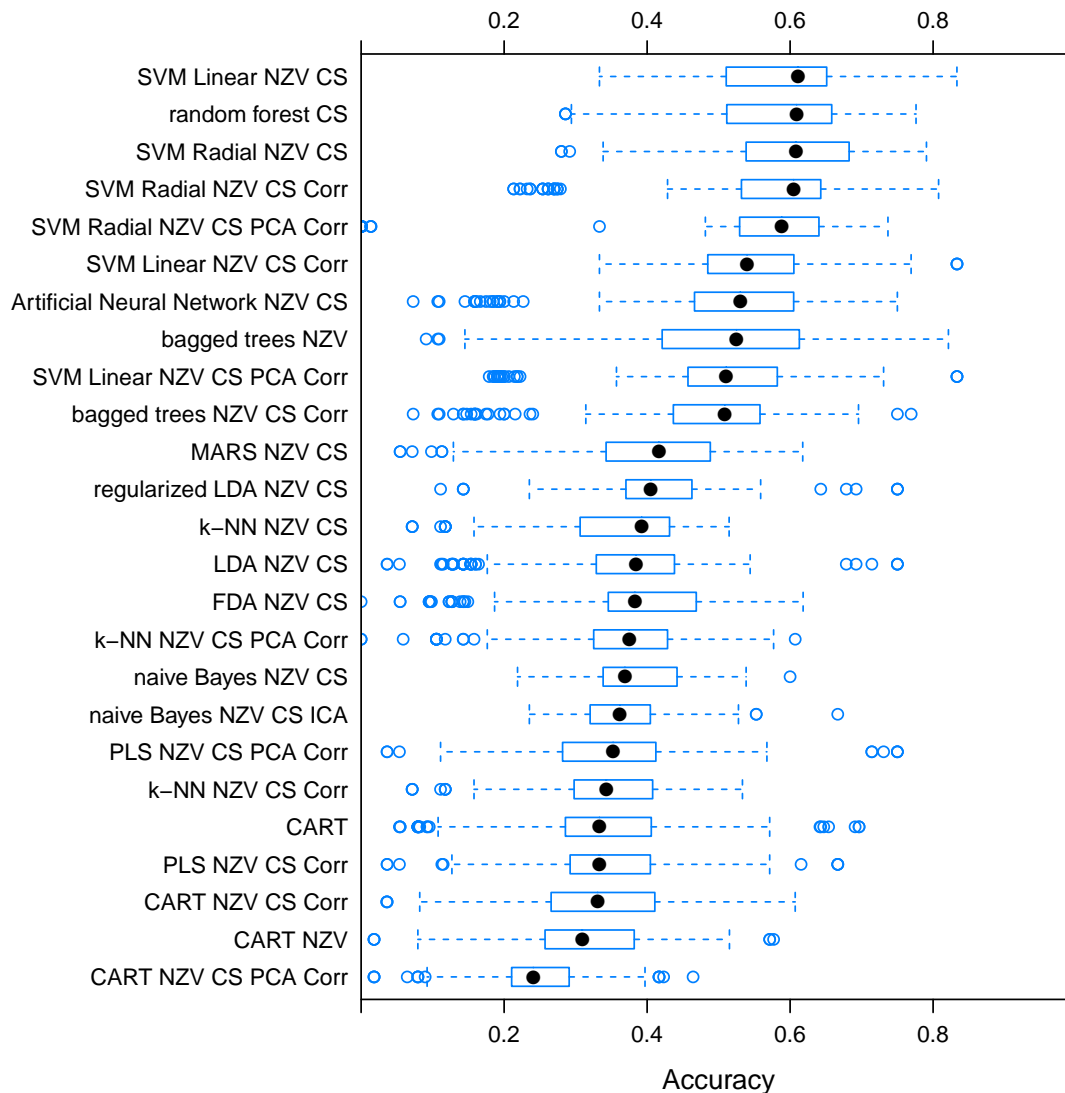
Figure 4: Distribution of cross-validation accuracies for all classifiers (Table IV) for the SpCV-SMOTE-b run for the Sacramento basin (Table VI). CS indicates that Box-Cox transformations were applied when required; NZV means that the predictors with near-zero-variance were filtered; Corr indoicates that highly correlated predictors were filtered; ICA and PCA mean that independant component analysis and principal component analysis were performed on the predictors, respectively.

ures 2 and 3 and Table VI); (ii) while RF was selected as the best algorithm in this study, training and validating multiple algorithms select the model with the best performance in a more reliable way (Figure 4); (iii) when spatial correlation is expected, spatial cross-validation should be applied to limit label leaking from the training set to the testing set; (iv) investigating the spatial structure of the uncertainty associated with the predictions informs on the prediction skill and potential data gaps (Figure 7); and (v) when possible, assessing predictor importance provides insight in the physical validity of the model (Figure 8). In this study, the variable importance for the RF model shows that valley confinement, drainage area, and stream order are the most useful predictors of channel type (Figure 8). The presence of local drainage density confirms its relevance for describing geomorphic processes [51].

Fractal dimension – the statistical roughness of topography – exhibits more control on model predictions than traditional Terrain Analysis Metrics (e.g., elevation and curvature) based on 10-m topographic data. Interestingly, fractal dimension predictors are not significantly correlated to other predictors (T4, Figure 2) suggesting that they capture unique characteristics of the landscape at different spatial scales. The integrative nature of fractal dimension in terms of land sculpting processes appears here to provide a holistic landscape description that is more useful for predicting channel form than traditional near-channel terrain predictors.

The spatial resolution of the predictors may significantly impact their discriminative power and affect model performance. A major challenge of this study is the mismatch between the scales of the predictors inputed to the ML models and the scale at which labels
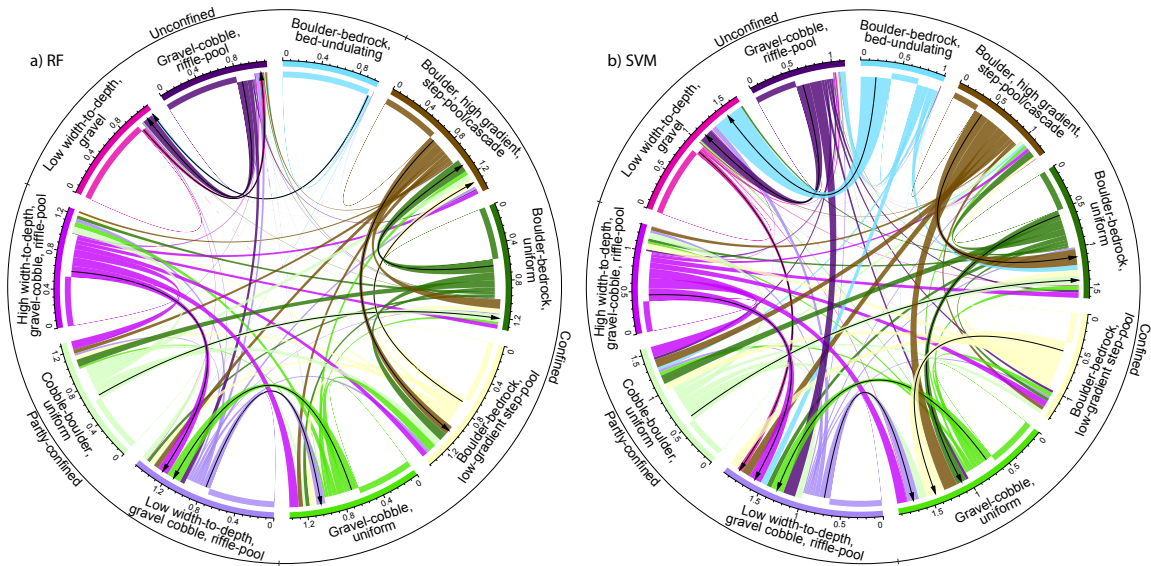
Figure 5: Chord diagram of the transition probabilities from one channel type to another in the Sacramento basin for (a) RF and (b) SVM models.
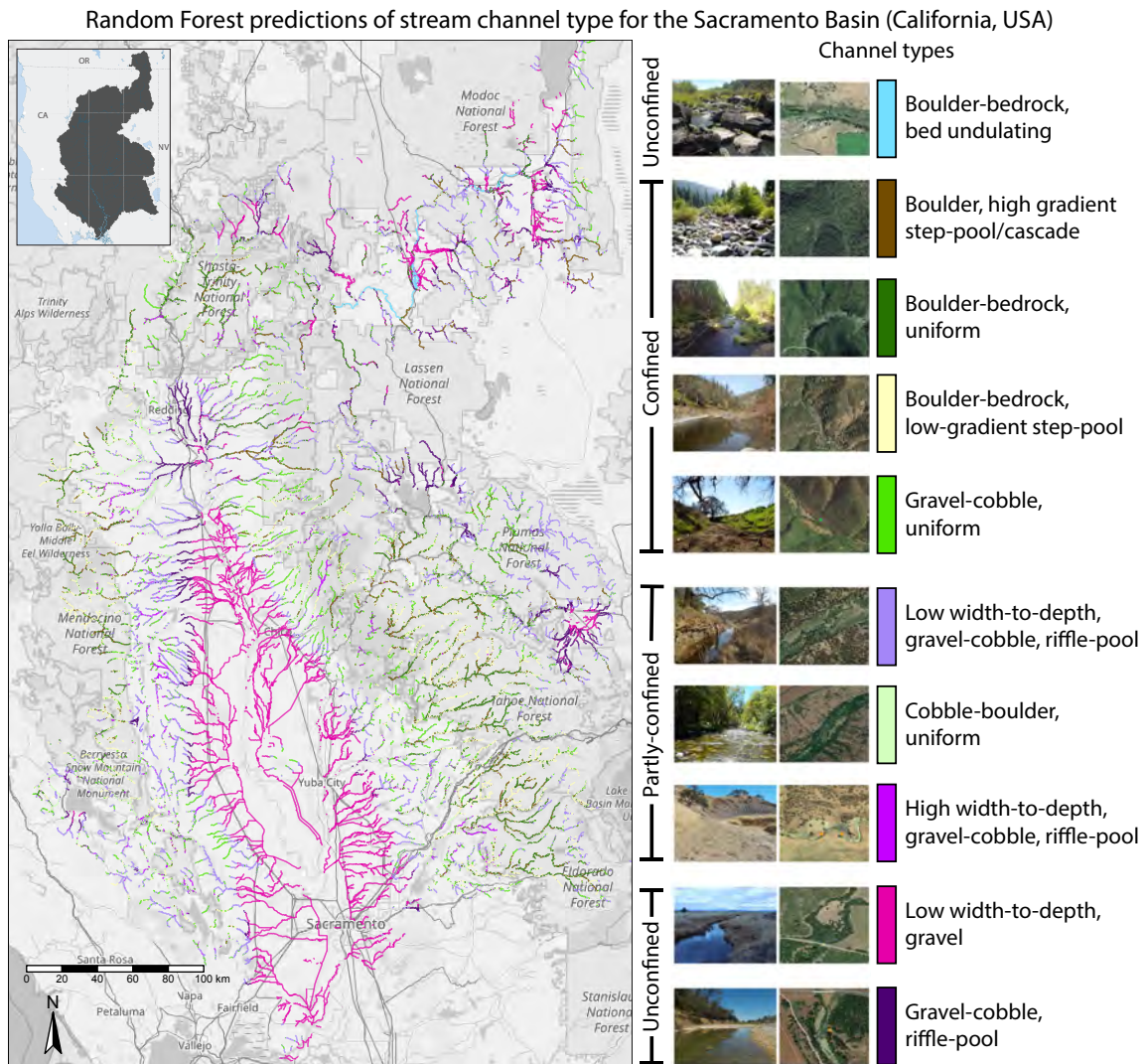

Figure 6: Map of the Sacramento Basin with the spatial predictions of the 10 types of channel. Inset shows location of the study area in California and in the general setting of the Pacific Coast of the United States.
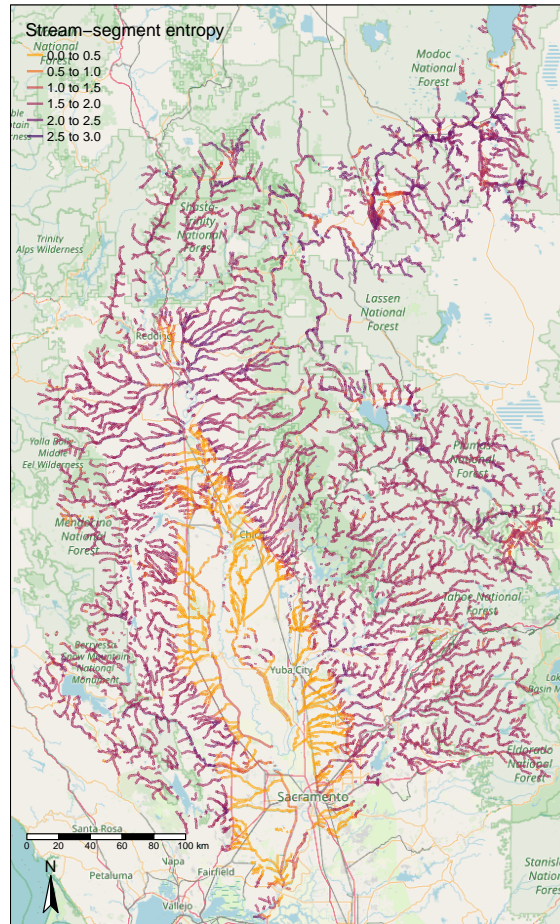
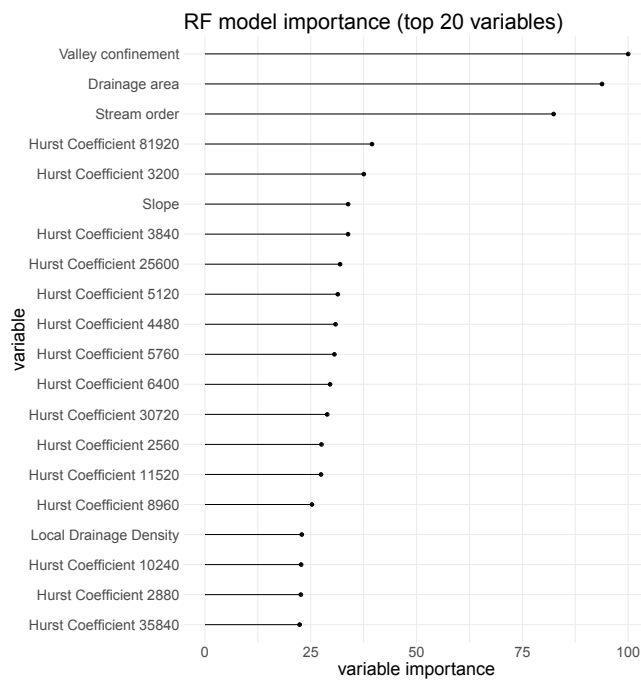Figure 7: Stream-interval posterior entropy in the Sacramento basin.

Figure 8: Variable importance of the RF model for the Sacramento basin.

have been defined. The labels correspond to channel types defined from field survey data at a 10^0 to 10^2-m scale [45]. In contrast, the spatial resolution of the predictors lies typically between 10^2 and 10^5 meters (Table II). The DCM analysis removes the coarsest scale (usually $> 10^3$ m) contextual predictors such as lithology, land cover and climate (Figure 2). However, it is unclear if these predictors are simply not strongly correlated to channel type or if they are just not available at sufficiently fine scales to distinguish important relationships.

The 10-m DEM used in this study may be less appropriate to decipher channel types than higher-resolution topography (e.g., 3 m) that better captures features linked with Earth surface processes [e.g., 122]. However, even with high-resolution topographic data, identifying the appropriate scale to describe local morphology remains challenging [123] and coarser resolution (5–10 m) may be best suited in some cases. For example, Lisenby and Fryirs [124] found that sediment connectivity was best represented using a 25-m rather than a 5-m or a 1-m DEM. Calculation of channel slope using the 10-m DEM may lead to substantial error, especially for short stream intervals [21]. Such an error-prone estimate of channel slope might explain why such a dominant control on geomorphic processes [e.g., 118] is not a more prominent predictor in the RF model (Figure 8). Further investigations are ongoing with topographic lidar data, but much of the world is still only covered with 10-m topographic data making this study a useful and extensible exercise.

The spatial scale mismatch between labels and predictors partly explains why RF outperforms deep learning algorithms (Figure 4). RF may benefit from its characteristic ensemble decision process [125] and its robustness to predictor noise [e.g., 96], but deep learning approaches have been very successful at predicting complex patterns [e.g., 126]. In particular, a finite deep learning network can approximate any function between input and output [127, 128] while avoiding local optimization minima [129] and keeping a relatively limited number of parameters [49]. As much as such a remarkable ability is rooted in mathematics and physics, complex structure often results from a sequence of simpler steps in a hierarchical generative process. Such a hierarchical process is efficiently reversed by stacked deep learning architectures when they can approximate near-perfect information distillation from one step of the generative process to the next [49]. This suggests that the gap between the scale of input data and the scale of output labels may be too wide in the Sacramento River basin to be bridged by deep learning. In addition, the difference in performance between traditional ML models (e.g., RF) and deep learning models might diagnose scale mismatch and be used to compare hierarchical bottom-up classifications established in vastly different settings.

B.   Geomorphic Implications of Entropy and Fractal Dimension

While we used entropy to select models with the best information content, the concepts of entropy in information theory and in statistical mechanics are closely related [130, 131]. In information theory, entropy measures the uncertainty of the information transmitted by a noisy process [114]. In statistical mechanics, entropy quantifies the number of accessible microscopic configurations for particles in a gas [e.g., 132]. Importantly, both in information theory and in statistical mechanics, entropy measures the inverse of degree of correlation in a system. Taking statistical mechanics as example, at low temperature, mean energy is low, limiting the number of accessible microscopic configurations, which leads to significant correlation in the system and to a lower entropy. Conversely, at high temperature, mean energy is high, implying a higher number of accessible microscopic configurations, which leads to a lower degree of correlation in the system and to a higher entropy.

In the Sacramento area, channel type diversity, stream interval entropy, and frequency of channel type transitions are the highest in mountainous areas, indicating that there are many possible channel type configurations (Figures 5-7). As a consequence, stream interval entropy correlates albeit weakly with slope and elevation which are both associated with higher energy to do geomorphic work [e.g., 133]. Such a correlation harks back to early comparisons of entropy in geomorphic and thermodynamic system which suggested a complete analogy between elevation and temperature [134–136]. While our results are more nuanced than these theoretical studies, our findings are reminiscent of the Boltzmann distribution which indicates that microscopic configurations of higher energy are less probable than microscopic configurations of lower energy. For example, the steepest channel type, step-pool/cascade is the most unstable (Figure 5). In addition, as this channel type is not limited to first-order streams, transition probabilities likely identify these stream intervals as high-energy unstable forms associated with knick zones of limited spatial extension [e.g., 14]. Recently, entropy was used to characterize the diversity of physical typology of river networks and define areas with similar hydrogeomorphic characteristics [18]. In a similar fashion, stream interval entropy in this study highlights distinct areas in terms of available energy.

Similar to entropy, fractal dimension is a measure of the degree of correlation of the topography. Low fractal dimension indicates a positively correlated surface (i.e., smooth) while high fractal dimension indicates negative correlation (i.e., jagged). In the Sacramento basin, mountainous areas generally have high fractal dimension while the Central Valley is characterized by lower fractal dimension. This suggests that the correlation structure of the topography as described by fractal dimension is closely linked to the correlation of channel types as described by stream interval entropy. In other words, fractal dimension relates to the distribution of available energy which is linked to the

distribution of accessible channel types. Conforming to physical and geomorphic intuition, the number of channel types and the energy of the hydro-geomorphic processes they represent are both expected to be higher in areas with high fractal dimension. This relationship between fractal dimension, geomorphic work and channel types explains in part the importance of fractal dimension predictors in the RF model (Figure 8). Further investigations are ongoing to decipher the links between fractal dimension, entropy and channel types as well as its physical meaning, mainly understood through its small-scale and large-scale end-members describing erosion processes and tectonism, respectively [e.g., 56–58, 137–139].

VI.   CONCLUSION

While ML and deep learning are increasingly harnessed in the natural sciences, their application in hydrologic sciences has been slow. This is partially explained by a preference for physics-derived, process-based modeling. This study exemplifies how some of the black box aspects of ML can be clarified through a rigorous approach both in model design and in the evaluation of model performance in statistical learning and predictive modeling. Our three-tier framework is transferable to other bottom-up spatial prediction problems prevalent in hydrology and tackles the significant challenge of predicting reach-scale field-derived channel types using publicly available coarse-scale predictors. Random Forest predictions coherently capture the large-scale geomorphic organization of the landscape, and entropy derived from posterior probabilities maps the predictive performance of the model and underlines uncertain and stable areas. Channel types appear partially controlled by the statistical roughness of the topography, which relates to the distribution of energy available to generate distinct channel types. Avenues for future research include comparison of bottom-up hierarchical classifications across diverse study areas, interpreting the physical meaning of fractal dimension at different scales and incorporating aerial imagery as input data given the demonstrated image recognition performance of convolution neural networks.

Appendix A: Estimating Entropy for Point Patterns Predictions

In this study, entropy is used to select the model with the expected lowest uncertainty. While we use entropy rate to prognosticate such uncertainty, this constrains our method to predictive modeling where transition probabilities can be computed. However, a kernelized entropy can be derived in a more general case with an interpretation similar to entropy rate. In particular, this estimation of model performance is model-agnostic, addressing the issue that ML models may exhibit different shapes for their posterior probability distributions even after posterior calibration.

Such a kernelized entropy is calculated using spatial statistics derived from the most probable class – in our case, channel types. Kernel smoothing then maps the relative risk, that is the spatially varying estimates of the probability of occurrence of each channel type. The resulting spatial probabilities are then aggregated by calculating their entropy and the spatial distribution of entropy matches the network-scale evaluation of the instability from entropy rate. Then, similar to entropy rate, comparing the distribution of kernelized entropy across models selects the ML model with the best information content [40, 117].

One key point in this approach is the selection of the bandwidth which determines the amount of smoothing introduced by the Gaussian kernel: Large bandwidth values correspond to a high degree of smoothing and vice-versa. Kernel bandwidth is selected as the mean value between: (i) the median length of the stream-intervals as defined in the NHDPlusV2, (ii) the mean length of the stream-intervals as defined in the NHDPlusV2, (iii) the inflexion point of the Ripley's K and Besag's L^* function, (iv) the inflexion point of the pair correlation function $g(r)$. The Ripley's K, Besag's L^* and the pair correlation $g(r)$ functions characterize a point pattern [e.g., 140] and identify how clustering evolves with increasing scale. An inflection point marks the scale at which the clustering starts to slow down with increasing scale and was estimated using a segmented linear regression. Such a kernel bandwidth selection process is sound but introduces a parametric component.

[1] W. M. Davis, The Geographical Journal **14**, 481 (1899).

[2] D. L. Rosgen, CATENA **22**, 169 (1994).

[3] D. R. Montgomery and J. M. Buffington, Geological Society of America Bulletin **109**, 596 (1997).

[4] G. J. Brierley and K. A. Fryirs, *Geomorphology and river management: Applications of the river styles framework* (John Wiley & Sons, 2013).

[5] J. Alexander, R. Zelt, and N. Schaepe, *Geomorphic segmentation, hydraulic geometry, and hy-*

draulic microhabitats of the Niobrara River, Nebraska—Methods and initial results*, Tech. Rep. (U.S. Geological Survey Scientific Investigations Report, 2009).

[6] G. O'Brien and J. Wheaton, *River Styles Report for the Middle Fork John Day Watershed, Oregon*, Tech. Rep. (Ecogeomorphology and Topographic Analysis Lab, Utah State University, 2014).

[7] G. R. O'Brien, J. Wheaton, K. Fryirs, P. McHugh, N. Bouwes, G. Brierley, and C. Jordan, Journal of Maps **13**, 369 (2017).

[8] M. Rinaldi, B. Belletti, F. Comiti, L. Nardi, M. Bussettini, L. Mao, and A. Gurnell, "The Geomorphic Units survey and classification System (GUS), Deliverable 6.2, Part 4, of REFORM (REstoring rivers FOR effective catchment Management), a Collaborative project (large-scale integrating project) funded by the European Commission within the 7 th Framework Programme under Grant Agreement 282656," (2015).

[9] R. Jha and P. Diplas, Journal of Hydraulic Research **56**, 299 (2017).

[10] K. A. Fryirs, J. M. Wheaton, and G. J. Brierley, Earth Surface Processes and Landforms **41**, 701 (2016).

[11] A. Kasprak, N. Hough-Snee, T. Beechie, N. Bouwes, G. Brierley, R. Camp, K. Fryirs, H. Imaki, M. Jensen, G. O'Brien, *et al.*, PloS one **11**, e0150293 (2016).

[12] J. Buffington and D. Montgomery, in *Treatise on Geomorphology* (Elsevier, San Diego, CA (USA), 2013) pp. 730–767.

[13] G. M. Kondolf, H. Piégay, L. Schmitt, and D. R. Montgomery, "Geomorphic classification of rivers and streams," in *Tools in Fluvial Geomorphology* (John Wiley Sons, Ltd, 2016) Chap. 7, pp. 133–158.

[14] Y. S. Hayakawa and T. Oguchi, Geomorphology **78**, 90 (2006).

[15] A. Alber and H. Piégay, Geomorphology **125**, 343 (2011).

[16] C. Roux, A. Alber, M. Bertrand, L. Vaudor, and H. Piégay, Geomorphology **242**, 29 (2015).

[17] A. Golly and J. M. Turowski, Earth Surface Dynamics **5**, 557 (2017).

[18] M. Thoms, M. Scown, and J. Flotemersch, JAWRA Journal of the American Water Resources Association (2018), 10.1111/1752-1688.12649.

[19] R. A. Hill, M. H. Weber, S. G. Leibowitz, A. R. Olsen, and D. J. Thornbrugh, JAWRA Journal of the American Water Resources Association **52**, 120 (2015).

[20] B. A. Lane, H. E. Dahlke, G. B. Pasternack, and S. Sandoval-Solis, JAWRA Journal of the American Water Resources Association **53**, 411 (2017).

[21] T. M. Neeson, A. M. Gorman, P. J. Whiting, and J. F. Koonce, North American Journal of Fisheries Management **28**, 722 (2008).

[22] D. Michie, Nature **218**, 19 (1968).

[23] C. Shen, Water Resources Research (2018), 10.1029/2018wr022643.

[24] C. Shen, E. Laloy, A. Elshorbagy, A. Albert, J. Bales, F.-J. Chang, S. Ganguly, K.-L. Hsu, D. Kifer, Z. Fang, K. Fang, D. Li, X. Li, and W.-P. Tsai, Hydrology and Earth System Sciences **22**, 5639 (2018).

[25] K. J. Bergen, P. A. Johnson, V. Maarten, and G. C. Beroza, Science **363**, eaau0323 (2019).

[26] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, Nature **566**, 195 (2019).

[27] P. Bugnicourt, S. Guitet, V. F. Santos, L. Blanc, E. D. Sotta, N. Barbier, and P. Couteron, Geomorphology **317**, 23 (2018).

[28] G. L. W. Perry and M. E. Dickson, Journal of Geophysical Research: Earth Surface **123**, 2954 (2018).

[29] V. M. Santos, T. Wahl, J. W. Long, D. L. Passeri, and N. G. Plant, Journal of Geophysical Research: Earth Surface (2019), 10.1029/2019jf005016.

[30] J.-H. Lee, M. I. Sameen, B. Pradhan, and H.-J. Park, Geomorphology **303**, 284 (2018).

[31] B. T. Pham, I. Prakash, and D. T. Bui, Geomorphology **303**, 256 (2018).

[32] O. Rahmati, N. Tahmasebipour, A. Haghizadeh, H. R. Pourghasemi, and B. Feizizadeh, Geomorphology **298**, 118 (2017).

[33] M. Bassiouni, C. W. Higgins, C. J. Still, and S. P. Good, Hydrology and Earth System Sciences **22**, 3229 (2018).

[34] M. A. Hassan, S. Bird, D. Reid, C. Ferrer-Boix, D. Hogan, F. Brardinoni, and S. Chartrand, Earth Surface Processes and Landforms **44**, 736 (2018).

[35] F. J. Clubb, B. Bookhagen, and A. Rheinwalt, Journal of Geophysical Research: Earth Surface (2019), 10.1029/2019jf005025.

[36] A. A. Vaughan, P. Belmont, C. P. Hawkins, and P. Wilcock, Journal of Geophysical Research: Earth Surface **122**, 1901 (2017).

[37] S. D. Hamshaw, M. M. Dewoolkar, A. W. Schroth, B. C. Wemple, and D. M. Rizzo, Water Resources Research **54**, 4040 (2018).

[38] T. Beechie and H. Imaki, Water Resources Research **50**, 39 (2014).

[39] R. A. McManamay, M. J. Troia, C. R. DeRolph, A. O. Sheldon, A. R. Barnett, S.-C. Kao, and M. G. Anderson, PLOS ONE **13**, e0198439 (2018).

[40] G. S. Nearing and H. V. Gupta, Water Resources Research **51**, 524 (2015).

[41] G. S. Nearing, D. M. Mocko, C. D. Peters-Lidard, S. V. Kumar, and Y. Xia, Journal of hydrometeorology **17**, 745 (2016).

[42] B. L. Ruddell, D. Drewry, and G. S. Nearing, Water Resources Research (2019).

[43] F. Kratzert, M. Herrnegger, D. Klotz, S. Hochreiter, and G. Klambauer, arXiv preprint arXiv:1903.07903 (2019).

[44] J. F. Mount, *California rivers and streams: the conflict between fluvial process and land use* (Univ of California Press, 1995).

[45] C. Byrne, G. Pasternack, H. Guillon, B. Lane, and S. S. Solis, (2019), 10.1002/essoar.10501068.1.

[46] B. A. Lane, G. Pasternack, E. Dahlke, Helen, and S. Sandoval-Solis, Progress in Physical Geography (2017).

[47] B. A. Lane, G. B. Pasternack, and S. S. Solis, Ecohydrology **11**, e1969 (2018).

[48] B. A. Lane, S. Sandoval-Solis, E. D. Stein, S. M. Yarnell, G. B. Pasternack, and H. E. Dahlke, Environmental Management (2018), 10.1007/s00267-018-1077-7.

[49] H. W. Lin, M. Tegmark, and D. Rolnick, Journal of Statistical Physics **168**, 1223 (2017).

[50] A. N. Strahler, Transactions, American Geophysical Union **38**, 913 (1957).

[51] M. Danesh-Yazdi, A. Tejedor, and E. Foufoula-Georgiou, Geomorphology **295**, 16 (2017).

[52] R. Pastor-Satorras and D. H. Rothman, Journal of Statistical Physics **93**, 477–500 (1998).

[53] P. S. Dodds and D. H. Rothman, Annual Review of Earth and Planetary Sciences **28**, 571–610 (2000).

[54] T. H. Wilson and J. Dominic, Earth Surface Processes and Landforms **23**, 509 (1998).

[55] N. A. Lifton and C. G. Chase, Geomorphology **5**, 77 (1992), fractals in Geomorphology.

[56] Q.-C. Sung and Y.-C. Chen, Geomorphology **62**, 181 (2004).

[57] A. Faghih and A. Nourbakhsh, Journal of Mountain Science **12**, 177 (2015).

[58] L. Liucci and L. Melelli, Earth Surface Processes and Landforms (2017), 10.1002/esp.4206.

[59] C. Duclut and B. Delamotte, Physical Review E: Statistical, Nonlinear, and Soft Matter Physics **96** (2017).

[60] C. Teutschbein, T. Grabs, H. Laudon, R. H. Karlsen, and K. Bishop, Journal of Hydrology **561**, 160 (2018).

[61] D. Mercado-Bettín, J. F. Salazar, and J. C. Villegas, Ecohydrology , e2072 (2019).

[62] M. D. Hurst, S. M. Mudd, R. Walcott, M. Attal, and K. Yoo, Journal of Geophysical Research: Earth Surface **117**, n/a (2012).

[63] J. P. Prancevic and J. W. Kirchner, Geophysical Research Letters **46**, 2084 (2019).

[64] S. Bonetti and A. Porporato, Geophysical Research Letters **44**, 5531 (2017).

[65] D. Gesch, M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler, Photogrammetric engineering and remote sensing **68**, 5 (2002).

[66] L. McKay, T. Bondelid, T. Dewald, J. Johnston, R. Moore, , and A. Rea, *NHDPlus Version 2: User Guide*, United States Environmental Protection Agency (EPA) (2012).

[67] S. A. Archfield, M. Clark, B. Arheimer, L. E. Hay, H. McMillan, J. E. Kiang, J. Seibert, K. Hakala, A. Bock, T. Wagener, W. H. Farmer, V. Andréassian, S. Attinger, A. Viglione, R. Knight, S. Markstrom, and T. Over, Water Resources Research **51**, 10078 (2015).

[68] P. W. Cazenave, J. K. Dix, D. O. Lambkin, and L. C. McNeill, Earth Surface Processes and Landforms **38**, 221 (2012).

[69] D. Buscombe, Computers & Geosciences **86**, 92 (2016).

[70] A. Agarwal, N. Marwan, M. Rathinasamy, B. Merz, and J. Kurths, Nonlinear Processes in Geophysics **24**, 599 (2017).

[71] D. Newman, J. Lindsay, and J. Cockburn, Geomorphology **312**, 40 (2018).

[72] S. Borut, G. Dragan, and L. Nada, Frontiers in Artificial Intelligence and Applications **215**, 1105–1106 (2010).

[73] L. P. F. Garcia, A. C. Lorena, and A. C. Carvalho, in *2012 Brazilian Symposium on Neural Networks* (IEEE, 2012).

[74] B. Sluban, D. Gamberger, and N. Lavrač, Data Mining and Knowledge Discovery **28**, 265 (2013).

[75] L. P. Garcia, A. C. de Carvalho, and A. C. Lorena, Neurocomputing **160**, 108 (2015).

[76] A. C. Lorena, L. P. F. Garcia, J. Lehmann, M. C. P. Souto, and T. K. Ho, "How complex is your classification problem? a survey on measuring classification complexity," (2018), arXiv:1808.03591.

[77] T. K. Ho and M. Basu, IEEE Transactions on Pattern Analysis and Machine Intelligence **24**, 289 (2002).

[78] A. C. Lorena, A. I. Maciel, P. B. C. de Miranda, I. G. Costa, and R. B. C. Prudêncio, Machine Learning **107**, 209 (2017).

[79] R. J. Hijmans, J. van Etten, J. Cheng, J. A. Greenberg, O. P. Lamigueiro, A. Bevan, *et al.*, *Package 'raster'* (2018), version 2.6-7.

[80] I. V. Florinsky, International Journal of Geographical Information Science **12**, 47 (1998).

[81] ESRI, *ArcGIS Desktop*, Environmental Systems Research Institute, Redlands, CA (2016).

[82] C. T. Haan, B. J. Barfield, and J. C. Hayes, *Design hydrology and sedimentology for small catchments* (Elsevier, 1994).

[83] K. G. Renard, G. R. Foster, G. Weesies, D. McCool, and D. Yoder, *Predicting soil erosion by water: a guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE)*, Vol. 703 (United States Department of Agriculture Washington, DC, 1997).

[84] J. Cress, D. Soller, R. Sayre, P. Comer, and H. Warner, *Terrestrial ecosystems – Surficial lithology of the conterminous United States* (2010), u.S. Geological Survey Scientific Investigations Map 3126, scale 1:5,000,000, 1 sheet.

[85] G. E. Schwarz and R. Alexander, *State soil geographic (STATSGO) data base for the conterminous United States*, Tech. Rep. (U.S. Geological Survey, 1995).

[86] C. Homer, J. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. Herold, J. Wickham, and K. Megown, Photogrammetric Engineering & Remote Sensing **81**, 345 (2015).

[87] PRISM Climate Group, *PRISM Gridded Climate Data*, Oregon State University (2004).

[88] D. J. Thornbrugh, S. G. Leibowitz, R. A. Hill, M. H. Weber, Z. C. Johnson, A. R. Olsen, J. E. Flotemersch, J. L. Stoddard, and D. V. Peck, Ecological Indicators **85**, 1133 (2018).

[89] J. Luengo and F. Herrera, Knowledge and Information Systems **42**, 147 (2013).

[90] M. Kuhn, Journal of Statistical Software, Articles **28**, 1 (2008).

[91] M. Kuhn, *Package 'caret'* (2018), version 6.0-80.

[92] H2O.ai, *R interface for H2O* (2018), r package version 3.20.0.8.

[93] C. Cortes and V. Vapnik, Machine Learning **20**, 273 (1995).

[94] A. C. Lorena, A. C. P. L. F. de Carvalho, and J. M. P. Gama, Artificial Intelligence Review **30**, 19 (2008).

[95] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees* (CRC press, 1984).

[96] E. W. Fox, R. A. Hill, S. G. Leibowitz, A. R. Olsen, D. J. Thornbrugh, and M. H. Weber, Environmental Monitoring and Assessment **189** (2017), 10.1007/s10661-017-6025-0.

[97] O. Csillik, I. Evans, and L. Drăguţ, Geomorphology **232**, 65–77 (2015).

[98] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, Journal of artificial intelligence research **16**, 321 (2002).

[99] P. Burman, Biometrika **76**, 503 (1989).

[100] P. Schratz, J. Muenchow, J. Richter, and A. Brenning, arXiv preprint arXiv:1803.11266 (2018).

[101] M. H. DeGroot and S. E. Fienberg, Journal of the Royal Statistical Society: Series D (The Statistician) **32**, 12 (1983).

[102] B. Zadrozny, in *Advances in neural information processing systems* (2002) pp. 1041–1048.

[103] A. Niculescu-Mizil and R. Caruana, in *Proceedings of the 22nd international conference on Machine learning - ICML '05* (ACM Press, 2005).

[104] J. Platt, Advances in large margin classifiers **10**, 61 (1999).

[105] B. Zadrozny and C. Elkan, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2002) pp. 694–699.

[106] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, in *Advances in neural information processing systems* (2000) pp. 547–553.

[107] B. Kijsirikul and N. Ussivakul, in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)* (IEEE, 2002).

[108] L. Dong, E. Frank, and S. Kramer, in *European Conference on Principles of Data Mining and Knowledge Discovery* (Springer, Berlin, Heidelberg, 2005) pp. 84–95.

[109] A. C. Lorena and A. C. de Carvalho, Neurocomputing **73**, 2837 (2010).

[110] T. M. Quiterio and A. C. Lorena, in *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on* (IEEE, 2016) pp. 115–120.

[111] M. N. Adnan and M. Z. Islam, in *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (Bruges (Belgium)* (2015) pp. 385–390.

[112] V. Melnikov and E. Hüllermeier, Machine Learning (2018), 10.1007/s10994-018-5733-1.

[113] T. Gneiting and A. E. Raftery, Journal of the American Statistical Association **102**, 359 (2007).

[114] C. E. Shannon, Bell system technical journal **27**, 379 (1948).

[115] D. B. Stephenson and F. J. Dolas-Reyes, Tellus A: Dynamic Meteorology and Oceanography **52**, 300 (2000).

[116] M. S. Roulston and L. A. Smith, Monthly Weather Review **130**, 1653 (2002).

[117] D. J. Daley and D. Vere-Jones, Journal of Applied Probability **41**, 297 (2004).

[118] G. E. Grant, F. J. Swanson, and M. G. Wolman, Geological Society of America Bulletin **102**, 340 (1990).

[119] D. R. Montgomery and J. M. Buffington, "River Ecology and Management: Lessons from the Pacific Coastal Ecoregion," (New York, Springer-Verlag, 1998) Chap. Channel processes, classification, and response, pp. 13–42.

[120] P. Probst, M. Wright, and A.-L. Boulesteix, arXiv

[121] J. Yang, J. Griffiths, and C. Zammit, River Research and Applications **35**, 932 (2019).

[122] P. Passalacqua, P. Belmont, D. M. Staley, J. D. Simley, J. R. Arrowsmith, C. A. Bode, C. Crosby, S. B. DeLong, N. F. Glenn, S. A. Kelly, and et al., Earth-Science Reviews **148**, 174–193 (2015).

[123] P. Tarolli, Geomorphology **216**, 295 (2014).

[124] P. E. Lisenby and K. A. Fryirs, Earth Surface Processes and Landforms **42**, 1588 (2017).

[125] T. K. Ho, in *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1 (IEEE, 1995) pp. 278–282.

[126] Y. LeCun, Y. Bengio, and G. Hinton, nature **521**, 436 (2015).

[127] G. Cybenko, Mathematics of control, signals and systems **2**, 303 (1989).

[128] K. Hornik, M. Stinchcombe, and H. White, Neural networks **2**, 359 (1989).

[129] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Proceedings of the National Academy of Sciences **113**, E7655 (2016).

[130] E. T. Jaynes, Physical review **106**, 620 (1957).

[131] E. T. Jaynes, Physical review **108**, 171 (1957).

[132] J. W. Gibbs, *Elementary Principles in Statistical Mechanics* (Univ. Press, New Haven, Conn, Yale, 1902).

[133] R. A. Bagnold, USGS Professional Paper **422-I**, 37 (1966).

[134] L. B. Leopold and W. B. Langbein, *The concept of entropy in landscape evolution*, Vol. 500 (US Government Printing Office, 1962) pp. 1–19.

[135] A. Scheidegger, International Association of Scientific Hydrology. Bulletin **9**, 12 (1964).

[136] A. Scheidegger, International Association of Scientific Hydrology. Bulletin **12**, 57 (1967).

[137] T. Xu, I. D. Moore, and J. C. Gallant, Geomorphology **8**, 245 (1993).

[138] J. R. Carr, Engineering geology **48**, 269 (1997).

[139] Q.-C. Sung, Y.-C. Chen, and P. Chao, Terrestrial, Atmospheric and Oceanic sciences journal **9**, 655 (1998).

[140] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical analysis and modelling of spatial point patterns*, Vol. 70 (John Wiley & Sons, 2008).

preprint arXiv:1804.03515 (2018).