Received: 12 September 2023 | Revised: 13 July 2024 | Accepted: 22 August 2024

DOI: 10.1002/esp.5984

RESEARCH ARTICLE

Mind the information gap: How sampling and clustering impact the predictability of reach-scale channel types in California (USA)

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

[1]University of California Davis, Davis, CA, United States

[2]Utah State University, Logan, UT, United States

Correspondence
Corresponding author Hervé Guillon,
Email: herve@guillon.xyz

Abstract

Clustering and machine learning-based predictions are increasingly used for environmental data analysis and management. In fluvial geomorphology, examples include predicting channel types throughout a river network and segmenting river networks into a series of channel types, or groups of channel forms. However, when relevant information is unevenly distributed throughout a river network, the discrepancy between data-rich and data-poor locations creates an information gap. Combining clustering and predictions addresses this information gap, but challenges and limitations remain poorly documented. This is especially true when considering that predictions are often achieved with two approaches that are meaningfully different in terms of information processing: decision trees (e.g., RF: random forest) and deep learning (e.g., DNN: deep neural networks). This presents challenges for downstream management decisions and when comparing clusters and predictions within or across study areas. To address this, we investigate the performance of RF and DNN with respect to the information gap between clustering data and prediction data. We use nine regional examples of clustering and predicting river channel types, stemming from a single clustering methodology applied in California, USA. Our results show that prediction performance decreases when the information gap between field-measured data and geospatial predictors increases. Furthermore, RF outperforms DNN, and their difference in performance decreases when the information gap between field-measured and geospatial data decreases. This suggests that mismatched scales between field-derived channel types and geospatial predictors hinder sequential information processing in DNN. Finally, our results highlight a sampling trade-off between uniformly capturing geomorphic variability and ensuring robust generalization.

KEYWORDS

Fluvial geomorphology, Channel type classification, Random forest, Deep learning, Information theory

1 | INTRODUCTION

Machine learning (ML) is gaining rapid popularity in the natural sciences due to its ability to identify patterns in large and complex datasets (Valentine and Kalnins 2016, Bergen et al. 2019, Reichstein et al. 2019). ML corresponds to pattern recognition that self-improves through experience (Michie 1968), and can be broadly categorized into clustering and prediction. Clustering, such as hierarchical clustering, identifies groups of similar patterns, or clusters, and assigns a cluster membership, or label, to each observation. For example, in watershed sciences, including geomorphology and hydrology, ML has been used to cluster patterns of channel form and hydrologic response at regional, continental, and global scales (e.g., Lane et al. 2017a, Gaucherel et al. 2017, Wolfe et al. 2019, Henshaw et al. 2019, Sergeant et al. 2020, Merritt et al. 2021, Lane et al. 2017b, McManamay et al. 2018, Byrne et al. 2020b, Clubb et al. 2019, Dallaire et al. 2019, Walley et al. 2020, Rabanaque et al. 2021). Alternatively, ML prediction leverages a model, such as a decision tree, to approximate the relationship between input data and pre-labeled output. For example, groups of channel forms were predicted by Flores et al. (2006) and Beechie and Imaki (2014) using prior labels from Montgomery and Buffington (1997) and Leopold and Wolman (1957), respectively.

Combining ML clustering and prediction is valuable when relevant information is unevenly distributed throughout the

studied system. For example, in river networks, locations with surveyed geometry and measured discharge are data-rich for predicting sediment transport compared to ungauged, unsurveyed locations where only coarser-scale remote sensing data is available. This disparity between data-rich and data-poor locations is an information gap that has been addressed by two main approaches: predict-first and cluster-first.

The predict-first approach predicts missing information at data-poor locations based on data-rich locations then clusters the resulting enhanced dataset to identify large-scale patterns. For example, McManamay et al. (2018) used ML to predict prior hydrologic classes, sediment size and stream temperature. The enhanced dataset of stream temperature was then clustered and combined with other binned datasets to explore the diversity in stream habitats throughout the entire network. In another example, Rabanaque et al. (2021) identified fluvial landforms from orthophotographs, predicted their occurrence throughout the network from coarser-scale satellite imagery, then used that information to extract geomorphic variables for clustering. In contrast, the cluster-first approach clusters at data-rich locations then predicts the identified patterns throughout the network, including at data-poor locations.

The cluster-first approach leverages local domain knowledge, which is crucial when facing ground-truthing requirements for stakeholder buy-in (Abrahart et al. 2012). For example, Byrne et al. (2020b) clustered field-measured channel attributes then the resulting clusters were predicted throughout the network by Guillon et al. (2020). In another example, Merritt et al. (2021) used a three-step approach: predicting daily streamflow at gauges with incomplete records from those with complete records, clustering the enhanced dataset, and then predicting the resulting clusters throughout the network. Regardless of the approach, obtaining evenly-distributed information throughout a river network has numerous eco-geomorphic applications such as automating the river segmentation problem described by Nardini et al. (2020).

Despite the growing use of combining ML clustering and predictions in watershed sciences, challenges and limitations remain poorly documented. The work of Peñas et al. (2014) in hydrology and Kasprak et al. (2016) in geomorphology are notable exceptions. Peñas et al. (2014) found weak agreement between clusters from cluster-first and predict-first approaches. This incongruence can significantly impact management decisions. Similarly, Kasprak et al. (2016) reported loosely comparable clusters when using empirical approaches for identifying channel forms in a single watershed.

Further, the predict-first approach is only tractable if relevant information can be reliably predicted. When information is evenly distributed (i.e. no information gap), the problem is straightforward, and a single-step approach can be used (e.g., Walley et al. 2020). However, if data is sparse, the information gap between data-rich and data-poor locations is significant, or both, predicting first may not be effective. For example, channel forms defined from surveyed topography cannot be reliably predicted from coarse-scale data, but clusters of these forms can be. Here clustering essentially compresses information by grouping similar patterns, reducing the information gap and making the predictions tractable. Importantly, clustering at data-rich locations results in clusters that are trivially separable using clustering data but unlikely to be equally separable using predictions data.

Having discussed how cluster-first and predict-first approaches differ, we hereafter focus on the predict-first approach and turn to another aspect rarely addressed in the watershed sciences literature: the impact of differing information processing on the performance of ML predictions. Information processing refers to the specific algorithmic steps applied to transform data and recognize patterns. In particular, the information processing is meaningfully different between the two current main ML predictions approaches: decision trees and deep learning.

A decision tree (DT) sequentially thresholds an input variable, or predictor, in a tree-like structure that segments the predictor space. The resulting subspaces correspond as exclusively as possible to a unique label, a characteristic called purity. At each split of a DT, the predictor maximizing purity is chosen by an information selection process based either on the Gini coefficient (Gini 1936) or an information theory measure (e.g., entropy, see Appendix A). While other variations of DTs are gaining traction (e.g., Chen and Guestrin 2016, Ke et al. 2017), the most popular remains random forest (RF, Breiman et al. 1984). In RF, the forest is an ensemble of DTs with each one repeating the information selection process on a random subset of predictors. Repeating this process while subsampling observations leads to (mostly) uncorrelated trees, making the ensemble decision process robust to noise, resistant to outliers, and allowing generalization. The popularity of RF is further explained by its conceptual simplicity, explainability, and speed. DTs have been used in geosciences for over half a century (Newendorp 1976) and some recent applications with RF include predicting sediment transport (Bhattacharya et al. 2007), global seafloor sediment porosity (Martin et al. 2015), sediment rating curves (Vaughan et al. 2017), stream habitats (McManamay et al. 2018), and mapping fluvial landforms (Rabanaque et al. 2021), subaerial fluvial sediment facies (Gómez et al. 2022), and channel forms (Guillon et al. 2020).

Deep learning, the second main approach for ML predictions, repeats and stacks the basic structure of an artificial neural network, and is widely used in complex engineering applications with image, video, text, and time series data (LeCun et al. 2015). While deep learning was first applied in geosciences decades ago (Zhao and Mendel 1988, Dowla et al.

1990), its applications are recently reemerging in a variety of areas, including predicting discharge (e.g., Kratzert et al. 2019, Worland et al. 2019, Tennant et al. 2020), rainfall (Pan et al. 2019, Gauch et al. 2021, Adewoyin et al. 2021), landslide susceptibility (Ermini et al. 2005), river width (e.g., Ling et al. 2019), forecasting (e.g., Fleming et al. 2015) or reconstructing floods (e.g., Bomers et al. 2019), detecting sediment grain size (Chen et al. 2022), and mapping topographic features (Valentine et al. 2013), drainage networks (Mao et al. 2021), riverscape (Alfredsen et al. 2022), and riverbed sediment size (Marchetti et al. 2022). An artificial neural network consists of a succession of layers of connected neurons. Each neuron holds a weight describing its connection to neurons in the next layer and some form of activation functionally combining inputs from neurons in the previous layer. The first, last, and intermediate layers correspond to input, output, and hidden layers, respectively.

A deep neural network (DNN) has more than one hidden layer and numerous architectures exist to arrange the hidden layers and their connections (see Shen (2018) for examples in hydrologic sciences). Importantly, DNNs learn distributed representations: different patterns of neural activity correspond to different labels, but each neuron and its connections store patterns for many non-neighboring observations (Hinton 1984, Bengio et al. 2013). Such distributed representations efficiently capture irregular patterns stemming from multiple, interacting, hierarchical inputs (Bengio et al. 2013). Specifically, the sequential processing of data through the hidden layers hierarchically discovers meaningful abstractions by optimally decoupling dependent inputs, extracting relevant information from noise and compressing it for generalization (Lin et al. 2017, Tishby and Zaslavsky 2015, Shwartz-Ziv and Tishby 2017, Bény 2013, Mehta and Schwab 2014, Li and Wang 2018, Koch-Janusz and Ringel 2018, Yang et al. 2023). For example, DNNs have been shown to sequentially learn meaningful visual (Bau et al. 2020) or mathematical (Amey et al. 2021) representations leading to increasing performance through the successive layers (Schilling et al. 2021, Erdmenger et al. 2021, Cao et al. 2022, Fischer et al. 2022, Yang et al. 2023, Hartmann et al. 2021). In contrast, sequential information compression is absent from DTs, which learn global patterns for neighboring observations; an efficient approach when observations are limited and with tabular data (i.e. a table with columns and rows for variables and observations, respectively). In watershed sciences, where limited, tabular data is common, it remains unclear whether RF or DNN perform better.

We now consider the differing information processing of RF and DNN in the context of using a cluster-first approach to address an information gap. We ask the following three research questions:

Q1 How does clustering at data-rich locations impact subsequent predictions at data-poor locations?

Q2 In such case, how does the difference in information processing between RF and DNN relate to their respective statistical learning performance?

Q3 As some clusters may be more adequately described by prediction data than others, the information gap is cluster-dependent: for which cluster(s) is the information gap the largest?

We hypothesize that evaluating the information gap between clustering data and prediction data is central to answering these questions and that the difference in information processing between DNN and RF prognosticates their performance. To answer these questions, we leverage nine regional examples of clustering and predicting river channel forms, stemming from a single clustering methodology (Byrne et al. 2020b) applied in California, USA. We characterize each regional set of clusters of channel form by the information present in field-measured channel attributes and by the performance of subsequent DNN and RF predictive models using geospatial predictors. Outcomes from this analysis yield general implications for the sampling strategies at the core of ML predictions in watershed sciences and associated stratified management decisions.

2 | METHODS

We formalize hereafter this study's experimental design, describe the prior clustering of California channel forms, and explain our approach for evaluating the information gap between clustering and predictions data, assessing the performance of ML predictive models, and investigating the linkages between clustering, information gap, and ML performance.

2.1 | Experimental Design

This study's experimental design focuses on the interaction between clustering data and prediction data (Fig. 1a), setting it apart from related works by the authors (Byrne et al. 2020b, Lane and Byrne 2021, Guillon et al. 2020, Lane et al. 2021). Byrne et al. (2020a) used a stratified sampling process on stratification data $\mathbf{x_s}$ to obtain field sampling data $\mathbf{x_f}$ and clustering $\mathcal{M}_{\text{cl}}$ was used to estimate the joint probability distribution $p(\mathbf{x}, \mathbf{y})$ between clustering data $\mathbf{x}$ and labels $\mathbf{y}$:

$$\mathcal{M}_{\text{cl}} : \mathbf{x} \xrightarrow{p(\mathbf{x},\mathbf{y})} \mathbf{y} \tag{1}$$

Labels $\mathbf{y}$ take discrete values over a set of clusters to indicate membership. For example, with clusters $\{c_1, c_2\}$, the first

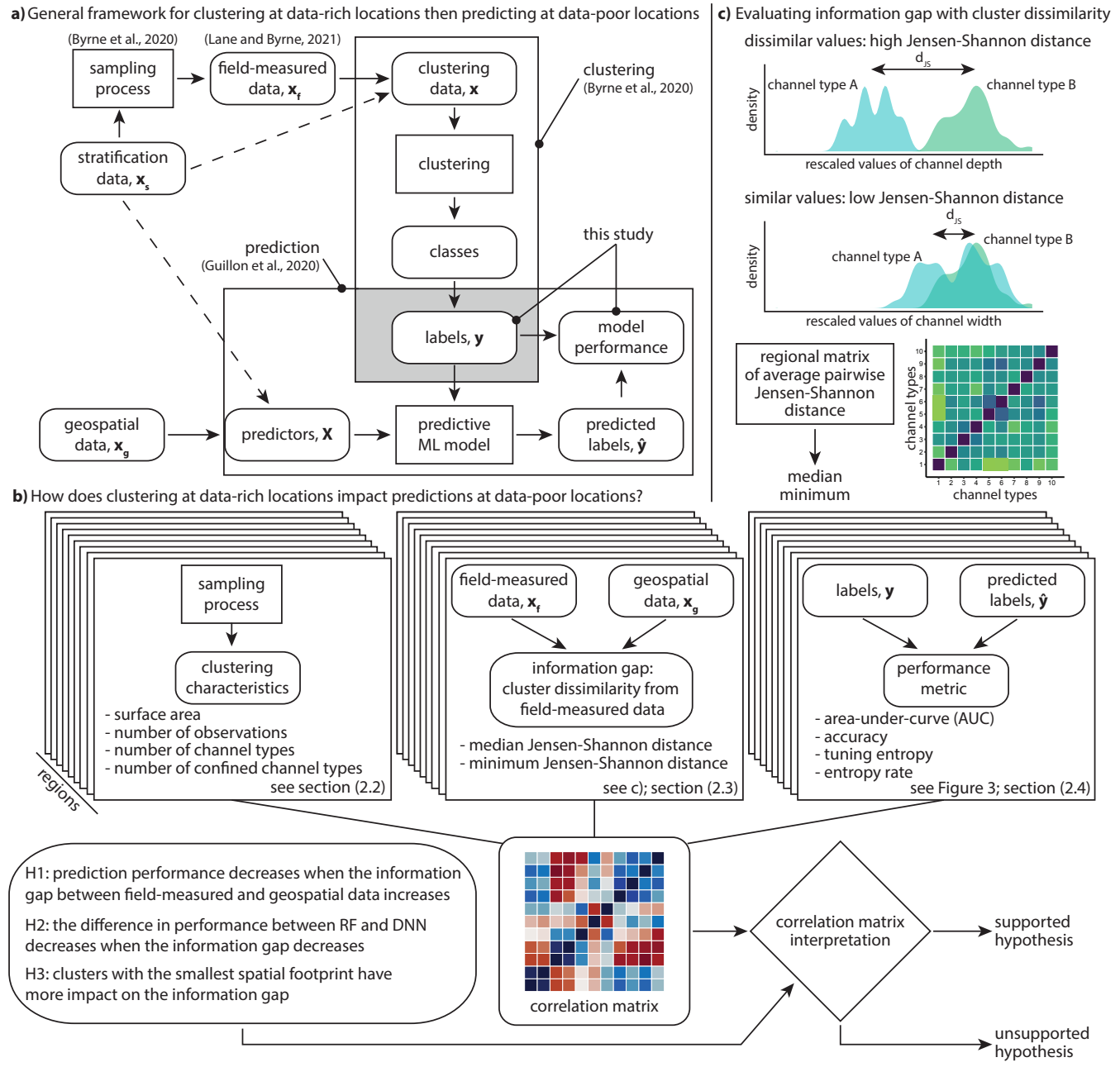

FIGURE 1 Experimental design

five labels could be: $\{c_1, c_2, c_2, c_1, c_1\}$. This means that the second and third observations belong to c_2, whereas the others belong to c_1. Guillon et al. (2020) used geospatial data $\mathbf{x_g}$ and a predictive model $\mathcal{M}_{\text{pred}}$, was used to estimate the probability distribution of the labels $\mathbf{y}$ given the value taken by predictors $\mathbf{X}$, $p(\mathbf{y} \mid \mathbf{X})$:

$$\mathcal{M}_{\text{pred}} : \mathbf{X} \xrightarrow{p(\mathbf{y} \mid \mathbf{X})} \mathbf{y} \tag{2}$$

However, when labels used for predicting stem from prior clustering, predicting is conditioned on the clustering and $\mathcal{M}_{\text{pred}}$ follows the conditional probability distribution $p(\mathbf{y} \mid \mathbf{X}, p(\mathbf{x}, \mathbf{y}))$. Consequently, predicted labels depend on both the clustering $\mathcal{M}_{\text{cl}}$ and predictive models $\mathcal{M}_{\text{pred}}$, as does any performance metrics of $\mathcal{M}_{\text{pred}}$ comparing observed and predicted labels.

Byrne et al. (2020a) combined field sampling and stratification data to obtain clustering data ($\mathbf{x} = \{\mathbf{x_f}, \mathbf{x_s}\}$) and Guillon et al. (2020) combined geospatial and stratification data to obtain predictors ($\mathbf{X} = \{\mathbf{x_g}, \mathbf{x_s}\}$). Thus, the explicit form of the

predictive model takes into account both clustering (1) and predictor data (2):

$$\mathcal{M}_{\text{pred}} : \mathbf{X} \xrightarrow{p(\mathbf{y} \mid \mathbf{x_g}, \mathbf{x_s}, p(\mathbf{x_s}, \mathbf{x_f}, \mathbf{y}))} \mathbf{y} \quad (3)$$

Catchment geospatial data (included in **X**) typically has 10-30 m resolution, which is notably coarser than field-measured data (included in **x**) with 0.1-1.0 m resolution. Therefore, an information gap exists between clustering data and prediction data. Further, if spatial resolution drives the difference in relevant information, clusters with the smallest spatial footprint are more likely be inadequately described by coarser-scale data and would therefore be disproportionately impacted. The more dissimilar the clusters are with respect to field-measured data, the smaller the information gap is. For example, coarse-scale data likely contains relevant information to discriminate between a river meandering in a wide valley and a mountain stream in narrow valley, but may not contain relevant information to discriminate between two types of mountain streams both occuring in narrow valleys.

Taking the above into consideration, we pose three hypotheses linked to our research questions:

H1 Prediction performance decreases when the information gap between field-measured and geospatial data increases;

H2 The difference in performance between RF and DNN decreases when the information gap between field-measured and geospatial data decreases;

H3 Clusters with the smallest spatial footprint, here valley-confined channels, have a disproportionate impact on the information gap between field-measured and geospatial data.

A correlation analysis was conducted to test these hypotheses across nine regions by examining three groups of variables detailed hereafter: clustering characteristics, information gap, and prediction performance (Fig. 1b).

2.2 | Prior Clustering of California Channel Forms

As a scientific testbed, we used nine independent sets of clusters of channel forms established for nine government-defined water management regions in California (USA) (SWRCB 2019). Each cluster groups observations with similar channel forms therefore defining a channel type. Here, a channel type refers to a stream channel interval with relatively uniform characteristics over lengths of 10-20 channel widths. This scale is useful in relating channel morphology to watershed and channel processes, as well as habitat characteristics (Montgomery and Buffington 1997, Byrne et al. 2020a). California faces many natural resource challenges (e.g., Lane et al. 2018) and has diverse physiographic features (Mount 1995). The nine study regions vary in terms of size, hydro-climate, physiography, and geology (Fig. 2, Table 1). The adjacent Sacramento (SAC) and San-Joaquin-Tulare (SJT) regions mainly pertain to their namesake river basins, spanning the alluvial Central Valley and the granitic Sierra Nevada mountain range. The northern, wetter SAC region includes the volcanic Modoc Plateau. The South Coast (SC), South Central Coast (SCC), North Central Coast (NCC), South Fork Eel (SFE) and North Coast (NC) regions are all coastal, winter rain-dominated regions, transitioning from drier to wetter climates from south to north. These regions correspond to the Southern and Northern Coast Ranges along the San Andreas Fault. The Klamath (K) region is the wettest region and marks the southern extent of the Cascade Range. Conversely, the South East (SECA) region is overall the driest region and marks the inception of the Basin and Range Province. The regions vary in area from 1,785 km^2 for SFE to 107,622 km^2 for SECA (Table 1), spanning two orders of magnitude.

Previous studies had varying sampling densities across different regions due to financial and logistical constraints, as well as the scarcity of unaltered examples for certain types of channels (see Table 2). For instance, small, low-order, valley-unconfined streams in mountain meadows or valley floors are often altered by land and water management activities. Each region had between 63 to 290 observations, with 5 to 10 channel types per region. To capture natural variability, field sampling locations were selected based on stratified random sampling using four GIS-derived variables: 10-m digital elevation model channel slope, valley confinement, drainage area and sediment supply, as detailed in Byrne et al. (2020b), Lane and Byrne (2021). Valley confinement was calculated as the perpendicular distance between the stream interval centerline and valley walls on both sides of the interval (Byrne et al. 2020b). For every 200-m stream interval in California, valley confinement was averaged over four cross-sections. Sediment supply was

Region ID	Geographical region	Observations	Channel types	Area (km^2)
K	Klamath	105	7 (3)	27,747
NC	North Coast	201	8 (6)	12,504
NCC	North Central Coast	103	6 (4)	13,263
SAC	Sacramento Basin	290	10 (4)	70,130
SC	South Coast	67	5 (2)	36,982
SCC	South Central Coast	119	8 (3)	26,595
SECA	South East California	63	5 (2)	107,622
SFE	South Fork Eel	96	7 (5)	1,785
SJT	San-Joaquin-Tulare	65	6 (4)	83,498

TABLE 1 Regional clustering characteristics. The number of confined channel types is reported between parenthesis.

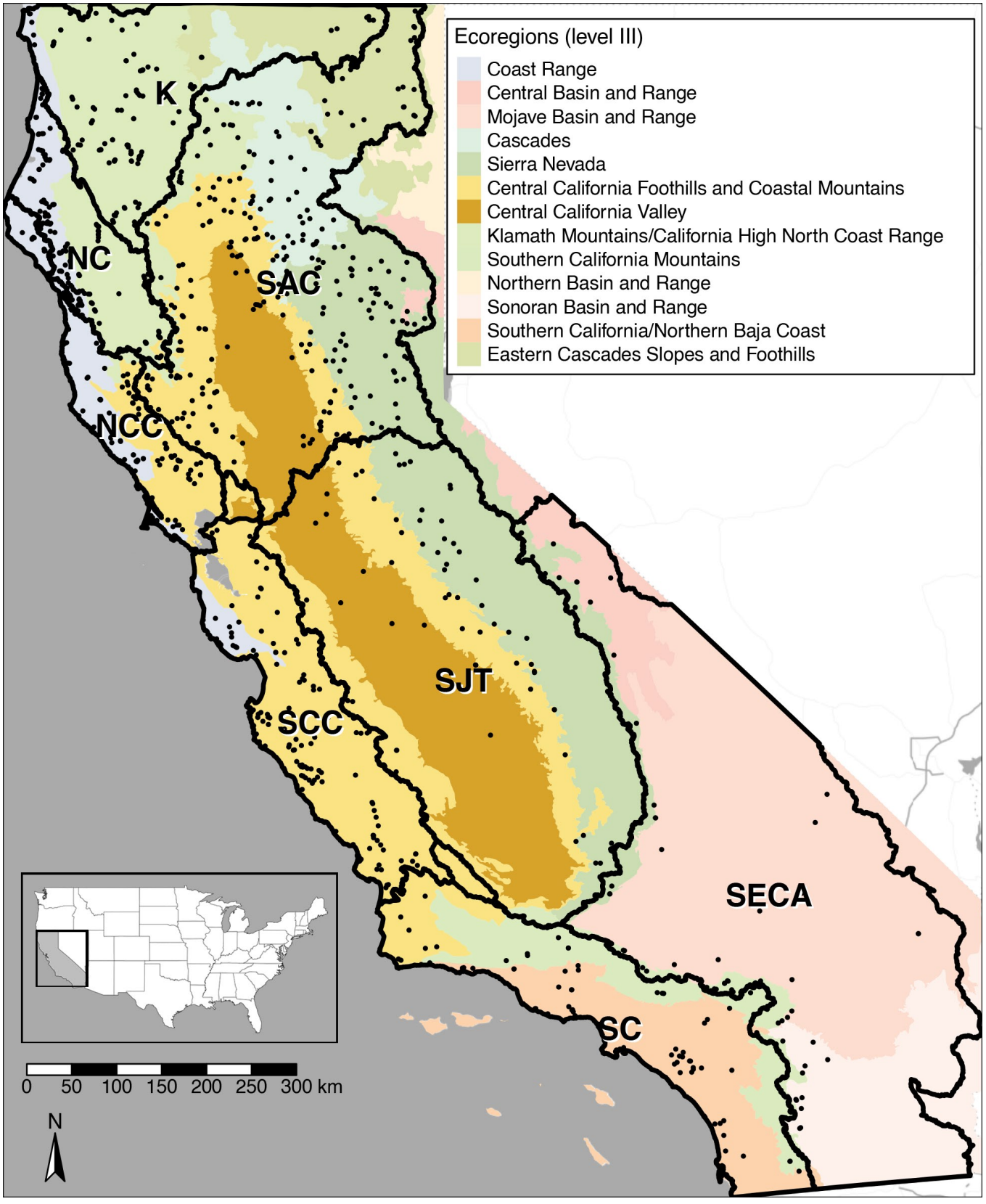

FIGURE 2 Field sites location in California (USA) across nine distinct regions (black dots). Ecoregions are displayed as a proxy combining geology, soils, vegetation, climate, and hydrology Omernik and Griffith (2014). Inset shows the general location. K: Klamath; NC: North Coast ; NCC: North Central Coast; SAC: Sacramento Basin ; SC: South Coast ; SCC: South Central Coast; SECA: South East California; SFE: South Fork Eel; SJT: San-Joaquin-Tulare.

	Region								
Channel type	K	NC	NCC	SAC	SCC	SC	SECA	SFE	SJT
1	18	8	23	6	9	9	14	12	5
2	4	32	21	27	7	8	8	4	19
3	1	17	9	36	21	6	8	12	6
4	5	14	21	33	27	23	19	28	9
5	14	5	24	43	18	21	14	30	4
6	16	28	24	45	8	–	–	4	22
7	47	–	36	33	16	–	–	6	–
8	–	–	43	24	13	–	–	–	–
9	–	–	–	27	–	–	–	–	–
10	–	–	–	16	–	–	–	–	–
Total	105	104	201	290	119	67	63	96	65
Min	1	5	9	6	7	6	8	4	4
St. Dev.	15.56	10.76	10.27	11.85	7.00	7.96	4.67	10.98	7.73

TABLE 2 Distribution of the number of observations across all regions.

estimated using the revised universal soil loss equation (Renard et al. 1997) using statewide datasets for rainfall-runoff erosivity, soil erodibility, slope characteristics, and land cover (SWRCB 2017).

Prior to the present study, for each region, channel forms were clustered into channel types using the analytical methodology of Lane et al. (2017b), which was updated by Byrne et al. (2020b) and reported in Byrne et al. (2019), Guillon et al. (2019), Byrne et al. (2020a) (Fig. 1a). This approach relied on hierarchical clustering based on field-measured channel attributes, including bankfull channel width and depth, width and depth variability, grain size metrics, and channel slope. In addition, two stratification data variables, valley confinement and catchment area, were included. All measurements of channel attributes were made using the same standardized procedure by field technicians trained together to yield consistent and reliable results (Lane and Byrne 2021). The resulting clusters correspond to the channel types and were named *a posteriori* using expert-knowledge in terms of valley confinement, bed morphology, and sediment size. In particular, streams were categorized as located in confined, partly-confined, and unconfined valleys based on valley confinement distances of < 100, $100 - 1,000$, and $> 1,000$ m, respectively. When analyzing channel types obtained using the same methodology but over all of California, Lane et al. (2021) found that the resulting channel types covered and even exceeded the entire range of channel types considered by Montgomery and Buffington (1997) (see Fig. 4b in Lane et al. (2021)). A total of 1,110 sampling sites were included across all regions, resulting in nine regional sets of channel types (Fig. 2, Table 2).

2.3 | Evaluating the Information Gap between Field-measured and Geospatial Data

We evaluated the information gap between field-measured and geospatial data by computing cluster dissimilarity with respect to field-measured data (see Fig. 1b;c). A dissimilarity measure is a statistical distance that quantifies how different two objects are. In our case, these objects are clusters resulting from clustering in each region, and the dissimilarity measure is the average Jensen-Shannon distance, $\bar{d}_{JS}$, with respect to the distributions of seven field-measured attributes: bankfull depth, bankfull width, bankfull width-to-depth ratio, coefficients of variation for width and depth, and the 50th and 84th percentiles of grain size (D50, D84). The Jensen-Shannon distance (A6) is a symmetric measure of discrimination between two probability distribution functions, and it is a proper distance metric for constructing distance matrices (Lin 1991, Topsoe 2000, Endres and Schindelin 2003). The Jensen-Shannon distance derives from the Jensen-Shannon divergence (A5) (Lin 1991, Topsoe 2000), which is a symmetric version of the Kullback-Leibler divergence (A4) (Kullback and Leibler 1951). Appendix A provides links between the Jensen-Shannon distance and other information theory metrics. Channel types with high $\bar{d}_{JS}$ are defined from more dissimilar underlying information and require, on average, less information to discriminate between them (see Fig. 1c). Conversely, channel types with low $\bar{d}_{JS}$ are defined from more similar underlying information and require, on average, more information to discriminate between them (see Fig. 1c). For each region, we constructed a Jensen-Shannon distance matrix between each possible pair of channel types (see Fig. 1b). To compare regions, regional matrices of average Jensen-Shannon distance across field attributes $\bar{d}_{JS}$ were summarized by their median, mean and minimum values. We also calculated the same metrics for valley-confined channel types.

2.4 | Assessing Performance of Machine Learning Models

Our approach to training and evaluating the performance of ML models buils on the work of Guillon et al. (2020) with two important modifications (Fig. 1a;3). First, we selected predictors (Table 3) using mutual information prior to statistical learning. Second, we evaluated performance of statistical learning using nested resampling. By training models with an increasing number of predictors, these changes allowed us to evaluate tuning stability (see below), and select an optimal number of predictors to fairly compare DNN and RF. For each region, we tuned DNN and RF models with nested resampling and compared them to three baseline models trained with default hyperparameters: featureless (a mean model always predicting the most frequent label), naive Bayes (a model using Bayes rule with strong independence assumptions (Laplace 1820)), and *k*-nearest-neighbor (a dissimilarity-based model (Cover and Hart 1967)).

2.4.1 | Selecting Predictors

To prevent over-fitting and promote robust generalization, we selected predictors before their use in RF and DNN models. In predicting fine-scale channel types from coarser-scale geospatial predictors, Guillon et al. (2020) utilized data complexity measures to select groups of predictors. For example, data sparsity, which divides the number of observations by the number of predictors, is used to assess problem complexity (Lorena et al. 2018). In contrast, we selected individual predictors using a mutual information-based method (Guyon and Elisseeff 2003, Bommert et al. 2020), maximizing the relevance of predictors for identifying channel types based on their statistical relationship with the predictors (Fig. 3, step 2; (A3)). However, as this selection is based on field-measured data at sampling site locations, it may be biased according to the observed distribution of channel types (Table 2). To address this and derive a more robust selection, we averaged predictor selection over 500 iterations for each region, each using 80% of the training data in a stratified subsampling scheme. This process selected predictors with the highest degree of statistical dependence with respect to channel type distributions for a given region.

When selecting predictors, mutual information is algorithm-agnostic and maximizes predictor relevance. However, it does not consider redundancy, which may impact RF and DNN differently. Although only perfectly correlated variables are truly redundant, creating new predictors from highly correlated but complementary predictors may improve performance (Guyon and Elisseeff 2003). Despite this, RF and DNN have distinct information processing and may be impacted differently. RF's ensemble decision process implicitly combines predictors and robustly removes irrelevant ones. In contrast, DNN uses multiple hidden layers of neurons to combine input, which can act as latent predictors. Thus, removing highly correlated but complementary predictors may negatively impact DNN performance by hindering the discovery of relevant latent predictors. To reduce near-perfect redundancy without substantially impacting DNN performance, we removed predictors with an absolute pairwise correlation greater than 0.95 prior to mutual information selection. The removed predictor was the one with the largest average absolute correlation across all predictors. In total, we ran 49 models with a number of selected predictors ranging from 2 to 50.

2.4.2 | Measuring Performance

We evaluated the performance of machine learning models across the nine regions of study by benchmarking them using the area under the receiver operating characteristic curve (AUC) and hyperparameter tuning entropy. Prior to learning, we balanced the observations using the synthetic minority oversampling technique (Chawla et al. 2002). We also checked the input data for no-variance predictors, centered and scaled them, and imputed missing values with median imputation.

While the parameters of ML models are adjusted during training, performance depends on hyperparameters that define model architecture or algorithmic behavior. Hyperparameter selection, or tuning, involves training models for different sets of hyperparameters and comparing their resulting performance. For RF, we tuned optimum number of predictors available for splitting by performing a discrete search among 16 regularly-spaced, discrete values between 2 and the number of predictors. For DNN, we tuned seven hyperparameters by choosing them from the following sets of common values: number of hidden layers $\in \{2, 3, 4, 5\}$, number of neurons in each hidden layers $\in \{5, 10, 20, 30, 50, 100, 200\}$, learning rate, controlling the size of the steps in gradient descent optimization, $\in \{0.5, 0.1, 0.05, 0.01, 0.005\}$, batch size, controlling the size of the subset of data used to update one gradient step $\in \{16, 32, 64\}$, momentum, inertially controlling the influence of previous step on the current update, $\in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, hidden and visible layers dropouts, acting as regularization (i.e. limiting over-fitting) by de-activating some neurons, $\in \{0, 0.1, 0.2\}$. We performed a 100-iterations random discrete search for tuning the DNN's hyperparameters, as a strictly discrete search is prohibitively expensive. We trained the DNN for 20 epochs, or cycles through the training dataset, and with a batch number between 120 and 560, depending on the size of the training dataset for each region of study.

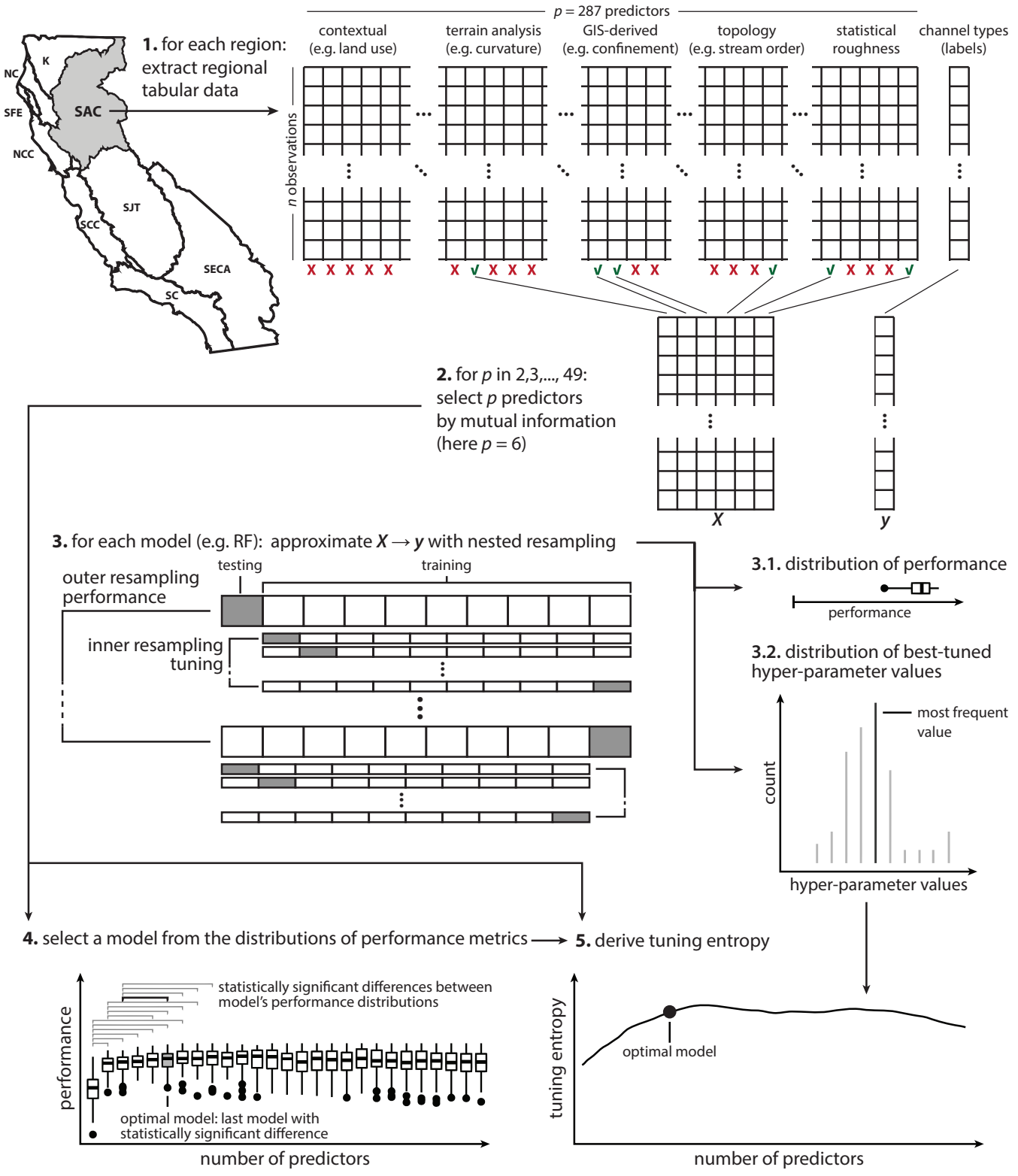

FIGURE 3 Schematic of the machine-learning framework. The complete list of predictors is provided in Table 3. RF: Random Forest.

To estimate tuning robustness and limit over-fitting, we used nested resampling which separates model evaluation from tuning (calibration) by using two nested loops: an outer loop for model evaluation and an inner loop for model tuning (Bischl et al. 2012, Fig. 3, step 3). We chose 10 repeats of 10-fold stratified cross-validation as the outer resampling, and 10-fold

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. (2020b)
	Sediment supply	-	Haan et al. (1994)	Renard et al. (1997)
Network topology (4)	Drainage area	-	McKay et al. (2012)	Hill et al. (2015)
	Strahlers 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)

TAM-DM : Terrain Analysis Metrics - Distribution Metrics

TABLE 3 Predictors Used in the Machine Learning Framework. The 10-m National Elevation Data Set (Gesch et al. 2002, NED) and the Stream-Catchment Data Set (StreamCat; Hill et al. 2015) 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 (McKay et al. 2012, NHDPlusV2) is publicly available on both platforms. TAM-DM: Terrain Analysis Metrics-Distribution Metrics (see Guillon et al. (2020)).

stratified cross-validation as the inner resampling. Specifically, the training dataset was randomly split into 10 subsets or folds with equally-distributed classes between each fold. The model was trained on nine folds and tested on the one remaining fold. This outer process was repeated 10 times, with each fold serving as the testing data once. During model training, the data of the nine outer training folds was randomly split into 10 new folds with equally-distributed classes. Hyperparameter tuning was performed on these nine inner folds and tested on the one remaining inner fold. This inner process was repeated 10 times, with each fold serving as the testing data once. The hyperparameter values that maximize model performance across the data of the 10 inner folds, i.e., the data of the nine outer training folds, were selected. These best-performing hyperparameters were then used to train the model on the nine outer folds and test it on the one remaining outer fold. The outer fold selection and thus the entire process was repeated 10 times. Nested resampling provides a distribution of best-tuned hyperparameters for each iteration of the outer resampling in addition to the distribution of model performance obtained by traditional resampling (Fig. 3, step 3.2). Here, 10 repeats of outer 10-fold cross-validation produced 100 values of best-tuned hyperparameters. However, nested resampling has a high computational cost: together with 49 runs for predictor selection, these benchmark parameters resulted in training a total of 51,534,000 tuned models and 132,300 baseline models (5,726,000 per region).

Model performance was evaluated using AUC and hyperparameter tuning entropy, which was calculated from the distribution of the best-tuned hyperparameters (Fig. 3, step 5). AUC measures the ability of a model to distinguish between positive and negative observations, striking a balance between maximizing true positives and minimizing false positives (Rosset 2004). In contrast, accuracy maximizes both true positives and true negatives. We chose to optimize for AUC instead of accuracy due to its higher discrimination performance, its relation to dissimilarity, and its suitability for limited datasets (Rosset 2004, Huang and Ling 2005, Ferri et al. 2009). Hyperparameter tuning entropy was calculated using Shannon's entropy (Equation A1) by considering the probability of a given hyperparameter value being selected as best-tuned and represents the uncertainty in selecting optimal values for hyperparameters with more uncertain hyperparameter selection leading to higher tuning entropy. To account for the different range of possible values, each tuning entropy was normalized by the maximum possible tuning entropy. For example, if only three out of 16 possible values for the number of predictors

available for splitting were reported from the nested resampling best-tuned hyperparameters for RF, the resulting tuning entropy was normalized by $\log_2(16)$. For DNN, the reported tuning entropy is an average of the tuning entropies of its seven hyperparameters.

For each region, we selected the optimal model based on the statistical differences between AUC distributions for different numbers of predictors (Fig. 3, step 4). The selection was performed in a sliding window of seven models, with one model with $\mathcal{M}(n)$ with n predictors being compared to the following models: $\{\mathcal{M}(n+1) \ldots \mathcal{M}(n+6)\}$. We considered the optimal model to be the last model exhibiting a significant statistical difference between its performance distributions. We performed a statistical comparison using Dunn's test, with a Bonferroni correction of the p-value to account for multiple comparisons. We considered a difference to be significant if the test p-value was less than $0.05/7 \simeq 0.007$.

As in Guillon et al. (2020), we assessed the performance of predictive modeling for each region at the network scale using entropy rate. This measure leverages the network structure of the predictions by estimating the stability of the predictions from the transition probabilities between each channel type. Entropy rate prognosticates the prediction skill of a model (Stephenson and Dolas-Reyes 2000, Roulston and Smith 2002), and helps select models that provide the best information (Daley and Vere-Jones 2004, Nearing and Gupta 2015). We computed both metrics from predictions after a cross-validated multinomial calibration that corrects the potential distortion of posterior probabilities and improves model performance (DeGroot and Fienberg 1983, Zadrozny 2002, Niculescu-Mizil and Caruana 2005).

2.5 | Correlation Analysis

Finally, a correlation analysis was performed to investigate potential linkages between the regional clustering characteristics, information gap estimated by cluster dissimilarity, and measures of machine learning (ML) model performance, as described in the previous subsections (Figure 1b). The clustering characteristics included the number of observations, study area, observation density, number of channel types, and number of valley-confined channel types. Since channel types in confined valleys were expected to be the most difficult clusters to predict due to their narrower spatial footprint being imperfectly captured by large scale geospatial predictors (H3; Guillon et al. 2020), they were considered independently. For each region, the matrix of average pairwise $\bar{d}_{JS}$ was summarized by the median and minimum values over all channel types and valley-confined channel types only, respectively. The measures of ML models performance included: AUC, accuracy, hyperparameter tuning entropy, entropy rate and the difference in performance between RF and DNN. Both Pearson and Spearman correlations were computed on scaled data and yielded similar results. Due to the limited dataset (n = 9), we present average results from 500 correlations performed with 80% subsampling.

Region ID	Median $\bar{d}_{JS}$	Mean $\bar{d}_{JS}$	Minimum $\bar{d}_{JS}$
K	0.54 (0.45)	0.57 (0.44)	0.38 (0.38)
NC	0.47 (0.45)	0.48 (0.43)	0.34 (0.34)
NCC	0.52 (0.52)	0.52 (0.51)	0.33 (0.35)
SAC	0.47 (0.44)	0.49 (0.43)	0.27 (0.34)
SC	0.53 (0.52)	0.53 (0.52)	0.37 (0.52)
SCC	0.51 (0.45)	0.50 (0.45)	0.33 (0.39)
SECA	0.54 (0.62)	0.53 (0.62)	0.37 (0.62)
SFE	0.61 (0.58)	0.59 (0.58)	0.34 (0.42)
SJT	0.62 (0.62)	0.61 (0.60)	0.40 (0.44)

TABLE 4 Cluster dissimilarity with respect to field-measured data estimated by the Jensen-Shannon distance ($\bar{d}_{JS}$). Values for valley-confined channel types are reported between parenthesis.

3 | RESULTS

In the following subsections, we present the results regarding: (i) the information gap as estimated by cluster dissimilarity with respect to field-measured data; (ii) the performance of the ML models; (iii) channel type predictions and river segmentation throughout California and (iv) the correlation analysis between regional clustering characteristics, cluster dissimilarity, and prediction performance.

3.1 | Cluster Dissimilarity with Respect to Field-measured Data

Cluster dissimilarity with respect to field-measured data varied between the different regional clusterings of channel forms, and was estimated with the Jensen-Shannon distance. The Jensen-Shannon distance matrices for all nine regions are summarized in Table 4, and an example of the derivation of Jensen-Shannon distance is provided for SAC, the only region with ten channel types, in Figure 4. The two regions with the highest cluster dissimilarity were SFE and SJT, while the two regions with the lowest cluster dissimilarity were NC and SAC. In seven regions, the minimum $\bar{d}_{JS}$ was not between two confined channel types. Nonetheless, in eight out of nine regions, $\bar{d}_{JS}$ decreased when considering confined channel types, indicating that these channel types are less dissimilar with respect to field-measured data. In the odd region, SECA, there are only two valley-confined channel types and thus only one value of

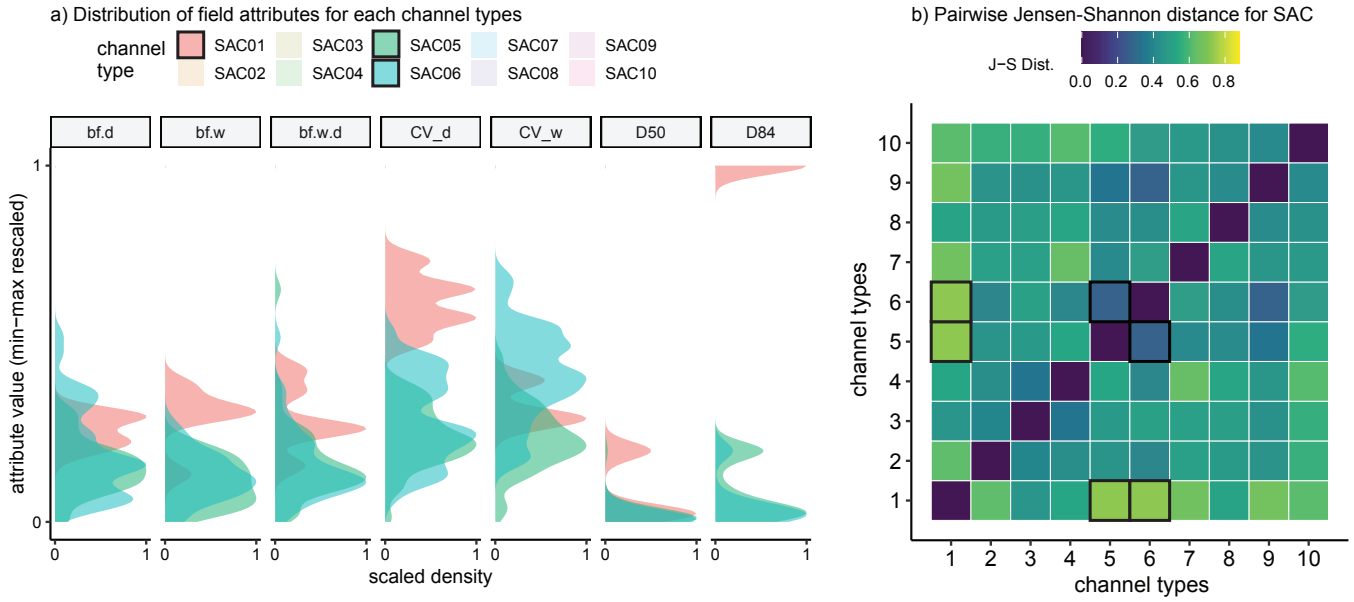

FIGURE 4 Example of Jensen-Shannon distance derivation. For clarity, only the distributions corresponding to the channel types SAC01, SAC05 and SAC06 are shown on pannel a). The corresponding Jensen-Shannon distances are highlighted by black squares on pannel b). SAC: Sacramento region; bf.d: bankfull depth; bf.w: bankfull width; bf.w.d: bankfull width to depth ratio; CV_d: coefficient of variation of bankfull depth; CV_w: coefficient of variation of bankfull width; D50: 50th percentile of grain size distribution; D84: 84th percentile of grain size distribution.

Model	Predictors	AUC	Accuracy	Training time	Normalized Tuning Entropy
DNN	30	0.922	0.663	2450	0.792
RF	18	0.949	0.749	54	0.757

TABLE 5 Summary table of the average model performance across all areas of study. Training time is given here in seconds for one iteration of the learning process and does not correspond to the total CPU-hours required for training. DNN: Deep Neural Network; RF: Random Forest.

pairwise $\bar{d}_{JS}$ leading to an equal median, mean, and minimum $\bar{d}_{JS}$ (Table 4).

3.2 | Performance in Statistical Learning and Predictive Modelling

RF outperformed DNN in terms of AUC even as the number of predictors increased (Fig. 5). The performance of all models improved with additional predictors, but RF consistently exhibited a greater and more rapid increase in performance. In certain regions (NC, SC, and SFE), the performance of the naive Bayes model decreased with additional predictors, indicating the progressive inclusion of irrelevant or noisy predictors in the learning process. Furthermore, DNN significantly underperformed, only outperforming the nearest neighbor baseline model in two out of the nine regions (SC and SFE).

Across all regions, the tuning entropy remained high for all models and increased with the number of predictors (Fig. 6). This effect was generally more pronounced for DNN than for RF. DNNs tuning entropy was high yet stable with respect to the number of predictors. Tuning entropies of RF optimal models were high in regions SAC, SC, and SECA (Fig. 6). In all regions, the RF tuning entropy increased with the initial addition of predictors. However, after this initial increase, the evolution of RF tuning entropy with the number of predictors was nuanced and noisy, but either dominantly decreased or increased. In regions K, NC, SAC, SECA, SFE and SJT, tuning entropy tended to decrease with additional predictors, whereas it tended to increase in regions NCC, SC and SCC. The optimal RF models used, in general, a lower number of predictors than DNN while maintaining a relatively high tuning entropy and clearly outperformed DNN in terms of AUC and accuracy (Table 5, Fig. 5). Consequently, RF was selected for performing network-scale predictions. This finding is similar to RF's performance in statistical learning reported by Guillon et al. (2020), who focused only on region SAC and used a different approach for predictor selection, training procedure, and model selection.

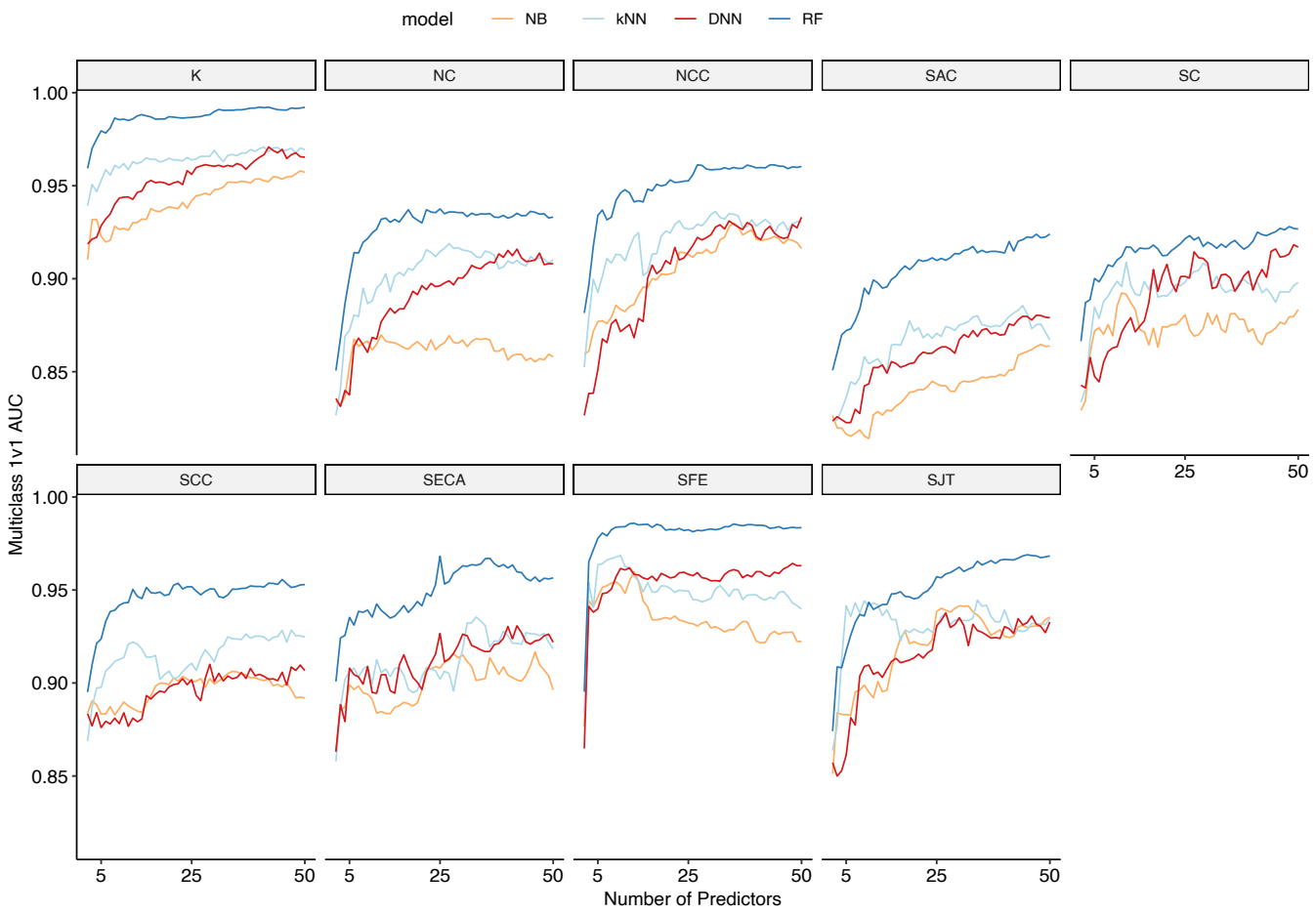

FIGURE 5 Evolution of the performance of ML models measured by multiclass 1v1 Area Under Curve (AUC) with an increasing number of predictors. The featureless model is not pictured: its AUC is constant at 0.5. NB: Naive Bayes; kNN: *k*-Nearest Neighbors; DNN: Deep Neural Network; RF: Random Forest.

3.3 | Channel Type Predictions and River Segmentation Throughout California

The spatial distribution of RF predictions was generally consistent with the expected distribution of channel types (Figure 7). The RF algorithm predicted the channel type of 689,029 individual 200-m stream intervals present in the NHD Plus V2 dataset (McKay et al. 2012). Adjacent intervals with the same type may be merged to yield a river segmentation into reaches of non-uniform length.

Valley confinement was most often selected as a predictor in the optimal RF models across all regions (see Figure 8), which is not surprising given its dominant control over channel setting (Fryirs et al. 2016, Lane et al. 2021). Relevant predictors in over half of the regions also included the standard deviation of elevation, the statistical roughness of topography at small spatial scales (Hurst coefficients), median slope, and curvature metrics. Drainage areas at the watershed and stream segment scale appeared to be relevant, albeit only in less than half of the regions. Contextual predictors only appeared in the optimal set of predictors in the SC region, where, in addition to valley confinement and drainage area metrics, they correspond to nine predictors describing lithology (6) and land use (3).

3.4 | Correlation Analysis

The correlation analysis indicated that clustering resulted in different numbers of channel types and had an impact on ML performance due to several sampling design factors (Tables 1,4-5; Fig. 5,9). The number of channel types was strongly linked to the number of observations for clustering characteristics ($r = 0.90$, $p = 0.008$). Observation density was negatively correlated with catchment area ($r = -0.63$, $p = 0.14$), but neither the number of channel types nor the number of observations were definitively linked to area and observation density. Area and observation density were negatively and positively

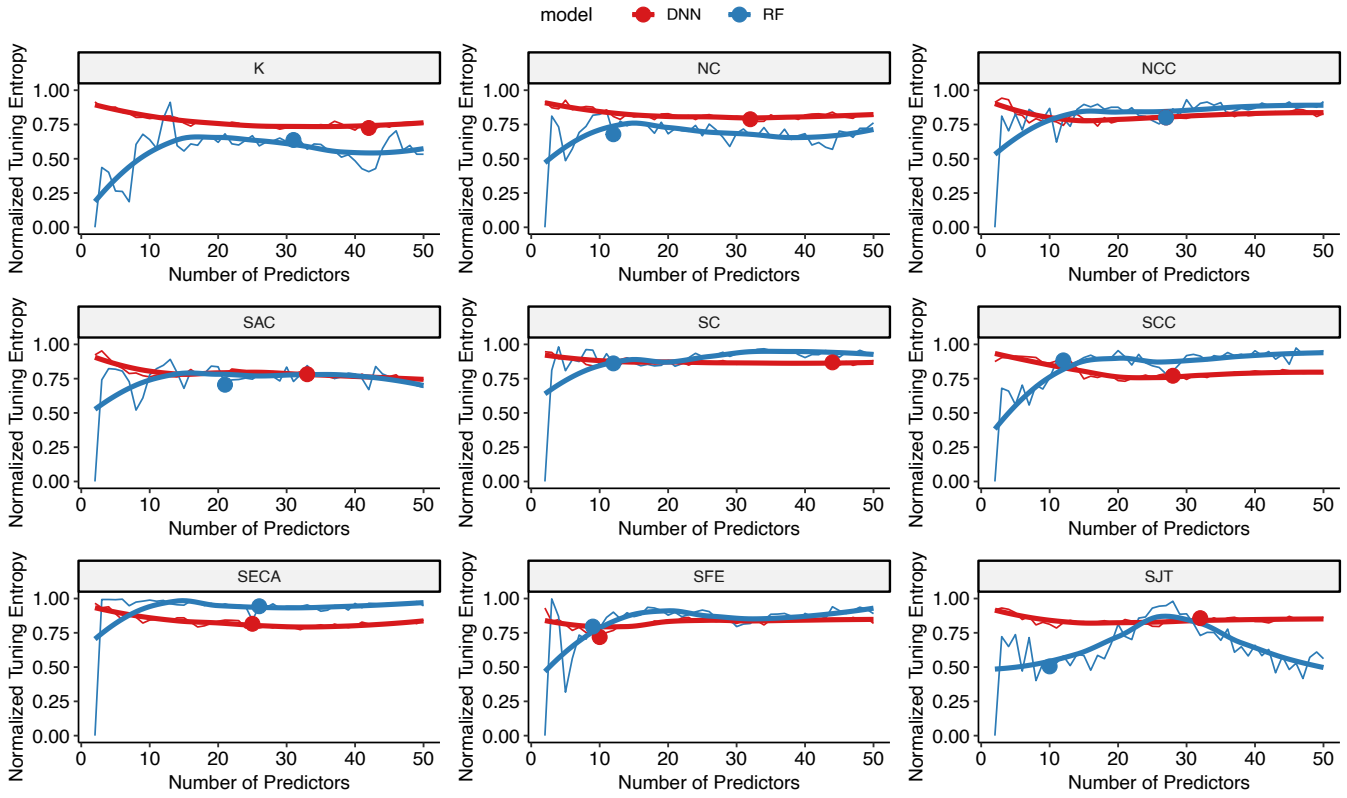

FIGURE 6 Evolution of tuning entropies with an increasing number of predictors. Solid circles represent the optimal model. DNN: Deep Neural Network; RF: Random Forest. The solid circle represents the value of tuning entropy for the optimal models.

correlated with the number of valley-confined channel types, respectively (r = –0.48, p = 0.30; and r = 0.66, p = 0.13). Observation density and area were not correlated with the statistical learning performance for DNN or RF (Fig. 9a-b). Instead, statistical learning performance metrics were negatively correlated with the number of observations, the number of channel types, and area: for DNN's area under the curve (AUC), r = –0.60, p = 0.19, r = –0.45, p = 0.33, r = –0.31, p = 0.50, respectively; for RF's AUC, r = –0.44, p = 0.32, r = –0.21, p = 0.50, r = –0.35, p = 0.48, respectively. Accuracy and AUC were positively correlated with each other. All variations of $\bar{d}_{JS}$ were positively correlated with one another and negatively correlated with the number of channel types and observations. Specifically, the minimum $\bar{d}_{JS}$ was negatively correlated with the number of observations (r = –0.82, p = 0.04). Median and minimum $\bar{d}_{JS}$ across all channel types were positively correlated with statistical learning performance metrics, but more so for DNN than RF: for DNN's metrics, $r \gtrsim 0.72$, $p \lesssim 0.08$, $r \gtrsim 0.69$, $p \lesssim 0.12$, respectively; for RF's metrics, $r \gtrsim 0.59$, $p \lesssim 0.19$, $r \gtrsim 0.47$, $p \lesssim 0.32$, respectively. Correlations were generally lower for median and minimum $\bar{d}_{JS}$ over valley-confined channel types only (Fig. 9). However, performance metrics were negatively correlated with the number of observations and the number of classes ($r \gtrsim$ –0.70, $p \lesssim 0.10$).

The more dissimilar the clusters were with respect to field-measured data, the more stable the predictions were. The entropy rate of RF predictions was generally negatively correlated with the $\bar{d}_{JS}$ metrics. This was especially true for the minimum $\bar{d}_{JS}$ for confined channel types (r = –0.80, p = 0.04). In addition, the entropy rate was weakly correlated with regional metrics that increase in complexity, such as the number of observations, number of channel types, and number of confined channel types. For RF, the entropy rate and hyperparameter tuning entropy were only weakly linked (r = –0.32, p = 0.48, Fig. 9b). Both were negatively correlated, albeit weakly, with statistical learning performance metrics. Hyperparameter tuning entropy appeared mostly disconnected from statistical learning performance metrics (r = –0.02, p = 0.76). In general, hyperparameter tuning entropy showed a weak correlation with the other variables, with the exception of the minimum $\bar{d}_{JS}$ for confined channel types (r = 0.47, p = 0.31) and the number of confined channel types (r = –0.47, p = 0.29). This suggests that hyperparameter tuning entropy increases with decreasing complexity, while entropy rate increases with increasing complexity.

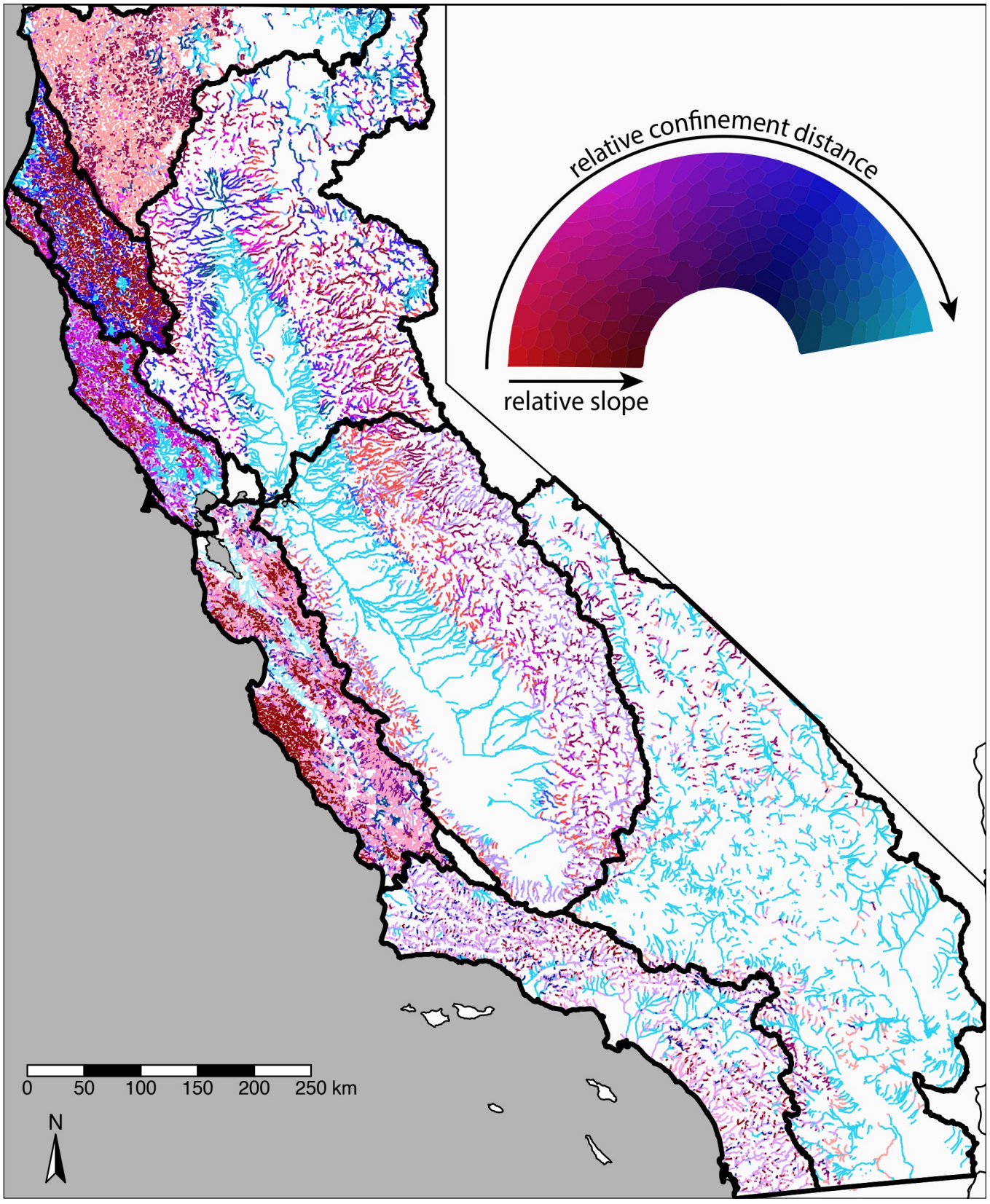

FIGURE 7 Map of all RF predictions of reach-scale channel types. In each region, hue maps to confinement so that cyan (red) corresponds to the most unconfined (confined) channel type, and lightness maps to slope so that the channel type with low (high) slope are drawn in lighter (darker) colors.

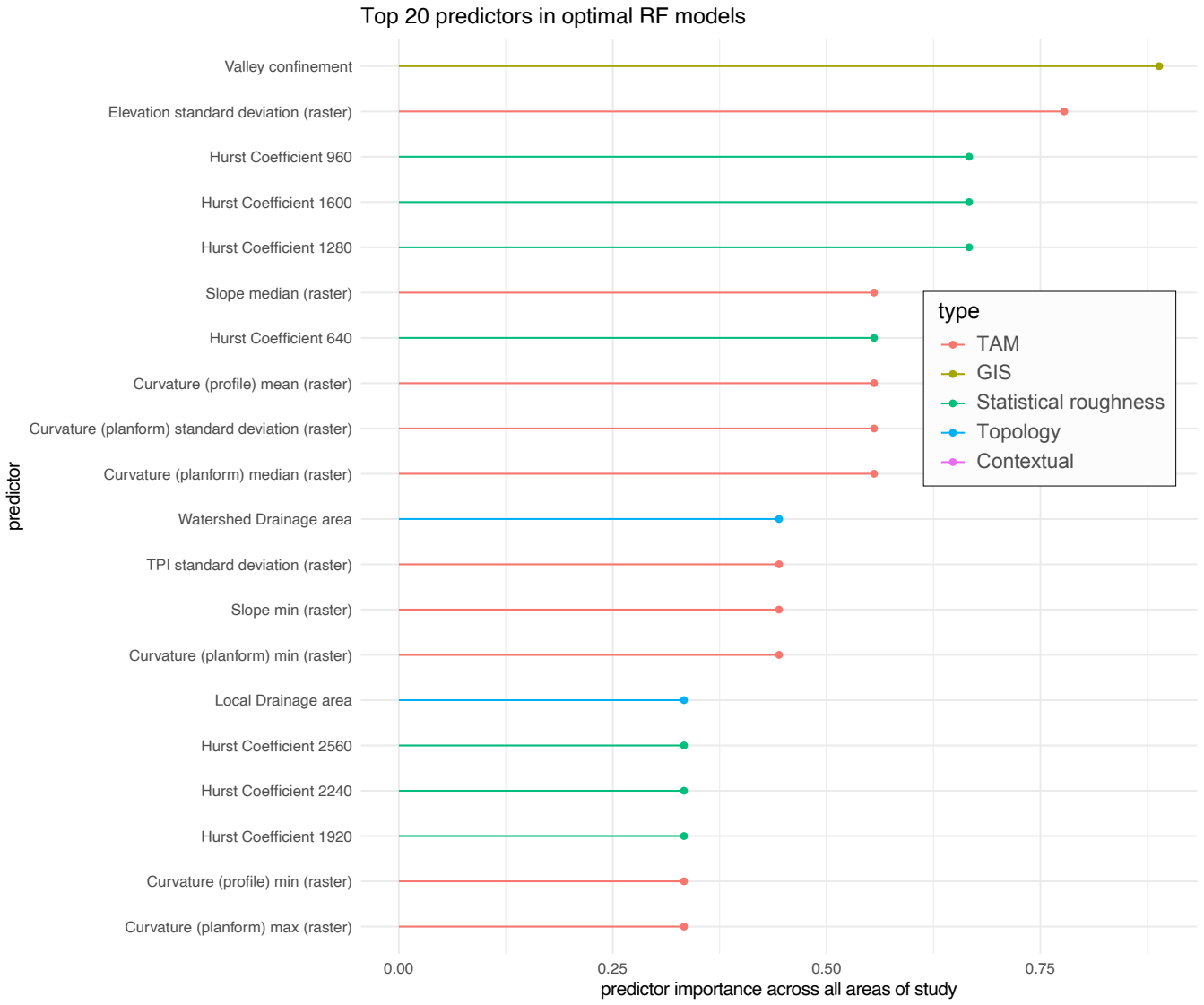

FIGURE 8 Regional variable importance. TAM: Terrain Analysis Metrics predictors (see Guillon et al. (2020)); GIS: GIS-derived predictors. The Hurst coefficient predictors represent a measure of statistical roughness at different spatial scale. TAM were calculated at two spatial scales: a 512-m tile centered on the 200-m stream line midpoint (*raster*) and a riparian buffer of 100-m (*near*).

The performance metrics of RF and DNN models prognosticated the information gap between clustering data and prediction data (Fig. 9c), and their correlations were inverted compared to DNN and RF correlations (Fig. 9a-b). Moreover, the difference in statistical learning performance between RF and DNN models was positively correlated with the number of observations and the number of channel types, as well as the number of valley-confined channel types. In contrast, it was negatively correlated with all $\bar{d}_{JS}$ metrics, particularly the minimum $\bar{d}_{JS}$ ($r = -0.58$, $p = 0.18$; $r = -0.61$, $p = 0.15$).

4 | DISCUSSION

Our correlation analysis related three groups of variables across nine regions – regional clustering characteristics, information gap, and performance metrics (Fig. 1b), In doing so, it answers our research questions and corroborates our three hypotheses. The information gap between clustering data and prediction data, and here specifically between field-measured and geospatial data, was estimated by the cluster dissimilarity with respect to field-measured data, and is central to evaluate the impact of clustering at data-rich locations on predictions at data-poor locations. We found a positive relationship between

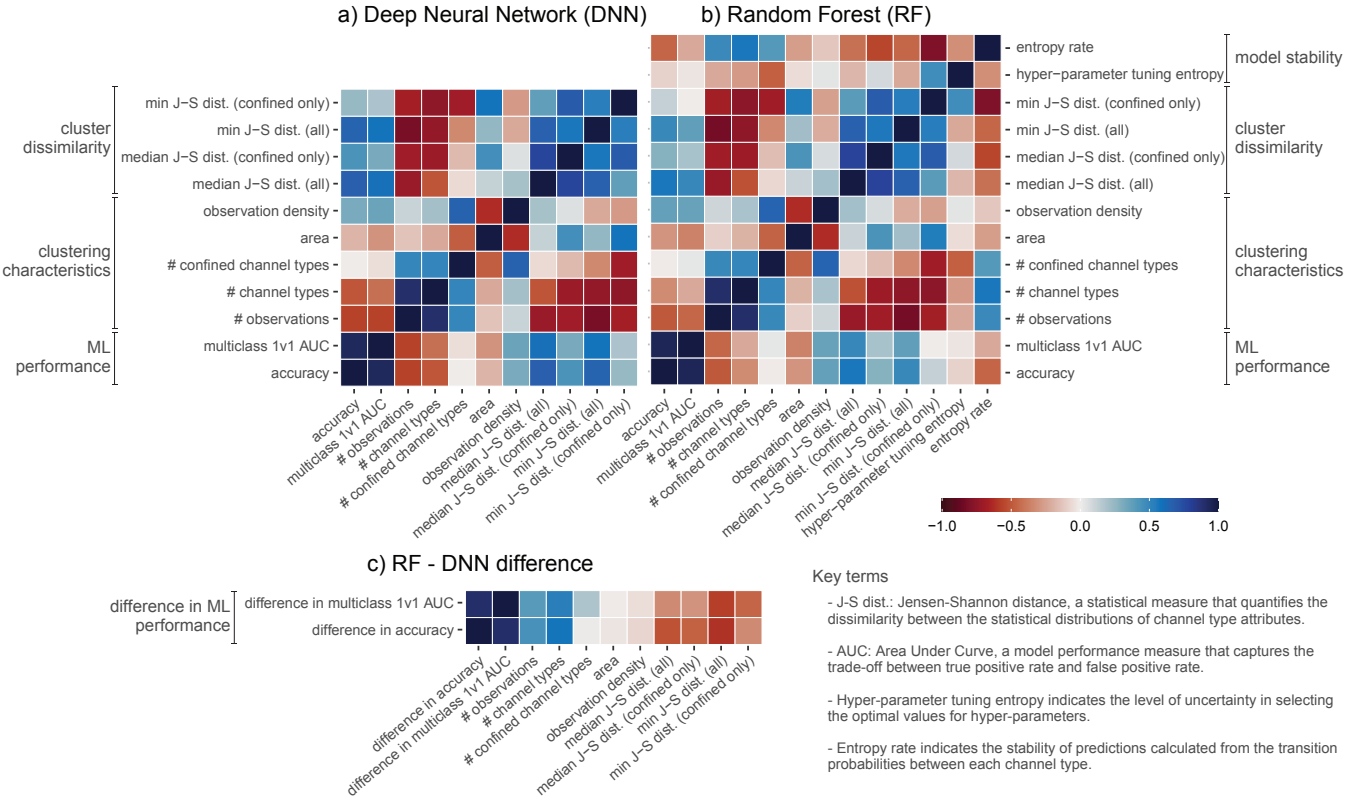

FIGURE 9 Correlation matrix for a) DNN; b) RF and c) difference between DNN and RF.

cluster dissimilarity measured by the Jensen-Shannon distance and ML performance (Fig. 9a-b), indicating that a smaller information gap leads to better performance, responding to Q1 and corroborating H1. Hence, even when following identical sampling and analysis methods, clustering impacts subsequent ML predictions. Moreover, this effect is stronger for DNN than for RF (Fig. 9a-b), and the difference in performance between RF and DNN decreases with cluster dissimilarity (Fig. 9c), responding to Q2 and corroborating H2. We also found that clusters with the smallest spatial footprint, or valley-confined channel types, have a lower average cluster dissimilarity in a given region while often not the lowest (Table 4), responding to Q3 and supporting H3. In the following section, we explain and discuss these results, their limitations and implications.

4.1 | Information Gap between Clustering and Prediction Data Explains why RF Outperforms DNN

Evaluating the information gap between clustering data and prediction data from cluster dissimilarity measured by the Jensen-Shannon distance (Fig. 4, Table 4) helps with interpreting and comparing labels derived from clustering and with assessing their impact on statistical learning performance. In particular, differences in average cluster dissimilarity can be interpreted as differences in the scale at which the clusters are inherently defined within and between the different regions. Further, in the application to clustering channel forms in nine regions of California, our results suggest that cluster dissimilarity is linked to the scale mismatch between labels and geospatial predictors and explains deep learning underperformance. As in Guillon et al. (2020), DNN underperforms relative to RF in most regions of California (Fig. 5, Table 5).

Two combined reasons help explain this under-performance of DNN. First, in general, DNN performance increases with the number of available observations, and the current data deluge partly explains their increasing popularity (LeCun et al. 2015). For example, Nearing et al. (2021) noted that while deep learning has had transformative results in branches of hydrology benefiting from rich observational data from sensor networks and remote sensing (e.g., catchment hydrology, streamflow predictions, hydrometeorology), its success has been more limited in branches with sparser data (e.g., groundwater hydrology). Similarly, Kirstain et al. (2021) showed that performance was more efficiently increased by collecting more training data than by training larger models when experimenting with a deep learning model whose architecture is used in some hydrologic tasks (e.g., Sahoo et al. 2017, Adewoyin

et al. 2021). In addition, while there have been recent breakthroughs in using DNN for tabular datasets ranging in size from 2.5 10^3 to 10^7 observations (Shavitt and Segal 2018, Arık and Pfister 2020), this remains at least an order of magnitude larger than the datasets in our research study (63-290 observations, Table 1). This is a common issue in ML predictions of patterns of channel forms and hydrologic response.

The under-performance of DNNs can also be attributed to their sequential compression of information through successive layers, which filters out irrelevant information to estimate the relationship between input and output. To better understand this issue, we discuss relevant research on statistical learning and information processing in DNNs. The combination of feature engineering and information compression in DNNs is similar to a central concept in modern statistical and particle physics: the renormalization group (Stuckelberg 1953, Gell-Mann and Low 1954). The renormalization group summarizes local random variables by coarse-grained random variables, such as averages of neighboring points, when considering a system at increasing length scales (see Turcotte (1997) and Sornette (2006) for an introduction in geosciences). This sequential coarse-graining from a microscopic scale to a macroscopic scale parallels DNNs' information processing (Bény 2013, Mehta and Schwab 2014, Lin et al. 2017, Li and Wang 2018, Koch-Janusz and Ringel 2018, Yang et al. 2023): where a DNN learns and engineers relevant features in its layers, and the renormalization group's coarse-graining retains only relevant degrees of freedom and integrates out the irrelevant ones. A complete theory for such sequential information processing and compression (Saxe et al. 2019, Gabrié et al. 2019, de Mello Koch et al. 2020), and more generally for DNNs' ability to learn and generalize complex patterns, still remains elusive (Zhang et al. 2016, Sejnowski 2020, Poggio et al. 2020). However, recent analysis of DNNs' internal dynamics and learned representations substantiate its relevance. For example, Bau et al. (2020) dissected a convolutional DNN trained on image recognition and showed that individual units matched human-interpretable concepts. Similarly, Amey et al. (2021) descrambled the weight matrices of fully connected DNNs, which are usually harder to interpret, and found that each layer had learned recognizable mathematics and communications engineering concepts, such as a band-pass filter. Observing the dynamics of DNN internal representations, Schilling et al. (2021) introduced a dissimilarity measure computed from neural activations and showed that performance increases over time and through the successive layers of the network. A similar process was described from the perspective of information flow using the Kullback-Leibler divergence (Erdmenger et al. 2021), Wasserstein distance (Cao et al. 2022), correlation functions (Fischer et al. 2022) and with explicit coarse-graining operations (Yang et al. 2023). Although Hartmann et al. (2021) found similar patterns, their study of neural activation statistics provides a more nuanced view that focuses on state-of-the-art architectures rather than learning theory. Finally, Fang et al. (2021) isolated the topmost layer of a DNN, which predicts observations from the learned representations, and analyzed it in terms of neural collapse (Papyan et al. 2020). Neural collapse is a common empirical pattern in DNN where the last-layer representations collapse close to the cluster mean, allowing the network to attribute clusters based on the closest cluster mean. While neural collapse improves generalization, interpretability, and robustness, Fang et al. (2021) showed that it does not occur in widely imbalanced datasets leading to indistinguishable under-represented clusters.

While we broadly agree with the findings from a recent empirical comparison of DT-based models and DNNs (Grinsztajn et al. 2022), we further connect the underperformance of DNNs to the difference in information processing between RF and DNN. Grinsztajn et al. (2022) showed that DT-based models outperformed DNNs on 45 datasets with balanced classes and more than 3,000 observations. In one particular experiment, Grinsztajn et al. (2022) smoothed the target variable during training but not during validation, essentially preventing the model from learning potentially irregular patterns. This creates an information gap between the target seen in training and in validation. With increasing smoothing, such an information gap increases, which degrades the performance of both RF and DNN. Note that the approach from Grinsztajn et al. (2022) is applied within one dataset and applied to 45 datasets from various domains, whereas we discuss here a comparison between nine datasets from one domain.

In this study, we observe that the information gap between clustering data and prediction data leads to missing or overly noisy information, which hinders efficient information processing in the DNN and reverse engineering of the hierarchical generative process between input and output (Tishby and Zaslavsky 2015, Lin et al. 2017). In addition, our datasets are limited in size (see Table 1), and the channel types are defined from field scale data (< 200 m) while the input predictors are defined at a coarser scale (> 500 m) (see Table 3). The impact of the information gap stemming from these mismatched spatial scales is illustrated by the correlation between cluster dissimilarity with respect to field-measured data and the performance of DNN relative to RF. The more dissimilar the channel types, the lower the difference between RF and DNN (see Fig. 9c). This indicates that the difference in information processing between DNN and RF predicts their performance. With additional observations, DNNs' information compression could better filter out noisy information, reducing the

performance gap between RF and DNN (see Supporting Information for an example using 10-m and 1-m data in NCC region). However, in the case of limited, noisy, tabular datasets with a potential information gap stemming from a scale mismatch between labels and predictors – a common issue in geosciences – our results suggest that algorithms without sequential information compression (e.g., RF) may consistently outperform DNN-inspired algorithms.

4.2 | RF Limitations

Although random forest (RF) performs well in estimating patterns between channel types and predictors (see Figure 5), generalizing the learned pattern in predictive modeling and expert assessment of the geomorphic relevance of the resulting predictions by the authors led to the implementation of post-hoc heuristics for predictions in region K. While geomorphologists often use field experience and expert knowledge to form channel-type expectations, this approach has proven to be too subjective, opaque, and non-repeatable as a general practice. Nevertheless, we hoped it could be useful for interpreting prediction performance and adjusting results for region K. For example, when comparing the predicted spatial distribution of channel types with the expected one, the mainstem channel type K03 appeared difficult to predict. Consequently, we implemented a stream-order-based heuristic. Similarly, expert assessment played an important role in evaluating the clustering results from Byrne et al. (2019), Guillon et al. (2019), Byrne et al. (2020a). In another type of performance review, we thoroughly evaluated an individual site's curious cluster assignment while standing on-site with all the data in hand. In these cases, we found that observational values close to a threshold but just over the line in the "wrong" direction at a decision-tree node could send the site to a different, less sensible cluster. This highlights the difference between human expert opinion, which prefers to classify rivers on a top-down basis, prioritizing fundamental geological controls, and a pattern recognition algorithm that is teasing out similarities and differences among all available variables.

Limited sampling is the primary reason for the need of a post-hoc heuristic to adjust predictions with expectations in region K. Channel types have significantly different natural abundances, with some types being rare and challenging to isolate in the sampling scheme, while others are so altered that natural examples are scarcely available. As the clustering is data-driven from fine-scale field measurements, one cannot accurately discern which sites are which type a priori. In an attempt to mitigate this issue, this study relied on mindful and intensive sampling designs using GIS-derived stratification data at a coarser scale to seek equal-effort sampling among the most likely different channel types (Lane et al. 2017b, Byrne et al. 2020b). Despite this unusual effort compared to past clustering studies, the resulting clusters and the number of observations per cluster ended up being quite different from the forecasted experimental design (Lane et al. 2017b, Byrne et al. 2020b). In all regions, the resulting unequal sampling of channel types is addressed with the commonly used synthetic minority oversampling technique (Chawla et al. 2002), which generates synthetic observations to aid the statistical learning of channel types with a lower number of observations. However, the random generation of synthetic observations is handled with a k-Nearest Neighbour algorithm (with $k \leq 5$ in our case depending on the number of available observations). Consequently, fewer field observations of a channel type resulted in less diversity in the corresponding synthetic data, hindering robust learning of the patterns between under-sampled channel types and predictors. In region K, the mispredicted channel type has the lowest possible value for prevalence, only one out of 105 observations, and thus no diversity in the associated synthetic data (Table 2). The next rarest channel type, represented by four out of 105 observations, appeared frequently enough to enable robust pattern learning when compared to expert evaluation. Interestingly, most of the under-sampled channel types fall into two categories tied to the logistics of field sampling: high-order mainstem rivers and low-order steep cascade/step-pool channels. High-order mainstem rivers are often highly channelized and altered while displaying larger channel dimensions that hinder field sampling. Low-order steep cascade/step-pool channels are difficult to access through private land and remote, dangerous terrain, leading to sampling a specific subset of the most accessible channels that may bias resulting clusters.

4.3 | Implications for clustering then predicting

This study's results have general implications for sampling strategies to cluster and predict information in watershed sciences. The correlation analysis (Fig. 9) underlines a positive correlation between the number of field observations and the number of clusters, and a negative correlation between the number of clusters and cluster dissimilarity. This translates into better ML performance. With fewer observations, the likelihood of finding statistically significant groupings decreases, resulting in fewer, more dissimilar clusters that can be separated with coarser information, which reduces the complexity of the problem. Conversely, an increasing number of observations leads to fine-scale, less dissimilar clusters, at least for some channel types, and to a more complex problem. This effect is amplified by the information gap between clustering data and prediction data stemming from mismatched spatial scales. In other words, with fewer observations, one

can get away with a simpler, coarse-scale problem to solve. For example, riffle-pool reaches are common in California, and random sampling may oversample them despite our effort to stratify sampling using four meaningful catchment scale variables. This results in more representation and therefore diversity of riffle-pool reach channel types. In contrast, there could be an equal diversity of cascade reach channel types, but if there are fewer of these sites in the geographical study area, are more difficult to access and survey, or are randomly less sampled, then the clustering is more likely to group them into a unique cluster. This can result in mismatched scales of clustering between broader channel types with varying representation and diversity. Our study shows that, even when statistical learning performs well, generalizing the learned pattern beyond the training datasets may be hindered by insufficient or unequal sampling. Consequently, there exists a sampling trade-off between uniformly capturing natural variability and robustly learning a generalizable pattern. A small number of observations captures some of the natural variability in the study area at a uniform level of detail across broad channel types. Conversely, a larger number of observations ensures that a generalizable pattern is robustly learned across all broad channel types but can lead to an equivalent diversity of fine-scale clusters. This trade-off is likely an ubiquitous problem in watershed sciences where clustering and prediction contends with a mix of rare and common types, multi-scalar typologies, uneven human disturbances across types, limited sampling capability, and high uncertainty in design of the sampling strategy. Characterizing this trade-off space and optimal sampling design is beyond the scope of this study, but likely depend on the definition of the clustering, which then conditions the performance in statistical learning. Additionally, to the authors knowledge, there is no framework in watershed sciences that factors in the cost for variable acquisition in the total cost function of a clustering problem (e.g., Andrade and Okajima 2021).

The increasing popularity of clustering in watershed sciences raises the question of how to compare them. With diversity in purposes, data types, approaches, and instances for the same environmental systems, science needs to synthesize and interpret that diversity and complexity to enable broader understanding and societal benefit beyond each original application. However, applying one geomorphic clustering approach to multiple regions highlights that comparison is not straightforward. This likely remains true when comparing different clustering approaches in a single region of interest. In particular, the information content of the clustering, interpreted as the overall spatial scale at which channel types are defined, varies widely, hindering direct comparison across regions. This leads to a fuzzy correspondence between clusters across regions, akin to the loose agreement between empirical classifications of channel types reported by Kasprak et al. (2016). A better strategy to robustly compare areas of study or combine results may be to assemble a dataset spanning geographical areas of interest and perform a new clustering pooling all data into one set, as was done in Lane et al. (2021). This better ensures that clusters are defined with a similar level of information. However, such an approach is only tractable if the underlying sampling methods, raw data, and data processing steps of clustering are reasonably similar. In addition, the human preference for few clusters, statistical metrics that constrain the number of clusters to a small set in light of data limitations, and the unequal sampling of common versus rare channel types pose the risk that important, diverse phenomena will be lost in data-driven approaches. While we focused here on the cluster-first approach, this remains true for the predict-first approach as both approaches are predicted on sampling. Notwithstanding, Peñas et al. (2014) showed that under-represented hydrological patterns were more often predicted throughout the network with a predict-first approach than with a cluster-first approach. Regardless of the approach, these concerns increase as study area increases for the same sampling density and may lead to under-representing natural diversity which could endanger rare environments and species. Accordingly, these concerns are magnified at the global scale (Meyer and Pebesma 2022).

An objective constraint on the specific scale of the set of statistical classification labels, as presented in this study through cluster dissimilarity (Fig. 4) and deep learning relative performance (Fig. 9c), is likely to benefit a wide variety of applications across watershed sciences. In particular, the same label is often used by different scientists to represent a range of spatial or temporal scales. For example, in fluvial geomorphology, a common label used to describe a site on a river is a "riffle-pool reach". However, this label has no inherent spatial scale: some studies use it to refer to lengths as short as 1-5 times channel width, while others use the same label to refer to lengths as long as 100-1000 times channel width. Further complicating the definition of scale, channel form results from multiscale interactions, both in time and space (Lane and Richards 1997), between confinement (Fryirs et al. 2016), transport rate (Singh et al. 2009), sediment supply (Attal and Lavé 2006), and biota (Corenblit et al. 2011). Having more explicitly defined scales associated with statistically derived labels (here, channel types) would yield a more transparent and universal lexicon and facilitate a better understanding of eco-physical processes intertwined with spatio-temporal patterns represented by labels. sTo that end, our results shows that mismatched scales are linked to the information gap between field-measured clustering data and geospatial prediction data which is equally evaluated from clustering data (Fig. 4, Table

4) and from the relative performance of traditional and deep learning approaches (Fig 9c).

5 | CONCLUSION

Natural resources managers increasingly rely on machine learning to inform regional decision-making. For example, river restoration or water management strategies may be selectively applied based on channel type predictions. However, when unevenly-distributed information leads to an information gap between data-rich and data-poor locations, it was unclear how clustering at data-rich locations impacts subsequent predictions at data-poor locations. This study characterized the impact of prior clustering on the statistical learning performance of two leading prediction approaches with distinct information processing – decision trees and deep learning. We estimated the information gap between clustering data and prediction data based on cluster dissimilarity with respect to field-measured data and related it to the performance of DNN and RF models using geospatial predictors. We leveraged nine examples of regional river channel form predictions stemming from a single clustering methodology applied in California, USA. Our findings suggest that clustering at data-rich locations impacts subsequent predictions at data-poor predictions and that this effect is stronger for DNN than for RF. Moreover, our results show that there is a trade-off between collecting the minimum number of observations to uniformly capture natural variability across channel types and collecting enough data to ensure that a generalizable pattern is learned. An increasing number of observations results in finer-scale clusters (at least for some channel types), which leads to a more complex problem, including mismatched scales between clusters and between labels and predictors. This mismatch in the spatial scale between clusters derived from field-measured attributes and geospatial predictors likely hinders efficient information processing and explains why RF outperforms DNN. Therefore, our results suggest that algorithms without sequential information compression (e.g., RF) will consistently outperform DNN-inspired algorithms in the case of limited and noisy tabular datasets with potential information gap resulting from a scale mismatch between labels and predictors.

Future research directions include investigating the optimal number of field sites to develop a regional clustering, while balancing the information content of the clustering, statistical learning performance, and cost of field attribute acquisition. Additionally, a quantitative framework could be developed for comparing clusterings in different geographical areas when the underlying data are dissimilar or unavailable. This framework could build on the results of this study and use information theory metrics, the difference in performance between deep and shallow learning, or both. Another direction is defining a more transparent, data-driven lexicon for the inherent scale of clusters derived from expert opinion to better understand the spatio-temporal patterns of eco-physical processes while accurately representing natural diversity. A formal comparison of cluster-first and predict-first approaches and of the associated trade-offs is missing in geomorphology. In general, deriving geospatial predictors at a finer spatial scale to reduce the potential scale mismatch between field-derived labels and geospatial predictors and associated information gap remains an area of future research. For example, in fluvial geomorphology, tools to extract meaningful sub-reach scale geomorphic attributes from remote sensing products are still limited, error-prone, and constrained by data availability. Similarly, leveraging a different data modality, such as remote sensing imagery, to replace or augment available coarse geospatial predictors would increase the information content in geospatial predictors, likely favoring sequential information processing and compression from deep learning. Finally, recent breakthroughs in self-supervised and semi-supervised learning would likely improve the performance of deep learning approaches by leveraging data over the entire prediction domain, rather than relying solely on labeled data (Shwartz-Ziv and LeCun 2023) even with tabular data (Yoon et al. 2020).

ACKNOWLEDGMENTS

This research was supported by the California State Water Resources Control Board under grant number 16-062-300. We also acknowledge the U.S. Department of Agriculture, Hatch project number CADLAW7034H, and the Utah Water Research Laboratory. Data sources are reported in Table 3. Code, long form documentation and data are available throught the open source `R` package `RiverML v1.0.0` archived at https://doi.org/10.5281/zenodo.4062525.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

REFERENCES

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M. et al. (2012) Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography*, 36(4), 480–513.

Adewoyin, R.A., Dueben, P., Watson, P., He, Y. & Dutta, R. (2021) TRU-NET: a deep learning approach to high resolution prediction of rainfall. *Machine Learning*,, 1–28.

Alfredsen, K., Dalsgård, A., Shamsaliei, S., Halleraker, J.H. & Gundersen, O.E. (2022) Towards an automatic characterization of riverscape development by deep learning. *River Research and Applications*, 38(4), 810–816.

Amey, J.L., Keeley, J., Choudhury, T. & Kuprov, I. (2021) Neural network interpretation using descrambler groups. *Proceedings of the National Academy of Sciences*, 118(5).

Andrade, D. & Okajima, Y. (2021) Adaptive covariate acquisition for minimizing total cost of classification. *Machine Learning*, 110(5), 1067–1104.

Arık, S.O. & Pfister, T. (2020) Tabnet: Attentive interpretable tabular learning. *arXiv*,.

Attal, M. & Lavé, J. (2006) Changes of bedload characteristics along the Marsyandi River (central Nepal): Implications for understanding hillslope sediment supply, sediment load evolution along fluvial networks, and denudation in active orogenic belts. In: *Special Paper 398: Tectonics, Climate, and Landscape Evolution*Geological Society of America, pp. 143–171.
URL http://dx.doi.org/10.1130/2006.2398(09)

Bau, D., Zhu, J.Y., Strobelt, H., Lapedriza, A., Zhou, B. & Torralba, A. (2020) Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48), 30071–30078.

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(1), 39–57.

Bengio, Y., Courville, A. & Vincent, P. (2013) Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.

Bény, C. (2013) Deep learning and the renormalization group. *arXiv preprint arXiv:1301.3124*,.

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.

Bhattacharya, B., Price, R. & Solomatine, D. (2007) Machine learning approach to modeling sediment transport. *Journal of Hydraulic Engineering*, 133(4), 440–450.

Bischl, B., Mersmann, O., Trautmann, H. & Weihs, C. (2012) Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, 20(2), 249–275. doi:10.1162/evco_a_00069.
URL https://doi.org/10.1162/evco{_}a{_}00069

Bomers, A., van der Meulen, B., Schielen, R. & Hulscher, S. (2019) Historic flood reconstruction with the use of an artificial neural network. *Water resources research*, 55(11), 9673–9688.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J. & Lang, M. (2020) Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143, 106839.

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

Byrne, C.F., Guillon, H., Lane, B.A., Pasternack, G.B. & Solis, S.S. (2019) *Sacramento River Basin Geomorphic Classification: Final Report – Submitted to the California State Water Resources Control Board*. University of California, Davis.
URL https://watermanagement.ucdavis.edu/download_file/view_inline/258

Byrne, C.F., Guillon, H., Lane, B.A., Pasternack, G.B. & Solis, S.S. (2020) *Coastal California Regional Geomorphic Classification: Final Report – Submitted to the California State Water Resources Control Board*. University of California, Davis.
URL https://watermanagement.ucdavis.edu/download_file/view_inline/509

Byrne, C.F., Pasternack, G.B., Guillon, H., Lane, B.A. & Sandoval-Solis, S. (2020) Reach-scale bankfull channel types can exist independently of catchment hydrology. *Earth Surface Processes and Landforms*, 45(9), 2179–2200.

Cao, J., Li, J., Hu, X., Wu, X. & Tan, M. (2022) Towards interpreting deep neural networks via layer behavior understanding. *Machine Learning*, 111(3), 1159–1179.

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.

Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016*, pp. 785–794.

Chen, X., Hassan, M.A. & Fu, X. (2022) Convolutional neural networks for image-based sediment detection applied to a large terrestrial and airborne dataset. *Earth Surface Dynamics*, 10(2), 349–366.

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

Corenblit, D., Baas, A.C., Bornette, G., Darrozes, J., Delmotte, S., Francis, R.A. et al. (2011) Feedbacks between geomorphology and biota controlling Earth surface processes and landforms: A review of foundation concepts and current understandings. *Earth-Science Reviews*, 106(3-4), 307–331. doi:10.1016/j.earscirev.2011.03.002.
URL https://doi.org/10.1016%2Fj.earscirev.2011.03.002

Cover, T. & Hart, P. (1967) Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21–27.

Cress, J., Soller, D., Sayre, R., Comer, P. & Warner, H. (2010) *Terrestrial ecosystems – Surficial lithology of the conterminous United States*. U.S. Geological Survey Scientific Investigations Map 3126, scale 1:5,000,000, 1 sheet.
URL https://pubs.usgs.gov/sim/3126/

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.

Dallaire, C.O., Lehner, B., Sayre, R. & Thieme, M. (2019) A multidisciplinary framework to derive global river reach classifications at high spatial resolution. *Environmental Research Letters*, 14(2), 024003.

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. doi:10.1016/j.geomorph.2017.06.009.
URL https://doi.org/10.1016%2Fj.geomorph.2017.06.009

de Mello Koch, E., de Mello Koch, A., Kastanos, N. & Cheng, L. (2020) Short-sighted deep learning. *Physical Review E*, 102(1), 013307.

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.

Dowla, F.U., Taylor, S.R. & Anderson, R.W. (1990) Seismic discrimination with artificial neural networks: preliminary results with regional spectral data. *Bulletin of the Seismological Society of America*, 80(5), 1346–1373.

Endres, D.M. & Schindelin, J.E. (2003) A new metric for probability distributions. *IEEE Transactions on Information theory*,.doi:10.1109/TIT.2003.813506.

Erdmenger, J., Grosvenor, K.T. & Jefferson, R. (2021) Towards quantifying information flows: relative entropy in deep neural networks and the renormalization group. *arXiv preprint arXiv:2107.06898*,.

Ermini, L., Catani, F. & Casagli, N. (2005) Artificial neural networks applied to landslide susceptibility assessment. *Geomorphology*, 66(1-4), 327–343.

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

Fang, C., He, H., Long, Q. & Su, W.J. (2021) Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43).

Ferri, C., Hernández-Orallo, J. & Modroiu, R. (2009) An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1), 27–38.

Fischer, K., René, A., Keup, C., Layer, M., Dahmen, D. & Helias, M. (2022) Decomposing neural networks as mappings of correlation functions. *Physical Review Research*, 4(4), 043143.

Fleming, S.W., Bourdin, D.R., Campbell, D., Stull, R.B. & Gardner, T. (2015) Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific Northwest river. *JAWRA Journal of the American Water Resources Association*, 51(2),

502–512.

Flores, A.N., Bledsoe, B.P., Cuhaciyan, C.O. & Wohl, E.E. (2006) Channel-reach morphology dependence on energy, scale, and hydroclimatic processes with implications for prediction using geospatial data. *Water Resources Research*, 42(6).

Florinsky, I.V. (1998) . *International Journal of Geographical Information Science*, 12(1), 47–62. doi:10.1080/136588198242003.
URL https://doi.org/10.1080%2F136588198242003

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. doi:10.1002/esp.3893.
URL https://doi.org/10.1002%2Fesp.3893

Gabrié, M., Manoel, A., Luneau, C., Barbier, J., Macris, N., Krzakala, F. et al. (2019) Entropy and mutual information in models of deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 124014.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J. & Hochreiter, S. (2021) Rainfall–runoff prediction at multiple timescales with a single long short-term memory network. *Hydrology and Earth System Sciences*, 25(4), 2045–2062.

Gaucherel, C., Frelat, R., Salomon, L., Rouy, B., Pandey, N. & Cudennec, C. (2017) Regional watershed characterization and classification with river network analyses. *Earth Surface Processes and Landforms*, 42(13), 2068–2081. doi:10.1002/esp.4172.
URL https://doi.org/10.1002/esp.4172

Gell-Mann, M. & Low, F.E. (1954) Quantum electrodynamics at small distances. *Physical Review*, 95(5), 1300.

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.

Gini, C. (1936) On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208(1), 73–79.

Gómez, R.D., Pasternack, G.B., Guillon, H., Byrne, C.F., Schwindt, S., Larrieu, K.G. et al. (2022) Mapping subaerial sand-gravel-cobble fluvial sediment facies using airborne lidar and machine learning. *Geomorphology*, 401, 108106.

Grinsztajn, L., Oyallon, E. & Varoquaux, G. (2022) Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520.

Guillon, H., Byrne, C.F., Lane, B.A., Pasternack, G.B. & Solis, S.S. (2019) *South fork of the Eel river Basin geomorphic Classification: Final Report – Submitted to the California State Water Resources Control board*. University of California, Davis.
URL https://watermanagement.ucdavis.edu/download_file/view_inline/144

Guillon, H., Byrne, C.F., Lane, B.A., Solis, S.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*,.

Guyon, I. & Elisseeff, A. (2003) An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157–1182.

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

Hartmann, D., Franzen, D. & Brodehl, S. (2021) Studying the evolution of neural activation patterns during training of feed-forward relu networks. *Frontiers in Artificial Intelligence*, 4, 642374.

Henshaw, A.J., Sekarsari, P.W., Zolezzi, G. & Gurnell, A.M. (2019) Google Earth as a data source for investigating river forms and processes: Discriminating river types using form-based process indicators. *Earth Surface Processes and Landforms*,.

Hijmans, R.J., van Etten, J., Cheng, J., Greenberg, J.A., Lamigueiro, O.P., Bevan, A. et al. (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. doi:10.1111/1752-1688.12372.
URL https://doi.org/10.1111/1752-1688.12372

Hinton, G.E. (1984) Distributed representations,.

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.

Huang, J. & Ling, C.X. (2005) Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3), 299–310.

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.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. et al. (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 3146–3154.

Kirstain, Y., Lewis, P., Riedel, S. & Levy, O. (2021) A few more examples may be worth billions of parameters. *arXiv preprint arXiv:2110.04374*,.

Koch-Janusz, M. & Ringel, Z. (2018) Mutual information, neural networks and the renormalization group. *Nature Physics*, 14(6), 578–582.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. & Nearing, G. (2019) Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12), 5089–5110.

Kullback, S. & Leibler, R.A. (1951) On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79–86. doi:10.1214/aoms/1177729694.

Lane, B., Guillon, H., Byrne, C., Pasternack, G.B., Kasprak, A. & Sandoval-Solis, S. (2021) Channel-reach morphology and landscape properties are linked across a large heterogeneous region. *Earth Surface Processes and Landforms*,.

Lane, B.A. & Byrne, C.F. (2021) *California river classification field survey protocols*.
URL https://doi.org/10.4211/hs.023f24c1a62f48f496e10b7cbafe6b86

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.B., Dahlke, Helen, E. & Sandoval-Solis, S. (2017) The role of topographic variability in river channel classification. *Progress in Physical Geography*,.

Lane, B.A., Pasternack, G.B. & Sandoval Solis, S. (2018) Integrated analysis of flow, form, and function for river management and design testing. *Ecohydrology*, 11(5), e1969.

Lane, S. & Richards, K. (1997) Linking river channel form and process: time, space and causality revisited. *Earth Surface Processes and Landforms*, é2(1), 249–260.

Laplace, P.S. (1820) *Théorie analytique des probabilités*. Vol. 7. : Courcier.

LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *nature*, 521(7553), 436. doi:10.1038/nature14539.

Leopold, L.B. & Wolman, M.G. (1957) *River channel patterns: braided, meandering, and straight*. : US Government Printing Office.

Li, S.H. & Wang, L. (2018) Neural network renormalization group. *Physical review letters*, 121(26), 260601.

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. doi:10.1007/s10955-017-1836-5.
URL https://doi.org/10.1007/s10955-017-1836-5

Lin, J. (1991) Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1), 145–151. doi:10.1109/18.61115.

Ling, F., Boyd, D., Ge, Y., Foody, G.M., Li, X., Wang, L. et al. (2019) Measuring river wetted width from remotely sensed imagery at the subpixel scale with a deep convolutional neural network. *Water Resources Research*, 55(7), 5631–5649.

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*,.doi:10.1002/esp.4206.
URL https://doi.org/10.1002%2Fesp.4206

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*.

Mao, X., Chow, J.K., Su, Z., Wang, Y.H., Li, J., Wu, T. et al. (2021) Deep learning-enhanced extraction of drainage networks from digital elevation models. *Environmental Modelling & Software*, 144, 105135.

Marchetti, G., Bizzi, S., Belletti, B., Lastoria, B., Comiti, F. & Carbonneau, P.E. (2022) Mapping riverbed sediment size from sentinel-2 satellite data. *Earth Surface Processes and Landforms*, 47(10), 2544–2559.

Martin, K.M., Wood, W.T. & Becker, J.J. (2015) A global prediction of seafloor sediment porosity using machine learning. *Geophysical Research Letters*, 42(24), 10–640.

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., et al. (2012) *NHDPlus Version 2: User Guide*. United States Environmental Protection Agency (EPA).

McManamay, R.A., Troia, M.J., DeRolph, C.R., Sheldon, A.O., Barnett, A.R., Kao, S.C. et al. (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. doi:10.1371/journal.pone.0198439.
URL https://doi.org/10.1371/journal.pone.0198439

Mehta, P. & Schwab, D.J. (2014) An exact mapping between the variational renormalization group and deep learning. *arXiv preprint arXiv:1410.3831*,.

Merritt, A.M., Lane, B. & Hawkins, C.P. (2021) Classification and Prediction of Natural Streamflow Regimes in Arid Regions of the USA. *Water*, 13(3). doi:10.3390/w13030380.
URL https://www.mdpi.com/2073-4441/13/3/380

Meyer, H. & Pebesma, E. (2022) Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1), 2208.

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.

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

Nardini, A., Yépez, S., Mazzorana, B., Ulloa, H., Bejarano, M.D. & Laraque, A. (2020) A systematic, automated approach for river segmentation tested on the magdalena river (colombia) and the baker river (chile). *Water*, 12(10), 2827.

Nearing, G.S. & Gupta, H.V. (2015) The quantity and quality of information in hydrologic models. *Water Resources Research*, 51(1), 524–538. doi:10.1002/2014wr015895.
URL https://doi.org/10.1002/2014wr015895

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M. et al. (2021) What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57(3), e2020WR028091.

Newendorp, P.D. (1976) *Decision analysis for petroleum exploration*. : Penn Well Books, Tulsa, OK.

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

Omernik, J.M. & Griffith, G.E. (2014) Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. *Environmental management*, 54(6), 1249–1266.

Pan, B., Hsu, K., AghaKouchak, A. & Sorooshian, S. (2019) Improving Precipitation Estimation Using Convolutional Neural Network. *Water Resources Research*,.doi:10.1029/2018wr024090.
URL https://doi.org/10.1029/2018wr024090

Papyan, V., Han, X. & Donoho, D.L. (2020) Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40), 24652–24663.

Peñas, F., Barquín, J., Snelder, T., Booker, D. & Álvarez, C. (2014) The influence of methodological procedures on hydrological classification performance. *Hydrology and Earth System Sciences*, 18(9), 3393–3409.

Poggio, T., Banburski, A. & Liao, Q. (2020) Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*,.

PRISM Climate Group (2004) *PRISM Gridded Climate Data*. Oregon State University.
URL http://prism.oregonstate.edu

Rabanaque, M.P., Martínez-Fernández, V., Calle, M. & Benito, G. (2021) Basin-wide hydromorphological analysis of ephemeral streams using machine learning algorithms. *Earth Surface Processes and Landforms*,.doi:10.1002/esp.5250.
URL https://doi.org/10.1002/esp.5250

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

Renard, K.G., Foster, G.R., Weesies, G., McCool, D. & Yoder, D. (1997) *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.

Rosset, S. Model selection via the AUC. In: *Proceedings of the twenty-first international conference on Machine learning*. *ACM, 2004*, p. 89.

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

Sahoo, S., Russo, T., Elliott, J. & Foster, I. (2017) Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. *Water Resources Research*, 53(5), 3878–3895.

Saxe, A.M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B.D. et al. (2019) On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12), 124020.

Schilling, A., Maier, A., Gerum, R., Metzner, C. & Krauss, P. (2021) Quantifying the separability of data classes in neural networks. *Neural Networks*, 139, 278–293.

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

Sejnowski, T.J. (2020) The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*,.

Sergeant, C.J., Falke, J.A., Bellmore, R.A., Bellmore, J.R. & Crumley, R.L. (2020) A classification of streamflow patterns across the coastal gulf of alaska. *Water Resources Research*, 56(2), e2019WR026127.

Shannon, C.E. (1948) A mathematical theory of communication. *Bell system technical journal*, 27(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

Shavitt, I. & Segal, E. (2018) Regularization learning networks: deep learning for tabular datasets. *arXiv preprint arXiv:1805.06440*,.

Shen, C. (2018) A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*,.doi:10.1029/2018wr022643.
URL https://doi.org/10.1029/2018wr022643

Shwartz-Ziv, R. & LeCun, Y. (2023) To compress or not to compress–self-supervised learning and information theory: A review. *arXiv preprint*

arXiv:2304.09355,.

Shwartz-Ziv, R. & Tishby, N. (2017) Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*,.

Singh, A., Fienberg, K., Jerolmack, D.J., Marr, J. & Foufoula-Georgiou, E. (2009) Experimental evidence for statistical scaling and intermittency in sediment transport rates. *Journal of Geophysical Research*, 114(F1). doi:10.1029/2007jf000963.
URL http://dx.doi.org/10.1029/2007JF000963

Sornette, D. (Ed.) (2006) *Critical Phenomena in Natural Sciences*. : Springer-Verlag.
URL https://doi.org/10.1007%2F3-540-33182-4

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. doi:10.1029/tr038i006p00913.
URL https://doi.org/10.1029/tr038i006p00913

Stuckelberg, E. (1953) La normalisation des constantes dans la theorie des quanta. *Helv. Phys. Acta*, 26, 499–520.

SWRCB (2017) *RUSLE K, LS, and R Factors Data and Methodology*. California State Water Resources Control Board.
URL https://ftp.waterboards.ca.gov/?u=GIS_Shared&p=GIS_Download&path=/swrcb/dwq/cgp/Risk/

SWRCB (2019) *Cannabis Cultivation Policy: Principles and Guidelines for Cannabis Cultivation. Sacramento, CA*. California State Water Resources Control Board.
URL https://www.waterboards.ca.gov/water_issues/programs/cannabis/docs/policy/final_cannabis_policy_with_attach_a.pdf

Tennant, C., Larsen, L., Bellugi, D., Moges, E., Zhang, L. & Ma, H. (2020) The utility of information flow in formulating discharge forecast models: A case study from an arid snow-dominated catchment. *Water Resources Research*, 56(8), e2019WR024908.

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. doi:10.1016/j.ecolind.2017.10.070.
URL https://doi.org/10.1016/j.ecolind.2017.10.070

Tishby, N. & Zaslavsky, N. Deep learning and the information bottleneck principle. In: *2015 IEEE Information Theory Workshop (ITW), Apr. 2015*. : IEEE.
URL https://doi.org/10.1109/itw.2015.7133169

Topsoe, F. (2000) Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4), 1602–1609. doi:10.1109/18.850703.

Turcotte, D.L. (1997) *Fractals and chaos in geology and geophysics*. : Cambridge university press.

Valentine, A. & Kalnins, L. (2016) An introduction to learning algorithms and potential applications in geomorphometry and earth surface dynamics. *Earth surface dynamics*, 4(2), 445–460.

Valentine, A.P., Kalnins, L.M. & Trampert, J. (2013) Discovery and analysis of topographic features using learning algorithms: A seamount case study. *Geophysical Research Letters*, 40(12), 3048–3054.

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(10), 1901–1923.

Walley, Y., Henshaw, A.J. & Brasington, J. (2020) Topological structures of river networks and their regional-scale controls: a multivariate classification approach. *Earth Surface Processes and Landforms*,.

Wolfe, J.D., Shook, K.R., Spence, C. & Whitfield, C.J. (2019) A watershed classification approach that looks beyond hydrology: application to a semi-arid, agricultural region in Canada. *Hydrology & Earth System Sciences*, 23(9).

Worland, S.C., Steinschneider, S., Asquith, W., Knight, R. & Wieczorek, M. (2019) Prediction and inference of flow duration curves using multioutput neural networks. *Water Resources Research*, 55(8), 6850–6868.

Yang, X.C., Xie, Z. & Yang, X.T. (2023) Exploring explicit coarse-grained structure in artificial neural networks. *Chinese Physics Letters*, 40(2), 020501.

Yoon, J., Zhang, Y., Jordon, J. & van der Schaar, M. (2020) Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33, 11033–11043.

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

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2016) Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*,.

Zhao, X. & Mendel, J.M. (1988) Minimum-variance deconvolution using artificial neural networks. In: *SEG Technical Program Expanded Abstracts 1988*Society of Exploration Geophysicists, pp. 738–741.

□

APPENDIX

A INFORMATION THEORY METRICS

Below we describe a set of established information theory metrics considered in this study and their relations: entropy, conditional entropy, mutual information, Kullback-Leibler divergence, Jensen-Shannon divergence and Jensen-Shannon distance.

Shannon's entropy describes the predictability of a random variable X with discrete probability mass function P over n outcomes (Shannon 1948):

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i) \tag{A1}$$

with b, the base of the logarithm function; when $b = 2$, information theory metrics have units of bit. If the distribution is biased towards a specific outcome, entropy is low. Conversely, entropy is maximum when all outcomes are equally probable. Following the rules of statistics, entropy can be conditioned on the distribution of another random variable Y. Then, conditional entropy, $H(X|Y)$, represents the uncertainty left in X after learning the outcome of Y(Shannon 1948):

$$H(X|Y) = -\sum_{i=1}^{n} P(y_i) \sum_{i=1}^{n} P(x_i|y_i) log_b P(x_i|y_i) \tag{A2}$$

From the definitions of entropy and conditional entropy stem mutual information, a measure of the degree of information shared between X and Y (Shannon 1948):

$$MI[X;Y] = H(X) - H(X|Y) \tag{A3}$$

where the right-hand side is the difference between the uncertainty in X before and after the outcome of Y becomes known. Mutual information is symmetric, $MI[X;Y] = MI[Y;X]$, and zero if X and $Y$are statistically independent.

The Kullback-Leibler divergence describes the mean information for discriminating between discrete probability distributions P and Q by observing P only (Kullback and Leibler 1951):

$$D_{KL}(P,Q) = \sum_{i=1}^{n} P(x_i) log_b \frac{P(x_i)}{Q(x_i)} \tag{A4}$$

Formally, the Kullback-Leibler divergence is the expectation of the logarithmic difference between discrete probability distributions P and Q with respect to probability distribution P. Because of this, the Kullback-Leibler divergence is asymmetric and, in non-trivial cases, $D_{KL}(P,Q) \neq D_{KL}(Q,P)$.

The Jensen-Shannon divergence is a measure of discrimination between two probability distribution functions and is directly related to the Kullback-Leibler divergence (Lin 1991, Topsoe 2000):

$$D_{JS}(P,Q) = \frac{1}{2}[D_{KL}(P,R) + D_{KL}(Q,R)] \tag{A5}$$

$$d_{JS} = D_{JS}^{1/2} \tag{A6}$$

with $R = \frac{1}{2}(P + Q)$ the midpoint probability. The Jensen-Shannon distance, $d_{JS} = D_{JS}^{1/2}$ retains the advantageous symmetric property of the Jensen-Shannon divergence, while satisfying the triangular inequality and being a proper distance metric (Endres and Schindelin 2003) which allows for constructing distance matrices, a common tool in data analysis (e.g., correlation matrix).

Received: 12 September 2023 | Revised: 13 July 2024 | Accepted: 22 August 2024

DOI: 10.1002/esp.5984

SUPPORTING INFORMATION

Supporting Information for "Mind the information gap: How sampling and clustering impact the predictability of reach-scale channel types in California (USA)"

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

[1]University of California Davis, Davis, CA, United States

[2]Utah State University, Logan, UT, United States

Correspondence
Corresponding author Hervé Guillon,
Email: herve@guillon.xyz

Contents of this file

1 | TEXT S1

Impact of spatial resolution on RF and DNN performance

In the following, we present a comparison of the performance of Random Forest (RF) and Deep Neural Network (DNN) when using terrain analysis predictors derived from topographic data at 10-m and 1-m resolutions, respectively. We focus hereafter on the North Central Coast (NCC, Figure 1) of California (USA) wherein both types of data are available along with channel types derived from Byrne et al. (2020). Using the available 1-m lidar data, we then re-calculate the 108 Terrain Analysis Metrics-Distribution Metrics (TAM-DM, Table 4). These predictors correspond to terrain analysis predictors (e.g., slope, curvature), summarized by various metrics (e.g., mean, standard deviation) over two spatial scales: a 512-m by 512-m tile centered on the labeled reach location and a 100-m riparian buffer along the streamline. All other predictors are kept the same between the two runs of this benchmark: `NCC` with 10-m TAM-DM predictors, and `NCC1m` with 1-m TAM-DM predictors. As we focus on comparing RF and DNN, the support vector machine model present in the main text was not trained for the `NCC1m`. Our machine learning (ML) framework is then carried out with predictor selection and nested resampling as described in the main text (Figure 2).

We now review the results of that comparison.

The evolution of the performance of ML models underscores the greater impact of finer-scale predictors on DNN than on RF (Figure S1). In particular, with an increasing number of predictors, RF performance appears roughly similar between using 10-m or 1-m TAM-DM predictors. Conversely, DNN's performance significantly increases when using 1-m TAM-DM predictors, outperforming baseline models (Figure S1). Both models achieve similar or better performance using 1-m data yet with fewer predictors (Table S1). For DNN, the Area Under Curve (AUC) of the optimal model increases from 0.92 using 27 predictors to 0.93 using 18 predictors. Similarly, for RF optimal model, AUC remains at 0.96 using 14 instead of 27 predictors.

For RF, the main impact of using finer-scale predictors is increasing the stability of the learning process (Figure S2). Here, the stability of the learning process is measured by deriving the entropy of the distribution of the best-tuned hyper-parameters resulting from the nested resampling (e.g., Figure S3). This measures the uncertainty related to the selection of the best-tuned

hyper-parameter(s). While there is little impact of the finer resolution TAM-DM predictors on DNN's tuning entropy, when using 1-m TAM-DM, RF tuning entropy is significantly improved. In particular, once reaching the maximum size of the hyper-parameter grid of 16, RF's tuning entropy appears to be plateauing with the 1-m data instead of increasing in the case of the 10-m data. This is further shown by the more constrained distribution of the best-tuned hyper-parameter(s) for the optimal RF and DNN models (Figure S3). Between 10 and 1-m data, DNN's tuning entropy changes from 0.80 to 0.82 while RF's tuning entropy decreases from 0.80 to 0.71 (Table S1).

Changing the resolution of the terrain analysis predictors, changes the importance of the predictors for predicting channel types (Figure S4). As expected, the selected predictors that were unchanged (i.e. non-TAM-DM) remain selected in the optimal RF model: valley confinement, Hurst coefficients and drainage area. Using 1-m data appears to filter out predictors based on curvature or Topographic Position Index (TPI). Importantly, of the 41 predictors removed for high correlation, only two were present in the optimal RF's set of predictors: standard deviation of TPI and maximum roughness. This means that mutual information alone filters out a significant number of TAM-DM predictors when using 1-m data. As a result, GIS-slope and contextual predictors appear higher on the variable importance list and the following terrain analysis metrics are promoted (Figure S4): median slope, median planform curvature, standard deviation of roughness, skewness of the topographic ruggedness index (TRI), standard deviation of elevation, standard deviation of planform curvature, minimum slope and maximum TRI.

FIGURE S1

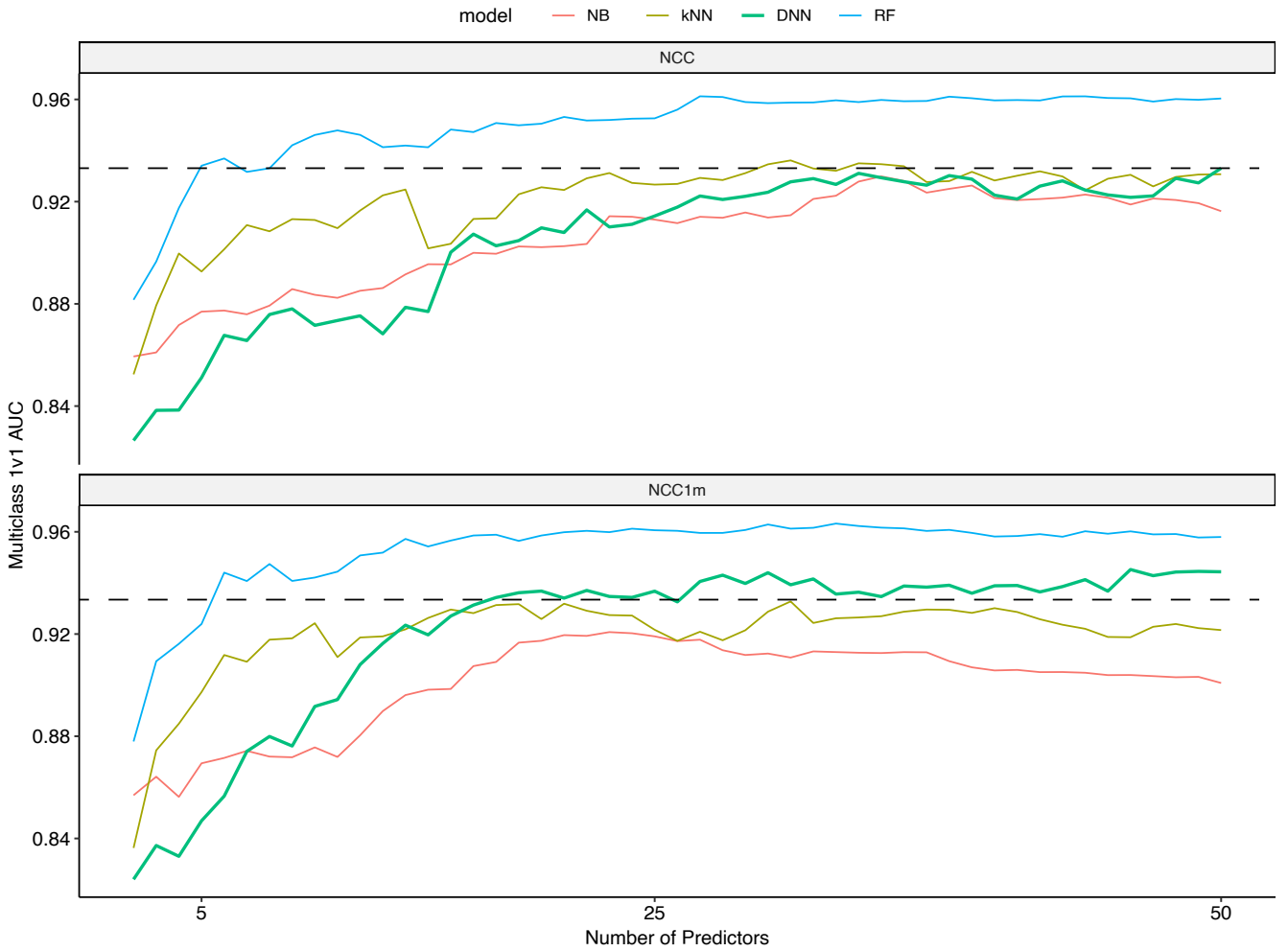

FIGURE S1 Evolution of the performance of ML models measured by multiclass 1v1 Area Under Curve (AUC) with an increasing number of predictors. The dashed line represents the maximum value for DNN's AUC in the `NCC` benchmark. The featureless model is not pictured: its AUC is constant at 0.5. NCC: North Central Coast; NCC1m: NCC with 1-m terrain analysis predictors. NB: Naive Bayes; kNN: *k*-Nearest Neighbors; DNN: Deep Neural Network; RF: Random Forest.

FIGURE S2

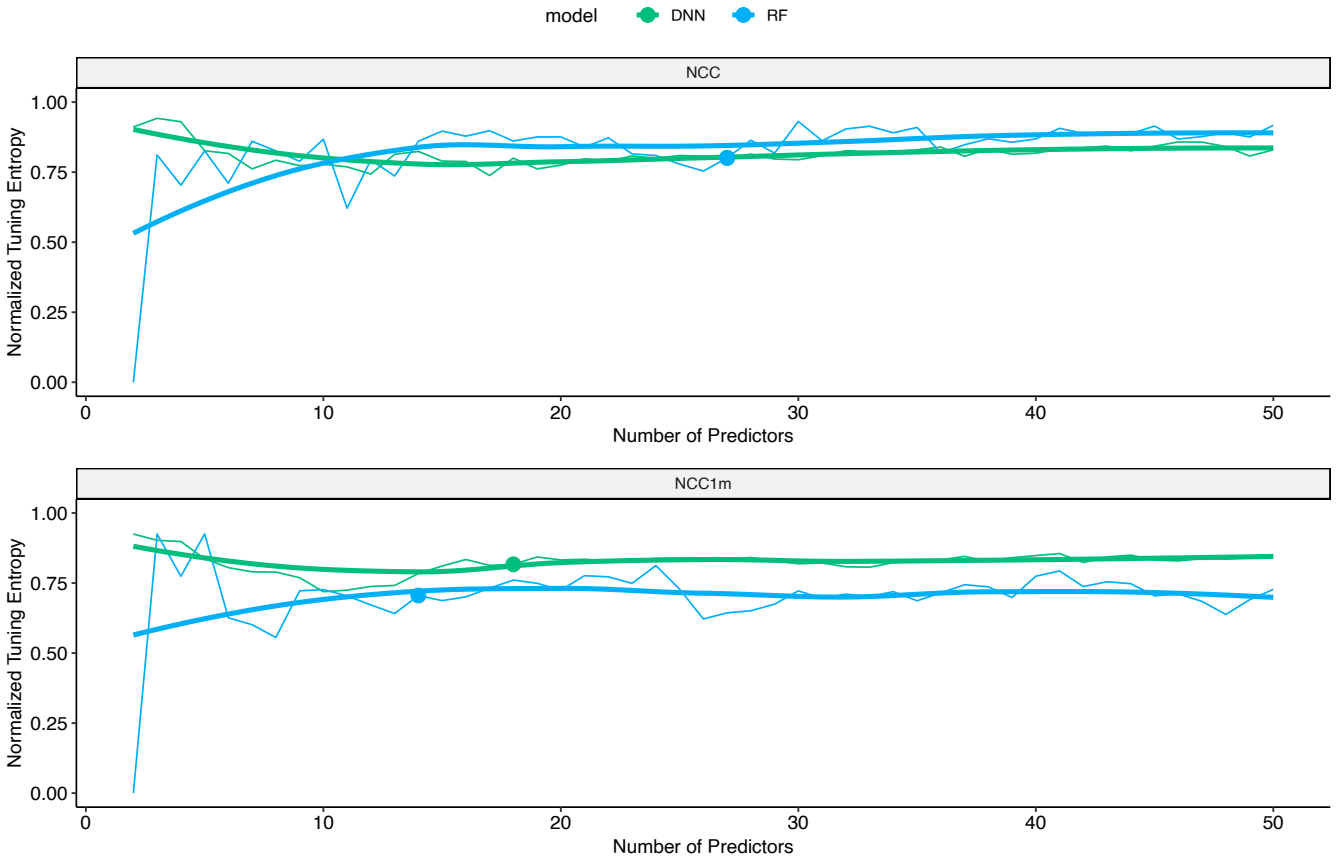

FIGURE S2 Evolution of tuning entropies with an increasing number of predictors. NCC: North Central Coast; NCC1m: NCC with 1-m terrain analysis predictors.

FIGURE S3

FIGURE S3 Distributions of best-tuned hyper-parameters for each region and each optimal ML model. a) Deep Neural Network; b) Random Forest. NCC: North Central Coast; NCC1m: NCC with 1-m terrain analysis predictors

FIGURE S4

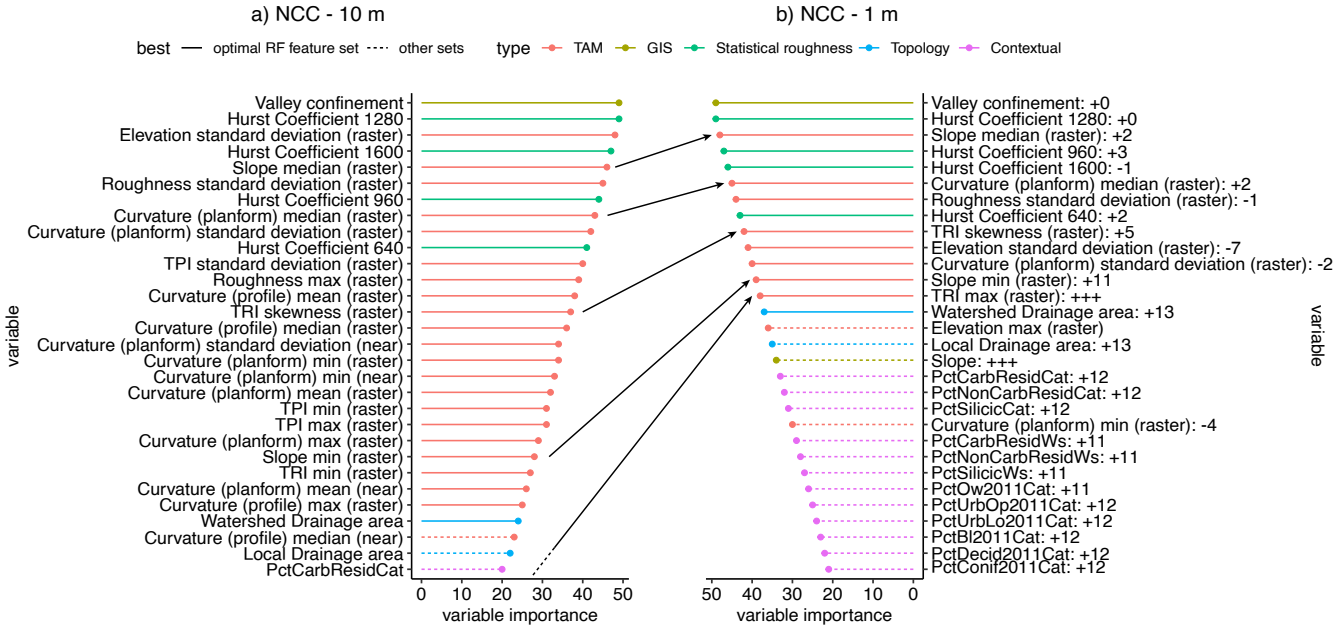

FIGURE S4 Variable importance derived from mutual information. TAM: Terrain Analysis Metrics predictors; GIS: GIS-derived predictors. TRI: Topographic Ruggedness Index; TPI: Topographic Position Index. The Hurst coefficient predictors represent a measure of statistical roughness at different spatial scale. The solid lines mark the predictors included in the optimal RF model. a) Results for NCC with 10-m data; b) results for NCC with 1-m terrain analysis predictors; the number in parenthesis after the variable name indicates its change in position between the two runs.

TABLE S1

Run	Model	Predictors	AUC	Accuracy	Training time	Normalized tuning entropy
NCC	DNN	27	0.92	0.66	1941	0.80
NCC	RF	27	0.96	0.76	65	0.80
NCC1m	DNN	18	0.93	0.69	1851	0.82
NCC1m	RF	14	0.96	0.74	37	0.71

TABLE S1 Summary table of the average performance of learners across all areas of study. Training time is given here in seconds for one iteration of the learning process and does not correspond to the total CPU-hours required for training.

REFERENCES

Byrne, C.F., Guillon, H., Lane, B.A., Pasternack, G.B. & Solis, S.S. (2020) *Coastal California Regional Geomorphic Classification: Final Report – Submitted to the California State Water Resources Control Board*. University of California, Davis.
URL https://watermanagement.ucdavis.edu/download_file/view_inline/509