



AUGUST 23 2023

# Feature selection for a continental-scale geospatial model of environmental sound levels

Katrina Pedersen ; Mark K. Transtrum; Kent L. Gee ; Shane V. Lympany; Michael M. James; Alexandria R. Salton



*J Acoust Soc Am* 154, 1168–1178 (2023)

<https://doi.org/10.1121/10.0020659>



View  
Online



Export  
Citation

CrossMark

## Related Content

K-Means clustering of inputs to a geospatial model for optimizing acoustic data collection

*Proc. Mtgs. Acoust* (November 2018)

A geospatial model of the global ambient soundscape

*J Acoust Soc Am* (October 2019)

Clustering analysis of inputs to a geospatial model of outdoor ambient sound

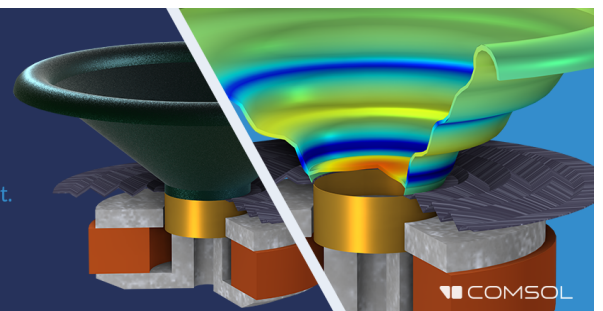
*J Acoust Soc Am* (September 2018)

25 August 2023 14:55:07

## Take the Lead in Acoustics



The ability to account for coupled physics phenomena lets you predict, optimize, and virtually test a design under real-world conditions – even before a first prototype is built.

» Learn more about COMSOL Multiphysics®



COMSOL

## Feature selection for a continental-scale geospatial model of environmental sound levels

Katrina Pedersen,<sup>1,a)</sup>  Mark K. Transtrum,<sup>1</sup> Kent L. Gee,<sup>1</sup>  Shane V. Lympany,<sup>2</sup> Michael M. James,<sup>2</sup> and Alexandria R. Salton<sup>2</sup>

<sup>1</sup>Physics and Astronomy, Brigham Young University, Provo, Utah 84602, USA

<sup>2</sup>Blue Ridge Research and Consulting, LLC, Asheville, North Carolina 28801, USA

### ABSTRACT:

Modeling environmental sound levels over continental scales is difficult due to the variety of geospatial environments. Moreover, current continental-scale models depend upon machine learning and therefore face additional challenges due to limited acoustic training data. In previous work, an ensemble of machine learning models was used to predict environmental sound levels in the contiguous United States using a training set composed of 51 geospatial layers (downselected from 120) and acoustic data from 496 geographic sites from Pedersen, Transtrum, Gee, Lympany, James, and Salton [JASA Express Lett. **1**(12), 122401 (2021)]. In this paper, the downselection process, which is based on factors such as data quality and inter-feature correlations, is described in further detail. To investigate additional dimensionality reduction, four different feature selection methods are applied to the 51 layers. Leave-one-out median absolute deviation cross-validation errors suggest that the number of geospatial features can be reduced to 15 without significant degradation of the model's predictive error. However, ensemble predictions demonstrate that feature selection results are sensitive to variations in details of the problem formulation and, therefore, should elicit some skepticism. These results suggest that more sophisticated dimensionality reduction techniques are necessary for problems with limited training data and different training and testing distributions.

© 2023 Acoustical Society of America. <https://doi.org/10.1121/10.0020659>

(Received 31 January 2023; revised 9 June 2023; accepted 30 July 2023; published online 23 August 2023)

[Editor: D. Keith Wilson]

Pages: 1168–1178

## I. INTRODUCTION

### A. Geospatial acoustics modeling

Environmental sound is the accumulation of all sounds in an outdoor environment and is often described via statistical or exceedance sound levels, such as the  $L_{10}$ ,  $L_{50}$ , and  $L_{90}$ , where the  $L_{NN}$  is the sound level exceeded NN% of the time. Quantifying anthropogenic noise, or unwanted or harmful sound from human activity, is of particular importance because such noise has been associated with adverse health effects in humans and wildlife.<sup>1,2</sup> Indeed, public health studies have found that increased noise may be associated with changes in blood pressure, heart rate, and stress<sup>3,4</sup> as well as mental health,<sup>5</sup> cognitive function,<sup>6</sup> and mental illnesses, such as depression and anxiety.<sup>7</sup> Ecologists have also studied the effects of environmental noise on animal behavior,<sup>8–12</sup> with a marked focus on animals that respond to sound, such as birds,<sup>10,11,13,14</sup> marine life,<sup>8,9,12,15,16</sup> and anurans (i.e., frogs and toads).<sup>17</sup> Given the negative impacts associated with anthropogenic noise, accurate modeling of continental-scale environmental sound levels has many potential applications, including aiding ecologists and public health workers in further identifying relationships between ecological and public health trends, respectively, and environmental noise. Additionally, accurate prediction of outdoor sound levels

may assist the National Park Service (NPS) in their charge to protect and restore natural acoustic environments within parks<sup>18,19</sup> and have applications in real estate, urban planning, and social justice.<sup>20–22</sup>

Direct, physical modeling of continental-scale environmental sound levels is difficult due to the multitude of possible acoustic effects, including diverse sources, barriers to propagation, etc., and variation in sound levels with time of day or season. Alternatively, one could conceive of modeling outdoor environmental sound levels utilizing geospatial data and physics-based acoustical models alone. Indeed, remote sensing has increased the amount of geospatial data available for modeling. However, these data are not free from error<sup>23</sup> and may not contain sufficient information to characterize an acoustic environment. Additionally, physics-based models do not exist for all acoustic effects (e.g., river sound). Therefore, current continental-scale modeling approaches rely on machine-learning techniques.<sup>24–29</sup>

In previous work, we validated two continental-scale machine learning models: one published by the NPS<sup>24,30</sup> and one generated by the current authors.<sup>29</sup> To the best of our knowledge, these are the only two models of continental-scale environmental sound levels. Both models use geospatial layers as inputs and the summer daytime A-weighted  $L_{50}$  as a validation metric over the contiguous United States (CONUS). Interestingly, the models differ by more than 20 dBA (roughly a factor of four in loudness) at some test

<sup>a)</sup>Email: [katrina.pedersen@gmail.com](mailto:katrina.pedersen@gmail.com)

locations in the CONUS, despite both models having comparable leave-one-out (LOO) validation metrics and considerable overlap between the geospatial and acoustic datasets used to create them. Additionally, holdout validation errors for both models are much larger than LOO validation metrics would predict. These large errors are attributed to limited acoustic training data (fewer than 500 unique geographic sites for either model), forcing models to make predictions in extrapolation regions (i.e., regions geospatially dissimilar to training environments). Indeed, although some initial results are promising,<sup>24–29</sup> the problem of continental-scale environmental sound level modeling is bottlenecked by limited acoustic data (due to the cost of data collection), making accurate modeling and model validation studies<sup>29</sup> more challenging.

We note that model overfitting and an incomplete set of relevant features may also contribute to discrepancies between both different model predictions (e.g., the NPS model and our model described in Ref. 29) and LOO and holdout errors. However, it is difficult to address these issues without first obtaining a larger and more diverse acoustic data set from which to train and validate models. More generally, we acknowledge that we do not have sufficient acoustic data to validate the model over the entire space for which we would like to make predictions (i.e., the CONUS). Despite this, modeling sound levels over continental scales is an important problem and it is worthwhile to investigate how we can improve current modeling efforts and help refine the problem.

Dimensionality reduction methods are commonly used in machine learning applications to reduce the dimensionality of the feature space (i.e., reduce the number of features). Such methods may reduce the size of extrapolation regions and lead to better agreement between training and test errors. In this paper, we explore one type of dimensionality reduction (feature selection) to determine appropriate approaches for dimensionality reduction for continental-scale modeling of environmental sound levels with limited acoustic data. In the remainder of Sec. I, we first describe the motivation for dimensionality reduction. We then describe previous work in continental-scale outdoor sound level modeling, taking note of any feature selection processes. Last, we provide a discussion of the validation error metric used in previous continental-scale sound level models (LOO cross-validation) and give an overview of the remainder of this paper.

## B. Dimensionality reduction

Dimensionality reduction, or the process of reducing the number of features in a dataset, has several benefits for machine learning, such as minimizing the curse of dimensionality, improving model accuracy and interpretability, decreasing computational and data requirements, and reducing uncertainty. The curse of dimensionality refers to challenges that occur when analyzing data in high-dimensional spaces that do not occur in low-dimensional spaces.<sup>31</sup>

Although data may be dense in low-dimensional spaces, data become increasingly sparse as the number of dimensions (or features) increases. Machine learning identifies patterns and trends in data, so problems have higher data requirements in high-dimensional spaces. Since acoustic data are limited in the case of outdoor sound level modeling on continental scales, it is likely that the sparsity of data is a challenge and contributes to the large differences between model predictions and holdout errors in Ref. 29. Model overfitting and an incomplete description of relevant features may also contribute to these large differences.

There are additional benefits to dimensionality reduction when training data are limited, as is the case for continental-scale environmental sound models. For example, models are more sensitive to noise (i.e., meaningless variation in data) when data are limited, so removing geospatial data that have large errors may improve model accuracy and reduce uncertainty estimates. Another potential challenge of modeling limited data is that models may be prone to use correlated, rather than causal, features for predictions. If both a causal feature and a correlated feature (i.e., a feature highly correlated with the causal feature) exist in the dataset, it is beneficial to use the causal feature, which will likely generalize better in extrapolation regions. If a causal feature does not exist in the dataset, it is desirable to use a correlated feature whose correlation extends to extrapolation regions (i.e., the CONUS in our case). However, with limited data, machine learning can latch onto spurious correlations that do not hold over the entire domain of interest. This likely contributed significantly to large discrepancies between the two continental-scale sound level models validated in Ref. 29 in some regions atypical of training data, such as the Great Salt Lake in Utah.

Dimensionality reduction can be divided into two types: feature extraction and feature selection.<sup>32</sup> Feature extraction methods, such as principal component analysis or manifold learning (e.g., diffusion maps or t-distributed stochastic neighbor embedding), transform feature vectors into a lower-dimensional space while attempting to minimize information loss, and feature selection methods identify a subset of the original features to use. Feature extraction methods have the benefit that they can identify lower-dimensional representations without removing entire features; however, they lose physical interpretability. As an initial investigation into dimensionality reduction for continental-scale environmental sound modeling, this work focuses on feature selection.

## C. Previous work

The two primary attempts to generate continental-scale environmental sound level maps were independently undertaken by the NPS<sup>24–27,30</sup> and the present authors.<sup>28,29</sup> Both methods applied some rudimentary feature selection as part of the model-building process. The NPS model was trained after applying feature selection to a set of 115 geospatial layers by first removing features with high Pearson

correlation coefficients.<sup>24,27</sup> Features were then removed one at a time, using the out-of-bag error due to permuting features to measure relative feature importance and removing the least important feature each iteration. The optimal number of features was determined by calculating the LOO root mean square error (RMSE) for all feature subsets and identifying the subset with the lowest LOO RMSE. After identifying the reduced feature set that minimized the LOO RMSE using default hyperparameters, five random forest hyperparameters were tuned. It was noted that although this process may not produce the best feature subset, it is computationally tractable.

Our previous models<sup>28,29</sup> (and those described in this paper) consist of an ensemble of six different machine learning models, each from a different model class: gradient-boosted regression trees (GBRs), neural networks (NNs), k-nearest neighbors (KNN), support vector machines (SVMs), kernel ridge regression (KRR), and Gaussian process regression (GPR). Most recently, we used a feature subset of 51 geospatial features<sup>29</sup> (downselected from 120), and we use that same subset in this paper. Model hyperparameters were tuned to minimize the LOO median absolute deviation (MAD). All six models had comparable LOO MAD errors. The median of the ensemble model predictions was used to make environmental sound level predictions for the CONUS. We provide additional information explaining the downselection process and apply further feature selection to the set of 51 features in this paper.

#### D. LOO cross-validation

Both the NPS model<sup>24–27</sup> and models generated by the present authors<sup>28,29</sup> utilize LOO cross-validation to measure model performance. The LOO cross-validation error is computed by removing each site from the training data (one at a time), training a model on the remaining data, and then calculating the residual for the corresponding site. The LOO RMSE and MAD are the RMSE and MAD of all residuals, respectively. Although LOO cross-validation is more computationally expensive than other validation metrics, it is often more appropriate for small datasets, particularly when each training site may provide unique information to the model. Additionally, since data are limited, the computational costs of computing LOO cross-validation errors are not unreasonably large. However, using the LOO cross-validation to estimate model performance for extrapolation regions is not advisable.

The LOO cross-validation error assumes new input data are drawn from the same distribution as the training data. For the case of geospatial environmental acoustic modeling, this assumption does not hold due to a combination of limited acoustic data and a biased distribution of acoustic data. For example, many training sites (65%) are from national parks; however, national parks comprise only a small percentage of the total CONUS land area. Hence, LOO errors are not likely to be a good estimate of model uncertainty for much of the CONUS. Indeed, we previously found that

LOO statistical metrics may differ significantly from hold-out validation errors.<sup>29</sup> The problem of estimating model transferability, or the ability of a model to make accurate predictions for data statistically different from the training data, is an open area of research.<sup>33,34</sup>

Although LOO validation metrics are not good indicators of model performance in extrapolation regions (i.e., model transferability), they are computationally tractable and can indicate model accuracy on data drawn from a similar distribution as the training data. Hence, in the absence of more acoustic training data, it is reasonable to select model hyperparameters which minimize the LOO MAD or RMSE. Note that we minimize the LOO MAD rather than the LOO RMSE because it is less sensitive to outliers; however, either is a reasonable choice.

#### E. Paper overview

In this paper, we extend our ensemble approach for continental-scale environmental sound modeling and explore four feature selection methods. Results of feature selection may improve model accuracy and interpretability, as well as determine appropriate approaches for dimensionality reduction for continental-scale modeling of environmental sound levels with limited acoustic data.

We consider 120 geospatial features in the modeling process and explain the motivation for the manual removal of 69 of those features. With the remaining 51 features, we then compare four feature selection methods. Each method identifies a reduced set of 15 features which give ensemble LOO MAD errors similar to those of all 51 features. Finally, we use an ensemble approach to calculate environmental sound level predictions in the CONUS from the reduced-order models. Our results show that the predictions of the reduced-order models depend strongly upon details in the problem formulation, including the feature importance metric (i.e., the feature selection method). This analysis indicates that more sophisticated dimensionality reduction techniques are required to take advantage of the benefits of dimensionality reduction (e.g., minimizing the curse of dimensionality) and improve model accuracy.

## II. METHODS

### A. Datasets

Data used in the modeling process were composed of both geospatial and acoustic data. The initial set of geospatial data contained 120 geospatial features described in Table I of the supplementary material in Ref. 29. Acoustic data for the summer daytime A-weighted  $L_{50}$  were obtained for 496 unique sites and are also described in the supplementary material for Ref. 29. These acoustic data and the geospatial data corresponding to the training sites are used to train the supervised machine learning models of environmental sound levels. We focus on predicting the summer daytime A-weighted  $L_{50}$  in this paper since much of the previous work in environmental sound level modeling has used this metric.



TABLE I. Subset of 51 geospatial layers used for environmental sound level modeling and feature selection in this paper.

Barren (200 m)	DistCoast	Herbaceous (5 km)	Slope
Barren (5 km)	DistMilitary	MilitarySum (40 km)	TdewAvgSummer
Cultivated (200 m)	DistRailroads	MixedForest (200 m)	TdewAvgWinter
Cultivated (5 km)	DistRoadsAll	MixedForest (5 km)	TMaxSummer
Deciduous (200 m)	DistRoadsMaj	PopDensity	TMaxWinter
Deciduous (5 km)	DistStreamO1	PPTSummer	TMinSummer
Developed (200 m)	DistStreamO3	PPTWinter	TMinWinter
Developed (5 km)	DistStreamO4	RddAll	VIIRSMean (270 m)
DistAirpHeli	Elevation	RddAll (5 km)	Water (200 m)
DistAirpHigh	Evergreen (200 m)	RddMajor	Water (5 km)
DistAirpLow	Evergreen (5 km)	RddMajor (5 km)	Wetlands (200 m)
DistAirpMod	FlightFreq (25 km)	Shrubland (200 m)	Wetlands (5 km)
DistAirpMoto	Herbaceous (200 m)	Shrubland (5 km)	

We note that the geospatial and acoustic datasets are the result of considerable time and effort and it is generally not trivial to add acoustic training data or geospatial features. Acoustic data are obtained by averaging sound level meter measurements at a given location (i.e., training site) over a minimum of two to three days, and often close to two weeks, depending on the variability of sound levels at the site. Geospatial features, on the other hand, require some processing (to map values to the same raster points used in the geospatial dataset), may be difficult or expensive to acquire, and often require significant quality review.

### B. Initial reduction to 51 features and feature scaling

Prior to utilizing feature importance metrics for feature selection, a quality review was performed for the existing 120 geospatial features. Feature processing, areas of analyses, sources of error, correlations with other features, and possible correlations to ambient sound levels were considered. Features were removed if they had large errors or uncertainties or considerable correlations with other features. Features were also removed if their quality was questionable due to poor documentation, or if physical intuition indicated they should be weak predictors of outdoor sound (e.g., latitude and longitude). After this initial feature selection, 51 of the original features remained; these are the features used to produce the ensemble models in Ref. 29. Limited explanation was given in Ref. 29 for why features were removed, so we provide a more detailed explanation here. We note that this initial downsampling is an essential part of the overall feature selection process because it removes features with known issues/noise.

The 120 feature names are described in Table 1 of the supplementary material for Ref. 29, and the reduced set of 51 features is listed in Table 2 of the same supplementary material. For convenience, we repeat Table 2 from Ref. 29 here as Table I. For further explanation of the features, we refer the reader to the supplementary material of Ref. 29.

The original set of 120 features included features that describe the land use, land cover, nighttime lights, climate, transportation noise, distances to potential sound sources, and geographic location. All land-use features were removed due

to possible errors in the layers, significant correlation with many of the land cover layers, and poor documentation. Many land-use layers have sharp, unphysical discontinuities, possibly due to errors in the layers. For example, the Cropland layers have some unphysical-looking discontinuities in southeastern North Dakota. All VIIRS layers (i.e., layers describing the amount of light at night), except the VIIRS mean upward radiance at night layer with a 270 m area of analysis, were removed due to large correlations with each other. (The area of analysis is the radius of the circular or cylindrical region from which data were processed to produce the value at a given site.) The annual precipitation, minimum and maximum temperatures, and dew points were removed due to high correlations with the corresponding summer and winter layers. The RoadNoise and AviationNoise layers were removed because both are discontinuous for values below 35 dB. The Forest land cover layer was omitted because it is further divided into the Deciduous, MixedForest, and Evergreen forest land cover layers, which are included in the subset of 51 features. The DistWaterBody layer, which gives the distance to the nearest body of water, and the PhysicalAccess layer, which indicates how accessible an area is, given transportation infrastructure and off-trail conditions, were removed due to concerns about errors in the layers. More specifically, some parts of rivers appear to be classified as bodies of water while others do not, and some values in the DistWaterBody layer, particularly in California, Nevada, and Arizona, appear suspicious. The PhysicalAccess layer, on the other hand, had some extremely large unphysical values. The DistAirpSea layer was removed due to high correlations with the other DistAirp features. Last, Latitude and Longitude were removed since they should generally not have a physical effect on the acoustic environment.

In addition to reducing the feature set to 51 features, we rescaled geospatial features based on physical arguments. We scaled most features (i.e., all that do not depend on distance) based on their distribution within the CONUS as opposed to their distribution in the training data. For these features, we used min-max scaling, which scales data to be between zero and one and preserves the shape of the data distribution. For some feature vector  $\mathbf{x}$ , the scaled vector is given by

$$\mathbf{x}_{scaled} = \frac{\mathbf{x} - x_{min}}{x_{max} - x_{min}}, \tag{1}$$

where  $x_{min}$  and  $x_{max}$  are the minimum and maximum values of the feature  $\mathbf{x}$ , respectively. For geospatial features that rely on distances, however, such as DistAirpHigh and DistCoast, we used an arctangent function to scale data to be between zero and one,

$$\mathbf{x}_{scaled} = \frac{2}{\pi} \arctan \frac{\mathbf{x}}{x_0}, \tag{2}$$

where  $x_0$  varies for different features and determines how quickly the scaling function approaches one. Arctangent functions were selected for scaling of distance-dependent features because it is expected that after the distance exceeds some threshold, the feature’s effect upon the environmental sound levels will not change. For example, after reaching 40 km from the nearest road, it is unlikely that the distance to the road will provide relevant information for environmental sound level predictions. Note that there is some ambiguity in the choice of appropriate distance thresholds ( $x_0$ ) and further refinements to scaling methods may be considered in the future. However, we believe these scaling methods are an improvement to those used in Ref. 28, which scaled all features to have a mean of zero and standard deviation of one.

### C. Feature importance metrics

All feature importance metrics have pros and cons—some of which are described here—and there is no single “best” way to calculate feature importance. Previously, the NPS model utilized the change in out-of-bag error due to permuting feature values in random forest models to estimate feature importance.<sup>27</sup> One disadvantage of permutation methods is that they often force models to focus on extrapolation regions that may not correspond to allowed areas of the feature space.<sup>35</sup> Additionally, feature importance metrics, including permutation methods, tend to over-emphasize correlated features.<sup>36</sup> To investigate the stability of feature selection results for different feature importance metrics, we compare four different feature importance metrics, namely, Gini importance, Gini importance with a correlation penalty, neural network weights, and expert intuition.

The Gini importance metric, or mean decrease impurity, is a common feature importance measure for random forest or GBR models. The Gini importance for a given feature is calculated using the error reduction and number of instances split at each node corresponding to that feature. Although we do not provide an explanation of how to compute the Gini importance here, the interested reader is referred to Refs. 37 and 38 for further details. The Gini importance is fast to calculate, but it is often biased in favor of features with higher cardinality or variability.<sup>39</sup> Additionally, the Gini importance has sometimes been shown to be biased towards correlated features.<sup>36</sup> For further information regarding the pros and cons of feature importance metrics for decision tree models, including the Gini importance and

permutation method used for the NPS model, the reader is referred to Refs. 40 and 41.

To reduce bias due to correlations among features, which can be quite large, we introduce a correlation penalty to modify the Gini importance metric. After the calculation of the Gini importance for a given GBR model, we iterated through all features. For each feature  $x_i$ , the most strongly correlated feature  $x_{corr}$  was identified. If the given feature  $x_i$  had a higher Gini importance than the correlated feature  $x_{corr}$ , its Gini importance was unchanged. Otherwise, its Gini importance was decreased by a factor of  $(1 - \text{corr}_{max})$ , where  $\text{corr}_{max}$  was the correlation corresponding to the maximally correlated feature  $x_{corr}$ . If two features are almost identical, this metric will strongly penalize the feature with the lower Gini importance, giving it an importance score near zero, while leaving the Gini importance of the other feature unchanged. This feature importance metric has similar pros and cons to the original Gini importance metric, but it penalizes correlated features to guide feature selection toward a set of independent (rather than strongly correlated, redundant) features.

The third feature importance metric is determined using the trained NN weights. There is no standard way to measure feature importance in a NN, but many methods have been suggested.<sup>42</sup> We quantified the NN feature importance by first identifying all paths from an input feature to the output and calculating the product of all weights along each path. Then, for each feature, the absolute value of all paths originating at that feature were summed together. Finally, these sums were normalized, and the results were used as a feature importance measure. For the case of zero hidden layers, the feature importance was determined by the magnitude of the weights from the input features to the output.

The last feature importance metric is purely subjective. An expert familiar with environmental acoustic modeling used maps of the geospatial features, information about their processing methods, areas of analysis, correlations to other features, etc., to select which features would be most important for determining CONUS environmental sound levels.

### D. Feature selection process

After reducing the geospatial feature set from 120 to 51 features, we tuned model hyperparameters for the six supervised machine learning model classes (GBRs, NNs, KNN, SVMs, KRR, and GPR) to minimize the LOO MAD. Model hyperparameters are settings that a user selects for the learning process, such as the learning rate or activation function in a neural network. We used the tree-structured Parzen estimator approach implemented in hyperopt,<sup>43,44</sup> a Python library for automatic hyperparameter tuning, to determine appropriate hyperparameters. This approach tunes hyperparameters with minimal supervision so that we can periodically retune hyperparameters at different stages of feature selection. Hyperparameter search spaces were adjusted occasionally to account for varying feature subsets.



FIG. 1. (Color online) LOO MAD errors for the summer daytime A-weighted  $L_{50}$  as a function of the number of features. Model hyperparameters were tuned at 51, 40, 30, 20, 15, 10, 5, 4, 3, 2, and 1 feature(s).

After tuning hyperparameters to minimize the LOO MAD and training all six members of the ensemble for the 51-feature model, we applied the four feature importance metrics described in the previous subsection to remove one feature at a time. For each metric, feature importance was calculated and the least important feature was removed to create four different feature subsets of size 50. All six model classes were retrained (using the hyperparameters identified from the 51-feature model) and the LOO MAD was calculated for all 24 models (6 models per subset of 50 features). At this point, there were four ensembles, each corresponding to a feature importance metric. For each ensemble, the corresponding feature importance metric was used to identify and then remove the least important feature again. This process was repeated every time a feature was removed.

Varying the number of geospatial features will change the optimal hyperparameters; therefore, model hyperparameter tuning was performed at 51, 40, 30, 20, 15, 10, 5, 4, 3, 2, and 1 feature(s).

### III. RESULTS

#### A. Changes in the LOO MAD error

Figure 1 shows the LOO MAD vs the number of features for each of the four feature importance metrics and the

six models that compose each ensemble model. All models were trained to predict the summer daytime A-weighted  $L_{50}$ . Figure 2 similarly shows the LOO MAD for the ensemble models, which are determined by the median prediction of all six members. Recall that hyperparameters were tuned at

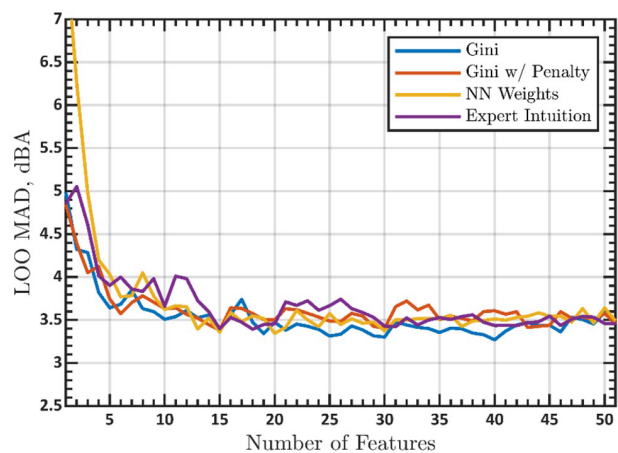


FIG. 2. (Color online) LOO MAD ensemble errors for the four metrics of determining feature importance as a function of the number of features. All models were trained to predict the summer daytime A-weighted  $L_{50}$  and ensemble predictions were determined by the median predicted level of the six ensemble members.



51, 40, 30, 20, 15, 10, 5, 4, 3, 2, and 1 feature(s). Models corresponding to the four feature importance metrics perform similarly, especially for larger feature sets. It is interesting that the number of features can be reduced significantly from the 51 initial features without much change to the LOO MAD, possibly due to large correlations between the geospatial features. However, the LOO MAD estimates the expected error for predictions made on statistically similar inputs, so, although models may perform well on the training data, LOO MAD makes no guarantees as to how the models will generalize to data drawn from a different distribution (i.e., the CONUS).

To further investigate the models generated during feature selection, we analyzed the reduced feature sets of 15 features and their corresponding CONUS ensemble predictions. We selected feature subsets of size 15 because all four ensembles had relatively low LOO MAD values there (likely in part due to hyperparameter tuning) and LOO MAD errors tended to start increasing as features continued to be removed.

### B. Comparison of top 15 features

Table II lists the top fifteen features identified by the four feature importance metrics in order of importance. The features in the bottom row would have been removed next and the features on the top row were the last remaining features used to train the 1-feature models. Interestingly, all four feature subsets include a feature that gives information about the distance to the nearest road, the distance to the nearest stream, and the amount of evergreen land cover. Two of the three subsets determined by feature importance metrics also include information about the mean upward radiance at night (VIIRS layer), road density, and the amount of shrubland and herbaceous land cover, all of which are represented in the expert’s list. It is encouraging that all feature subsets share some similarities with the

expert’s list, indicating that the individual feature lists are not *a priori* unreasonable.

Despite the similarities in the feature subsets, there are many significant differences. Each of the four subsets contains at least five unique feature layers, with the subset corresponding to the Gini importance with a correlation penalty containing the most (eight) unique layers. The subset corresponding to the Gini importance metric only contains one land cover feature while the three other subsets contain between five and seven land cover features each. Additionally, the expert favors land cover layers with a 200 m area of analysis while the Gini importance with a correlation penalty favors land cover layers with a 5000 m area of analysis. Given the many differences between the reduced feature sets, it is interesting that they all give comparable ensemble LOO MAD errors. This is likely due in part to limited training data and correlations among the geospatial features.

Further information regarding model behavior can be gained by looking at ensemble predictions. Figure 3 shows the ensemble predictions for the summer daytime A-weighted  $L_{50}$  for the 15-feature reduced feature sets corresponding to the features listed in Table II. Training sites are indicated by small circles and colored according to measured levels. Even though the four 15-feature ensemble models give similar LOO MAD error measures, CONUS ensemble predictions vary significantly among the four ensembles, indicating possible overfitting or training on feature sets that are not able to characterize the whole feature space relevant for outdoor sound level modeling. To emphasize the differences, the spread of ensemble model predictions for all four feature importance metrics at all sites in the CONUS is plotted in a histogram in Fig. 4. A histogram showing the spread of the four ensemble model fit predictions (i.e., predictions at training sites) is overlaid for comparison. The spread of ensemble predictions in the CONUS has a mean, median, and maximum of 6.9, 6.5, and 29.8 dBA,

TABLE II. Top 15 features identified by various feature importance metrics (re-ranking after removing each lowest-ranked feature using the 15-, 10-, 5-, 4-, 3-, and 2-feature tuned models).

Gini metric	Gini metric with correlation penalty	Neural network weights	Expert intuition
TdewAvgSummer	VIIRSMean (270 m)	TMinWinter	VIIRSMean (270 m)
VIIRSMean (270 m)	DistCoast	Water (5 km)	RddAll
Slope	DistRoadsMaj	Barren (5 km)	DistRoadsMaj
DistRoadsMaj	Shrubland (5 km)	RddAll (5 km)	DistStreamO3
DistStreamO3	PopDensity	Evergreen (5 km)	FlightFreq (25 km)
Evergreen (5 km)	Slope	DistRoadsMaj	PopDensity
DistMilitary	DistMilitary	Developed (200 m)	DistRailroads
PPTWinter	TMaxWinter	Barren (200 m)	Cultivated (200 m)
DistStreamO1	DistStreamO3	RddAll	Deciduous (200 m)
Elevation	Wetlands (5 km)	DistStreamO1	Wetlands (200 m)
RddAll (5 km)	Evergreen (5 km)	DistAirmoto	Herbaceous (200 m)
DistAirmoto	DistRoadsAll	Shrubland (200 m)	Shrubland (200 m)
DistAirmoto	Herbaceous (5 km)	FlightFreq (25 km)	Evergreen (200 m)
PPTSummer	TMaxSummer	DistStreamO4	Developed (200 m)
DistAirmoto	Deciduous (5 km)	Herbaceous (200 m)	TdewAvgSummer



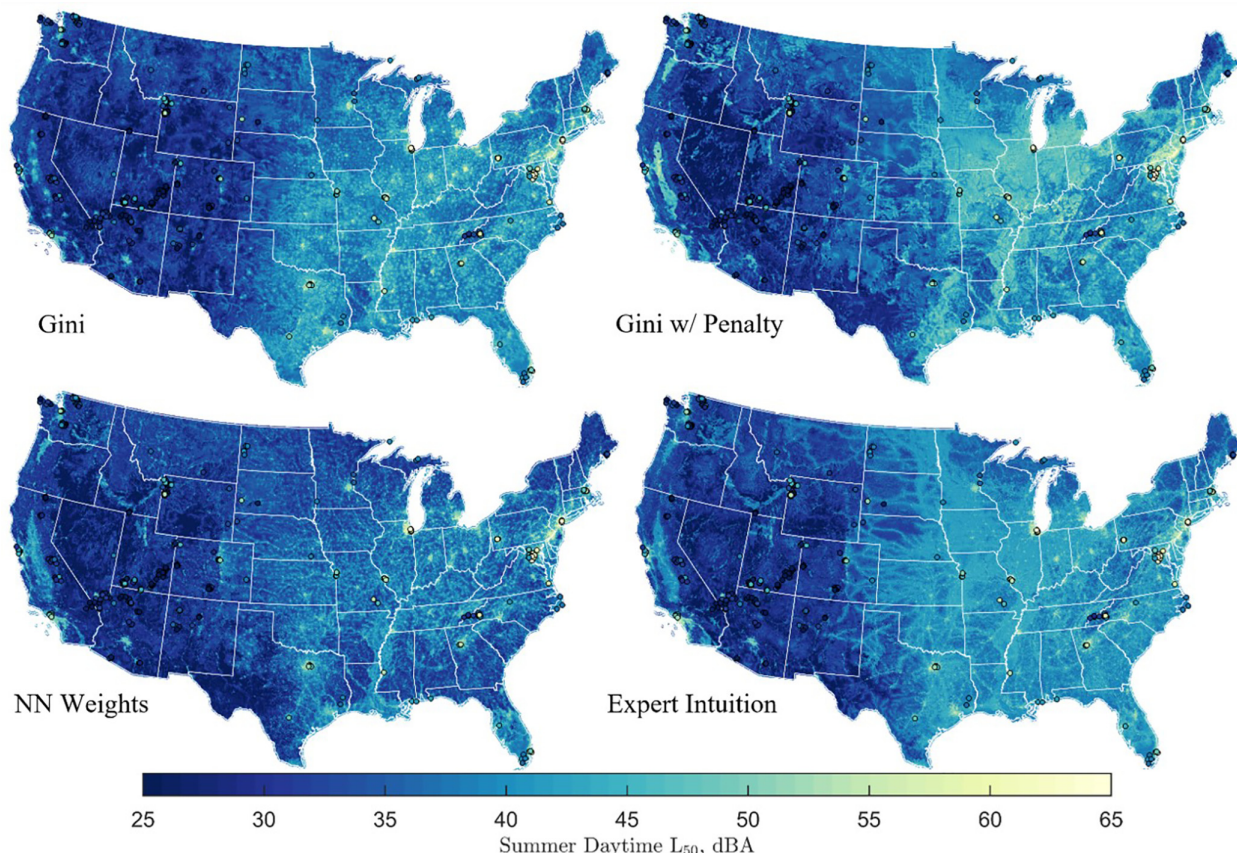


FIG. 3. (Color online) CONUS ensemble predictions of the summer daytime A-weighted  $L_{50}$  for models trained using the top 15 features identified from four different feature importance metrics. Training sites are indicated by small circles and are colored according to measured levels.

respectively, while the spread of ensemble predictions at training sites has a mean, median, and maximum of 2.1, 1.6, and 11.4 dBA, respectively. For reference, in our previous comparison of the NPS model and 51-feature model, the mean, median, and maximum absolute difference between the two models across the CONUS were 2.1, 1.6, and 21.3 dBA,

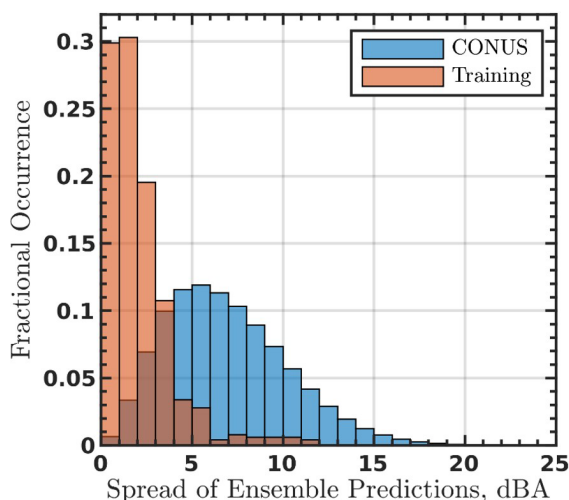


FIG. 4. (Color online) Histogram of the spread of CONUS predictions and training site predictions of the summer daytime A-weighted  $L_{50}$  for models trained using the top 15 features identified from four different feature importance metrics.

respectively, indicating a better agreement between the NPS and 51-feature model than the four 15-feature models.<sup>29</sup> Note that a difference of 6 dBA corresponds to a doubling of sound pressure level, so the differences in the CONUS predictions between the four ensemble models are not small. A difference of 1 dBA on the other hand is about the smallest perceivable change for a human being, so the spread of predictions at most training sites is relatively small. These results emphasize that the LOO MAD is not a reliable indicator of model accuracy in extrapolation regions.

More generally, these results suggest it is not beneficial to compare validation errors (e.g., LOO MAD and LOO RMSE) to select feature subsets for environmental sound level modeling on continental scales because such error measures do not describe model accuracy in extrapolation regions, and current training data are not representative of the CONUS. This points to the need for better dimensionality reduction methods that do not rely on the model’s predictive performance. Rather, dimensionality reduction techniques that attempt to characterize intrinsic dimensions of the data (i.e., feature extraction techniques<sup>32</sup>) and/or minimize shifts in data distributions between the training and test data<sup>45</sup> may be better suited to environmental sound level modeling.

**C. Sensitivity to feature selection process**

We also examined the sensitivity of results to changes in the problem formulation (e.g., how often hyperparameters

were tuned or the choice of the random seed used for hyperparameter tuning). Recall that the first two feature importance metrics are dependent upon the trained GBR model and the third metric is dependent upon the trained NN model. The last metric is independent of all models and training data, providing an interesting contrast to the other three data-driven methods of computing feature importance.

Since the Gini importance, Gini importance with correlation penalty, and importance calculated from the neural network weights all rely on trained models, varying hyperparameters affects feature importance estimates and the feature subsets identified by these metrics. Indeed, a comparison of the reduced feature sets generated using model hyperparameters tuned fewer times showed that results for these three data-driven feature importance metrics are sensitive to how often hyperparameters are tuned. For example, we found that the top 15 ranked feature subsets vary both in features and rankings when hyperparameters are only tuned at 51 and 40 features rather than at 51, 40, 30, 20, and 15 features. For both the Gini importance and Gini importance with correlation penalty, 11 of the top 15 features are the same. However, for the importance calculated from neural network weights, only seven of the top 15 features are the same. Therefore, tuning hyperparameters more or less often would result in different feature rankings and subsets. This is an example of how the feature selection results are sensitive to the details of the procedure.

There is also a certain amount of randomness in tuning hyperparameters and training models. Hence, there is some randomness in determining feature importance estimates for the three data-driven metrics. We found that both changing the random seed used to sample the hyperparameter spaces as well as making small changes to the hyperparameter search spaces resulted in different optimal hyperparameters and feature rankings.

Additionally, given the limited acoustic data, the three quantitative methods are biased by the distribution of training data and are therefore likely to select features that correlate with environmental sound levels in the training data, regardless of whether or not those correlations hold true for most of the CONUS. More particularly, we found that when Longitude was added to the set of 51 geospatial features and hyperparameters were tuned for that set of 52 features, both model predictions and feature rankings changed. Note that longitude is strongly correlated (Pearson correlation coefficient of 0.59) with the training data due to sampling bias. The mean, median, and maximum absolute differences of ensemble model predictions in the CONUS for the summer daytime A-weighted  $L_{50}$  for the sets of 51 and 52 features were 0.9, 0.7, and 11.1 dBA, respectively. Moreover, Longitude was ranked in the top 11 features for all three metrics when using hyperparameters tuned for the set of 52 features.

These results demonstrate that the feature selection results are not only sensitive to the choice of feature importance metric but also to relatively small changes in the feature selection process (e.g., the size of the search space and

frequency of hyperparameter tuning). This is further evidence that feature extraction methods may be better suited to environmental sound level modeling because they do not rely upon supervised machine learning models, hyperparameter optimization, validation error metrics, etc. Additionally, feature extraction can utilize information from all geospatial features while still reducing the dimensionality of feature space.

#### IV. CONCLUSION AND FUTURE WORK

Continental-scale environmental sound level modeling is an important but challenging problem with potential applications, including aiding the preservation of natural acoustic environments within national parks and informing ecological and public health studies. This paper has explored the viability of dimensionality reduction via feature selection for continental-scale environmental sound level modeling with limited data.

A feature set of 120 geospatial features was reduced to 51 by removing features with large errors or uncertainties, considerable correlations with other features, poor documentation, or lack of physical effect on environmental sound levels. Following the reduction to 51 features, we further reduced features using four feature importance metrics: Gini importance, Gini importance with a correlation penalty, neural network weights, and expert intuition. Feature selection was performed iteratively by training an ensemble model, determining the least important feature, as measured by each of the four feature importance metrics, and removing that feature. Hyperparameters were tuned occasionally to minimize the leave-one-out median absolute deviation. All models were trained to predict the summer daytime A-weighted  $L_{50}$ .

Leave-one-out median absolute deviation measures indicated that the cardinality of feature space could be reduced to 15 using all four feature importance metrics before error started to increase noticeably. The four feature sets differed significantly (i.e., they did not generally contain the same geospatial features). Additionally, ensemble model predictions for the contiguous United States indicated large variability in extrapolation regions among the four models. More specifically, the spread between the four ensemble models of predicted summer daytime A-weighted  $L_{50}$  levels in the contiguous United States had a mean, median, and maximum of 6.9, 6.5, and 29.8 dBA, respectively.

These results further demonstrate that traditional validation metrics, such as the leave-one-out median absolute deviation, are poor indicators of model transferability as discussed in Ref. 29. Additionally, results show that feature selection is strongly dependent upon the feature importance metric. An investigation of the sensitivity of feature selection results also showed that reduced feature sets are sensitive to details of the problem formulation. In particular, results are sensitive to the frequency of hyperparameter tuning, the hyperparameter search space, and the random seed



used to identify optimal hyperparameters. This should be cause for suspicion of feature selection for the problem of continental-scale environmental sound level modeling. Indeed, since the results of feature selection depend so strongly on variations in the problem formulation, they should not be taken seriously. This motivates the need for more sophisticated dimensionality reduction techniques. In particular, feature extraction methods that describe the intrinsic dimensionality of the data and do not rely on a model may be better suited to continental-scale environmental sound level modeling, especially if they are able to minimize data shifts between training and test data.

More generally, the results of this paper suggest that selecting feature sets to minimize training error for machine learning problems in which the training and unlabeled target data are drawn from different distributions, may result in model performance in extrapolation regions which is highly sensitive to details of the feature selection process. This is especially relevant in the case of limited training data. For such problems, it is necessary to consider dimensionality reduction methods that take into account the distribution of both the training and test data, either through feature extraction techniques or more sophisticated feature selection techniques.

#### ACKNOWLEDGMENTS

This research was supported by a U.S. Army Small Business Innovation Research (SBIR) contract to Blue Ridge Research and Consulting, LLC. We thank Dr. James Stephenson, technical monitor for the SBIR, for his thoughtful comments and feedback.

<sup>1</sup>A. V. Moudon, "Real noise from the urban environment: How ambient community noise affects health and what can be done about it," *Am. J. Prev. Med.* **37**(2), 167–171 (2009).  
<sup>2</sup>C. R. Kight and J. P. Swaddle, "How and why environmental noise impacts animals: An integrative, mechanistic review," *Ecol. Lett.* **14**(10), 1052–1061 (2011).  
<sup>3</sup>T. Bodin, M. Albin, J. Ardö, E. Stroh, P. Östergren, and J. Björk, "Road traffic noise and hypertension: Results from a cross-sectional public health survey in southern Sweden," *Environ. Health* **8**, 38 (2009).  
<sup>4</sup>T. Münzel, F. P. Schmidt, S. Steven, J. Herzog, A. Daiber, and M. Sørensen, "Environmental noise and the cardiovascular system," *J. Am. Coll. Cardiol.* **71**(6), 688–697 (2018).  
<sup>5</sup>S. A. Stansfeld, B. Berglund, C. Clark, I. Lopez-Barrio, P. Fischer, E. Öhrström, M. Á. M. Haines, J. Head, S. Hygge, I. van Kamp, and B. F. Berry, "Aircraft and road traffic noise and children's cognition and health: A cross-national study," *Lancet* **365**, 1942–1949 (2005).  
<sup>6</sup>S. P. Banbury, W. J. Macken, S. Tremblay, and D. M. Jones, "Auditory distraction and short-term memory: Phenomena and practical implications," *Hum. Factors* **43**(1), 12–29 (2001).  
<sup>7</sup>M. E. Beutel, C. Junger, E. M. Klein, P. Wild, K. Lackner, M. Blettner, H. Binder, M. Michal, J. Wiltink, E. Braehler, and T. Munzel, "Noise annoyance is associated with depression and anxiety in the general population—The contribution of aircraft noise," *PLoS One* **11**(5), e0155357 (2016).  
<sup>8</sup>P. A. Hastings and A. Sirovic, "Soundscapes offer unique opportunities for studies of fish communities," *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5866–5867 (2015).  
<sup>9</sup>L. Ruppé, G. Clément, A. Herrel, L. Ballesta, T. Décamps, L. Kéver, and E. Parmentier, "Environmental constraints drive the partitioning of the soundscape in fishes," *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6092–6097 (2015).

<sup>10</sup>B. C. Pijanowski, L. T. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napolitano, S. H. Gage, and N. Pieretti, "Soundscape ecology: The science of sound in the landscape," *BioScience* **61**(3), 203–216 (2011).  
<sup>11</sup>E. P. Derryberry, R. M. Danner, J. E. Danner, G. E. Derryberry, J. N. Phillips, S. E. Lipshutz, K. Gentry, and D. A. Luther, "Patterns of song across natural and anthropogenic soundscapes suggest that white-crowned sparrows minimize acoustic masking and maximize signal content," *PLoS One* **11**(4), e0154456 (2016).  
<sup>12</sup>S. M. Haver, H. Klinck, S. L. Nieuwkerk, H. Matsumoto, R. P. Dziak, and J. L. Miksis-Olds, "The not-so-silent world: Measuring Arctic, Equatorial, and Antarctic soundscapes in the Atlantic Ocean," *Deep Sea Res. Part I: Oceanogr. Res. Pap.* **122**, 95–104 (2017).  
<sup>13</sup>C. D. Francis, C. P. Ortega, and A. Cruz, "Noise pollution changes avian communities and species interactions," *Curr. Biol.* **19**(16), 1415–1419 (2009).  
<sup>14</sup>H. Slabbekorn and W. Halfwerk, "Behavioural ecology: Noise annoys at community level," *Curr. Biol.* **19**(16), R693–R695 (2009).  
<sup>15</sup>G. Buscaino, M. Ceraulo, N. Pieretti, V. Corrias, A. Farina, F. Filiciotto, V. Maccarrone, R. Grammauta, F. Caruso, A. Giuseppe, and S. Mazzola, "Temporal patterns in the soundscape of the shallow waters of a Mediterranean marine protected area," *Sci. Rep.* **6**, 34230 (2016).  
<sup>16</sup>F. Bertucci, E. Parmentier, G. Lecellier, A. D. Hawkins, and D. Lecchini, "Acoustic indices provide information on the status of coral reefs: An example from Moorea Island in the South Pacific," *Sci. Rep.* **6**, 33326 (2016).  
<sup>17</sup>S. Goutte, A. Dubois, and F. Legendre, "The importance of ambient sound level to characterise anuran habitat," *PLoS One* **8**(10), e78020 (2013).  
<sup>18</sup>National Park Service, "NPS Director's Order #47: Soundscape preservation and noise management," National Park Service, Washington, DC, 2000.  
<sup>19</sup>National Academy of Engineering, "Protecting national park soundscapes," National Park Service, Washington, DC, 2013.  
<sup>20</sup>E. Murphy and E. King, *Environmental Noise Pollution: Noise Mapping, Public Health, and Policy* (Newnes, Oxford, UK, 2014).  
<sup>21</sup>K. Kaliski, E. Duncan, and J. Cowan, "Community and regional noise mapping in the United States," *Sound Vib.* **41**(9), 14–17 (2007).  
<sup>22</sup>D. Łowicki and S. Piotrowska, "Monetary valuation of road noise. Residential property prices as an indicator of the acoustic climate quality," *Ecol. Indic.* **52**, 472–479 (2015).  
<sup>23</sup>R. S. Lunetta, R. G. Congalton, L. K. Fenstermaker, J. R. Jensen, K. C. McGwire, and L. R. Tinney, "Remote sensing and geographic information system data integration: Error sources and research issues," *Photogramm. Eng. Remote Sens.* **57**(6), 677–687 (1991).  
<sup>24</sup>D. Mennitt and K. Fristrup, "Influential factors and spatiotemporal patterns of environmental sound levels in the contiguous United States," *Noise Control Eng. J.* **64**(3), 342–353 (2016).  
<sup>25</sup>D. J. Mennitt and K. M. Fristrup, "Influential factors and spatiotemporal patterns of environmental sound levels," *INTER-NOISE NOISE-CONGR. Conf. Proc.* **250**(5), 2029–2040 (2015).  
<sup>26</sup>D. J. Mennitt, K. Fristrup, K. Sherrill, and L. Nelson, "Mapping sound pressure levels on continental scales using a geospatial sound model," in *Proceedings of InterNoise13*, Innsbruck, Austria (September 15–18, 2013).  
<sup>27</sup>D. Mennitt, K. Sherrill, and K. Fristrup, "A geospatial model of ambient sound pressure levels in the contiguous United States," *J. Acoust. Soc. Am.* **135**(5), 2746–2764 (2014).  
<sup>28</sup>K. Pedersen, M. K. Transtrum, K. L. Gee, B. A. Butler, M. M. James, and A. R. Salton, "Machine learning-based ensemble model predictions of outdoor ambient sound levels," *Proc. Mtgs. Acoust.* **35**, 022002 (2018).  
<sup>29</sup>K. Pedersen, M. K. Transtrum, K. L. Gee, S. V. Lympny, M. M. James, and A. R. Salton, "Validating two geospatial models of continental-scale environmental sound levels," *JASA Express Lett.* **1**(12), 122401 (2021).  
<sup>30</sup>National Park Service, "Geospatial sound modeling." <https://irma.nps.gov/Datastore/Reference/Profile/2217356> (Last viewed January, 2020).  
<sup>31</sup>R. E. Bellman, *Adaptive Control Processes: A Guided Tour*, Vol. 2045 (Princeton University Press, Princeton, NJ, 2015).  
<sup>32</sup>R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *J. Appl. Sci. Technol. Trends* **1**(2), 56–70 (2020).

- <sup>33</sup>S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.* **22**(10), 1345–1359 (2010).
- <sup>34</sup>F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE* **109**(1), 43–76 (2021).
- <sup>35</sup>G. Hooker and L. Mentch, "Please stop permuting features: An explanation and alternatives," [arXiv:1905.03151](https://arxiv.org/abs/1905.03151) (2019).
- <sup>36</sup>C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinf.* **9**(1), 307 (2008).
- <sup>37</sup>L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
- <sup>38</sup>L. Breiman, *Manual on Setting up, Using, and Understanding Random Forests v3.1* (University of California, Berkeley, CA, 2002), Vol. 1, p. 58.
- <sup>39</sup>C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinf.* **8**(1), 25 (2007).
- <sup>40</sup>G. Louppe, *Understanding Random Forests* (Cornell University Library, Ithaca, NY, 2014).
- <sup>41</sup>G. Louppe, "Understanding random forests: From theory to practice," [arXiv:1407.7502](https://arxiv.org/abs/1407.7502) (2014).
- <sup>42</sup>M. Gevrey, I. Dimopoulos, and S. Lek, "Review and comparison of methods to study the contribution of variables in artificial neural network models," *Ecol. Modell.* **160**, 249–264 (2003).
- <sup>43</sup>J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International Conference on Machine Learning*, Atlanta, GA (June 16–21, 2013), pp. 115–123.
- <sup>44</sup>J. Bergstra, "Hyperopt: Distributed asynchronous hyperparameter optimization in python," <http://jaberg.github.com/hyperopt> (Last viewed May 15, 2021).
- <sup>45</sup>C. Persello and L. Bruzzone, "Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning," *IEEE Trans. Geosci. Remote Sens.* **54**(5), 2615–2626 (2016).