Received: 31 March 2021 | Accepted: 15 July 2021

DOI: 10.1111/2041-210X.13686

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Study becomes insight: Ecological learning from machine learning

Qiuyan Yu[1] | Wenjie Ji[1,2] | Lara Prihodko[3] | C. Wade Ross[4] | Julius Y. Anchang[1] | Niall P. Hanan[1]

[1]Plant and Environmental Sciences, New Mexico State University, Las Cruces, New Mexico, USA

[2]Department of Geography, California State University Long Beach, Long Beach, California, USA

[3]Animal and Range Sciences, New Mexico State University, Las Cruces, New Mexico, USA

[4]Tall Timbers Research Station, Tallahassee, Florida, USA

Correspondence
Qiuyan Yu
Email: qiuyanyu@nmsu.edu

Funding information
US National Science Foundation, Grant/Award Number: 1832194 and 2025166; National Aeronautics and Space Administration, Grant/Award Number: NNX17AI49G and 80NSSC20K0976

Handling Editor: Saras Windecker

Abstract

1. The ecological and environmental science communities have embraced machine learning (ML) for empirical modelling and prediction. However, going beyond prediction to draw insights into underlying functional relationships between response variables and environmental 'drivers' is less straightforward. Deriving ecological insights from fitted ML models requires techniques to extract the 'learning' hidden in the ML models.
2. We revisit the theoretical background and effectiveness of four approaches for deriving insights from ML: ranking independent variable importance (Gini importance, GI; permutation importance, PI; split importance, SI; and conditional permutation importance, CPI), and two approaches for inference of bivariate functional relationships (partial dependence plots, PDP; and accumulated local effect plots, ALE). We also explore the use of a surrogate model for visualization and interpretation of complex multi-variate relationships between response variables and environmental drivers. We examine the challenges and opportunities for extracting ecological insights with these interpretation approaches. Specifically, we aim to improve interpretation of ML models by investigating how effectiveness relates to (a) interpretation algorithm, (b) sample size and (c) the presence of spurious explanatory variables.
3. We base the analysis on simulations with known underlying functional relationships between response and predictor variables, with added white noise and the presence of correlated but non-influential variables. The results indicate that deriving ecological insight is strongly affected by interpretation algorithm and spurious variables, and moderately impacted by sample size. Removing spurious variables improves interpretation of ML models. Meanwhile, increasing sample size has limited value in the presence of spurious variables, but increasing sample size does improves performance once spurious variables are omitted. Among the four ranking methods, SI is slightly more effective than the other methods in the

'In the right light, study becomes insight'—Take the Power Back, by Rage Against the Machine (lyrics by Commerford, Morello, De La Rocha, Wilk; © Wixen Music Publishing)

presence of spurious variables, while GI and SI yield higher accuracy when spurious variables are removed. PDP is more effective in retrieving underlying functional relationships than ALE, but its reliability declines sharply in the presence of spurious variables. Visualization and interpretation of the interactive effects of predictors and the response variable can be enhanced using surrogate models, including three-dimensional visualizations and use of loess planes to represent independent variable effects and interactions.

4. Machine learning analysts should be aware that including correlated independent variables in ML models with no clear causal relationship to response variables can interfere with ecological inference. When ecological inference is important, ML models should be constructed with independent variables that have clear causal effects on response variables. While interpreting ML models for ecological inference remains challenging, we show that careful choice of interpretation methods, exclusion of spurious variables and adequate sample size can provide more and better opportunities to 'learn from machine learning'.

KEYWORDS

bivariate functional relationship, boosted regression tree (BRT), ecological inference, interpretation of machine learning models, random forest (RF), variable importance

1 | THE NEED TO IMPROVE MACHINE LEARNING INTERPRETABILITY IN ECOLOGY

Ecologists and environmental scientists often find themselves searching for tools to model and predict complex, nonlinear and high-dimensional systems. In recent years, our ability to predict complicated systems has greatly improved through the development of machine learning (ML) algorithms. A substantial body of literature has accumulated on applications of ML in the ecological and environmental sciences (e.g. Cutler et al., 2007; Elith & Leathwick, 2017; Elith et al., 2008; Marmion et al., 2009). However, when prediction is not the primary goal, most ML methods tend to behave like 'black boxes', meaning that it can be challenging to derive new understanding or ecological insights from these statistical models, irrespective of their predictive abilities (Roscher et al., 2020). For example, if we are interested in understanding the distribution of species richness sampled across a study domain (the 'response' variable), we might use ML methods to fit relationships with a set of candidate independent ('predictor') variables (e.g. variables describing the spatial and temporal variation in climate, soils and disturbance). The resulting ML model may closely fit the training data and provide accurate predictions for locations between sample plots. However, without appropriate visualization or interpretation tools, that same model may provide little or no insight into the functional (or causal) relationships between species diversity and underlying climatic, edaphic and biotic interactions.

The ability of ML approaches to predict accurately is valuable, but ecological interpretation of the underlying functional relationships can be challenging (Lucas, 2020; Wenger & Olden, 2012). Nonetheless, the often superior predictive ability of ML approaches, relative to more traditional approaches (e.g. linear and nonlinear regression), suggests that the functional relationships are embodied in the fitted ML models. The challenge, therefore, is to provide tools to extract (quantify and/or visualize) those functional and ecological relationships from the black box. In particular, we would hope that ML will help answer three key ecological questions: (a) Which predictor variables are the most influential in determining the behaviour of the response variable? (b) What are the functional relationships between predictors and the response variable? And (c) how do interactions among predictors determine the complex and often nonlinear patterns in the response variable?

To extract ecological insights, we need a comprehensive understanding of interpretation approaches for ML models and how to generate reliable and accurate interpretation results (Brieuc et al., 2018; Cutler et al., 2007; Lucas, 2020). Earlier studies focused on the accuracy of variable importance estimations (Gini impurity and mean decrease in accuracy) related to impacts of correlated predictors (e.g. Gregorutti et al., 2017; Strobl et al., 2008), scale of measurement and number of categories (e.g. Nicodemus, 2011; Strobl et al., 2007) and their intrinsic stability (Calle & Urrea, 2011; Wang et al., 2016). Mean decrease in accuracy is sensitive to dataset noise (Calle & Urrea, 2011), while the performance of the Gini coefficient is affected by correlation between predictor variables and the number of categories for categorical predictors (Nicodemus, 2011). While agreement has not been reached for which of the two is more efficient (e.g. Calle & Urrea, 2011; Nicodemus, 2011), new variable importance measures have been developed (e.g. conditional importance by Strobl et al., 2007 and split importance [SI] by Elith et al., 2008), promoting the need for a comprehensive comparison among them. Meanwhile, compared to variable importance,

assessment for methods to extract bivariate relationships has received much less attention, although it is critical for ecological interpretation.

The need for sufficient training data (sample size) has long been a major concern for accurate ML model predictions (Perry & Dickson, 2018; Raudys & Jain, 1991; Stockwell & Peterson, 2002). However, it is less well-known how sample size impacts model interpretation, including variable importance and functional relationships. This is particularly important in ecology, where relatively small-scale experiments lead to ML models with small sample sizes. Assessment of variable importance measures (Gini and permutation importance) showed that both of them are sensitive to sample size (e.g. Strobl & Zeileis, 2008; Wang et al., 2016). Similarly, while the presence of spurious variables may have little negative impact on ML predictions, the covariance among true and spurious predictors may reduce our ability to effectively rank predictor importance. More particularly, the inclusion of spurious predictor variables will likely obscure our ability to retrieve 'true' functional relationships for influential predictors, as spurious variables alias a portion of the underlying correlations with true predictors and thus obscure the true relationships.

In this paper, we evaluate how ML interpretation approaches are impacted by sample size and the presence of spurious variables. We review interpretive tools that can shed new light on the ML black box, and provide opportunities for new and improved ecological and functional insights. We use a simulation dataset to illustrate key interpretation strategies, and demonstrate their abilities for reliable ecological interpretation and inference. In so doing, we provide some recommendations for efficient ecological interpretation.

2 | MATERIALS AND METHODS

We conducted a simulation, with known underlying relationships between response and predictor variables, to assess retrieval of ecological insights from ML models. We examine four algorithms to evaluate variable importance and two for visualization of bivariate functional relationships between each predictor and the response variable. We then introduce surrogate models to visualize the interactive (i.e. multivariate) effects of predictor variables on the response variable. We provide brief theoretical background for the interpretation approaches, compare their performance and analyse their sensitivity to sample size and spurious variables.

2.1 | Simulation design

We generated a pseudo dataset representing hypothetical variability in global species richness in response to three environmental predictors, with known underlying functional relationships (Figure 1a). To fulfil the goal of examining effectiveness of ML interpretation approaches, our simulation simplified and generalized the abiotic mechanisms of species richness by focusing on three primary factors: temperature (MAT, K; Antão et al., 2020; Fuhrman et al., 2008; Stegen et al., 2012), rainfall (MAP, mm; Frank et al., 2014; Gelfand et al., 2005) and disturbance (here we only used fire frequency, $year^{-1}$; He et al., 2019; Peterson & Reich, 2008). In this hypothetical example, we modelled species richness (SN) as a linear combination of the three environmental (predictor) variables:

$$\mathrm{SNnorm}_x = f(x) + \varepsilon, \tag{1}$$

$$\mathrm{SN} = 100 \times \left(w_1 \times \mathrm{SNnorm}_{\mathrm{MAT}} + w_2 \times \mathrm{SNnorm}_{\mathrm{MAP}} + w_3 \times \mathrm{SNnorm}_{\mathrm{Fire}}\right), \tag{2}$$

where x represents each of the three predictors (MAT, MAP and fire), and f is the associated deterministic function. SNnorm_x (each scaled 0–1) are the normalized species richness values determined by environmental predictors x. The three SNnorm_x (Figure 1a) were generated with parabolic, sigmoid and negative sigmoid relationships (f) for MAT, MAP and fire frequency, respectively, with added Gaussian error (ε). Weights (w) for MAT, MAP and fire frequency were set as $w_1 = 0.6$, $w_2 = 0.3$ and $w_3 = 0.1$, respectively, representing the relative importance of the three predictors in determining eventual SN. We constrained the range of SN to 0–100 by removing points (166 samples) out of the 0–100 range.

In addition, we simulated two environmental variables, 'V1' and 'V2', which in our simulation have varying degree of correlation with the three ('true') predictor variables, but no direct impact on species richness. These 'spurious variables' were included in the ML analysis to mimic common practice of including numerous variables when constructing ML models, many of which are mechanistically uninfluential in determining the response, but may nevertheless be correlated with it due to environmental covariance. The combination of correlation among predictor variables and noise obscures the underlying bivariate relationships (Figure 1).

Simulation of the five environmental variables incorporated hypothetical spatial gradients in a domain representing a hypothetical terrestrial world, with partial covariance among environmental variables, as typically found in real systems (Figure 2). Thus among the three 'true' predictor variables, MAT follows a latitudinal gradient peaking near the equator (Figure 2a); MAP peaks in the centre of the domain and declines outwards (Figure 2b); while fire frequency interacts with MAP and MAT, peaking in mesic systems with moderate rainfall and higher temperature (Figure 2c). Among the two 'spurious variables', V1 (Figure 2d) increases from west to east with no direct dependence on other predictors, while V2 (Figure 2e) is partially correlated with the other predictor variables.

Weight terms (w) in Equation 2 provide the underlying variable importance, ranked MAT > MAP > fire frequency. The deterministic relationships in Figure 1a represent the underlying functional relationships between species richness and the three deterministic predictors, while the spatial interactions among predictors (Figure 2) simulate covariance among predictors that tends to confound the ability of statistical modelling approaches to derive accurate functional relationships, particularly when examined using bivariate methods (Figure 1b,c). As is common in real-world situations, the

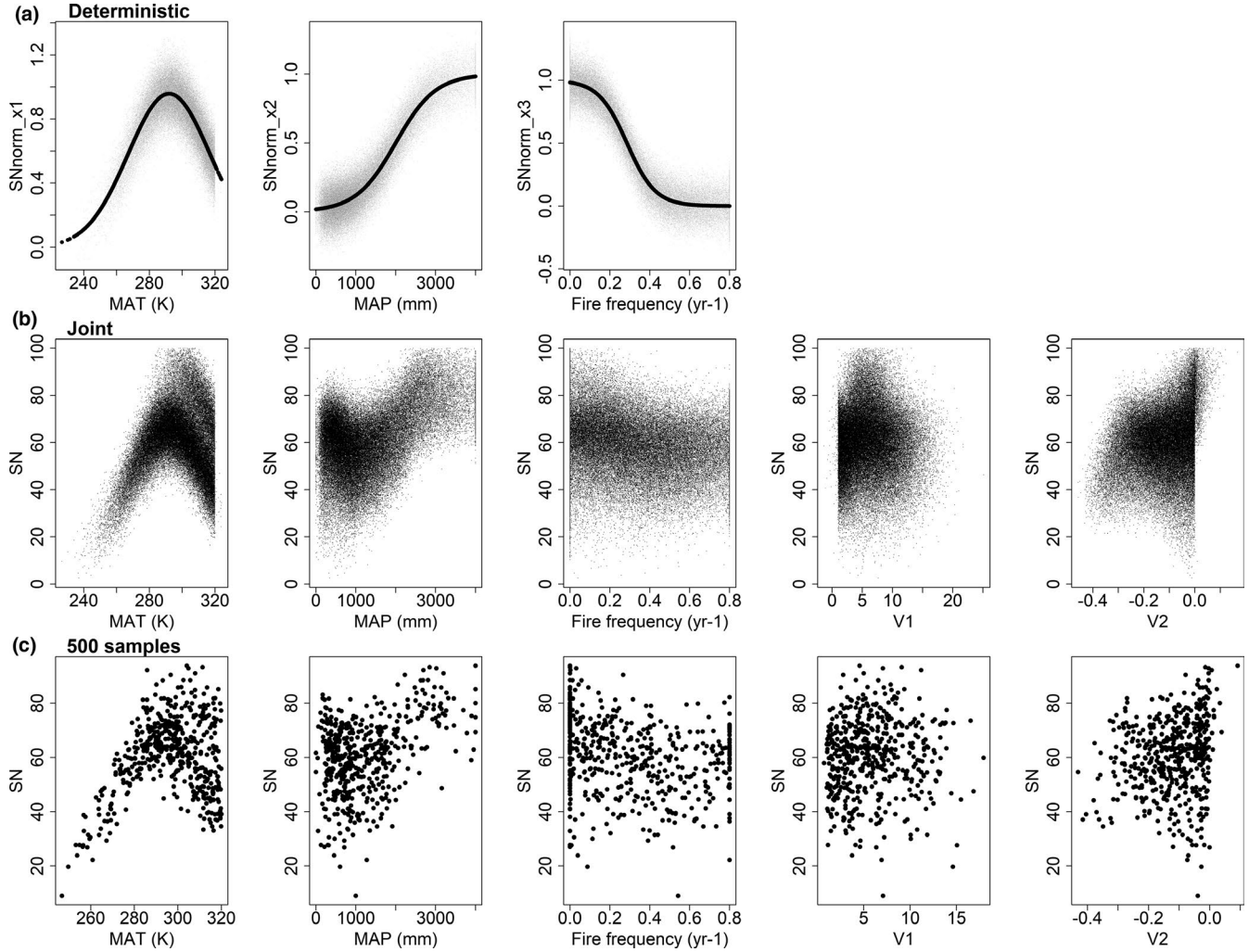

FIGURE 1 Simulated relationships between species richness for a hypothetical taxon and five environmental predictor variables. The upper row (a) shows the deterministic bivariate relationships (dark line) used to simulate species richness (Equation 1), with additional Gaussian noise (grey points). Row (b) shows the final species richness (SN, Equation 2) along the five predictors under 'data-rich' scenario ($N = 64{,}454$). Row (c) illustrates the additional challenge inherent in modelling data-poor situations, showing 500 samples randomly selected from row (b)

spatial interactions between predictor variables distort our ability to visualize bivariate relationships (Figure 1b,c), particularly for variables with lower influence such as mean annual precipitation and fire frequency ($w = 0.3$ and 0.1 respectively). Our goal, therefore, is to be able to work with complex, correlated and noisy data, such as simulated here, and discern which predictor variables are ecologically important, and then gain insights into the functional relationships with the response variable.

2.2 | Interpretation methods

We investigated interpretation approaches for extracting (a) variable importance, (b) bivariate functional relationships and (c) multivariate functional relationships and predictor interactions. For each interpretation approach, we examined how they are affected by sample size and spurious variables.

We used random forest (RF) and boosted regression tree (BRT) models as our primary ML examples, although many of the insights and recommendations apply to other common ML approaches. RF and BRT are used extensively in ecology and environmental sciences, in large part due to their ability to deal with nonlinear interactions and remarkable predictive capabilities (e.g. Anchang et al., 2020; Brieuc et al., 2018; Cutler et al., 2007; De'Ath, 2007; Jevšenak & Skudnik, 2021; Molnar, 2019; Prasad et al., 2006; Ross et al., 2020; 2021). The hyperparameters used for ML models were consistent among sample size and bootstraps. We used 100 trees for RF, while the number of predictors (mtry) was determined using the tuneFR function. For BRT, the tree complexity was three, learning rate equalled 0.01 and bag fraction was 0.8. The predicted species richness by RF and BRT and associated predictor errors are shown in Figure S1. The pseudo dataset comprised 64,620 (359 × 180) samples, representing species richness across the hypothetical domain (Figure 2f). Since real-world studies can be severely data limited, we

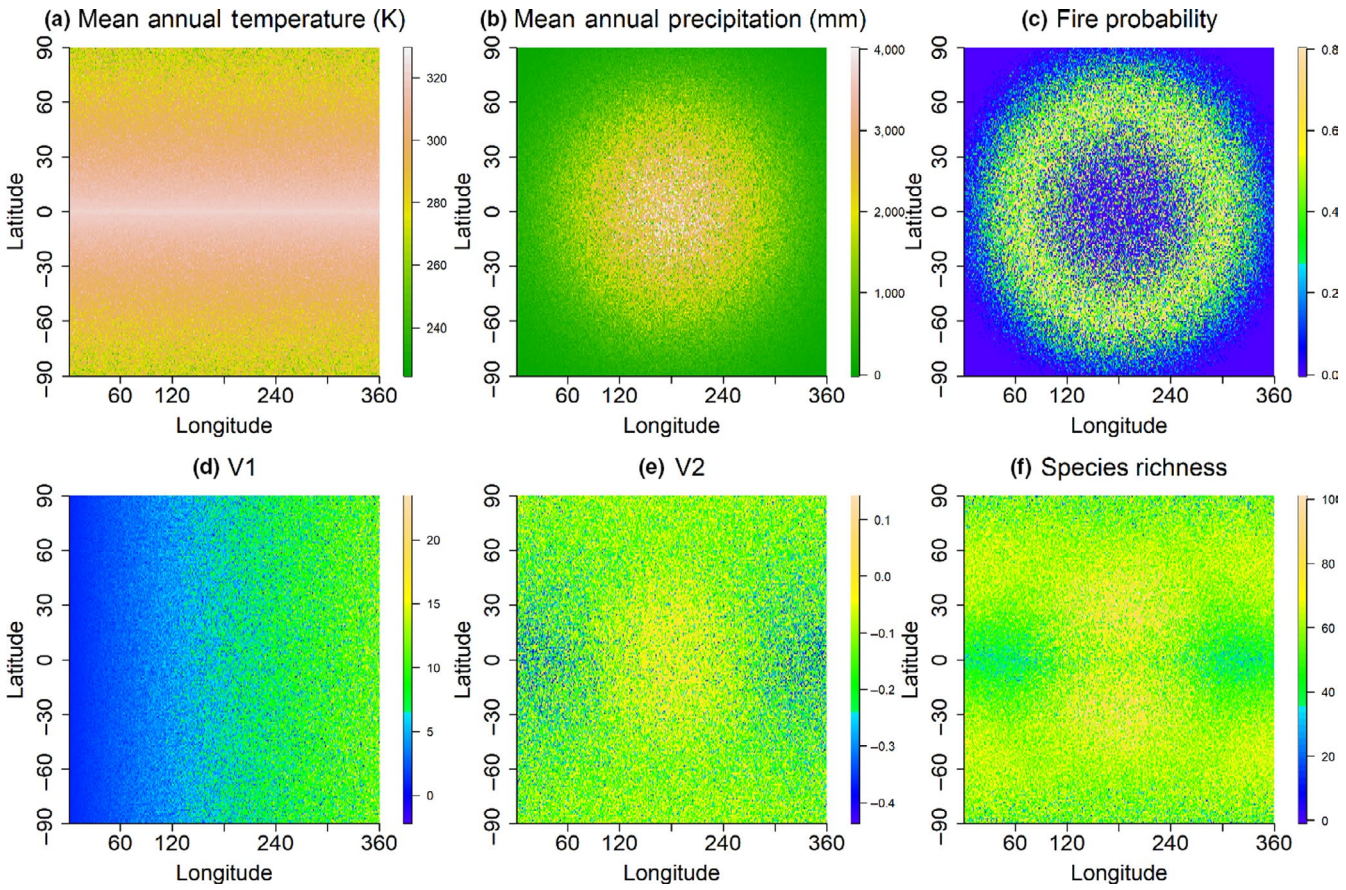

FIGURE 2 Spatial distribution of three influential environmental variables: (a) temperature, (b) precipitation and (c) fire, and two non-influential variables: (d) V1 and (e) V2 used to model (f) species richness generated using Equation 1

also sampled the full dataset (sample sizes: 100, and from 500 to 5,500 with an interval of 500) to examine how dataset size affects ML interpretation. To avoid spurious findings with random sample fluctuations, the random selection of samples was bootstrapped with 100 repetitions (the distribution of bootstrap samples with respect to predictor variables was checked for similarity with the overall dataset using Kolmogorov–Smirnov test).

2.2.1 | Variable importance

Variable importance (or feature importance) approaches are designed to rank the relative contribution of multiple predictor variables to response variables in ML models (Friedman, 2001). There is little consensus in the ML literature on how to calculate the relative importance of different independent variables in a fitted model. We tested four frequently used methods to rank predictor variables: Gini importance (GI; Breiman, 2001), permutation importance (PI; Cutler et al., 2012), conditional permutation importance (CPI; Strobl et al., 2007) and (SI; Elith et al., 2008). We estimated their accuracy in ranking predictor variables relative to simulated weights (w), and examined their sensitivity to sample size and spurious variables.

The first three importance measurements are associated with RF model, while the last one (SI) is commonly used with BRT models. GI quantifies the decreased variation of samples through all the nodes split by a predictor variable in a RF. PI estimates variable importance as the decrease in prediction accuracy (increase in error) when the target variable is randomized (permuted) and input to a previously trained model (Strobl et al., ,2007, 2008). GI and PI are implemented in the R package randomForest. R package party modifies the PI approach as the CPI, which computes the decrease in prediction accuracy following the permutation of portions of the range of a predictor variable (Strobl et al., 2008). SI considers how many times a predictor variable is used in splits across all trees in a boosted regression tree, and has been adopted by many packages such as LightGBM (Ke et al., 2017), XGBoost (Chen & Guestrin, 2016) and brt (Elith et al., 2008; Friedman & Meulman, 2003). See Table S1 for common R resources for ML.

2.2.2 | Bivariate functional relationships

We explored two methods, partial dependence plots (PDPs) and accumulated local effects (ALEs), for retrieval of underlying bivariate functional relationships. PDP has been a mainstay for ecological and environmental inference (e.g. Cutler et al., 2007; Friedman & Meulman, 2003; Galkin et al., 2018; Moya-Laraño & Corcobado, 2008; Sankaran et al., 2008), while the use of ALE method has been less frequently adopted (Apley & Zhu, 2016). These

two approaches are available in most ML programming packages (Table S1). Both PDP and ALE are designed to retrieve the functional relationships between response variables and predictive variables, somewhat analogous to a bivariate linear or nonlinear regression. PDP estimates the marginal effects of predictors by averaging the predicted outcomes from an ML model (Casalicchio et al., 2018; Friedman, 2001; Molnar et al., 2018; Zhao & Hastie, 2021). By contrast, ALE calculates the difference of local predictions across small intervals of the predictor variable range (Apley & Zhu, 2016). PDP and ALE can be applied to both RF and BRT.

2.2.3 | Multi-variate functional relationships and predictor interactions

While PDP and ALE are intended to show the bivariate relationship between the response variable and predictor variables, the multivariate effects of predictor variables are not readily apparent using bivariate visualizations. A less common approach for visualization of these underlying relationships in the environmental and ecological literature uses interpretable or 'surrogate' models (Molnar, 2019). A surrogate model uses predictions from an ML model that are re-analysed using a more easily interpretable (visualizable) model approach. Although any interpretable model (e.g. a simple decision tree, generalized linear or additive models) can be used as a surrogate model (Molnar et al., 2018), here we highlight the use of three-dimensional (3D) plane fitting (loess models) as a straightforward approach to visualization of multi-variate functional relationships and predictor interactions (Aho, 2013).

The surrogate model was generated from the fitted ML model (e.g. using the *predict* function in R) for all points in the calibration set, potentially enhanced with additional random points in data-poor situations. In this way, a surrogate model can represent the relationships embodied in the ML model, free of the noise associated with the original data. To explore surrogate model visualizations for ecological interpretation of fitted ML models, we examined the multivariate effect of the more influential variables on species richness using 3D loess planes to show the interactive effect of predictor pairs on the response variable.

3 | RESULTS

3.1 | Variable importance

We compared the estimated variable importance and the rank of the predictor variables with the simulated importance weights (Figures 3 and 4). In the presence of spurious variables (Figure 3),

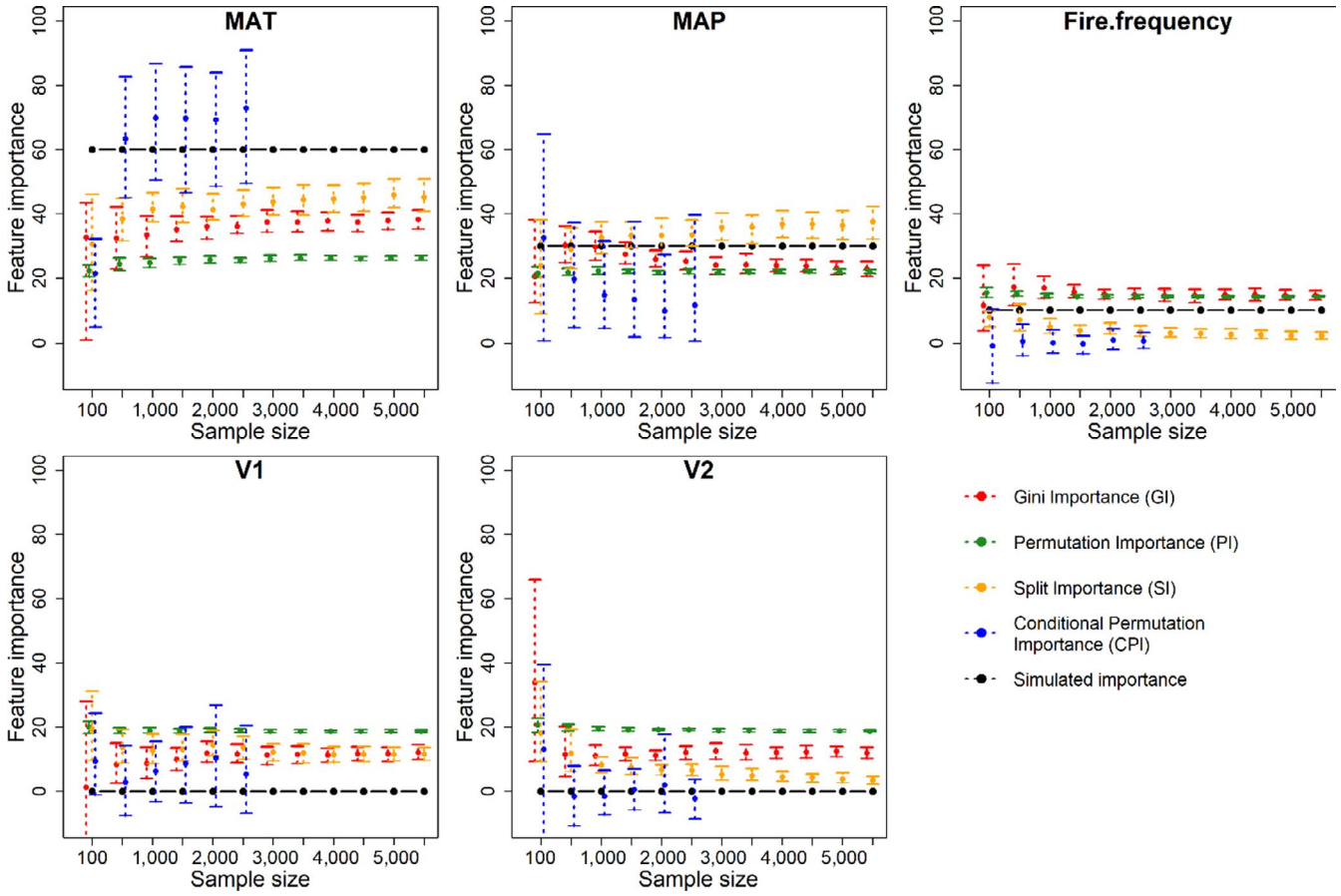

FIGURE 3 Estimation of feature importance (%) using four interpretation methods as a function of sample size (100, then from 500 to 5,500 with an interval of 500). Points represent mean value of 100 bootstraps, and error bars indicate variation

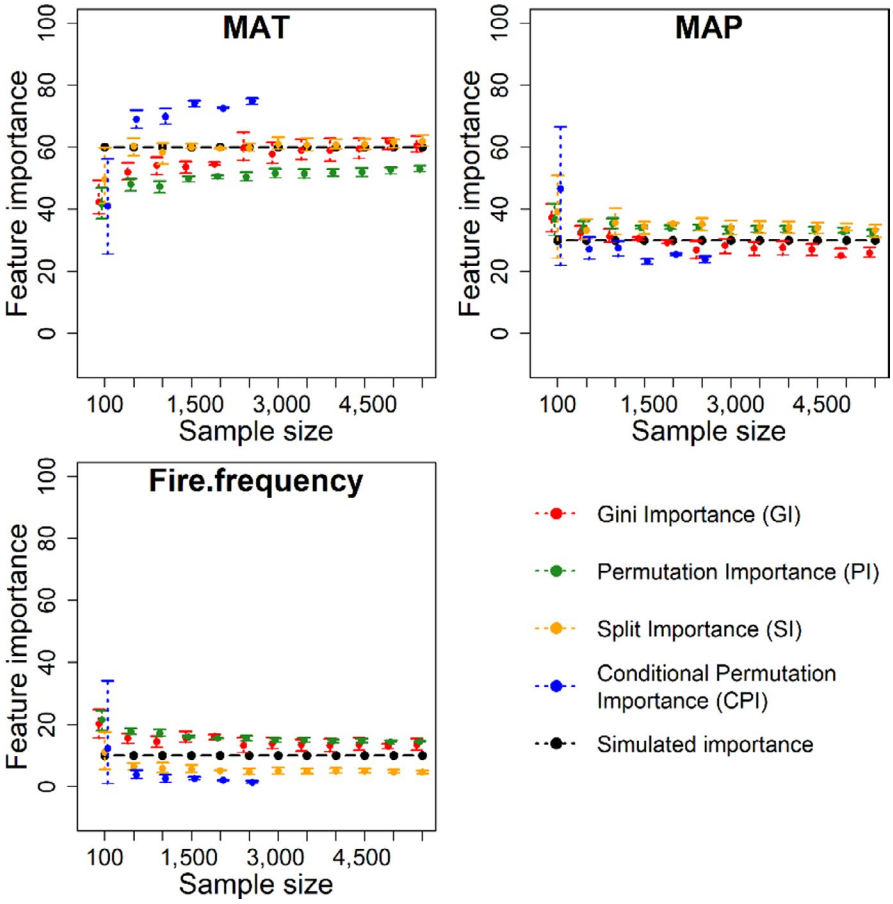

FIGURE 4 Estimated feature importance (%) for the three 'true' predictor variables using four interpretation methods with the two spurious variables omitted in the machine learning models

GI, SI and CPI were able to retrieve the correct rank order among influential variables (MAT > MAP > Fire), and SI showed slightly better result than the others. However, GI and SI assigned considerable importance to the spurious variables (V1, V2) and no method was able to consistently retrieve the applied weightings for the true predictors (w = 60%, 30% and 10% respectively) when spurious variables (V1, V2) were present in the models. PI was unable to correctly rank predictor variables (Figure 3), indicating particular sensitivity of this index to covariance and aliasing among true and spurious variables. Notably, CPI tends to exaggerate the importance of the more important variable (MAT), while assigning relatively low importance to less influential (but true) variables (MAP, fire) and spurious variables. CPI importance rankings were also much more variable than other importance methods. Larger sample size decreases the variation in importance estimates, but did not consistently improve accuracy indicating that, in the presence of spurious predictors, larger sample sizes will not necessarily improve predictor importance assessments.

Removal of spurious variables considerably improved the ability of GI, PI and SI to accurately rank and quantify predictor importance (Figure 4). However, CPI still tends to exaggerate importance of the more influential variable (MAT) while underestimating the less influential variables (MAP and fire). Larger sample sizes, in the absence of spurious variables, improved the ability of GI and PI to retrieve the importance of the three predictors. However, SI and CPI were insensitive to larger sample sizes. In the absence of spurious variables, GI and SI were the most reliable importance indices.

3.2 | Bivariate functional relationships

The accuracy assessment for retrieved bivariate functional relationships was conducted using a similarity measure (Kendall's Tau; Sen, 1968) between the retrieved curve and the simulated function for deterministic predictors. A high similarity value indicates that the retrieved functional relationship is similar to the deterministic function (Figure 1a).

Partial dependence plots were generally more accurate than ALE in retrieving the functional form of deterministic variables (Figure 1a), when computed using either the RF or BRT model (Figures S2 and S3). With the existence of spurious variables, increasing sample size has limited effect on the accuracy of either PDP or ALE to retrieve bivariate functions of the three deterministic variables (Figures 5 and 6). Retrieval of functional forms for the three deterministic variables is greatly improved following removal of spurious variables from the models (Figure 6). Figure S3 shows an example of largely improved retrieved curves for the three deterministic variables (either by PDP or ALE) with the spurious variables omitted. The similarity between the retrieved function (by PDP) and the simulated function increased from 0.5 to 0.8 for MAT, from 0.7 to 0.9 for MAP and from 0 to 0.8

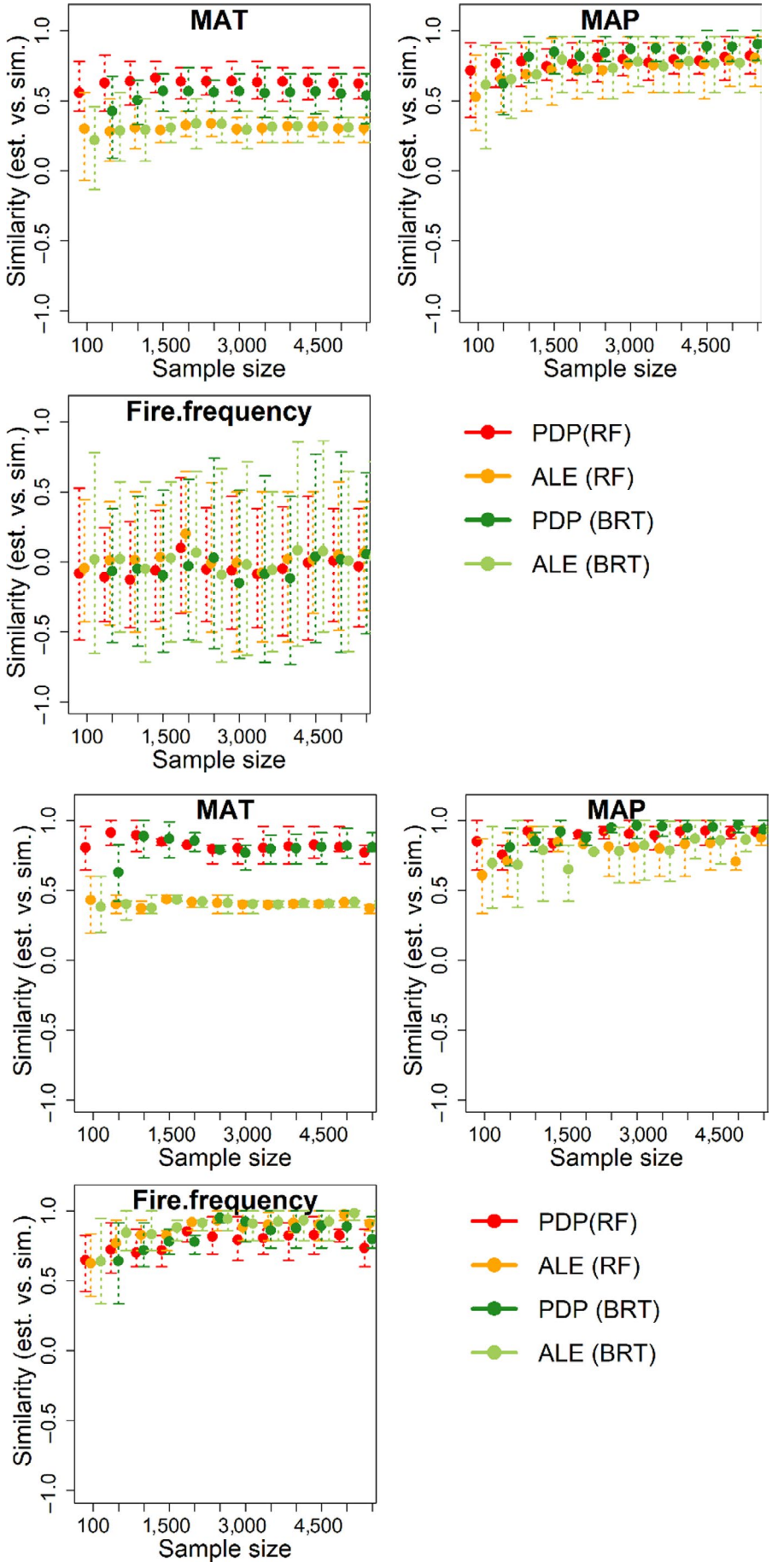

FIGURE 5 Similarity (Kendall's Tau between retrieved functional curve and the simulated function for the deterministic variables: MAT, MAP and fire frequency) assessment as a function of sample size. At each sample size, the distribution of similarity was generated by the 100 bootstraps. Points represent mean value of similarity among the 100 bootstraps, and error bars indicate variation

FIGURE 6 Error/similarity assessment of retrieved functional relationships compared to simulated functional relationships for the three 'true' predictor variables, with spurious variables removed

for fire frequency, indicating the strong impact of spurious variables on retrieval of bivariate functions. Increasing sample size also tends to improve the accuracy of PDP and ALE, when the spurious variables are excluded, up to a threshold of ~2,000 points beyond which accuracy reaches an asymptote in these simulations.

3.3 | Multi-variate functional relationships and predictor interactions

We used 3D loess planes to show how the response variable in our simulations varies with interacting pairs of the three influential variables (MAT, MAP and fire frequency). The simulated data are shown as points and the predictions from the fitted RF model (i.e. the surrogate model) are the planes. As might often occur in real-world data, the full representation of a predictor variable's functional range does not always occur, leading to concentration of data in parts of the feature space, and less well-supported loess predictions in under-sampled regions of feature space. However, the separate and interactive effects of the predictor variables begin to emerge in the shape of the surrogate models. For example, in Figure 7a, which shows simulated species richness response to the two most influential variables (MAT and MAP), the underlying effects of MAT are seen as a bell-shape distribution, with the asymptotic relationship between species richness and MAP. In our simulated dataset, the fire effect leads to a reduction in species richness but the relationship is complex reflecting realistic correlations between fire frequency and climate that were incorporated into the pseudo-data simulations (Kahiu & Hanan, 2018).

The surrogate model predictions for data-poor models (Figure 7d–f) are broadly similar to data-rich models but with less reliable predictions of separate and interactive effects, particularly in the already data-limited extremes of feature space. With limited data samples (1,000 in this case), the interactive effects emerging for the three influential variables are less clear. For example, the effect of fire frequency on species richness at low MAT shifted from the original nearly no effect (Figure 7b) to a conspicuous linear increasing curve (Figure 7e). Meanwhile, fire frequency, simulated with limited effect on specie richness under heavy rainfall due to high moisture (Figure 7c), presented significant negative impact on species richness with small sample size (Figure 7f).

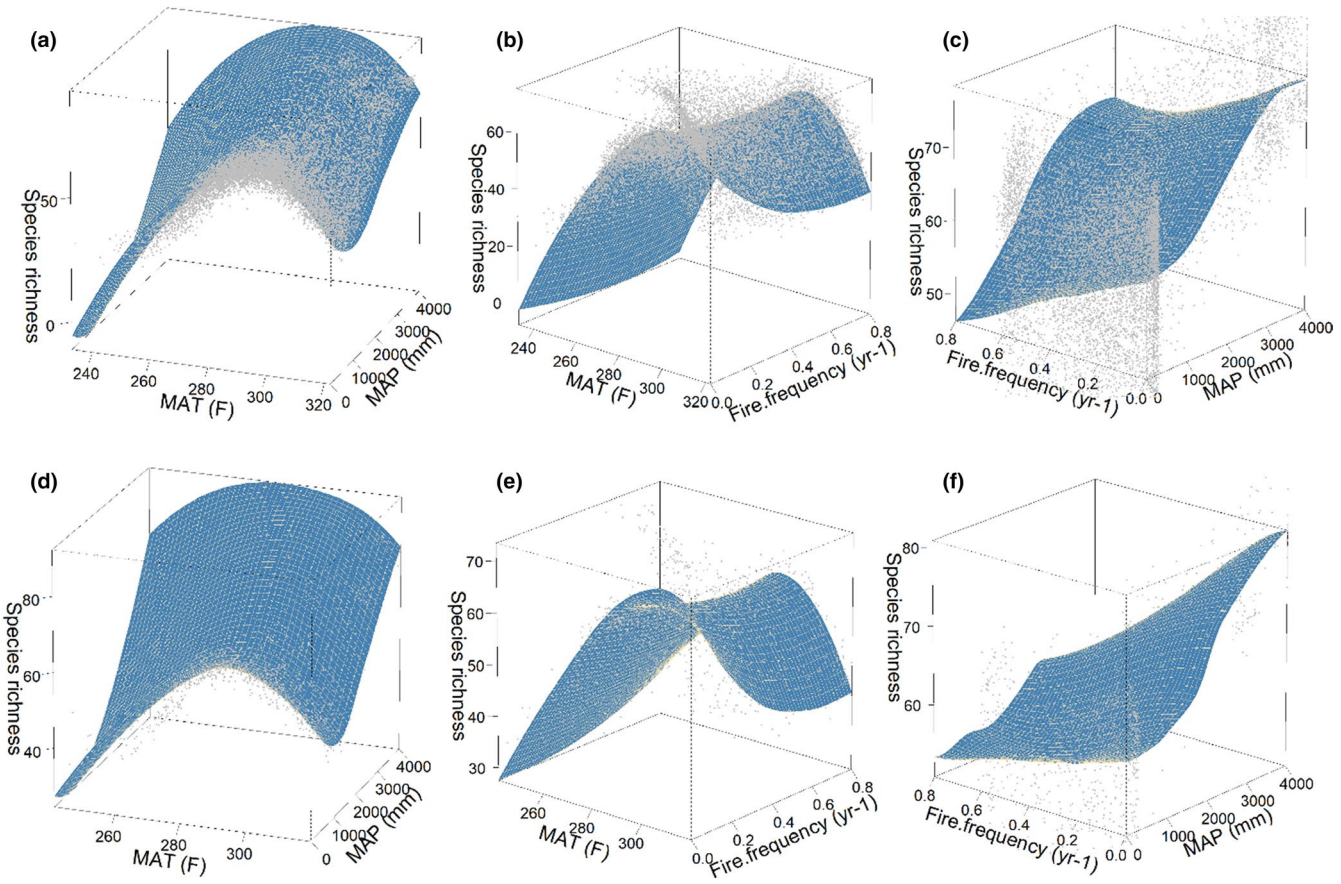

FIGURE 7 3D loess planes showing how a response variable (species richness) varies with the three deterministic predictor variables (MAT, MAP and fire frequency) with prediction from random forest for a data-rich scenario (a–c with 5,500 samples) and a data-poor scenario (d–f with 1,000 samples). Shaded blue planes are the surrogate models. Points are the original stochastically simulated data with known underlying functional relationships (Figure 1a)

4 | DISCUSSION: TOWARDS EFFICIENT ML MODEL INTERPRETATION

Machine learning models have been used extensively in ecological and environmental studies due to their simplicity in implementation and remarkable predictive ability. However, the 'black box' nature of most ML models limits ecological inference, process understanding and interpretation of the dynamics underlying the system being studied. In this paper, we reviewed several ML interpretation methods (four variable importance measurements, two functional relationship methods and a surrogate model) and examined their response to sample size and spurious variables, in an effort to improve ecological/process inferences, and make recommendations for use of ML models. We found that the performance of interpretation approaches used to identify which variables are most influential, and retrieve underlying functional relationships from ML models, is sensitive to methods selected and the presence of spurious variables. Increasing sample size improves interpretation when spurious variables are omitted from ML model fits. However, increasing sample size does not, on its own, overcome model confusion caused by spurious variables.

The inclusion of spurious variables (i.e. variables that are correlated with, but do not have a causal relationship with the target variable) severely impacts variable importance ranking and retrieval of functional curves. In particular, removing spurious variables can largely improve variable ranking using GI and SI, likely due to their sensitivity to within-predictor correlation (Nicodemus, 2011). In our simulations, the PDP approach was considerably more successful than ALE in retrieving underlying functional relationship. However, both methods were highly sensitive to the presence of spurious variables. As found for variable importance ranking, increasing sample size provided little benefit in the ability of PDP and ALE when the spurious variables were present, but increasing sample size did result in improved functional relationship retrieval when spurious variables were excluded.

Our results confirm previous research that careful selection of independent variables is essential for successful ML (e.g. Alizadeh et al., 2018; Seyedzadeh et al., 2019; Vellido et al., 2012). Although inclusion of numerous independent variables in ML models can yield improvements in predictive ability, this approach is not helpful when ecological interpretation is the goal. We also showed that the use of surrogate models (i.e. analysis of predictions from fitted ML models) can provide additional insights into multi-variate relationships and predictor interactions using, in this study, 3D loess planes. However, the surrogate model predictions were also very sensitive to dataset size, requiring larger datasets (>1,000 in this case) to characterize the interactive effects of predictor variables.

This study compared different interpretation methods for estimating variable importance and functional relationships and analysed the factors that may influence the interpretation of ML models. ML analysts should be aware that including correlated independent variables in ML models with no clear causal relationship to response variables can interfere with ecological inference. When ecological inference is important, ML models should be constructed with independent variables that have clear causal effects on response variables. While interpretation of ML models for ecological inference remains challenging, careful choice of interpretation methods, exclusion of spurious variables and sufficient sample size can provide ML users with more and better opportunities to 'learn from machine learning'.

ACKNOWLEDGEMENTS

This research was supported in part by the US National Aeronautics and Space Administration (NASA) as part of the NASA Carbon Cycle Science Program (Grant # NNX17AI49G) and the NASA ICESat-2 Program (Grant # 80NSSC20K0976) and the US National Science Foundation, Jornada Basin Long-Term Ecological Research (LTER) Program (Grant #1832194 and #2025166).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

N.P.H. and Q.Y. conceived the ideas and designed the simulation model; Q.Y. did the analysis; Q.Y. and W.J. led the writing of the manuscript. All authors contributed critically to writing and editing the drafts and gave the final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13686.

DATA AVAILABILITY STATEMENT

The simulation and analysis scripts in R are made available for public at Zenodo https://zenodo.org/badge/latestdoi/387275079 (Yu et al., 2021).

ORCID

Qiuyan Yu https://orcid.org/0000-0002-4003-0541
Wenjie Ji https://orcid.org/0000-0003-2554-4373

REFERENCES

Aho, K. A. (2013). *Foundational and applied statistics for biologists using R*. CRC Press.

Alizadeh, M. J., Kavianpour, M. R., Danesh, M., Adolf, J., Shamshirband, S., & Chau, K. W. (2018). Effect of river flow on the quality of estuarine and coastal waters using machine learning models. *Engineering Applications of Computational Fluid Mechanics*, *12*(1), 810–823. https://doi.org/10.1080/19942060.2018.1528480

Anchang, J. Y., Prihodko, L., Ji, W., Kumar, S. S., Ross, C. W., Yu, Q., Lind, B., Sarr, M. A., Diouf, A. A., & Hanan, N. P. (2020). Toward operational mapping of woody canopy cover in tropical savannas using Google Earth Engine. *Frontiers in Environmental Science*, *8*, 4. https://doi.org/10.3389/fenvs.2020.00004

Antão, L. H., Bates, A. E., Blowes, S. A., Waldock, C., Supp, S. R., Magurran, A. E., Dornelas, M., & Schipper, A. M. (2020). Temperature-related biodiversity change across temperate marine and terrestrial systems. *Nature Ecology & Evolution*, *4*(7), 927–933. https://doi.org/10.1038/s41559-020-1185-7

Apley, D. W., & Zhu, J. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *ArXiv*. 1612.08468.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brieuc, M. S., Waters, C. D., Drinan, D. P., & Naish, K. A. (2018). A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, *18*(4), 755–766. https://doi.org/10.1111/1755-0998.12773

Calle, M. L., & Urrea, V. (2011). Letter to the editor: Stability of random forest importance measures. *Briefings in Bioinformatics*, *12*(1), 86–89. https://doi.org/10.1093/bib/bbq011

Casalicchio, G., Molnar, C., & Bischl, B. (2018). Visualizing the feature importance for black box models. In M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, & G. Ifrim (Eds.), *Joint European conference on machine learning and knowledge discovery in databases* (pp. 655–670). Springer.

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

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In C. Zhang & Y. Ma (Eds.), *Ensemble machine learning* (pp. 157–175). Springer. https://doi.org/10.1007/978-1-4419-9326-7_5

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, *88*(1), 243–251. https://doi.org/10.1890/0012-9658(2007)88[243:btfema]2.0.co;2

Elith, J., & Leathwick, J. (2017). *Boosted Regression Trees for ecological modeling*. R documentation. Retrieved from https://cran.r-project.org/web/packages/dismo/vignettes/brt.pdf

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Frank, A. S., Wardle, G. M., Dickman, C. R., & Greenville, A. C. (2014). Habitat-and rainfall-dependent biodiversity responses to cattle removal in an arid woodland–grassland environment. *Ecological Applications*, *24*(8), 2013–2028. https://doi.org/10.1890/13-2244.1

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*(9), 1365–1381. https://doi.org/10.1002/sim.1501

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., & Brown, J. H. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(22), 7774–7778. https://doi.org/10.1073/pnas.0803070105

Galkin, F., Aliper, A., Putin, E., Kuznetsov, I., Gladyshev, V. N., & Zhavoronkov, A. (2018). Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. *bioRxiv*, 507780.

Gelfand, A. E., Schmidt, A. M., Wu, S., Silander, J. A., Latimer, A., & Rebelo, A. G. (2005). Modelling species diversity through species level hierarchical modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(1), 1–20. https://doi.org/10.1111/j.1467-9876.2005.00466.x

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678. https://doi.org/10.1007/s11222-016-9646-1

He, T., Lamont, B. B., & Pausas, J. G. (2019). Fire as a key driver of Earth's biodiversity. *Biological Reviews*, *94*(6), 1983–2010. https://doi.org/10.1111/brv.12544

Jevšenak, J., & Skudnik, M. (2021). A random forest model for basal area increment predictions from national forest inventory data. *Forest Ecology and Management*, *479*, 118601. https://doi.org/10.1016/j.foreco.2020.118601

Kahiu, M. N., & Hanan, N. P. (2018). Fire in sub-Saharan Africa: The fuel, cure and connectivity hypothesis. *Global Ecology and Biogeography*, *27*(8), 946–957. https://doi.org/10.1111/geb.12753

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

Lucas, T. C. (2020). A translucent box: Interpretable machine learning in ecology. *Ecological Monographs*, *90*(4), e01422. https://doi.org/10.1002/ecm.1422

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., & Thuiller, W. (2009). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, *15*(1), 59–69. https://doi.org/10.1111/j.1472-4642.2008.00491.x

Molnar, C. (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Molnar, C., Casalicchio, G., & Bischl, B. (2018). Iml: An r package for interpretable machine learning. *The Journal of Open Source Software*, *3*(786), 10–21105. https://doi.org/10.21105/joss.00786

Moya-Laraño, J., & Corcobado, G. (2008). Plotting partial correlation and regression in ecological studies. *Web Ecology*, *8*(1), 35–46. https://doi.org/10.5194/we-8-35-2008

Nicodemus, K. K. (2011). Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics*, *12*(4), 369–373. https://doi.org/10.1093/bib/bbr016

Perry, G. L., & Dickson, M. E. (2018). Using machine learning to predict geomorphic disturbance: The effects of sample size, sample prevalence, and sampling strategy. *Journal of Geophysical Research: Earth Surface*, *123*(11), 2954–2970.

Peterson, D. W., & Reich, P. B. (2008). Fire frequency and tree canopy structure influence plant species diversity in a forest-grassland ecotone. *Plant Ecology*, *194*(1), 5–16. https://doi.org/10.1007/s11258-007-9270-4

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(3), 252–264. https://doi.org/10.1109/34.75512

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Ross, C. W., Grunwald, S., Vogel, J. G., Markewitz, D., Jokela, E. J., Martin, T. A., Bracho, R., Bacon, A. R., Brungard, C. W., & Xiong, X. (2020). Accounting for two-billion tons of stabilized soil carbon. *Science of the Total Environment*, *703*, 134615. https://doi.org/10.1016/j.scitotenv.2019.134615

Ross, C. W., Hanan, N. P., Prihodko, L., Anchang, J., Ji, W., & Yu, Q. (2021). Woody-biomass projections and drivers of change in sub-Saharan Africa. *Nature Climate Change*, *11*(5), 449–455. https://doi.org/10.1038/s41558-021-01034-5

Sankaran, M., Ratnam, J., & Hanan, N. (2008). Woody cover in African savannas: The role of resources, fire and herbivory. *Global Ecology and Biogeography*, *17*(2), 236–245. https://doi.org/10.1111/j.1466-8238.2007.00360.x

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934

Seyedzadeh, S., Rahimian, F. P., Rastogi, P., & Glesk, I. (2019). Tuning machine learning models for prediction of building energy loads.

Sustainable Cities and Society, *47*, 101484. https://doi.org/10.1016/j.scs.2019.101484

Stegen, J. C., Ferriere, R., & Enquist, B. J. (2012). Evolving ecological networks and the emergence of biodiversity patterns across temperature gradients. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1731), 1051–1060. https://doi.org/10.1098/rspb.2011.1733

Stockwell, D. R., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*(1), 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 1–21. https://doi.org/10.1186/1471-2105-8-25

Strobl, C. & Zeileis, A. (2008). Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance. Department of Statistics: Technical Reports, No.17. https://doi.org/10.5282/ubm/epub.2111

Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. (2012). Making machine learning models interpretable. *ESANN*, *12*, 163–172.

Wang, H., Yang, F., & Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics*, *17*(1), 1–18. https://doi.org/10.1186/s12859-016-0900-5

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*(2), 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x

Yu, Q., Ji, W., Prihodko, L., Ross, C. W., Anchang, J. Y., & Hanan, N. P. (2021). Data from: Study becomes insight: Ecological learning from machine learning. *Zenodo*. https://zenodo.org/badge/latestdoi/387275079

Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, *39*(1), 272–281.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Yu, Q., Ji, W., Prihodko, L., Ross, C. W., Anchang, J. Y., & Hanan, N. P. (2021). Study becomes insight: Ecological learning from machine learning. *Methods in Ecology and Evolution*, 00, 1–12. https://doi.org/10.1111/2041-210X.13686