Surface shape reconstruction from phaseless scattered acoustic data using a random forest algorithm

Michael-David Johnson, Anton Krynkin, Giulio Dolcetti, et al.

Surface shape reconstruction from phaseless scattered acoustic data using a random forest algorithm

Michael-David Johnson,[1,a)] Anton Krynkin,[1] Giulio Dolcetti,[2] Mansour Alkmim,[3,b)] (iD) Jacques Cuenca,[3] and Laurent De Ryck[3]

[1]*Department of Mechanical Engineering, University of Sheffield, Sheffield, United Kingdom*

[2]*Department of Civil and Structural Engineering, University of Sheffield, Sheffield, United Kingdom*

[3]*Siemens Digital Industries Software, Interleuvenlaan 68, B-3001 Leuven, Belgium*

ABSTRACT:

Recent studies have demonstrated that acoustic waves can be used to reconstruct the roughness profile of a rigid scattering surface. In particular, the use of multiple microphones placed above a rough surface as well as an analytical model based on the linearised Kirchhoff integral equations provides a sufficient base for the inversion algorithm to estimate surface geometrical properties. Prone to fail in the presence of high noise and measurement uncertainties, the analytical approach may not always be suitable in analysing measured scattered acoustic pressure. With the aim to improve the robustness of the surface reconstruction algorithms, here it is proposed to use a data-driven approach through the application of a random forest regression algorithm to reconstruct specific parameters of one-dimensional sinusoidal surfaces from airborne acoustic phase-removed pressure data. The data for the training set are synthetically generated through the application of the Kirchhoff integral in predicting scattered sound, and they are further verified with data produced from laboratory measurements. The surface parameters from the measurement sample were found to be recovered accurately for various receiver combinations and with a wide range of noise levels ranging from 0.1% to 30% of the average scattered acoustical pressure amplitude.

© 2022 Author(s). *All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (*http://creativecommons.org/licenses/by/4.0/*).* https://doi.org/10.1121/10.0013506

(Received 9 September 2021; revised 20 July 2022; accepted 25 July 2022; published online 18 August 2022)

[Editor: Zoi-Heleni Michalopoulou]   Pages: 1045–1057

I. INTRODUCTION

Inverse acoustic scattering is concerned with the recovery of information about an object or a surface based on scattered acoustic data collected using sound sources and receivers. It has applications in fields such as non-intrusive damage testing as well as surface recovery.

A numerical method based on the boundary integral equations and Kirchhoff approximation to reconstruct the shape of a scattering surface was recently outlined.[1,2] This approach was found to be highly sensitive to uncertainties, partly because of the strong dependence on the phase of the scattered signal.[3] The errors in the inversion results were associated with the underdetermined and ill-posed nature of the problem.[3] The range of applicability in reconstructing a surface is also limited by the validity of a partial linearisation of the scattering problem, which is required to make the numerical inversion feasible.

Another recent approach includes recovering a rough surface at grazing angles using single-frequency, phaseless acoustic pressure through the use of an iterative marching method approach derived from the parabolic wave approximation (forward-scattered wave propagation assumption).[4] Although the inversion results are found to be relatively

accurate, it is assumed that the forward-scattered approach is not applicable in the context of this paper research due to significant differences in the assumptions and experimental setup and, in particular, acoustic remote sensing applications where the sound field is best described by a solution of the full Helmholtz equation [Eq. (3)].

Recent work[3] solved the reconstruction problem using matrix inversion, where the forward model of scattered acoustic pressure was linearised, resulting in the linear system of equations resolving the unknown profile of a rough surface. The use of an iterative approach such as machine learning is an appealing alternative to linearisation and as such is one of the central motivations of the present work.

As opposed to deterministic model-based inversion approaches, machine learning methods in wave scattering problems allow more flexibility. Machine learning and its applications in various fields of acoustics were discussed in detail by Bianco *et al.*[5] It has been shown that a combination of random forest and neural networks results in a robust method enabling reconstruction of geometrical features against noise.[6] This was achieved by first classifying training shapes with a random forest and then inverting the far-field scattered signal using neural networks to obtain geometrical features of different scattering objects.

Fan *et al.*[7] successfully applied deep learning methodologies utilising the Helmholtz equation in the recovery of the

a)Electronic mail: mjohnson9@sheffield.ac.uk

b)Also at: KU Leuven, Department of Mechanical Engineering, Celestijnenlaan 300 B, B-3001 Leuven, Belgium.

shape and placement of multiple scatterers in two different settings, including a seismic imaging setting where the source and receivers were above the scatterers and receivers were in a linear array. The scatterers were placed randomly and formed from a number of shapes such as triangles, squares, or ellipses. It was shown that for a large number of receivers and sources, the locations and orientations of the scatterers were successfully recovered with various amounts of noise in the dataset, while the recovered boundaries of the scatterers became blurred as the noise level increased.

Successful use of machine learning methodologies in acoustics was also demonstrated when identifying parameters such as the porosity and tortuosity of a porous material with an acoustic signal perturbed by noise.[8] It was shown that "acceptable accuracy with wide variety of noise levels"[8] can be achieved in recovering material properties.

Other works that recovered the parameters of surfaces instead of the full surface include recovering parameters of a sum of sine waves forming a rough surface,[9] recovering coefficients of a parametric curve of an obstacle,[10] and using a convolutional neural network to recover the root-mean-squared height and correlation length from a Gaussian rough surface through synthetic aperture radar.[11] The flexibility of data-driven approaches as compared to classical model inversion, in the presence of noise, stands as a central motivation of the present work.

The Kirchhoff approximation is still an active part of reconstruction efforts as in Ref. 12, and other methods have risen in inverse scattering especially in the near-field, such as recovering the far-field pattern given the near-field measurements[13] and obtaining the scattering coefficient from near-field measurements.[14]

The choice to use phaseless data as input was driven mainly by the characteristics of the random forest approach, which cannot handle coupled complex data. In Dolcetti et al.,[3] phase uncertainty was found to have a stronger impact than amplitude uncertainty on the accuracy of the surface reconstruction, and imperfect wrapping of the phase was found to cause a multi-modal distribution of the reconstruction error, especially at large roughness amplitudes (relative to the acoustic wavelength).

This work studies the feasibility of a machine learning approach to characterise a parametric rough rigid surface. Phaseless acoustical data were chosen due to the relatively simple amplitude only calibration technique compared to phase calibration, as well as to avoid the easily corruptible nature of phase measurements that are sensitive to uncertainties, such as uncertainties in position. Scattered phaseless acoustical pressure defined by a single-frequency source excitation will be synthetically generated through an application of the Kirchhoff approximation. Specifically, the estimation of the wavelength, amplitude, and offset of a sinusoidal acoustically rigid scattering surface is considered by means of a random forest algorithm trained on synthetic noisy data and tested on synthetic and experimental data.

The paper is organised in the following way: Sec. II presents the random forest model used for the estimation of the surface parameters. Section III contains relevant information regarding the methodology, including the selection of the forward model as the Kirchhoff approximation, the incorporation of noise, the way in which data were split into training and testing sets, the experimental setup, and the convergence of random forest as the number of trees increase. Section IV contains the relevant results and evaluation of the performance of the testing set and experimental data set, including a comparison between the method proposed in this paper and the short array method[1] and model performance in recovering unseen parameters. Section V contains the conclusions of the paper.

II. RANDOM FOREST ALGORITHM

The purpose of machine learning algorithms in this work is to allow the estimation of a parameter set that uniquely defines the shape of a sinusoidal surface using the modulus of the scattered acoustic pressure measured at a finite number of locations, defining $\psi_s(\mathbf{R})$ as the acoustic pressure field at a point $\mathbf{R}$ produced by a source with coordinates (x_1, y_1), scattered by a sinusoidal rigid surface with profile

$$\zeta(x) = \zeta_1 \cos\left[\frac{2\pi}{\zeta_2}(x + \zeta_3)\right],$$ (1)

where the parameters ζ_i, $i = 1, 2, 3$ indicate the amplitude, wavelength, and offset of the sinusoidal surface, respectively. The signal was recorded at a set of M microphones with coordinates of the jth microphone given by $\mathbf{R}^{(j)} = (x_2^{(j)}, y_2^{(j)})$, $j = 1, \ldots, M$; the aim is to estimate at least one of the parameters ζ_i, given $|\psi_s(\mathbf{R})|$. The general setup of sound scattering by a rough surface in Oxy plane is presented in Fig. 1, where the source and receivers are located in the acoustic far-field above the surface.

The choice to reconstruct the parameter set instead of the surface shape at each location (as was done in, e.g., Ref. 1) was made to limit the complexity of the problem and to develop a method applicable for surfaces of higher complexity, while still allowing a relatively compact parametric representation based on Fourier series, as are typical for examples of water waves in some applications.[15,16]
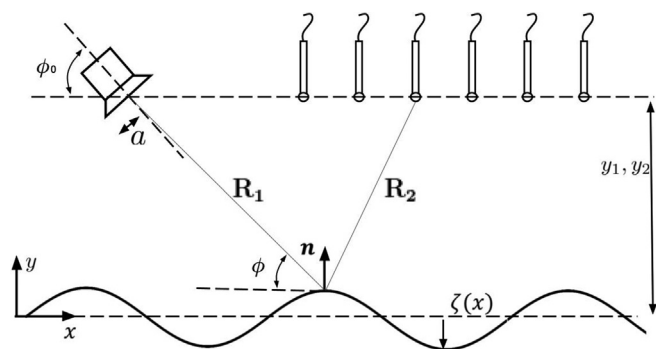

FIG. 1. The geometry of the problem where the rough surface is defined by a function $\zeta(x)$ from Eq. (1). Surface is not to scale.

Among the large number of existing machine learning approaches, here it was decided to employ a random forest approach for the recovery of the parameters of the surface. This is due to the simplicity of the model; the structure of a random forest is very strictly defined, giving fewer options to the user in its creation. This is different from a neural network approach, where the architecture needs to be carefully considered. It was also decided that not being able to extrapolate to parameters outside the range given by the training data would be a benefit for the problem investigated in this paper, to ensure the Kirchhoff condition is maintained. Random forests also benefit from not needing any input features to be scaled or standardised while also being able to measure feature importance through the use of Gini importance, although this can be biased when "input variables vary in their scale of measurement of their number of categories,"[17] although this is not investigated in this paper.

Random forests are a classification and regression model where the model predictions are formed from an average of multiple decision trees. A decision tree is a supervised machine learning model in which the resulting model is a tree-like structure, where queries on the data define the branches and predictions define the leaves. As the values of parameters in this study are real numbers, regression random forests are used. The composition of decision trees is discussed significantly by Breiman *et al.*[18] However, a brief derivation based on Hastie *et al.*[19] is included in this paper.

A. Decision tree

Due to random forests being formed from many decision trees, it is important to have an understanding of how decision trees partition data to make predictions. Suppose a dataset $D = (\mathbf{\Omega}, \mathbf{Z})$, where $\mathbf{\Omega} = (\omega_{lj}) \in \mathbb{R}^{N_D \times M}$ and $\mathbf{Z} = (\zeta_{lj}) \in \mathbb{R}^{N_D \times N}$ are the matrices of input and output values, respectively, with M the number of receivers and N the number of outputs to recover, and N_D denotes the number of samples in a dataset, with $l = 1, \dots, N_D$. Next, it is necessary to define a splitting criterion (j, S), where j is a column in $\mathbf{\Omega}$ and S is some value defined on an interval $[\min(\mathbf{\Omega}), \max(\mathbf{\Omega})]$. It is noted that throughout the paper, indices j, i, and l are reserved for the columns in the input dataset (referred to as features and associated with the receiver location in the receiver array), for the columns in the output dataset (referred to as surface parameters), and for the rows in input/output datasets (referred to as samples), respectively.

Then a binary partition of D forms two datasets, $D_1(j, S)$ and $D_2(j, S)$, via the following form:

$$D_1(j, S) = \left\{ (\mathbf{\Omega}, \mathbf{Z}) | \omega_{lj} \leq S, \quad l = 1, \dots, N_D, j = 1, \dots, M \right\}, \quad (2)$$

$$D_2(j, S) = D \backslash D_1(j, S). \quad (3)$$

where $\backslash$ is the set minus operator. To find the best splitting criteria in the case of $\mathbf{Z}$ having one output (i.e., $i = 1$),

one must minimise the following equation,[19] modified with weights:

$$\min_{(j, S)} \left[\frac{N_{D_1}}{N_D} \sum_{D_1} (\zeta_l - \bar{\zeta}_{D_1})^2 + \frac{N_{D_2}}{N_D} \sum_{D_2} (\zeta_l - \bar{\zeta}_{D_2})^2 \right], \quad (4)$$

where $\bar{\zeta}_D$ is an average of all outputs for a dataset. This process is then repeated recursively on D_1 and D_2 until the nodes are "pure," i.e., $\sum_{\omega_{lj} \in D_{\text{leaf}}} (\zeta_l - \bar{\zeta}_{D_{\text{leaf}}})^2 = 0$ for some partition D_{leaf}, or another stopping criterion is reached. The final partitions are called leaves and are where the predictions are measured from. The decision tree's prediction at a leaf node will be the mean of all the outputs in that leaf node. In reality, for regression implementations, a value of zero will never be obtained, so other stopping criteria to make a leaf must be considered, such as the sum of squared errors becoming lower than a threshold or defining a minimal number of elements (no fewer than 2) required to be in a sub-dataset.

Essentially, decision trees recursively split a dataset into grouped subsets via Eqs. (2) and (3) until the tree is fully formed. Common algorithms to generate the decision trees include the iterative dichotomiser (ID) algorithm[20] and the classification and regression trees algorithm (CART).[18] This paper uses a modified version of the CART algorithm contained in the PYTHON package scikit-learn.[21]

Due to the nature of the decision tree algorithm, decision trees tend to overfit the given training data. This can be improved with methods such as cost complexity pruning.[19] Another method to improve performance is through the application of random forests,[22,23] where the overfitting issue is tackled through incorporating randomness. In this approach, a "forest" is made from many decision trees, where each decision tree is individually trained on a random subset of the dataset, and the overall random forest prediction is an average of all the decision trees in the forest. Comparisons of the performance over decision trees have been studied previously, i.e., in Ref. 23, where in classification, random forests yielded an improved performance in 17 of 20 datasets tested for the same number of attributes over decision trees. Random forests have also been used in regression problems, such as in Ref. 24, where random forests were compared against support vector machines and a partial least squares (PLS) method to identify heavy metal content in soil from hyperspectral modelling. Zhou *et al.* found that both support vector machines and random forests were "significantly better than that of PLS,"[24] and random forests had an improved performance over support vector machines.

Classification often performs well in recovering from a discrete set of labels. For example, the amplitudes could be binned to [0.0–0.0005, 0.0005–0.001, 0.001–0.0015,…], and the same can be adopted for the wavelength and offset. However, to increase the resolution of the predictions, the number of bins has to increase. Therefore, by tending the number of labels to infinity, regression appears to be the reasonable approach for the present work.

J. Acoust. Soc. Am. **152** (2), August 2022

Johnson *et al.*   1047

The random forests in this work were created using the PYTHON library scikit-learn.[21]

III. ESTIMATION OF THE SURFACE PARAMETERS

As is very common with machine learning, large datasets must be produced to be applied in the training of the models to obtain useful results. For this paper, a dataset of the scattered wavefield is generated for the rigid surface given by Eq. (1) using the Kirchhoff approximation following Ref. 25. These data were then used for the training of the random forest by using the scattered phase-removed acoustic pressure $|\psi_s(\mathbf{R})|$ as inputs to the random forest ($\mathbf{\Omega}$) and surface parameters as the output ($\mathbf{Z}$), effectively simulating an inverse problem. Finally, the trained learner was applied to two sets of test data (one synthetic and one experimental) to evaluate its performance.

A. Generation of synthetic data

The Kirchhoff approximation was chosen due to its explicit form, obtained through the approximation of a scattered signal via an assumption based on reflections from a tangent plane. This makes it suitable for the calculation of large amounts of data, which is required for training and testing the random forests, while being fast to compute [approximately 0.01 s for 34 receiver Kirchhoff simulations on an AMD Ryzen 93 900X central processing unit (CPU) with 32 GB RAM]. The suitability of this approximation for the conditions analysed in this paper was presented in Krynkin et al.[1,2]

Let the rough surface be defined by a function $\zeta(x)$ as shown in Eq. (1), which satisfies the following condition:[26]

$$\sin(\phi) > \frac{1}{(kh)^{1/3}},$$ (5)

where h is the radius of curvature of the surface, k is the acoustic wavenumber, and ϕ is the incident angle of the acoustic wave. With the condition of Eq. (5) satisfied, it is possible to use the Kirchhoff approximation to model reflections of an acoustic wave from a rough surface using a tangent plane approach. In two dimensions, for a source with a given directivity pattern, the scattered acoustic pressure ψ_s can be expressed in the following equation:[2]

$$\psi_s(\mathbf{R}) = \frac{1}{2k\pi\mathrm{i}} \int_{-\infty}^{\infty} \frac{A(x,0)}{\sqrt{R_1 R_2}} e^{\mathrm{i}k(R_1+R_2)}(q_y - q_x\gamma)\mathrm{d}x,$$ (6)

where, as shown in Fig. 1, the values R_1 and R_2 are the Euclidean distance from the source at (x_1, y_1) and receiver at (x_2, y_2) to a given point $[x, \zeta(x)]$ on the surface, respectively,

$$R_1 = \sqrt{(x_1 - x)^2 + (y_1 - \zeta(x))^2},$$ (7)

$$R_2 = \sqrt{(x_2 - x)^2 + (y_2 - \zeta(x))^2}.$$ (8)

In Eq. (6), $\mathbf{R} = (x_2, y_2)$, $\gamma = \mathrm{d}\zeta(x)/\mathrm{d}x$; q_x and q_y are the x and y components of $\mathbf{q} = -k\nabla_S(R_1 + R_2)$ with the gradient

$\nabla_S = (\partial/\partial x, \partial/\partial y)$. The directivity term $A(\mathbf{r})$ is defined in this work as the far-field radiation from a baffled piston, which is given by[27]

$$A(\mathbf{r}) = \frac{2J_1(ka \sin(\phi(\mathbf{r}) - (-\phi_0 + \pi/2)))}{ka \sin(\phi(\mathbf{r}) - (-\phi_0 + \pi/2))},$$ (9)

where a is the aperture, J_1 is the Bessel function of the first kind with $n = 1$, ϕ_0 is the angle of inclination of the source main axis to the Ox axis, and $\phi(\mathbf{r})$ is the angle between the vector produced from the location of the source and the point $\mathbf{r}$ with the Oy axis.

Following the application of the Kirchhoff approximation to simulate the scattered field, the phase is removed from Eq. (6) through the application of modulus,

$$p(\mathbf{R}) = |\psi_s(\mathbf{R})|.$$ (10)

Taking into account the receiver locations in an array of M receivers, phase-removed acoustic pressure used in the random forest algorithm is given by the following matrix:

$$\mathbf{p} = \{p_l(\mathbf{R}^{(j)}) | j = 1, ..., M, l = 1, ..., N\},$$ (11)

where the rows of the matrix are formed from p_l (an ensemble containing the absolute array pressure for a given ζ_l), and $\mathbf{R}^{(j)}$ form the columns (receiver locations defined with respect to the origin of the Oxy plane). For brevity, the dependence on $\mathbf{R}^{(j)}$ will be omitted, resulting in $p_{lj} = p_l(\mathbf{R}^{(j)})$, and if operations are row-wise, only the j superscript will be omitted, resulting in $\mathbf{p}_l$.

B. Noise

The Kirchhoff approximation model is deterministic; therefore, one set of model parameters maps to a given sound pressure field. However, in practical applications, noise is present in measured data. It is proposed to modify the solution of the Kirchhoff approximation via additive noise, calculated as

$$\tilde{\mathbf{p}}_l = \mathbf{p}_l + \epsilon_l,$$ (12)

where $\epsilon_{lj} \in \epsilon_l$, $\epsilon_{lj} \sim \mathcal{N}(0, \sigma)$ is drawn from a normal distribution independently for each receiver. For additive noise, the standard deviation, σ, was selected to be percentages chosen for investigation (0.1%, 1%, 5%, 7%, 8%, 9%, 10%, 12%, 15%, 17%, 20%, 25%, 30%; this relates to an approximate signal-to-noise ratio of 29.7, 19.7, 12.7, 11.2, 10.6, 10.1, 9.7, 8.9, 7.9, 7.3, 6.7, 5.7, and 4.9 dB, respectively) multiplied by $\bar{p}_l$—the average of the receiver's pressure magnitude taken across all receivers for the given surface in the absence of the noise. The acoustic pressure for each receiver is then normalised by the maximum value for the Kirchhoff approximation scattered from the flat surface so that it can be used in the random forest algorithm.

C. Synthetic training and testing datasets

A large number of datasets that correspond to different realisations of the parameters ζ_i, $i = 1, ..., 3$ were prepared

1048   J. Acoust. Soc. Am. **152** (2), August 2022

Johnson et al.

using Eqs. (6) and (12). For these calculations, the source and receiver locations were chosen in accordance with the existing experimental data[3] that were later used for validation in this paper. The source location was at (x_1, y_1) $= (-0.20, 0.22)$ m. The angle ϕ_0 of the source main axis to the Ox axis was 60°. The receivers were located at a height of approximately $y_2 = 0.28$ m in the y axis, and 34 receivers were distributed evenly with x_2 taking values from –0.13 to 0.53 m in the x axis, leading to an average distance between the receivers of 0.02 m. The data were generated numerically through the use of the Kirchhoff approximation, and the integration was done numerically through the application of Simpson's rule[28] over the integration range of $x \in [-3, 3]$ m, which was vectorised to improve the speed of data generation.

Multiple datasets were formed for one-parameter, two-parameter, and three-parameter surfaces defined in Eq. (1) to investigate the performance of the algorithms as the number of unknowns increased. The values for each parameter were generated with uniform spacing from a lower bound to an upper bound; the choices for parameter values and resolution are shown in Table I.

As well as the surface generation, multiple datasets were generated with different proportions of noise to the absolute acoustic pressure, with σ varied between $0.001\bar{p}_l$ and $0.3\bar{p}_l$ as described above. Noise was added by cloning a pair of pressures and surface parameters 20 times and independently adding noise to every receiver, yielding datasets with the following sizes for one-, two-, and three-parameter datasets: 630, 18 900, 567 000.

The models were trained using the training set and then evaluated on the testing set to provide an indication of the model's performance on unseen data. The same approach described above was applied to generate both the training set and the synthetic testing set. The two subsets were split randomly, with a proportion of 70% training set and 30% testing set, such that the intersection of the training and testing set would yield the empty set.

D. Experimental testing dataset

Experimental data used in this paper for validation purposes were collected with 34 1/4 in. microphones [G.R.A.S. (Holte, Denmark) 40PH] and with a loudspeaker [Visaton (Haan, Germany) G 25 FFL], arranged with the same geometry described in Sec. III C. A sinusoidal surface (with

amplitude $\zeta_1 = 0.0015$ m and wavelength $\zeta_2 = 0.05$ m) illustrated in Fig. 2 was machined from an aluminium block with a length of 0.55 m in the x direction and a width of 0.35 m.

A signal was produced at 14 kHz and recorded simultaneously at all microphones, with a sampling rate of 102.4 kHz. The amplitude at each microphone was calculated by a Fourier transform applied to 0.02 s long segments of the signal and then averaged over 2000 segments using a Hann window. The data were calibrated *in situ* by comparing measurements of the sound field reflected by a flat surface with the corresponding prediction calculated numerically, following the procedure outlined in Dolcetti *et al.*[3]

Even after calibration, the pressure field scattered by the sinusoidal surface differed from the one predicted with the Kirchhoff approximation model, due to the uncertainties in the measurements, especially at the microphones further away from the source. This difference could be seen as an equivalent random noise with σ given by

$$\sigma = \frac{\sqrt{\sum_{j=1}^{M} \left[\psi_{ref}(\boldsymbol{R}^{(j)}) - p(\boldsymbol{R}^{(j)}) \right]^2 / M}}{\bar{p}(\boldsymbol{R})}, \tag{13}$$

where p is the Kirchhoff approximation given in Eq. (6), with $\zeta(x) = 0.0015 \sin (2\pi x/0.05)$, and $|\psi_{ref}|$ is the experimental data. σ was found to be 0.195, which quantifies the deviations observed in Fig. 3.

E. Performance evaluation

In this section, the choices of metrics used to estimate the accuracy of the inversion algorithm are briefly defined. For evaluation against the synthetic testing set, the coefficient of determination ($\mathcal{R}^2$) is used in the following form:[29]

$$\mathcal{R}^2(i) = 1 - \sum_{l=1}^{N} \frac{\left[\zeta_{li} - f(\mathbf{p}_l)_i \right]^2}{(\zeta_{li} - \bar{\zeta}_i)^2}, \tag{14}$$

where $\mathbf{p}_l = |\psi_s(\mathbf{R}^{(j)})|, j = 1..M$ are the inputs for a given surface with $\mathbf{Z}_l$ parameters, N is the number of samples, $\bar{\zeta}_i$ is the mean of the ith output, ζ_{li} are outputs at the lth row for the ith

TABLE I. Bounds for each parameter in the data generation stage as well as the number of samples generated within those bounds.

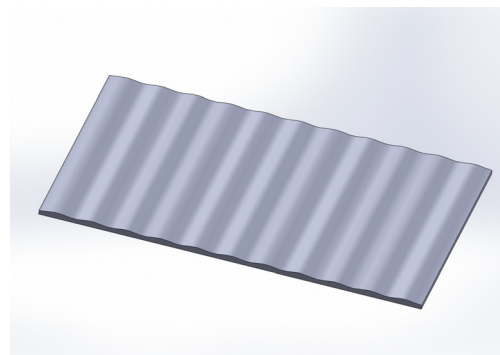| Number of recovered parameters	Amplitude bounds	Wavelength bounds	Offset bounds	Number of samples
1 parameter	−0.01 m, 0.01 m	N/A[a]	N/A	30
2 parameters	−0.01 m, 0.01 m	0.035 m, 0.15 m	N/A	30
3 parameters	−0.01 m, 0.01 m	0.035 m, 0.15 m	−0.02 m, 0.02 m	30

[a]Not available (N/A).

FIG. 2. (Color online) Three-dimensional (3D) rendering of the surface used in the acquisition of the experimental sample.

J. Acoust. Soc. Am. **152** (2), August 2022
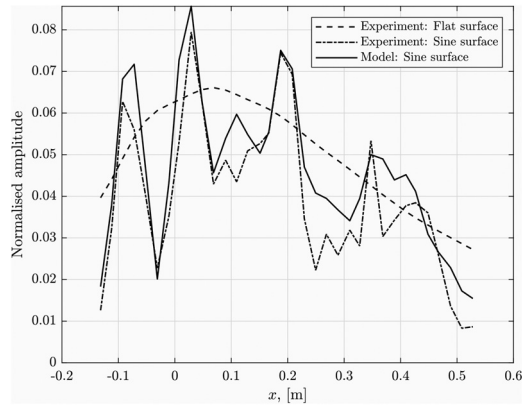
Johnson *et al.* 1049

FIG. 3. Overlay of the Kirchhoff approximation solution in comparison to the experimental example, which is calibrated for 14 kHz.

value from $\mathbf{Z}_l$ set of parameters, and f is the estimator, which is the random forest explained in Sec. II. Therefore, $f(\mathbf{p}_l)_i$ is the prediction of the ith value from the random forest given $\mathbf{p}_l$. For the two- and three-parameter estimation, the $\mathcal{R}^2$ score is calculated for each parameter and then averaged.

To evaluate the accuracy of the model predictions when predicting the surface given exclusively from the experimental data, the absolute error (AE) was also used, given by

$$E_i = |f(\mathbf{p})_i - \zeta_i|, \quad i = 1, \ldots, 3. \tag{15}$$

As there is only a single surface measurement available from the experimental data, $N = 1$, and there is no averaging in Eq. (15) compared to that in Eq. (14).

When plotting the results, the value of E_i is also normalised by the corresponding surface parameter, except for the offset, which is normalised by the wavelength.

It is noted that in the two- and three-parameter recovery, the AE of each output parameter was considered. This allows for an investigation into the change in parameter prediction as the number of parameters increases, while not allowing the overall AE to be dominated by the highest scale—the wavelength.

F. Convergence

A key hyperparameter of consideration is the number of decision trees used in the construction of the random forest. There are instances where an increase in the number of trees in a random forest only increases computational cost without much improvement in performance.[30] It was shown that both the errors in classification and regression forests are monotonically decreasing functions with respect to the number of trees.[31] These results also highlight that the most performance improvement was seen from random forests built from 10 trees to 250 trees.

Convergence testing has been performed for random forests generated with 1% and 15% added noise for three-parameter recovery. Three-parameter recovery was chosen due to the size of the dataset as well as the complexity of recovering three parameters. The change of the coefficient

of determination as the number of trees increases is presented in Fig. 4. In the case of 1% noise shown in Fig. 4, the increase in the coefficient of determination slows significantly after five trees in the random forest algorithm. For the case of 15% added noise, the coefficient of determination increases rapidly for random forests created from 1 tree to 50 trees as shown in Fig. 4, which signifies a better performance of the algorithm in predicting the testing set. The coefficient of determination demonstrated in Fig. 4 decreases in gradient rather significantly for the random forests as the number of trees passes 200, with a relative percentage increase in coefficient of determination of 0.18%. When the number of trees reaches 750, the increase in the number of trees has a smaller effect on the coefficient of determination while also significantly affecting computational time. In this study, with the hardware described in Sec. III A, the computational time for the random forests with 750 trees increased to 53.67 times that of the random forest with 1 tree. Increasing the number of trees to 1200 results in a further increase in computational time to 81.7 times. Therefore, the approach used in this paper was to generate random forests consisting of 700 trees, which is substantially inside the convergence factor to ensure convergence for all datasets while not having a significantly negative effect in terms of computation time.

IV. RESULTS

This section is formed from four subsections. First, the coefficient of determination is considered when evaluating model performance using the testing subset of the synthetic data. Following this, a comparison between the random forest approach against the short array method[1] is showcased. Then a study on the performance of random forests when evaluating parameters unseen by the model in training is showcased. Finally, the AE between model predictions and the surface used in the experimental sample is considered. For the AE when considering the two- and three-parameter surfaces, the AE is presented such that the AE per parameter is separated.
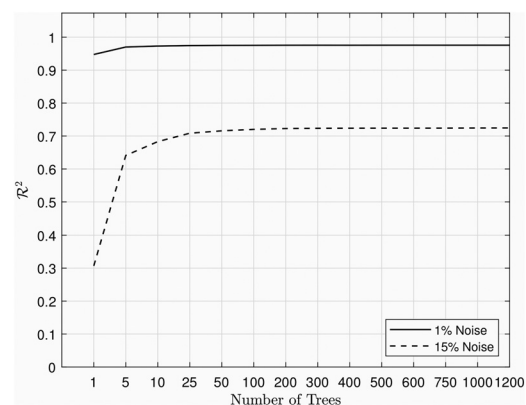

FIG. 4. Convergence of $\mathcal{R}^2$ when number of trees increase for 1% and 15% added noise.

A. Testing on synthetic data

Testing was done initially based on synthetic data, for a large number of surface realisations, and for various amounts of noise and numbers of receivers, where the first N receivers were considered. For one-parameter estimations, the surface wavelength and offset are assumed to be known, while the amplitude parameter ζ_1 is estimated. In this case, the coefficient of determination was found to be close to 1 for noise levels below 17%, except for 9% noise for random forests generated with ten receivers where $\mathcal{R}^2 \approx 0.888$, and then dropping to slightly below 0.9 at 30% noise, as shown in Fig. 5. It is noted that for ten receivers at 17% noise, $\mathcal{R}^2 \approx 0.861$. The decrease in coefficient of determination for the data generated by ten receivers at 9% and 17% noise is interesting due to the drop in value of $\mathcal{R}^2$. This behavior was deemed to be an outlier from the specific shuffle of training and testing sets. Randomly reshuffling the training and testing sets 1000 times for the 9% and 17% noise cases gave average coefficients of determination of 0.986 and 0.959, with standard deviations of 0.014 and 0.026, respectively.

When the surface amplitude ζ_1 and surface wavelength ζ_2 are estimated, the coefficient of determination lowers throughout all the noise percentages, as well as showing a smoother decay of $\mathcal{R}^2$ value with the noise level. For random forests generated with 15, 20, and 34 receivers, the minimal value of the coefficient of determination is above 0.7 at the dataset with 30% noise as shown in Fig. 6. For the random forest based on 10 receivers, the $\mathcal{R}^2$ is much worse than with a higher numbers of receivers, dropping to approximately 0.5 at 30% noise.

When all three surface parameters (amplitude ζ_1, wavelength ζ_2, and offset ζ_3) are estimated, the coefficient of determination decreases faster compared to the one- and two-parameter problems. This is highlighted at random forests generated with 10 receivers, where the lowest value $\mathcal{R}^2$ is approximately 0.3 at 30% noise compared to approximately 0.5 and 0.9 in one and two parameters, respectively.

A key feature to note is that the coefficient of determination changes only slightly for random forests generated with
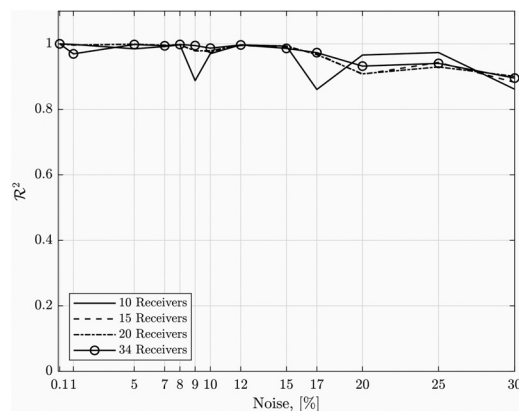

FIG. 6. Coefficient of determination for the estimation of the surface amplitude, ζ_1, and the surface wavelength, ζ_2, at varying noise levels and considering different numbers of receivers.

different receiver combinations when the number of receivers is greater than ten as shown in Figs. 5–7. This highlights that favourable model performance using machine learning to aid the inversion process can be obtained with fewer receivers. Having favourable performance with fewer receivers makes it possible to set up cost-efficient and practical applications.

B. Investigation of change of angle and frequency

To judge performance in different setup conditions and to verify the choice of angle and frequency used in the experiment, varying source angles ϕ_0 and source excitation frequencies were tested. The source angles were varied from 20° to 90°, with 70 equally spaced samples. The source excitation frequencies varied from 1000 to 45 000 Hz with 45 equally spaced samples. To reduce computation time, two-parameter datasets were created. The offset of the sinusoidal surface was fixed to 0. Datasets were created in the same way as described in Sec. III C, where noise was not added to the datasets. Figure 8 showcases the results of the coefficient of determination for the resulting testing sets. In the results, the method does not perform well below 5000 Hz for all angles tested. This is concluded to be due to

FIG. 5. Coefficient of determination for the estimation of the surface amplitude, ζ_1, at varying noise levels and considering different numbers of receivers.
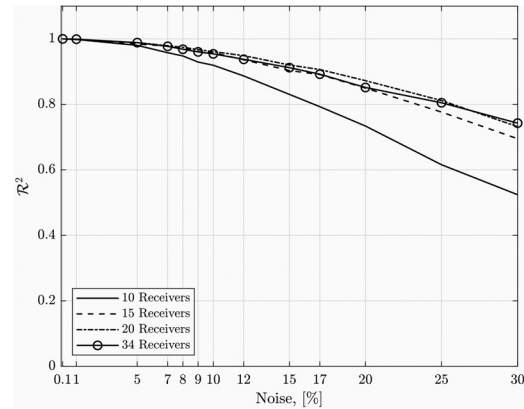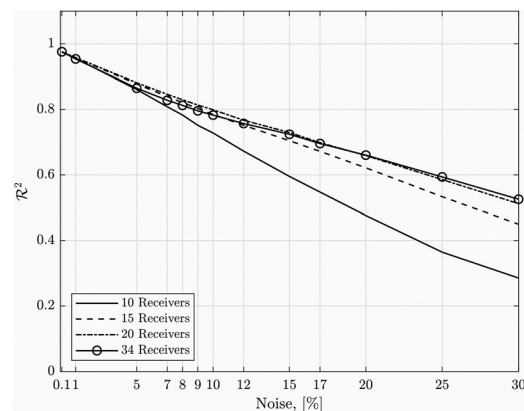

FIG. 7. Coefficient of determination for the estimation of the surface amplitude, ζ_1, the surface wavelength, ζ_2, and the offset, ζ_3, at varying noise levels and considering different numbers of receivers.

J. Acoust. Soc. Am. **152** (2), August 2022

Johnson *et al.* 1051

the scale of the acoustic wavelength being much larger than the amplitude and wavelength of the surface, causing the scattered absolute acoustic pressure to have no substantially different features. Between 5000 and 15 000 Hz, the coefficient of determination is at near maximum for all angles tested. As the frequency increases past 15 000 Hz, an ideal angle region is found between 30° and 60°.

C. Comparison with short array method

A comparison between the method proposed in this paper and the so-called short array method[1] in the unbiased version proposed by Dolcetti et al.,[3] hereafter called the SA0 method, is discussed in this section. The comparison was made between the three-parameter random forest generated with 5% noise and the SA0 model. Both models had the same initial conditions on setup and were evaluated with 141 750 samples taken from the testing set. To match the parameter-based recovery from the random forest to the surface-based recovery from the SA0 model, a set of surfaces was reconstructed from the parameters provided by the random forest by populating the values into Eq. (1) in the range $x \in [-0.11, 0.17]$, which is the specular range of the source and the receiver array. The two methods were applied to the same set of synthetic pressure data and compared in terms of the spatial root mean square difference between the target and reconstructed surfaces averaged over all surfaces with the same amplitude and normalised by the surface true amplitude parameter. The results can be seen in Fig. 9. The deviation in methods appears to increase rapidly when $\zeta_1/\lambda > 0.1$, where the SA0 method begins to increase in error significantly. This is mainly because of the loss of validity for the linearisation of the Kirchhoff integral, which is the basis of the SA0 method, whereas the random forest approach uses the Kirchhoff approximation directly without linearisation. A direct comparison to the previous methods can be made by analysing Fig. 13(c) in Dolcetti et al.[3] Calculating the
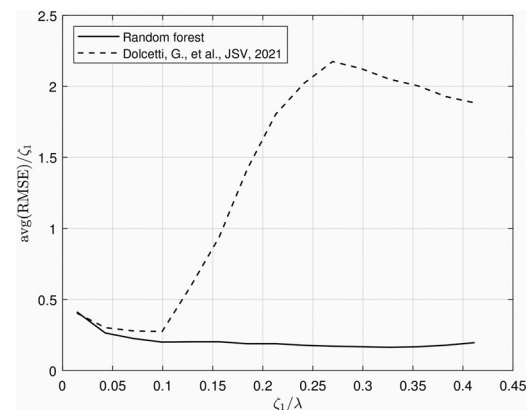
FIG. 9. Comparison between the approach offered in the paper against the SA0 approach: averaged root-mean-square difference between target and reconstructed surfaces normalised by the surface amplitude ζ_1, plotted against the surface amplitude normalised by the acoustic wavelength, ζ_1/λ.

root-mean-square error (RMSE) factored by the acoustic wavelength for the surface recovered from Table II within the specular points of the receivers and the microphones (following Dolcetti et al.[3]) yields a value of 0.0165, 0.027, and 0.169 for one, two, and three parameters, respectively. This outperforms the SA0 method in recovering the experimental sample and is comparable to reconstruction using multiple frequencies, while also being close to the reconstruction based on synthetic data without noise. The exception is the two-parameter recovery, which performs approximately the same as the SA0 method. Although this is one sample, the results, highlighted from both synthetic recovery as shown in Fig. 9 and experimental recovery earlier in this paragraph, showcase improved performance with the method proposed in this paper, especially as the amplitude increases.

D. Evaluating surfaces not seen by the model in training

To highlight the generality of using a random forest-based approach, predictions on surfaces that would have never been seen in the training and testing sets in Sec. III were evaluated. Two-parameter recovery was chosen to highlight the generalisation. Surface parameters were chosen by doubling the number of samples shown in Table I and removing samples that correspond to the training and testing set used to train the random forest. This leads to 841 surfaces to be predicted. To investigate generalisation further, noise was added to modify the receiver pressure at the
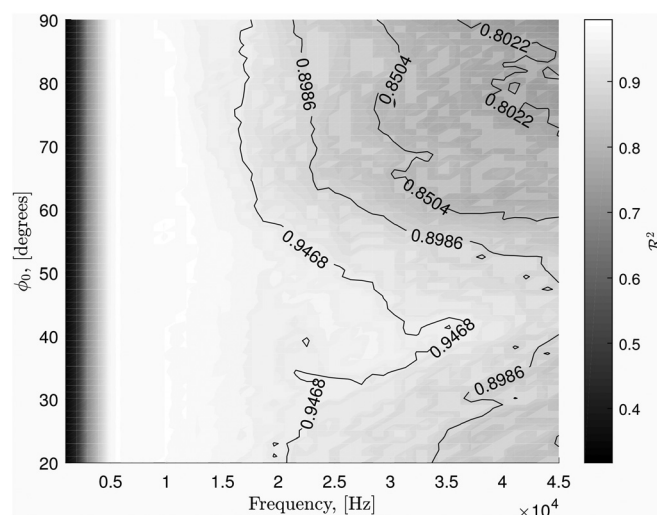
FIG. 8. Contour plot of the coefficient of determination of a testing set for the two-parameter recovery for varying angle of incidence ϕ_0 [see Eq. (9)] and frequency.

TABLE II. Results from parameter recovery for surface prediction using 15% noise random forest.

Number of recovered parameters	Amplitude	Wavelength	Offset	$\mathcal{R}^2$
Actual parameter	0.0015 m	0.05 m	0 m	N/A
1 parameter	0.002 07 m	N/A	N/A	0.986
2 parameters	0.001 58 m	0.0531 m	N/A	0.912
3 parameters	0.001 42 m	0.0516 m	−0.001 94 m	0.724

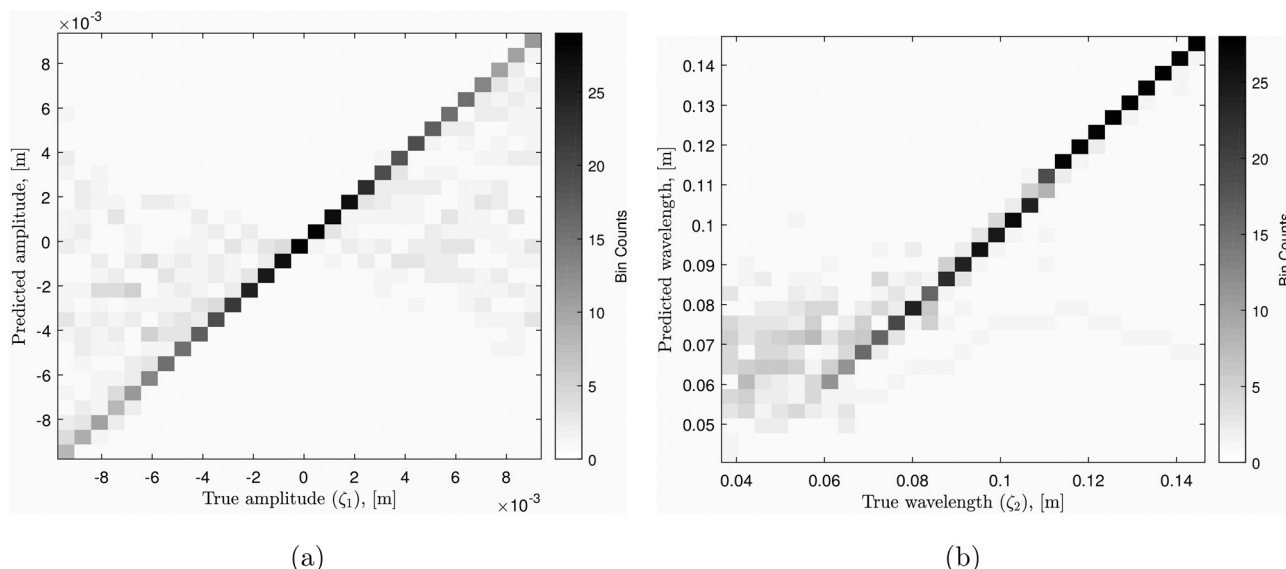1052   J. Acoust. Soc. Am. **152** (2), August 2022

Johnson et al.

FIG. 10. The predicted values versus true unseen parameters: (a) surface amplitude ζ_1 and (b) surface wavelength ζ_2.

unseen model parameters, where the percentage chosen to modify the pressure was different compared to that seen by the random forest during training. The random forest was trained on acoustic pressure linked to the two-parameter dataset with 5% added noise and evaluated on a dataset generated with 7% noise. In these data, the coefficient of determination given by Eq. (14) when training on the surface amplitude and surface wavelength is 0.6 and 0.84, respectively. Figures 10(a), 10(b), and 11 highlight the results of the recovered two-surface parameters.

There is a clear trend in Figs. 10(a) and 10(b) along the identity line, where true and predicted values are equal for both the amplitude and the wavelength. This can be seen from the density shown in the figures. The amplitude component has a spread of predictions that widen in proportion to the surface elevation height. The reconstruction of the wavelength begins to deteriorate for values starting below $\zeta_2 = 0.08$ m. It appears that the performance decreases when recovering surfaces with high amplitude and low wavelength, which corresponds to a higher Kirchhoff parameter.

Figure 11, representing the distribution of RMSE (described in Sec. IV C) defined along the specular region by the difference between the true surfaces and the surfaces reconstructed with the random forest predictions, demonstrates that the majority (approximately 74%) of the reconstructed surfaces predicted with the random forest algorithm fall within $\pm 0.41 \zeta_1$, which is the standard deviation of the RMSE deviation.

E. Testing on experimental data

The coefficient of determination was used as a metric to measure model performance with a synthetic testing subset of the synthetic data when discussing the results in Sec. IV A. In this section, predictions were obtained and compared against the experimental sample given in Sec. III C. Using the experimental phase-removed acoustic pressure as an input to the random forests, the prediction was then compared with the true surface parameters from the experiment using AE given in Eq. (15). The value of AE was then normalised by the actual surface parameters and expressed as a percentage point, converting it to a relative error.

Figure 12 highlights the relative error from one parameter's prediction of the surface amplitude ζ_1 based on experimental data. For the random forest generated with the full set of 34 receivers, the maximal AE divided by the actual surface amplitude rises from below 26% to 50% as the noise increases from 0.1% to 9%. As the noise level increases above 9%, the relative error decreases under 50% reaching its minimum 5% at 12% noise. This behavior is closely matched to the random forest generated by 20 receivers. For the random forest generated with 15 receivers, there is a large spike of relative error for noise values under 7% noise, and then the error for these random forests matches the error curve of the random forests generated with 20 and 34 receivers. For random forests generated with 10 receivers,
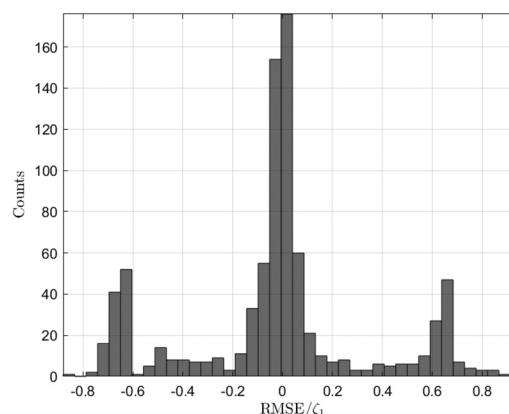

FIG. 11. Histogram of the RMSE, measuring the difference between the true surfaces produced with the unseen parameters and the reconstructed surfaces produced using the random forest predictions of the unseen parameters, normalised by the true parameter's amplitude.

J. Acoust. Soc. Am. **152** (2), August 2022
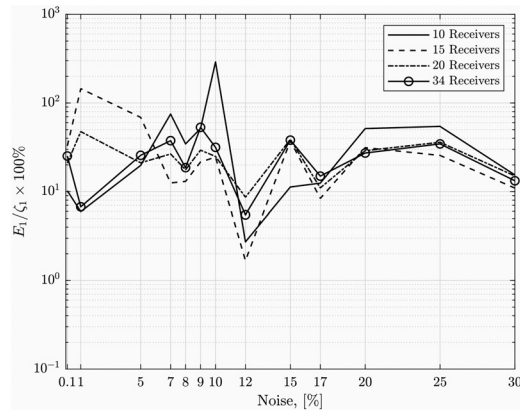
Johnson *et al.* 1053

FIG. 12. The change of relative error values for the random forests generated with different numbers of receivers and noise for one parameter surface recovery of the experimental sample.

the relative error tends to stay with the other error curves except for noise levels of 10%, where there is a spike in relative error of approximately 300%.

Figures 13(a) and 13(b) contain the relative errors of the amplitudes and wavelengths, respectively, for the two-parameter recovery. The random forests generated with 10 and 15 receivers have relatively high errors in comparison to 20 and 34 receivers. The wavelength and amplitude relative error for random forests generated with 20 and 34 receivers are close to the actual wavelength and amplitude values of the surface, staying mostly under 20% of the AE factored by the actual amplitude and under 25% of the AE factored by the actual wavelength. There are exceptions at 9% and 10% noise for random forest generated with 34 receivers and 12% for random forests generated with 20 receivers. For over 20% noise, the random forests generated with 34 receivers also exceed 25% relative error. There is a clear separation for the random forests generated with 10 and 15 receivers in comparison to 20 and 34 receivers in amplitude and wavelength, with the smaller number of receivers producing errors that are a factor of 10 larger for the wavelength component.

Figure 14 contains the relative error of all three parameters. For the full set of receivers, the relative error of the recovered surface amplitude decreases from 200% to settle at approximately 50% for random forests generated with 7% or more noise—except for 15% and 17% noise, where the relative error is found within the 3%–5% range. Interestingly, the random forest generated with 20 receivers has a lower error curve than the random forest generated with the full set of receivers. This could be due to an increased fit to the Kirchhoff approximation solution with the experimental results, as shown in Fig. 3. For the wavelength parameter, every choice of the number of receivers except for 34 receivers yielded a relative error greater than 10%. The offset AE, divided by the actual wavelength, stayed below 10% for every receiver combination and noise level. The random forest generated with 34 receivers kept the lowest AE value compared to the random forests generated with fewer receivers throughout.

The major problem with the three-parameter model is the prediction of the amplitude, where the relative error is high even for the full set of receivers. For the random forest generated with 34 receivers, the relative error was less than 10% at only 15% and 17% noise levels and then over 50% at all other noise levels. Although this error is in the submillimeter scale, the consistent underestimation can make the prediction invalid. It is noted that this underestimation could be due to the difference in scale between the parameters, with the wavelength of the surface being significantly larger than the amplitude. It is important to note that 15% and 17% noise is close to the estimated deviation between the measurements and the predictions by the Kirchhoff approximation model calculated in Sec. III D estimated with Eq. (13), which could explain the improved performance of the inversion at these noise levels.

Figure 15 and Table II highlight the model's prediction of the surface given the experimental acoustical pressure in comparison to the actual surface shape at datasets with 15% noise and with 34 receivers. 15% noise was selected due to the improved performance in the two- and three-parameter
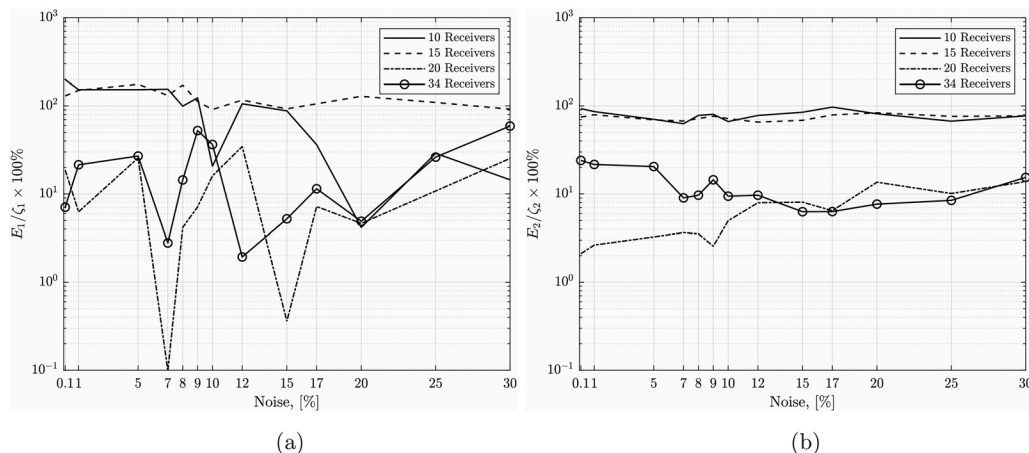

(a)

(b)

FIG. 13. The change of relative error values for the random forests generated with different numbers of receivers and noise for two-parameter surface recovery of the experimental sample, separated by parameter: (a) amplitude, (b) wavelength.
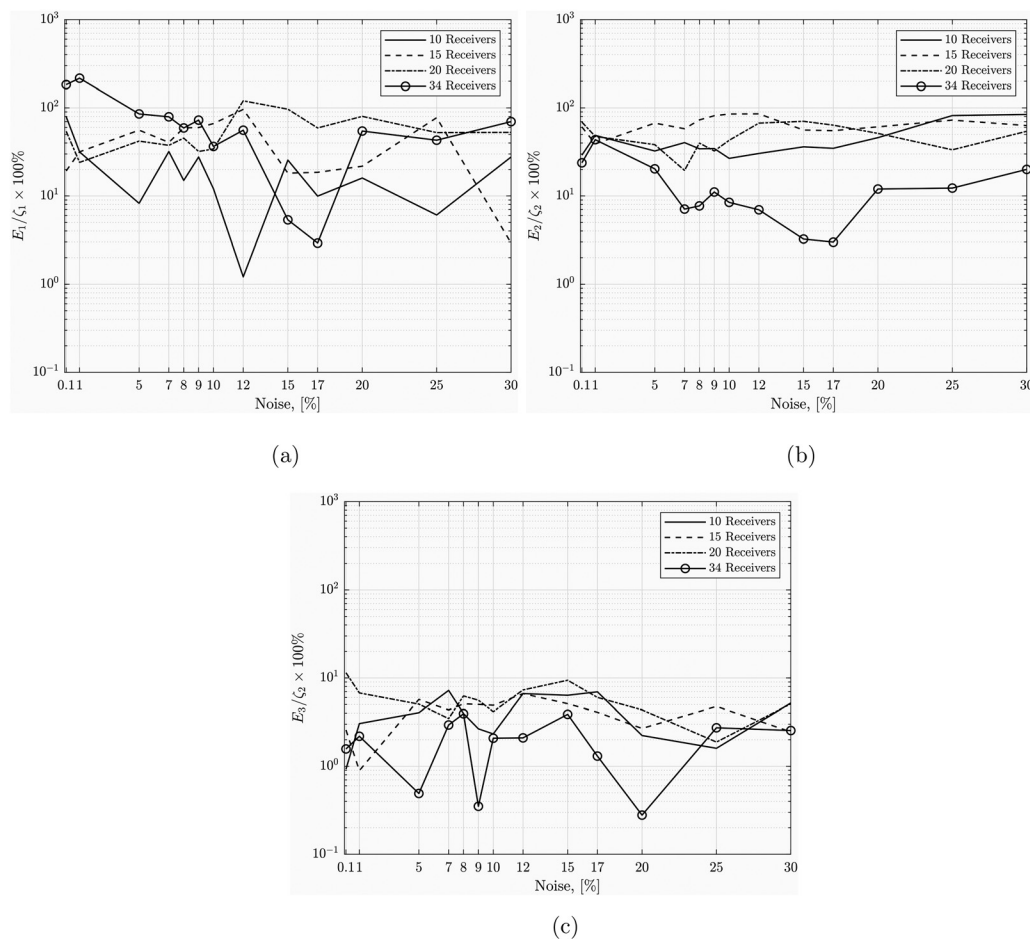
FIG. 14. The change of relative error values for the random forests generated with different numbers of receivers and noise for three-parameter surface recovery of the experimental sample, separated by parameter: (a) amplitude, (b) wavelength, and (c) offset.

models and similarity to the expected error from the Kirchhoff approximation model to the experimental data sample. The x-limits of the plots ranging from –0.15 to 0.15 m are defined by the width of the mainlobe in the source directivity pattern given by Eq. (9). The dominating component of difference to the experimental sample is the wavelength due to the scale differences in amplitude and wavelength as shown in Table II.

V. CONCLUSION

While training and testing the random forest regression algorithm, it was observed that for one-parameter datasets, the coefficient of determination is highly favourable, staying significantly above 0.8 over the range of added noise. For the random forests generated with the two-parameter datasets, all receiver subsets except for ten receivers slowly decay from 1 to just under 0.8. When random forests are trained with the three-parameter datasets, the coefficient of determination is observed to be above 0.7 for added noise levels between 0.1% and 15%, except for the random forest generated with ten receivers. With ten receivers, the coefficient of determination decreases significantly for noise levels above 10%.

The relative error has been used as a metric to evaluate model performance when predicting the surface parameters from experimental data. It has been noted that the relative error varies significantly between 0% and 100% depending on the added noise and number of receivers used in the recovery of parameters. In the three-parameter recovery, the lowest relative error values for 34 receivers have been consistently observed when noise levels are at 15% and 17%. It has been noted that these noise levels are comparable with the estimated discrepancy between the analytical solution used to generate synthetic data and the experimental results.

The approach of using random forests to predict the surface parameters of a rough surface provides favourable results even with high proportions of noise present, indicating that the proposed methodology performs considerably better than the SA0 method for single-frequency excitation analysed in Dolcetti et al.[3] for surfaces with amplitude/wavelength ≥ 0.1. The reasons are the loss of validity of the linearisation of the Kirchhoff integral, which is the basis of the method introduced in Krynkin et al.,[1] and the strong impact of phase uncertainties. It should be noted that once training of random forests has been completed, surface parameter recovery is predicted nearly instantly, allowing for recovery of measured surface in real-time. This paper

J. Acoust. Soc. Am. **152** (2), August 2022
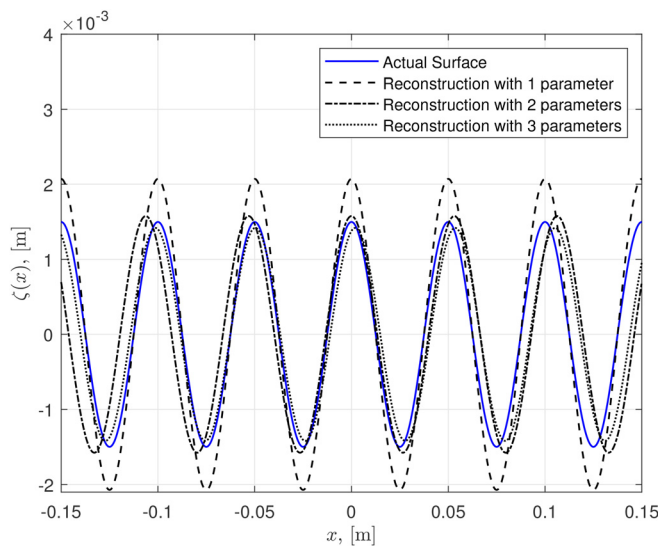
Johnson et al.   1055

FIG. 15. (Color online) Surface prediction for 15% noise random forest generated with 34 receivers.

has shown that this method works well for a simple sinusoidal surface.

The results in this paper have shown that the method works well for a simple sinusoidal surface and as such stands as an initial proof of concept that can be generalised to complex surfaces, for example, through Fourier series decomposition. Another extension to the results of this paper would be to incorporate a measure of confidence in the model's predictions, such as Bayesian inference, which could also benefit from flexibility and could provide complementary insights.

[1]A. Krynkin, K. V. Horoshenkov, and T. V. Renterghem, "An airborne acoustic method to reconstruct a dynamically rough flow surface," J. Acoust. Soc. Am. **140**, 2064–2074 (2016).

[2]A. Krynkin, G. Dolcetti, and S. Hunting, "Acoustic imaging in application to reconstruction of rough rigid surface with airborne ultrasound waves," Rev. Sci. Instrum. **88**, 024901 (2017).

[3]G. Dolcetti, M. Alkmim, J. Cuenca, L. De Ryck, and A. Krynkin, "Robust reconstruction of scattering surfaces using a linear microphone array," J. Sound Vib. **494**, 115902 (2021).

[4]Y. Chen, O. R. Spivack, and M. Spivack, "Rough surface reconstruction from phaseless single frequency data at grazing angles," Inverse Probl. **34**(12), 124002 (2018).

[5]M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C. A. Deledalle, "Machine learning in acoustics: Theory and applications," J. Acoust. Soc. Am. **146**(5), 3590–3628 (2019).

[6]W. Yin, W. Yang, and H. Liu, "A neural network scheme for recovering scattering obstacles with limited phaseless far-field data," J. Comput. Phys. **417**, 109594 (2020).

[7]Y. Fan and L. Ying, "Solving inverse wave scattering with deep learning," arXiv:1911.13202 (2019).

[8]T. Lähivaara, L. Kärkkäinen, J. M. Huttunen, and J. S. Hesthaven, "Deep convolutional neural networks for estimating porous material parameters with ultrasound tomography," J. Acoust. Soc. Am. **143**(2), 1148–1158 (2018).

[9]B. Engquist, C. Frederick, Q. Huynh, and H. Zhou, "Seafloor identification in sonar imagery via simulations of Helmholtz equations and discrete optimization," J. Comput. Phys. **338**, 477–492 (2017).

[10]Y. Gao, H. Liu, X. Wang, and K. Zhang, "On an artificial neural network for inverse scattering problems," J. Comput. Phys. **448**, 110771 (2022).

[11]T. Song, L. Kuang, L. Han, Y. Wang, and Q. H. Liu, "Inversion of rough surface parameters from SAR images using simulation-trained convolutional neural networks," IEEE Geosci. Remote Sens. Lett. **15**(7), 1130–1134 (2018).

[12]N. J. Joslyn and D. R. Dowling, "Recovery of coherent reflection from rough-surface scattered acoustic fields via the frequency-difference autoproduct," J. Acoust. Soc. Am. **151**(1), 620–633 (2022).

[13]M. Kleiner, H. Gustafsson, and J. Backman, "Measurement of directional scattering coefficients using near-field acoustic holography and spatial transformation of sound fields," J. Audio Eng. Soc. **45**(5), 331–346 (1997).

[14]A. Richard, D. Fernández Comesaña, J. Brunskog, C. H. Jeong, and E. Fernandez-Grande, "Characterization of sound scattering using near-field pressure and particle velocity measurements," J. Acoust. Soc. Am. **146**(4), 2404–2414 (2019).

[15]G. Dolcetti, A. Krynkin, and K. V. Horoshenkov, "Doppler spectra of airborne sound backscattered by the free surface of a shallow turbulent water flow," J. Acoust. Soc. Am. **142**(6), 3387–3401 (2017).

[16]G. Dolcetti and A. Krynkin, "Doppler spectra of airborne ultrasound forward scattered by the rough surface of open channel turbulent water flows," J. Acoust. Soc. Am. **142**(5), 3122–3134 (2017).

[17]C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics **8**(1), 25 (2007).

[18]L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees* (CRC, Boca Raton, FL, 1984).

[19]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer, New York, 2009), pp. 307-309.

[20]J. R. Quinlan, "Induction of decision trees," Mach. Learn. **1**(1), 81–106 (1986).

[21]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).

[22]L. Breiman, "Random forests," Mach. Learn. **45**(1), 5–32 (2001).

[23]J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," Int. J. Comput. Sci. Issues **9**(5), 272–278 (2012).

[24]W. Zhou, H. Yang, L. Xie, H. Li, L. Huang, Y. Zhao, and T. Yue, "Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model," Catena **202**, 105222 (2021).

[25]A. Krynkin, K. V. Horoshenkov, A. Nichols, and S. J. Tait, "A non-invasive acoustical method to measure the mean roughness height of the free surface of a turbulent shallow water flow," Rev. Sci. Instrum. **85**, 114902 (2014).

[26]F. G. Bass and I. F. Fuks, *Wave Scattering from Statistically Rough Surfaces* (Pergamon, Oxford, UK, 1979), pp. 220–227.

[27]P. M. Morse and K. U. Ingard, *Theoretical Acoustics* (Princeton University, Princeton, NJ, 1986), p. 381.

[28]E. A. Kendall, *An Introduction to Numerical Analysis* (Wiley, New York, 1989), pp. 37-45.

[29]D. Zhang, "A coefficient of determination for generalized linear models," Am. Stat. **71**(4), 310–316 (2017).

[30]T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *International Workshop on Machine Learning and* *Data Mining in Pattern Recognition* (Springer, New York, 2012), pp. 154–168.

[31]P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," J. Mach. Learn. Res. **18**(1), 6673–6690 (2017).