MAY 06 2024

Underwater sound speed profile estimation from vessel traffic recordings and multi-view neural networks

Joseph L. Walker ; Zheng Zeng; Vanessa M. ZoBell; Kaitlin E. Frasier

J. Acoust. Soc. Am. 155, 3015–3026 (2024)
https://doi.org/10.1121/10.0025920

JASA ARTICLE

Underwater sound speed profile estimation from vessel traffic recordings and multi-view neural networks

Joseph L. Walker,[1,a)] Zheng Zeng,[2] Vanessa M. ZoBell,[1] and Kaitlin E. Frasier[1]

[1]*Scripps Institution of Oceanography, University of California San Diego, San Diego, California 92093-0238, USA*

[2]*Department of Electrical and Computer Engineering, University of California San Diego, San Diego, California 92093-0238, USA*

ABSTRACT:

Sound speed is a critical parameter in ocean acoustic studies, as it determines the propagation and interpretation of recorded sounds. The potential for exploiting oceanic vessel noise as a sound source of opportunity to estimate ocean sound speed profile is investigated. A deep learning-based inversion scheme, relying upon the underwater radiated noise of moving vessels measured by a single hydrophone, is proposed. The dataset used for this study consists of Automatic Identification System data and acoustic recordings of maritime vessels transiting through the Santa Barbara Channel between January 2015 and December 2017. The acoustic recordings and vessel descriptors are used as predictors for regressing sound speed for each meter in the top 200 m of the water column, where sound speeds are most variable. Multiple (typically ranging between 4 and 10) transits were recorded each day; therefore, this dataset provides an opportunity to investigate whether multiple acoustic observations can be leveraged together to improve inversion estimates. The proposed single-transit and multi-transit models resulted in depth-averaged root-mean-square errors of 1.79 and 1.55 m/s, respectively, compared to the seasonal average predictions of 2.80 m/s.

https://doi.org/10.1121/10.0025920

(Received 21 September 2023; revised 9 April 2024; accepted 13 April 2024; published online 6 May 2024)

[Editor: Oleg A. Godin] Pages: 3015–3026

I. INTRODUCTION

Acoustic inversion is frequently employed in oceanography for the purpose of inferring ocean sound speed profile (SSP): a parameter that characterizes the dependence of the speed of sound on the temperature, salinity, and pressure of water (Chen *et al.*, 2018; Lovett, 1978). Reliably estimating sound speeds is of interest due to the profound effect of these profiles on acoustic propagation. Knowledge of local SSPs is important for improving the performance of underwater acoustic systems such as sonar and for various oceanographic studies involving ocean currents, internal waves, and underwater topography. Oceanic SSPs can be directly measured using autonomous underwater vehicles or surface vessel-based instruments, such as a conductivity-temperature-depth (CTD) sensor. The CTD sensor is lowered into the water column, recording the temperature, salinity, and pressure at regular intervals on both the descent and ascent paths. The recorded variables are then related to sound speed using a polynomial expression such as the Chen and Millero equation (Chen and Millero, 1977). These direct measurements are typically conducted during periodic field efforts. Hindcast models are used to interpolate spatially and temporally between these local measurements, ingesting observations to estimate oceanographic conditions across a region or period of interest. These models use observational data (opportunistic *in situ* measurements, satellite observations, and buoy data) and detailed physical oceanographic models, often developed for a specific region of interest (Stammer *et al.*, 2002; Zaba *et al.*, 2018).

In traditional acoustic inversion research, an active source is used alongside vertical hydrophone arrays for the inversion process. However, this setup can be costly and requires specialized equipment that is not widely available or easily deployed. Moreover, these advanced systems may not be suitable for long-term (month- to year-long) deployments desired for extended studies of temporal SSP variability, particularly in remote or deep-sea locations. The technical constraints of active source systems have limited the scale of data collection. To overcome these challenges, there is a demand for inversion strategies that leverage more easily accessible single-sensor passive acoustic recordings and opportunistic sound sources.

The introduction of the Automatic Identification System (AIS), which provides precise locations of vessels, has made it possible to use vessel traffic noise as a source of opportunity. This approach presents three main advantages: (1) vessels produce low frequency noise that can be detected at long distances, (2) maritime vessels are found in almost all ocean regions, making them a widely accessible source of data, and (3) the regular and frequent movement of vessels makes them a consistent and reliable sound source for long-term studies. Numerous studies have demonstrated the use of propeller noise from passing vessels received by seafloor hydrophones as acoustic sources of opportunity for estimating characteristics of the ocean environment and

[a)]Email: jlwalker@ucsd.edu

seafloor through which the signals have traveled (Gemba *et al.*, 2018; Gervaise *et al.*, 2012; Koch and Knobles, 2005; Tollefsen *et al.*, 2020). This strategy has been used to estimate the waveguide invariant property, which represents the dispersive characteristics of the waveguide under variable oceanographic conditions as well as for geoacoustic parameter inversions (Park *et al.*, 2005; Stotts *et al.*, 2010; Verlinden *et al.*, 2017). Few studies have investigated the use of machine learning for opportunistic acoustic inversion in the water column, though SSP variability has been used as a predictor for estimation of seabed parameters with machine learning (Escobar-Amado *et al.*, 2021).

Using uncontrolled, opportunistic vessel traffic noise as an acoustic source in oceanographic applications poses several challenges. The two main challenges are signal variability and background noise. The acoustic signal radiated by vessels is highly variable, dependent on factors such as the size of the vessel, speed, load, and environmental conditions, including ocean currents and wind resistance (McKenna *et al.*, 2013). Moreover, the signal is anisotropic due to hull interference and potential secondary sound sources from ship systems other than the propeller, introducing additional variability dependent on vessel orientation relative to an acoustic receiver (Gassmann *et al.*, 2017). Some of these factors are provided by AIS, but transit-dependent factors such as load and actual draught are not. Incomplete information can limit our ability to explain observed acoustic variability. The underwater environment is inherently noisy, and vessel traffic noise can be masked by other sources of noise such non-target vessels or natural sounds from marine life, wind, and waves. Recording systems can also differ in their self-noise characteristics. These challenges are exacerbated when a single hydrophone is employed to sample the acoustic signal, as is typically done in long-term observational passive acoustic monitoring.

When the underwater radiated noise (URN) of a moving ship is recorded in a shallow water environment, the signal contains characteristic interference patterns when viewed in the time-frequency domain (Brekhovskikh *et al.*, 1991; Chuprov, 1982). Prior works have linked these striation patterns with interference between propagative modes and exploited them to perform geoacoustic inversion (Gervaise *et al.*, 2012). These works relied upon conventional signal processing tools to extract the dispersion patterns. However, these algorithms require a high signal-to-noise ratio in order to be reliably extracted.

In recent years, machine learning approaches have been shown to outperform conventional signal and image processing techniques in a wide range of spectrogram processing applications (Ferguson *et al.*, 2018; Kirsebom *et al.*, 2020; Liu *et al.*, 2021; Tréboutte *et al.*, 2023). One of the advantages of using deep learning for acoustic inversion is that it can learn relationships between the input data and the output properties, even when those relationships are highly nonlinear and difficult to model using traditional methods. Multi-view learning, a machine learning approach that leverages multiple sources of information (i.e., views), can be integrated to learn more robust and accurate models. In the context of acoustic inversion, multi-view learning could theoretically be used to combine multiple recorded transits from the same day to improve the estimation of daily SSPs.

In this study, we investigate whether passive recordings of transiting vessels from a single hydrophone can be used together with deep learning to estimate local SSPs. We conducted a comparative modeling analysis to evaluate our prediction that spectral striation patterns observed between 50 and 200 Hz during vessel transits are informative of the SSP. We also inspect the learned filters of the convolutional neural network used in this study to determine whether our model identified these features as informative during training.

II. MATERIALS AND METHODS

A. Study site

The dataset used in this study consists of acoustic recordings of maritime vessels transiting through the Santa Barbara Channel (SBC) between January 2015 and December 2017 (Fig. 1). The traffic separation scheme within the SBC is approximately 20 nautical miles wide, extending from Point Conception in the north to the Long Beach Harbor. The SBC experiences a high volume of vessel traffic throughout the year, with container ships making up approximately 60% of all transits (Frasier *et al.*, 2022). Vehicle carriers, bulk carriers, and tankers each constitute about 10% of the transits, and cruise ships, tugs, research vessels, law enforcement and military vessels combined make up less than 10%.

B. AIS dataset

Vessels have been identified through AIS records, collected at onshore stations located at Coal Oil Point (34.411°N, 119.877°W) from April 2014 to the present and Santa Ynez Peak (34.029°N, 119.784°W) from August 2016 to the present. The received AIS messages were time-stamped and continuously logged with an on-site computer.

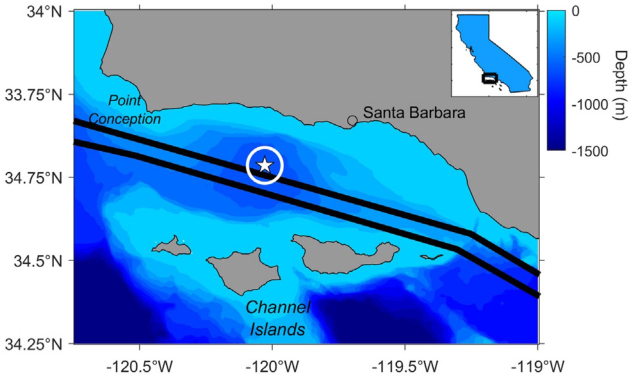

FIG. 1. (Color online) Map of the Santa Barbara Channel, using the World Geodetic Survey 1984 coordinate system. The traffic separation scheme is shown as black lines, and the HARP location is shown with a white pentagram. The white circle around the HARP denotes the 6 km boundary within which ship transits were tracked and acoustically delimited for this study.

These records were compared against time stamps of the same vessel transits matched based on the Maritime Mobile Service Identity (MMSI) extracted from Marine Cadastre[1] to ensure accurate position estimation in time. Time stamps are not broadcast over the AIS; therefore, the correct association of time and position depends on the time synchronization of the AIS receiving station. Where time stamps differed, Marine Cadastre positions, which are derived from U.S. Coast Guard records, were taken as the better estimate of position in time. All available AIS-derived information relevant to understanding vessel signature variability was used in this study. These variables are listed in Table I. The ship type (referred to as "type" in Table I) variable consists of seven different ship categories. The majority of the recorded transits were "cargo" ships, which comprised 78.5% of all transits. The remaining ship types and their relative frequency are as follows: "tanker" (7.2%), "tug" (5.1%), "research" (3.5%), "offshore supply" (3.5%), "passenger" (1.1%), and "recreational" (1.1%).

C. Vessel noise dataset

An existing database of 5865 recordings of identified ships transiting through the SBC between January 2015 and December of 2017 was used for this study. Acoustic recordings were collected in the SBC using a high-frequency acoustic recording package (HARP) (Wiggins and Hildebrand, 2007), which is a bottom-mounted recorder with a hydrophone tethered 10 m above the seafloor (580 m bottom depth). The location of the HARP (34.270°N, 120.030°W) relative to the study site is shown in Fig. 1.

The recordings were sampled at 200 kHz, which was then decimated by a factor of 20, resulting in a 10 kHz sampling rate, and a Nyquist frequency of 5 kHz. The data were low-pass filtered with an 8th order Chebyshev type I IIR filter to prevent aliasing during decimation. Each transit recording was segmented to consider only the time period in which the ship was within 6 km of the recording station. These audio clips were converted into spectrograms using a 10 000-point short-time Fourier transform with no overlap, resulting in a frequency resolution of 1 Hz and time resolution of 1.0 s. Spectrograms were cropped to limit the frequency range under consideration from to 10 to 300 Hz, the range over which local vessel noise is typically the dominant signal in this dataset and interference patterns are most apparent.

Multiple ship systems generate underwater noise; however, the highest amplitude source is typically generated by cavitation associated with the ship's propeller (Ross, 1976). The 2019 International Organization for Standardization estimates source depth for propeller-generated noise is as 70% of the draught (ISO 17208-1:2019, 2019). In our study, we include reported draught as a predictor, under the assumption that it is related to propeller depth and therefore to effective source depth. However, reported draught may not be routinely updated, may therefore differ from transit-specific operational draught, and may therefore have limited predictive utility. The effective source depth can be inferred to some degree from spectral data due to the Lloyd's mirror effect, in which surface reflections cause deconstructive interference at low frequencies (Gassmann *et al.*, 2017; Pereira *et al.*, 2020); or as part of a joint ship source-ocean parameter estimation problem using data recorded on hydrophone arrays (Tollefsen *et al.*, 2020).

D. Hindcast dataset

Estimating the near-surface region of the SSP is challenging because it experiences the highest level of variability. For this reason, daily SSPs were obtained for the top 200 m of the study region using the California State Estimation Short-Term State Estimation (CASE-STSE) model output (Zaba *et al.*, 2018). This model utilizes hindcast data and integrates the Massachusetts Institute of Technology general circulation model (MITgcm) through a least-square fitting solution. The data used in the integration include profiles from Spray gliders, high-resolution expendable bathythermographs, Argo, and satellite measurements of sea surface height and temperature. All CASE-STSE profiles used in this study are shown in Fig. 2. Although our presumed source depths are below the sea surface (approximately 3–5 m deep), near-surface portions of the SSP were included in the model because the minimum horizontal ranges between our sources and receiver are usually 8 or

TABLE I. Description of predictor variables used in statistical models.

Predictor variable	Abbreviation	Description
Ship design		
Length	LOA	Total length of ship in meters
Type	TYP	Numerical value that represents the general category of the vessel's type or purpose
Operational		
Draught	DRT	Depth of a vessel below the waterline
Heading	HDG	Direction that a vessel's bow is pointing
Course over ground	COG	Actual direction of progress of a vessel relative to the Earth's surface
Speed over ground	SOG	Speed of a vessel relative to the Earth's surface
Closest point of approach	CPA	Point at which the distance between the ship and receiver is smallest
Oceanographic		
Month	MTH	Month of the year

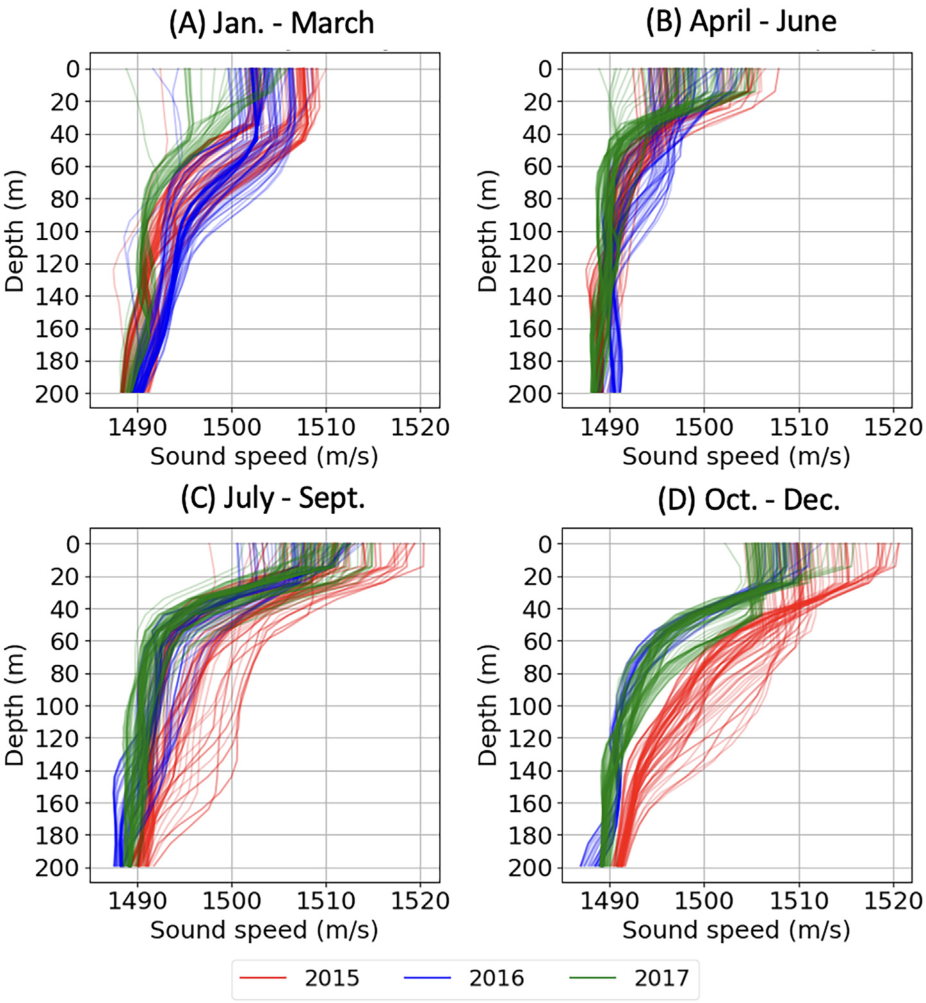

FIG. 2. (Color online) CASE-STSE generated SSPs for the study period in the grid cell nearest to the acoustic recording station. Panels (A) to (D) show all profiles generated for the months of January to March, April to June, July to September, and October to December, respectively. The associated year for each profile is color-coded.

more times the water depth. Therefore, we expect that indirect surface and bottom-interacting transmission paths are present in these recordings, providing some information on near-surface sound speeds.

E. Theoretical prediction

To investigate whether SSP differences might influence the appearance of vessel transit range-frequency spectrograms, an acoustic propagation model was used to simulate transits of an individual vessel with two different SSPs. Full water column SSPs were calculated using CTD cast data collected at a CalCOFI station located near the acoustic recording site (line 81.8, station 46.9) in summer (July 22, 2016) with a warm and highly stratified surface layer and in winter (January 16, 2017) with a cooler, deeper mixed layer. Bottom composition was derived from local sediment cores, obtained and measured by ZoBell *et al.* (2023). Propagation loss for a modeled source positioned 10 m below the sea surface was computed using the parabolic equation model RAMGeo (Collins, 1993) at 1 m vertical and 10 m horizontal resolution between 1 and 200 Hz in 1 Hz increments. Model parameters included a relative depth resolution of 0.05, relative range resolution of 2, and a reference phase velocity of 1500 m/s and used 6 terms for the Padé expansion. The source spectrum of a representative cargo ship was estimated by averaging estimated monopole source level spectra at the closest point of approach from three 2016 transits using a Lloyd's mirror correction and 1 Hz resolution [the source level estimation methodology using portions of the same dataset is detailed in ZoBell *et al.* (2023)]. Predicted range-frequency spectrograms were computed by subtracting propagation loss at each frequency from the vessel's estimated source spectrum at a series of range steps, computed at 10 m intervals along an actual transit path of the ship recorded on January 13, 2016. The difference between predicted spectrograms estimated using the summer versus winter profiles was visualized by computing the difference between the two spectrograms.

F. Models

1. Baseline model

Oceanic SSPs typically manifest seasonal patterns, primarily due to their significant dependence on temperature. Therefore, we first propose a model for SSP estimation that computes seasonal averages from previous years to estimate the SSP for all days within that specific season. This approach does not utilize any of the AIS or acoustic data.

This model is from here onwards referred to as the baseline model.

This baseline model is used to provide context for the neural network-based model performance. All neural network-based models used in this study use the same information as the baseline model (i.e., season) in addition to the transit data. Therefore, we can evaluate the informativeness of the transit recordings by comparing the performance of the neural network-based models with the baseline. If the transit data contain additional information regarding the local SSP, we would expect incorporation of the transit data to improve the SSP prediction estimate. Conversely, if the transit data are uninformative, we expect the estimation performance to remain unchanged.

2. Single-transit model

We designed a neural network-based model to produce an estimate for SSP from each of the recorded transits using the spectrograms and vessel descriptors. This model is from here onwards referred to as the single-transit model. The data used to train the single-transit model are denoted as $\mathcal{X}_s$ and can be formulated as follows. The data corpus $\mathcal{X}_s = \{(\mathbf{x}^{(1)}, \mathbf{v}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{v}^{(n)}, \mathbf{y}^{(n)})\}$, where $\mathbf{x}^{(i)}$ and $\mathbf{v}^{(i)}$ are the spectrogram representation of the audio recording and the vessel descriptors, respectively, for the ith recorded transit. We seek to learn a model, $f : (\mathbf{x}, \mathbf{v}) \rightarrow \mathbf{y}$, that maps input variables $\mathbf{x}$ and $\mathbf{v}$ to $\mathbf{y} \in \mathcal{R}^{200\times 1}$, where $\mathbf{y}$ is a vector containing the sound speed estimations for each meter in the upper 200 m of the water column [Fig. 3(A)]. We conceptualized f as comprising two functions that are applied in sequence: first an encoder function, E, and then a regression function, R.

The encoder function $E : (\mathbf{x}, \mathbf{v}) \rightarrow \mathbf{z}$ maps input variables $\mathbf{x}$ and $\mathbf{v}$ to a hidden variable, $\mathbf{z} \in \mathcal{R}^{128\times 1}$. Because the input variables are of different modalities (spectrogram image and AIS data), we divide E into two sub-encoders. Spectrograms are encoded using a convolutional neural network, denoted as E_S, while vessel descriptors are encoded using a fully connected network, denoted as E_V. E_S is composed of three convolutional blocks, with each block incorporating a sequence of convolutional, batch normalization, rectified linear unit (ReLU) activations, and a max-pooling layer. These convolutional blocks are followed by two fully connected blocks that each use a dense layer followed by dropout and ReLU activations. E_V comprises four fully connected blocks, each featuring a dense layer followed by batch normalization, dropout regularization, and ReLU activations. Both E_S and E_V return a vector that is then concatenated together [Fig. 3(B)]. The encoder function E refers to this integrated process of joint encoding and concatenation. Details regarding layer parameters as well as layer input and output dimensions for E_S and E_V are provided in Fig. 3(B).

The regression function $R : \mathbf{z} \rightarrow \mathbf{y}$ produces the SSP estimate from $\mathbf{z}$. R comprises three fully connected blocks, each featuring a dense layer followed by batch normalization, dropout regularization, and ReLU activations. Details regarding layer parameters as well as layer input and output dimensions for R are provided in Fig. 3(C).

Our proposed method hinges on leveraging recorded vessel noise in conjunction with AIS data to estimate SSP. To validate that our model genuinely learns pertinent features related to sound speed from the combined modalities and avoids relying on any spurious correlations that might exist between AIS data and ocean sound speed, we introduce a modified version of the single-transit model, referred to as single-transit (noAudio). In this variant, we set all spectrogram values to zero, effectively removing the vessel noise data while retaining only the AIS data.

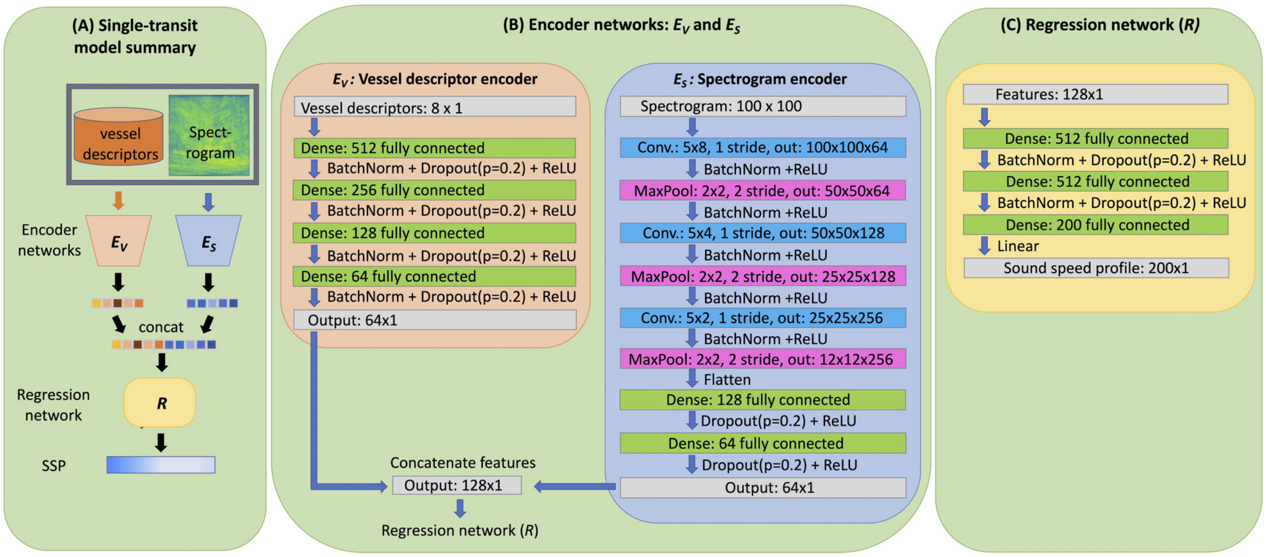
(A) Single-transit model summary
vessel descriptors
Spect-rogram
Encoder networks
E_V
E_S
concat
Regression network
R
SSP
(B) Encoder networks: E_V and E_S
E_V: Vessel descriptor encoder
Vessel descriptors: 8 x 1
Dense: 512 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 256 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 128 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 64 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Output: 64x1
E_S: Spectrogram encoder
Spectrogram: 100 x 100
Conv.: 5x8, 1 stride, out: 100x100x64
BatchNorm +ReLU
MaxPool: 2x2, 2 stride, out: 50x50x64
BatchNorm +ReLU
Conv.: 5x4, 1 stride, out: 50x50x128
BatchNorm +ReLU
MaxPool: 2x2, 2 stride, out: 25x25x128
BatchNorm +ReLU
Conv.: 5x2, 1 stride, out: 25x25x256
BatchNorm +ReLU
MaxPool: 2x2, 2 stride, out: 12x12x256
Flatten
Dense: 128 fully connected
Dropout(p=0.2) + ReLU
Dense: 64 fully connected
Dropout(p=0.2) + ReLU
Output: 64x1
Concatenate features
Output: 128x1
Regression network (R)
(C) Regression network (R)
Features: 128x1
Dense: 512 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 512 fully connected
BatchNorm + Dropout(p=0.2) + ReLU
Dense: 200 fully connected
Linear
Sound speed profile: 200x1

FIG. 3. (Color online) Single-transit model architecture. (A) Summary of the single-transit model. (B) The spectrogram and vessel descriptor data are encoded separately using a convolutional neural network, E_S, and fully connected network, E_V. Both encoder networks output a vector, both of which are concatenated. (C) The concatenated vector is then forward propagated through a fully connected regression network, R, which produces an estimate for sound speed for each meter in the upper 200 m of the water column.

3. Multi-transit models

The last set of models we consider are designed to combine and/or contrast information from multiple transits to improve SSP estimation. The methodologies we examine are inspired by, or are direct implementations of, existing multi-view machine learning techniques. For our application, all transits recorded on the same day are considered as distinct "views" that contain information about the same SSP.

To train the multi-transit models, we organize the acoustic dataset into daily collections denoted as $\mathbf{C}_V$, where each collection consists of multiple transits. Notably, all transits within a collection share the same SSPs, as they were recorded on the same day. This corpus is denoted as $\mathcal{X}_m = \{(\mathbf{C}_V^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{C}_V^{(N)}, \mathbf{y}^{(N)})\}$, where each collection $\mathbf{C}_V^i$ consists of all spectrogram and vessel descriptor pairings that were recorded on the ith day and $N = 899$ is the number of sampling days. Hence, we reformulate the modeling task as $f : \mathbf{C}_V \rightarrow \mathbf{y} | \mathbf{C}_V = \{(\mathbf{x}, \mathbf{v})^{(1)}, (\mathbf{x}, \mathbf{v})^{(2)}, \ldots, (\mathbf{x}, \mathbf{v})^{(T_V)}\}$, where $(\mathbf{x}, \mathbf{y})^{(v)}$ represents one audio recording and vessel descriptor pair, $v \in \{1, \ldots, T_V\}$, and T_V denotes the number of transits in collection $\mathbf{C}_V$, which is variable across the collections. All variables in $\mathcal{X}_m$ are the same as the single-transit data collection $\mathcal{X}_s$.

The simplest way to leverage multiple transits is to average the estimations for each of the transits in a collection using the single-transit model. We refer to this approach as single-transit (avg). However, this approach is not able to leverage complementary information or weigh saliency differences from multiple transits to improve prediction accuracy. To address these concerns, we evaluate three "late fusion" techniques for combining information across the transits within each collection. Late fusion is a technique in multi-view learning that allows the combination of learned features from multiple views at a later stage in the learning process (Feng *et al.*, 2018; Lin and Kumar, 2018; Seeland and Mäder, 2021; Su *et al.*, 2015).

Two existing late fusion approaches we consider are (1) late fusion (max), where the maximum value is calculated for each of the features across the transits, and (2) late fusion (concat), where a fixed number of feature vectors are concatenated (Seeland and Mäder, 2021). Last, we implement a novel late fusion technique referred to as late fusion (token), which is described below.

The forward propagation of a transit collection, $\mathbf{C}_V$, into encoder E produces a matrix, $\mathbf{Z}_V$, whose columns are the feature vectors of length $D = 128$ from each transit in the collection:

$$\mathbf{Z}_V = \left[\mathbf{z}^{(1)}; \ldots; \mathbf{z}^{(T_V)}\right] = E(\mathbf{C}_V) \in \mathcal{R}^{D \times T_V}. \quad (1)$$

For late fusion (max), an element-wise maximum operation is applied for each of the D features, which produces the vector

$$\hat{\mathbf{z}}_V = \max_v \mathbf{Z}_V \in \mathcal{R}^{D \times 1}. \quad (2)$$

For late fusion (concat), we concatenate k columns in $\mathbf{Z}_V$ to form a vector,

$$\hat{\mathbf{z}}_V = \left[\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(k)}\right] \in \mathcal{R}^{Dk \times 1}. \quad (3)$$

If $k < T_V$, we subsample the transits by randomly selecting k columns without replacement. If $k > T_V$ we upsample the transits by randomly selecting $k - T_V$ columns with replacement to duplicate and concatenate all features' vectors. For each training fold, the value for k is empirically determined to optimize performance on a validation set, as outlined in Sec. II G.

Late fusion (token) combines ideas from scaled dot-product attention and prompt tuning with the goal of automatically weighting more informative transits (Jia *et al.*, 2022; Vaswani *et al.*, 2017). A weight matrix, $\mathbf{W} \in \mathcal{R}^{h \times D}$, is used to project the features in $\mathbf{Z}_V$ into a lower dimension of size $h = 64$. The projected features are then compared against a learnable token, $\mathbf{q} \in \mathcal{R}^{h \times 1}$. The similarity values are then normalized using the softmax function. The normalized values are then used to compute a weighted sum of the original features:

$$\hat{\mathbf{z}}_V = \text{softmax}\left(\frac{\mathbf{q}^T \mathbf{W}^T \mathbf{Z}_V}{\sqrt{D}}\right) \cdot \mathbf{Z}_V^T \in \mathcal{R}^{D \times 1}. \quad (4)$$

A model trained with this multi-view approach will reduce its loss by learning to assign larger weights (i.e., large similarity with $\mathbf{q}$) to transits that produce more reliable sound speed estimates. An illustration of the late fusion (token) method is shown in Fig. 4.

For all the aforementioned late-fusion methods, the fused feature vector $\hat{\mathbf{z}}_V$ is forward propagated through the regression network R to regress SSP.

G. Experimental setup

Our partitioning of the training and testing data was deliberately crafted to emulate a real-world scenario, where

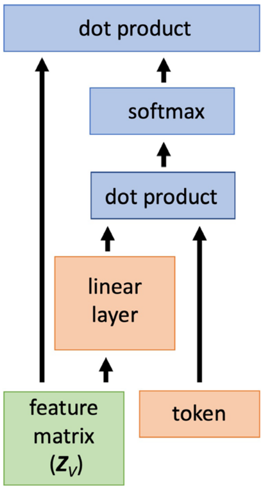

FIG. 4. (Color online) Our fusion method. Features $\mathbf{Z}_V$ are passed through a linear layer, followed by a dot product operation with a learnable token. Orange boxes indicate learnable values. The green box indicates features from encoder E. Blue boxes indicate a fixed mathematical operation.

the model is trained on historical data and subsequently deployed on future data. Data from the year 2017 were divided into four distinct testing sets, each corresponding to a specific season. These testing sets were created to ensure non-overlapping periods and were defined as follows: winter (January to March), spring (April to June), summer (July to September), and fall (October to December). This partitioning allowed for the assessment of model performance in relation to the different seasons of the year. We then perform fourfold cross-validation where for each fold, one season from 2017 is used for testing and the remaining data are used for training. Neural network models use vessel descriptor data, spectrograms, and hindcast profiles for training and testing, whereas the baseline models utilize only the hindcast profiles. The fourfold data partitioning methodology is illustrated in Fig. 5.

For all neural network-based models, 25% of each training set was allocated for validation-based early stopping with a patience of 30 epochs. Optimization was performed with the ADAM optimizer using a learning rate of 1×10^{-4} and a scheduler that decays this learning rate by a factor of 0.75 every 10 epochs. Regression loss is computed as root-mean-square error (RMSE). Unless stated otherwise, a batch size of 24 is used. The multi-transit models were trained using a two-stage approach, where each stage uses the same optimization process as described above: (1) for first-stage training, the encoder network and regression network are trained to estimate SSP from each transit, and (2) for second-stage training, the layers of the encoder network are fixed, and the parameters of the multi-view learning mechanism (if applicable) and the regression network are trained. To assess neural network model performance, we compare the RMSE of our models to that of the baseline model. We visualize the model's performance by plotting model residuals as a function of depth.

III. RESULTS

Our comparative modeling analysis revealed visual and measurable differences in predicted spectrograms that were produced by the same simulated vessel transit under different SSPs (Fig. 6). Differences included the angles of the predicted large and small scale interference patterns between approximately 5 and 200 Hz. This model is highly simplified and does not reflect the anisotropic radiation pattern of a real vessel; however, it supports the basic prediction that information regarding the SSP may be embedded in the recorded spectrograms when observing the evolution of the recorded signal as a broadband source moves relative to our acoustic receiver.

The proposed single-transit model provided an average error reduction of about 36% compared to the baseline model across the testing folds (Table II). We attribute the observed performance improvement to the inclusion of the acoustic data, as its exclusion (the noAudio model) led to a performance level comparable to historical averaging. The performance improvement of the single-transit model was variable across the folds. Specifically, during the summer and fall testing seasons, the single-transit model achieved substantial reductions in estimation error, with improvements of 44% and 43%, respectively. In contrast, its performance improvement was relatively modest during the spring testing season, with only a 13% reduction in error observed.

Model error was reduced by an additional 8% by averaging estimates obtained from multiple transits (Table II). The best performing multi-transit model was late fusion (token), which provided an average error reduction of 13.5% compared to the single-transit model and a 5.5% error reduction compared to the single-transit (avg) model. Late fusion (token) performed similarly to, or slightly better than, late fusion (concat). However, late fusion (concat) requires finding the optimal number of transits to select (k),

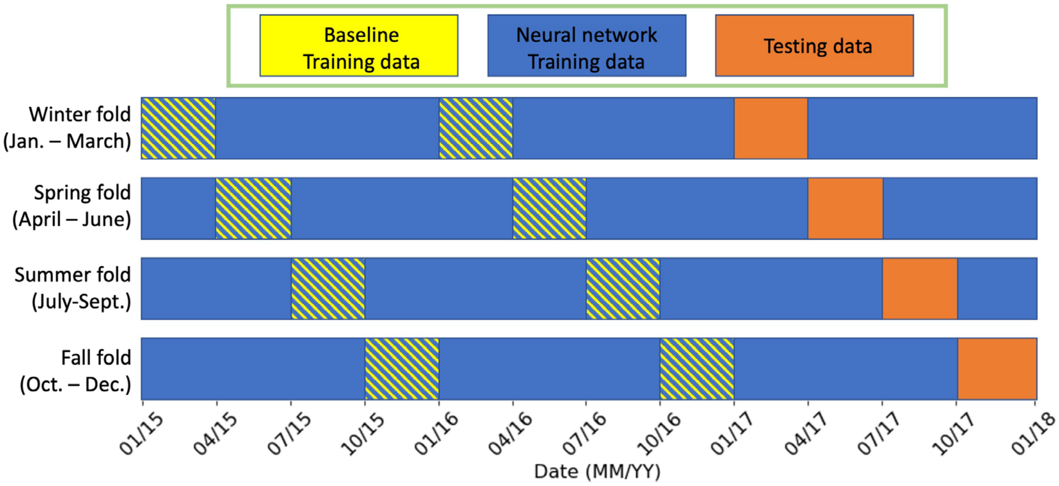

FIG. 5. (Color online) Illustration of the fourfold cross-validation approach for the baseline and neural network models. For each fold, models are tested on data from a single season in 2017, shown in orange. Neural network models use vessel descriptor data, spectrograms, and hindcast profiles from the time windows considered in each fold, whereas the baseline models utilize only the hindcast profiles. Note that time regions with mixed colors indicate that data were used to train both models.

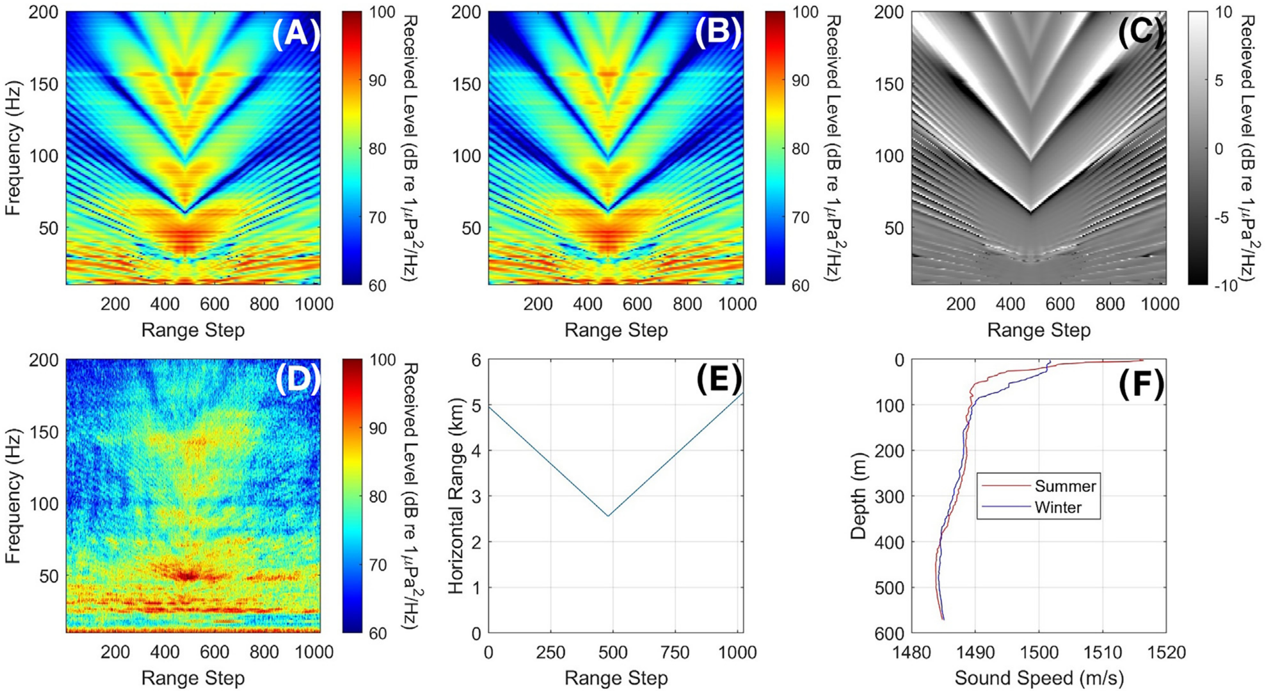

FIG. 6. (Color online) (A) Predicted range-frequency spectrogram using a representative winter SSP [blue line in panel (F)]. (B) Predicted range-frequency spectrogram using a representative summer SSP [red line in panel (F)]. (C) Difference between panels (A) and (B). (D) Real range-frequency spectrogram of the ship used in this example, recorded in January 2016. (E) Horizontal range at each range step for the ship transit used in this example. (F) SSPs.

which necessitated running seven times as many experiments (Fig. 7).

To understand what features the model uses for prediction, we produced a visual representation of the 16 filters learned by the single-transit model after being trained on the summer fold (Fig. 8). Noteworthy are seven filters, outlined in red, which display patterns highly reminiscent of the striation patterns associated with the waveguide invariant. This suggests that the network is leveraging information regarding the waveguide invariant to inform its predictions.

RMSE of the model predictions was evaluated as a function of four predictor variables and ship types (Fig. 9). The consistent vertical dispersion of points in each subplot highlights the absence of discernible performance trends, indicating stability and reliability across predictor variables.

IV. DISCUSSION

Neural network models were able to make improved SSP estimations when provided with spectrograms of ship noise and AIS data relative to baseline models using seasonal averages and AIS only data. The relatively lower estimation bias of the neural network models compared to the baseline suggests that the neural network is able to learn relevant patterns and relationships that can generalize across the seasons. The estimation error was found to be highest in the near-surface regions of the SSP. In order to mitigate these errors, we propose that future work should explore the integration of additional observational modalities, such as satellite-derived sea surface temperature estimates.

Oceanic SSPs have inherent seasonal regularity; however, year-to-year variability is strongly evident in this dataset. The reconstructed hindcast profiles at the study site

TABLE II. Model performance of SSP estimation across the four test seasons.

	RMSE (m/s) for[a]						
				Multi-transit			
Test fold	Baseline	Single-transit (noAudio)	Single-transit	Single-transit (avg)	Late fusion (max)	Late fusion (concat)	Late fusion (token)
January to March	2.4	2.43	1.69	1.59	1.59	1.58	**1.52**
April to June	2.11	2.07	1.71	1.48	1.52	1.52	**1.47**
July to September	2.84	2.1	1.58	**1.45**	1.55	1.55	1.48
October to December	3.83	3.91	2.18	2.04	1.93	**1.69**	1.72
Average	2.8	2.63	1.79	1.64	1.65	1.59	**1.55**

[a] Performance is reported in terms of RMSE in meters per second. The best performing model for each season is shown in boldface.

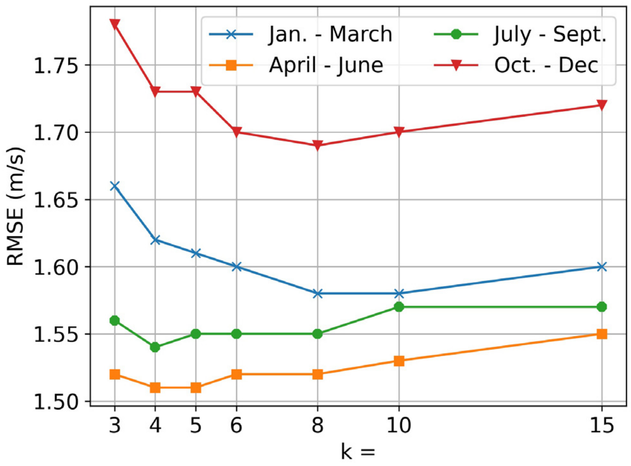

FIG. 7. (Color online) Late fusion (concat) prediction RMSE on seasonally aggregated testing data with varying transit sampling number k. The graph generally reveals distinct U-shape curves, demonstrating that low and high values of k result in suboptimal performance, while an intermediate range of k values yields the highest model performance.

reflect this variability, and it is particularly noticeable comparing profiles from the year 2015 to other years (Fig. 2). In 2015, El Niño conditions lead to the development of warm water "Blob" in the northeast Pacific (Tseng *et al.*, 2017). It is expected that the presence of strong interannual variability in the data would result in estimation bias in both the baseline and single-transit models. However, although both models exhibited estimation bias, the residual plots in Fig. 10 indicate that the deep learning-based approach experiences comparatively less estimation bias than the baseline model. For example, in Fig. 10(D), the prediction error distributions from the two models generally follow the same trend, but the estimations from the single-transit model are more closely centered on the zero line than the baseline.

As indicated by the performance of single-transit (avg), averaging multiple estimates helps to mitigate the effects of random errors or outliers by simply leveraging a larger sample. However, calculating an average considers all estimates to have equal weight in the final average and provides no mechanism for leveraging complimentary information or discard outliers. Multi-view learning techniques provide the opportunity for extracting such information by pooling features extracted across the different observations.

Our results using the traditional multi-view techniques (max and concat) were similar to those of Seeland and Mäder (2021), where the multi-view methods with learnable solutions provided the best results. However, using feature concatenation is complicated in this application because the number of transits is variable. This means that a fixed number, k, of transits need to be sampled, which introduces the trade-off: if k is too small, there is less information to leverage, but if k is too large, the number of trainable parameters grows linearly, potentially leading to over-fitting. This produces a U-shape error curve with variable k where the

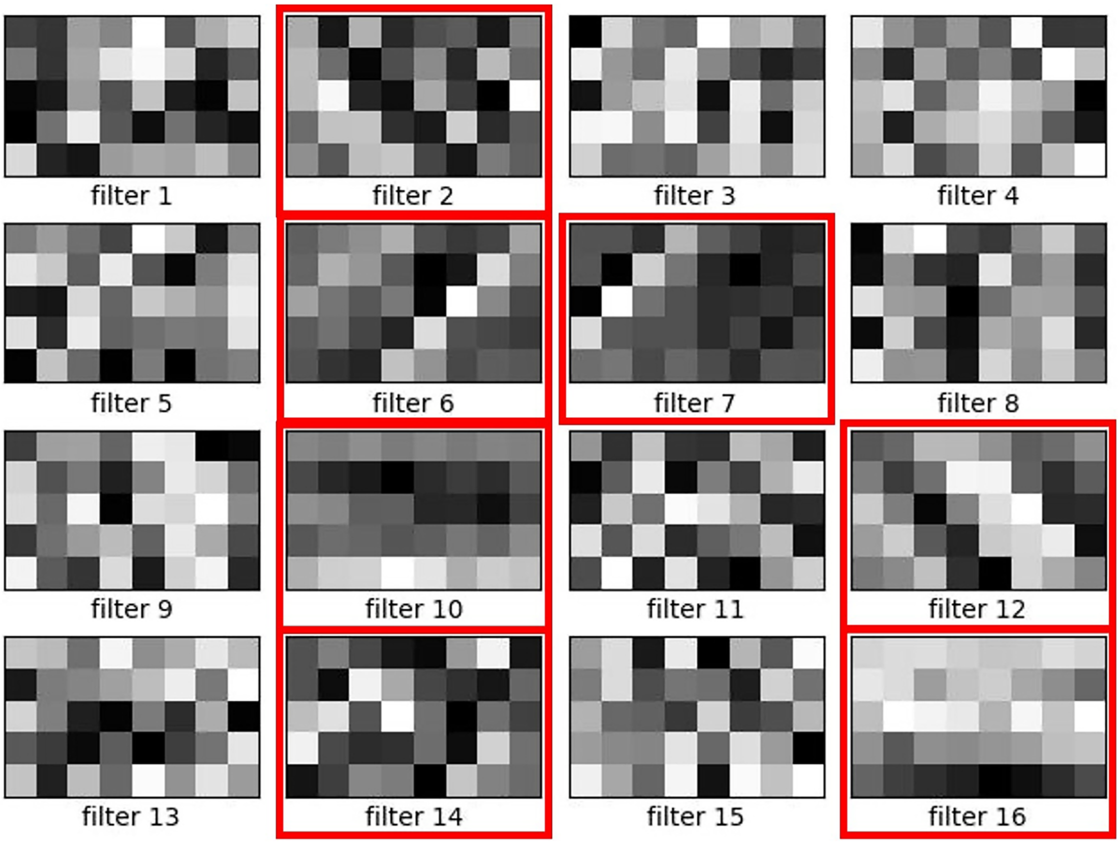

FIG. 8. (Color online) Visualization of 16 learned filters from the single-transit model trained on the summer testing fold. Notably, seven filters (highlighted with red boxes) exhibit patterns reminiscent of the striation pattern associated with the waveguide invariant property.

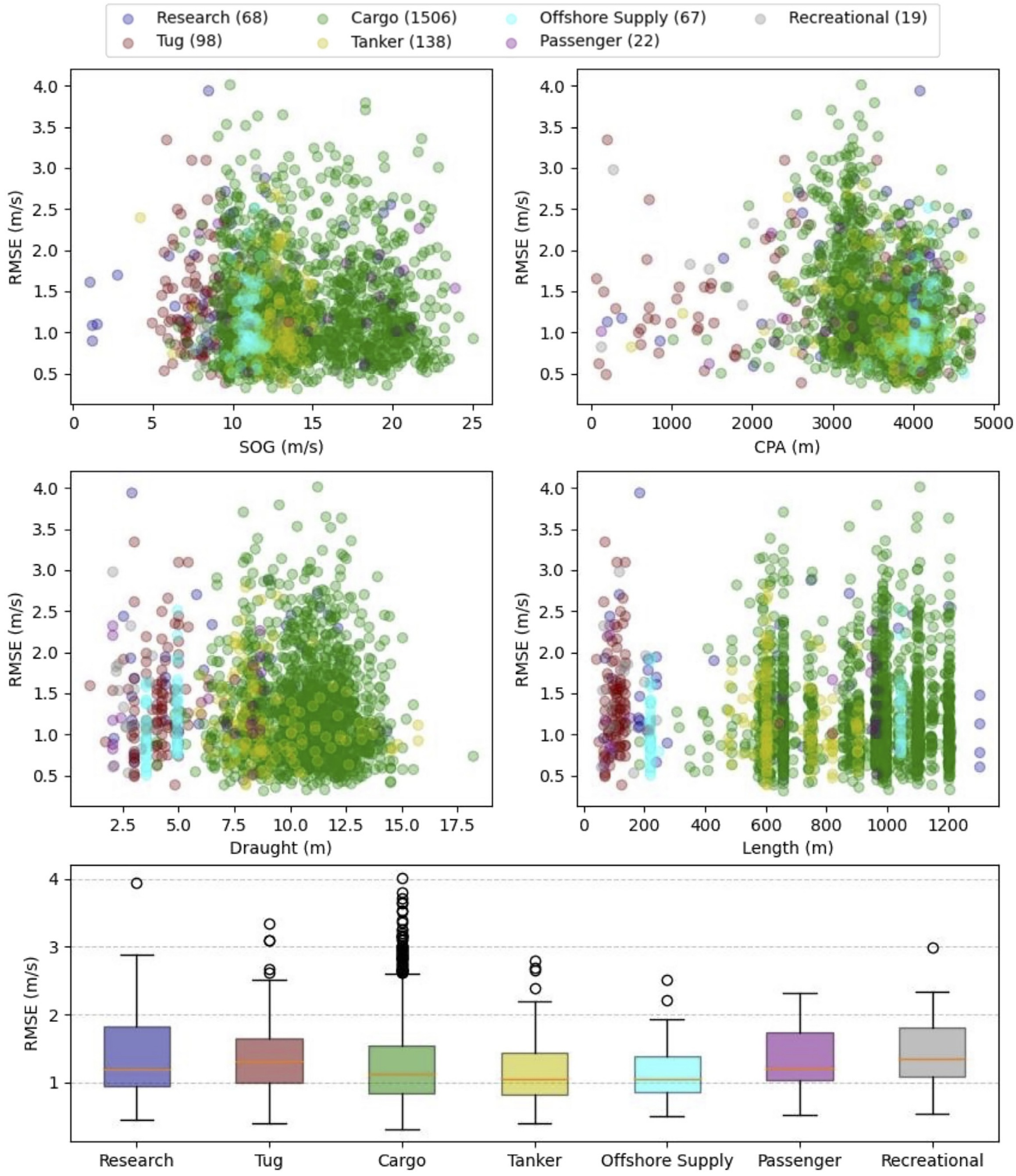

FIG. 9. (Color online) (Top) Scatterplots illustrating RMSE of the single-transit model across the four continuous valued predictor variables, vessel speed over ground (SOG), closest point of approach (CPA), draught, and length, with color-coding representing the seven ship types. (Bottom) A box plot shows RMSE by ship type. This visualization incorporates aggregated testing data from all four testing folds. The number of data points from each ship type is shown in parentheses next to each ship name in the legend. The consistent vertical dispersion of points in each subplot highlights the absence of discernible performance trends, indicating stability and reliability across predictor variables.

optimal value for *k* needs to be found through experimentation (Fig. 7).

Our proposed token learning method has the advantage of scaling to arbitrary input size while maintaining a fixed and relatively small number of trainable parameters, which may improve generalizability. Moreover, in this application, we anticipate that employing a weighted sum, where all features within the transit receive the same weight, rather than using feature-specific fusion, will improve performance. Most multi-view models are developed under the implicit assumption that each observation is uniquely informative regarding the target variable. In other words, each observation contains predictive information that the other observations do not contain (e.g., consider two images of the same plant, where one image captures the detail of the leaf and the other captures the flower). For our application, it is unlikely that different transits contain this kind of complementary information. Instead, some transits exhibit higher

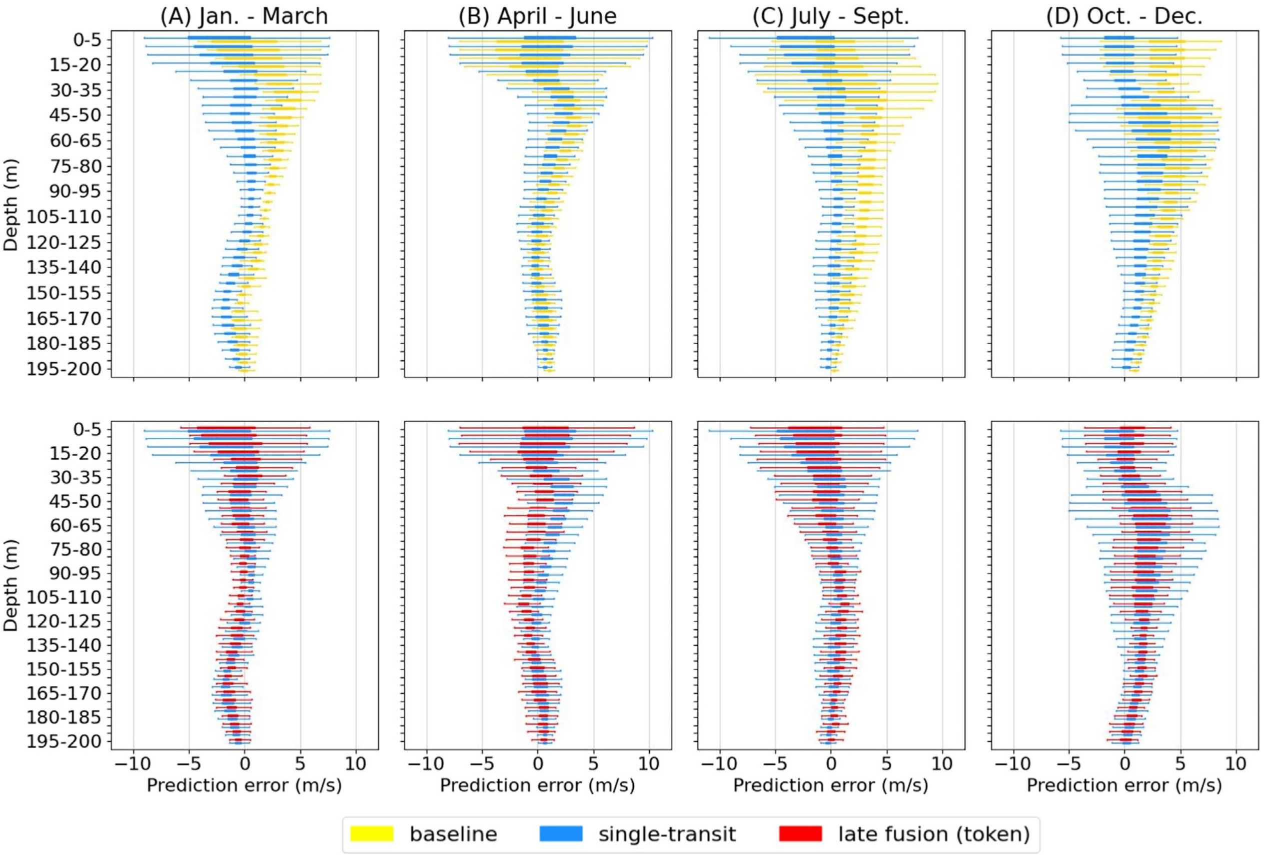

FIG. 10. (Color online) Comparison of baseline, single-transit, and late fusion (token) model prediction residuals by season (columns). The box plots show the distribution of model residuals with respect to depth, grouping residuals for every 5-m depth interval. Points below the lower fence and above the upper fence are not shown. (Top) Comparison of baseline (yellow) and single-transit (blue). (Bottom) Comparison of single-transit (blue) and late fusion (token [red]).

saliency compared to others, and the goal of leveraging multiple observations is to rank the salience, in contrast to pooling information across the transits. This approach may have broad applicability in oceanographic acoustic observation problems involving large amounts of weakly curated data in which feature salience is variable in time and space, particularly if the salience of relevant features is difficult to estimate *a priori*. If multiple sensors were available, fusion approaches could be used to incorporate simultaneous views.

An important limitation of this approach is the availability of SSP estimates for model training. Although quarterly *in situ* measurements were available from a nearby CalCOFI station and periodic local glider transits, these were determined to be too infrequent for training; therefore, this study used data assimilative hindcasts for training. Agreement between these regionally specific hindcasts and the available *in situ* measurements was high for this well-sampled, highly studied region. Further experimentation is needed to evaluate whether this approach could be used to refine or improve hindcast estimates, particularly in under-sampled regions. Additionally, the proposed method represents a preliminary exploration aimed at evaluating the feasibility of extracting sound-speed relevant features from single sensor acoustic recordings. Further development will be required to adapt this method for use across different recording environments.

V. CONCLUSION

In this paper, a neural network-based model, which uses acoustic recordings of URN of transiting ships and their transit metadata, is proposed to predict SSPs. Additionally, we propose a data fusion strategy suitable for large observational acoustic datasets, in which data are weakly curated and feature salience differs between observations used for prediction. Our results show that the addition of vessel transit recordings markedly improved the estimation of SSPs compared to the use of historical averages. We show that multiple transit recordings can be leveraged together to improve SSP estimation and compare multiple techniques for combining available information. We note that this work serves as a first approach in estimating oceanic SSPs from vessel URN, and there still exist sources of error in the estimations of the best performing model. Future work incorporating other data modalities and alternative hydrophone configurations may help further reduce this estimation error.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research and Task Force Ocean program manager Robert Headrick (Grant No. 200249). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1] Available at https://marinecadastre.gov/.

Brekhovskikh, L. M., Lysanov, Y. P., and Beyer, R. T. (**1991**). "Fundamentals of ocean acoustics," J. Acoust. Soc. Am. **90**(6), 3382–3383.

Chen, C., Ma, Y., and Liu, Y. (**2018**). "Reconstructing sound speed profiles worldwide with sea surface data," Appl. Ocean Res. **77**, 26–33.

Chen, C.-T., and Millero, F. J. (**1977**). "Speed of sound in seawater at high pressures," J. Acoust. Soc. Am. **62**(5), 1129–1135.

Chuprov, S. (**1982**). "Interference structure of a sound field in a layered ocean," in *Ocean Acoustics, Current Status*, edited by L. M. Brekhovskikh and I. B. Andreevoi (Nauka, Moscow), pp. 71–91 (in Russian).

Collins, M. D. (**1993**). "A split-step Padé solution for the parabolic equation method," J. Acoust. Soc. Am. **93**(4), 1736–1742.

Escobar-Amado, C. D., Neilsen, T. B., Castro-Correa, J. A., Van Komen, D. F., Badiey, M., Knobles, D. P., and Hodgkiss, W. S. (**2021**). "Seabed classification from merchant ship-radiated noise using a physics-based ensemble of deep learning algorithms," J. Acoust. Soc. Am. **150**(2), 1434–1447.

Feng, Y., Zhang, Z., Zhao, X., Ji, R., and Gao, Y. (**2018**). "GVCNN: Group-view convolutional neural networks for 3D shape recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT (IEEE, New York), pp. 264–272.

Ferguson, E. L., Williams, S. B., and Jin, C. T. (**2018**). "Sound source localization in a multipath environment using convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada (IEEE, New York), pp. 2386–2390.

Frasier, K., ZoBell, V., MacGillivray, A., Dolman, J., Ainsworth, L., and Zhao, J. (**2022**). "Evaluation of ECHO Vessel Noise Correlation Models with a Novel Dataset Collected in the Santa Barbara Channel," Technical Report No. 658 (JASCO Applied Sciences, Silver Spring, MD).

Gassmann, M., Wiggins, S. M., and Hildebrand, J. A. (**2017**). "Deep-water measurements of container ship radiated noise signatures and directionality," J. Acoust. Soc. Am. **142**(3), 1563–1574.

Gemba, K. L., Sarkar, J., Cornuelle, B., Hodgkiss, W. S., and Kuperman, W. A. (**2018**). "Estimating relative channel impulse responses from ships of opportunity in a shallow water environment," J. Acoust. Soc. Am. **144**(3), 1231–1244.

Gervaise, C., Kinda, B. G., Bonnel, J., Stéphan, Y., and Vallez, S. (**2012**). "Passive geoacoustic inversion with a single hydrophone using broadband ship noise," J. Acoust. Soc. Am. **131**(3), 1999–2010.

ISO 17208-1:2019 (**2019**). "Underwater acoustics—Quantities and procedures for description and measurement of underwater sound from ships—Part 1: Requirements for precision measurements in deep water used for comparison purposes" (International Organization for Standardization, Geneva, Switzerland).

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. (**2022**). "Visual Prompt Tuning," arXiv:2203.12119.

Kirsebom, O. S., Frazao, F., Simard, Y., Roy, N., Matwin, S., and Giard, S. (**2020**). "Performance of a deep neural network at detecting North Atlantic right whale upcalls," J. Acoust. Soc. Am. **147**(4), 2636–2646.

Koch, R. A., and Knobles, D. P. (**2005**). "Geoacoustic inversion with ships as sources," J. Acoust. Soc. Am. **117**(2), 626–637.

Lin, C., and Kumar, A. (**2018**). "Contactless and partial 3D fingerprint recognition using multi-view deep representation," Pattern Recognit. **83**, 314–327.

Liu, F., Shen, T., Luo, Z., Zhao, D., and Guo, S. (**2021**). "Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation," Appl. Acoust. **178**, 107989.

Lovett, J. R. (**1978**). "Merged seawater sound-speed equations," J. Acoust. Soc. Am. **63**(6), 1713–1718.

McKenna, M. F., Wiggins, S. M., and Hildebrand, J. A. (**2013**). "Relationship between container ship underwater noise levels and ship design, operational and oceanographic conditions," Sci. Rep. **3**(1), 1760.

Park, C., Seong, W., and Gerstoft, P. (**2005**). "Geoacoustic inversion in time domain using ship of opportunity noise recorded on a horizontal towed array," J. Acoust. Soc. Am. **117**(4), 1933–1941.

Pereira, A., Harris, D., Tyack, P., and Matias, L. (**2020**). "On the use of the Lloyd's mirror effect to infer the depth of vocalizing fin whales," J. Acoust. Soc. Am. **148**(5), 3086–3101.

Ross, D. (**1976**). *Mechanics of Underwater Noise* (Pergamon Press, New York).

Seeland, M., and Mäder, P. (**2021**). "Multi-view classification with convolutional neural networks," PLoS One **16**(1), e0245230.

Stammer, D., Wunsch, C., Giering, R., Eckert, C., Heimbach, P., Marotzke, J., Adcroft, A., Hill, C. N., and Marshall, J. (**2002**). "Global ocean circulation during 1992–1997, estimated from ocean observations and a general circulation model," J. Geophys. Res. **107**(C9), 1-1–1-27, https://doi.org/10.1029/2001JC000888.

Stotts, S. A., Koch, R. A., Joshi, S. M., Nguyen, V. T., Ferreri, V. W., and Knobles, D. P. (**2010**). "Geoacoustic inversions of horizontal and vertical line array acoustic data from a surface ship source of opportunity," IEEE J. Ocean. Eng. **35**(1), 79–102.

Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. (**2015**). "Multi-view convolutional neural networks for 3D shape recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile (IEEE, New York), pp. 945–953.

Tollefsen, D., Dosso, S. E., and Knobles, D. P. (**2020**). "Ship-of-opportunity noise inversions for geoacoustic profiles of a layered mud-sand seabed," IEEE J. Ocean. Eng. **45**(1), 189–200.

Tréboutte, A., Carli, E., Ballarotta, M., Carpentier, B., Faugère, Y., and Dibarboure, G. (**2023**). "KaRIn noise reduction using a convolutional neural network for the SWOT ocean products," Remote Sens. **15**(8), 2183.

Tseng, Y.-H., Ding, R., and Huang, X.-m. (**2017**). "The warm Blob in the northeast Pacific—the bridge leading to the 2015/16 El Niño," Environ. Res. Lett. **12**(5), 054019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (**2017**). "Attention is all you need," in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY), Vol. 30.

Verlinden, C. M. A., Sarkar, J., Cornuelle, B. D., and Kuperman, W. A. (**2017**). "Determination of acoustic waveguide invariant using ships as sources of opportunity in a shallow water marine environment," J. Acoust. Soc. Am. **141**(2), EL102–EL107.

Wiggins, S. M., and Hildebrand, J. A. (**2007**). "High-frequency acoustic recording package (HARP) for broad-band, long-term marine mammal monitoring," in *2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies*, Tokyo, Japan (IEEE, New York), pp. 551–557.

Zaba, K. D., Rudnick, D. L., Cornuelle, B. D., Gopalakrishnan, G., and Mazloff, M. R. (**2018**). "Annual and interannual variability in the California current system: Comparison of an ocean state estimate with a network of underwater gliders," J. Phys. Oceanogr. **48**(12), 2965–2988.

ZoBell, V. M., Gassmann, M., Kindberg, L. B., Wiggins, S. M., Hildebrand, J. A., and Frasier, K. E. (**2023**). "Retrofit-induced changes in the radiated noise and monopole source levels of container ships," PLoS One **18**(3), e0282677.