***Feature importance:*** Feature importance is needed before developing scaling relationships. This is because the number of extrinsic features is significant and uninformative features may not be predictive indicators in developing scaling laws. Hence, this feature database needs to be distilled to get informative extrinsic variables that influence species richness for all ten compounds. For instance, the table below describes the dataset type and the total number of extrinsic features within this database. We explore six different state-of-the-art approaches to identify and rank important features (see supplementary material and hydrosheds\_vs\_SR\_v7.pptx). This ranking using six different methods allows us to reduce the bias of discarding important features when performing feature selection. As the dataset is imbalanced, we used all 54 samples to perform feature importance. The inputs to the sensitivity analysis methods are the values of extrinsic variables. The corresponding output on which the extrinsic features are assessed is the species richness of all 10 compounds. The feature importance values are then aggregated and averaged to get a single value upon which the extrinsic variables are ranked.

Table-XXX: Summary of dataset type, the total number of extrinsic features, and a number of essential features synthesized using six different sensitivity analysis methods.

|  |  |  |
| --- | --- | --- |
| **Dataset name/type** | **Total number of extrinsic features** | **Total number of important features** |
| WHONDRS | 45 | 12 |
| StreamStats | 8 | 8 |
| HydroSheds | 294 | 8 |
| EPAWaters-Catchment | 137 | 12 |
| EPAWaters-Watersheds | 137 | 22 |

The six sensitivity analysis methods we used for feature importance include the Pearson correlation coefficient, Spearman correlation coefficient, F-test, Mutual information (MI), Random Forest (RF), and SHAPley values. Pearson correlation measures the linearity between extrinsic descriptors and species richness, while the Spearsman correlation measures this relationship's non-parametric measure of monotonicity. F-test is a univariate feature selection that allows us to select the best features based on univariate statistical tests. MI measures the strength of non-linear dependency between the variables and species richness. It is equal to zero if species richness is independent of extrinsic variables, and higher values mean higher dependence. RF model with 100 decision trees is trained, and essential extrinsic descriptors are evaluated using permutation-based feature importance. This technique allows us to measure the increase in the prediction error of the RF model after we permuted the extrinsic descriptor's values. This permutation method breaks the relationship between the feature and the actual outcome. SHAPley is a method based on cooperative game theory and is used to increase transparency and interpretability of model-agnostic feature selection models such as RF. It provides an average of all the marginal contributions of an extrinsic variable considering all possible combinations.

***Scaling laws –*** A power law function (y = bxz) describing the relationship between extrinsic descriptors and species richness for each compound is used to develop scaling laws. Here, ‘y’ is the species richness, ‘x; is the extrinsic descriptor, b is the power-law coefficient, and z is the scaling exponent. We use Scipy Python non-linear least squares to fit this power law scaling function to input-output data. Statistical tests are performed to compute the p-value and check if the estimated parameters of the power law (i.e., b and z) are significant.

***PCA analysis –*** We have also performed PCA analysis to understand the data variation across the 54 samples. A total of 11 features among the important extrinsic variables across all the datasets are used in the PCA analysis (slide-19 in hydrosheds\_vs\_SR\_v19.pptx). A total of two PCA components are extracted with scores and loadings matrix. Positive PCA loadings indicate that an extrinsic feature and a principal part are positively correlated. That is an increase in one value results in an increase in the PC. Negative loadings indicate a negative correlation between a feature and its PC. A significant loading value (either positive or negative) indicates that an extrinsic descriptor strongly affects that specific principal component.