SCEC Report: Proposal 24116

A Statewide Sediment Velocity Model: Development and Implementation of a Conditional Random Field

PI: Domniki Asimaki
Postdoctoral Researcher: Grigorios (Greg) Lavrentiadis
Graduate Student: Yi Liu

July 27, 2025

Abstract

We present a non-parametric, data-driven near-surface velocity model for CA that can be used to populate the basin structures of the velocity models. The CA_SVM trained here for Southern California, is developed as a conditional random field of uncertain trend function and uncertain fluctuations about the trend of the residuals relative to the SCEC CVM-S4.26, using Gaussian process regression with the superposition of multiple Gaussian random fields. In the first part of the project, we aggregated geotechnical and geophysical data from Southern California in a 1D model with hyperparameters that we calibrated using a Bayesian estimation; and in the second part, we extended the aforementioned 1D model to 3D by means of the *separability* assumption (**?**), namely separating the fluctuations in the horizontal and vertical directions; and successively separating the latter into stationary and spatially varying kernels. In the future, we plan to integrate datasets with higher spatial resolution such as DAS inversion data to better capture the spatially varying components of the model, and work in collaboration with the SCEC IT personnel to implement the Statewide SVM on UCVM for future testing of basin effects.

1 Data description

We assembled a database consisting of 658 profiles from the web portal shear-wave velocity profile database (VSPDB) [Kwak et al., 2021], and 33 sets of sonic log profiles provided Andreas Plesch and John Shaw (personnal communication). All profiles are located within the Los Angeles Basin, as shown in Figure 1.

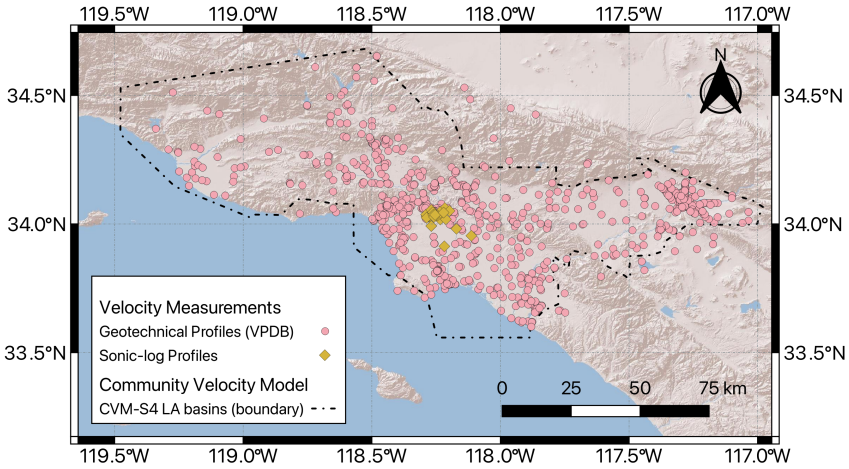

Figure 1: Locations of the profiles used to develop the model.

1.1 Geotechnical Profiles

The geotechnical profiles were obtained using invasive (downhole and cross-hole) and non-invasive (surface wave testing) methods. It has been shown that, with the exception of the very near surface of the profiles, both classes of methodologies yield similar results. We did not implement any correction factors to further unify the data across the two sets and instead, used the two classes of data as is. The marginal distribution of V_{S30} and maximum depth of the geotechnical profiles is shown in Figure 2. As can be seen, the V_{S30} range of most profiles is between 200 and 600 m/s, with a minimum V_{S30} of 171 m/s and a maximum V_{S30} of 1214 m/s. The maximum depth of most profiles is less than 100 m, concentrated between 20 and 70 m. To extend out model to deeper layers of the shallow crust, we supplemented our database with sonic log profiles obtained from the oil and gas industry exploration.

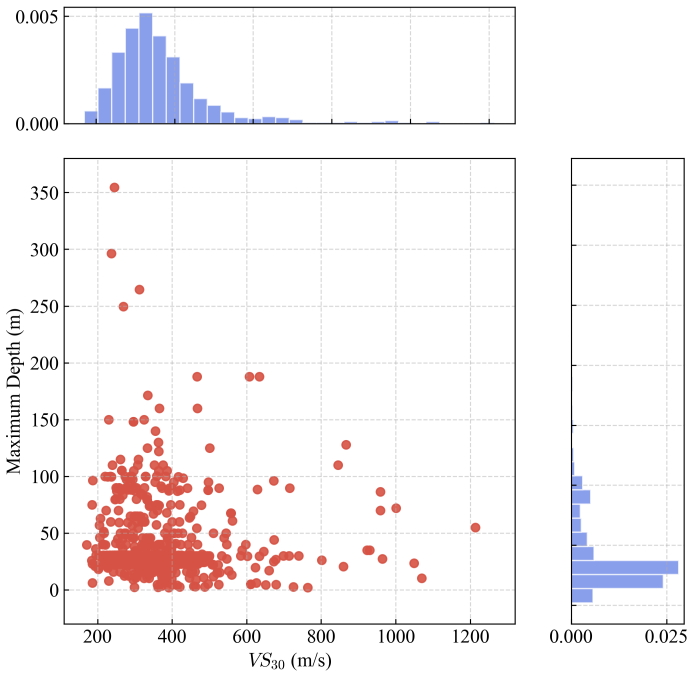

Figure 2: Distributions of V_{S30} and maximum depth of the profiles used in the model development.

1.2 Sonic Log Profiles

Shear wave velocity cannot be directly measured in sonic logs; instead, it is estimated through empirical relationships, most often by estimating the Poisson's ratio of the medium. In this case, we used the implied Poisson's ratio of CVM-S at the location of the sonic log and the measured V_P to backcalculate the measured V_S. To preprocess the data for model training, we used the moving median function and fit it with 100 spline basis functions to remove high-frequency noise. An example of a preprocessed sonic log profile is shown in Figure 3.

1.3 Combined Dataset

As mentioned above, our model is based on the residual:

$$\delta = ln(V_S) - ln(V_S^{CVM}) \tag{1}$$

where V_S is the measurement data obtained from geotechnical or sonic log profiles and (V_S^{CVM} is the velocity profile of CVM at the same location.

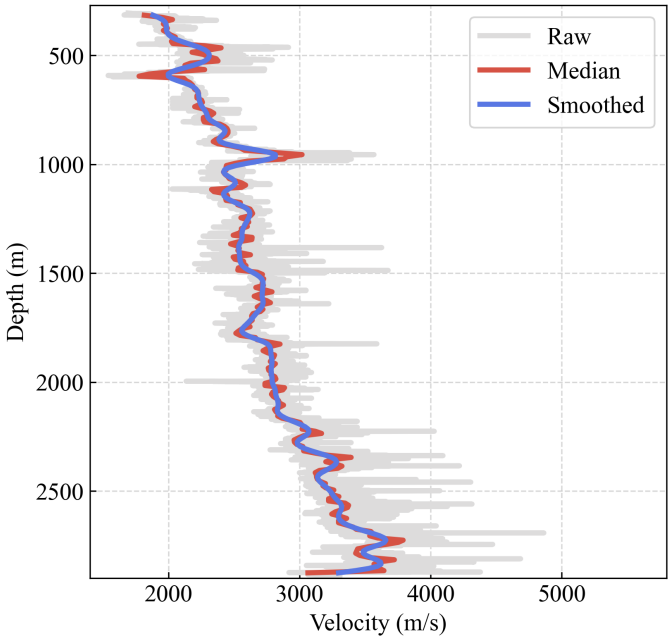

Figure 3: Preprocessing result of a sonic log profile.

The residuals corresponding to the two datasets (geotechnical and sonic logs) are shown in Figure 4 and Figure 5. As can be seen, the spread of the geotechnical profiles lies between -3 and 1 down to 400 m depth, and the majority of the data is centered around 0, which is the foundation for establishing a Gaussian regression model. On the other hand, the residuals of the sonic log profiles extend down to 2000 m depth, which helped us characterize the correlation structure at depth, where we didn't have geotechnical data.

Finally, Figure 6 depicts the combination of residuals calculated for the two different profile datasets, where the wider, shallow spread of the geotechnical profiles, and the narrow zero-centered deep extent of the sonic logs can be readily compared.

2 Model development

2.1 Model Formulation

Gaussian Process (GP) is a non-parametric Bayesian regression method widely used to model data with spatial or temporal correlations. Given a set of observation data, Gaussian processes can not only predict the mean response, but can also be used to quantify

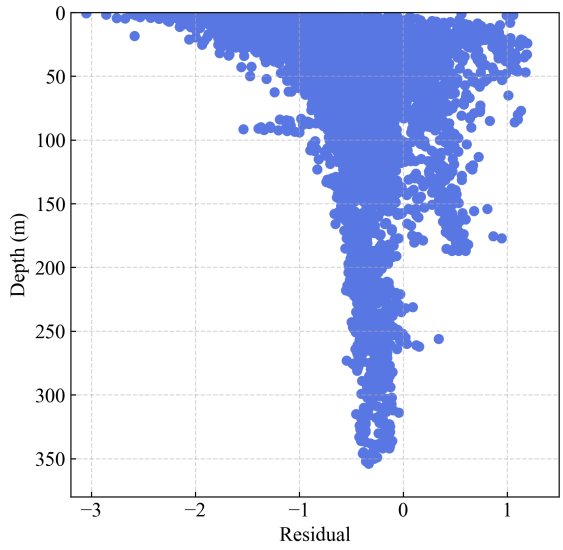

Figure 4: Geotechnical profile residuals.

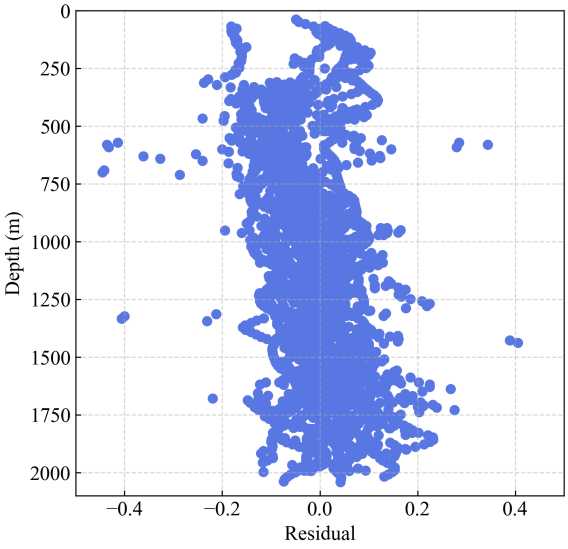

Figure 5: Sonic log profile residuals

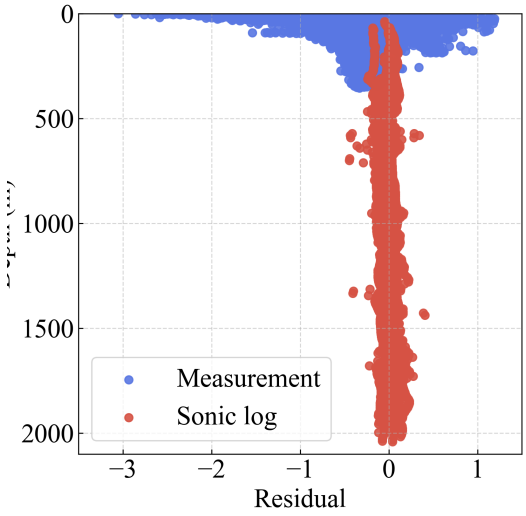

Figure 6: Residuals in the total datasets.

uncertainty. GP models are fully described by mean and covariance functions, which include prior assumptions about the smoothness, periodicity, and/or spatial structure of the objective function. A GP model is defined as follows:

$$f(x) = GP(m(x), k(x, x')) \tag{2}$$

in which $m(x)$ is the mean function and $k(x, x')$ is the covariance, where:

$$m(x) = E(f(x)) \tag{3}$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))] \tag{4}$$

In this study, we use GP to estimate the velocity residual relative to the SCEC CVM4.26 (δ) defined in Equation (1), taking the mean as 0 and the covariance, also known as the kernel function, consisting of two parts, which we elaborate on in the next section. More specifically, the residual between V_s and V_s^{CVM} is expressed as:

$$\delta \sim GP(0, k(x, x')) \tag{5}$$

Goal of the optimization is to maximize the logarithmic marginal likelihood function, that is:

$$log[p(y|X)] = -\frac{1}{2}y^T(k + \sigma_n^2 I)^{-1}y - \frac{1}{2}log|k + \sigma_n^2 I| - \frac{n}{2}log(2\pi) \tag{6}$$

where σ_n is the variance of the observed noise.

2.2 Kernel Development

In our study, the input parameters are: X, Y, d, *elevation*, $surf\ elevation$ and $Z_{2.5}$, where X, Y are Easting and Northing in Universal Transverse Mercator (UTM) coordinates, $surf\ elevation$ is the difference between depth (d) and elevation (*elevation*) as extracted from a digital elevation model, and $Z_{2.5}$ is the depth at which the shear wave velocity becomes equal to 2500 m/s. To capture the regional and site-specific spatial correlation, we formulate the kernel of the GP model as a composite of a stationary and a spatially varying one. The composite kernel showed optimal performance compared to each kernel alone. Below we describe the structure of each term in detail.

2.2.1 Stationary Kernel

To capture the spatial variability of the regional structural trend, we introduce the following stationary kernel terms:

$$k_1(x, x') = (k_r(Z_{2.5}, surf\ elevation) + k_c) \times k_r(d) \tag{7}$$

which k_c is the constant kernel, and k_r is the radial basis function (RBF) kernel shown below:

$$k_r(r) = \sigma^2 \exp\left(-\frac{r^2}{2\ell^2}\right) \tag{8}$$

This term consists of two radial basis function kernels and one constant kernel, specifically designed to model velocity variability within geological formations. We combine the target depth and surface elevation as inputs to the kernel function to characterize the coupling of spatial variability that manifests at depth as a result of the overburden pressure. At the same time, the depth parameter d defined below sea level is used as an independent input to further refine the spatial variation of the disturbance correlation.

2.2.2 Spatially Varying Kernel

For local spatial variability scaling, we use the product of two Matern kernel functions, which act on: (X, Y), which represents the correlation in the plane space (horizontal direction); and *elevation*, which represents the correlation in the terrain elevation. The expression for the non-stationary kernel is of the form:

$$k_1(x, x') = k_m(X, Y) \times k_m(elevation) \tag{9}$$

where k_m is the Matern kernel function shown below:

$$k_m(r) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}r}{\ell}\right) \tag{10}$$

where r is the Euclidean distance between X and Y, ν is the smoothing parameter that controls the smoothness of the kernel function that was assigned a value of 2.5 in this study, l is the length scale parameter, $\Gamma(\nu)$ is the Gamma function, and K_ν is the Bessel function of second kind.

3 Training results and validation

3.1 Training results

We finally fitted the GP model using a training set and optimized the marginal likelihood function to obtain the optimal model hyperparameters. During the training process, we used the Adam optimizer and recorded the loss values for each iteration. The loss function, which stabilizes at approximately -0.28 after 200 iterations, is shown in Figure 7.

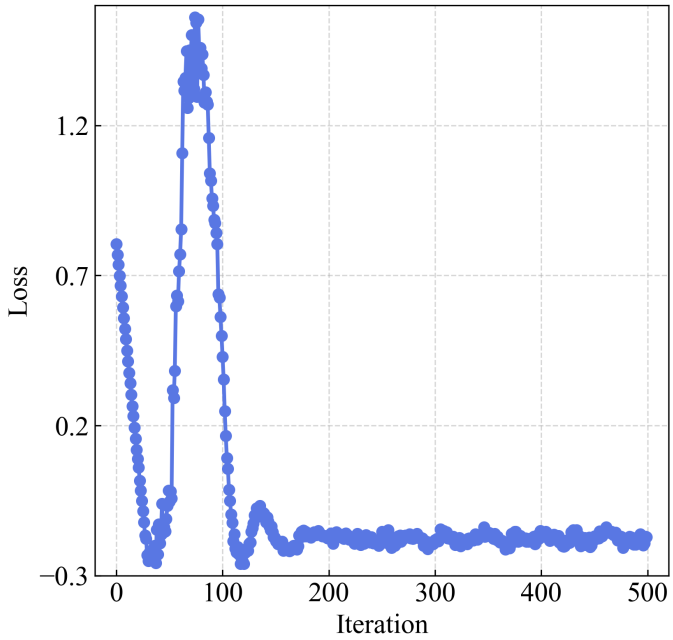

Figure 7: Loss function.

The kernel function used consists of two parts: the first part is the Matern kernel for modeling geological morphology, the second part is the RBF and constant term kernel functions used to model the variability of geological formations. Furthermore, the Matern kernel function consists of two lengthscales, the horizontal corresponding to the variables X, Y and the *elevation* one. The combination of RBF and constant term contains three lengthscales.

4 Model implementation in Los Angeles Basin

We finally use the trained model to compute the shallow crustal velocity model at six cross sections across the Los Angeles area, and compare them to the corresponding cross sections of UCVM and the tapered GTL by Olsen and collaborators. The cross sections are shown in Figure 8 and each is then compared to the state of the art tapered GTL and standard CVM-S4.26 in the Figures that follow. Results show differences that are potentially impactful for wave propagation simulations. We plan in the future to implement our model in the UCVM and simulate recorded events in Southern California to compare with previous shallow crustal models, including the tapered GTL and SVM previously developed by the authors.

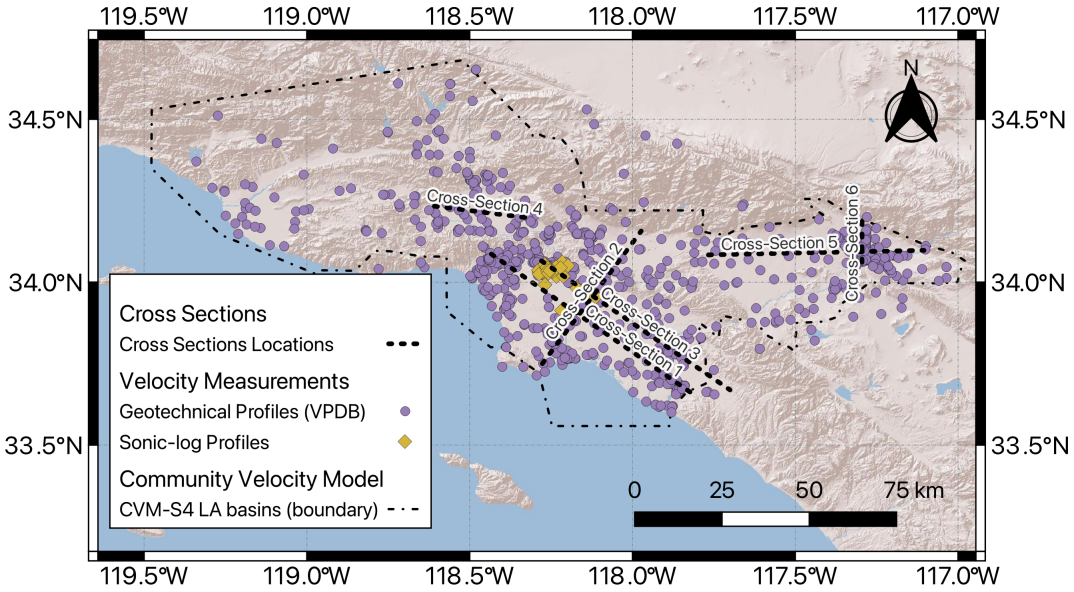

Figure 8: Cross sections depicted in the figures below.

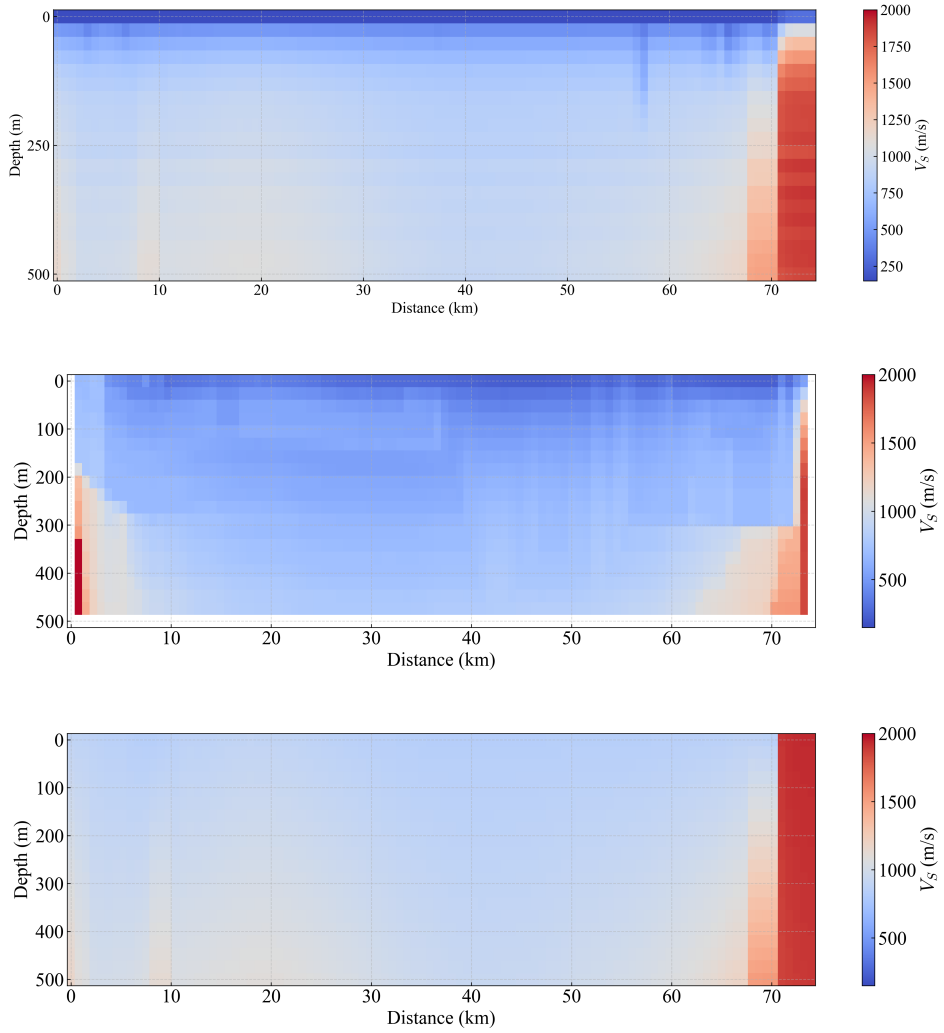

Figure 9: top) GP model (this work); (middle) tapered GTL model; (bottom) UCVM model for cross section 1

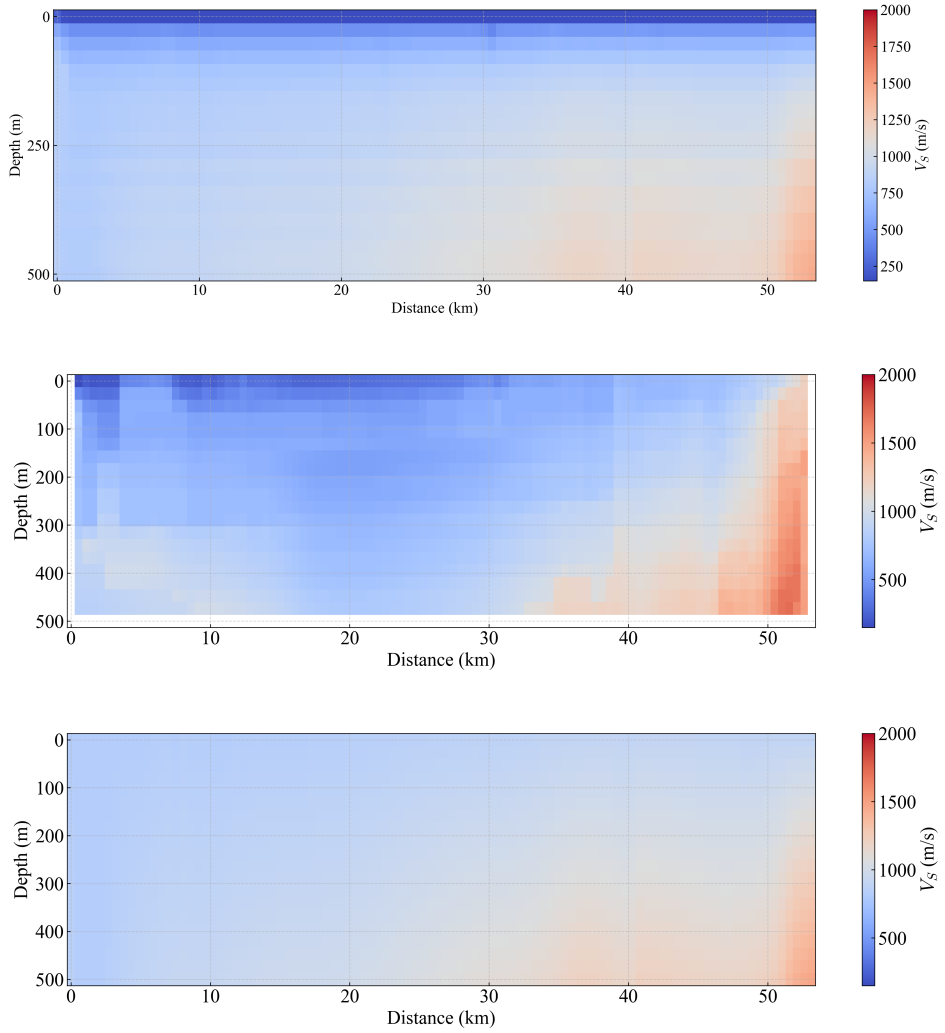

Figure 10: top) GP model (this work); (middle) tapered GTL model; (bottom) UCVM model for cross section 2

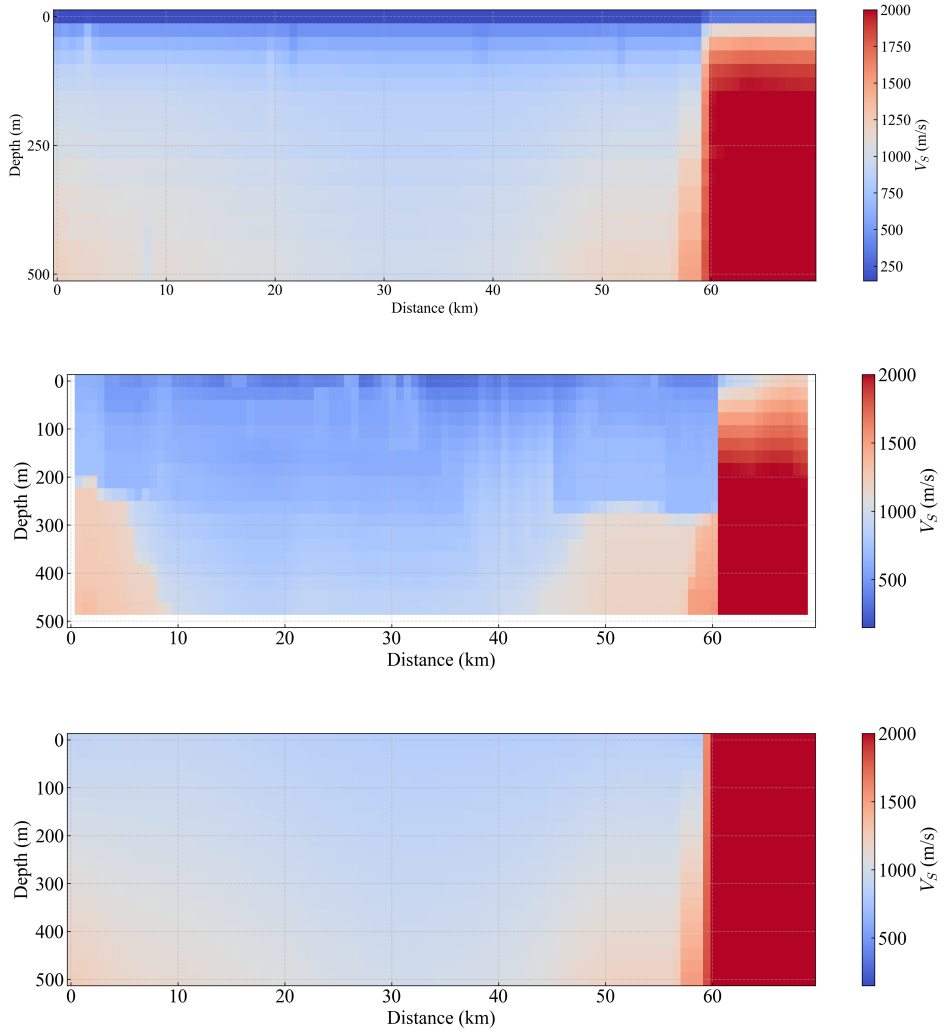

Figure 11: top) GP model (this work); (middle) tapered GTL model; (bottom) UCVM model for cross section 3

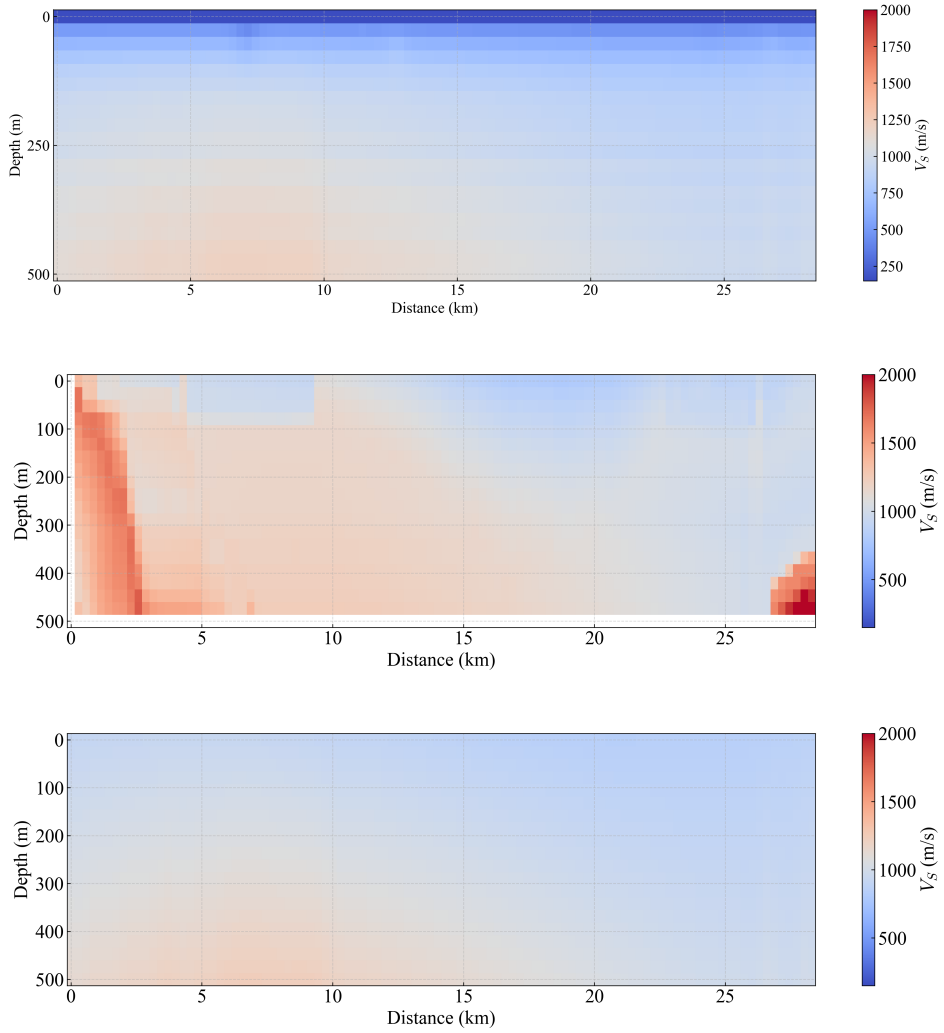

Figure 12: top) GP model (this work); (middle) tapered GTL model; (bottom) UCVM model for cross section 4

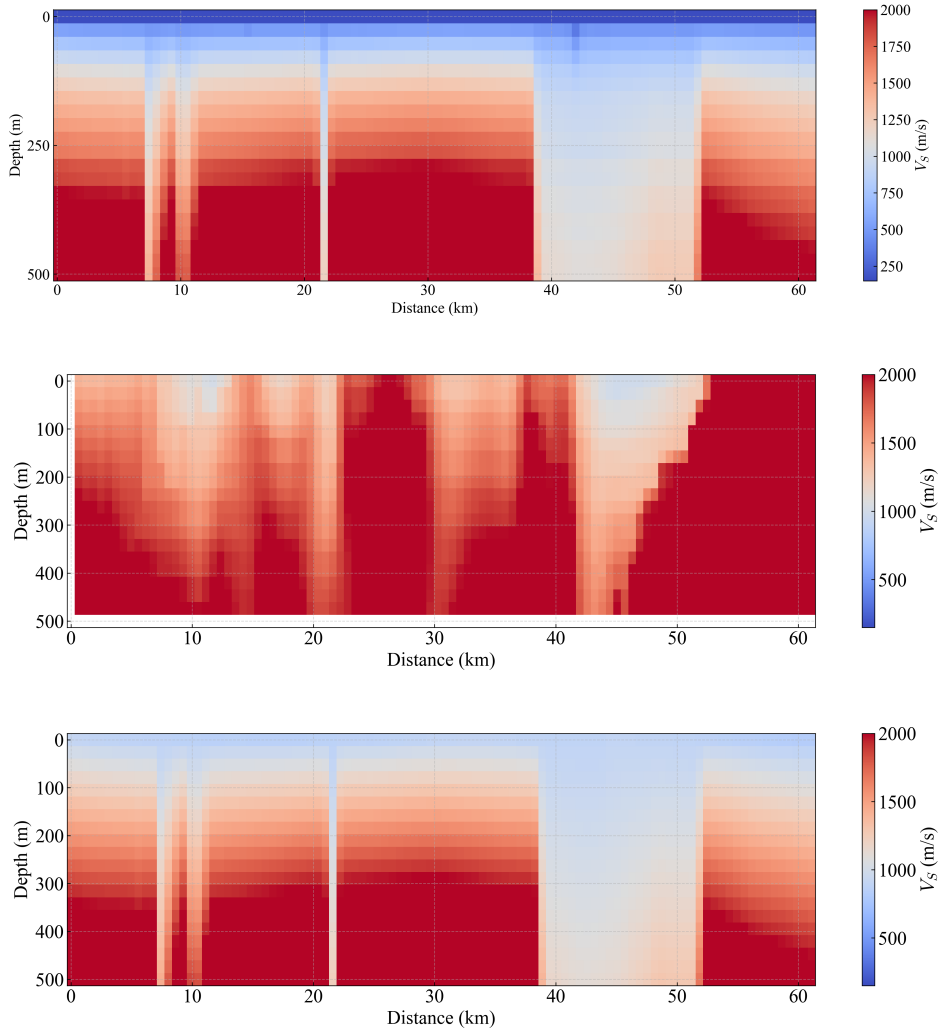

Figure 13: top) GP model (this work); (middle) tapered GTL model; (bottom) UCVM model for cross section 5

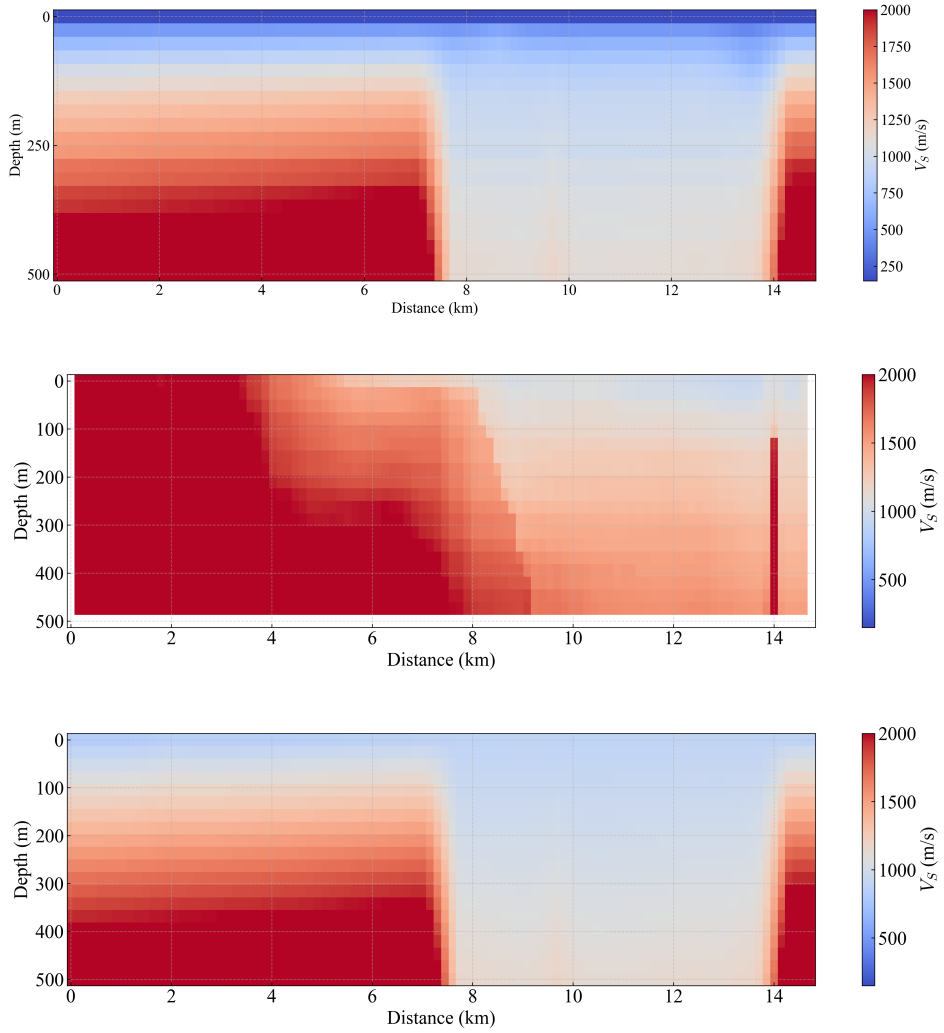

Figure 14: top) GP model (this work); (middle) tapered GTL model; (bottom) UCVM model for cross section 6