A quantitative approach for comparing statistical classifications founded in machine learning and information theory

Hervé Guillon[1], Belize Lane[2], Colin Francis Byrne[3], Gregory Brian Pasternack[1], and Samuel Sandoval Solis[1]

[1]University of California, Davis
[2]Utah State University
[3]University of California Davis

November 22, 2022

Abstract

Statistical classifications and machine-learning-based predictive models are increasingly used for environmental data analysis and management. There now exist numerous classifications on the same topic but applied to different regions or spatial scales, such as geomorphic classifications. However, no quantitative meta-analysis framework exists to compare and reconcile across multiple classifications. To fill this gap, we jointly characterize statistical classifications and predictions by combining information theory and machine learning in three novel ways by: (i) measuring the degree of discriminatory information underlying a statistical classification; (ii) estimating the stability of the learning process with tuning entropy; and (iii) leveraging the sequential coarse-graining of information inherent to deep neural networks but absent from traditional machine learning models. This framework is applied through a benchmark of 59 millions models on a unique example of a single statistical classification methodology applied to nine different regions of California, USA. Regional results show that random forest consistently outperforms deep neural networks. In addition, a correlation analysis between regional characteristics, the level of discriminatory information of each classification, and the performance in statistical learning explains variations in performance and reveals the decisive role of the spatial scale of classification outputs. Because such a spatial scale is itself linked to the common situation of limited field sampling, directly comparing findings from statistical classifications and associated predictions appears seldom justified. A more desirable avenue to compare findings lies in combining data underlying statistical approaches in an interpretable and justifiable environmental data science.

A quantitative approach for comparing statistical classifications founded in machine learning and information theory

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

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

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

Key Points:

- The information ingrained in statistical classifications explains the performance of their machine-learning predictions
- The difference in traditional and deep learning performance prognosticates the minimum degree of information needed to separate classes
- Sampling trade-off occurs between capturing natural variability at a uniform level of detail and ensuring robust generalization

Corresponding author: Hervé Guillon, `hguillon@ucdavis.edu`

Abstract

Statistical classifications and machine-learning-based predictive models are increasingly used for environmental data analysis and management. There now exist numerous classifications on the same topic but applied to different regions or spatial scales, such as geomorphic classifications. However, no quantitative meta-analysis framework exists to compare and reconcile across multiple classifications. To fill this gap, we jointly characterize statistical classifications and predictions by combining information theory and machine learning in three novel ways by: (i) measuring the degree of discriminatory information underlying a statistical classification; (ii) estimating the stability of the learning process with tuning entropy; and (iii) leveraging the sequential coarse-graining of information inherent to deep neural networks but absent from traditional machine learning models. This framework is applied through a benchmark of 59 millions models on a unique example of a single statistical classification methodology applied to nine different regions of California, USA. Regional results show that random forest consistently outperforms deep neural networks. In addition, a correlation analysis between regional characteristics, the level of discriminatory information of each classification, and the performance in statistical learning explains variations in performance and reveals the decisive role of the spatial scale of classification outputs. Because such a spatial scale is itself linked to the common situation of limited field sampling, directly comparing findings from statistical classifications and associated predictions appears seldom justified. A more desirable avenue to compare findings lies in combining data underlying statistical approaches in an interpretable and justifiable environmental data science.

1 Introduction

Machine Learning (ML) is becoming increasingly prevalent in natural sciences because of its ability to identify and predict patterns in large and complex datasets (Shen, 2018; Bergen et al., 2019; Reichstein et al., 2019). In hydrologic sciences, this popularity leads to an increasing number of statistical classifications and ML predictions of patterns of hydrologic response and channel forms at regional, continental and global scales (Table 1). A statistical classification (e.g., hierarchical clustering) identifies latent patterns directly from data, and a predictive ML model (e.g., random forest, neural networks) estimates a relationship between data and already labelled patterns, or labels. In addition, statistical classification and predictive modelling may augment one another. For example, in Byrne et al. (2019) and Guillon et al. (2020), the labels outputted by a statistical classification were fed into a predictive model using more-readily available data (e.g., remote sensing) than the one used for classifying. Alternatively, in McManamay et al. (2018), the dataset used in classification was completed using predictive models.

Classification and pattern recognition are routine human activities in both science and management, but the ability to integrate across many studies to reveal underlying natural phenomena hinges on the compatibility of numerous and likely divergent methodological choices. There are many classification purposes, data types, approaches, and instances for the same environmental systems, yet science needs to synthesize and interpret that diversity and complexity to enable broader understanding and societal benefit beyond each original classification application. Hence, while the growing number of statistical classifications provides material for meta-analysis at increasingly larger scales, a quantitative approach to compare such statistical findings is missing, limiting the potential for scientific synthesis. As a result, meta-analysis dominantly relies on expert knowledge. For example, in a rare instance of comparison of geomorphic classifications, Kasprak et al. (2016) compared the results from three conceptual and one statistical approaches to classifying river channels within a single watershed of the Columbia River Basin, USA, and found generally comparable outputs. Kasprak et al. (2016) heuristically attributed the fuzzy correspondence across classifications to the geomorphic linkage between form and process (Davis, 1899) inherent to the empirical classifications tested, which differed in their required input data and spatio-temporal scale of their labels. Yet, both for empirical and statistical classifications, it is unclear how to integrate findings beyond one individual geographic area into knowledge in other regions or at larger scales. In addition, there has been no attempt to quantitatively compare one or multiple statistical classifications within or between study areas.

In this study, we develop a quantitative approach for comparing statistical classifications which, albeit developed in the context of fluvial geomorphology, is intended for use across all environmental sciences. The developed framework is capable of answering important scientific synthesis questions in the testbed context of fluvial geomorphology: why do different statistical classifications end up with different numbers of classes, and how does this relate to ML performance? We leverage nine statistical classifications of river channel forms, stemming from a single classification methodology (Byrne et al., 2019), and recently developed for nine distinct regions of California, USA. Our approach, rooted in information theory and machine learning, increases the interpretability of each individual classification and enables the direct comparison of distinct classifications established in different or the same region into an overarching unified classification to facilitate analysis and management. Specifically, we characterize each classification by its information content and by the performance of the associated predictive models. Study results yield general implications for the sampling strategies at the core of statistical classifications and subsequent predictions. The rest of the article is organized as follows. The next section introduces necessary background on information theory and machine learning. Section 3 details our case study and methods. Section 4 presents results and section 5 discusses their implications and limitations. Section 6 concludes by summarizing our findings.

2 Background on Information Theory and Artificial Intelligence

Since their inception in the 1950s, artificial intelligence (AI) and information theory have been closely related. In fact, at the 1956 Dartmouth Workshop on Artificial Intelligence, a landmark event for the foundation of AI, one of the attendants was Claude Shannon, the founder of information theory with his seminal paper "A Mathematical Theory of Communication" (Shannon, 1948). Broadly, information theory is concerned with estimating (and preserving) the statistical structure of a message communicated with noise and AI is concerned with mimicking (human) cognition processes. Seventy years later, both fields span a vast scope and, because of their shared roots in statistics and computer science, are present to some extent in most fields of science. In particular, machine learning (ML), the subfield of AI concerned with pattern recognition with self-improving algorithms (Michie, 1968), is increasingly popular.

Hydrological sciences have increasingly incorporated information theory and AI. For information theory, this is best exemplified by a recent series of articles debating its relation with statistical physics (Perdigão et al., 2020), inference (Nearing et al., 2020) and model parsimony (Weijs & Ruddell, 2020) and for AI, by a recent review of the hydrologic applications and associated challenges of deep learning (detailed below, Shen, 2018). Some specific examples of information theoretic approaches include evaluating causality in dynamical systems (Jiang & Kumar, 2019), identifying hydrologic response times (Tennant et al., n.d.), interpolating spatial data (Thiesen et al., 2020), and diagnosing the performance (Ruddell et al., 2019) and structure (Bennett et al., 2019) of physics-based hydrologic models. Furthermore, applications of deep learning are becoming pervasive in hydrology, for example to predict streamflow (e.g. Kratzert et al., 2019; Worland et al., 2019; Tennant et al., n.d.), downscale satellite products (e.g. Alemohammad et al., 2018) or model outputs (e.g. Pan et al., 2019), reconstruct historic flood (e.g. Bomers et al., 2019) and classify images (e.g. Ling et al., 2019).

Deep learning repeats and stacks the basic structure of an artificial neural network (LeCun et al., 2015). An artificial neural network is constituted by a succession of layers of connected neurons. Each neuron holds a weight, describing its connection with neurons in the next layer, and some form of activation, functionally combining inputs from neurons in the previous layer. The first, last and in-between layers correspond to input, output, and hidden layers, respectively. A shallow neural network has one hidden layer, 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 example in hydrologic sciences).

While deep learning can predict complex patterns (LeCun et al., 2015), its performance is only partially explained by how a deep neural network processes information. DNN can approximate any function (Cybenko, 1989; Hornik et al., 1989) without settling in local optima (Baldassi et al., 2016) or suffering from over-parameterization (Belkin et al., 2019; Geiger et al., 2019). Notwithstanding, a complete theoretical explanation of DNN's ability to generalize learned patterns is still missing (Zhang et al., 2016). Nonetheless, deep learning success is tied to the stacked architecture of DNN, which sequentially reverses the hierarchical generative process between output and input (H. W. Lin et al., 2017). In particular, the sequential processing of the data through the hidden layers optimally decouple dependent inputs, extract relevant information from noise and compress it to allow generalization (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017). Such information distillation, or the sequential coarse-graining of information from a microscopic scale to a macroscopic scale, led to decisive cross-pollination between statistical learning and statistical physics (H. W. Lin et al., 2017; Carleo et al., 2019).

Information distillation is absent from Support Vector Machine (SVM) and Random Forest (RF), two of the most used traditional machine learning methods (i.e. non-deep learning). SVM is a maximum margin classifier where the width of the margin between classes is defined by the distance between each class' closest points forming the support vectors of the class boundary (Cortes & Vapnik, 1995). While filtering the information inputted to SVM is a common practice, SVM itself omits explicit information distillation. RF is an ensemble of classification and regression trees (Breiman et al., 1984) and includes at each split of each tree an information selection process based on the Gini coefficient or on an information theory measure. The repetition of this information selection process in each tree of the forest is combined with internal bagging and leads to (mostly) uncorrelated trees which makes the ensemble decision process robust to noise and yields good generalization. While this repeated and random predictor selection explains the performance of RF when the training dataset is reduced, noisy or both (Fox et al., 2017), it is distinct from the information distillation present in DNN. In this study, we relate this distinction in information processing between deep learning and traditional ML to the information ingrained in the labels from statistical classifications

3 A Quantitative Approach for Comparing Statistical Classifications

To characterize each individual classification and allow for cross-comparison, we aim to evaluate (i) the degree of discriminatory information ingrained in each classification, that is the amount of information needed to separate class examples, and (ii) the performance of ML models. Then, we investigate the potential linkages between the degree of discriminatory information and ML model performance by performing a correlation analysis. For completeness, we also include in this correlation regional classification characteristics like the number of observations. The following sections detail the derivation of each variable included in the correlation analysis.

3.1 Statistical Classifications of California Channel Types

As a case study, we use independent channel classifications for nine regions of California (USA), a physiographically diverse state (Mount, 1995) with numerous integrated management challenges (e.g. Lane et al., 2018). The nine regions vary in terms of size, hydro-climate, physiography, and geology (Fig. 1, Table 2). Each region received a different intensity of sampling due to financial and logistical constraints, rareness of some natural conditions, and the limited remains of sufficiently natural sites for some channel types (Table 3). For example, small, low-order, unconfined streams in mountain meadows and on valley floors are ubiquitously plowed over for various land uses. Among selected sites for all regions, all observations were made using the same procedures by people trained together to yield standardized results. Sampling locations for each region were randomly selected in a stratified scheme aimed at capturing the existing natural variability in terms of four GIS-derived attributes: 10-m digital elevation model channel slope, valley confinement, drainage area and sediment supply. The channel classification for each region was made using the same analytical methodology (Byrne et al., 2019) involving hierarchical clustering mostly based on field-measured channel attributes (e.g. bankfull channel width and depth, width and depth variability, grain size metrics, and channel slope), but also including GIS-derived valley confinement and catchment area. In total, 1,110 observations were used to produce nine regional statistical classifications (Fig. 1, Table 3). The labels resulting from the classifications, the channel types, are identified and named in terms of valley confinement, bed morphology and sediment size (e.g. unconfined riffle-pool sand-bedded river, confined cascade/step-pool stream with boulders).

3.2 Estimating Information Needed to Discriminate between Classes

The degree of information inherent to each statistical classification relates to the information needed to discriminate between each label of each classification and is derived from information theory metrics. We detail below the relations between 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):

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

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):

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

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

$$MI[X;Y] = H(X) - H(X|Y)$$

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 & Leibler, 1951):

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

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 (J. Lin, 1991; Topsoe, 2000):

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

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 & Schindelin, 2003) which allows for constructing distance matrices; a common tool in data analysis (e.g. correlation matrix).

In the California channel classification case, for each regional classification and between each pair of its channel types, we calculate the average $\bar{d}_{JS}$ from the distributions of seven channel attributes measured in-situ: 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). Channel types with high $\bar{d}_{JS}$ are defined from more distinct underlying information, requiring on average a lower degree of information (i.e. less information) for discriminating between them. Conversely, channel types with low $\bar{d}_{JS}$ are defined from less distinct underlying information, requiring on average a higher degree of information (i.e. more information) to discriminate between them. Importantly, the channel attributes used to construct the Jensen-Shannon distance matrix excludes the four coarser-scale GIS metrics used to stratify the random sampling, even though two of them (area and confinement) were prominent in classifying some regions. In consequence, the $\bar{d}_{JS}$ represents here the degree of discriminatory information potentially missing from coarser-scale predictors. Each regional $\bar{d}_{JS}$ matrix is summarized by its median Jensen-Shannon distance, $\tilde{d}_{JS}$, representing the typical degree of information needed to separate class examples in a given regional classification.

3.3 Assessing Performance of Machine Learning Models

Given a regional channel classification created by identifying patterns from data mainly measured in-situ, ML prediction aims to assign labels to the remaining unsampled sites throughout the region, on the basis of coarser scale remote sensing and GIS data. Our three-tiered ML framework is based on Guillon et al. (2020) with the following modifications (Fig. 2). Predictors (Table 4) are selected prior to statistical learning by an information filter and 49 runs are performed with a number of predictors between 2 and 50. For each region, three baseline models (naive bayes, featureless and k-nearest-neighbor) are trained with default hyper-parameters and DNN, SVM and RF models are tuned with nested resampling. This allows for evaluating the stability of the tuning process and selecting an optimal number of predictors. We detail in the following how predictors are filtered and how ML model performance is measured.

3.3.1 Filtering Predictors

The inclusion of a high number of irrelevant predictors may lead to over-fitting, hindering a robust generalization. In the complex problem of predicting fine-scale labels with coarser-scale geospatial predictors, Guillon et al. (2020) selected groups of predictors using data complexity measures (Lorena et al., 2018). Here, we select individual predictors using a filtering method based on mutual information (Guyon & Elisseeff, 2003). Such a filter maximizes the relevance of the predictors used to identify the channel types based on the statistical relationship between predictor and channel types distribution. As the filtering is based on field-measured data at observation locations, it may be biased according to the observed distribution of channel types (Table 3). To address this issue and derive a more robust filter, the filtering is averaged over 500 iterations, each using 80% of the training data in a stratified subsampling scheme. This selects predictors that have the highest degree of statistical dependence with respect to the channel types distribution.

Filtering predictors with mutual information is algorithm-agnostic but maximizes predictor relevance without consideration for redundancy. Nonetheless, only perfectly correlated variables are truly redundant with no additional information gained by adding them. In fact, engineering new predictors from highly correlated but complementary predictors may increase class separation (Guyon & Elisseeff, 2003). Yet, the three ML models tested here have a different inherent relationship to the predictor space and might be impacted differently. SVM calculates its maximum margin at once for the entire predictor space and likely benefits from removing redundant predictors. In RF, each decision tree is built sequentially by comparing at each split a subset of individual predictors against a subset of observations. The ensemble decision process implicitly combines predictors while being able to robustly filter out irrelevant predictors. In DNN, multiple hidden layers of neurons act as latent predictors by combining input. In consequence, removing highly correlated but complementary predictors may impact DNN performance negatively by hindering the discovery of relevant latent predictors. Thus, to reduce near-perfect redundancy without potentially impacting DNN performance, predictors with correlation greater than 0.95 are filtered out before entering the information filter (Fig. 2).

3.3.2 Measuring Performance

We assess ML model performance in both statistical learning and predictive modeling. The performance in statistical learning is assessed by benchmarking ML models across the nine regions of study using area-under-curve (AUC) and hyper-parameter tuning entropy. The observations are balanced using Synthetic Minority Oversampling TEchnique (Chawla et al., 2002) and the input data are filtered for no-variance predictors, centered and scaled, and missing values are imputed with a median imputation. The tuning is discrete with length 16 for RF and SVM, and random with 100 iterations for DNN. DNN are trained during 20 epochs and with a batch number between 120 and 560 depending on the size of the training dataset for each region of study. The benchmark is performed with nested resampling that estimates the robustness of the tuning process and limits over-fitting by using two nested loops: an inner loop for model tuning and an outer loop for model selection (Bischl et al., 2012, Fig. 2). Here, the outer resampling is a 10-fold stratified cross-validation repeated 10 times, and the inner resampling is a 10-fold stratified cross-validation. While traditional resampling leads to a distribution of model performance, nested resampling additionally provides a distribution of best-tuned hyper-parameters. In consequence, in addition to AUC, the performance of the model is assessed by estimating the hyper-parameter tuning entropy from the distribution of their best-tuned hyper-parameters. AUC is preferred here to accuracy for its higher discrimination performance, its relation to class-separability and its suitability for limited dataset (Rosset, 2004; Huang & Ling, 2005; Ferri et al., 2009). Combined with 49 runs of predictor selection, these benchmark parameters lead to training 57,267,000 tuned models and 132,300 baseline models (6,510,000 per region).

For each region, the selection of the optimal model is based on the statistical differences between AUC distributions for different numbers of predictors. The selection is performed in a sliding window of 7 models, meaning that one model with $\mathcal{M}(n)$ with n predictors is compared to the following models $\{\mathcal{M}(n+1) \ldots \mathcal{M}(n+6)\}$. The statistical comparison is performed by a Dunn's test with a Bonferroni correction of the p-value to account for multiple comparisons. In consequence, a difference is considered significant if the test p-value is lower than $0.05/7 \simeq 0.007$.

Similar to Guillon et al. (2020), the performance in predictive modeling is assessed for each region at the network-scale using entropy rate and an expert evaluation of the geomorphic relevance of the predictions (Fig. 2). Entropy rate leverages the network structure of the predictions and estimates the stability of the predictions from the transition probabilities between each channel type. Such entropy rate prognosticates the prediction skill of a model (Stephenson & Dolas-Reyes, 2000; Roulston & Smith, 2002) and helps select models providing the best information (Daley & Vere-Jones, 2004; Nearing & Gupta, 2015). Both metrics are computed from predictions after a cross-validated multinomial calibration that corrects the potential distortion of posterior probabilities and improves model performance (DeGroot & Fienberg, 1983; Zadrozny, 2002; Niculescu-Mizil & Caruana, 2005).

3.4 Correlation Analysis

The previous subsections explain how the three main types of data characterizing each regional classification were obtained to enable cross-classification comparison. A correlation analysis then elucidates the potential linkages between variables describing each region, measures of the information ingrained in statistical classifications, and measures of the performance of traditional ML models and deep learning models. The regional variables included are: number of observations, area, observation density, number of classes, and number of confined classes. Confined channel types are likely the most difficult classes to correctly identify because their more limited spatial imprint is imperfectly captured by large scale geospatial predictors (Guillon et al., 2020). The degree of required discriminatory information $\bar{d}_{JS}$ in each regional classification is aggregated by a median value over all classes and by minima $\min\left\{\bar{d}_{JS}\right\}$ over all classes and confined classes only. The ML model performance metric variables include AUC, accuracy, hyper-parameter tuning entropy, entropy rate and the relative difference in performance between traditional and deep learning models. Both Pearson and Spearman correlation were performed on scaled data and yielded similar results. Because of the limited dataset ($n = 9$), we present the average of 500 correlations performed with a 80% subsampling.

4 Results

In the following section, we report results for: (i) the discriminatory information ingrained in each statistical classification; (ii) the performance of ML models in statistical learning and predictive modeling; and (iii) the correlation analysis between regional characteristics and ML model performance.

4.1 Information Needed to Discriminate between Classes

The median degree of information required to discriminate between classes, $\tilde{d}_{JS}$, is varied between the different regional classifications. An example of derivation of Jensen-Shannon distance is provided for the Sacramento region (Fig. 3) while being directly summarized for all nine Jensen-Shannon distance matrices (Table 5). The two regions requiring the least amount of discriminatory information are SJT and SFE. The two regions requiring the largest amount of discriminatory information are NC and SAC. In most regions,

d_{JS} decreases when considering confined channel types, indicating that these channel types require more information to be identified. Nonetheless, in seven regions, the minimum d_{JS} is not between two confined channel types (Table 5).

4.2 Performance in Statistical Learning and Predictive Modelling

RF outperforms other models in terms of AUC, even with an increasing number of predictors (Fig. 4). The performance of all models increases with additional predictors, but RF consistently displays a greater and faster increase in its performance. In NC, SC and SFE regions, additional predictors decrease the performance of the naive bayes model, suggesting the progressive inclusion of irrelevant or noisy predictors in the learning process. Furthermore, DNN significantly underperforms, only outperforming the default nearest neighbour baseline model in two of nine regions: SC and SFE.

Tuning entropy remains high for all models and increases with the number of predictors (Fig. 5). This effect is generally more marked for SVM or DNN than for RF. DNN's tuning entropy is high yet stable with respect to the number of predictors. For DNN, the tuning entropy is an average of the tuning entropies of its seven hyper-parameters (Fig. 6a). Interestingly, across all regions, the same trend is observed for these hyperparameters, pointing to a more stable learning process for DNN than for SVM or RF. The RF learning process appears relatively stable in most regions with exception of SCC and SECA (Fig. 6b). Tuning entropies are high and with similar values for both RF and SVM in SAC, SC, and SECA regions (Fig. 5). RF tuning entropy is clearly higher than SVM's one in the SCC region. RF tuning entropy is, however, clearly lower than SVM's one in K, NC, NCC, SFE and SJT regions. In all regions, RF tuning entropy rapidly increases with the initial addition of predictors before either reaching a plateau. However, after this initial increase, the evolution of RF tuning entropy with the number of predictors is nuanced. In K, NC, SAC, SECA, SFE and SJT regions, tuning entropy tend to then decreases with additional predictors, whereas it increases in NCC, SC and SCC regions. While maintaining a relatively high tuning entropy, the optimal RF models use in general a lower number of predictors than SVM or DNN (Table 6) and clearly outperforms the other two models in terms of AUC and accuracy (Fig. 4). In consequence, RF is selected for performing the predictions. Such performance in statistical learning of RF precludes using entropy rate and stream-segment entropy to select one final predictive model as done by Guillon et al. (2020).

RF predictions generally conform with expert-based expectations of the regional distribution of channel types (Fig. 7). Across all regions, valley confinement is most often selected as a predictor in the optimal RF models (Fig. 8). In more than half of the nine regions, the standard deviation of elevation, the statistical roughness of topography at short spatial scales (Hurst coefficients), median slope and curvature metrics are selected as relevant predictors. Drainage area at the watershed and stream interval scales appears relevant albeit only in less than half of the regions. Contextual predictors only appear in the optimal set of predictors in SC region where, after valley confinement and drainage area metrics, they correspond to nine predictors describing lithology (6) and land use (3).

4.3 Correlation Analysis

Correlation analysis revealed that different classifications yield a varying number of channel types and a varying ML performance as a result of a few sampling design factors (Tables 2,5-6; Fig. 4,9). For regional characteristics, the number of channel types is strongly linked to the number of observations ($r = 0.90$). As expected, observation density is negatively correlated with catchment area ($r = -0.62$), but the number of channel types and the number of observations are inconclusively linked to area and observation density. Area and observation density are negatively correlated with the number of confined channel types ($r =$

-0.47 and $r = 0.65$, respectively). The observation density and area are not correlated with the statistical learning performance for DNN or RF (Fig. 9a-b). Instead, statistical learning performance metrics are anti-correlated with the number of observations, the number of channel types and the number of confined channel types. They are also positively correlated with one another. Similarly, all variations of d_{JS} are positively correlated with one another. All d_{JS} metrics positively correlate with statistical learning performance metrics, more so with AUC. However, they are negatively correlated with the number of observations and the number of classes.

The less information needed to separate classes, the more stable the label predictions are. Entropy rate is generally anti-correlated with d_{JS} metrics, especially with the minimum d_{JS} for confined channel types ($r = -0.80$), and weakly correlated with regional metrics increasing the complexity of the classification task: number of observations, number of channel types and number of confined channel types. For RF, entropy rate and hyper-parameter tuning entropy are only weakly linked ($r = -0.32$, Fig. 9b) and both are weakly anticorrelated with statistical learning performance metrics. Hyper-parameter tuning entropy appears mostly disconnected from statistical learning performance metrics ($r = -0.10$, $r \simeq 0$). In general, hyper-parameter tuning entropy shows weak correlation with the other variables with exception of the minimum d_{JS} for confined channel types ($r = 0.48$) and the number of confined channel types ($r = -49$). This suggests that hyper-parameter tuning entropy increases with decreasing complexity while entropy rate increases with increasing complexity.

The difference in performance between traditional ML models and deep learning models prognosticates the required degree of discriminatory information (Fig. 9c). The correlation of the statistical learning metrics are inverted with respect to DNN and RF correlations (Fig. 9a-b). The difference in statistical learning performance between RF and DNN correlates with the number of observations, the number of channel types and the number of confined channel types while being negatively correlated with all of the d_{JS} metrics quantifying the degree of discriminatory information needed to separate the channel types, in particular minimum d_{JS}.

5 Discussion

5.1 Discriminatory Information Explains why RF Outperforms DNN

Our proposed framework characterizes each individual classification by estimating the typical amount of information needed to discriminate between classes. This measure provides major insights into the meaning of the labels derived from statistical classifications, importantly helping their interpretation and comparison. In particular, the difference in discriminatory information (Table 5) is interpreted as differences in the scale at which the labels are inherently defined between the different regional classifications. In the application to statistical classifications of channel types in California, we suggest that the degree of discriminatory information is linked to the scale mismatch between labels and geospatial predictors and explains deep learning under-performance.

Even when selecting predictors inputted to ML models, RF outperforms SVM and DNN. Filtering the predictors makes for a fair benchmark for SVM which does not include any predictor selection process like RF or DNN. Yet, the tuning entropy results (Fig. 6) underline that SVM exhibits a noisy statistical learning without one defined value for its hyper-parameter. Nonetheless, the SVM included in the benchmark is a linear SVM and a kernelized SVM may display a better performance by being able to capture non-linear patterns.

As in Guillon et al. (2020), DNN underperforms relative to other ML models in most regions (Fig. 4) while exhibiting stable statistical learning with a more defined choice of hyper-parameters than RF or SVM, and across all regions (Fig. 6). Two combined reasons explain DNN under-performance. First, in general, DNN performance increases with the number of available observations, and the current data deluge helps explain their increasing popularity (LeCun et al., 2015). Second, the performance of DNN is tied to information distillation through the successive layers of the networks, which filters out irrelevant information from the input to predict the output. In the California channel classification case, the datasets are limited in size (Table 2) and the output labels are defined from field scale data ($<$ 200 m) while the input predictors are defined at a coarser scale ($>$ 500 m, Table 4). This scale mismatch corresponds to missing or overly noisy information, which precludes efficient information processing in the DNN and the reverse engineering of the hierarchical generative process between input and output (Tishby & Zaslavsky, 2015; H. W. Lin et al., 2017). The effect of the scale mismatch is exemplified by the correlation between required discriminatory information from the datasets used to generate the labels and the performance of DNN relative to RF: the coarser the label, the lower the difference between RF and DNN (Fig. 9c). With additional observations, DNN information distillation is likely to better filter out noisy information, reducing the gap in performance between RF and DNN. However, in the case of limited and noisy datasets with potential scale mismatch in the definition of labels and predictors, a common issue in environmental sciences, it is likely that RF-inspired algorithms will consistently outperform DNN-inspired algorithms.

An objective constraint on the specific scale of the set of statistical classification labels, such as found in this study from discriminatory information (Fig. 3) and deep learning relative performance (Fig. 9c), is likely beneficial for a wide variety of classifications across the hydrologic 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 applied to 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. Having more explicitly defined scales associated with statistically derived labels would yield a more universal lexicon and facilitate a better understanding of eco-physical processes intertwined with spatio-temporal patterns represented by labels. For classifications based on time series analysis, we suggest that the information content of the core data (i.e. temporal resolution, number of stations) defines the spatial or temporal scale of the resulting labels. Interestingly, our correlation analysis shows that discriminatory information, or scale, is equally evaluated either from the data used for the classification (Fig. 3, Table 5) or from the relative performance of traditional and deep learning approaches (Fig 9c). Below, we further discuss the main implications of our results in terms of limitations, analysis and management implications.

5.2 Limitations

While statistical learning performs well to estimate patterns between channel types and predictors (Fig. 4), generalizing the learned pattern in predictive modeling and assessing the geomorphic relevance of the resulting predictions lead to implementing a post-hoc heuristics to predictions in one of nine regions. The geomorphic relevance is qualitatively assessed by comparing the predicted spatial distributions of channel types with their expert-based expected spatial distribution. In the K region, the mainstem channel type K03 is hardly predicted to occur in the mainstem where it ought to and a stream-order-based heuristic was implemented.

Limited sampling mainly explains the need for a post-hoc heuristic to conform ML predictions with expert-knowledge expectations in K region. Channel types exist in significantly different natural abundances, with some types quite rare and difficult to isolate in the sampling scheme, and other types so anthropogenically impacted as to be all but unavailable to sample despite their potential importance for aquatic and riparian ecology. In all regions, the unequal sampling of channel types (Table 3), that is the imbalance in the classification labels, is addressed with the commonly used SMOTE (Chawla et al., 2002) which generates synthetic observations and helps the statistical learning of channel types with a lower number of examples. However, the random generation of synthetic observations is handled with a k-Nearest Neighbour algorithm (with in our case $k \leq 5$ depending on the number of available observations). In consequence, fewer field observations of a channel type lead to less diversity in the corresponding synthetic data, hindering a robust learning of the patterns between under-sampled channel types and predictors. In the K regions, the mispredicted channel type has the lowest possible value for prevalence, 1 over 105 observations, and thus for the diversity in the associated synthetic data (Table 3). The next lowest value of prevalence, 4, appears high enough across all regions to enable robust pattern learning when compared to expert evaluation. Interestingly, the most of the under-sampled channel types fall into two categories tied to the logistics of in-situ sampling: high-order main stem rivers and low-order steep cascade/step-pool channels. High-order main stem rivers are often highly channelized and far from natural conditions while displaying dimensions and water depth 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 most accessible channels.

5.3 Implications

The results of this study have some general implications for the sampling strategies at the core of statistical classifications and subsequent predictions. Our correlation analysis underlines a positive correlation between the number of field observations and the number of classes and a negative correlation between the number of classes and the required discriminatory information. This then translates into better ML performance. This is somewhat paradoxical yet explicable. With fewer observations, the likelihood of finding statistically significant groupings decreases resulting in fewer classes, which can be separated with coarser information, reducing the complexity of the problem. In other words, with fewer observations, one can get away with a simpler, coarse-scale problem to solve. An increasing number of observations leads to fine-scale labels, at least for some channel types, and to a more complex problem including mismatched scales. For example, riffle-pool reaches are so common that random sampling is likely to oversample them (even with our effort to stratify sampling using four meaningful catchment scale variables), yielding more variety of riffle-pool reach channel types. In contrast, there could be an equal diversity of cascade reach types, but if there are fewer of these sites in the geographical study area or there are randomly fewer sampled, then the classification is more likely to lump them together into one class. Thus, the outcome can be mismatched scales of classification between broader channel types, and even when statistical learning performs well, generalizing the learned pattern beyond the training datasets may be hindered by insufficient sampling. Consequently, there exists a sweet spot between sampling enough to capture some of the natural variability in the study area at a uniform level of detail across broad channel types and sampling more but not enough to ensure that a generalizable pattern is learned across all broad channel types to yield an equivalent diversity of fine-scale labels. This is likely an ubiquitous problem in natural sciences where classification and prediction contend with a mix of rare and common types, multi-scalar typologies, uneven anthropogenic disturbance across types, limited sampling capability, and high uncertainty in design of the sampling strategy. The characterization of this sampling optimum is beyond the scope of this study as it likely depends on the definition of the statistical classification which then conditions the performance in statistical learning.

The increasing popularity of statistical classifications begs the question of how one would compare them. Our application of one statistical classification methodology to multiple regions indicates that it is not a straightforward task. This likely remains true when comparing different statistical classification methodologies in a single region of interest. In particular, the information content of the classification, interpreted as the overall spatial scale at which channel types are defined, vastly differs, impacting performance in statistical learning and predictive modeling. This hinders direct comparison between statistical classification outputs, as the statistical classification results in a fuzzy correspondence akin to the loose agreement between empirical classifications of channel types (Kasprak et al., 2016). A better strategy to robustly compare areas of study or combine results from statistical classifications is to assemble a dataset spanning geographical areas and perform a new bottom-up statistical classification pooling all data into one set. This better ensures that labels 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 statistical classification are reasonably similar. All data also has to be publicly available from their authors, further bolstering reproducible, open, transparent, interpretable and justifiable environmental data science (Murdoch et al., 2019; Yu & Kumbier, 2020).

6 Conclusion

Machine learning is becoming more prevalent to inform decision-making, in particular with the increasing popularity of machine-learning-enabled classifications and predictions in environmental sciences. In this study, we thoroughly investigated the previously unexplored robustness of the statistical learning process and evaluated the often unknown spatial scale of statistical classification outputs. Our proposed approach combines information theory and machine learning in three novel ways: (i) measuring the degree of discriminatory information underlying a statistical classification; (ii) estimating the stability of the learning process with tuning entropy; and (iii) leveraging the sequential coarse-graining of information inherent to deep neural networks but absent from traditional machine learning models. While applied to a unique example of a single statistical classification framework applied to nine distinct regions of California, the developed approach is relevant for numerous classification and prediction problems in environmental sciences and underlines the importance of limited sampling on classification outputs and associated predictions. Importantly, the approach characterizes and compares different statistical classifications, providing an estimate of the spatial scale of their outputs and paving the way for a reconciliation of findings across or within study areas. In addition, we found that the difference in traditional and deep learning performance identifies the minimum degree of information needed to separate classes, providing a proxy for spatial scale.

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 CA-D-LAW-7034-H. Data sources are reported in Table 4. 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`.

7 Tables

Reference	Target	Classification data	Prediction data
McManamay et al. (2018)	physical habitat diversity	6 discrete-valued physical habitat layers	–
McManamay et al. (2018)	reach-scale hydrologic class	–	land cover, climate, topography, soils
McManamay et al. (2018)	summertime temperature	–	land cover, climate, topography, soils
McManamay et al. (2018)	mean substrate diameter	–	land cover, climate, topography, soils
McManamay et al. (2018)	bankfull width	–	land cover, climate, topography, soils
Wolfe et al. (2019)	watershed hydrologic class	climate, geology, topography, land cover	–
Yang et al. (2019)	surface-groundwater interactions	–	topography, hydrology, geology, land cover
Henshaw et al. (2019)	channel form	channel dimensions, morphological features	–
Beechie & Imaki (2014)	channel pattern	–	topography, hydrology, land cover
Clubb et al. (2019)	geomorphic domains	river profiles	–
Lane, Dahlke, et al. (2017)	reach-scale hydrologic class	topography, geology, climate	topography, geology, climate
Lane, Pasternack, et al. (2017)	channel types	field-measured attributes, topography	–
Byrne et al. (2019)	channel types	field-measured attributes, topography	–
Guillon et al. (2020)	channel types	–	topography, climate, geology, land cover
Sergeant et al. (n.d.)	watershed hydrologic class	daily streamflow statistics	–
Gaucherel et al. (2017)	watershed hydrologic class	topography, land cover, network topology	–
Dallaire et al. (2019)	river types	hydrology, climate, topography	–
Flores et al. (2006)	channel types	–	topography, hydrology, climate
Walley et al. (2020)	watershed river networks	network topology	–

Table 1: Recent examples of statistical classifications and machine-learning predictions in hydrologic sciences.

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 2: Regional characteristics. The number of confined channel types is reported between parenthesis.

	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 3: Distribution of observations across all regions.

Predictor name	Spatial scale	Original data	Methodology
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)
Channel slope	200 m	Gesch et al. (2002)	ESRI (2016)
Confinement	-	Gesch et al. (2002)	Byrne et al. (2019)
Sediment supply	-	Haan et al. (1994)	Renard et al. (1997)
Drainage area	-	McKay et al. (2012)	Hill et al. (2015)
Strahler's stream order	-	McKay et al. (2012)	Strahler (1957)
Local drainage density	-	McKay et al. (2012)	Danesh-Yazdi et al. (2017)
Hurst coefficients	640 m to 82 km	Gesch et al. (2002)	Liucci & Melelli (2017)
Lithology	>1 km	Cress et al. (2010)	Hill et al. (2015)
Soil characteristics	1 km	Schwarz & 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 4: 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.

Region ID	Median JSd	Mean JSd	Minimum JSd
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 5: Degree of discriminatory information estimated from the Jensen-Shannon distance. Values for confined channel types are reported between parenthesis.

Model	Predictors	AUC	Accuracy	Training time	Normalized tuning entropy
DNN	30	0.931	0.668	4.510	0.792
RF	18	0.949	0.740	0.290	0.757
SVM	31	0.943	0.743	0.366	0.844

Table 6: 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.

8 Figures

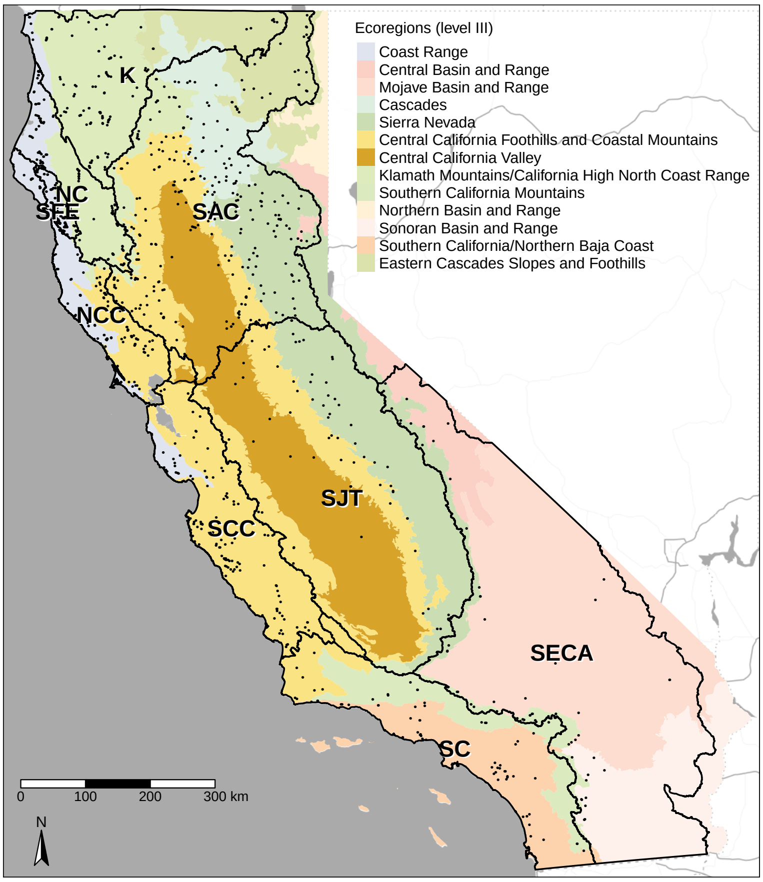

Figure 1: Field sites location in California (USA) across nine distinct regions. Ecoregions are displayed as a proxy combining geology, soils, vegetation, climate, and hydrology Omernik & Griffith (2014). Detailed information about each region is in Table 2.

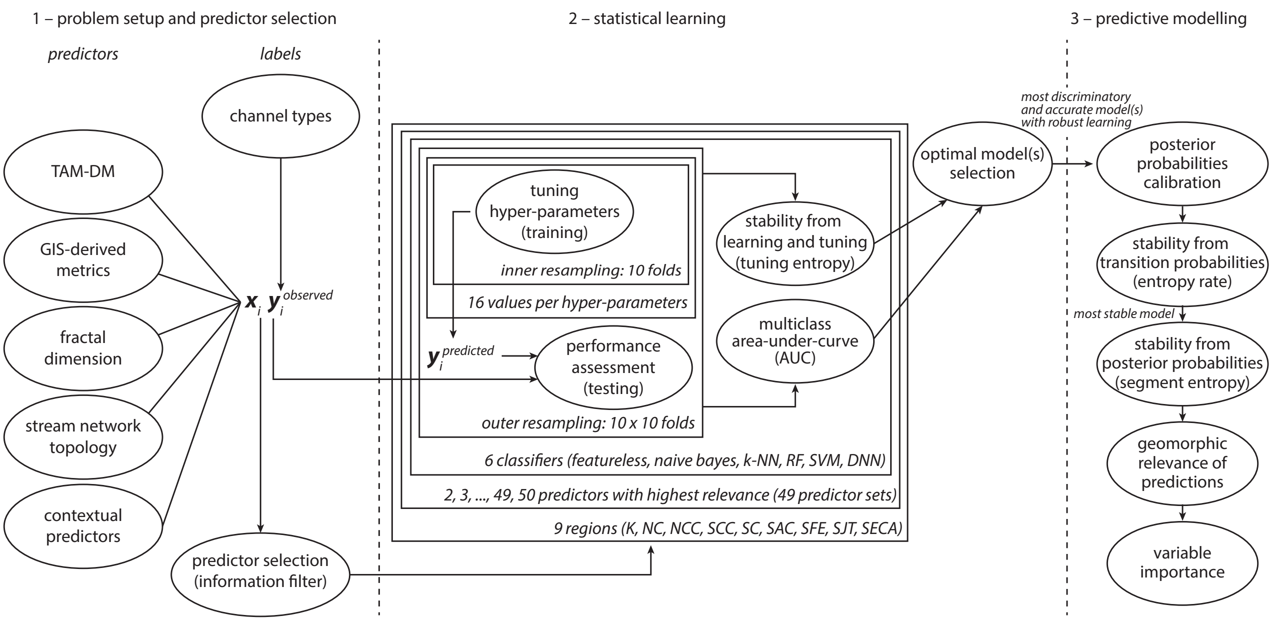

Figure 2: Schematic of the machine-learning framework.

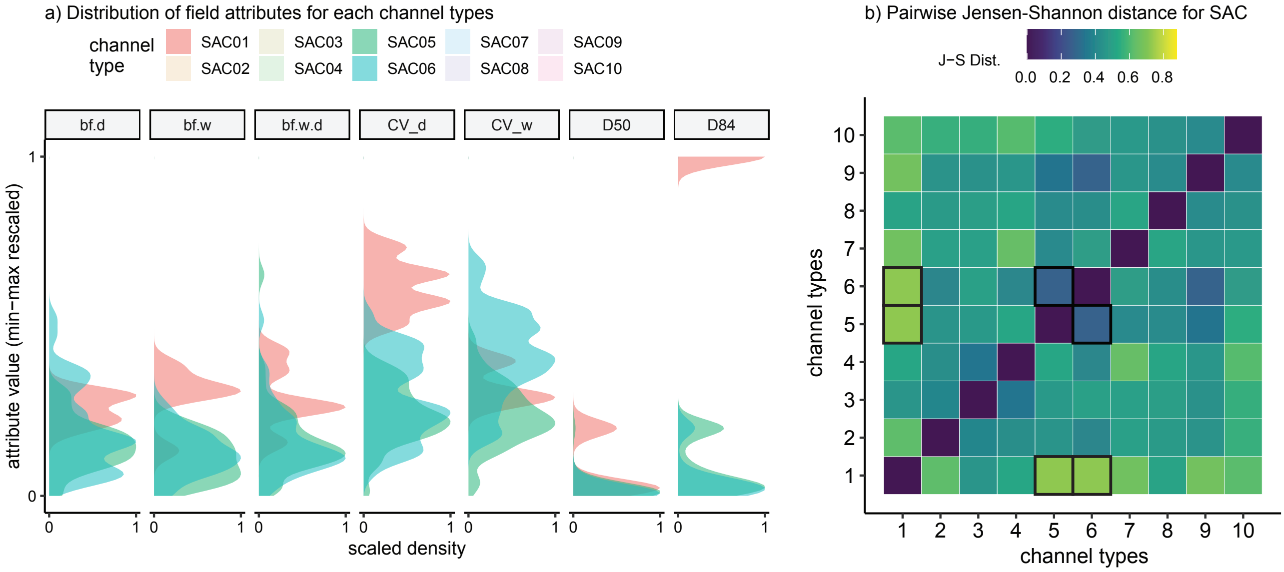

Figure 3: Example of Jensen-Shannon distance derivation

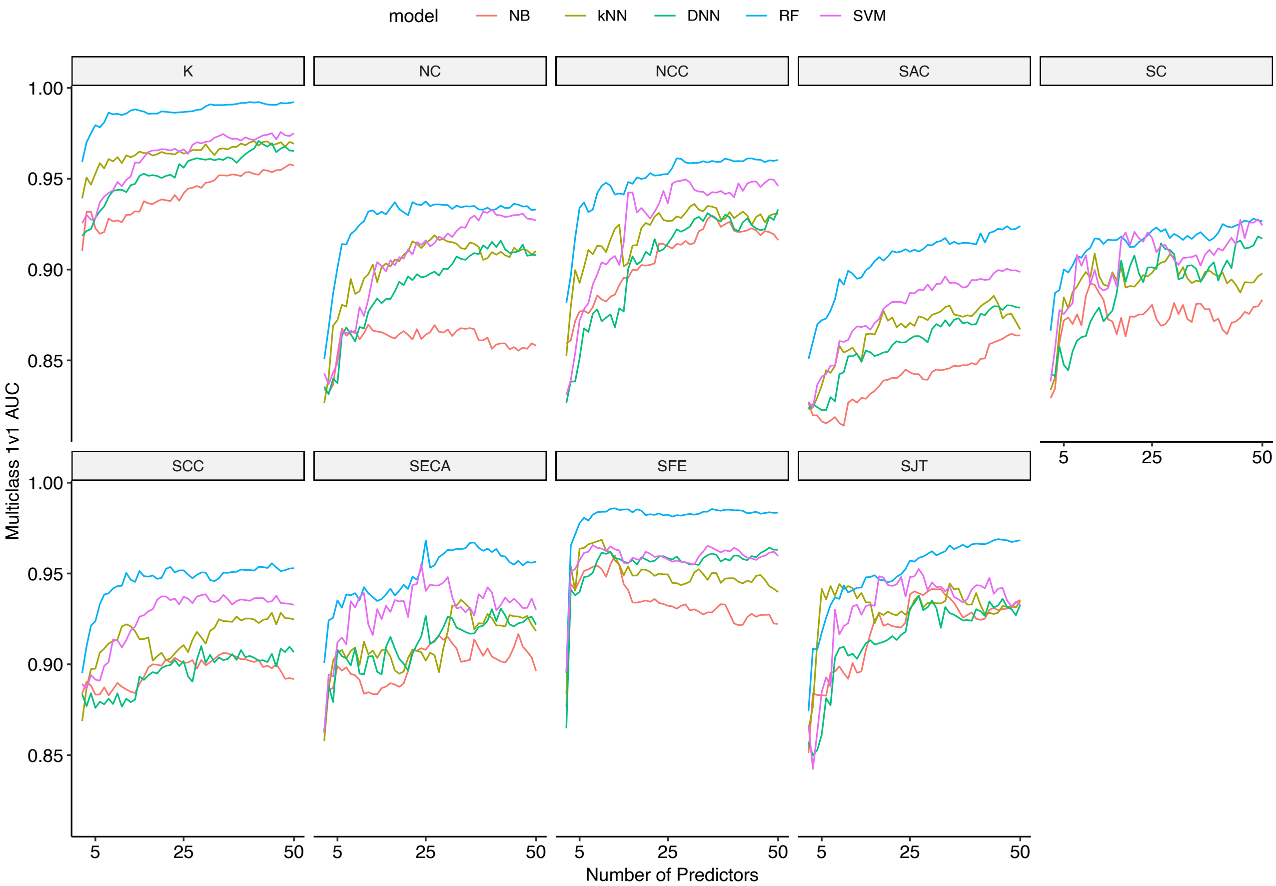

Figure 4: Performance of ML models. The featureless model is not pictured: its AUC is constant at 0.5.

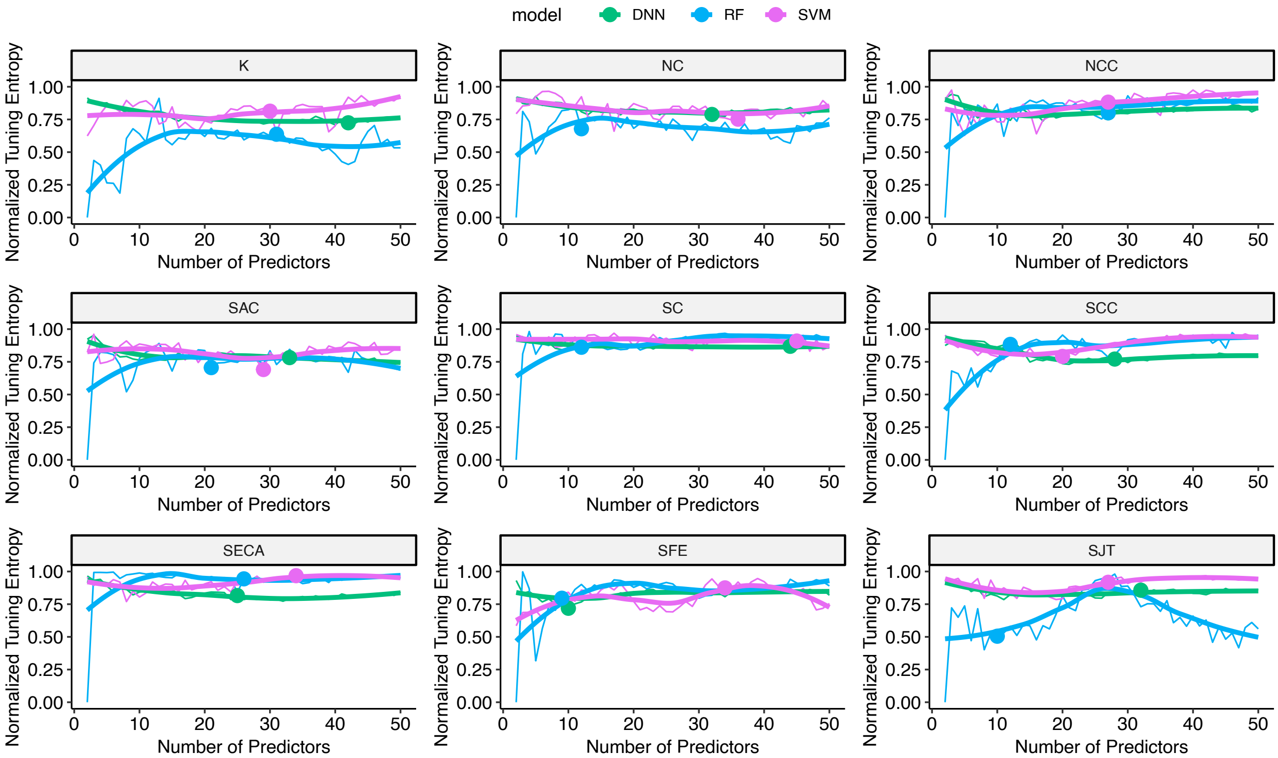

Figure 5: Evolution of tuning entropies with the number of predictors

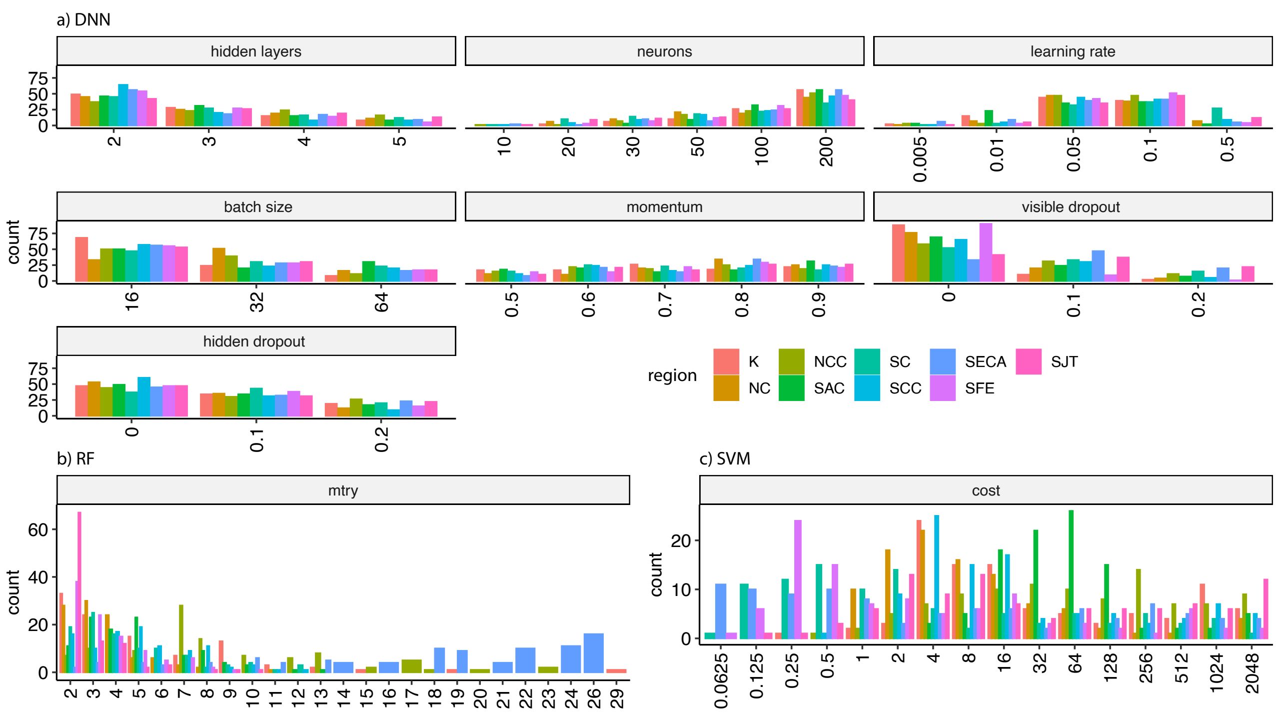

Figure 6: Tuning entropy for each region and each optimal ML model

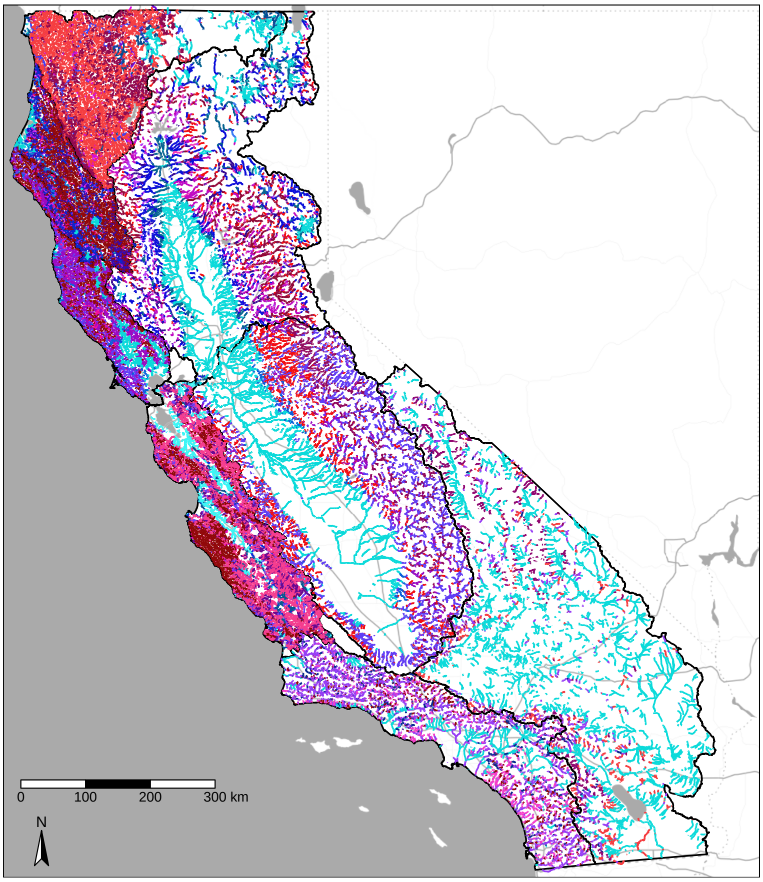

Figure 7: Map of all predictions. In each region, hue maps to confinement so that cyan (red) corresponds to the most unconfined (confined) channel type, and lightness maps to slope so that the channel type with low (high) slope are drawn in lighter (darker) colors.

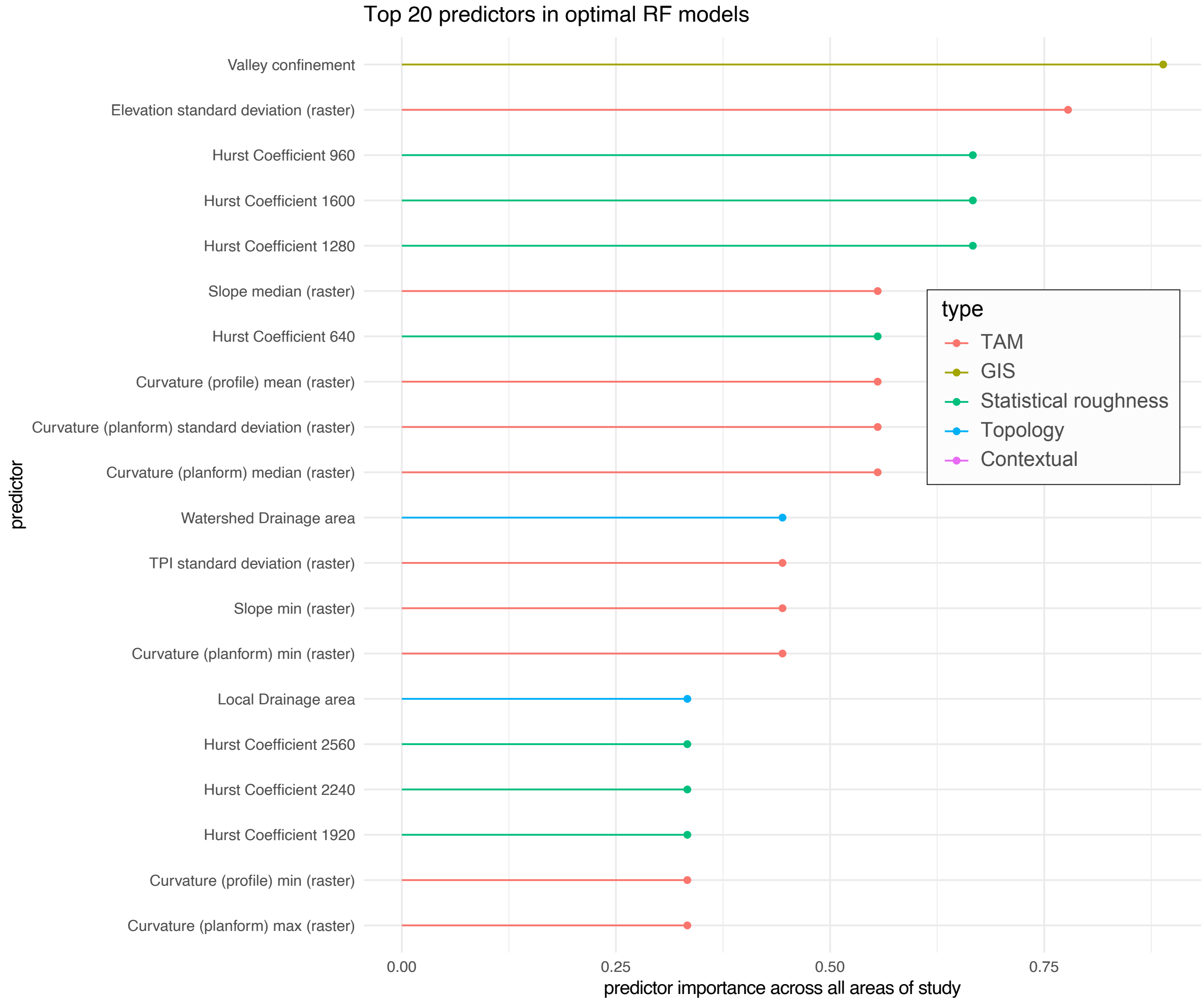

Figure 8: Regional variable importance

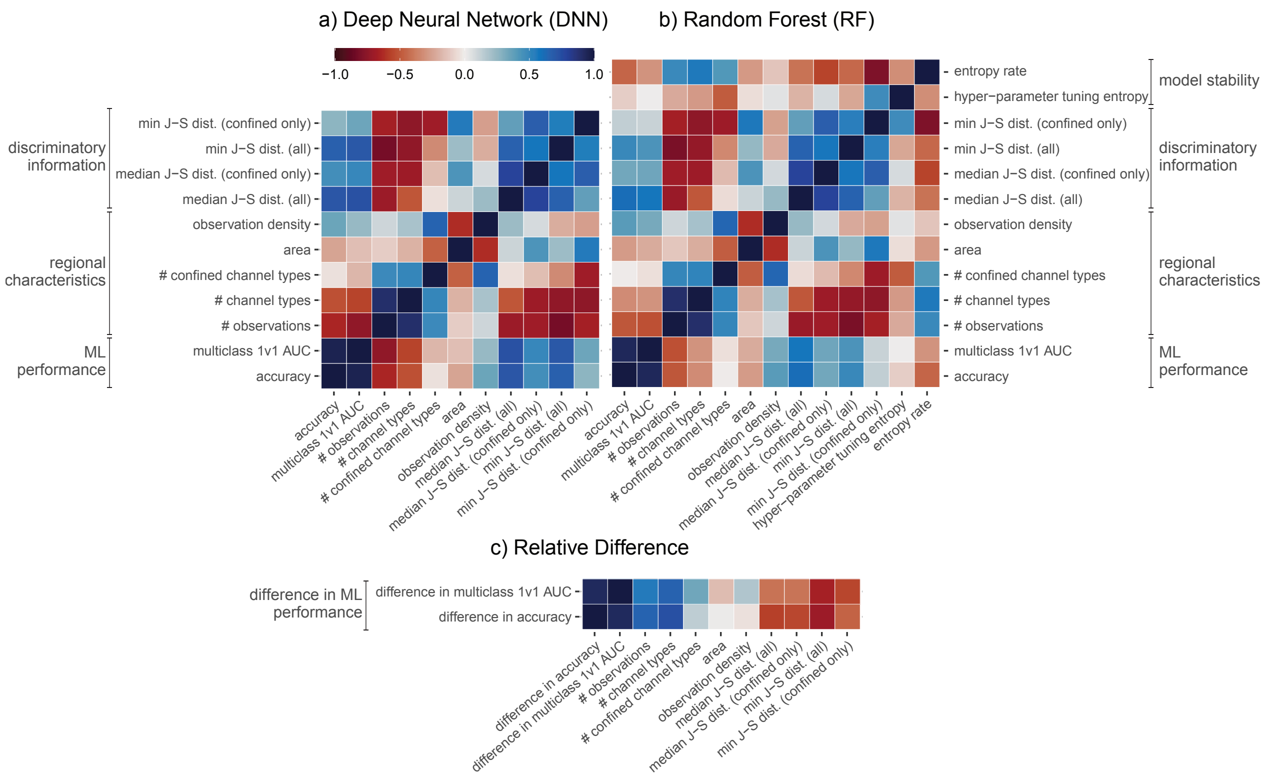

Figure 9: Correlation matrix for a) DNN; b) RF and c) the relative difference between DNN and RF.

References

Alemohammad, S. H., Kolassa, J., Prigent, C., Aires, F., & Gentine, P. (2018, October). Global downscaling of remotely sensed soil moisture using neural networks. *Hydrology and Earth System Sciences*, *22*(10), 5341–5356. Retrieved from `https://doi.org/10.5194/hess-22-5341-2018` doi: 10.5194/hess-22-5341-2018

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

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.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019, July). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854. Retrieved from `https://doi.org/10.1073/pnas.1903070116` doi: 10.1073/pnas.1903070116

Bennett, A., Nijssen, B., Ou, G., Clark, M., & Nearing, G. (2019). Quantifying process connectivity with transfer entropy in hydrologic models. *Water Resources Research*, *55*(6), 4613–4629.

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.

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012, June). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Com-*

putation, *20*(2), 249–275. Retrieved from https://doi.org/10.1162/evco_a_00069 doi: 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.

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

Byrne, C. F., Pasternack, G. B., Guillon, H., Lane, B. A., & Sandoval-Solis, S. (2019). Reach-scale bankfull channel types can exist independently of catchment hydrology. *Earth Surface Processes and Landforms*.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., … Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4), 045002.

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

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

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

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

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

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

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, oct). Self-dissimilar landscapes: Revealing the signature of geologic constraints on landscape dissection via topologic and multi-scale analysis. *Geomorphology*, *295*, 16–27. Retrieved from `https://doi.org/10.1016%2Fj.geomorph.2017.06.009` doi: 10.1016/j.geomorph.2017.06.009

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

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

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

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

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

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, jan). Accuracy of local topographic variables derived from digital elevation models. *International Journal of Geographical Information Science*, *12*(1), 47–62. Retrieved from https://doi.org/10.1080%2F136588198242003 doi: 10.1080/136588198242003

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

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

Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., & Wyart, M. (2019, July). Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, *100*(1). Retrieved from https://doi.org/10.1103/physreve.100.012115 doi: 10.1103/physreve.100.012115

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.

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.

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' [Computer software manual]. (version 2.6-7)

Hill, R. A., Weber, M. H., Leibowitz, S. G., Olsen, A. R., & Thornbrugh, D. J. (2015, dec). 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. Retrieved from `https://doi.org/10.1111/1752-1688.12372` doi: 10.1111/1752-1688.12372

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., ... Megown, K. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*, *81*(5), 345–354.

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

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.

Jiang, P., & Kumar, P. (2019). Using information flow for whole system understanding from component dynamics. *Water Resources Research.*

Kasprak, A., Hough-Snee, N., Beechie, T., Bouwes, N., Brierley, G., Camp, R., ... others (2016). The blurred line between form and process: A comparison of stream channel classification frameworks.

PloS one, *11*(3), e0150293.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019, August). Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling. *Hydrology and Earth System Sciences Discussions*, 1–32. Retrieved from https://doi.org/10.5194/hess-2019-368 doi: 10.5194/hess-2019-368

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. A., Dahlke, H. E., Pasternack, G. B., & Sandoval-Solis, S. (2017). Revealing the diversity of natural hydrologic regimes in california with relevance for environmental flows applications. *JAWRA Journal of the American Water Resources Association*, *53*(2), 411–430.

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

Lane, B. A., Pasternack, G. B., & Solis, S. S. (2018, mar). Integrated analysis of flow, form, and function for river management and design testing. *Ecohydrology*, *11*(5), e1969. Retrieved from https://doi.org/10.1002/eco.1969 doi: 10.1002/eco.1969

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

Lin, H. W., Tegmark, M., & Rolnick, D. (2017, jul). Why does deep and cheap learning work so well? *Journal of Statistical Physics*, *168*(6), 1223–1247. Retrieved from https://doi.org/10.1007/s10955-017-1836-5 doi: 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., ... others (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, aug). The fractal properties of topography as controlled by the interactions of tectonic, lithological, and geomorphological processes. *Earth Surface Processes and Landforms*. Retrieved from `https://doi.org/10.1002%2Fesp.4206` doi: 10.1002/esp.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.*

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., , & Rea, A. (2012). Nhdplus version 2: User guide [Computer software manual].

McManamay, R. A., Troia, M. J., DeRolph, C. R., Sheldon, A. O., Barnett, A. R., Kao, S.-C., & Anderson, M. G. (2018, jun). 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. Retrieved from `https://doi.org/10.1371/journal.pone.0198439` doi: 10.1371/journal.pone.0198439

Michie, D. (1968). "memo" functions and machine learning. *Nature*, *218*(5136), 19.

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

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

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

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & V. Gupta, H. (2020). Does information theory provide a new paradigm for earth science? hypothesis testing. *Water Resources Research*, e2019WR024918.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning - ICML '05.* ACM Press. Retrieved from https://doi.org/10.1145/1102351.1102430 doi: 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, jan). Improving precipitation estimation using convolutional neural network. *Water Resources Research.* Retrieved from https://doi.org/10.1029/2018wr024090 doi: 10.1029/2018wr024090

Perdigão, R. A., Ehret, U., Knuth, K. H., & Wang, J. (2020). Debates: Does information theory provide a new paradigm for earth science? emerging concepts and pathways of information physics. *Water Resources Research*, *56*(2), e2019WR025270.

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

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019, feb). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. Retrieved from https://doi.org/10.1038/s41586-019-0912-1 doi: 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. (2004). Model selection via the auc. In *Proceedings of the twenty-first international conference on machine learning* (p. 89).

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

Ruddell, B. L., Drewry, D., & Nearing, G. S. (2019). Information theory for model diagnostics: tradeoffs between functional and predictive performance in ecohydrology models. *Water Resources Research*.

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

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

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

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

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. Retrieved from https://doi.org/10.1029/tr038i006p00913 doi: 10.1029/tr038i006p00913

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

Thiesen, S., Vieira, D. M., Mälicke, M., Wellmann, J. F., & Ehret, U. (2020). Her: an information theoretic alternative for geostatistics. *Hydrology and Earth System Sciences Discussions*, 1–30.

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

Tishby, N., & Zaslavsky, N. (2015, April). Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (ITW).* IEEE. Retrieved from https://doi.org/10.1109/itw.2015.7133169 doi: 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

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.*

Weijs, S. V., & Ruddell, B. L. (2020). Debates: Does information theory provide a new paradigm for earth science? sharper predictions using occam's digital razor. *Water Resources Research*, *56*(2), e2019WR026471.

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

Yu, B., & Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences*, *117*(8), 3920–3929.

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