RESEARCH ARTICLE

Key Points:
- Clustering is used to improve the retrieval of offshore-onshore deconvolution functions (DFs)
- Typhoons and storms play an important role in the retrieval of clear offshore-onshore DFs
- DFs retrieved with our method can be used to better simulate the long-period ground motions from subduction earthquakes

Supporting Information:
- Supporting Information S1

Correspondence to:
L. Viens,
viens.loic.58r@st.kyoto-u.ac.jp

Improving the Retrieval of Offshore-Onshore Correlation Functions With Machine Learning

Loïc Viens[1] [ID] and Tomotaka Iwata[1]

[1]Disaster Prevention Research Institute, Kyoto University, Gokasho, Uji, Kyoto, Japan

Abstract The retrieval of reliable offshore-onshore correlation functions is critical to improve our ability to predict long-period ground motions from megathrust earthquakes. However, localized ambient seismic field sources between offshore and onshore stations can bias correlation functions and generate nonphysical arrivals. We present a two-step method based on unsupervised learning to improve the quality of correlation functions calculated with the deconvolution technique (e.g., deconvolution functions, DFs). For a DF data set calculated between two stations over a long time period, we first reduce the data set dimensions using the principal component analysis and cluster the features of the low-dimensional space with a Gaussian mixture model. We then stack the DFs belonging to each cluster together and select the best stacked DF. We apply our technique to DFs calculated every 30 min between an offshore station located on top of the Nankai Trough, Japan, and 78 onshore receivers. Our method removes spurious arrivals and improves the signal-to-noise ratio of DFs. Most 30-min DFs selected by our clustering method are generated during extreme meteorological events such as typhoons. To demonstrate that the DFs obtained with our method contain reliable phases and amplitudes, we use them to simulate the long-period ground motions from a M_w 5.8 earthquake, which occurred near the offshore station. Results show that the earthquake long-period ground motions are accurately simulated. Our method can easily be used as an additional processing step when calculating offshore-onshore DFs and offers a new way to improve the prediction of long-period ground motions from potential megathrust earthquakes.

Plain Language Summary Seismic waves from subduction earthquakes are generally characterized by a strong and elongated long-period component due to their propagation through complex velocity structures such as accretionary wedges. Seismic interferometry, which consists of cross-correlating continuous ambient seismic field signals at two seismic stations, can be used to retrieve the wave propagation between the two sensor's locations. However, the retrieval of clear wave propagation between offshore and onshore stations is difficult due to the characteristics of the ambient seismic field. We develop a method based on unsupervised learning to improve the quality of correlation functions between offshore and onshore sites. We apply our method to correlation functions calculated between an offshore station on top of the Nankai Trough, Japan, and surrounding onshore stations. The correlation functions retrieved with our method can be used to better simulate the ground motions from a M_w 5.8 earthquake, which occurred along the Nankai Trough. Improving our ability to retrieve accurate wave propagation between offshore and onshore stations is critical to better predict the long-period ground motion from potential megathrust earthquakes, which are likely to happen along subduction zones worldwide in the near future.

1. Introduction

Seismic interferometry is a well-established method used to gain geophysical information about the Earth's subsurface. By cross-correlating ambient seismic field time series recorded by a pair of seismometers, the seismic wave propagation between the two sensor's locations can be retrieved. Theoretical studies demonstrated that for homogeneously distributed ambient seismic field sources and/or a fully diffuse medium, the cross-correlation function (CCF) should yield the true Green's function of the medium (Fichtner & Tsai, 2019; Weaver & Lobkis, 2001). However, such conditions are rarely fulfilled on Earth as the ambient seismic field is primarily generated by ocean waves at long periods (>1 s) and by human activities at short periods.

Station-to-station CCFs are generally calculated over short ambient seismic field time windows ranging from a few minutes to hours and are then linearly stacked over a longer time period to increase their

signal-to-noise ratios (SNRs). To improve the retrieval of the phase information of the CCFs, the short ambient seismic field records are routinely preprocessed using 1-bit normalization and/or prewhitening (Bensen et al., 2007). The phase information of preprocessed CCFs has been extensively used to image the Earth's subsurface (Lin et al., 2008; Shapiro et al., 2005) and to monitor temporal changes occurring in the Earth through time (Brenguier, Campillo, et al., 2008; Brenguier, Shapiro, et al., 2008). However, the preprocessing steps generally involve nonlinear operations which can bias the amplitude information of the CCFs.

Empirical studies showed that seismic interferometry by deconvolution with no preprocessing can be used to retrieve both the amplitude and phase information of CCFs (Viens et al., 2017). Deconvolution functions (DFs) have been used to simulate the long-period ground motions from moderate (Denolle et al., 2013; Prieto & Beroza, 2008; Sheng et al., 2017; Viens et al., 2014; Viens, Koketsu, et al., 2016) and large (Denolle et al., 2014, 2018; Viens, Miyake, & Koketsu, 2016) crustal earthquakes as well as mine collapse events (Kwak et al., 2017). However, the retrieval of reliable amplitudes is still debated as it strongly depends on the location and characteristics of ambient seismic field sources (Stehly & Boué, 2017; Stehly et al., 2006; Tsai, 2011).

The recent release of continuous data recorded by ocean bottom seismometers deployed on top of subduction zones worldwide offers new opportunities to better understand the complex seismic wave propagation through accretionary wedges. However, the retrieval of unbiased DFs between offshore and onshore stations is challenging as the ocean bottom environment is generally noisier than continental sites (Webb, 1998). Moreover, localized ambient seismic field sources between station pairs, such as ocean storms, can corrupt the DFs with spurious arrivals (Retailleau et al., 2017; Shapiro et al., 2006). Along the Nankai Trough, Japan, offshore-onshore DFs have been calculated and used to successfully simulate the long-period ground motions from moderate (Viens et al., 2015) and large (Viens & Denolle, 2019) subduction earthquakes. Nevertheless, the computed DFs are noisier than that retrieved between onshore station pairs and tend to contain strong spurious arrivals, which do not simply vanish by stacking DFs over a longer time period.

Several stacking schemes have been developed to improve the retrieval of the wave propagation in case of noisy and/or corrupted CCFs. For example, Chaput et al. (2015) introduced an optimization approach that uses a Markov chain Monte Carlo scheme to improve the convergence of correlation functions. Weaver and Yoritomo (2018) proposed multiple techniques to improve the retrieval of CCFs and applied them to synthetic data. Recently, Xie et al. (2020) developed a root-mean-square ratio selection stacking method that only choose short-duration CCFs that contribute constructively to the retrieval of clear CCFs. All these techniques significantly improve the retrieval of the phase information of the CCFs, but little attention has been given to the retrieval of accurate amplitudes.

In this study, we introduce a two-step technique based on unsupervised learning to improve the recovery of offshore-onshore DFs with accurate phases and amplitudes. The first step consists of computing the principal component analysis (PCA) of a prestack DF data set calculated from short ambient seismic field time windows (e.g., a few minutes to hours) over a long period of time (e.g., 1 year) between two seismic stations. We only keep the output of the first n principal components (PCs), with n being much smaller than the duration of each DF, in order to reduce the dimensions of the data set. Second, we cluster the data from the low-dimensional space and linearly stack the DFs belonging to each cluster together. Such a two-step method has been used to cluster different types of high-dimensional data sets, such as DNA gene expression and internet newsgroups (Ding & He, 2004). In seismology, a similar approach using dimensionality reduction with autoencoders and clustering has been developed to classify earthquake waveforms (Mousavi et al., 2019).

We first present our two-step clustering method and apply it to a synthetic data set. We then introduce the computation of offshore-onshore DFs between seismic stations located on top and in the vicinity of the Nankai Trough. We then compare the DFs obtained with our method to that obtained with linear stacking and with a more sophisticated stacking method. To validate the retrieval of DFs with accurate phases and amplitudes, we finally transform the DFs into velocity waveforms and compare them to the long-period (4–10 s) velocity waveforms from a moderate M_w 5.8 subduction earthquake, which occurred in the vicinity of an offshore station.

2. Methods: A Two-Step Clustering Technique and Its Application to a Synthetic Data Set

Clustering is a machine learning method that is used to partition a data set into different groups with similar characteristics in an unsupervised manner. While clustering algorithms, such as k-means and Gaussian mixture models (GMMs), perform well on low-dimensional data sets, their performance rapidly decreases as the dimension of the data increases (Steinbach et al., 2004). In seismic interferometry, thousands of DFs with a duration of a few hundred seconds can be computed from one year of continuous data recorded by a station pair, resulting in high-dimensional data sets.

2.1. PCA and GMM

The PCA is a popular statistical approach to reduce the number of dimensions of a large data set into a low-dimensional set of features. This is achieved by transforming the input data into a set of uncorrelated, orthogonal, PCs. The PCs are ordered so that the first PC explains the largest data variance, the second PC retains the second largest variance, and so forth. For more details about the PCA, we refer the reader to the extensively literature about the method (Jolliffe, 2002, and references therein).

In this study, our goal is to reduce the dimension of prestack DF data sets, which are calculated from short 30-min ambient seismic field time windows recorded over 1 year by station pairs (more details about the computation of the DFs are given in section 3.1). Each DF data set is first standardized (e.g., removal of the mean and amplitude scaling to unit variance) prior to compute its PCA. For the offshore-onshore DFs of this study, we keep the output of the first 20 PCs (e.g., $n = 20$) and discuss our choice in supporting information Text S1 and Table S1. The data from the low-dimensional space of the first n PCs are then clustered using a GMM. This probabilistic model clusters the data by assuming that they are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. To learn the parameters of each Gaussian distribution, we use the expectation-maximization (EM) algorithm (Dempster et al., 1977).

One of the main problems of clustering algorithms is that the number of clusters needs to be accurately determined in advance to reduce potential underfitting or overfitting of the data (Figueiredo & Jain, 2002). To automatically determine the optimal cluster number for a given DF data set, we cluster the output of its first n PCs using GMMs with different numbers of clusters and compute a Bayesian Information Criterion (BIC, Schwarz, 1978) score for each model. The optimal model is generally selected as the one with the lowest BIC score. Note that the formulation of the BIC considers a trade-off between model fitting and model complexity, with a penalty term to penalize more complex models which are most likely to better fit the data.

As stated above, the optimal number of clusters is generally identified with the minimum BIC value. However, it has been argued that the location of a knee (also called kink or elbow) in a BIC versus cluster number plot represents better the optimal number of clusters (Murphy, 2012; Zhao et al., 2008). While several methods have been developed to determine the knee location in BIC curves, we use that from Satopaa et al. (2011), which is based on the mathematical definition of curvature for a continuous function. For all the DF data sets considered in this study, the optimal number of clusters ranges between 2 and 6.

GMM clustering is performed on the output of the first n PCs using the optimal number of clusters and the DFs belonging to each cluster are linearly stacked together. Finally, one only needs to select the stacked DF that minimizes spurious arrivals and maximizes the symmetry between the anticausal (negative) and causal (positive) parts among the two to six stacked DFs. We provide more information about the automatic selection of the best DF in section 3.3. The main advantage of using unsupervised learning to improve the retrieval of DFs is that only a few easily tunable parameters are required to select the waveforms, which need to be stacked together. Moreover, applying this method to a DF data set with thousands of waveforms is fast and can easily be used as an additional step when processing ambient seismic field time series.

2.2. Synthetic Data Set

We first apply our method to a simple synthetic data set made of waveforms with known signal and noise properties (Figure 1). We consider the propagation of surface waves between two hypothetical stations (A and B). To reproduce the dispersive property of surface waves, we use a chirp function with initial and final frequencies of 0.05 and 0.25 Hz, respectively. The duration of the chirp signal is 70 s and starts at a lag time of 10 s. The absolute maximum amplitude of the chirp signal is constant through time with a value of 0.5, and
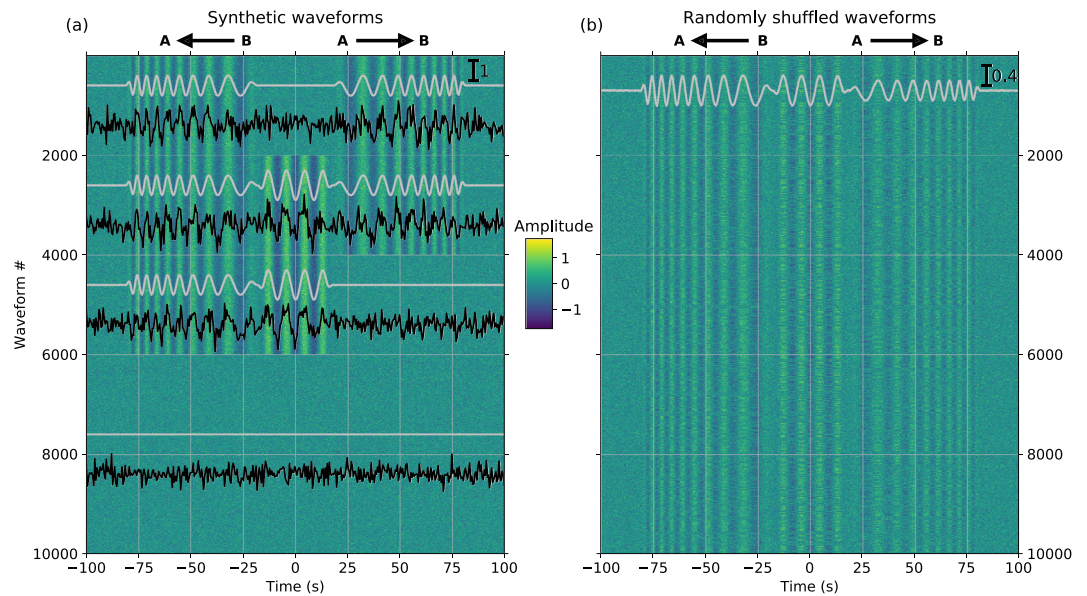
Figure 1. (a) Synthetic waveforms propagating from Station A to Station B (positive part) and from Station B to Station A (negative part). For the four different types of waveforms, an example of the clean and noisy traces is shown in gray and black, respectively. The background color represents the noisy waveforms. (b) Randomly shuffled waveforms. The gray trace represents the raw stack of the 10,000 waveforms. The amplitude scale of the gray waveform in the top right-hand corner is different than that in (a).

both ends of the signal are slightly tapered. The sampling rate of the waveforms is 2 Hz. In this synthetic example, we do not intend to replicate real correlation functions, which are difficult to simulate due to the complexity of the ambient seismic field source mechanisms and the wave propagation through the Earth.

To mimic a correlation function data set that could be biased by an uneven distribution of noise sources, we construct 10,000 waveforms with four different types of signals. The first 2,000 waveforms represent correlation functions obtained with a uniform distribution of the noise sources around the two stations. Therefore, both the anticausal and causal parts of the CCFs can be retrieved (Figure 1a, top gray waveform). The next 2,000 waveforms (Waveform Number 2,001 to 4,000) represent the case where noise sources are still uniformly distributed around the two stations, but local noise sources (e.g., ocean storms) are also active between the station pair. This leads to clear spurious arrivals in addition to the anticausal and causal signals. The spurious arrivals are created using a 0.11 Hz cosine function with a higher amplitude than the chirp signal. From Waveform Number 4,001 to 6,000, we consider the case where noise sources are located between the two stations and in the stationary phase zone of Station B. Therefore, the CCFs only contain the anticausal part of the signal (e.g., propagation from Station B to Station A) as well as spurious arrivals. The last of the four groups is composed of 4,000 waveforms with no signal. We then add some random noise drawn from Gaussian distributions with a mean of zero and a variance of 1 to all the waveforms. Note that the random noise values are normalized so that the absolute maximum noise value is equal to 1.0, which corresponds to twice the absolute maximum amplitude of the chirp signal, before being added to the waveforms (Figure 1a, background color and black waveforms). Finally, we randomly shuffle the 10,000 noisy waveforms and show them in Figure 1b. The raw stack of all the waveforms cancels the random noise but strong spurious arrivals can be observed and the anticausal and causal signals have different amplitudes (Figure 1b, gray waveform).

2.3. Clustering the Synthetic Waveforms

We standardize the shuffled data set, compute its PCA, and show its projection on the first two PCs in Figure 2a. For this synthetic example, we only keep the output of the first two PCs, which explain 20.4% of the cumulative data variance. This value is very close to the 26.0% cumulative data variance explained by the first 20 PCs. This can be explained by the fact that all waveforms in this synthetic example are constructed with similar chirp and cosine signals to which white noise is added. Therefore, the chirp and cosine signals
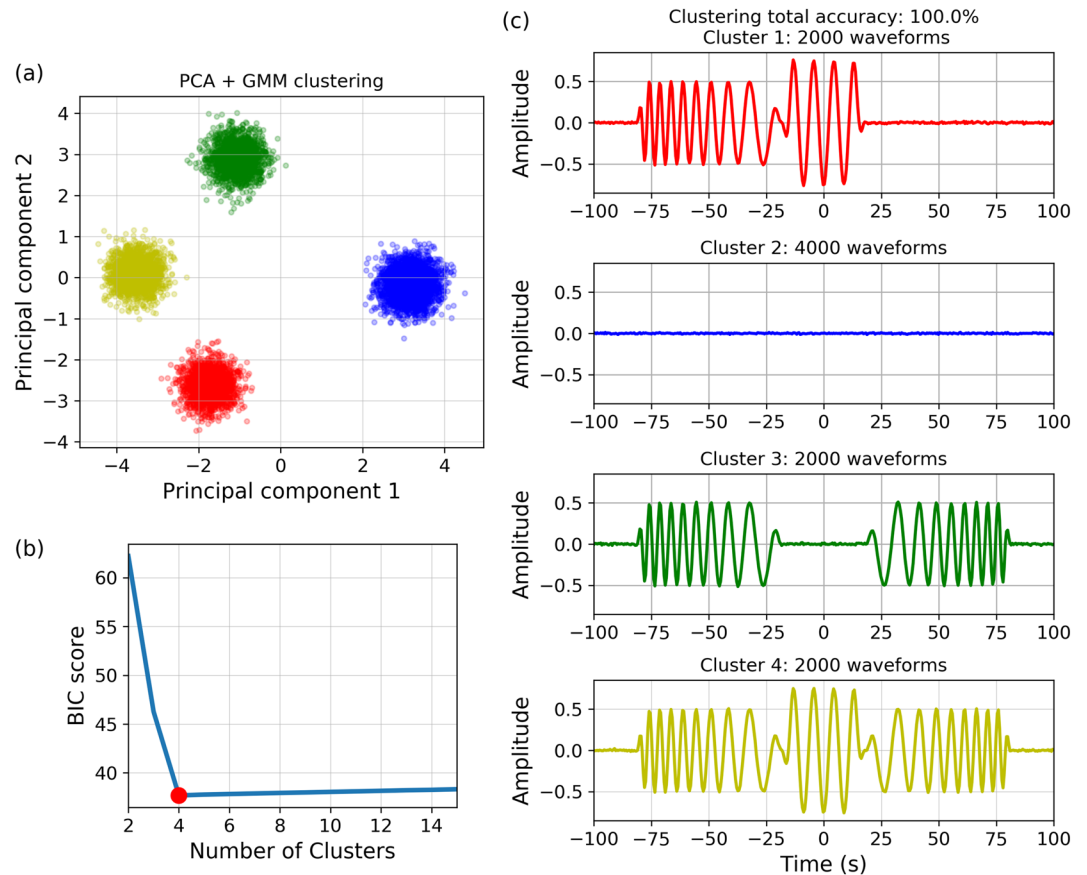
Figure 2. (a) Projection of the synthetic data on the first two principal components of the PCA. The four colors correspond to the four clusters obtained with the Gaussian mixture model (GMM). (b) BIC score for GMMs performed on the output of the first two PCs with different numbers of clusters. The lowest and optimal BIC value obtained with the knee method is found for four clusters. (c) Stack of the waveforms from each cluster. The random noise added to the data is canceled by the stacking of the waveforms and the four types of waveforms originally created are retrieved. The total accuracy of the clustering is also indicated on top of (c) and is 100% for this synthetic example.

are likely defined by the first two PCs and the white noise by the following PCs. For real offshore-onshore DF data sets, however, the output of more PCs is required due to the complexity of the waveforms.

To determine the appropriate number of clusters for the GMM, we cluster the output of the first two PCs using GMMs with 2 to 15 clusters and compute the BIC of each model (Figure 2b). We then use the knee method to determine that the optimal number of clusters is four (red dot in Figure 2b). For this synthetic case, the lowest value of the BIC corresponds to the optimal number of clusters. The clustering performed with the GMM using four clusters is represented by the colors in Figure 2a. We finally stack the waveforms belonging to each cluster and show them in Figure 2c. In this example, the four types of noisy waveforms are clustered with an accuracy of 100%, meaning that no waveform is misclassified. The linear stack of the data from each group cancels the random noise and allows us to retrieve the four types of initial signals. As the goal of seismic interferometry is to retrieve unbiased correlation functions with no spurious arrivals to infer the physical properties of the Earth, one only needs to select the waveform from Cluster 3 in this synthetic example. Note that the cluster number might change if the clustering is performed multiple times.

3. Application to Offshore-Onshore DFs

3.1. Computation of DFs

We focus on 1 year of continuous data recorded by the KME18 DONET sensor and 78 high-sensitivity Hi-net seismometers from 1 April 2015 to 31 March 2016. All the stations in this study record with a sampling rate of 100 Hz and are shown in Figure 3. The raw data are first corrected for their instrument responses, band-pass
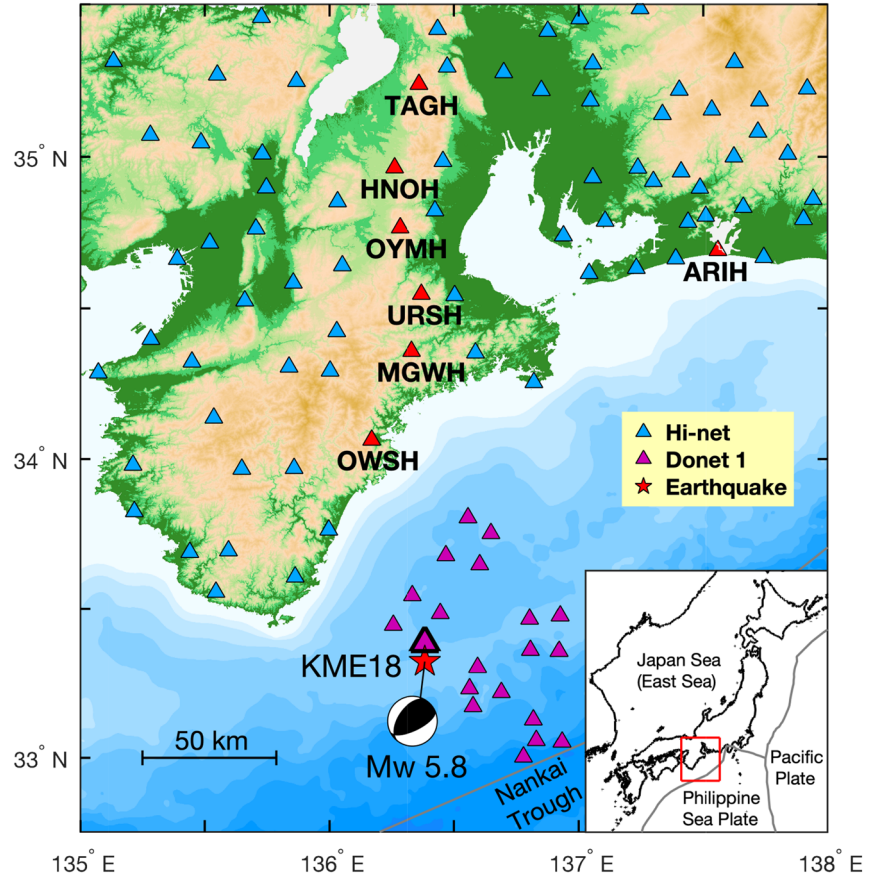
Figure 3. Topographic map of the region of interest, including the 20 offshore DONET 1 stations (purple) and the 78 onshore Hi-net receivers (blue). The location of the 2016 M_w 5.8 earthquake, which occurred 7 km away from the KME18 station, is shown by the red star with its focal mechanism. The seven Hi-net stations used in this study are highlighted by red triangles, and their names are also indicated. The inset map shows the Japan Islands, the plate boundaries (gray lines), and the location of the region of interest (red rectangle).

filtered between 1 and 20 s using a two-pass four-pole Butterworth filter and downsampled to 4 Hz to speed up the computation process. As the KME18 virtual source is located in a 1-m deep borehole and the 78 Hi-net receivers are buried in boreholes with depths ranging from 100 to 3,000 m, the two horizontal components are not always aligned along the north-south (N) and east-west (E) directions. Therefore, we rotate the two horizontal components of Hi-net and DONET stations to the true north and east directions using the orientations determined by Shiomi (2013) and Nakano et al. (2012), respectively. We then divide the data set into 30-min-long time series and discard windows with spikes larger than 10 times the standard deviation of the window to remove the effect of potential earthquakes. We finally compute the DFs between offshore and onshore stations as

$$\text{DF}_{i,j}(x_r, x_s, t) = \widetilde{\mathfrak{F}}^{-1}\left(\frac{\hat{v}_i(x_r, \omega)\hat{v}_j^*(x_s, \omega)}{\{|\hat{v}_j(x_s, \omega)|\}^2}\right), \tag{1}$$

where $\hat{v}_j(x_s, \omega)$ and $\hat{v}_i(x_r, \omega)$ are the Fourier transforms of 30-min-long velocity records at the offshore virtual source (x_s) and the onshore receiver (x_r) for the jth and ith components (e.g., N, E, or vertical Z). Note that we zero-pad the 30-min velocity records to 5 times their original length before computing the Fourier transform. The $*$ symbol denotes the complex conjugate, ω represents the frequency domain, $|\cdot|$ is the absolute value, and $\{\cdot\}$ represents a smoothing of the spectrum using a 10-point moving average to stabilize the denominator term. The inverse Fourier transform ($\widetilde{\mathfrak{F}}^{-1}$) is applied to retrieve the DFs between the two stations in the time domain (denoted by t). For each station pair, we rotate the nine-component DF tensor from the east-north-vertical (ENZ) coordinate system to the radial-transverse-vertical (RTZ) system, where
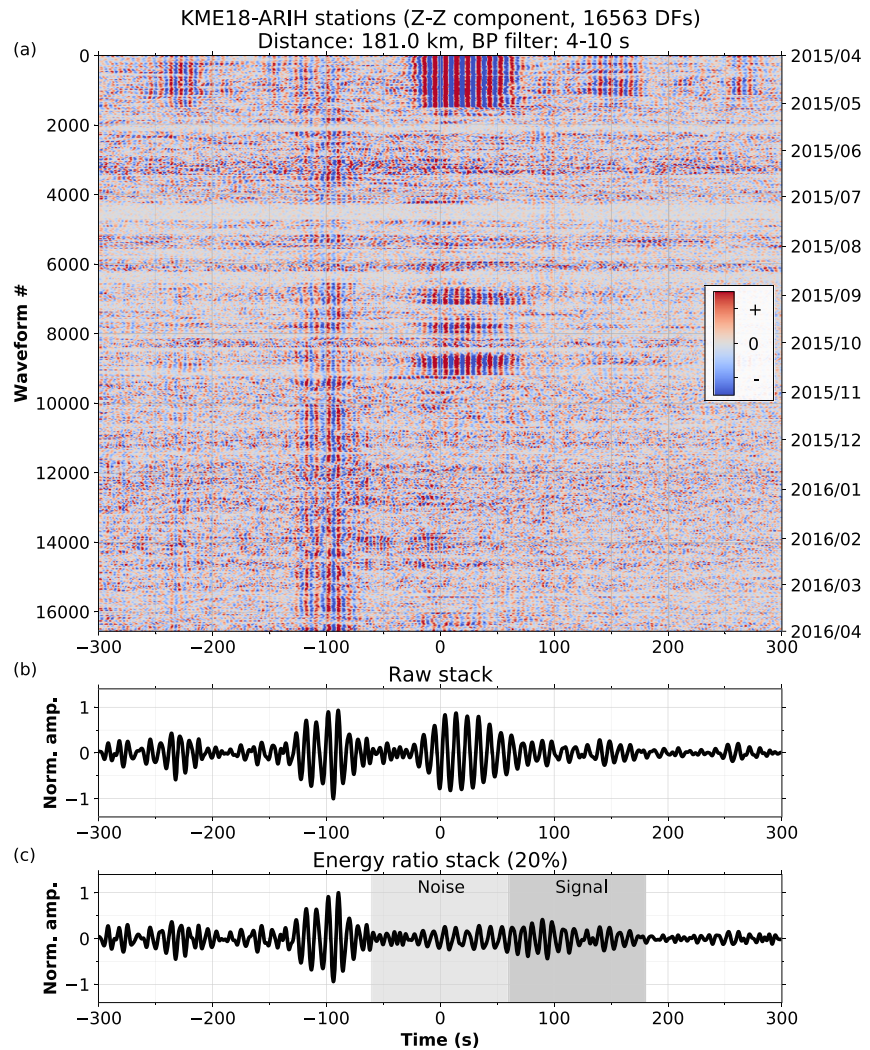
Figure 4. (a) The 16,563 thirty-minute deconvolution functions (DFs) calculated between 1 April 2015 and 31 March 2016 between the KME18 and ARIH stations (locations in Figure 3) for the *Z-Z* component. (b) Raw stack over the year of all the 30-min DFs. (c) DF obtained by stacking 20% of the 30-min DFs with the highest energy ratios (R_E). The light gray and dark gray areas represent the time windows over which the energy of the noise (E_{noise}) and that of the signal (E_{signal}) are calculated, respectively. The amplitude of the waveforms in Figure 4b and 4c is normalized by the highest absolute value. All the waveforms are band-pass filtered between 4 and 10 s.

R and *T* are the radial and transverse directions from the virtual source, respectively. In the following, we assume that Love waves are retrieved on the *T-T* DFs and that Rayleigh waves are captured by the *Z-Z* and *R-R* DFs. Finally, all the waveforms are band-pass filtered between 4 and 10 s using a two-pass four-pole Butterworth filter.

We show an example of the DFs calculated every 30 min between the KME18 and ARIH stations for the *Z-Z* component in Figure 4a. For this station pair, we obtain a total number of 16,563 waveforms, which is less than the total number of waveforms over 366 days (e.g, 17,568 waveforms) as some time windows are removed during the preprocessing step. In Figure 4b, we show the raw stack of the offshore-onshore 30-min DFs over the year. Assuming a theoretical Rayleigh wave velocity of 3.0 km/s, the first physical signals should arrive after 60 s given the 181-km interstation distance. Therefore, the clear arrivals in the anticausal part between −60 and −140 s are likely Rayleigh waves propagating between the two stations. However, strong near-zero time lag arrivals can also be observed and probably overlap with Rayleigh waves arrivals in the causal part of the DF. Such nonphysical arrivals are likely generated by ambient seismic field sources located between the two stations, as the interstation path is mainly under the ocean (station locations in
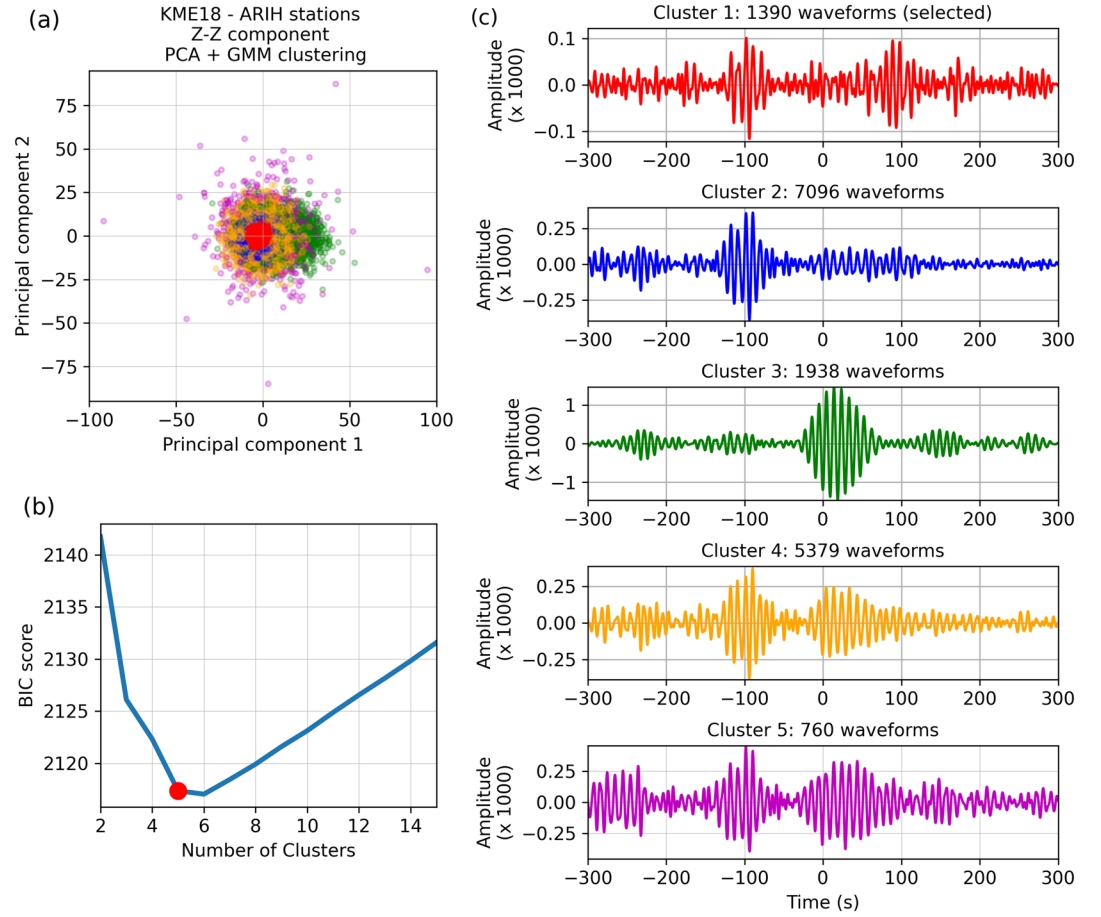
Figure 5. (a) Projection of the 16,563 *Z-Z* DFs between the KME18 and ARIH stations on the first two principal components of the PCA. The five colors correspond to the clusters obtained with the GMM. (b) BIC score for GMMs performed on the output of the first 20 PCs with different numbers of clusters. The knee method determines that five clusters is optimal (red dot). (c) Stack of the 30-min DFs belonging to each cluster shown in (a). The colors of the waveforms correspond to the colors in (a). Our method automatically selects the *Z-Z* DF from Cluster 1 for this station pair.

Figure 3). Finally, we can observe in Figure 4a that the amplitude of the spurious arrivals varies through the year. For this station pair, the spurious arrivals are especially strong for the DFs calculated from the data recorded between April and mid-May 2015 and between September and November 2015.

3.2. Energy Ratio Stack Method

In the following, we also compare the clustering results to a more sophisticated stack of the DFs based on the ratio of the energy of a signal window over that of a noise window. We refer to this technique as the energy ratio stack method. For each 30-min DF, the energy of the signal and noise windows is computed as

$$E_{\text{signal}} = \sum_{t=t_s}^{t_s \times 3} \text{DF}(t)^2,$$ (2)

and

$$E_{\text{noise}} = \sum_{t=-t_s}^{t_s} \text{DF}(t)^2,$$ (3)

where t_s is the time of seismic waves arriving to the onshore receiver with a velocity of 3.0 km/s. E_{noise} is computed from $-t_s$ to t_s, which covers both the anticausal and causal parts of the DFs. E_{signal} is calculated for the causal part of the signal from t_s to $t_s \times 3$ so that the signal window duration is the same as for E_{noise}.

Note that we only focus on the causal part of the DFs, which is later used to simulate earthquake waveforms. For the KME18 and ARIH station pair, t_s is equal to 60.3 s and the two time periods are illustrated in Figure 4c.

The ratio of the two quantities is finally computed as $R_E = E_{\text{signal}}/E_{\text{noise}}$, and we linearly stack 20% of the 30-min DFs with the highest R_E values. Our choice of keeping 20% of the DFs is further discussed in supporting information Text S2 and Figure S1, where we compare the results for other thresholds (e.g., stack of the DFs with the highest 5, 15, 20, and 25% R_E values). In Figure 4c, we show the waveform obtained with the energy ratio stack method with a threshold of 20% between the KME18 and ARIH stations. The energy ratio stack method allows us to significantly reduce the amplitude of the spurious arrivals. However, the Rayleigh wave in the causal part does not appear as clearly as in the anticausal part for this station pair.

3.3. Clustering Offshore-Onshore DFs

We apply the two-step clustering method to the 16,563 Z-Z DFs between the KME18 and ARIH stations and show the results in Figure 5. We first standardize the data set, compute its PCA, and keep the output of the first 20 PCs which explain 36.9% of the cumulative data variance. Similarly to the synthetic data, we cluster the output of the first 20 PCs using GMMs with 2 to 15 clusters and compute the BIC score of each model (Figure 5b). We then use the knee method to determine that the optimal number of clusters is five. Contrary to the synthetic example, the BIC value for five clusters is slightly higher than the minimum BIC value, which is found for six clusters. The projection of the data on the first two PCs is shown in Figure 5a together with the clustering results. We only present the projection of the data on the first two PCs as visualizing the data over 20 dimensions is impossible. Unlike the synthetic case, no clear clusters can be observed in the plot of the first two PCs nor in any other PC combinations. Nevertheless, the waveforms obtained from the stack of the 30-min DFs belonging to each cluster have different characteristics (Figure 5c). This can be explained by the fact that a GMM is, fundamentally, an algorithm for density estimation. Therefore, the fact that the data point density changes in the first 2-PC space in Figure 5a, and in the first 20-PC space in general, explains why a GMM can cluster our DF data set, even if there are no well-separated groups of data.

In Figure 5c, the stacked DFs from the fourth and fifth clusters are very similar to that from the raw stack over the year and contain strong spurious arrivals as well as clear Rayleigh waves in their anticausal parts. The DF from Cluster 5, however, contains stronger spurious arrivals compared to that from Cluster 4. The waveform from the third cluster only contains spurious arrivals. The stacked DF from the second cluster contains clear Rayleigh wave arrivals in its anticausal part and weak spurious arrivals but does not contain clear Rayleigh waves in its causal part. Finally, the waveform from the first cluster, which is obtained by stacking 1,390 thirty-minute DFs, contains clear anticausal and causal Rayleigh waves with almost similar amplitudes and only weak spurious arrivals. As our goal is to improve the retrieval of offshore-onshore DFs, we select the waveform from Cluster 1 for the Z-Z component between the KME18 and ARIH stations.

To automate the selection of the best stacked DF, we use the fact that the data of the corresponding cluster lay near the origin of the first two-PC plot and have the lowest variance. This property is consistent for all the stations and all components of the Green's tensor and can be observed in Figure 5a for the data from the first cluster. Therefore, we simply compute the variance of the data from each cluster on the first two PCs and automatically select the stacked DF from the cluster with the lowest variance in the following.

We show the DFs between the KME18 station and 78 onshore Hi-net stations for the T-T, R-R, and Z-Z components calculated with the raw stack over the year in Figure 6a, with the energy ratio stack method in Figure 6b, and with the two-step clustering method in Figure 6c. In the 4- to 10-s period range, spurious arrivals can be observed for the three components of the raw stack DFs and are especially strong for the Z-Z component (Figure 6a). The energy ratio stack and two-step clustering methods significantly reduce spurious arrivals of the DFs for the three components.

To quantify the effect of our method on the retrieval of clear DFs, we compute SNR values for each component (e.g., T-T, R-R, and Z-Z) and each station pair. The SNR values are computed as the ratio of the root-mean-square (RMS) of a signal window over that of a noise window. The noise window is defined as the part of the DF between −25 to 25 s and the signal window starts at the time of seismic waves arriving with a velocity of 3.0 km/s and ends after 50 s for the causal part of the DFs. The two time periods are highlighted in
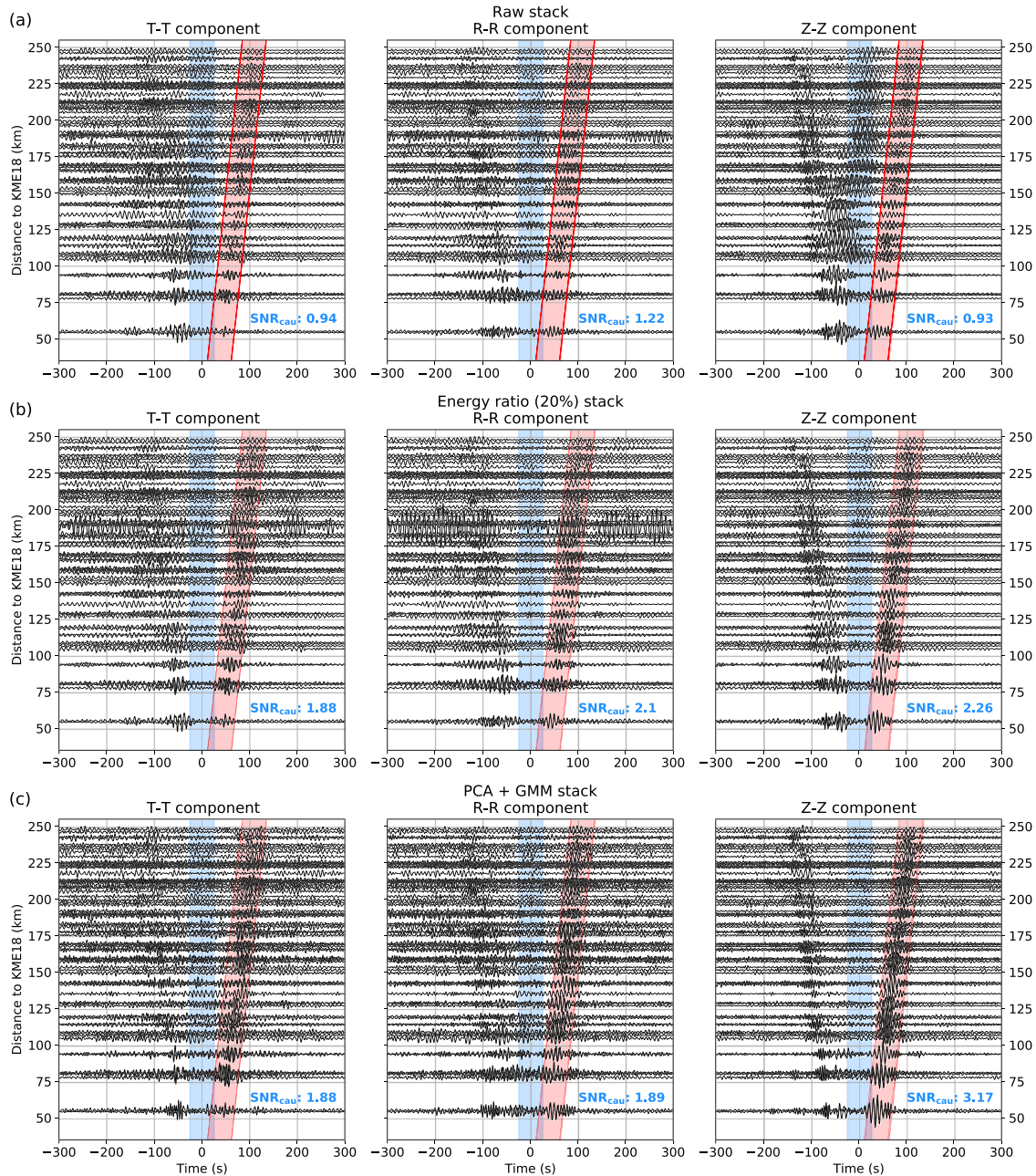
Figure 6. (a) Moveout of the raw stack of the deconvolution functions for the *T-T*, *R-R*, and *Z-Z* components. (b) and (c) are similar to (a) for the DFs obtained with the energy ratio stack method and with the two-step clustering method using the first 20 PCs. For each panel, the SNR_{cau} value calculated using the RMS ratio of the signal window (red area) over the noise window (blue area) is also indicated. The SNR_{cau} values are obtained by averaging the SNR of each DF over the 76 stations located at more than 75 km from the KME18 station. All the waveforms are band-pass filtered between 4 and 10 s.

Figure 6. For each component, we then average all the SNR values over the 76 Hi-net stations located at more than 75 km from the KME18 station to obtain the SNR_{cau} values shown in each panel of Figure 6. We exclude two stations located at less than 75 km from the virtual source as the first physical wave arrivals are likely to arrive during the first 25 s of the DFs. For the three components, the SNR_{cau} values of the DFs calculated with our method and with the energy ratio stack method are higher than that obtained with the raw stack over the year. For the *Z-Z* component, which contains strong spurious arrivals in the raw stack over the year, the highest SNR_{cau} value is obtained with the clustering method. For the two horizontal components, which contain less spurious arrivals, the energy ratio stack method
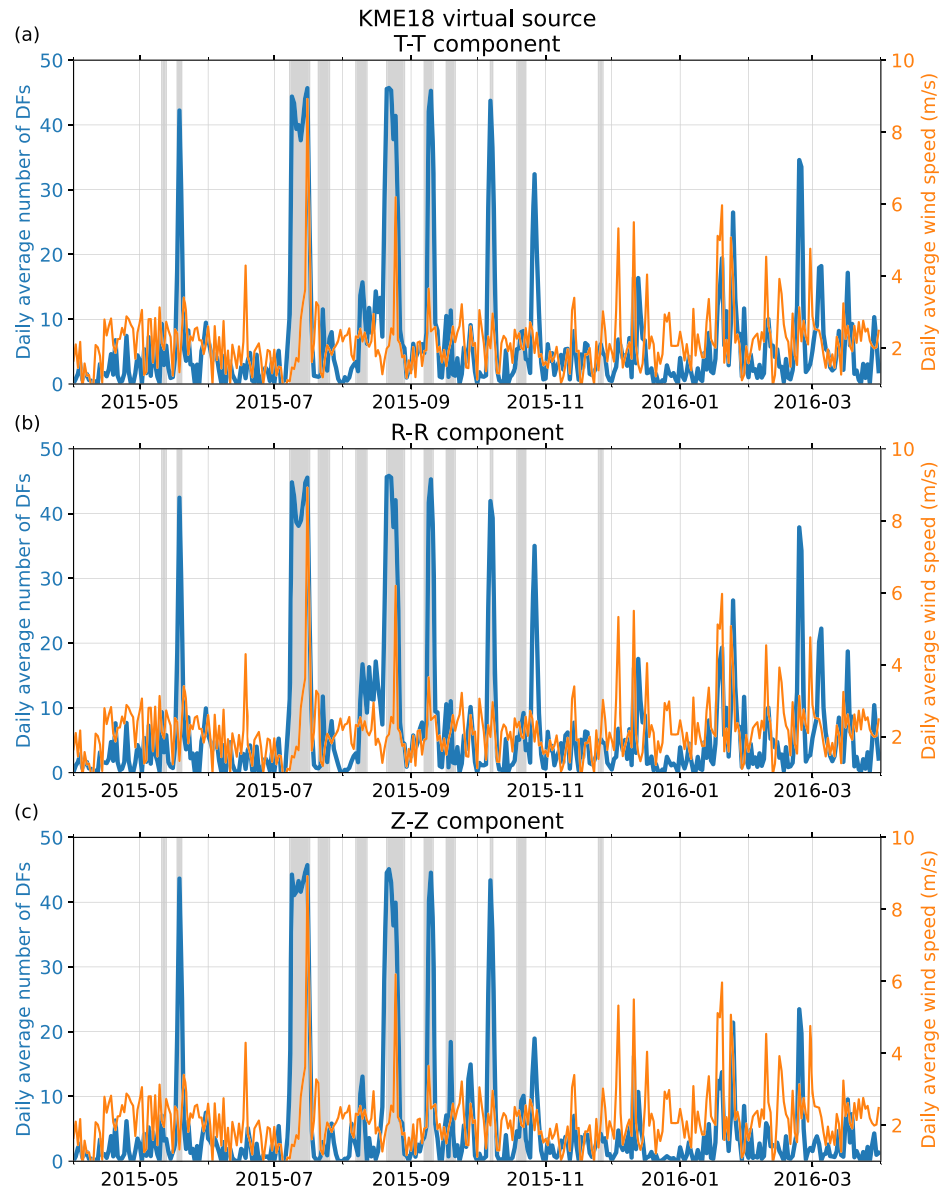
Figure 7. (a) Daily average number of *T-T* DFs obtained with the two-step clustering method between the KME18 and 78 Hi-net stations (blue). The daily average wind speed at the Owase weather station, located close to the OWSH station (location in Figure 3), is shown in orange. The gray areas represent time periods where tropical storms or typhoons occurred or passed within a 1,500-km radius around the KME18 station. (b) and (c) are same as (a) for the *R-R* and *Z-Z* DFs, respectively.

allows us to retrieve higher and equal SNR_{cau} values for the *R-R* and *T-T* components compared to the clustering method. The SNR_{cau} values suggest that our method is especially powerful for correlation data set with very strong spurious arrivals (e.g., *Z-Z* component), which might overlap with the arrival of physical waves. Moreover, the clustering method allows us to significantly reduce the strong amplitudes of the *R-R* and *T-T* DFs that can be observed in Figure 6b for one station located at approximately 190 km from the virtual source.

To investigate the mechanisms involved in the retrieval of clear DFs with the two-step clustering method between the KME18 virtual source and the 78 onshore Hi-net stations, we show the daily number of selected DFs averaged over all the receiver stations between 1 April 2015 and 31 March 2016 in Figure 7. While the

data of each component are processed independently, time periods with high numbers of selected DFs are consistent through the year. Between May and November 2015, there are five distinct time periods where the two-step clustering method selects more than forty 30-min DFs per day for the three components. Extreme meteorological events, such as storms, typhoons, and cyclones, are well known to efficiently excite the ambient seismic field and favor the retrieval of correlation functions (Nishida, 2017, and references therein). To quantify the effect of powerful storms on the selection of DFs, we show the time periods where tropical storms (wind speed between 89 and 117 km/hr), typhoons (wind speed ≥ 118 km/hr), and extratropical cyclones passed or occurred within a 1,500-km radius around the virtual source. In 2015, the five time periods where the clustering method selects more than forty 30-min DFs correlate with the occurrence of nearby extreme meteorological events. For the end of 2015 and 2016, the 30-min DFs selected by the two-step clustering method cannot be explained by typhoons, which generally occur during summer months. To demonstrate that local storms can also contribute to the selection of DFs, we also show the daily average wind speed at the Owase weather station, which is located near the OWSH seismic station, in Figure 7. During winter months, strong winds recorded at the weather station, which can be interpreted as the passing local storms, generally correlate with an increase of selected DFs. Finally, strong winds are not always recorded during the highlighted storm time periods in Figure 7 as some typhoons do not make landfall.

Local storms and typhoons seem to efficiently excite the ambient seismic field and favor the retrieval of clear offshore-onshore DFs near the Nankai Trough. However, all the time periods with a higher rate of selected DFs cannot be explained by local storms and typhoons. In the region of interest, Takagi et al. (2018) showed that Rayleigh waves in the 4- to 8-s period range are mainly generated in the Japan Sea (also known as East Sea) during winter months. Therefore, the occurrence of storms in the Japan Sea, which might not be visible on the wind speed data recorded at the Owase weather station, could potentially be the cause of the time periods over which a relatively large number of DFs is selected. However, additional work, which is outside the scope of this study, is required to fully understand the mechanisms contributing to the retrieval of clear offshore-onshore DFs.

4. M_w 5.8 Earthquake Simulation

4.1. Earthquake Data

A M_w 5.8 earthquake occurred on 1 April 2016 at 11:39:07 Japan Standard Time in the vicinity of the KME18 station (Figure 3). The F-net solution of the National Research Institute for Earth Science and Disaster Resilience (NIED) locates the earthquake at a depth of 12 km, which is close to the plate interface, with a subduction dominant focal mechanism. The occurrence of the earthquake on the plate interface was later confirmed by further studies (Nakano et al., 2018; Takemura et al., 2018; Wallace et al., 2016). We correct the earthquake velocity records at the 78 Hi-net stations for their instrument responses and rotate the horizontal waveforms to the radial and transverse directions from the epicenter. The three-component velocity data are then band-pass filtered between 4 and 10 s using a four-pole and two-pass Butterworth filter and are downsampled from 100 to 10 Hz.

4.2. Simulating Velocity Waveforms With DFs

To demonstrate that the DFs obtained with our two-step clustering method have reliable phases and amplitudes, we use them to simulate the velocity waveforms from the M_w 5.8 earthquake, which can be considered as a point source for the period range of interest (e.g., 4–10 s). Similarly to Viens and Denolle (2019), we consider the causal part of the *T-T*, *R-R*, and *Z-Z* DFs in the following as it likely better captures the site amplification and attenuation effects compared to the anticausal part (Bowden et al., 2015; Liu et al., 2016). The causal *T-T* DFs are considered to simulate Love waves (e.g., *T* component from the earthquake) and the causal *R-R* and *Z-Z* DFs are used to simulate Rayleigh waves, which are the *R* and *Z* components from the earthquake.

We first take the time derivative of the causal DFs to retrieve the proportionality between the correlation function and the Green's function (Sabra et al., 2005). We then resample the DFs from 4 to 10 Hz and convolve them with a source time function to simulate velocity waveforms. The source time function is a Gaussian function with a duration of 1 s, and its amplitude is set so its integral over its duration is equal to 4.9×10^{17} Nm, which is the seismic moment of the earthquake determined by the F-net NIED solution.
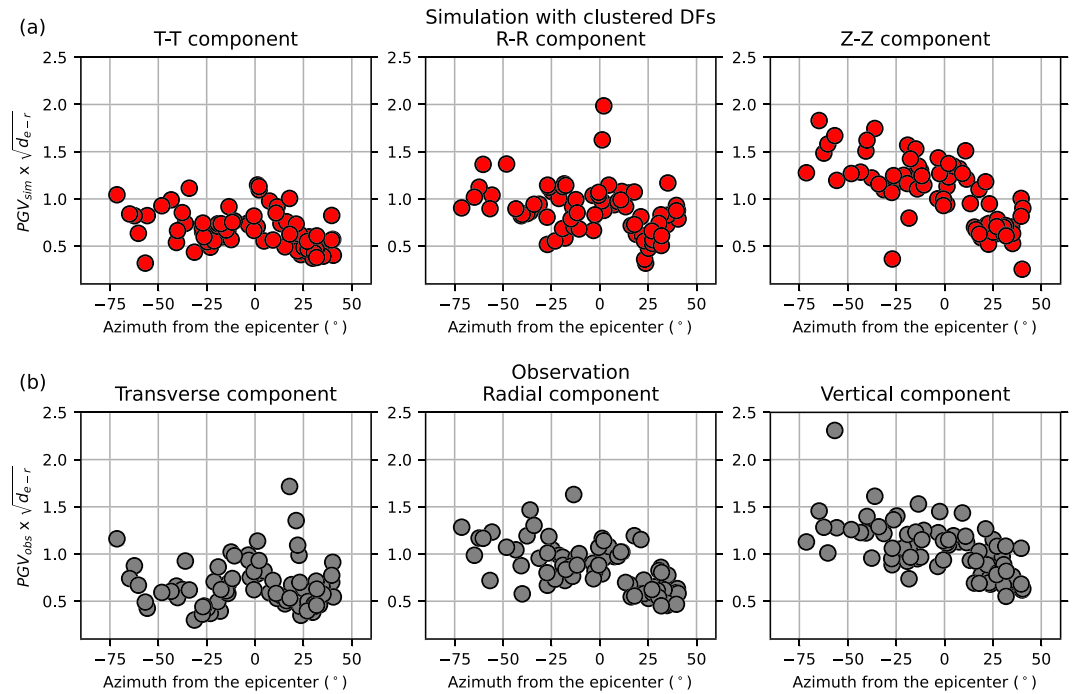
Figure 8. (a) Surface wave long-period peak ground velocities of the simulated waveforms (PGV_{sim}) using the *T-T*, *R-R*, and *Z-Z* DFs obtained with the PCA and GMM clustering method (clustered simulations), as a function of the azimuth from the epicenter. The PGVs are corrected for the surface wave geometrical spreading between the epicenter and receiver locations (e.g., multiplication by $\sqrt{d_{e-r}}$). (b) Long-period peak ground velocities of the 2016 M_{w} 5.8 earthquake (PGV_{obs}) for the transverse, radial, and vertical components after surface wave geometrical spreading correction, as a function of the azimuth from the epicenter. The PGV values are computed for seismic waves traveling slower than 3.0 km/s to focus on surface wave amplitudes and the zero azimuth is north.

To account for the fact that the earthquake epicenter is located 7 km away from the KME18 station, we multiply the amplitude of the simulated waveform by the difference of surface wave geometrical spreading (e.g., multiplication by $\sqrt{d_{v-r}}/\sqrt{d_{e-r}}$, with d_{v-r} and d_{e-r} being the KME18-receiver and epicenter-receiver distances, respectively). We also time shift the simulated waveforms considering a constant local surface wave velocity of 3.0 km/s, assuming that surface wave dispersion is negligible in the 4- to 10-s period range. Finally, as only the relative, rather than absolute, amplitude is preserved by the DFs, the simulated velocity waveforms need to be calibrated with the velocity waveforms from the earthquake. We compute a calibration factor common to the 78 Hi-net stations but different for each component by taking average of the ratio of the simulated over recorded surface wave long-period peak ground velocity (PGV) over all stations. The surface wave long-period PGV of each waveform is defined as the maximum absolute amplitude of the waves traveling slower than 3.0 km/s.

Surface wave radiation pattern effects should also be taken into account when simulating earthquakes with DFs (Denolle et al., 2013). However, Viens and Denolle (2019) showed that for a M_{w} 5.5 event, which occurred near the trench in the Tonankai region, the surface wave propagation effect through the accretionary wedge is stronger than radiation pattern effects. To demonstrate that a similar effect can be observed for the M_{w} 5.8 earthquake, we correct the simulated and observed velocity waveforms for the surface wave geometrical spreading effect by multiplying the waveforms by $\sqrt{d_{e-r}}$. The observed and simulated surface wave long-period PGV after geometrical spreading correction are shown in Figure 8 as a function of the azimuth from the epicenter. Similar long-period PGV variations with the azimuth can be observed for the three components. For the observed radial (*R*) and vertical (*Z*) components and the simulated waveforms with *R-R* and *Z-Z* components, the amplitude of the long-period PGVs decreases with increasing azimuth. For the recorded transverse (*T*) and simulated with *T-T* components, a peak of maximum PGV values can be observed around the zero azimuth and minimum values are located near the −30° and 30° azimuths from
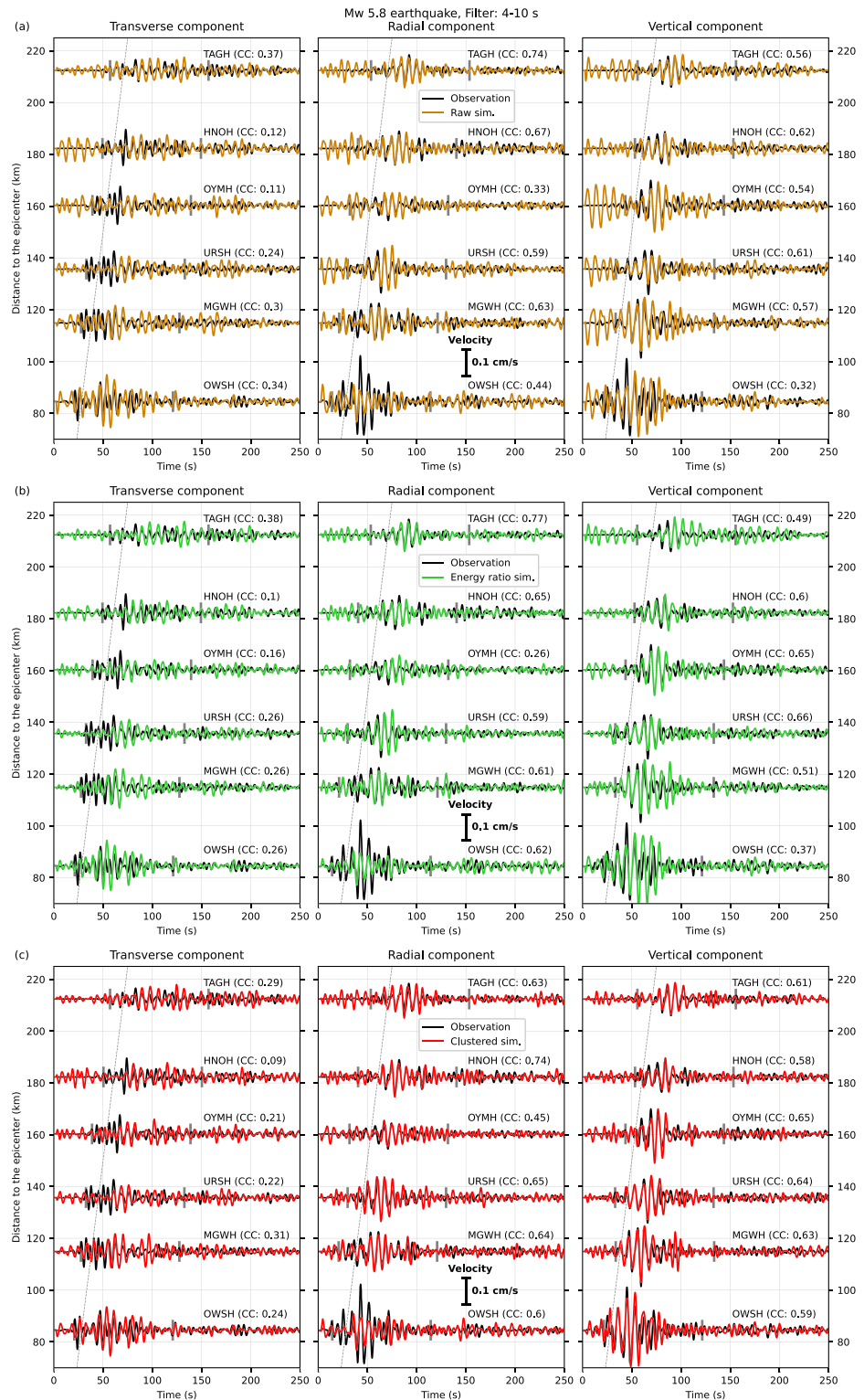
Figure 9. (a) Comparison between raw simulations (orange traces) and recorded (black traces) velocity waveforms for the M_{w} 5.8 earthquake for the transverse, radial, and vertical components in the 4- to 10-s period range. (b) Same as (a) for the simulations using the DFs obtained with the energy ratio stack method (green traces). (c) Same as (a) for the clustered simulations (red traces). The location of the stations is shown in Figure 3. For each station, the correlation coefficient (CC) between the simulated and observed waveforms is calculated between the two vertical gray lines and its value indicated between parenthesis. The dashed lines represent the 3.0-km/s moveout.
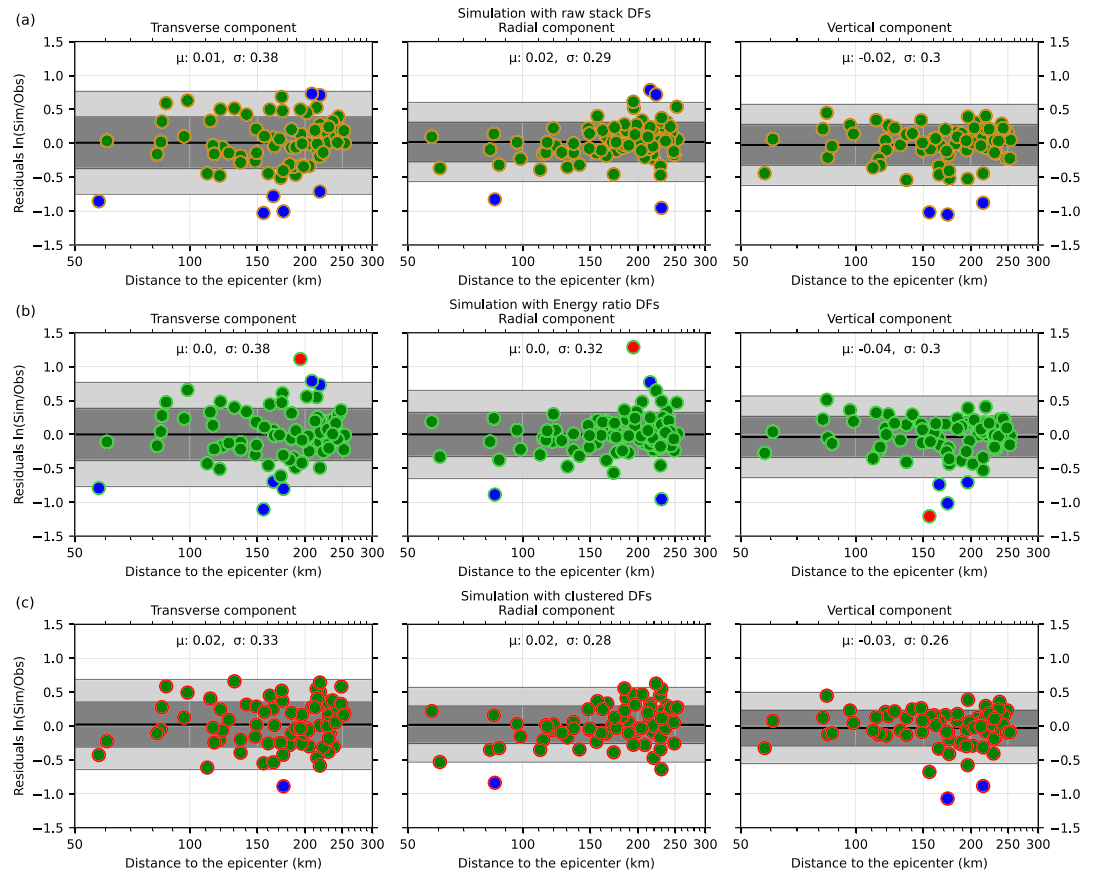
Figure 10. Surface wave long-period PGV residuals for the transverse, radial, and vertical components as a function of the distance to the epicenter of the M_w 5.8 event. The raw, energy ratio, and clustered simulations are used in Figures 10a–10c, respectively. Green filled circles indicate that the ratio between the simulated and observed PGVs is within a factor of 2, and blue filled circles show ratios larger than a factor of 2 but within a factor of 3. The red filled circles represent a ratio larger than a factor of 3. The thick black line represents the mean of the data, and the 1 and 2 standard deviations to the mean are shown by the dark gray and light gray areas, respectively. The mean of the PGV residuals (μ) and the 1 standard deviation to the mean (σ) value are shown on top of each panel.

the north. As the simulated waveforms only contain the signature of the seismic wave propagation between the KME18 station and onshore stations, similar azimuthal variations as the earthquake suggest that propagation effects have a dominant effect on the amplitude of surface waves and are stronger than radiation pattern effects for the M_w 5.8 earthquake in the 4- to 10-s period range.

Therefore, we simply consider the causal *T-T*, *R-R*, and *Z-Z* DFs convolved with the source time function and corrected for the fact that the KME18 station and the epicenter are not colocated to simulate the transverse, radial, and vertical earthquake waveforms. In the following, we consider three types of simulated waveforms. The first type of simulations uses the raw stack of the DFs over the year and are called raw simulations. The second and third types of simulated waveforms use the DFs obtained with energy ratio stack method and the two-step clustering method and are called energy ratio and clustered simulations, respectively.

4.3. Simulation Results

In Figure 9, we show the simulated and observed velocity waveforms for the transverse, radial, and vertical components at six Hi-net stations (locations in Figure 3). The raw, energy ratio, and clustered simulations are shown in Figures 9a–9c, respectively. For the three components, the main wave packet travels with a velocity of approximately 3.0 km/s and is relatively well retrieved by the three simulation methods. However, the energy ratio and clustered simulations have less spurious arrivals, which allows us to better identify the true wave arrivals.
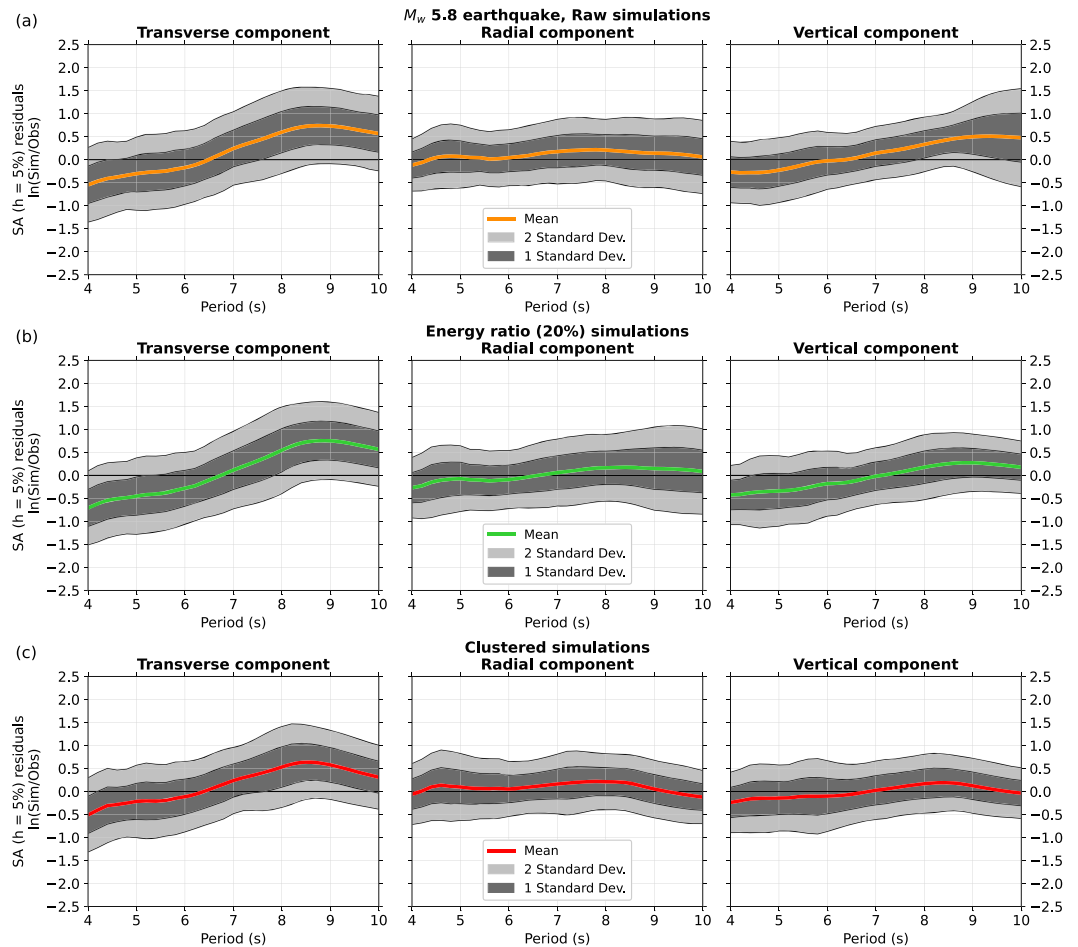
Figure 11. (a) Five-percent damped spectral acceleration (SA) residuals computed between the raw simulations and the observed waveforms of the $M_{\rm w}$ 5.8 earthquake over the 78 Hi-net stations for the transverse, radial, and vertical components. (b) Same as (a) for the energy ratio simulations. (c) Same as (a) for the clustered simulations. For each panel, the mean of the SA residuals (colored line) is shown together with the 1 (dark gray area) and 2 (light gray area) standard deviations to the mean. The zero bias is highlighted by black straight lines.

To quantitatively compare the phase of the observed and simulated waveforms, we compute a correlation coefficient (CC) for each waveform pair. The CC is calculated over 100 s from the time 1% of the earthquake cumulative energy is reached (vertical gray ticks in Figure 9). This metric varies between −1, when the two waveforms are out of phase by 180°, and 1 when the two waveforms are identical. We also allow a 1-s phase shift when calculating CCs to account for potential earthquake location errors. For the six stations shown in Figure 9, the clustered simulations generally reproduce better the phase of the earthquake waveforms than the raw and energy ratio simulations, as shown by generally higher CCs. Over the 234 waveforms (e.g., 78 receiver stations and 3 components), 149 CCs (e.g., 64%) calculated between the observed waveforms and clustered simulations are higher or equal than if the raw simulations are used. For the 85 smaller CCs, 65 of them are smaller by less than a value of 0.1, which indicates that the raw and clustered simulations are very similar (e.g., *T* and *R* components of the TAGH station in Figure 9a and 9c). The energy ratio and raw simulations have an equal number of higher CCs (e.g, 117 waveforms each). Finally, the clustered simulations allows us to obtain higher CCs for 145 waveforms (e.g., 62%) compared to the energy ratio simulations.

To quantify the difference between the observed and simulated waveform amplitudes, we use the surface wave long-period PGVs and compute their residuals as the natural logarithm of the simulated over observed PGV ratios. The residuals are shown in Figure 10 as a function of the distance to the earthquake epicenter. For the three types of simulated waveforms, the mean of the residuals is close to the zero bias. Moreover, there is no clear variation of the residual distribution with the distance to the epicenter in

Figure 10. This indicates that the attenuation of the waves with distance is relatively well preserved by the raw, energy ratio, and clustered simulations. However, the PGVs from the clustered simulations (Figure 10c) reproduce better the observed PGVs as shown by smaller standard deviations to the mean for the three components compared to that shown in Figures 10a and 10b for the PGVs from the raw and energy ratio simulations. Over the 234 PGV ratios calculated between the clustered simulations and observed waveforms, 230 ratios are smaller than a factor of 2 and 4 ratios are larger than a factor of 2 but smaller than a factor of 3 (green and blue circles in Figure 10c). For the raw simulations, 220 PGV ratios are within a factor of 2 and 14 ratios are larger than a factor of 2 but smaller than a factor of 3. For the energy ratio simulations, 219 PGV ratios are within a factor of 2, 3 are larger than a factor of 3, and the other 12 ratios are larger than a factor of 2 but smaller than a factor of 3.

We finally compute 5% damped spectral accelerations (SAs) for the observed and simulated velocity waveforms. We first take the time derivative of the velocity waveforms to retrieve the corresponding acceleration time series and compute the SA using the Duhamel's integral technique (Chopra, 2015). We then calculate the SA residuals as the natural logarithm of the simulated over observed SAs for each period. Finally, we calculate the mean of the residuals over the 78 stations as well as the 1 and 2 standard deviations to the mean for each period and show the results in Figure 11. For the radial and vertical components, the clustered simulations perform better than the raw and energy ratio simulations as the mean of the SA residuals is closer to the zero bias and the standard deviation values are smaller. Moreover, the zero bias line in Figure 11c is always within 1 standard deviation, which is not the case for the vertical component in Figures 11a and 11b. For the transverse component, the SA residuals are not as good as for the radial and vertical components for the three simulation methods. Nevertheless, the clustered simulations perform better than the raw and energy ratio simulations to reproduce the earthquake SA. The variations observed for the transverse component can potentially be caused by the fact that Love waves are not as well retrieved as Rayleigh waves in the offshore-onshore setting. This can be observed in Figure 6 where the T-T DFs have the lowest SNR$_{cau}$ values among the three components for the three methods.

5. Conclusions

We introduced a method based on unsupervised learning to improve the retrieval of offshore-onshore correlation functions calculated with the deconvolution technique (DF). Our method works in two steps: first, the dimension of a DF data set calculated between two seismic stations is reduced using the PCA; and second, the data from the low-dimensional space are clustered with a GMM. The waveforms belonging to each cluster are finally stacked together and the clustered DF that improves the symmetry between the anticausal and causal parts and removes spurious arrivals is selected.

We applied our method to DFs calculated between the offshore KME18 station and 78 onshore Hi-net stations in Japan. The selected DFs obtained with the clustering method have higher signal-to-noise ratios than that obtained with the raw stack of the DFs over the year. Moreover, our method is especially efficient for selecting and stacking DFs with strong spurious arrivals (e.g., Z-Z DFs), as we obtained higher SNR values than with the energy ratio stack method.

To demonstrate that the DFs calculated with the clustering method contain reliable phases and amplitudes, we transformed the DFs into velocity waveforms and compared them to the recorded waveforms of a M_w 5.8 earthquake, which occurred close to the virtual source. The simulated waveforms obtained with the clustered DFs reproduced the earthquake waveforms better than the raw and energy ratio simulations.

Our two-step clustering method offers a new way to easily improve the quality of correlation functions with strong spurious arrivals without requiring any high performance computing. By improving the retrieval of reliable DFs between offshore and onshore stations, we hope to improve the prediction of long-period ground motions from potential future megathrust earthquakes that could occur along subduction zones worldwide, such as the Nankai Trough or the Cascadia subduction zone.

Data Availability Statement

The data from the NIED Hi-net (https://doi.org/10.17598/NIED.0003) and DONET 1 (https://doi.org/10.17598/NIED.0008) networks can be downloaded from the website (https://www.hinet.bosai.go.jp). The

clustering algorithm developed in this study together with the codes to reproduce Figures 1, 2, 4, and 5 are available online (at https://doi.org/10.5281/zenodo.3873790). Most figures were made using the Matplotlib library (Hunter, 2007).

References

Bensen, G. D., Ritzwoller, M. H., Barmin, M. P., Levshin, A. L., Lin, F., Moschetti, M. P., et al. (2007). Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements. *Geophysical Journal International*, *169*, 1239–1260. https://doi.org/10.1111/j.1365-246X.2007.03374.x

Bowden, D. C., Tsai, V. C., & Lin, F. C. (2015). Site amplification, attenuation, and scattering from noise correlation amplitudes across a dense array in Long Beach, CA. *Geophysical Research Letters*, *42*, 1360–1367. https://doi.org/10.1002/2014GL062662

Brenguier, F., Campillo, M., Hadziioannou, C., Shapiro, N. M., Nadeau, R. M., & Larose, E. (2008). Postseismic relaxation along the San Andreas fault at Parkfield from continuous seismological observations. *Science*, *321*, 1478–1481. https://doi.org/10.1126/science.1160943

Brenguier, F., Shapiro, N. M., Campillo, M., Ferrazzini, V., Duputel, Z., Coutant, O., & Nercessian, A. (2008). Towards forecasting volcanic eruptions using seismic noise. *Nature Geoscience*, *1*, 126–130. https://doi.org/10.1038/ngeo104

Chaput, J., Clerc, V., Campillo, M., Roux, P., & Knox, H. (2015). On the practical convergence of coda-based correlations: A window optimization approach. *Geophysical Journal International*, *204*, 736–747. https://doi.org/10.1093/gji/ggv476

Chopra, A. K. (2015). *Dynamics of structures: Theory and applications to earthquake engineering* (4th ed.). Upper Saddle River, NJ: Pearson Education Limited.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society, Series B*, *39*(1), 1–38.

Denolle, M. A., Boué, P., Hirata, N., & Beroza, G. C. (2018). Strong shaking predicted in Tokyo from an expected M7+ Itoigawa-Shizuoka earthquake. *Journal of Geophysical Research: Solid Earth*, *123*, 3968–3992. https://doi.org/10.1029/2017JB015184

Denolle, M. A., Dunham, E. M., Prieto, G. A., & Beroza, G. C. (2013). Ground motion prediction of realistic earthquake sources using the ambient seismic field. *Journal of Geophysical Research: Solid Earth*, *118*, 2102–2118. https://doi.org/10.1029/2012JB009603

Denolle, M. A., Dunham, E. M., Prieto, G. A., & Beroza, G. C. (2014). Strong ground motion prediction using virtual earthquakes. *Science*, *343*, 399–403. https://doi.org/10.1126/science.1245678

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on machine learning* (p. 29). *ICML '04*. New York, NY, USA: ACM.

Fichtner, A., & Tsai, V. C. (2019). Theoretical foundations of noise interferometry. In N. Nakata, L. Gualtieri, & A. Fichtner (Eds.), *Seismic ambient noise*. Cambridge, UK: Cambridge University Press.

Figueiredo, M. A. T., & Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 381–396. https://doi.org/10.1109/34.990138

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Kwak, S., Song, S. G., Kim, G., Cho, C., & Shin, J. S. (2017). Investigating the capability to extract impulse response functions from ambient seismic noise using a mine collapse event. *Geophysical Research Letters*, *44*, 9653–9662. https://doi.org/10.1002/2017GL075532

Lin, F.-C., Moschetti, M. P., & Ritzwoller, M. H. (2008). Surface wave tomography of the western United States from ambient seismic noise: Rayleigh and Love wave phase velocity maps. *Geophysical Journal International*, *173*, 281–298. https://doi.org/10.1111/j.1365-246X.2008.03720.x

Liu, X., Ben-Zion, Y., & Zigone, D. (2016). Frequency domain analysis of errors in cross-correlations of ambient seismic noise. *Geophysical Journal International*, *207*, 1630–1652. https://doi.org/10.1093/gji/ggw361

Mousavi, S. M., Zhu, W., Ellsworth, W., & Beroza, G. (2019). Unsupervised clustering of seismic signals using deep convolutional auto-encoders. *IEEE Geoscience and Remote Sensing Letters*, *16*(11), 1693–1697. https://doi.org/10.1109/LGRS.2019.2909218

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Nakano, M., Hyodo, M., Nakanishi, A., Yamashita, M., Hori, T., Kamiya, S., et al. (2018). The 2016 Mw 5.9 earthquake off the southeastern coast of Mie Prefecture as an indicator of preparatory processes of the next Nankai Trough megathrust earthquake. *Progress in Earth and Planetary Science*, *5*, 30. https://doi.org/10.1186/s40645-018-0188-3

Nakano, M., Tonegawa, T., & Y., K. (2012). Orientations of donet seismometers estimated from seismic waveforms. *JAMSTEC Report of Research and Development*, *15*, 77–89.

Nishida, K. (2017). Ambient seismic wave field. *Proceedings of the Japan Academy. Series B, Physical and biological sciences*, *93*, 423–448. https://doi.org/10.2183/pjab.93.026

Prieto, G. A., & Beroza, G. C. (2008). Earthquake ground motion prediction using the ambient seismic field. *Geophysical Research Letters*, *35*, L14304. https://doi.org/10.1029/2008GL034428

Retailleau, L., Boué, P., Stehly, L., & Campillo, M. (2017). Locating microseism sources using spurious arrivals in intercontinental noise correlations. *Journal of Geophysical Research: Solid Earth*, *122*, 8107–8120. https://doi.org/10.1002/2017JB014593

Sabra, K. G., Gerstoft, P., Roux, P., Kuperman, W. A., & Fehler, M. C. (2005). Extracting time-domain Green's function estimates from ambient seismic noise. *Geophysical Research Letters*, *32*, L03310. https://doi.org/10.1029/2004GL021862

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *Proceedings of the 2011 31st international conference on distributed computing systems workshops, Minneapolis, Minnesota, USA* (pp. 166–171). https://doi.org/10.1109/ICDCSW.2011.20

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Shapiro, N. M., Campillo, M., Stehly, L., & Ritzwoller, M. H. (2005). High-resolution surface-wave tomography from ambient seismic noise. *Science*, *307*, 1615–1618. https://doi.org/10.1126/science.1108339

Shapiro, N. M., Ritzwoller, M. H., & Bensen, G. D. (2006). Source location of the 26 sec microseism from cross-correlations of ambient seismic noise. *Geophysical Research Letters*, *33*, L18310. https://doi.org/10.1029/2006GL027010

Sheng, Y., Denolle, M. A., & Beroza, G. C. (2017). Multicomponent C3 Green's functions for improved long period ground-motion prediction. *Bulletin of the Seismological Society of America*, *107*, 2836. https://doi.org/10.1785/0120170053

Shiomi, K. (2013). New measurements of sensor orientation at NIED Hi-net stations. *Report NIED*, *80*, 1–20.

Stehly, L., & Boué, P. (2017). On the interpretation of the amplitude decay of noise correlations computed along a line of receivers. *Geophysical Journal International*, *209*, 358. https://doi.org/10.1093/gji/ggx021

Stehly, L., Campillo, M., & Shapiro, N. M. (2006). A study of the seismic noise from its long-range correlation properties. *Journal of Geophysical Research*, *111*, B10306. https://doi.org/10.1029/2005JB004237

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In L. T. Wille (Ed.), *New directions in statistical physics: Econophysics, bioinformatics, and pattern recognition* (pp. 273–309). Berlin, Heidelberg.

Takagi, R., Nishida, K., Maeda, T., & Obara, K. (2018). Ambient seismic noise wavefield in Japan characterized by polarization analysis of Hi-net records. *Geophysical Journal International*, *215*, 1682–1699. https://doi.org/10.1093/gji/ggy334

Takemura, S., Kimura, T., Saito, T., Kubo, H., & Shiomi, K. (2018). Moment tensor inversion of the 2016 southeast offshore Mie earthquake in the Tonankai region using a three-dimensional velocity structure model: Effects of the accretionary prism and subducting oceanic plate. *Earth Planets Space*, *70*, 50. https://doi.org/10.1186/s40623-018-0819-3

Tsai, V. C. (2011). Understanding the amplitudes of noise correlation measurements. *Journal of Geophysical Research*, *116*, B09311. https://doi.org/10.1029/2011JB008483

Viens, L., & Denolle, M. A. (2019). Long-period ground motions from past and virtual megathrust earthquakes along the Nankai Trough, Japan. *Bulletin of the Seismological Society of America*, *109*, 1312–1330. https://doi.org/10.1785/0120180320

Viens, L., Denolle, M., Miyake, H., Sakai, S., & Nakagawa, S. (2017). Retrieving impulse response function amplitudes from the ambient seismic field. *Geophysical Journal International*, *210*, 210. https://doi.org/10.1093/gji/ggx155

Viens, L., Koketsu, K., Miyake, H., Sakai, S., & Nakagawa, S. (2016). Basin-scale Green's functions from the ambient seismic field recorded by MeSO-net stations. *Journal of Geophysical Research: Solid Earth*, *121*, 2507–2520. https://doi.org/10.1002/2016JB012796

Viens, L., Laurendeau, A., Bonilla, L. F., & Shapiro, N. M. (2014). Broad-band acceleration time histories synthesis by coupling low-frequency ambient seismic field and high-frequency stochastic modelling. *Geophysical Journal International*, *199*, 1784–1797. https://doi.org/10.1093/gji/ggu362

Viens, L., Miyake, H., & Koketsu, K. (2015). Long-period ground motion simulation of a subduction earthquake using the offshore-onshore ambient seismic field. *Geophysical Research Letters*, *42*, 5282–5289. https://doi.org/10.1002/2015GL064265

Viens, L., Miyake, H., & Koketsu, K. (2016). Simulations of long-period ground motions from a large earthquake using finite rupture modeling and the ambient seismic field. *Journal of Geophysical Research: Solid Earth*, *121*, 2507–2520. https://doi.org/10.1002/2016JB012796

Wallace, L. M., Araki, E., Saffer, D., Wang, X., Roesner, A., Kopf, A., et al. (2016). Near-field observations of an offshore Mw 6.0 earthquake from an integrated seafloor and subseafloor monitoring network at the Nankai Trough, southwest Japan. *Journal of Geophysical Research: Solid Earth*, *121*, 8338–8351. https://doi.org/10.1002/2016JB013417

Weaver, R. L., & Lobkis, O. I. (2001). Ultrasonics without a source: Thermal fluctuation correlations at MHz frequencies. *Physical Review Letters*, *87*, 134,301. https://doi.org/10.1103/PhysRevLett.87.134301

Weaver, R. L., & Yoritomo, J. Y. (2018). Temporally weighting a time varying noise field to improve green function retrieval. *The Journal of the Acoustical Society of America*, *143*(6), 3706–3719. https://doi.org/10.1121/1.5043406

Webb, S. C. (1998). Broadband seismology and noise under the ocean. *Reviews of Geophysics*, *36*, 105–142. https://doi.org/10.1029/97RG02287

Xie, J., Yang, Y., & Luo, Y. (2020). Improving cross-correlations of ambient noise using a RMS-ratio selection stacking method. *Geophysical Journal International*, *222*, 989–1002. https://doi.org/10.1093/gji/ggaa232

Zhao, Q., Hautamaki, V., & Fränti, P. (2008). Knee point detection in BIC for detecting the number of clusters. In J. Blanc-Talon, S. Bourennane, W. Philips, D. Popescu, & P. Scheunders (Eds.), *Advanced concepts for intelligent vision systems* (pp. 664–673). Berlin, Heidelberg.