Title

Data-Driven Velocity Model Evaluation Using K-Means Clustering

Journal

Geophysical Research Letters, 48(23)

ISSN

0094-8276

Authors

Xiong, Neng
Qiu, Hongrui
Niu, Fenglin

Publication Date

2021-12-16

DOI

10.1029/2021gl096040

Peer reviewed

Data-driven Velocity Model Evaluation using K-means Clustering

Neng Xiong[1], Hongrui Qiu[1,2*], and Fenglin Niu[1]

[1]Department of Earth, Environmental and Planetary Sciences, Rice University, Houston, TX, USA

[2]Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corresponding author: Hongrui Qiu (qiuhonrui@gmail.com; hongruiq@mit.edu)

Key Points:

- We develop a data-driven method that evaluates a velocity model using the K-means clustering and Rayleigh wave phase velocity dispersion
- The model evaluation method is applied to community velocity models, CVM-S4.26 and CVM-H15.1, in Southern California
- The result suggests that CVM-S4.26 gets an evaluation score ~3 times higher than that of CVM-H15.1 for structures in the top ~20 km

Abstract

We develop a data-driven clustering method to evaluate a velocity model using surface wave velocity dispersion. This is done by first computing theoretical dispersion curves for 1-D velocity profiles of all the grid locations and then splitting the resulting dispersion curves into a certain number of groups via the K-means clustering. The observed dispersion curves are also clustered following the same procedure and the velocity model is assessed by comparing the spatial patterns obtained for the observed and synthetic datasets. The method is applied to evaluate two community velocity models in southern California, CVM-S4.26 and CVM-H15.1, using phase velocity maps derived for 3-16s Rayleigh waves. We found a good correlation in the spatial distribution of clusters between the result of CVM-S4.26 and that of the observed data, suggesting that the CVM-S4.26 fits the observed dispersion maps better than the CVM-H15.1 in terms of features extracted from the clustering analysis.

Plain Language Summary

With increasing volume of recorded seismic data, various velocity models are often derived for the same region using different datasets and seismic networks with different spatial coverage and resolution. Therefore, evaluating all the existing velocity models in the overlapping region can provide crucial information to future development of tomographic models, such as constructing a standard model by merging all the velocity models. As a machine learning technique, clustering analysis has proven its ability to extract hidden grouping features from large unlabeled datasets. In this study, we develop a simple workflow that utilizes a specific (K-means) clustering method to evaluate velocity model. Instead of applying the clustering method directly to the velocity model, we first calculate theoretical predictions for a certain measurable parameter (phase velocity of Rayleigh wave) using the input model and assess the model by comparing the clustering results obtained for the synthetic and observed datasets. The proposed model evaluation method is applied to the well-maintained community velocity models, CVM-H15.1 and CVMS-4.26, in Southern California. The result suggests that CVM-S4.26 is much better than CVM-H15.1 for structures in the top ~20 km.

1 Introduction

With the increasing volume of data recorded by regional and global seismic networks, seismic tomography has become an important and powerful tool for understanding earth interior structure in the past decades. Southern California (SC; Fig. 1) is one of the most active and imaged plate boundary regions. Velocity models that cover various depth from near surface to upper mantle and spatial ranges with different resolutions were derived for this area (e.g., Berg et al., 2018; Lee et al., 2014; Lin et al., 2013; Roux et al., 2016). This is done by using different types of datasets, such as surface waves (e.g., Zigone et al., 2015) and teleseismic body waves (e.g., Schmandt and Humphreys, 2010), and inversion schemes, for example, by fitting travel-time (e.g., Fang et al., 2016) and full-waveform (e.g., Tape et al., 2010). Among these velocity models, the community velocity models (CVMs), CVM-H15.1 (Shaw et al., 2015) and CVM-S4.26 (Lee et al., 2014), are well maintained and often used as the starting model in travel-time based tomography studies (e.g., Qiu et al., 2019; Share et al., 2019).

Although CVM-H15.1 and CVM-S4.26 are both constructed through full-waveform inversion, differences between the two models are obvious (e.g., Figs. 2a-b and 2d-e). This inconsistency is related to the choice of input dataset (e.g., frequency range, station coverage, usage of ambient noise data), inversion parameters (e.g., regularization, smoothing), and the starting model. Although some large-scale features (e.g., major geologic provinces; Fig. 1) are seen in both models, it is still challenging to determine which model to use in studies that aim at improving or interpreting the velocity structure in SC. For instance, Qiu et al. (2019) demonstrated that the synthetic dispersion curves calculated using either CVM-H15.1 or CVM-S4.26 match poorly with the observed dispersion maps. However, through the 1-D Vs inversion (Herrmann, 2013) at each grid location, the misfit values are significantly reduced to a level comparable to the estimated uncertainties for both CVMs. This is likely due to the non-uniqueness of the inversion problem, which makes it difficult to evaluate which model is more realistic through the analysis of data misfit.

One way to assess the quality of a velocity model is through forward 3-D waveform simulation based on the wave equation. This is usually done by comparing earthquake recordings or empirical Green's functions retrieved from ambient noise data with synthetic waveforms simulated using the same source-receiver configuration (Ma et al., 2008; Imperatori and Gallovič, 2017). However, the application of such a model validation method is limited by two main factors: (1) complicated evaluation scheme, i.e., any inaccurate information in the velocity model along the ray path can contribute to the mismatch between synthetic and observed waveforms; and (2) intensive computational costs, particularly for observations at high frequencies (e.g., > 1 Hz).

In recent years, machine learning has become more and more popular in extracting hidden features from large datasets in seismology (e.g., Bergen et al., 2019; Kong et al., 2018). Clustering analysis, as an unsupervised learning method, found success in mining different types of noise sources in continuous seismic recordings (Johnson et al., 2020; Snover et al., 2020). The nature of dividing data into groups with similar pattern makes clustering analysis suitable in dealing with large unlabeled datasets, such as seismic waveforms and velocity models. Eymold and Jordan (2019) applied the K-means clustering algorithm to the 1-D velocity profiles of CVM-S4.26 and discovered good correlation between surface geology features in SC and the resulting clustering pattern. However, it is important to note that, by directly clustering the 1-D velocity profiles, the obtained spatial pattern highly depends on the depth range of the input model (e.g., 0-50 km in Eymold and Jordan, 2019). Moreover, clustering results of the same region can also change with different input velocity models, and such difference is often hard to interpret, as the comparison does not involve data fitting to field measurements.

In this study, we propose a data-driven evaluation scheme for velocity models based on the K-means clustering method. This is done by first calculating synthetic surface wave velocity dispersion curves for all 1-D velocity profiles of an input velocity model, and then clustering the synthetic and observed velocity dispersion curves independently into a certain number of groups through clustering analysis. The velocity model is rated by estimating the similarity between spatial patterns obtained from the synthetic and observed dispersion data. The proposed method is applied to two velocity models in SC (CVM-H15.1 and CVM-S4.26). The two velocity models and the Rayleigh wave phase velocity dispersion maps measured by Qiu et al. (2019) that are used to assess the models are described in section 2. The theoretical basis and workflow of the K-means algorithm are reviewed and illustrated in section 3. In section 4, we show the spatial patterns of the clustering analysis for CVM-H15.1 and CVM-S4.26, and the evaluation of each model based on the observed phase velocity maps.

2 Data

The two community velocity models, CVM-H15.1 and CVM-S4.26, analyzed in this study cover the SC plate boundary region (Figs. 2a-b). Both models were extracted using the same grid size (0.05°×0.05°) as the Rayleigh wave phase velocity dispersion maps. Depth of both CVMs were sampled with an interval of 500 m. Except basin regions, the CVM-H15.1 was built upon an initial model derived from a local earthquake tomographic inversion (Shaw et al., 2015). The model in basin areas is derived from more precise studies using borehole measurements and seismic reflection data (Süss and Shaw, 2003; Tabora et al, 2016), and held fixed during the wave-equation-based tomographic inversion. The CVM-H15.1 was derived utilizing data from 143 regional earthquakes recorded by 203 seismic stations (Shaw et al., 2015 and references therein). The CVM-S4.26 was constructed based on a different starting model via a similar tomographic inversion scheme (Lee et al., 2014) but using more earthquakes (160) and seismic stations (258). It is important to note that, in addition to earthquake data, ambient noise cross correlations calculated for pairs of stations are included in the inversion of CVM-S4.26.

The Rayleigh wave phase velocity dispersion maps used to evaluate the CVMs are discretized on a 0.05°×0.05° grid and derived via Eikonal tomography from Qiu et al. (2019). The dispersion maps contain a total of 4076 phase velocity dispersion curves ranging from 3s to 16s. Figure 2c shows the phase velocity map at 7s period. Clear phase velocity contrast can be seen across geologic provinces, such as high velocities in Peninsular Ranges and low velocities in Salton Trough. In order to evaluate the CVMs via clustering analysis, the theoretical phase velocity dispersion curves are also calculated for all 1-D velocity profiles using the CPS package developed by Herrmann (2013). Both the observed and synthetic Rayleigh wave phase velocity dispersion curves are discretized into 17 data points from 3s to 16s.

3 K-means clustering

In this study, we utilize the K-means clustering method to group a series of 1-D curves into a predetermined number of clusters. Let $\boldsymbol{n}$ be the number of the input 1-D curves, and $\boldsymbol{K}$ be the number of clusters. First, $\boldsymbol{K}$ 1-D profiles are randomly chosen from the input dataset as the initial centroids $\{\mu_1,\ \mu_2,\ \mu_3, \dots, \mu_k\}$. The Euclidean distances between each 1-D curve to all centroids are then calculated as the L2 norm between the two vectors:

$$D_k = ||x - \mu_k||_2, \tag{1}$$

where $\boldsymbol{x}$ is the target velocity profile vector, and $\boldsymbol{D}$ is the distance vector that contains $\boldsymbol{K}$ number of values. Then, the target profile is assigned to its closest cluster, i.e., the cluster yields the smallest distance. After all the profiles are assigned to a cluster, the centroid profile of each cluster is then updated as the average of all the profiles that belong to the cluster:

$$\mu'_k = \frac{\sum_{i=1}^{N_k} x_{ik}}{N_k}, \tag{2}$$

where $\boldsymbol{N_k}$ is the number of profiles in the k-th cluster. If $\mu'_k \neq \mu_k$ for any k-th cluster, a new iteration of clustering process described by equations (1) and (2) is performed, in which all data profiles are reassigned based on the updated centroids.

We note that the result of clustering analysis is sensitive to the choice of $\boldsymbol{K}$ value. The Elbow Method is often used to optimize the determination of $\boldsymbol{K}$ value (Eymold and Jordan 2019). This is done by calculating the total distance of all data profiles to the corresponding centroid (of the cluster they assigned to), which is given by:

$$J(\boldsymbol{K}) = \sum_{i=1}^{n} \sum_{k=1}^{\boldsymbol{K}} \delta_i^k ||x_i - \mu_k||^2 \tag{3a}$$

where,

$$\delta_i^k = \begin{cases} 1, & if \ \min_j \left(|x_i - \mu_j|\right)^2 = |x_i - \mu_k|^2 \\ 0, & otherwise \end{cases}. \tag{3b}$$

The optimal ***K*** value is determined as the knee of the objective function ***J(K)***, where the gradient of the total variance flattens, indicating a diminishing return for increasing number of centroids. The clustering result may also be sensitive to the initial centroids if the objective function ***J(K)*** reaches a local minimum. Here, we run the clustering analysis 10 times using initial centroid locations generated randomly and keep the result with the lowest ***J(K)***.

4 Results

Figure 3 shows results of K-means clustering analysis performed directly on the Vs profiles of the CVM-H15.1 and CVM-S4.26 with an optimized K value of 3. This is similar to Eymold and Jordan (2019) but with the K-means clustering applied only to Vs in the top 50 km and grid cells covered by the phase velocity maps of Qiu et al. (2019). Similar large-scale spatial patterns can be seen for clustering results of both velocity models (Figs. 3a and 3c). Cluster #1 (colored in red) covers regions with extremely low velocities at shallow depth, including sedimentary basins like Salton Trough and LA basin. For Clusters #2 (in blue) and #3 (in green), we overlay the 31 km Moho depth contour resolved from Tape et al. (2012) onto the clustering maps (Figs. 3a and 3c) and find a good correlation between the contour lines (white dashed curves) and boundaries between the two clusters.

The contour lines of the Moho interface at 31 km and the boundaries of Cluster #3 matches particularly well for CVM-H15.1. In this case, the Moho depth variation dominates the clustering results (Figs. 3b and 3d). This result is different from the more complicated pattern obtained in Eymold and Jordan (2019), which is likely because the 1-D Vp and Vs profiles at each grid cell are combined first before clustering and their study area is much larger. We note that the distribution of clusters could vary significantly if the depth range of the input Vs is changed, as the result would have no sensitivity to the Moho depth variation if structures only in the top 10 km are analyzed.

Although both models yield similar spatial patterns of the resulting clusters, obvious differences are still observed and difficult to interpret. In this study, however, we apply the clustering analysis to the synthetic phase dispersion curves calculated at all available grid cells for each CVM. Different from clustering of Vs profiles, the resulting spatial pattern of clusters from the synthetic phase velocity dispersion curves can be evaluated quantitatively using the observed phase velocity maps. Therefore, we first present the clustering analysis for the phase velocity maps derived by Qiu et al. (2019) and then evaluate each CVM by comparing the corresponding clustering result with that of the observed phase velocity maps.

4.1. Clustering of the observed phase velocity maps

Figures 2c and 2f show a map-view and a 1-D profile of the Rayleigh wave phase velocity dispersion data obtained from Qiu et al. (2019), respectively. Compared to the Vs model (e.g., Figs. 2d-e), the phase velocity profile (Figure 2f) is much smoother (e.g., no sharp velocity gradient due to Moho discontinuity) and sensitive to Vs values in a wide range of depth (Fig. S1). Since the number of clusters K is a hyperparameter, we apply the Elbow Method and obtain the optimal K value as 4 (Fig. S2). The clustering result of the observed phase velocity maps (Fig. 4a) shows that Clusters #1 (in orange) and #2 (in red) mainly occupy the basin areas (e.g., LA basin and Salton Trough) with a very relatively low phase velocity at short period (3-6s), Cluster #3 (in blue) appears mostly in the Peninsular Ranges region, and Cluster #4 (in green) covers the Mojave Desert area.

4.2. Clustering of synthetic phase velocity maps for CVM-H15.1

Similar to section 4.1, we use K = 4 in the clustering analysis of synthetic phase velocity dispersion curves calculated for CVM-H15.1 and the result is shown in Figure 4b. We note that, for a direct comparison, only dispersion curves calculated for grid cells covered by the data of Qiu et al. (2019) are included in the analysis. To ensure the colors assigned to clusters obtained for the CVM-H15.1 are consistent with those of the observed data, we use the centroid dispersion curve to label each cluster (Figs. 4d and 4e). Although the resulting spatial pattern also highlights the Salton Trough, Los Angeles basin, and Ventura basin (i.e., low velocity anomalies at shallow depth; Fig. 4e) with Clusters #1 and #2, the area is much smaller compared to those in Figure 4a. For each cluster label, we calculated the Jaccard index (Halkidi et al., 2002), the ratio between the sizes of intersection and union of two datasets, to estimate the similarity between two datasets and get the overall Jaccard index of 18.6% accounting for all clusters. We also compute the corresponding true positive rate (TPR) that is adopted in Eymold & Jordan (2020) for each cluster (Table S1).

4.3. Clustering of synthetic phase velocity maps for CVM-S4.26

Clustering result of CVM-S4.26 using K = 4 is shown in Figure 4c. A good spatial correlation is observed between Clusters #1 and #2 and basin areas. Moreover, the size of these two clusters agrees well with those in Figure 4a. Consistent with clustering pattern for the observed phase velocity maps, majority of the grid cells in Cluster #3 is also well confined within the Peninsular Ranges region (Fig. 4c). Both Jaccard index and TPR for all clusters obtained from CVM-S4.26 are significantly higher than (~2-4 times of) those of CVM-H15.1. More specifically, the overall Jaccard index of CVM-S4.26 is 57.4%, which is ~3 times than that of CVM-H15.1 (Table S1).

5 Discussion

In this study, we develop an alternative method to rate a velocity model via the K-means clustering method. This technique is applied to Community Velocity Models (CVMs) in SC using Rayleigh wave phase velocity maps derived from Qiu et al. (2019). Here, we further investigate the results by analyzing the K value, depth sensitivity kernel, and data misfit.

5.1. Selection of K value

The K-means clustering analysis assigns similar data samples or profiles into the same cluster and is effective in extracting grouping features from large unlabeled datasets. However, the clustering result is dependent on the input number of clusters, i.e., the K value. In section 3, the optimal K is 3 for clustering of 1-D Vs profiles (Fig. 3), whereas an optimal K = 4 is used in clustering of phase velocity dispersion curves in section 4 (Fig. 4). Since the effect of K value on the clustering result of 1-D Vs profiles is well discussed in Eymold and Jordan (2019), we focus on how the choice of K value alters the clustering of phase velocity dispersion curves and illustrate the results using K = 3 in Figure S3 and K = 5 in Figures S4-S6.

For K = 3, the number of extracted features from the clustering analysis is reduced compared to the case with K = 4. As expected, the spatial pattern shown in Fig. S3a for the observed phase velocity dispersion curves is almost identical to that shown in Fig. 4a after merging Clusters #1 (center of the basins) and #2 (edge of the basins) together. However, for the clustering of synthetic phase velocity dispersion curves (Figs. S3b-c), Cluster #3 in Figs. 4a and 4c that primarily occupies the Peninsular Ranges region is missing from the K = 3 results, indicating the difference between synthetic dispersion curves in the center and at the edge of basins is much larger than the difference between basin and non-basin. This may be caused by the anomalously low phase velocities (< 2 km/s) in the period range of 3-5s within LA basin, Ventura basin, and Salton Trough (red color areas in Fig. S3b-c), where significantly low Vs (< 1 km/s) at shallow depth are observed in both CVMs (Figs. S3e-f).

On the other hand, for K = 5, the clustering result of the observed phase velocity is highly dependent on the initialization, i.e., the choice of starting centroids (section 3). This is illustrated in Figure S4, where two different clustering patterns are obtained when two different starting centroids are randomly initialized. Such observed difference is greatly suppressed if we reduce the number of clusters from 5 to 4 by attributing the cluster in maroon to red and blue in Fig. S4a and Fig. S4c, respectively. The clustering result for the synthetic phase dispersion curves of CVMH-15.1 is also dependent on the centroid initialization (Fig. S5), whereas the clustering result for CVM-S4.26 is less sensitive to the choice of starting centroids (Fig. S6).

In conclusion, clustering results using K = 5 are less stable than those of K = 3 and K = 4, and the result of K = 3 can be easily reproduced by merging two specific clusters obtained using K = 4. This likely suggests a maximum number of four dominating groups that can be extracted from the Rayleigh wave phase velocity dispersion curves between 3-16s in the study area through clustering analysis, which justifies our choice of K = 4 based on the Elbow Method result.

5.2. Depth sensitivity

Figures 3a and 3c show the clustering results for 1-D Vs profiles in the top 50 km extracted from CVM-H15.1 and CVM-S4.26, respectively. The resulting spatial pattern yields two dominating structural features: basins (in red) with low velocities in the top 10 km and regions (in green) with a deep (> 31 km) Moho discontinuity. The clusters obtained using the observed phase velocity dispersion curves between 3s and 16s, on the other hand, exhibit a different spatial pattern (Figs. 4a and 4d). While the basins still stand out from the clustering results in Fig. 4a, the other dominating structural feature outlined by the clustering analysis of dispersion curves is the Peninsular Ranges.

Considering Rayleigh wave phase velocities at periods < 16s are most sensitive to structures in the top 20 km (Fig. S1), the variation in Moho depth likely has little contribution to dispersion curves between 3s and 16s. This is supported by the observation that the spatial pattern in Fig. S7 derived using 1-D Vs profiles of CVM-S4.26 only in the top 20 km is consistent with that of the observed dispersion curves (Fig. 4a). Therefore, we mainly evaluate the CVMs only in the top ~20 km via clustering analysis of dispersion curves between 3s and 16s. It is important to note that, in addition to extending the period range, we can also evaluate the velocity model at shallower depth by incorporating H/V ratio measurements from Berg et al. (2018).

5.3. Comparison with data misfit

We compute the data misfit of Rayleigh wave phase velocity for each CVM as the L2 norm between the observed and synthetic dispersion curves (Fig. S8). The resulting misfit maps show similar patterns for both CVM-H15.1 and CVM-S4.26 with median values of ~0.5 km/s. In general, basin regions yield large misfit values (> 0.6 km/s), while smaller values (< 0.3 km/s) are observed in Mojave Desert and Peninsular Ranges. This suggests both models are similar in terms of fitting the phase velocity dispersion data. In contrast, our clustering-analysis-based evaluation method aims at comparing spatial patterns of the dominating structural features extracted independently from the observed and synthetic datasets, rather than focusing directly on the difference between them that is predominated by basin areas, and clearly shows that CVM-S4.26 is a better choice for structures in the top ~20 km.

6 Conclusions

We develop, for the first time, a simple workflow to evaluate velocity model via the K-means clustering method using observed surface wave phase velocity dispersion maps. This is done by first applying the K-means clustering analysis to synthetic phase velocity dispersion curves calculated for CVM-H15.1 and CVM-S4.26, and then validating each synthetic dataset against the observed phase velocity maps obtained by Qiu et al. (2019). The resulting clustering pattern of both models is dominated by the distribution of sedimentary basins and major geologic provinces (e.g., Mojave Desert and Peninsular Ranges). Based on the comparison between clustering results of synthetic and observed dispersion curves, the Jaccard similarity coefficient averaged over all clusters is 57.4% for CVM-S4.26, which is more than 3 times higher than that of CVM-H15.1 (18.6%), suggesting the spatial pattern of clusters obtained from CVM-S4.26 matches much better with that of the observed data than CVM-H15.1. This is consistent with the fact that ambient noise cross correlation data is included in the inversion of CVM-S4.26 but not incorporated in the construction of CVM-H15.1.

Since the observed phase velocity maps between 3s and 16s are likely only sensitive to velocity structures in the top 20 km, other types of seismic data (e.g., H/V ratio, receiver function) that have higher sensitivity to a different depth range could be incorporated into the evaluation scheme to assess the part of velocity model at shallower or greater depth. The proposed clustering-based model evaluation method provides a simple and first-order rating system for any existing velocity models that complements the more sophisticated model validation studies based on 3-D full-waveform simulations and can provide crucial information to future development of tomographic models, such as merging velocity models (e.g., determine the weighting of each velocity model in overlapping regions).

Data Availability Statement

The Rayleigh wave phase velocity maps are obtained from Qiu et al. (2019) and accessible at https://doi.org/10.17632/dt9x54dtrr.1. The community velocity models were extracted using UCVMC (https://github.com/SECcode/UCVMC). The Python module Scikit-Learn version 1.01 (Pedregosa et al., 2011) is used to perform the K-means clustering.

Acknowledgements

The authors thank W. K. Eymold for useful discussions and are grateful to all members of the Seismology and Tectonics at Rice University for comments and discussions. We thank the Editor Dr. Daoyuan Sun, an anonymous reviewer and Dr. Weisen Shen for their constructive comments that help improve this paper. This study was supported by Rice University.

References

Barak, S., Klemperer, S. L., & Lawrence, J. F. (2015). San Andreas Fault dip, Peninsular Ranges mafic lower crust and partial melt in the Salton Trough, Southern California, from ambient-noise tomography. *Geochemistry, Geophysics, Geosystems*, 16, 3946–3972. doi:10.1002/2015GC005970.

Berg, E. M., Lin, F.-C., Allam, A., Qiu, H., Shen, W., & Ben-Zion, Y. (2018). Tomography of Southern California via Bayesian joint inversion of Rayleigh wave ellipticity and phase velocity from ambient noise cross-correlations. *Journal of Geophysical Research: Solid Earth* 123, 9933–9949. https://doi.org/10.1029/2018JB016269

Bergen, K. J., Johnson, P. A., de Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, 363(6433), eaau0323. https://doi.org/10.1126/science.aau0323

Eymold, W. K., & Jordan, T. H. (2019). Tectonic regionalization of the Southern California crust from tomographic cluster analysis. *Journal of Geophysical Research: Solid Earth*, 124, 11,840–11,865. https://doi.org/10.1029/2019JB018423

Fang, H., Zhang, H., Yao, H., Allam, A., Zigone, D., Ben-Zion, Y, et al. (2016). A new algorithm for three-dimensional joint inversion of body wave and surface wave data and its application to the Southern California plate boundary region. *Journal of Geophysical Research: Solid Earth*, 121, 3557–3569. doi:10.1002/2015JB012702

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster validity methods: part I. *ACM SIGMOD Record*, **31**(2). https://doi.org/10.1145/565117.565124

Herrmann, R. B. (2013). Computer programs in seismology: An evolving tool for instruction and research. *Seismological Research Letters*, 84(6), 1081–1088. https://doi.org/10.1785/0220110096

Imperatori, W., & Gallovič, F. (2017). Validation of 3D Velocity Models Using Earthquakes with Shallow Slip: Case Study of the 2014 Mw 6.0 South Napa, California, Event. *Bulletin of the Seismological Society of America* 107 (2): 1019–1026. doi: https://doi.org/10.1785/0120160041

Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020). Identifying different classes of seismic noise signals using unsupervised learning. *Geophysical Research Letters*, 47, e2020GL088353. https://doi.org/10.1029/2020GL088353

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2018). Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, 90(1), 3–14. https://doi.org/10.1785/0220180259

Lee, E. J., Chen, P., Jordan, T. H., Maechling, P. B., Denolle, M. A., & Beroza, G. C. (2014). Full-3-D tomography for crustal structure in southern California based on the scattering-integral and the adjoint-waveform methods. *Journal of Geophysical Research: Solid Earth*, 119, 6421–6451. https://doi.org/10.1002/2014JB011346

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(1), 281–298. https://doi.org/10.1111/j.1365-246X.2008.03720.x

Lin, F.-C., Li, D., Clayton, R. W., & Hollis, D. (2013). High-resolution 3D shallow crustal structure in Long Beach, California: Application of ambient noise tomography on a dense seismic array. *Geophysics*, 78: Q45-Q56. https://doi.org/10.1190/geo2012-0453.1

Ma, S., Prieto, G. A., & Beroza, G. C. (2008). Testing Community Velocity Models for Southern California Using the Ambient Seismic Field. *Bulletin of the Seismological Society of America*, 98 (6): 2694–2714. doi: https://doi.org/10.1785/0120080947

Magistrale, H., McLaughlin, K., & Day, S. (1996). A geology-based 3D velocity model of the Los Angeles basin sediments. *Bulletin of the Seismological Society of America*, 86(4): 1161–1166.

Magistrale, H., Day, S., Clayton, R.W., & Graves, R. (2000). The SCEC southern California reference three-dimensional seismic velocity model version 2. *Bulletin of the Seismological Society of America*, 90(6B), S65–S76. https://doi.org/10.1785/0120000510

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Qiu, H., Lin, F.-C., & Ben-Zion, Y. (2019). Eikonal Tomography of the Southern California Plate Boundary Region. *Journal of Geophysical Research: Solid Earth*, 124, 9755–9779. https://doi.org/10.1029/2019JB017806

Roux, P., Moreau, L., Lecointre, A., Hillers, G., Campillo, M., Ben-Zion, Y., Zigone, D., & Vernon, F. (2016). A methodological approach toward high-resolution seismic imaging of the San Jacinto Fault Zone using ambient noise recordings at a spatially-dense array. *Geophysical Journal International*, 206, 980–992.

Schmandt, B., & Humphreys, E. (2010). Seismic heterogeneity and small-scale convection in the southern California upper mantle. *Geochemistry, Geophysics, Geosystems*, 11, Q05004. https://doi.org/10.1029/2010GC003042

Share, PE., Guo, H., Thurber, C. H., Zhang, H., & Ben-Zion, Y. (2019). Seismic Imaging of the Southern California Plate Boundary around the South-Central Transverse Ranges Using Double-Difference Tomography. *Pure and Applied Geophysics*, 176, 1117–1143. https://doi.org/10.1007/s00024-018-2042-3

Shaw, J. H., Plesch, A., Tape, C., Süss, M. P., Jordan, T. H., Ely, G., et al. (2015). Unified structural representation of the southern California crust and upper mantle. *Earth and Planetary Science Letters*, 415, 1. https://doi.org/10.1016/j.epsl.2015.01.016

Snover, D., Johnson, C. W., Bianco, M. J., & Gerstoft, P. (2020). Deep Clustering to Identify Sources of Urban Seismic Noise in Long Beach, California. *Seismological Research Letters*, 92, 1011–1022. https://doi.org/10.1785/0220200164.

Süss, M. P., & Shaw, J. H. (2003). P wave seismic velocity structure derived from sonic logs and industry reflection data in the Los Angeles basin, California. *Journal of Geophysical Research*, 108(B3), 2170. https://doi.org/10.1029/2001JB001628

Taborda, R., Azizzadeh-Roodpish, S., Khoshnevis, N., & Cheng, K. (2016). Evaluation of the southern California seismic velocity models through simulation of recorded events. *Geophysical Journal International*, 205(3), 1342–1364. https://doi.org/10.1093/gji/ggw085

Tape, C., Liu, Q., Maggi, A., & Tromp, J. (2010). Seismic tomography of the southern California crust based on spectral-element and adjoint methods. *Geophysical Journal International*, 180(1): 433–462. https://doi.org/10.1111/j.1365-246X.2009.04429.x

Tape, C., Plesch, A., Shaw, J. H., & Gilbert, H. (2012). Estimating a Continuous Moho Surface for the California Unified Velocity Model. *Seismological Research Letters*, 83(4): 728–735. doi: https://doi.org/10.1785/0220110118

Yang, Y., & Forsyth, D. W. (2006). Rayleigh wave phase velocities, small-scale convection, and azimuthal anisotropy beneath southern California. *Journal of Geophysical Research*, 111, B07306. https://doi.org/10.1029/2005JB004180

Zigone, D., Ben-Zion, Y., Campillo, M., & Roux, P. (2015). Seismic tomography of the Southern California plate boundary region from noise-based Rayleigh and Love waves. *Pure and Applied Geophysics*, 172(5), 1007–1032. https://doi.org/10.1007/s00024-014-0872-1

Figure caption

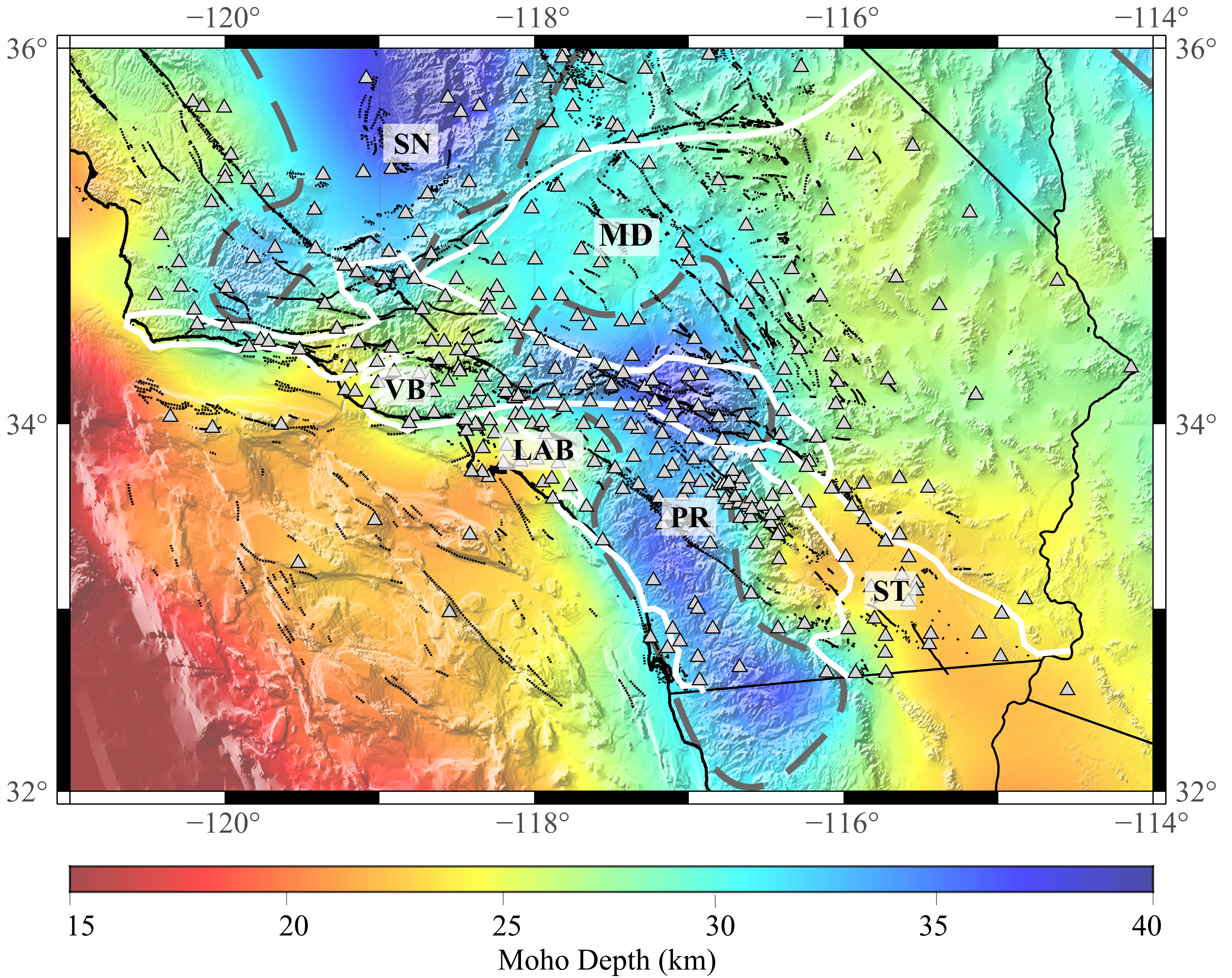
Figure 1. Map of Southern California plate boundary region. Background color indicates the Moho depth (Tape et al., 2012). Grey dashed line depicts the 31 km Moho depth contour. Grey triangles are the stations used in Qiu et al. (2019). White solid line outlines the boundaries of major geological provinces. MD: Mojave Desert, PR: Peninsular Ranges, LAB: LA Basin, VB: Ventura Basin, SN: Sierra Nevada, ST: Salton Trough.

Figure 2. Vs maps at 8 km extracted from (a) CVM-H15.1 and (b) CVM-S4.26. (c) Phase velocity map at 7 s from Qiu et al. (2019). Grey square in (a)-(c) indicates the location of the vertical velocity profile shown in (d)-(f). Grey dashed line in (d)-(f) is the average profile of the entire study region.

Figure 3. K = 3 clustering results of (a) CVM-H15.1 and (c) CVM-S4.26. White dashed line is the 31 km Moho depth contour. Average Vs profile of each cluster of (b) CVM-H15.1 and (d) CVM-S4.26.

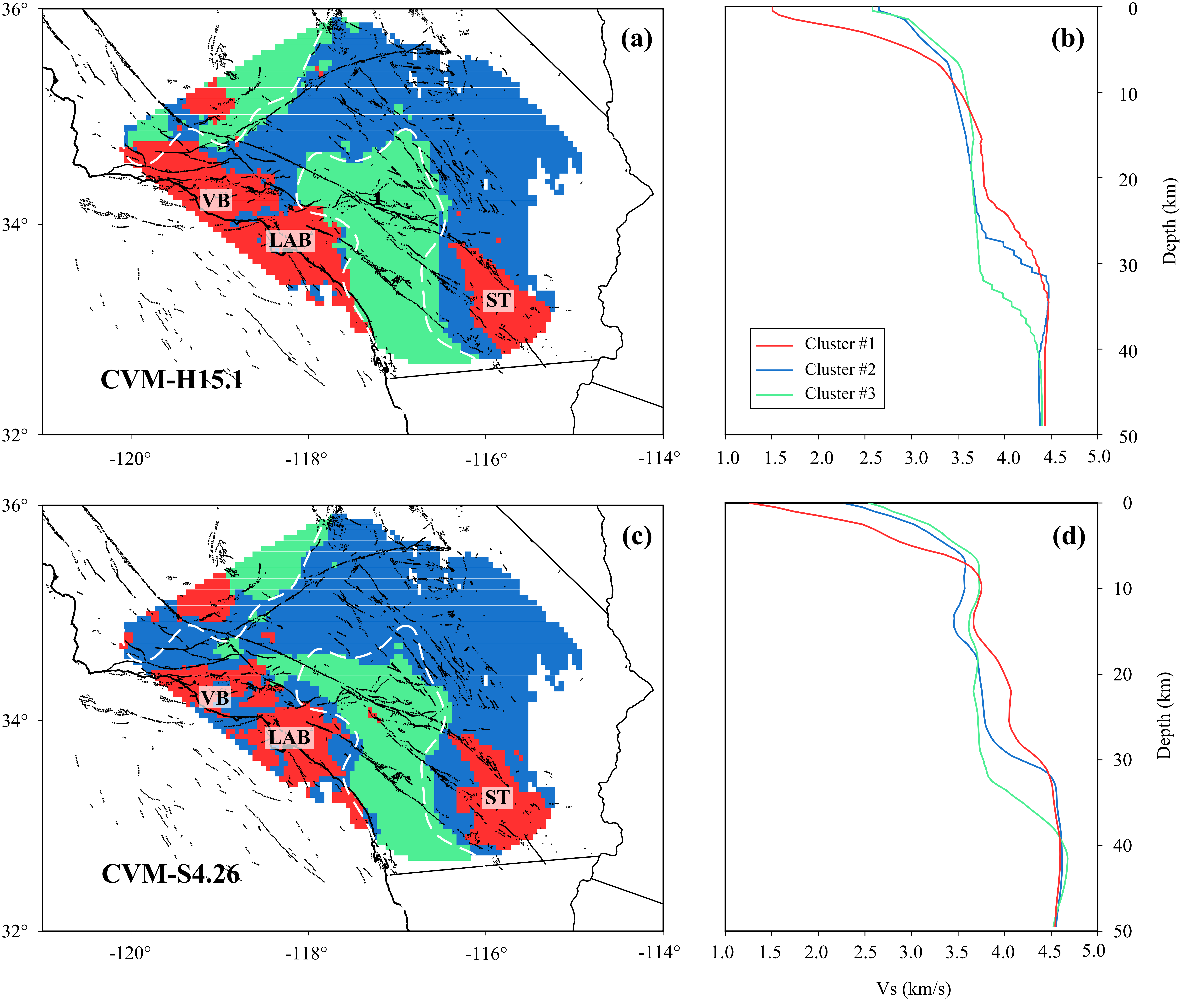
Figure 4. K = 4 clustering result computed for (a) observed phase velocity and synthetic phase velocity of (b) CVM-H15.1 and (c) CVM-S4.26. Corresponding average phase velocity profile for each cluster (d)-(f). Dashed lines in (e) and (f) are average phase velocity profiles of each cluster shown in (d).

Figure 1.



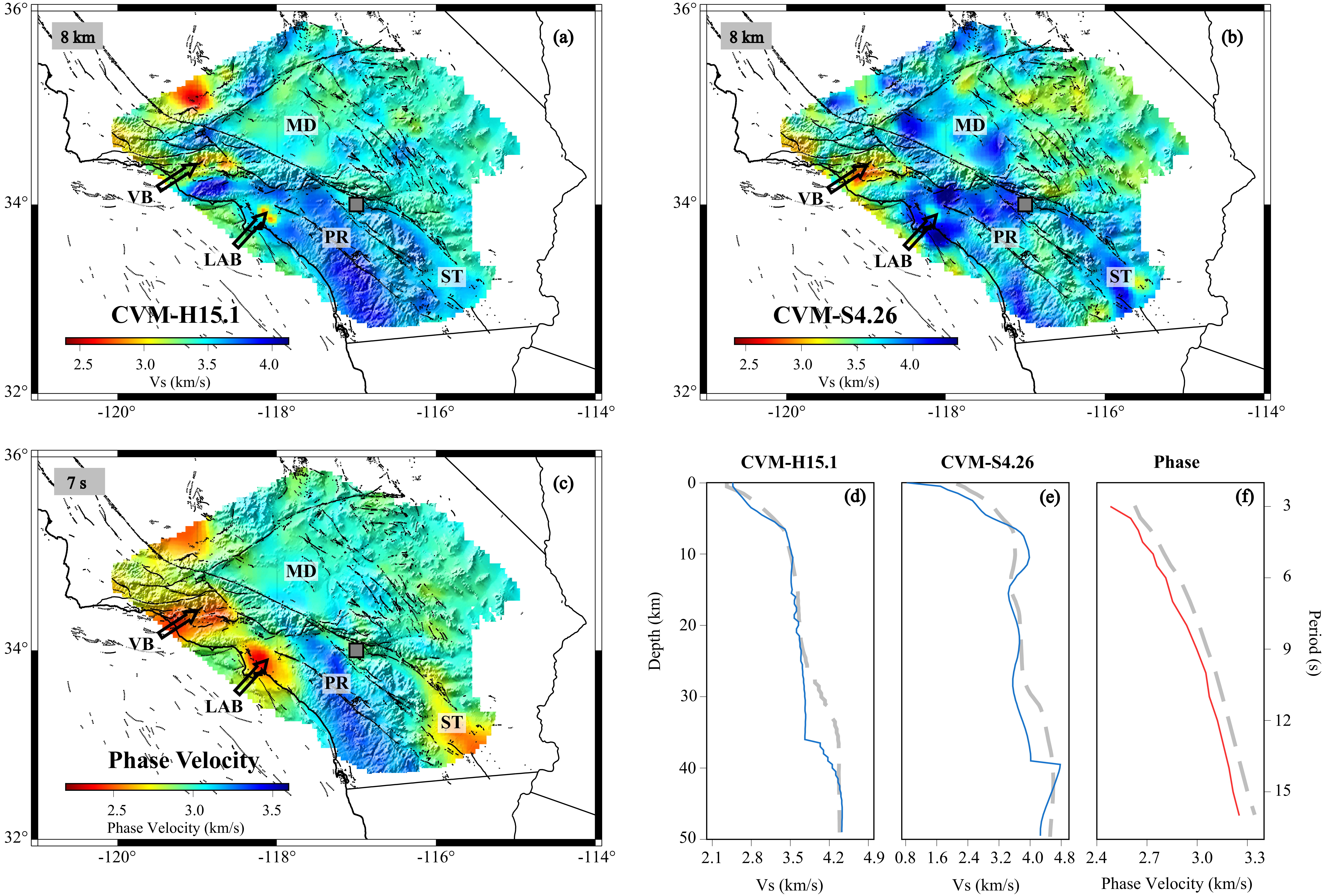
−120°
−118°
−116°
−114°
36°
34°
32°
SN
MD
VB
LAB
PR
ST
15
20
25
30
35
40
Moho Depth (km)

Figure 2.



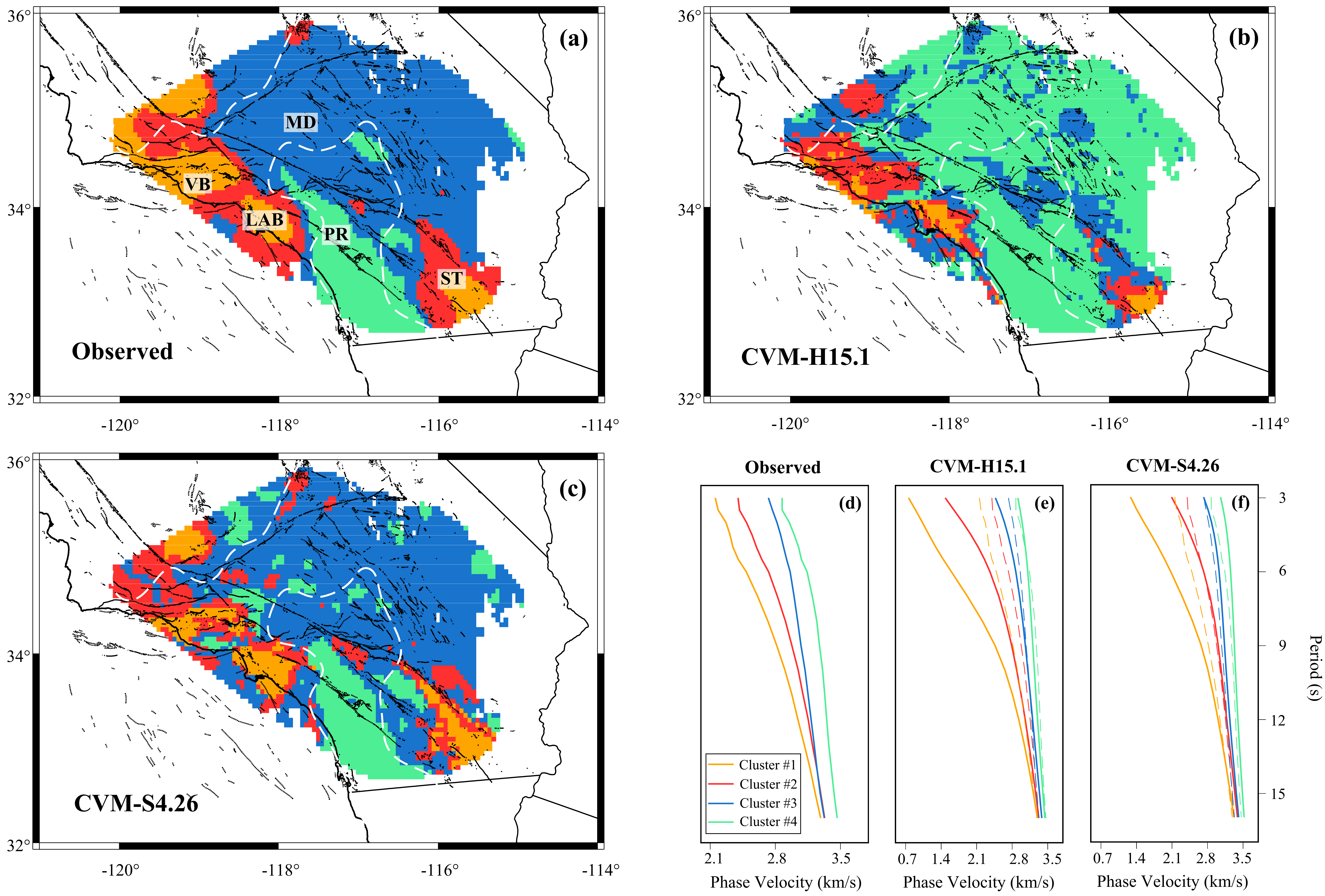
(a)
VB
LAB
ST
CVM-H15.1
(b)
Cluster #1
Cluster #2
Cluster #3
Depth (km)
(c)
VB
LAB
ST
CVM-S4.26
(d)
Depth (km)
Vs (km/s)

Figure 3.



8 km
(a)
MD
VB
LAB
PR
ST
CVM-H15.1
2.5 3.0 3.5 4.0
Vs (km/s)
8 km
(b)
MD
VB
LAB
PR
ST
CVM-S4.26
2.5 3.0 3.5 4.0
Vs (km/s)
7 s
(c)
MD
VB
LAB
PR
ST
Phase Velocity
2.5 3.0 3.5
Phase Velocity (km/s)
36°
34°
32°
-120°
-118°
-116°
-114°
CVM-H15.1
CVM-S4.26
Phase
(d)
(e)
(f)
Depth (km)
0 10 20 30 40 50
2.1 2.8 3.5 4.2 4.9
Vs (km/s)
0.8 1.6 2.4 3.2 4.0 4.8
Vs (km/s)
2.4 2.7 3.0 3.3
Phase Velocity (km/s)
Period (s)
3 6 9 12 15

Figure 4.



(a)
MD
VB
LAB
PR
ST
Observed
(b)
CVM-H15.1
(c)
CVM-S4.26
36°
34°
32°
-120°
-118°
-116°
-114°
Observed
CVM-H15.1
CVM-S4.26
(d)
(e)
(f)
Cluster #1
Cluster #2
Cluster #3
Cluster #4
2.1
2.8
3.5
0.7
1.4
Phase Velocity (km/s)
3
6
9
12
15
Period (s)