Transferring elastic low frequency extrapolation from synthetic to field data

Introduction
The presence of low-frequency content in acquired seismic data is a known remedy for full-waveform inversion (FWI). In particular, the low-frequency data illuminates large scale structures in the subsurface and, thus, might compensate for the missing good starting model for optimization (Long et al., 2014). However, due to hardware and natural limitations, the frequency content below 3-4 Hz is usually absent in the real-world seismic data. While difficult to reconstruct these missing data in conventional ways, data-driven approaches showed promise in relevant applications. Low-frequency extrapolation of acoustic data was discussed in Ovcharenko et al. (2019); Sun and Demanet (2019); Fabien-Ouellet (2020). Wang et al. (2020) proposed a self-supervised approach to link successive frequency bands. Another self-supervised approach, but in model domain was proposed by Hu et al. (2020). Aharchaou et al. (2020) showed a field data application on ocean-bottom node data. Elastic land data was extrapolated in synthetic setup in (Ovcharenko et al., 2020; Sun and Demanet, 2020). Low-wavenumber velocity model was recovered from early FWI gradients and directly from the data in Plotnitskii et al. (2020) and Kazei et al. (2019b), respectively.

While S-waves cannot be directly recorded by a streamer, three elastic parameters can theoretically be reconstructed from P-waves alone, provided that sufficient aperture and frequency content is available (Kazei and Alkhalifah, 2019). We consider elastic parametrization of the subsurface in a marine streamer setup and attempt to reconstruct the missing low-frequency content in the acquired field data. For this purpose, we train a deep learning model exclusively on synthetic data where the full band is known and then predict the missing low-frequencies in field marine streamer data. However, the major challenge of machine learning applications is to transfer knowledge learned from synthetic datasets to target datasets acquired in reality. Source parameters, noise imprint and the physics of wave propagation are hard to match between synthetic and real-world seismic data. Here we describe the workflow where we start with processed field data and use it as a source of properties for generating synthetic waveforms, which closely match field observations.

Marine streamer data
The target field data for extrapolation are shot gathers acquired by CGG. The marine survey took place in the North-Western Australia Continental Shelf, where the data were recorded using the Broadseis acquisition system with a variable depth streamer (Soubaras and Dowle, 2010). The single-line data contain 1824 shots and it is high-pass filtered by the provider for frequencies above 2.5 Hz. We use data from half of the available hydrophones, which results in 324 channels spaced by 25 m for every shot. The near and far offsets are about 165 m and 8.3 km, respectively. The recording time is 7 sec with 2 ms sampling. We use these field data as a source of properties for synthetic data generation and as a target for inference in the end. In particular, the near-offset part of the field data serves as a donor of source signature, while the far offset part before first arrival delivers the noise signature. We also build a smooth 1D RMS velocity model by velocity analysis and use it as a seed for generation of random subsurface initializations.

Generation of the training dataset
The training dataset is an essential part of a deep learning solution. Given training data that are statistically similar to the target data distribution, even a simple neural network might deliver plausible results. This is why the generation of realistic datasets is the main focus of this work.

Synthetic subsurface models. Our set of training models is based on the local trends in the elastic parameters and represents an intermediate solution between search for general model (Ovcharenko et al., 2019) and solution tailored to a particular dataset (Hu et al., 2020). In particular, we use a 1D velocity produced by velocity analysis on field data as a seed for generating pseudo-random distributions of compressional velocities, $V_p$. Then we analytically create shear wave velocity, $V_s$, and density, $\rho$, distributions as $V_p/\sqrt{3}$ and $300*\sqrt[3]{V_p}$, respectively. To produce a random $V_p$ model we first sample a random reflectivity series as deep as desired model. Then we integrate it to produce a velocity log and subtract the mean from it. Then we replicate the produced log of velocity perturbations for desired width of the model and add the 1D seed trend from field data. Next, we follow Kazei et al. (2019a) and apply
an elastic transform to distort the layered model. Finally, we select the maximum value between the 1D seed trend and the produced model (Figure 1).

Figure 1: The generation of random subsurface models. Central logs for 256 random models (left), respective mean and standard deviation (central) and two examples of generated random models (right).

Wavefield simulation. Given a set of elastic models, parametrized by $V_p$, $V_s$ and $\rho$, we run elastic wave propagation in them, powered by Python API of DENISE-Black-Edition package (Köhn, 2011). Source signature of an airgun defines properties of the recorded streamer data, which matches our observation that using a Ricker wavelet as a source term for modeling is insufficient. Thus, we use the mean of source signatures from Kalita and Alkhalifah (2017) and inject it as signature for all sources. To avoid numerical instability caused by non-zero initiation of extracted source wavelet, we also pad the source signature with zeros and low-pass it below 10 Hz. This leads to a fair match of direct arrivals in water between synthetic wavefield and observed data (Figure 2, right). Finally, we fit shots into range [-1, 1] by dividing them by maximum of its absolute values.

Adding realistic noise. Matching direct arrivals and general frequency content is still insufficient to enable knowledge transfer from synthetic to field data. The missing component is realistic noise. Meaning that the training works well on synthetic data, but is not capable of inference on noisy field data. Adding realistic noise solves this issue. However, the frequency content of simple random Gaussian noise is skewed toward high frequencies and it is likely to disappear when the data is low-pass filtered at the training stage. The viable option for a streamer survey is to extract noise patterns directly from far offsets in the field data, before the direct wave arrival. We extract such triangles of clean noise (Figure 2, left) from 38 field shots, and then flip and duplicate them to randomly tile the entire synthetic shot.

Figure 2: Comparison of elastic synthetic and field data. Not-matching shot gathers (left), respective power spectra (central) and direct wave arrivals (right). Ambient noise is extracted from empty areas of field data (dashed box) and then added to synthetic data.

Overview. To sum up, starting from a 1D seed velocity trend extracted from field data, it takes us about 5 sec to generate 256 elastic pseudo-random models, measuring $[152 \times 500]$ grid points with 25 m spacing (Figure 1). We use the realistic source signature to simulate 6 seconds-long wavefield recordings for 3 shots in each of generated elastic models. This results in 768 synthetic shot gathers, split as 612/78/78 into training, validation and testing partitions, respectively. Every synthetic shot
gather initially measures \([324 \text{ receivers } \times 3000 \text{ time steps}]\), but we then low-pass it below 10 Hz and downsample to 375 samples along last dimension.

The final step of the training dataset preparation is to split the synthetic data into bands of input high-frequency (HF) and target low-frequency (LF) data. The input data are band-pass filtered between 4 and 10 Hz, while the target data are low-pass filtered below 4 Hz. Moreover, we hard-set zeros for any frequency content below 2.5 Hz in the input HF data to avoid low-frequency signal leakage. We repeat the procedure for field data shots and use the HF partition from all field shots as a donor of realistic noise. An example pair of input-target synthetic data are shown in Figure 3. Note, that the LF target is noise-free, so the side-product of the network training is denoising.

**Figure 3:** Sample of synthetic training data. Bandpass filters (left) and respective pair of high-frequency (>5 Hz) and low-frequency (<3 Hz) data, measuring \([324 \times 375]\).

**Extrapolation results**

The perceptual and statistical proximity of training (synthetic) and testing (field) datasets allows using a general-purpose U-Net architecture (Ronneberger et al., 2015) for image-to-image translation to address the task of low-frequency extrapolation. Otherwise, when datasets differ significantly, one should implement concepts of domain adaptation to match datasets at the training stage. We train the original model for 150 epochs, with a batch size of 4 and a learning rate of \(1\text{e}^{-4}\) using Adam optimizer (Kingma and Ba, 2014). After training on synthetic data, we run inference of the trained model on a field shot gather, band-pass filtered between 4 and 10 Hz. Figure 4 shows extrapolation results for, visualized for low-pass bands below 10 Hz (left) and 3 Hz (right). Obviously, the network succeeded in the side-task of removing noise, recovering underlying continuous waveforms. Signal in filed data < 3 Hz is almost indistinguishable, while prediction in this range looks meaningful. The proper quality control would be running FWI on extrapolated data.

**Figure 4:** The predicted low-frequency data for a shot gather recorded during marine streamer survey. Visualization of observed and predicted data in low-pass bands <10 Hz (left) and <3 Hz (right).
Conclusions
We trained a convolutional neural network to predict low frequencies on synthetic elastic data and applied it to field data. The same approach can be suitable for land data and shallow water regions as elastic effects are covered by the synthetic dataset. Synthetic data diversity, correct source estimation and realistic noise are crucial in bridging the gap between synthetic and field data domains. However, the drawback of the described workflow is that it is not automatic and requires substantial human efforts.

Acknowledgements
The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia. We also thank Mahesh Kalita and Abdullah Alali from KAUST for their help.

References