2-27-2020

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

Hervé Guillon
University of California, Davis

Colin F. Byrne
University of California, Davis

Belize A. Lane
Utah State University

Samuel Sandoval Solis
University of California, Davis

Gregory B. Pasternack
University of California, Davis

Recommended Citation

Guillon, H., Byrne, C. F., Lane, B. A., Sandoval Solis, S., & Pasternack, G. B. (2020). Machine learning predicts reach-scale channel types from coarse-scale geospatial data in a large river basin. Water Resources Research, 56, e2019WR026691. https://doi.org/10.1029/2019WR026691

Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR026691

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]Department of Land, Air and Water Resources, University of California, Davis, CA, USA, [2]Department of Civil and Environmental Engineering, Utah State University, Logan, UT, USA

Key Points:

- The developed framework selects a Random Forest model to predict 10 channel types determined from 290 field surveys over ~109,000 two hundred-meter reaches
- Imperfect training information and scale mismatch between labels and predictors explain changes in the entropy-based predictive performance
- The persistent roughness of the topography partially controls channel types

Correspondence to:
H. Guillon,
hguillon@ucdavis.edu

Citation:
Guillon, H., Byrne, C. F., Lane, B. A., Sandoval Solis, S., & Pasternack, G. B. (2020). Machine learning predicts reach-scale channel types from coarse-scale geospatial data in a large river basin. *Water Resources Research*, *56*, e2019WR026691. https://doi.org/10.1029/2019WR026691

Received 6 NOV 2019
Accepted 24 FEB 2020
Accepted article online 27 FEB 2020

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.

1. 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 (Davis, 1899) gave rise to a variety of reach-scale geomorphic classifications (e.g., Montgomery & Buffington, 1997; Rosgen, 1994). With increasing collaborations between scientists and stakeholders, basin-wide hierarchical classifications such as the River Styles framework have gained traction (e.g., Alexander et al., 2009; Brierley & Fryirs, 2013; Jha & Diplas, 2017; O'Brien & Wheaton, 2014; O'Brien et al., 2017; Rinaldi et al., 2015). Such classifications usually rely on field-based observations of valley confinement (Fryirs et al., 2016), channel geometry (e.g., width and sinuosity) and instream features (e.g., bar and pool). These mainly descriptive classifications have a fuzzy correspondence (Kasprak et al., 2016), suggesting that automation and standardization of reach-scale classification may not only be possible but beneficial (Buffington & Montgomery, 2013; Kondolf et al., 2016). To that end, reproducible, physically interpretable river characterization within a GIS framework represents an important step forward (e.g., Alber & Piégay, 2011; Golly & Turowski, 2017; Hayakawa & Oguchi, 2006; Roux et al., 2015; Thoms et al., 2018).

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., Hill et al., 2015) and streamflow response, extrapolating hydrologic classes to the stream network is well established (e.g.,

Lane et al., 2017). 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. (2017) 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., Neeson et al., 2008) 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., Michie, 1968). 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 with process-based deterministic approaches. For example, recent reviews underline a limited use of the latest developments of ML in hydrologic sciences (Shen, 2018; Shen et al., 2018), a slower adoption in geophysics (Bergen et al., 2019), and emerging applications in Earth System Science (Reichstein et al., 2019). 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 (Bugnicourt et al., 2018), predicting geomorphic disturbance (Perry & Dickson, 2018) or dune erosion (Santos et al., 2019), mapping susceptibility to landslide and gully erosion (Lee et al., 2018; Pham et al., 2018; Rahmati et al., 2017), inferring ecohydrological parameters (Bassiouni et al., 2018), analyzing model residuals (Hassan et al., 2018), clustering river profiles (Clubb et al., 2019), classifying and predicting sediment-discharge relationships (Hamshaw et al., 2018; Vaughan et al., 2017), and assessing stream diversity with large-scale top-down approaches (Beechie & Imaki, 2014; McManamay et al., 2018). In the face of such rising popularity, the interpretability and assessment of uncertainty in ML models remain key issues (e.g., Reichstein et al., 2019).

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 ~20,000 ML models in statistical learning, and (iii) assess the performance ~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., Kratzert et al., 2019; Nearing & Gupta, 2015; Nearing et al., 2016; Ruddell et al., 2019).

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.

2. Case Study

Spanning 70,130 km^2, the Sacramento River basin in California (USA) exhibits diverse physiography and hydrology (Mount, 1995). 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 (Lane et al., 2017).

Byrne et al. (2019) previously identified 10 dominant geomorphic channel types present in the Sacramento River basin from data acquired at 290 field sites (Table 1). 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

Table 1
Channel Types Identified in the Sacramento Basin by Byrne et al. (2019)

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

variability of rivers in the study basin (Byrne et al., 2019; Lane et al., 2017) makes it an ideal test bed for testing large-scale geomorphic classification needed for future integrative research (e.g., Lane, Pasternack & Solis 2018; Lane, Sandoval-Solis, et al., 2018).

3. 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., Lin et al., 2017). 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 (Byrne et al., 2019), 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 performance of models in predictive modeling.

3.1. 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. (2019) and 287 predictors (Table 2). Several categories of predictors with a documented influence on landscape and channel morphology were considered: (a) channel confinement (e.g., Fryirs et al., 2016); (b) stream network topology (e.g., Danesh-Yazdi et al., 2017; Strahler, 1957); (c) statistical roughness or fractal dimension of the topography (Dodds & Rothman, 2000; Duclut & Delamotte, 2017; Faghih & Nourbakhsh, 2015; Lifton & Chase, 1992; Liucci & Melelli, 2017; Pastor-Satorras & Rothman, 1998; Sung &

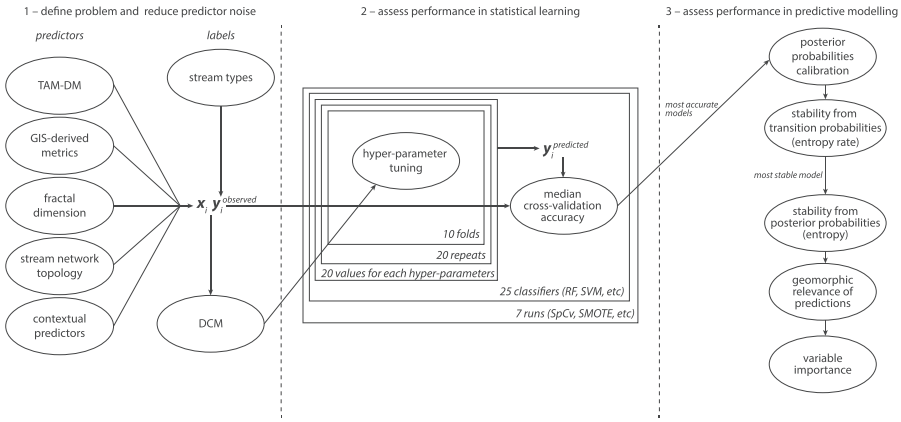

Figure 1. Schematic of the three-tiered machine-learning framework.

Chen, 2004; Wilson & Dominic, 1998); (d) contextual variables like geology, soils, land cover, and climate (e.g., Mercado-Bettín et al., 2019; Teutschbein et al., 2018); and (e) nine terrain analysis metrics (e.g., curvature, Hurst et al., 2012; Prancevic & Kirchner, 2019; Rosgen, 1994). Since the distribution of Terrain Analysis Metrics across the landscape differentiates stages of landscape maturity (Bonetti & Porporato, 2017) and channel types (Lane et al., 2017), 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 data set (NED ; Gesch et al., 2002), (ii) the stream network from the National Hydrology data set (NHDPlusV2; McKay et al., 2012), and (iii) the Stream-Catchment data set (StreamCat; Hill et al., 2015) 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 2). In addition, defining the relevant scales to distinguish controlling physical processes is often difficult (e.g., Archfield et al., 2015). To address this challenge, we considered 32 sets of five nested spatial scales with upper scales from ~0.6 to ~82 km when calculating the fractal dimension following the box-counting method from Liucci and Melelli (2017). This can be seen as an alternative to using multiscale topographic decomposition (e.g., Agarwal et al., 2017; Buscombe, 2016; Cazenave et al., 2012; Newman et al., 2018). 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 and six Distribution Metrics result in 108 Terrain Analysis Metrics-Distribution Metrics predictors (Table 2).

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. (2019), the labels are treated as noiseless and we refrained from implementing techniques that evaluate and correct for uncertain labeling (e.g., Borut et al., 2010; Garcia et al., 2012, 2015; Sluban et al., 2013). To filter potential predictor noise, the complexity of the problem was assessed by calculating data complexity measures (DCM; Lorena et al., 2018). DCMs offer a model agnostic tool (Ho & Basu, 2002; Lorena et al., 2017; Lorena et al., 2018) 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., Garcia et al., 2015). 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. (2018) (Table 3).

Table 2
Predictors Used in the Machine Learning Framework

Predictors group	Predictor name	Spatial scale	Original data	Methodology
TAM-DM (108)	Elevation	512 m; 100-m buffer	Gesch et al. (2002)	(Hijmans et al., 2018)
	Slope	512 m; 100-m buffer	Gesch et al. (2002)	(Hijmans et al., 2018)
	Aspect	512 m; 100-m buffer	Gesch et al. (2002)	(Hijmans et al., 2018)
	Roughness	512 m; 100-m buffer	Gesch et al. (2002)	(Hijmans et al., 2018)
	Flow direction	512 m; 100-m buffer	Gesch et al. (2002)	(Hijmans et al., 2018)
	Planform curvature	512 m; 100-m buffer	Gesch et al. (2002)	(Florinsky, 1998)
	Profile curvature	512 m; 100-m buffer	Gesch et al. (2002)	(Florinsky, 1998)
	Topographic position index	512 m; 100-m buffer	Gesch et al. (2002)	(Hijmans et al., 2018)
	Terrain ruggedness index	512 m; 100-m buffer	Gesch et al. (2002)	(Hijmans et al., 2018)
GIS metrics (3)	Channel slope	200 m	Gesch et al. (2002)	(ESRI, 2016)
	Confinement	—	Gesch et al. (2002)	(Byrne et al., 2019)
	Sediment supply	—	Haan et al. (1994)	Renard et al. (1997)
Network topology (4)	Drainage area	—	McKay et al. (2012)	(Hill et al., 2015)
	Strahler's stream order	—	McKay et al. (2012)	(Strahler, 1957)
	Local drainage density	—	McKay et al. (2012)	(Danesh-Yazdi et al., 2017)
Fractal dimension (32)	Hurst coefficients	640 m to 82 km	Gesch et al. (2002)	Liucci and Melelli (2017)
Contextual predictors (140)	Lithology	>1 km	Cress et al. (2010)	Hill et al. (2015)
	Soil characteristics	1 km	Schwarz and Alexander (1995)	Hill et al. (2015)
	Land cover	30-m initial resolution	Homer et al. (2015)	Hill et al. (2015)
	1981–2010 climatologies	800-m initial resolution	PRISM Climate Group (2004)	Hill et al. (2015)
	Indices of Catchment Integrity	—	Thornbrugh et al. (2018)	Hill et al. (2015)

Note. The 10-m National Elevation Data Set (NED; Gesch et al., 2002) and the Stream-Catchment Data Set (StreamCat; Hill et al., 2015) are publicly available on download platform from the U.S. Geological Survey and the U.S. Environmental Protection Agency, respectively. The stream network from the National Hydrology Data Set (NHDPlusV2; McKay et al., 2012) is publicly available on both platforms. TAM-DM: Terrain Analysis Metrics-Distribution Metrics.

3.2. Assess Performance in Statistical Learning

3.2.1. Statistical Learning

While numerous ML models exist, the best model for a given task is often unknown at the start (Luengo & Herrera, 2013), 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` (Kuhn, 2008; 2018) and `h2o` (H2O.ai, 2018).

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) (Cortes & Vapnik, 1995). 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., Lorena et al., 2008).

RF is an ensemble of classification and regression trees built from random subsets of predictors (Breiman et al., 1984). At each split of each tree, a predictor is chosen as splitting variable based on an information selection process (e.g., Gini coefficient) which ultimately provides a measure of variable importance. In

Table 3
DCMs Computed to Characterize the Problem Complexity and to Filter Predictor Noise

Category	Acronym	Name	Information
Dimensionality	T2	Average number of points per dimension	Sparsity of the data set
	T3	Average number of points per PCA dimension	Sparsity of the data set 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	Nonlinearity 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 interclass/extraclass 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

Note. We follow definitions from Lorena et al. (2018). PCA = principal component analysis; SVM = support vector machine; NN = nearest neighbor.

addition, the ensemble decision process from uncorrelated (random) trees leads to great performance when the training data set is reduced, noisy, or both (Fox et al., 2017).

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 environmental sciences (Bergen et al., 2019; Reichstein et al., 2019; Shen, 2018).

Prior to statistical learning, a set of transformations was applied to predictors (Table 4). 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., Csillik et al., 2015). 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 5). 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 1). In a base run, this data set imbalance was left untouched. The Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002) 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 (Burman, 1989). 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 (Schratz et al., 2018) and addresses the important issue of spatially correlated training data.

For each model (Table 4) and for each run (Table 5), the set of best hyperparameters was tuned by a grid search across 20 different values per parameter resulting in the training of ~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

Table 4
Types of Preprocessing

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	✓	✓	✗	✗	✗

Note. 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.

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., Breiman et al., 1984).

3.2.2. 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

Table 5
Type of Runs Performed to Evaluate the Influence of Predictors, Oversampling, and Resampling

Run	Predictors	Oversampling	Resampling
base-a	All	No	20 × 10-fold cross validation
base-b	No contextual predictors	No	20 × 10-fold cross validation
SpCv-b	No contextual predictors	No	20 × 10-fold spatial cross validation
SMOTE-a	All	SMOTE	20 × 10-fold cross validation
SMOTE-b	No contextual predictors	SMOTE	20 × 10-fold cross validation
SpCV-SMOTE-a	All	SMOTE	20 × 10-fold spatial cross validation
SpCV-SMOTE-b	No contextual predictors	SMOTE	20 × 10-fold spatial cross validation

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 (DeGroot & Fienberg, 1983; Niculescu-Mizil & Caruana, 2005; Zadrozny, 2002). Given the sigmoid shape of most distortions, Platt (1999) proposed a sigmoid calibration to address this effect. Other useful approaches include Bayesian calibration and isotonic scaling (Zadrozny & Elkan, 2002) which both require binarizing the problem (Adnan & Islam, 2015; Dong et al., 2005; Kijsirikul & Ussivakul, 2002; Lorena & de Carvalho, 2010; Melnikov & Hüllermeier, 2018; Platt et al., 2000; Quiterio & Lorena, 2016). 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 (Platt, 1999). The `R` package `glmnet` was used to fit a generalized linear model with an elastic net penalty and with a 10-fold cross validation.

3.3. 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 (Gneiting & Raftery, 2007). 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 (Shannon, 1948) and prognosticates the prediction skill of a model (Roulston & Smith, 2002; Stephenson & Dolas-Reyes, 2000). 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 (Daley & Vere-Jones, 2004) and indicates which model provides the best information (Nearing & Gupta, 2015).

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., Grant et al., 1990; Montgomery & Buffington, 1998), similar reasoning can be applied for point pattern using spatial statistics Appendix 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 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.

4. 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.

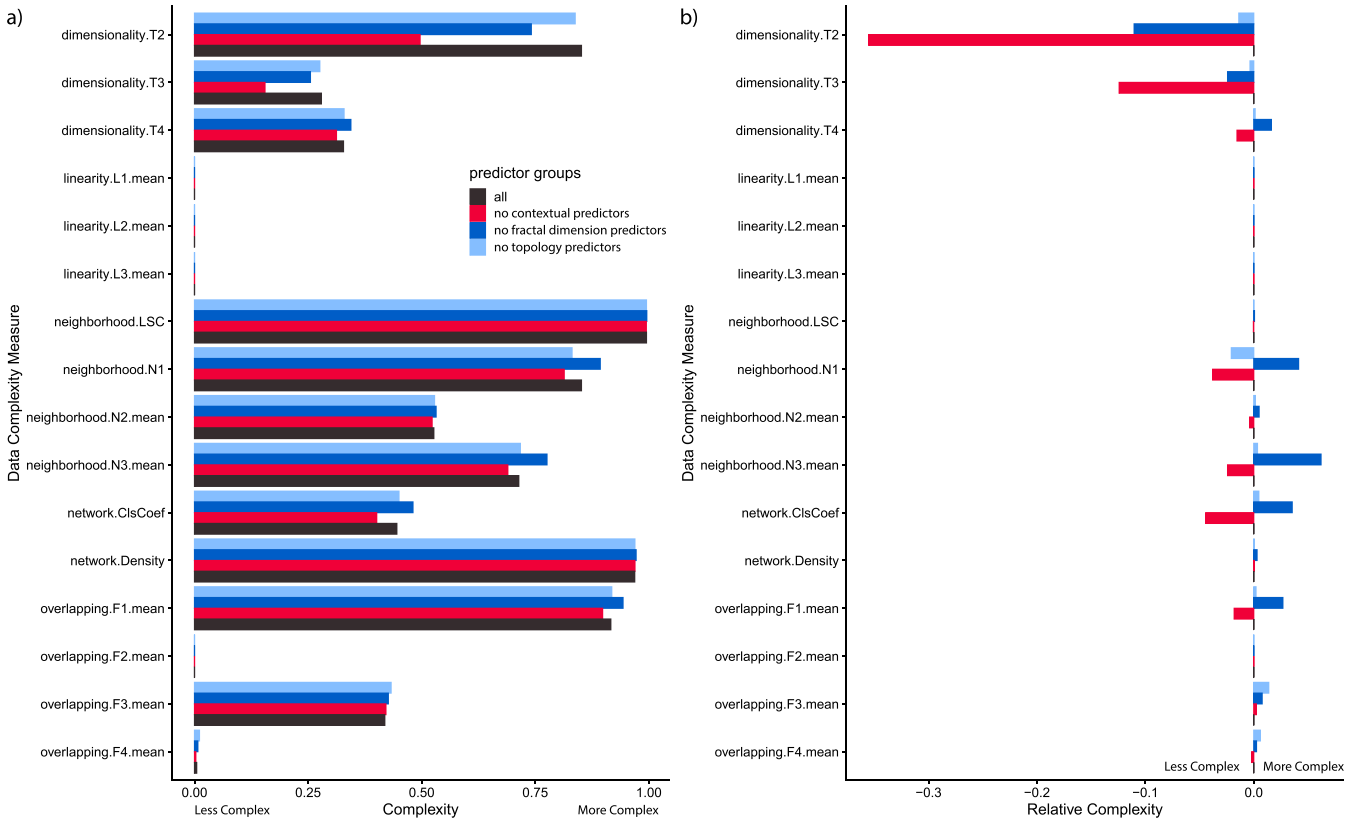

Figure 2. Summary of data complexity measures computed following Lorena et al. (2018) for four sets of predictors with the initial set of predictors in black: (a) DCMs values; (b) relative complexity to the initial predictor sets. Increasing values indicate increasing complexity. Except for F1, mean 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.

4.1. 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

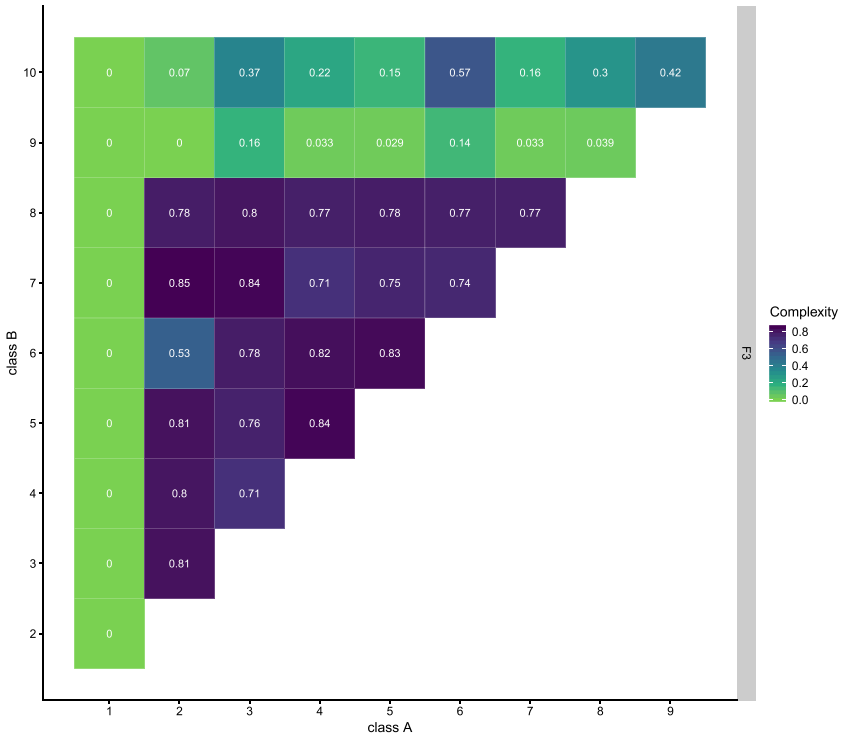

Figure 3. Maximum individual predictor efficiency (F3) for each pair of classes. For each pair of classes, values are between 0 and 1, with 0 indicating a simple problem and 1 a complex one.

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 1). 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.

Table 6
Results From the Different Runs Performed (Table 5)

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

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

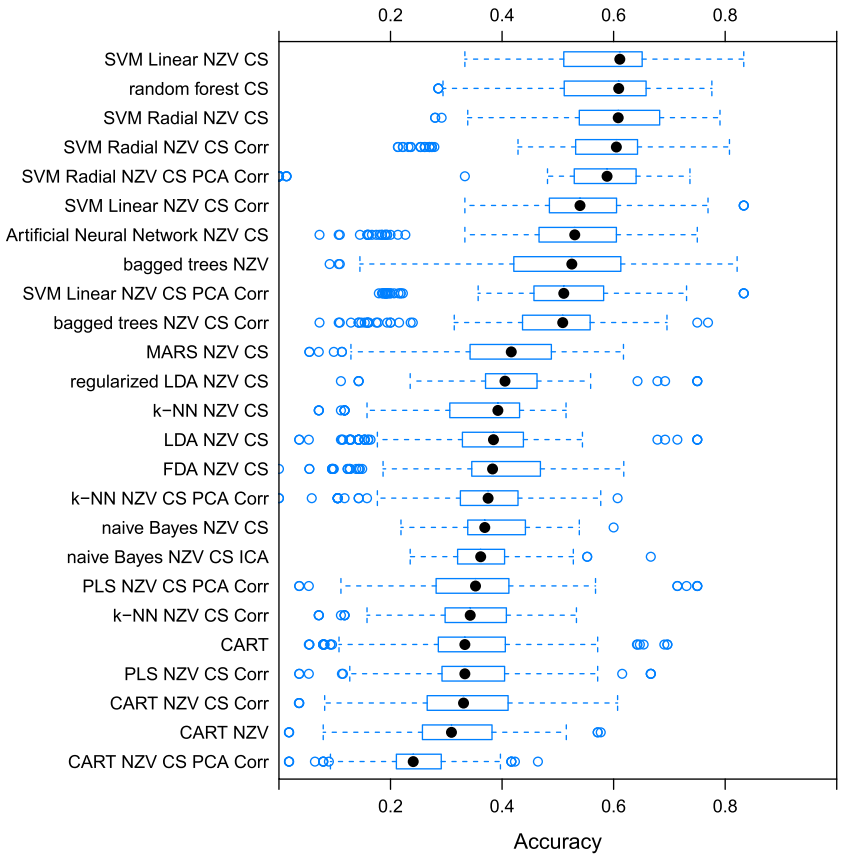

Figure 4. Distribution of cross-validation accuracies for all classifiers (Table 4) for the SpCV-SMOTE-b run for the Sacramento basin (Table 6). CS indicates that Box-Cox transformations were applied when required; NZV means that the predictors with near-zero variance were filtered; Corr indicates that highly correlated predictors were filtered; ICA and PCA mean that independent component analysis and principal component analysis were performed on the predictors, respectively.

4.2. Random Forest Outperforms Other Models

Seven different model runs were performed to assess the influence of predictors, oversampling, and resampling (Table 6). 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 6).

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 min), while R-SVM and RF were the most costly (>2 hr). Good performance of these models were expected from the 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 predictions

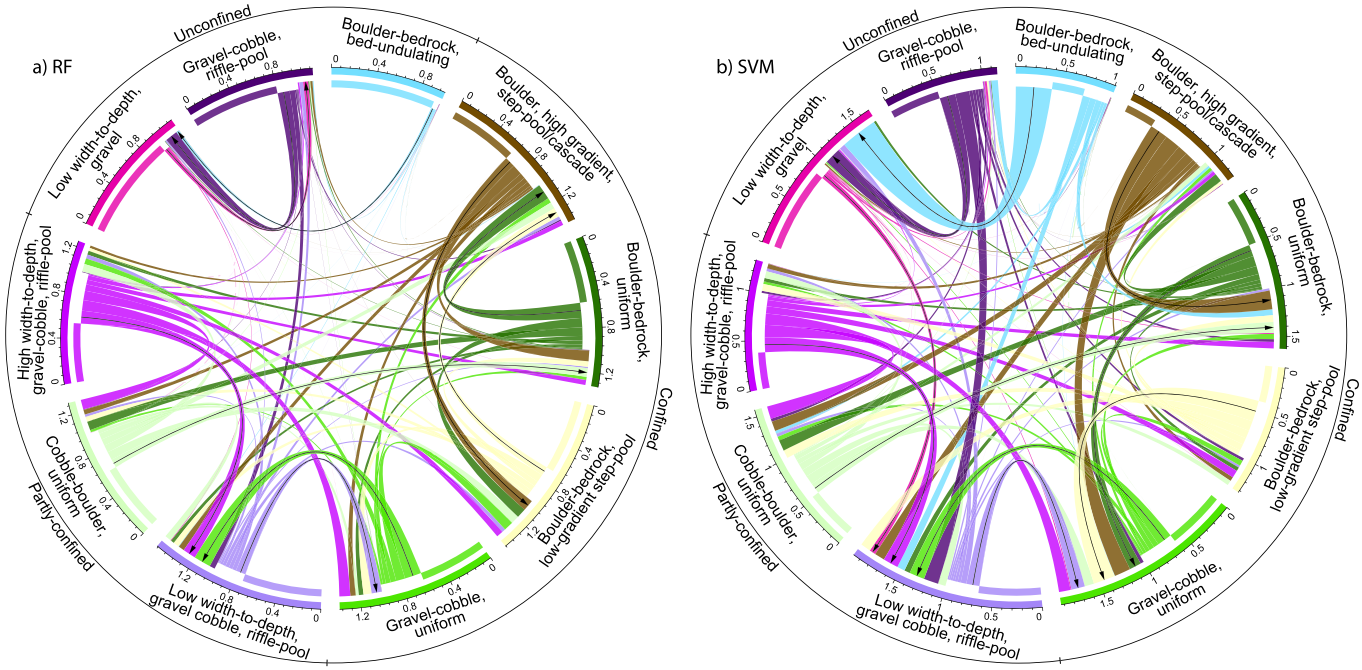

Figure 5. Chord diagram of the transition probabilities from one channel type to another in the Sacramento basin for (a) RF and (b) SVM models.

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 are not significantly correlated, making the decision process of the ensemble model more stable (Probst et al., 2018). 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).

4.3. 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.

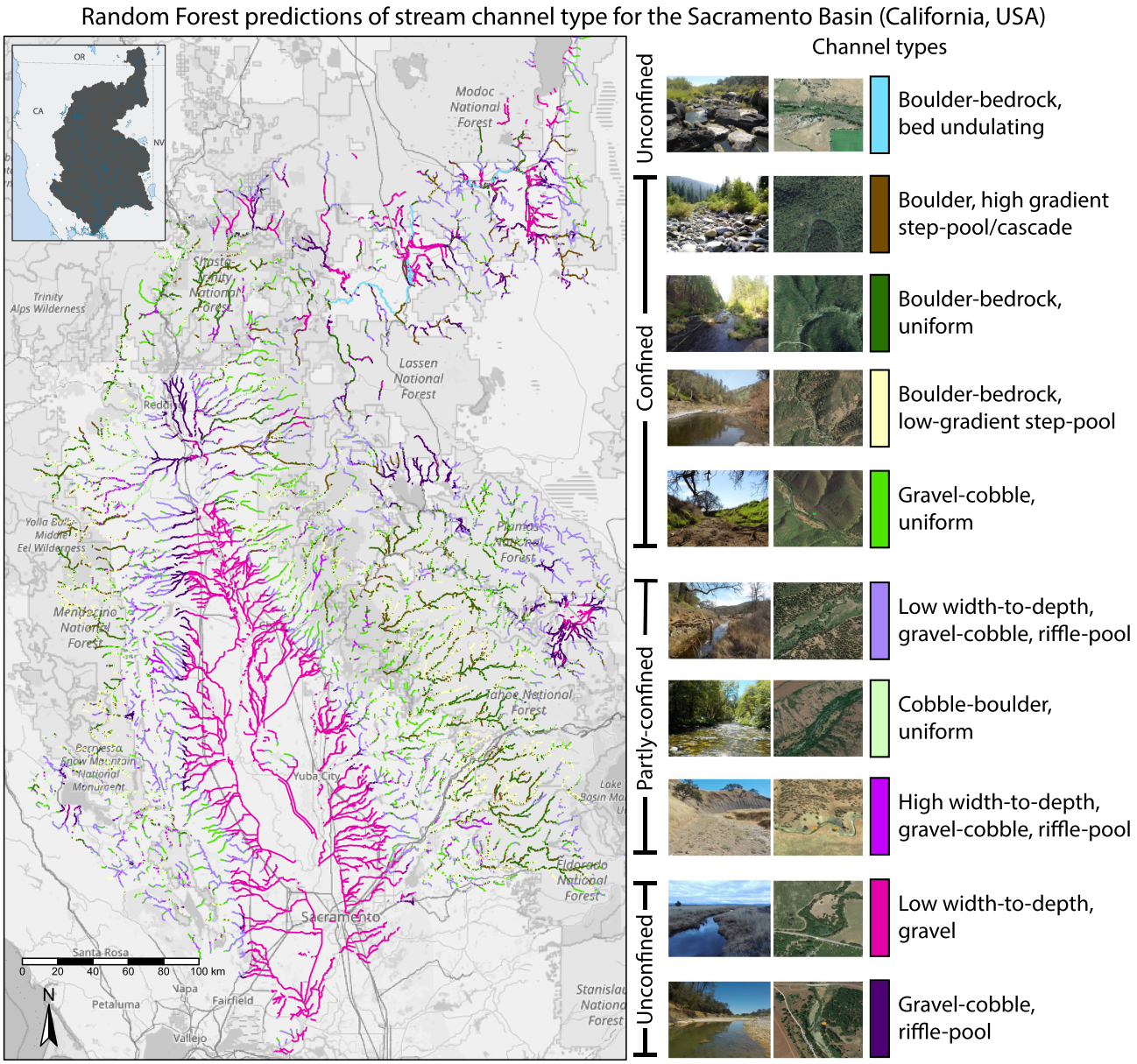

Figure 6. Map of the Sacramento basin with the spatial predictions of the 10 types of channel. Inset shows location of the study area in California and in the general setting of the Pacific Coast of the United States.

4.4. 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, and 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 and 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

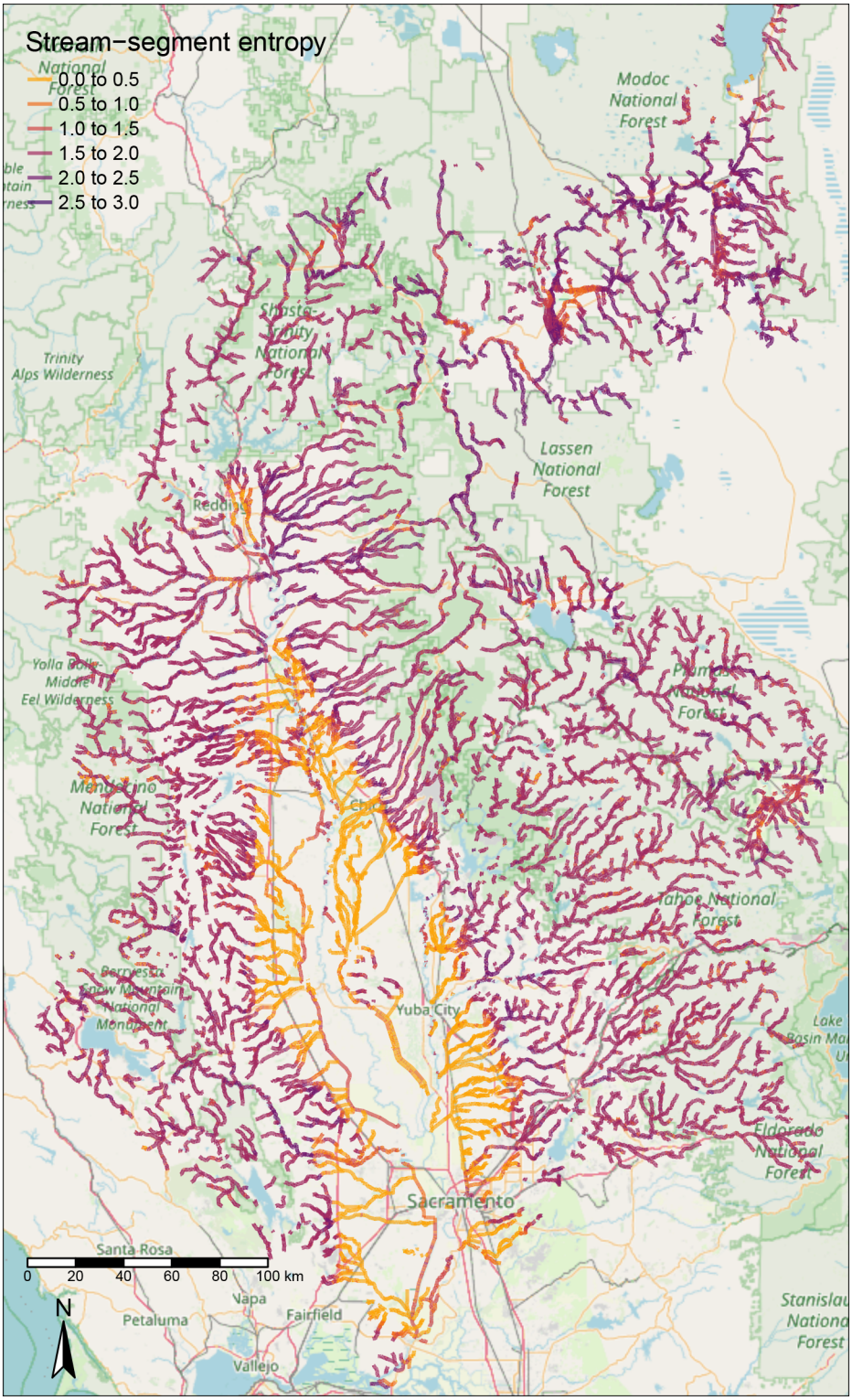

Figure 7. Stream-interval posterior entropy in the Sacramento basin.

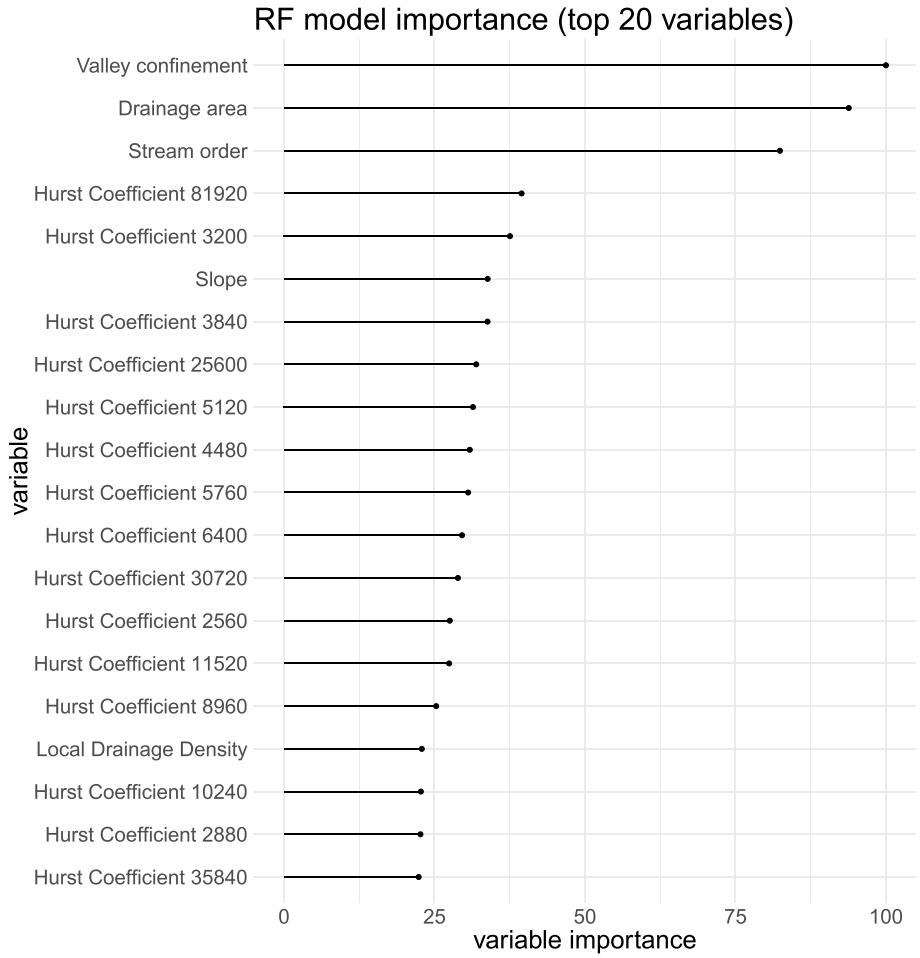

Figure 8. Variable importance of the RF model for the Sacramento basin.

and topology predictors is also suggested by the DCM analysis (Figures 2 and 3) and is supported by similar variable importance assessment for the ANN model.

5. Discussion

5.1. 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 (Figures 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 (Lane et al., 2017), surface-groundwater interactions (Yang et al., 2019), landsliding risk (Lee et al., 2018), or erosion processes (Rahmati et al., 2017): (i) using DCM to understand the characteristics of the problem and to reduce predictor noise significantly improves statistical learning accuracy (Figures 2 and 3 and Table 6); (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 (Danesh-Yazdi et al., 2017).

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 have been defined. The labels correspond to channel types defined from field survey data at a 10^0- to 10^2-m scale (Byrne et al., 2019). In contrast, the spatial resolution of the predictors lies typically between 10^2 and 10^5 m (Table 2). 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., Passalacqua et al., 2015). However, even with high-resolution topographic data, identifying the appropriate scale to describe local morphology remains challenging (Tarolli, 2014) and coarser resolution (5–10 m) may be best suited in some cases. For example, Lisenby and Fryirs (2017) 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 (Neeson et al., 2008). Such an error-prone estimate of channel slope might explain why such a dominant control on geomorphic processes (e.g., Grant et al., 1990) 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 (Ho, 1995) and its robustness to predictor noise (e.g., Fox et al., 2017), but deep learning approaches have been very successful at predicting complex patterns (e.g., LeCun et al., 2015). In particular, a finite deep learning network can approximate any function between input and output (Cybenko, 1989; Hornik et al., 1989) while avoiding local optimization minima (Baldassi et al., 2016) and keeping a relatively limited number of parameters (Lin et al., 2017). 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 (Lin et al., 2017). 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.

5.2. 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 (Jaynes, 1957a, 1957b). In information theory, entropy measures the uncertainty of the information transmitted by a noisy process (Shannon, 1948). In statistical mechanics, entropy quantifies the number of accessible microscopic configurations for particles in a gas (e.g., Gibbs, 1902). 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., Bagnold, 1966). Such a correlation harks back to early comparisons of entropy in geomorphic and thermodynamic system which suggested a complete analogy between elevation and temperature (Leopold & Langbein, 1962; Scheidegger, 1964; 1967). 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., Hayakawa & Oguchi, 2006). Recently, entropy was used to characterize the diversity of physical typology of river networks and define areas with similar hydrogeomorphic characteristics (Thoms et al., 2018). 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 hydrogeomorphic 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., Carr, 1997; Faghih & Nourbakhsh, 2015; Liucci & Melelli, 2017; Sung et al., 1998; Sung & Chen, 2004; Xu et al., 1993).

6. 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 (Daley & Vere-Jones, 2004; Nearing & Gupta, 2015).

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 inflection point of the Ripley's K and Besag's L^* function, and (iv) the inflection point of the pair correlation function $g(r)$. Ripley's K, Besag's L^*, and the pair correlation $g(r)$ functions characterize a point pattern (e.g., Illian et al., 2008) 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.

Acknowledgments
This research was supported by the California State Water Resources Control Board under Grant 16-062-300. We also acknowledge the U.S. Department of Agriculture, Hatch Projects CA-D-LAW-7034-H and CA-D-LAW-2243-H. Data sources are reported in Table 2.

References

Adnan, M. N., & Islam, M. Z. (2015). One-vs-all binarization technique in the context of random forest. In *Proceedings of the european symposium on artificial neural networks, computational intelligence and machine learning* (pp. 385–390). Bruges, Belgium.

Agarwal, A., Marwan, N., Rathinasamy, M., Merz, B., & Kurths, J. (2017). Multi-scale event synchronization analysis for unravelling climate processes: A wavelet-based approach. *Nonlinear Processes in Geophysics*, *24*(4), 599–611. https://doi.org/10.5194/npg-24-599-2017

Alber, A., & Piégay, H. (2011). Spatial disaggregation and aggregation procedures for characterizing fluvial features at the network-scale: Application to the Rhône basin (France). *Geomorphology*, *125*(3), 343–360. https://doi.org/10.1016/j.geomorph.2010.09.009

Alexander, J. S., Zelt, R. B., & Schaepe, N. J. (2009). Geomorphic segmentation, hydraulic geometry, and hydraulic microhabitats of the Niobrara River, Nebraska—Methods and initial results: U.S. Geological Survey Scientific Investigations Report.

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., et al. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, *51*, 10,078–10,091. https://doi.org/10.1002/2015wr017498

Bagnold, R. A. (1966). An approach to the sediment transport problem from general physics. *USGS Professional Paper*, *422-I*(I), 37. Retrieved from http://www.journals.cambridge.org/abstract_S0016756800049074

Baldassi, C., Borgs, C., Chayes, J. T., Ingrosso, A., Lucibello, C., Saglietti, L., & Zecchina, R. (2016). Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, *113*(48), E7655–E7662. https://doi.org/10.1073/pnas.1608103113

Bassiouni, M., Higgins, C. W., Still, C. J., & Good, S. P. (2018). Probabilistic inference of ecohydrological parameters using observations from point to satellite scales. *Hydrology and Earth System Sciences*, *22*(6), 3229–3243. https://doi.org/10.5194/hess-22-3229-2018

Beechie, T., & Imaki, H. (2014). Predicting natural channel patterns based on landscape and geomorphic controls in the Columbia river basin, USA. *Water Resources Research*, *50*, 39–57. https://doi.org/10.1002/2013WR013629

Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433), eaau0323.

Bonetti, S., & Porporato, A. (2017). On the dynamic smoothing of mountains. *Geophysical Research Letters*, *44*, 5531–5539. https://doi.org/10.1002/2017gl073095

Borut, S., Dragan, G., & Nada, L. (2010). Advances in class noise detection. *Frontiers in Artificial Intelligence and Applications*, *215*(ECAI 2010), 1105–1106. https://doi.org/10.3233/978-1-60750-606-5-1105

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

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

Buffington, J. M., & Montgomery, D. R. (2013). Geomorphic classification of rivers, *Treatise on geomorphology* (pp. 730–767). San Diego, CA: Elsevier. https://doi.org/10.1016/b978-0-12-374739-6.00263-3

Bugnicourt, P., Guitet, S., Santos, V. F., Blanc, L., Sotta, E. D., Barbier, N., & Couteron, P. (2018). Using textural analysis for regional landform and landscape mapping, Eastern Guiana Shield. *Geomorphology*, *317*, 23–44. https://doi.org/10.1016/j.geomorph.2018.03.017

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods<. *Biometrika*, *76*(3), 503. https://doi.org/10.2307/2336116

Buscombe, D. (2016). Spatially explicit spectral analysis of point clouds and geospatial data. *Computers & Geosciences*, *86*, 92–108. https://doi.org/10.1016/j.cageo.2015.10.004

Byrne, C., Pasternack, G., Guillon, H., Lane, B., & Solis, S. S. (2019). Reach-scale bankfull channel types can exist independently of catchment hydrology. Retrieved from https://doi.org/10.1002/essoar.10501068.1

Carr, J. R. (1997). Statistical self-affinity, fractal dimension, and geologic interpretation. *Engineering geology*, *48*(3-4), 269–282.

Cazenave, P. W., Dix, J. K., Lambkin, D. O., & McNeill, L. C. (2012). A method for semi-automated objective quantification of linear bedforms from multi-scale digital elevation models. *Earth Surface Processes and Landforms*, *38*(3), 221–236. https://doi.org/10.1002/esp.3269

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Clubb, F. J., Bookhagen, B., & Rheinwalt, A. (2019). Clustering river profiles to classify geomorphic domains. *Journal of Geophysical Research: Earth Surface*, *124*, 1417–1439. https://doi.org/10.1029/2019jf005025

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Cress, J., Soller, D., Sayre, R., Comer, P., & Warner, H. (2010). Terrestrial ecosystems surficial lithology of the conterminous United States. [Computer software manual]. Retrieved from https://pubs.usgs.gov/sim/3126/ (U.S. Geological Survey Scientific Investigations Map 3126, scale 1:5,000,000, 1 sheet).

Csillik, O., Evans, I. S., & Drăguţ, L. (2015). Transformation (normalization) of slope gradient and surface curvatures, automated for statistical analyses from DEMs. *Geomorphology*, *232*, 65–77. https://doi.org/10.1016/j.geomorph.2014.12.038

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303–314.

Daley, D. J., & Vere-Jones, D. (2004). Scoring probability forecasts for point processes: The entropy score and information gain. *Journal of Applied Probability*, *41*(A), 297–312.

Danesh-Yazdi, M., Tejedor, A., & Foufoula-Georgiou, E. (2017). Self-dissimilar landscapes: Revealing the signature of geologic constraints on landscape dissection via topologic and multi-scale analysis. *Geomorphology*, *295*, 16–27. https://doi.org/10.1016/j.geomorph.2017.06.009

Davis, W. M. (1899). The geographical cycle. *The Geographical Journal*, *14*(5), 481–504.

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *32*(1-2), 12–22.

Dodds, P. S., & Rothman, D. H. (2000). Scaling, universality, and geomorphology. *Annual Review of Earth and Planetary Sciences*, *28*(1), 571–610. https://doi.org/10.1146/annurev.earth.28.1.571

Dong, L., Frank, E., & Kramer, S. (2005). Ensembles of balanced nested dichotomies for multi-class problems, *European conference on principles of data mining and knowledge discovery* (pp. 84–95). Berlin, Heidelberg: Springer.

Duclut, C., & Delamotte, B. (2017). Nonuniversality in the erosion of tilted landscapes. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, *96*, 12149.

ESRI (2016). Arcgis desktop [Computer software manual]. Redlands, CA.

Faghih, A., & Nourbakhsh, A. (2015). Implication of surface fractal analysis to evaluate the relative sensitivity of topography to active tectonics, Zagros Mountains, Iran. *Journal of Mountain Science*, *12*(1), 177–185. https://doi.org/10..1007/s11629-014-3005-5

Florinsky, I. V. (1998). Accuracy of local topographic variables derived from digital elevation models. *International Journal of Geographical Information Science*, *12*(1), 47–62. https://doi.org/10.1080/136588198242003

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment*, *189*(7), 316. https://doi.org/10.1007/s10661-017-6025-0

Fryirs, K. A., Wheaton, J. M., & Brierley, G. J. (2016). An approach for measuring confinement and assessing the influence of valley setting on river forms and processes. *Earth Surface Processes and Landforms*, *41*(5), 701–710. https://doi.org/10.1002/esp.3893

Garcia, L. P., de Carvalho, A. C., & Lorena, A. C. (2015). Effect of label noise in the complexity of classification problems. *Neurocomputing*, *160*, 108–119.

Garcia, L. P. F., Lorena, A. C., & Carvalho, A. C. (2012). A study on class noise detection and elimination. In *2012 brazilian symposium on neural networks*, IEEE. https://doi.org/10.1109/sbrn.2012.49

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., & Tyler, D. (2002). The national elevation dataset. *Photogrammetric engineering and remote sensing*, *68*(1), 5–32.

Gibbs, J. W. (1902). *Elementary principles in statistical mechanics*. New Haven, Conn, Yale: University Press.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Golly, A., & Turowski, J. M. (2017). Deriving principal channel metrics from bank and long-profile geometry with the R package cmgo. *Earth Surface Dynamics*, *5*(3), 557–570. https://doi.org/10.5194/esurf-5-557-2017

Grant, G. E., Swanson, F. J., & Wolman, M. G. (1990). Pattern and origin of stepped-bed morphology in high-gradient streams, Western Cascades, Oregon. *Geological Society of America Bulletin*, *102*(3), 340–352.

H2O.ai (2018). R interface for h2o. [Computer software manual]. Retrieved from https://github.com/h2oai/h2o-3 (R package version 3.20.0.8).

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

Hamshaw, S. D., Dewoolkar, M. M., Schroth, A. W., Wemple, B. C., & Rizzo, D. M. (2018). A new machine-learning approach for classifying hysteresis in suspended-sediment discharge relationships using high-frequency monitoring data. *Water Resources Research*, *54*, 4040–4058. https://doi.org/10.1029/2017WR022238

Hassan, M. A., Bird, S., Reid, D., Ferrer-Boix, C., Hogan, D., Brardinoni, F., & Chartrand, S. (2018). Variable hillslope-channel coupling and channel characteristics of forested mountain streams in glaciated landscapes. *Earth Surface Processes and Landforms*, *44*(3), 736–751. https://doi.org/10.1002/esp.4527

Hayakawa, Y. S., & Oguchi, T. (2006). DEM-based identification of fluvial knickzones and its application to japanese mountain rivers. *Geomorphology*, *78*(1-2), 90–106.

Hijmans, R. J., van Etten, J., Cheng, J., Greenberg, J. A., Lamigueiro, O. P., & Bevan, A. (2018). Package 'raster'. version 2.6-7.

Hill, R. A., Weber, M. H., Leibowitz, S. G., Olsen, A. R., & Thornbrugh, D. J. (2015). The stream-catchment (StreamCat) dataset: A database of watershed metrics for the conterminous United States. *JAWRA Journal of the American Water Resources Association*, *52*(1), 120–128. https://doi.org/10.1111/1752-1688.12372

Ho, T. K. (1995). Random decision forests, *Proceedings of 3rd international conference on document analysis and recognition* (vol. 1, pp. 278–282). Montreal, Quebec, Canada, Canada: IEEE.

Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 289–300. https://doi.org/10.1109/34.990132

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., et al. (2015). Completion of the 2011 national land cover database for the conterminous United States—Representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*, *81*(5), 345–354.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Hurst, M. D., Mudd, S. M., Walcott, R., Attal, M., & Yoo, K. (2012). Using hilltop curvature to derive the spatial distribution of erosion rates. *Journal of Geophysical Research*, *117*, F02017. https://doi.org/10.1029/2011jf002057

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns* (Vol. 70). Hoboken, NJ: John Wiley & Sons.
Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical review*, *106*(4), 620.
Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical review*, *108*(2), 171.
Jha, R., & Diplas, P. (2017). Elevation: A consistent and physically-based framework for classifying streams. *Journal of Hydraulic Research*, *56*(3), 299–312. https://doi.org/10.1080/00221686.2017.1354928
Kasprak, A., Hough-Snee, N., Beechie, T., Bouwes, N., Brierley, G., Camp, R., et al. (2016). The blurred line between form and process: A comparison of stream channel classification frameworks. *PloS one*, *11*(3), e0150293.
Kijsirikul, B., & Ussivakul, N. (2002). Multiclass support vector machines using adaptive directed acyclic graph. In *Proceedings of the 2002 international joint conference on neural networks. IJCNN'02 (cat. no.02ch37290)*, IEEE. https://doi.org/10.1109/ijcnn.2002.1005608
Kondolf, G. M., Piégay, H., Schmitt, L., & Montgomery, D. R. (2016). Geomorphic classification of rivers and streams. In *Tools in fluvial geomorphology* (Chap. 7, pp. 133–158). John Wiley & Sons. https://doi.org/10.1002/9781118648551.ch7
Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019). Neuralhydrology—Interpreting LSTMs in hydrology. arXiv preprint arXiv:1903.07903.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, Articles*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05
Kuhn, M. (2018). Package 'caret'. [Computer software manual]. (version 6.0-80).
Lane, B. A., Dahlke, H. E., Pasternack, G. B., & Sandoval-Solis, S. (2017). Revealing the diversity of natural hydrologic regimes in california with relevance for environmental flows applications. *JAWRA Journal of the American Water Resources Association*, *53*(2), 411–430.
Lane, B. A., Pasternack, G., Dahlke, E., & Sandoval-Solis, S. (2017). The role of topographic variability in river channel classification. *Progress in Physical Geography*, *41*, 3091333177181.
Lane, B. A., Pasternack, G. B., & Sandoval-Solis, S. S. (2018). Integrated analysis of flow, form, and function for river management and design testing. *Ecohydrology*, *11*(5), e1969. https://doi.org/10.1002/eco.1969
Lane, B. A., Sandoval-Solis, S., Stein, E. D., Yarnell, S. M., Pasternack, G. B., & Dahlke, H. E. (2018). Beyond metrics? the role of hydrologic baseline archetypes in environmental water management. *Environmental Management*, *62*, 678–693. https://doi.org/10.1007/s00267-018-1077-7
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436. https://doi.org/10.1038/nature14539
Lee, J.-H., Sameen, M. I., Pradhan, B., & Park, H.-J. (2018). Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. *Geomorphology*, *303*, 284–298. https://doi.org/10.1016/j.geomorph.2017.12.007
Leopold, L. B., & Langbein, W. B. (1962). The concept of entropy in landscape evolution. In *Geological Survey Professional Paper* (Vol. 500, pp. 1–19).
Lifton, N. A., & Chase, C. G. (1992). Tectonic, climatic and lithologic influences on landscape fractal dimension and hypsometry: Implications for landscape evolution in the San Gabriel Mountains, California. *Geomorphology*, *5*(1), 77–114. https://doi.org/10.1016/0169-555X(92)90059-W
Lin, H. W., Tegmark, M., & Rolnick, D. (2017). Why does deep and cheap learning work so well? *Journal of Statistical Physics*, *168*(6), 1223–1247. https://doi.org/10.1007/s10955-017-1836-5
Lisenby, P. E., & Fryirs, K. A. (2017). 'Out with the old?' Why coarse spatial datasets are still useful for catchment-scale investigations of sediment (dis)connectivity. *Earth Surface Processes and Landforms*, *42*(10), 1588–1596. https://doi.org/10.1002/esp.4131
Liucci, L., & Melelli, L. (2017). The fractal properties of topography as controlled by the interactions of tectonic, lithological, and geomorphological processes. *Earth Surface Processes and Landforms*, *42*, 2585–2598. https://doi.org/10.1002/esp.4206
Lorena, A. C., & de Carvalho, A. C. P. L. F. (2010). Building binary-tree-based multiclass classifiers using separability measures. *Neurocomputing*, *73*(16-18), 2837–2845. https://doi.org/10.1016/j.neucom.2010.03.027
Lorena, A. C., de Carvalho, A. C. P. L. F., & Gama, J. M. P. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, *30*(1-4), 19–37. https://doi.org/10.1007/s10462-009-9114-9
Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P., & Ho, T. K. (2018). How complex is your classification problem? A survey on measuring classification complexity.
Lorena, A. C., Maciel, A. I., de Miranda, P. B. C., Costa, I. G., & Prudêncio, R. B. C. (2017). Data complexity meta-features for regression problems. *Machine Learning*, *107*(1), 209–246. https://doi.org/10.1007/s10994-017-5681-1
Luengo, J., & Herrera, F. (2013). An automatic extraction method of the domains of competence for learning classifiers using data complexity measures. *Knowledge and Information Systems*, *42*(1), 147–180. https://doi.org/10.1007/s10115-013-0700-4
McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., & Rea, A. (2012). Nhdplus version 2: User guide. [Computer software manual].
McManamay, R. A., Troia, M. J., DeRolph, C. R., Sheldon, A. O., Barnett, A. R., Kao, S.-C., & Anderson, M. G. (2018). A stream classification system to explore the physical habitat diversity and anthropogenic impacts in riverscapes of the eastern United States. *PLOS ONE*, *13*(6), e0198439. https://doi.org/10.1371/journal.pone.0198439
Melnikov, V., & Hüllermeier, E. (2018). On the effectiveness of heuristics for learning nested dichotomies: An empirical analysis. *Machine Learning*, *107*, 1537–1560. https://doi.org/10.1007/s10994-018-5733-1
Mercado-Bettín, D., Salazar, J. F., & Villegas, J. C. (2019). Long-term water balance partitioning explained by physical and ecological characteristics in world river basins. *Ecohydrology*, e2072. https://doi.org/10.1002/eco.2072
Michie, D. (1968). "Memo" functions and machine learning. *Nature*, *218*(5136), 19.
Montgomery, D. R., & Buffington, J. M. (1997). Channel-reach morphology in mountain drainage basins. *Geological Society of America Bulletin*, *109*(5), 596–611.
Montgomery, D. R., & Buffington, J. M. (1998). *River ecology and management: Lessons from the Pacific coastal ecoregion Edited by* R. Naiman, & R. Bilby, pp. 13–42. New York: Springer-Verlag.
Mount, J. F. (1995). *California rivers and streams: The conflict between fluvial process and land use.* United States: Univ of California Press.
Nearing, G. S., & Gupta, H. V. (2015). The quantity and quality of information in hydrologic models. *Water Resources Research*, *51*, 524–538. https://doi.org/10.1002/2014wr015895
Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., & Xia, Y. (2016). Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *Journal of hydrometeorology*, *17*(3), 745–759.
Neeson, T. M., Gorman, A. M., Whiting, P. J., & Koonce, J. F. (2008). Factors affecting accuracy of stream channel slope estimates derived from geographical information systems. *North American Journal of Fisheries Management*, *28*(3), 722–732.
Newman, D. R., Lindsay, J. B., & Cockburn, J. M. H. (2018). Evaluating metrics of local topographic position for multiscale geomorphometric analysis. *Geomorphology*, *312*, 40–50. https://doi.org/10.1016/j.geomorph.2018.04.003

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning - ICML '05*, ACM Press. https://doi.org/10.1145/1102351.1102430

O'Brien, G. O., & Wheaton, J. (2014). River styles report for the middle Fork John Day Watershed, Oregon: Ecogeomorphology and Topographic Analysis Lab, Utah State University.

O'Brien, G. R., Wheaton, J., Fryirs, K., McHugh, P., Bouwes, N., Brierley, G., & Jordan, C. (2017). A geomorphic assessment to inform strategic stream restoration planning in the middle Fork John Day Watershed, Oregon, USA. *Journal of Maps*, *13*(2), 369–381. https://doi.org/10.1080/17445647.2017.1313787

PRISM Climate Group (2004). Prism gridded climate data. [Computer software manual]. Retrieved from http://prism.oregonstate.edu

Passalacqua, P., Belmont, P., Staley, D. M., Simley, J. D., Arrowsmith, J. R., Bode, C. A., et al. (2015). Analyzing high resolution topography for advancing the understanding of mass and energy transfer through landscapes: A review. *Earth-Science Reviews*, *148*, 174–193. https://doi.org/10.1016/j.earscirev.2015.05.012

Pastor-Satorras, R., & Rothman, D. H. (1998). Scaling of a slope: The erosion of tilted landscapes. *Journal of Statistical Physics*, *93*(3/4), 477–500. https://doi.org/10.1023/B:JOSS.0000033160.59155.c6

Perry, GeorgeL. W., & Dickson, M. E. (2018). Using machine learning to predict geomorphic disturbance: The effects of sample size, sample prevalence, and sampling strategy. *Journal of Geophysical Research: Earth Surface*, *123*, 2954–2970. https://doi.org/10.1029/2018jf004640

Pham, B. T., Prakash, I., & Bui, D. T. (2018). Spatial prediction of landslides using a hybrid machine learning approach based on random subspace and classification and regression trees. *Geomorphology*, *303*, 256–270. https://doi.org/10.1016/j.geomorph.2017.12.008

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61–74.

Platt, J. C., Cristianini, N., & Shawe-Taylor, J. (2000). Large margin dags for multiclass classification. In *Advances in neural information processing systems*, pp. 547–553.

Prancevic, J. P., & Kirchner, J. W. (2019). Topographic controls on the extension and retraction of flowing streams. *Geophysical Research Letters*, *46*, 2084–2092. https://doi.org/10.1029/2018gl081799

Probst, P., Wright, M., & Boulesteix, A.-L. (2018). Hyperparameters and tuning strategies for random forest. arXiv preprint arXiv:1804.03515.

Quiterio, T. M., & Lorena, A. C. (2016). Determining the structure of decision directed acyclic graphs for multiclass classification problems, *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)* pp. 115–120). Recife, Brazil: IEEE.

Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H. R., & Feizizadeh, B. (2017). Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. *Geomorphology*, *298*, 118–137. https://doi.org/10.1016/j.geomorph.2017.09.006

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Renard, K. G., Foster, G. R., Weesies, G. A., McCool, D. K., & Yoder, D. C. (1997). *Predicting soil erosion by water: A guide to conservation planning with the revised universal soil loss equation (RUSLE)*, vol. 703. Washington, DC: United States Department of Agriculture.

Rinaldi, M., Belletti, B., Comiti, F., Nardi, L., Bussettini, M., Mao, L., & Gurnell, A. M. (2015). 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 7th framework programme under grant agreement 282656.

Rosgen, D. L. (1994). A classification of natural rivers. *Catena*, *22*(3), 169–199. https://doi.org/10.1016/0341-8162(94)90001-9

Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, *130*(6), 1653–1660.

Roux, C., Alber, A., Bertrand, M., Vaudor, L., & Piégay, H. (2015). "FluvialCorridor": A new ArcGIS toolbox package for multiscale riverscape exploration. *Geomorphology*, *242*, 29–37. https://doi.org/10.1016/j.geomorph.2014.04.018

Ruddell, B. L., Drewry, D., & Nearing, G. S. (2019). Information theory for model diagnostics: Tradeoffs between functional and predictive performance in ecohydrology models. *Water Resources Research*, *55*, 6534–6554. https://doi.org/10.1029/2018WR023692

Santos, V. M., Wahl, T., Long, J. W., Passeri, D. L., & Plant, N. G. (2019). Combining numerical and statistical models to predict storm-induced dune erosion. *Journal of Geophysical Research: Earth Surface*, *124*, 1817–1834. https://doi.org/10.1029/2019jf005016

Scheidegger, A. (1964). Some implications of statistical mechanics in geomorphology. *International Association of Scientific Hydrology. Bulletin*, *9*(1), 12–16. https://doi.org/10.1080/02626666409493650

Scheidegger, A. (1967). A complete thermodynamic analogy for landscape evolution. *International Association of Scientific Hydrology. Bulletin*, *12*(4), 57–62. https://doi.org/10.1080/02626666709493550

Schratz, P., Muenchow, J., Richter, J., & Brenning, A. (2018). Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. arXiv preprint arXiv:1803.11266.

Schwarz, G. E., & Alexander, R. (1995). State soil geographic (STATSGO) data base for the conterminous united states. U.S. Geological Survey.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*, 8558–8593. https://doi.org/10.1029/2018wr022643

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656. https://doi.org/10.5194/hess-22-5639-2018

Sluban, B., Gamberger, D., & Lavrač, N. (2013). Ensemble-based noise detection: Noise ranking and visual performance evaluation. *Data Mining and Knowledge Discovery*, *28*(2), 265–303. https://doi.org/10.1007/s10618-012-0299-1

Stephenson, D. B., & Dolas-Reyes, F. J. (2000). Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus A: Dynamic Meteorology and Oceanography*, *52*(3), 300–322.

Strahler, A. N. (1957). Quantitative analysis of watershed geomorphology. *Transactions, American Geophysical Union*, *38*(6), 913. https://doi.org/10.1029/tr038i006p00913

Sung, Q.-C., & Chen, Y.-C. (2004). Self-affinity dimensions of topography and its implications in morphotectonics: An example from taiwan. *Geomorphology*, *62*(3-4), 181–198. https://doi.org/10.1016/j.geomorph.2004.02.012

Sung, Q.-C., Chen, Y.-C., & Chao, P. C. (1998). Spatial variation of fractal parameters and its geological implications. *Terrestrial, Atmospheric and Oceanic sciences journal*, *9*(4), 655–672.

Tarolli, P. (2014). High-resolution topography for understanding earth surface processes: Opportunities and challenges. *Geomorphology*, *216*, 295–312. https://doi.org/10.1016/j.geomorph.2014.03.008

Teutschbein, C., Grabs, T., Laudon, H., Karlsen, R. H., & Bishop, K. (2018). Simulating streamflow in ungauged basins under a changing climate: The importance of landscape characteristics. *Journal of Hydrology*, *561*, 160–178. https://doi.org/10.1016/j.jhydrol.2018.03.060

Thoms, M., Scown, M., & Flotemersch, J. (2018). Characterization of river networks: A GIS approach and its applications. *JAWRA Journal of the American Water Resources Association*, *54*, 899–913. https://doi.org/10.1111/1752-1688.12649

Thornbrugh, D. J., Leibowitz, S. G., Hill, R. A., Weber, M. H., Johnson, Z. C., Olsen, A. R., et al. (2018). Mapping watershed integrity for the conterminous United States. *Ecological Indicators*, *85*, 1133–1148. https://doi.org/10.1016/j.ecolind.2017.10.070

Vaughan, A. A., Belmont, P., Hawkins, C. P., & Wilcock, P. (2017). Near-channel versus watershed controls on sediment rating curves. *Journal of Geophysical Research: Earth Surface*, *122*, 1901–1923. https://doi.org/10.1002/2016JF004180

Wilson, T. H., & Dominic, J. (1998). Fractal interrelationships between topography and structure. *Earth Surface Processes and Landforms*, *23*(6), 509–525. https://doi.org/10.1002/(SICI)1096-9837(199806)23:6<509::AID-ESP864>3.0.CO;2-D

Xu, T., Moore, I. D., & Gallant, J. C. (1993). Fractals, fractal dimensions and landscapes—A review. *Geomorphology*, *8*(4), 245–262. https://doi.org/10.1016/0169-555X(93)90022-T

Yang, J., Griffiths, J., & Zammit, C. (2019). National classification of surface-groundwater interaction using random forest machine learning technique. *River Research and Applications*, *35*(7), 932–943.

Zadrozny, B. (2002). Reducing multiclass to binary by coupling probability estimates. In *Advances in neural information processing systems*. (pp. 1041–1048).

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth acm international conference on knowledge discovery and data mining*. (pp. 694–699). New York: ACM Press.