Machine Learning Methods for Spaceborne GNSS-R Sea Surface Height Measurement From TDS-1

Yun Zhang, Shen Huang, Yanling Han, Shuhu Yang, Zhonghua Hong, Dehao Ma, and Wanting Meng

***Abstract*—Sea surface height (SSH) retrieval based on spaceborne Global Navigation Satellite System Reflectometry (GNSS-R) usually uses the GNSS-R geometric principle and delay-Doppler map (DDM). The traditional method condenses the DDM information into a single scalar measure and requires error model correction. In this article, the idea of using machine learning methods to retrieve SSH is proposed. Specifically, two widely-used methods, principal component analysis combined with support vector regression (PCA-SVR) and convolution neural network (CNN), are used for verification and comparative analysis based on the observation data provided by Techdemosat-1 (TDS-1). According to the DDM inversion method, ten features from TDS-1 Level 1 data are selected as inputs; The SSH verification model based on the Danmarks Tekniske Universitet (DTU) 15 ocean wide mean SSH model and the DTU global ocean tide model is used as output verification of SSH. For the hyperparameters in the machine learning model, a grid search strategy is used to find the optimal values. By analyzing the TDS-1 data from 31 GPS satellites, the mean absolute error (MAE), root-mean-square error (RMSE) and coefficient of determination (R^2) of the PCA-SVR inversion model are 0.61 m, 1.72 m, and 99.56%, respectively; and the MAE, RMSE, and R^2 of the CNN inversion model is 0.71 m, 1.27 m, and 99.76%, respectively. In addition, the time required to train the PCA-SVR and CNN inversion models is also analyzed. Overall, the technique proposed in this article can be confidently applied to SSH inversion based on TDS-1 data.**

***Index Terms*—Convolution neural network (CNN), Global Navigation Satellite System Reflectometry (GNSS-R), principal component analysis combined with support vector regression (PCA-SVR), sea surface height (SSH).**

I. Introduction

GLOBAL Navigation Satellite System Reflectometry (GNSS-R) is a rapidly developing remote sensing technology that can use GNSS signals reflected from the ocean surface to retrieve the sea surface height (SSH) [1]. Compared with airborne and shore-based receiving platforms [2], [3], spaceborne GNSS-R not only has the advantages of multiple signal sources and low cost but also has significant advantages in global coverage and space-time diversity. The data provided by the Disaster Monitoring Consortium (DMC) of the British National Space Center (BNSC), Techdemosat-1 (TDS-1) of Surrey Satellite Technology Limited (SSTL), and Cyclone Global Navigation Satellite System (CYGNSS) of National Aeronautics and Space Administration (NASA) provide valuable opportunities for studying the performance of spaceborne GNSS-R in retrieving SSH.

Clarizia *et al.* [4] conducted a preliminary study on spaceborne GNSS-R sea level retrieval using TDS-1 data, and the results were basically consistent with the Danmarks Tekniske Universitet (DTU) 10 ocean wide mean SSH model (DTU10). Mashburn *et al.* [5] investigated the performance of the TDS-1 data for SSH retrievals on a global scale. The analysis includes consideration of the transmitter and receiver orbits, time tag corrections, models for ionospheric and tropospheric delays, zenith to nadir antenna baseline offsets, ocean and solid Earth tides, and a comparison with DTU10. Li *et al.* [6] evaluated the ocean altimetry performance of GNSS-R by analyzing different retracking methods, such as HALF, DER, and FIT, using the original CYGNSS dataset. Qiu and Jin [7] estimated the global mean SSH by using the relationship between the waveform characteristics of the delay waveform obtained from the delay Doppler map (DDM) of CYGNSS Level 1 data, and the results have a strong correlation with the results of DTU10. Zhang *et al.* [8] used TDS-1 Level 1 data and DDM SSH inversion technology to focus on the analysis of errors in inversion, and the mean absolute error (MAE) was 6.05 m with respect to the DTU 15 ocean wide mean SSH model (DTU15) results. Mashburn *et al.* [9] presented a reflection-model-based approach for delay retracking that uses simulated DDMs to retrieve the specular delay from measured DDMs, aimed to account for the challenges including precise delay retracking, correction of ionospheric effects, and spacecraft receiver positioning. These spaceborne SSH inversion research abovementioned are based on the DDM method, which can be carried out on a wide area of the sea. The other traditional method is based on carrier phase, which theoretically has higher accuracy. Li *et al.* [10] used 40 ms TDS-1 Level 0 data on sea ice to achieve sea ice thickness inversion with an accuracy of 4.7 cm. Li *et al.* [11] used CYGNSS Raw IF data to realize the inversion of water level height of Qinghai Lake with an accuracy error of less than 10 cm. However, the reflected signal is weak, it is difficult to observe the continuous

Manuscript received August 9, 2021; revised November 18, 2021; accepted December 25, 2021. Date of publication December 31, 2021; date of current version January 20, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41871325 and in part by the National Key R&D Program of China under Grant 2019YFD0900805. *(Corresponding author: Shuhu Yang.)*

Yun Zhang, Shen Huang, Yanling Han, Shuhu Yang, Zhonghua Hong, and Dehao Ma are with the College of Information Technology, Shanghai Ocean University, Shanghai 201306, China, and also with the Key Laboratory of Fisheries Information, Ministry of Agriculture, Shanghai 201306, China (e-mail: y-zhang@shou.edu.cn; m190711283@st.shou.edu.cn; ylhan@shou.edu.cn; shyang@shou.edu.cn; zhhong@shou.edu.cn; gabriel0224@126.com).

Wanting Meng is with the Shanghai Spaceflight Institute of TT&C and Telecommunication, Shanghai 201109, China (e-mail: wanting_meng@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3139376

carrier phase component, so the application scope is such as calm lake and sea ice. In addition, the accuracy improvement of traditional SSH inversion method mainly depends on some appropriate error correction, such as the uncertainty of satellite orbit, the error of a signal passing through the ionosphere and troposphere, and deviations in antenna attitude. However, there are also other unknown errors in spaceborne observations, which are too difficult to correct completely. This article attempts to seek other methods, such as machine learning (ML) methods, to obtain high precision SSH inversion model.

In recent years, Yan and Huang [12] proposed a scheme using Convolution Neural Networks (CNN) for sea ice detection based on TDS-1 GNSS-R DDMs; Asgarimehr *et al.* [13] proposed a neural network scheme for spaceborne GNSS-R wind speed inversion from TDS-1 data; Li *et al.* [14] constructed an artificial neural network (ANN) model with five hidden layers and 200 neurons per layer by analyzing CYGNSS data, and they both achieved good wind speed retrieval performance. Senyurek *et al.* [15] used three widely used ML methods: ANN, random forest, and support vector machine (SVM) to carry out comparative analysis of soil moisture inversion. These studies have obtained exciting results and proved the strong potential and value of ML methods in the field of GNSS-R inversion research.

At present, the research of GNSS-R SSH retrieval based on ML methods is in its infancy. In 2021, Wang *et al.* [16] analyzed the airborne data in the Baltic Sea and constructed a new ML fusion model for SSH retrieval. Compared with the DTU15, the root-mean-square error (RMSE) is about 0.23 m, and the correlation coefficient is about 0.75, which is a successful achievement in the study of airborne SSH inversion based on ML methods. Different from traditional empirical model studies, ML mining inversion relationship from raw data, and the accuracy of SSH can be improved by increasing the available information of DDM. The essence of SSH retrieval based on machine learning is a nonlinear regression problem of supervised learning. This article focuses on the study of spaceborne SSH inversion, and proposes for the first time two different spaceborne SSH inversion models based on principal component analysis combined with support vector regression (PCA-SVR) and CNN. TDS-1 data does not have high temporal resolution (relative to CYGNSS), but it can cover global sea surface to be used as input to the model. The mean sea surface height data of DTU15 with the data of DTU global ocean tide model also with its advantage of global coverage will be used as model output and validation of the real-time SSH. PCA is a multivariate statistical method that is mainly used to reduce the dimensionality of data [17]. SVR is a machine learning algorithm with structural risk minimization as the regression objective and can be used for linear or nonlinear regression tasks, which is a branch of SVM [18]. The introduction of PCA can reduce TDS-1 data dimensionality while achieving the effect of removing data redundancy in order to try to obtain a more accurate model for the SVR. CNN is a feedforward neural network with convolution calculation and depth structure, which is one of the representative algorithms of deep learning [19]. In this article, one-dimensional (1-D) feature sequence extracted from TDS-1 data is taken as input, and 1-D convolution is used to process data to obtain deeper data information [20]. The analysis shows that the results of PCA-SVR and CNN inversion model are basically consistent with those of the verification model.

TABLE I
FILTER PARAMETERS AND SCOPE

Parameter	Limit
signal-to-noise ratio (SNR)	> -3dB
antenna gain	> 5dB
elevation angle	> 45°

II. DATASETS

A. TDS-1 Dataset

TDS-1 is a technology demonstration satellite launched by the SSTL on July 8, 2014. The orbit altitude of the satellite is 635 km, the inclination angle is 98°, and the detection node is 9:00 P.M. local time. TDS-1 carries a satellite GNSS remote sensing receiver called SGR-ReSI, which is used to generate 1-s incoherent accumulated DDMs of GPS L1 reflection data to verify GNSS-R-related technologies [21], [22]. In this article, several time periods of the TDS-1 Level 1 data are used. The discontinuous data were collected on November 4-6, November 28-30, December 6-8, December 14-16, 2016 and April 9-30, and June 2-10, 2018. The observed satellites included all 31 satellites of the GPS constellation.

The inversion of SSH from spaceborne GNSS-R is limited by many factors. According to the experience, the data set is filtered according to the relevant thresholds in Table I, in which the SNR is DDMSNRAtPeakSingleDDM of the TDS-1 data [8]. To eliminate the influence of polar sea ice and land surface, the observation data with latitudes higher than 70° and the observation data located on land are removed [5]. In addition, the power distribution of some reflected signals is bound to be seriously affected by ocean roughness, so it is necessary to remove the data with obvious abnormal waveforms, including: 1) the peak of the waveform is not unique and 2) the delay of the peak is less than 30 lag in the slice with a zero Doppler value of DDM [see Fig. 4(b)].

After screening, data were retained from approximately 1.4 million reflections and 31 GPS satellites. Fig. 1 shows the specular reflection points of our interested data in the dataset, and it can be intuitively seen that the track of the reflection points covers almost all sea areas in the world. According to the GPS pseudorandom noise (PRN) number, Fig. 2 shows the distribution of reflected data originating from different GPS satellite, with an average of about 45 000 data per satellite.

B. SSH Verification Model

Due to the lack of real SSH data, GNSS-R SSH inversion research mainly uses the mean SSH model developed by the DTU as the reference for the real SSH. Compared with an ellipsoid, DTU15 has drawn a map with a resolution of 1/60°

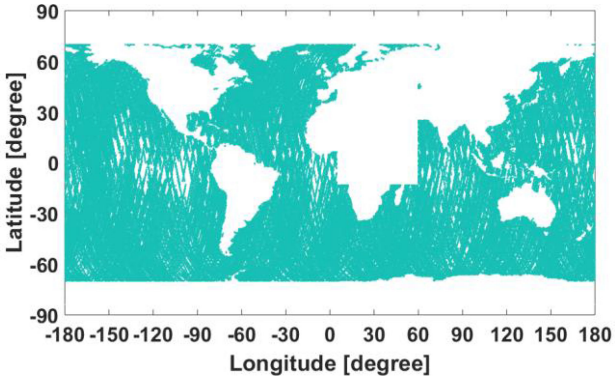


Fig. 1. TDS-1 specular reflection points of our interested data in the dataset.

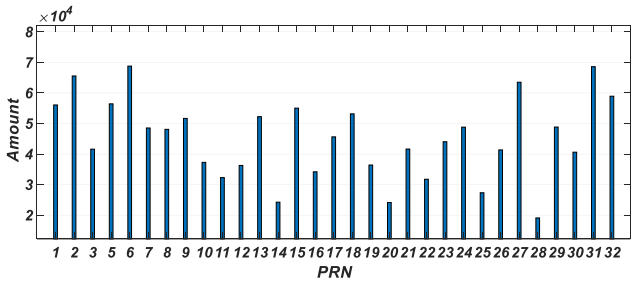


Fig. 2. Number of samples per GPS satellite.

by 1/60°, which is a grid mean SSH product with real global coverage [23].

The resolution of the DTU global ocean tide model is 0.125° by 0.125°, and the grid size is 2881 × 1441. It can calculate the instantaneous tidal height at a location corresponding to a certain latitude and longitude based on the Julian daytime. As shown in formula (1), this article selects DTU15 H_{MSSH} and the DTU global ocean tide model H_{Tide} to establish the DTU SSH verification model $\mathrm{SSH}_{\mathrm{DTU}}$ [8]. The resolution difference between DTU verification model and TDS-1 data is solved by linear interpolation

$$\mathrm{SSH}_{\mathrm{DTU}} = H_{\mathrm{MSSH}} + H_{\mathrm{Tide}}. \tag{1}$$

III. Features Extraction

In order to obtain the SSH inversion model with high accuracy by ML method, a key link is to consider what kind of parameters to be selected as training features. The specific strategy is to get inspiration from GNSS-R geometric relations and DDM SSH inversion methods. According to the following discussion, a total of ten features were selected, which are elevation angle, distance difference between delay at sea surface specular reflection point and delay at ellipsoid specular reflection point, antenna gain, SNR, the 3-D vector velocity of the signal transmitter GPS satellite, and the three-dimensional vector velocity of the signal receiver TDS-1 satellite.

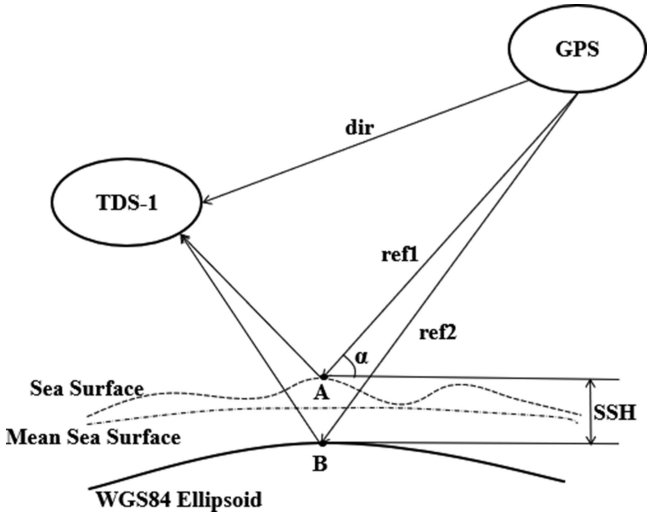


Fig. 3. Geometric relationship diagram of spaceborne GNSS-R.

A. Geometric Relationship of Spaceborne GNSS-R SSH Inversion

As shown in Fig. 3, the ellipsoid selects WGS84 as the reference and α as the elevation angle of GPS. Point A is the specular reflection point on the sea surface, and Point B is the specular reflection point on the ellipsoid. The direct signal is dir, the reflection signal passing through the specular reflection point on the sea surface is $ref1$, and the reflection signal passing through the ellipsoid specular reflection point is $ref2$. Under ideal conditions, the SSH ($\mathrm{SSH}_{\mathrm{ideal}}$) can be determined by the elevation angle α and the delay distance difference (DIFF) between $ref1$ and $ref2$ [24]

$$\mathrm{SSH}_{\mathrm{ideal}} = \frac{ref2 - ref1}{2 * \sin\alpha} = \frac{\mathrm{DIFF}}{2 * \sin\alpha}. \tag{2}$$

Fig. 4(a) shows a typical DDM generated by signal reflection at point A and Fig. 4(b) shows the delay waveform in a DDM when the Doppler value is 0. In Fig. 4(b), the center point of the delay dimension is the open-loop tracking point, which can be used to determine the delay of $ref2$ relative to dir in the waveform window ($\mathrm{Delay}_{\mathrm{OLTP}}$) [5]. The point on the leading edge of the waveform at 70% of the peak power is chosen as the retracking point, which can be used to determine the delay of $ref1$ relative to dir in the waveform window ($\mathrm{Delay}_{\mathrm{retrack}}$) [25]. Therefore, DIFF can be calculated by formula (3), where c is the speed of light

$$\mathrm{DIFF} = c * (\mathrm{Delay}_{\mathrm{retrack}} - \mathrm{Delay}_{\mathrm{OLTP}}). \tag{3}$$

According to the previous analysis, the DIFF and elevation angle α from the GNSS-R geometry relation can be used as characteristic parameters.

B. Factors Affecting DDM

As shown in Fig. 4(a), the distribution of the DDM is also affected by the Doppler effect. The Doppler frequency shift

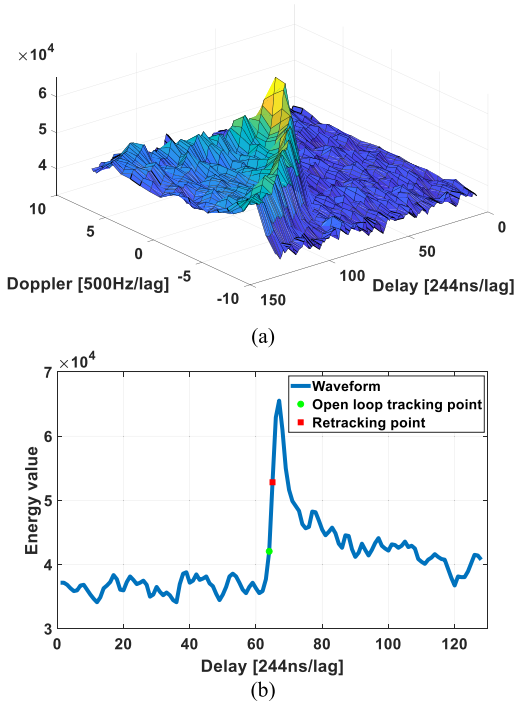


Fig. 4. (a) Example of a spaceborne DDM. (b) Slice with a zero Doppler value.

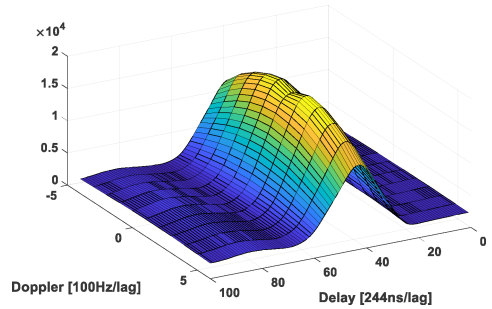


Fig. 5. Example of a ground-borne DDM.

reflects the instantaneous velocity of the receiver relative to the satellite at the time of the measurement. Fig. 5 shows a typical ground-borne DDM. Compared with Fig. 4(a), it is obvious that they differ significantly in the Doppler dimension. The signal energy values in the Doppler dimension of spaceborne DDM are more undulating, while the Doppler dimension of ground-borne DDM is smoother. The different motion patterns of spaceborne receivers and ground-borne receivers contribute to this phenomenon.

If both the transmitting and receiving sources of the signal are moving, then the Doppler shift of the satellite carrier signal received by the receivers f_d can be defined as formula (4), where $I^{(s)}$ is the unit observation vector in the direction from the transmitter GPS satellite to the receiver TDS-1 satellite, $v^{(s)}$ is the velocity of the transmitting source, v is the velocity of the receiving source, and λ is the wavelength of satellite signal [26]

$$f_d = \frac{(v - v^{(s)}) \cdot I^{(s)}}{\lambda}. \tag{4}$$

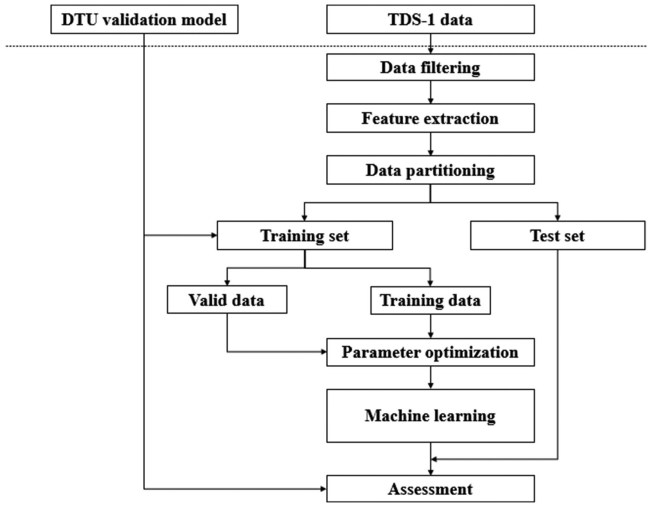


Fig. 6. SSH inversion process based on ML.

The result of the integration of formula (4) over time is equal to the amount of change in geometric distance divided by the wavelength, i.e., the integral Doppler reflects the magnitude and direction of the change in geometric distance. Therefore, the vector velocity (Tx, Ty, Tz) of the signal transmitter GPS satellite and the vector velocity (Rx, Ry, Rz) of the signal receiver TDS-1 satellite are considered characteristic parameters of the ML model.

In order to avoid the possibility that the velocity may carry additional position information, this article specifically verifies the correlation between the velocity and SSH, and the verification results show that the correlation between the velocity and height is almost zero. At last, considering that the distribution and numerical value of DDMs is directly affected by the quality of the reflected signal, antenna gain and SNR in the DDM information are also taken as the characteristic parameters of the ML model.

IV. Methodology

A. Retrieval Process of SSH

As is clearly shown in Fig. 6, the spaceborne GNSS-R height retrieval process based on the ML methods is as follows.

First, reprocessing operations including data filtering are carried out from TDS-1 dataset, and then relevant feature information is extracted. According to the close relationship between the abovementioned characteristics and GPS satellites, as well as the differences in the motion law of satellites and operation loss of satellites, this article considers that different satellite independent analyses can enhance the SSH retrieval ability of the model. To avoid the computational difficulty of large-scale satellite-borne data and to ensure the accuracy of the model as much as possible, this article randomly selects 20% of the samples in each sub dataset as the training set to train the model, and the remaining 80% is used as the test set to test the model. To avoid over fitting, in the training set, 50% of the samples are randomly selected to optimize the model parameters and

the remaining 50% are used to verify the results of parameter optimization, which means that the training sample size of the participating models is only 10% of the dataset.

Then, two ML methods, PCA-SVR and CNN, are used to establish the SSH inversion model. For performance comparison between two ML methods, the same data settings are used for both. All ten features are used as inputs. The two sections of this chapter will describe each of these two methods in detail. The experimental environment used was an Intel(R) Core (TM) i9-9820XCPU with 3.30 GHz and 64 GB of installed memory.

Finally, the inversion results are compared with the DTU verification model. As mentioned in Section II, the DTU verification model is also used as a comparison model to evaluate the inversion results. By introducing the ocean tide model and interpolation calculation, the DTU verification model has been able to obtain the approximate real-time SSH at specific locations. Therefore, the inversion results of PCA-SVR model and CNN model are also real-time SSH. After the inversion results of the model are obtained, the obvious abnormal points with a difference of more than 50 m from the model height are removed. The MAE, RMSE, coefficient of determination (R^2) and time spent on modeling are used as evaluation indexes.

B. PCA-SVR in Retrieval Process of SSH

Before the SVR analysis of the dataset extracted from TDS-1 data, PCA is used to extract more advantageous comprehensive features, which can be used to retrieve SSH more efficiently. Set the quantity of data as m and the dimension of data as n, and the dataset can be represented as matrix X_{m*n}. In this article, m represents the sample size of each satellite, and n represents the ten features. The specific algorithm of PCA is as follows.

1) Input matrix X_{m*n}, and zero mean processing of each column of data.
2) Calculate the covariance matrix $Y = XX^T$.
3) The eigenvalues and eigenvectors are calculated, and the corresponding eigenvectors $u_1, u_2, \ldots, u_l$ of the first l eigenvalues are selected to form the transformation matrix U.
4) The principal component matrix is calculated by multiplying the original matrix and the transformation matrix, $Z = U^T X$, and the output matrix is Z_{m*l}.

SVR uses l synthetic variables obtained by PCA as input parameters. Then, to improve the convergence efficiency, the principal component matrix Z_{m*l} is standardized.

The purpose of SVR is to find a mapping f such that the error between the function value f and the expected value y is not greater than a given value ε. Suppose $f(x) = \omega\varphi(x) + b$, where ω is the weight vector, b is the threshold value, and $\varphi(x)$ is the mapping function; hence, the goal of SVR can be formalized as

$$\begin{aligned} \text{Min} \quad & \tfrac{1}{2}||\omega||^2 + \quad C\sum_{i=1}^{n}(\xi_i + \xi_i{}^*) \\ \text{s.t.} \quad & \begin{cases} y_i - \omega\varphi(x) - b \le \varepsilon + \xi_i\xi_i \ge 0 \\ \omega\varphi(x) + b - y_i \le \varepsilon + \xi_i{}^*\xi_i{}^* \ge 0 \\ i = 1, 2, 3, \ldots, n \end{cases} . \end{aligned} \tag{5}$$

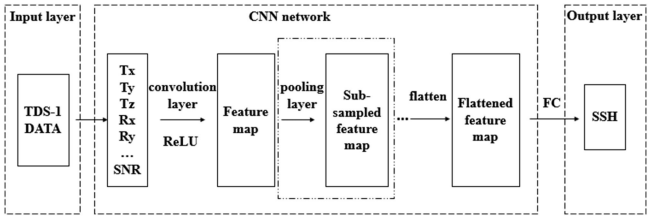


Fig. 7. Schematic diagram of the network structure of CNN.

In formula (5), $C > 0$ is the penalty parameter, and ξ_i and $\xi_i{}^*$ are relaxation variables. The duality principle is used, and the Lagrange multipliers β_i and $\beta_i{}^*$ are introduced to solve the abovementioned equation

$$f(x) = \sum_{i=1}^{n}(\beta_i{}^* - \beta_i)K(x_i, x_j) + b. \tag{6}$$

$K(x_i, y_j)$ in the formula is a kernel function introduced by nonlinear SVR to address the dimensional disaster [27]. In this article, the radial basis kernel function with moderate number of hyperparameters is selected, which can solve nonlinear problems well

$$K(x, y) = exp(-\gamma||(x - y)||^2). \tag{7}$$

C. CNN in Retrieval Process of SSH

Generally, CNN consists of an optional set of convolution layer, pooling layer, full connection layer, and output layer [19]. As shown in Fig. 7, convolution layer is composed of a certain number of convolution kernels of a certain size. Each convolution kernel can be regarded as a feature extractor, which is convoluted with input data. After that, the convolution data was processed by the Rectified Linear Unit activation function. The pooling layer can be selected to compress data to remove data redundancy. Convolution layers can be multiple. Then, use full connection layer (FC) for combining all the features learned earlier. Finally, the output is given to the classification result or regression result.

The analysis data extracted ten features from TDS-1 data is 1-D data, so the convolution kernel in convolution layers is also 1-D here. According to the classical CNN structure, two convolution layers of the same setting were chosen to process the data here. Since the extracted feature dataset has only ten features, the network structure does not consider the pooling layer in order to avoid the loss of data information as much as possible. For the inversion of the SSH, only one neuron is set in the output layer and the Linear function is chosen as the activation function. As for the number of convolution kernels and the size of convolution kernels are hyperparameters, for GNSS-R data without specific prior knowledge, this article will use some common empirical values to analyze and select the most suitable value.

The specific process of using CNN to train the SSH inversion model can be described as follows.

1) First, the training data are input into the network in batches for forward propagation.

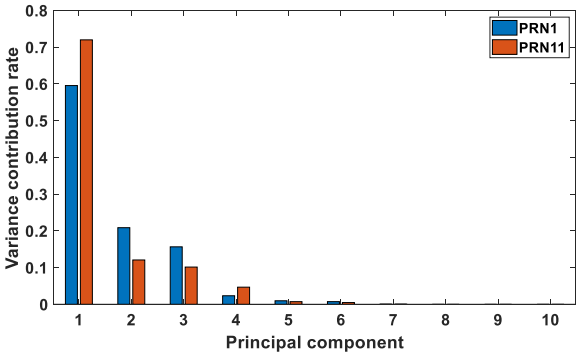


Fig. 8. Variance contribution rate of each principal component.

2) Then, the weights of network are modified by using back propagation learning [28].
3) The iterative training is carried out continuously and stops after 10 000.

V. Results and Discussion

A. Results of PCA-SVR SSH Inversion Model

Taking GPS PRN1 and PRN11 as examples, PCA is carried out on the original data. Section IV-B introduced that the principal component matrix could be established according to the eigenvalues and eigenvectors of the covariance matrix. In order to quantify the number of principal (l), the variance contribution rate of each principal component is calculated, i.e., the ratio of variance of principal component to total variance, then the cumulative variance contribution rate of selected principal components is calculated. The larger the cumulative variance contribution rate is, the more likely the original data information will be presented comprehensively [29]. Fig. 8 lists the variance contribution rate of each principal component. For PRN1, the first principal component provides a 59.53% variance contribution rate, the second principal component provides a 20.84% contribution rate, and the third principal component provides a 15.63% contribution rate. The cumulative contribution rate of the first five principal components can reach over 99%, and almost all the original feature information is retained. For PRN11, the first principal component provides 71.97% of the variance contribution, the second principal component provides 12.06%, while the third principal component provides 10.13%. And the cumulative contribution of the first five principal components can likewise reach more than 99%. Therefore, the first five ($l = 5$) principal components will be selected in the PCA-SVR model to replace the data of the original ten features to achieve the effect of removing data redundancy.

In the nonlinear SVR with radial basis function, the selection of penalty coefficient C and kernel function parameter γ largely determines the excellence of the model [30]. According to practical analysis experience, the range of SVR parameter optimization is set at C ($10^{-1\sim3}$) and γ ($10^{-1\sim2}$). According to the grid search, 20 sets of parameter combinations need to be searched.

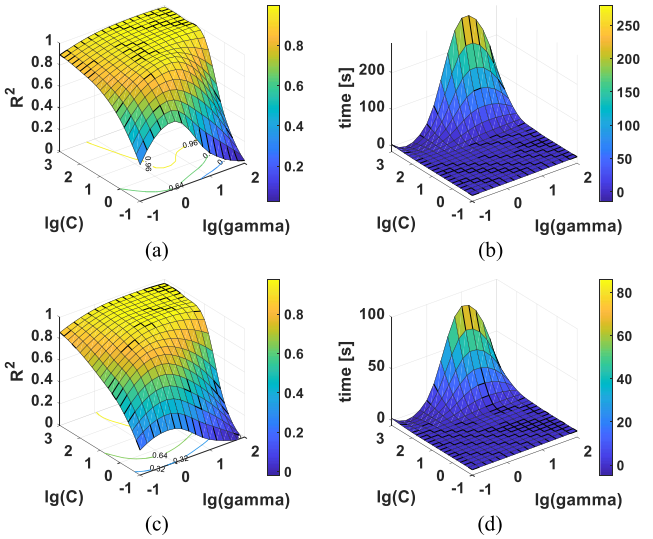


Fig. 9. Smoothed display of parameter optimization results of PRN1. (a) R^2, (b) time spent; PRN11: (c) R^2, (d) time spent. (lg = $\log_{10}$).

Fig. 9(a) and (c) shows the hyperparameter optimization process for PRN1 and PRN11. The choice of the optimal parameter combination is determined by the maximum R^2. Fig. 9(b) and (d) shows the time overhead required under each set of parameter combinations. It can be clearly observed that the optimization search process for the two satellite sub datasets has an extremely similar pattern. When γ is large and C is small, the accuracy of the model is very poor. When C is small, with the increase of γ, the accuracy of the model first increases and then decreases. When C is large, the size of γ is gradually insensitive to the impact of model accuracy. The R^2 peaks at C = 1000 and γ = 10. The R^2 of PRN1 at peak is 99.43% and the R^2 of PRN11 at peak is 97.72%. When C is small, the time spent to build the model is small and insensitive to γ. As C increases, the change in γ clearly determines the time overhead. The reason for this is that the training of the SVR does not depend on the size of the data volume, but on the number of support vectors involved in building the model. As C increases or γ decreases, the number of support vectors computed by SVR tends to increase. And to avoid the overfitting phenomenon, the dataset used to test the inversion model is independent from the training and validation datasets, as described in Section IV.

Fig. 10 shows the performance of the PCA-SVR inversion model on PRN1 and PRN11. It is clear that, except for a few points with large deviations, the inversion results are basically consistent with the DTU validation model, with MAE, RMSE, and R^2 of 0.56 m, 1.69 m, and 99.45% for PRN1, and MAE, RMSE, and R^2 of 0.88 m, 2.68 m and 98.57% for PRN11, respectively.

B. Results of CNN SSH Inversion Model

Before the analysis using the CNN method, the optimal settings of the number and size of the convolution kernels in the convolution operation were unknown. In this article, PRN1 and PRN11 are used as example analyses to determine the

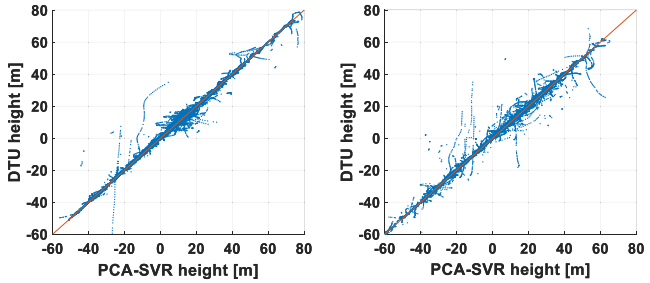


Fig. 10. PCA-SVR model inversion results and the DTU verification model on PRN1 (Left) and PRN11 (Right).

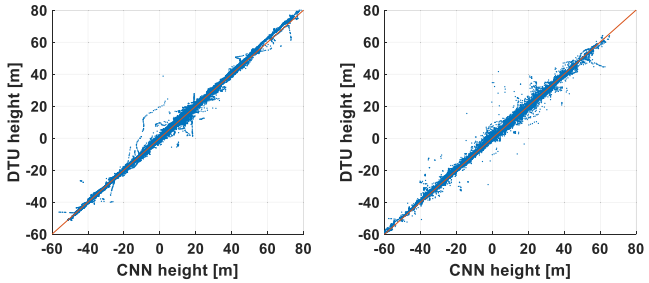


Fig. 12. CNN model inversion results and the DTU verification model on PRN1 (Left) and PRN11 (Right).

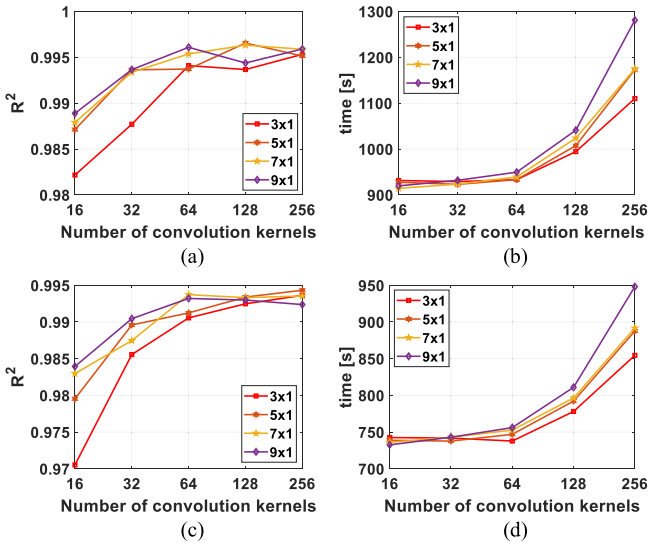


Fig. 11. Smoothed display of parameter optimization results of PRN1: (a) R^2, (b) time spent; PRN11: (c) R^2, (d) time spent.

optimal values of these two hyperparameters. According to the practical analysis experience, the number of convolutional kernels in CNN is selected as 16, 32, 64, 128, 256, and the size of convolutional kernels is selected as 3×1, 5×1, 7×1, and 9×1. In total, 20 sets of parameter combinations need to be searched.

Fig. 11(a) and (c) shows the effect of CNN with different hyperparameters for PRN1 and PRN11 inversion on the validation set. Again, R^2 is utilized as a reference standard to measure the goodness of the model. It can be roughly seen that the effect of different convolution kernel sizes on the inversion accuracy is less obvious. When the convolution kernel size is larger than 5×1, the inversion accuracy does not improve significantly as the convolution kernel size continues to increase. As the number of convolution kernels increases, the overall inversion accuracy tends to increase. When the number of convolution kernels reaches 128 and 256, the accuracy of the model with some convolution kernel size settings starts to fluctuate and does not improve significantly. Fig. 11(b) and (d) shows the time spent on training the model for CNN with different hyperparameters. The model training time is not sensitive to the size of the convolution kernels, but increases with the number of convolution kernels. The R^2 of PRN1 is maximum at the convolutional kernel size and number of 5×1 and 128, respectively, with a value of 99.65%. The R^2 of PRN11 is maximum at the convolutional kernel size and number of 5×1 and 256, respectively, with a value of 99.43%. Considering the apparent increase in time overhead, which does not bring a significant improvement in accuracy, the number of convolution kernels for PRN11 can be reduced to a choice of 128. Therefore, the size and number of CNN convolution kernel are set as 5×1 and 128 respectively, which are more suitable for the dataset in this article. According to the subsequent analysis and verification of this article, the inversion performance of the CNN in this setting is equally well for other subdatasets.

Fig. 12 shows the performance of the CNN inversion model on PRN1 and PRN11. It can be clearly seen that the inversion results are in general agreement with those of the DTU validation model. The MAE, RMSE, and R^2 of PRN1 are 0.91 m, 1.48 m, and 99.58%, respectively; the MAE, RMSE, and R^2 of PRN11 are 0.91 m, 1.92 m, and 99.27%, respectively. Relative to the inversion results of the PCA-SVR inversion model, there are fewer points with larger deviation values.

C. Comparison of PCA-SVR and CNN Inversion Model

Fig. 13(a) presents the performance of the PCA-SVR model for inversion of SSH on all GPS satellite sub datasets. Fig. 13(b) shows the performance of the CNN model on all GPS satellite sub datasets. It can be clearly seen that both have relatively average inversion performance on all satellite subdatasets, with MAE below 1 m, RMSE basically not exceeding 2 m, and R^2 close to 1. Such performance initially indicates the effectiveness of PCA-SVR and CNN for TDS-1 SSH inversion applications. Compared with the DDM SSH inversion method [8], the inversion error of the SSH inversion model obtained based on PCA-SVR and CNN is smaller, which further validates the superiority of the method in this article. The modeling time of the PCA-SVR model is within 500 s, with large differences among subdatasets. In contrast, the modeling time of the CNN model is less than 1000 s, with little variation among the subdatasets. Referring to Fig. 2, the time overhead of the CNN model has a significant correlation with the size of the dataset, while the PCA-SVR model does not. This is because the computational complexity of convolution and back propagation in CNN is proportional to the amount of data in the training set; whereas

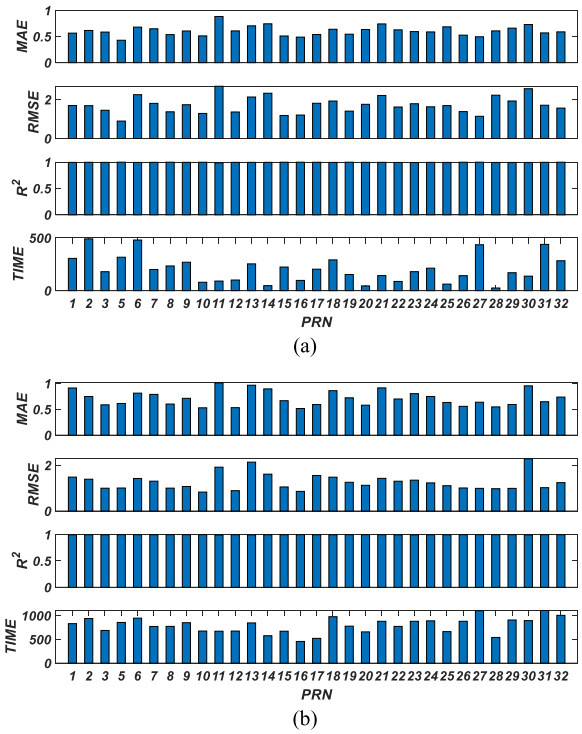


Fig. 13. Performance of two ML models on data from 31 GPS satellites. (a) PCA-SVR. (b) CNN.

TABLE II
AVERAGE OF THE INVERSION RESULTS OF THE 31 SUBDATASETS

Method	MAE (m)	RMSE (m)	R2 (%)	TIME (s)
PCA-SVR	0.61	1.72	99.56	205.84
CNN	0.71	1.27	99.76	796.91

the computational complexity of regression hyperplane in SVR depends on the number of support vectors, which is a subset of the training set.

Table II summarizes the average inversion performance on the 31 subdatasets of the two models. TIME in the table represents the average training time required to build the model. The MAE of the PCA-SVR model is 0.61 m, the RMSE is 1.72 m, the R^2 is 99.56%, and the time overhead is 205.84 s. The MAE of the CNN model is 0.71 m, the RMSE is 1.27 m, the R^2 is 99.76%, and the time overhead is 796.91 s.

In terms of model fit, the R^2 of the two are as high as 99.56% and 99.76%, respectively, which confirms that both can be successfully applied to SSH inversion studies. The MAE of the PCA-SVR model is 0.1 m lower than that of the CNN model, but the RMSE is almost 0.5 m higher. This indicates that the PCA-SVR model has a better average performance, but there are relatively more points with excessive deviation values in the inversion results; while the inversion performance of the CNN model is relatively more stable with fewer values of excessive deviation. Figs. 10 and 12 have shown the comparison of the inversion results with the validation results for the two

TABLE III
NUMBER OF SAMPLES IN DIFFERENT PERIODS

PRN	4.9-4.29	4.30
1	31228	1511
11	20332	1199

subdatasets, which is in accordance with the abovementioned analysis. The reason behind this phenomenon may be the introduction of relaxation factors in the SVR, which allows the SVR to ignore a small number of outlier points when building the regression hyperplane.

In terms of average time overhead, the CNN model is about four times longer than the PCA-SVR model, because CNN requires multiple iterations for training, while SVR builds the model directly according to support vectors. It is worth noting that this does not include the time for hyperparameter optimization. The time overhead of optimal parameter search is related to the range of parameters searched. The wider the parameter search range and the smaller the step size, the more time is spent. The disadvantage exists that the parameter range is usually set empirically, which is difficult to master.

PCA-SVR model and CNN model use multiple original physical quantities to retrieve SSH, and their main optimization scheme is the selection of hyperparameters. Compared with traditional method [4], [5], [8], the ML models can simplify the establishment of error models and improve the accuracy, which is due to the nonlinear fitting ability of ML methods.

D. Generalization Ability of ML Model

In order to further study the generalization ability of the above ML models, this article selects 21 continuous days of data from April 9-29, 2018 in the dataset for the establishment of the ML model for SSH retrieval, and then predicts the data following April 30. Similarly, taking PRN1 and PRN11 as examples, Table III lists the data sample size corresponding to each period. The proportion of samples participating in the training is 10%.

Fig. 14(a) and (b) shows the predictions of the PCA-SVR model and the CNN model for the April 30 data for PRN1 and PRN11, respectively. The errors are obtained by differencing the results of each ML model with the SSH values provided by the DTU validation model. The MAE of the prediction results of the PCA-SVR model is about 10 m on average, and the MAE of the prediction results of the CNN model is about 5 m on average. CNN has the advantage of deep learning capability, so it has better performance in prediction results than PCA-SVR model, however, compared with the results summarized in Table II, their accuracy of prediction for future data are both relatively poor overall. Compared with traditional method [4], [5], [8], the ML models are less generalizable. It may be due to that the data feature parameters and the number of samples are not enough, which leads to incomplete information for SSH retrieval.

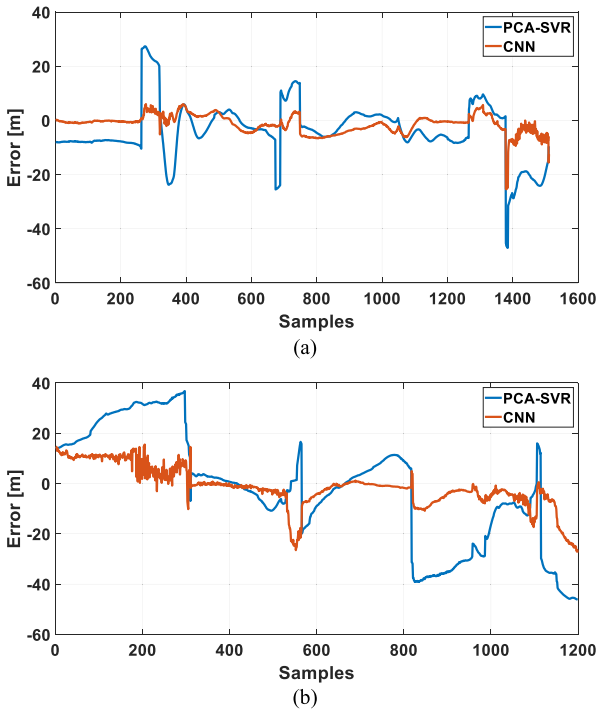


Fig. 14. Forecast errors of two ML models for the April 30 data. (a) PRN1. (b) PRN11.

VI. Conclusion

In this article, PCA-SVR and CNN were used to analyze the inverse relationship between TDS-1 data and SSH, and the corresponding SSH retrieval model was developed. The article aims to mine the SSH mapping relationships from the data itself. The spaceborne GNSS-R technology utilized has the great advantage of wide spatial and temporal coverage and low cost. Both SSH inversion models using PCA-SVR and CNN have exciting performance through independent analysis of data from different GPS satellites. In comparison with the SSH validation model, which is established by DTU mean SSH model and ocean tide model, the R^2 of both inversion models exceeds 99.5%. The MAE of the PCA-SVR model is relatively lower than that of the CNN model, and the RMSE of the PCA-SVR model is relatively higher than that of the CNN model. Therefore, the mean bias of the PCA-SVR model is lower, while the stability of the CNN model is somewhat higher. Moreover, the SSH inversion modeling using the PCA-SVR method is more time efficient in the dataset of this article.

Compared with traditional method for SSH retrieval, proposed ML methods have the advantages of not having to consider establishing error models, and the ability to make full use of the original physical quantities associated with the SSH, to get better accuracy. However, the ML models require a large amount of labeled data and should choose suitable features to build and train the models, if data characteristic parameters and the number of samples for ML model are not enough, generalization ability for SSH retrieval will be reduced. Currently, the combination of ML and GNSS-R is still in the exploratory stage. ML methods, while not providing a direct and clear explanation of the inversion process, provide a general insight that has been numerically analyzed and carefully validated. Such research is strongly recommended. In the future, as CYGNSS and other satellites would provide more and more sophisticated data, the accuracy of retrieval of SSH using ML techniques would likely be further improved.

Acknowledgment

The authors would like to thank the MERRByS website for the TDS-1 data and the DTU global tide model and the Danish Technical University for providing the DTU global mean sea level model. The authors would like to thank Prof. Y. Dongkai of Beijing University of Aeronautics and Astronautics and Dr. L. Weiqiang of CSIC-IEEC for their suggestions on GNSS-R satellite data analysis, and they also would like to thank Mr. Z. Bo and Dr. Q. Jin from Shanghai Institute of Aerospace Electronics for their suggestions on the receiver of reflected signals.

References

[1] M. Martin-Neira, "A passive reflectometry and interferometry system (PARIS): Application to ocean altimetry," *ESA J.*, vol. 17, no. 4, pp. 331–335, 1993.

[2] Y. Zhang, L. Tian, W. Meng, Q. Gu, Y. Han, and Z. Hong, "Feasibility of code-level altimetry using coastal beidou reflection (BeiDou-R) setups," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 4130–4140, Aug. 2015.

[3] Y. Zhang, B. Li, L. Tian, Q. Gu, Y. Han, and Z. Hong, "Phase altimetry using reflected signals from Beidou GEO satellites," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 10, pp. 1410–1414, Oct. 2016.

[4] M. P. Clarizia *et al.*, "First spaceborne observation of sea surface height using GPS-reflectometry," *Geophysical Res. Lett.*, vol. 43, no. 2, pp. 767–774, 2016.

[5] J. Mashburn, P. Axelrad, S. T. Lowe, and K. M. Larson, "Global ocean altimetry with GNSS reflections from TechDemoSat-1," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 4088–4097, Jul. 2018.

[6] W. Li, E. Cardellach, F. Fabra, S. Ribó, and A. Rius, "Assessment of spaceborne GNSS-R ocean altimetry performance using CYGNSS mission raw data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 238–250, Jan. 2020.

[7] H. Qiu and S. Jin, "Global mean sea surface height estimated from spaceborne Cyclone-GNSS reflectometry," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 356.

[8] Y. Zhang *et al.*, "Research on sea surface height inversion of GPS reflected signal based on TechDemoSat-1," *J. Beijing Univ. Aeronaut. Astronaut.*, vol. 47, no. 10, pp. 1941–1948, 2021.

[9] J. Mashburn, P. Axelrad, C. Zuffada, E. Loria, A. O'Brien, and B. Haines, "Improved GNSS-R ocean surface altimetry with CYGNSS in the seas of Indonesia," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6071–6087, Sep. 2020.

[10] W. Li *et al.*, "First spaceborne phase altimetry over sea ice using TechDemoSat-1 GNSS-R signals," *Geophysical Res. Lett.*, vol. 44, no. 16, pp. 8369–8376, 2017.

[11] W. Li *et al.*, "Lake level and surface topography measured with spaceborne GNSS-Reflectometry from CYGNSS mission: Example for the lake Qinghai," *Geophysical Res. Lett.*, vol. 45, 2018, Art. no. 24.

[12] Q. Yan and W. Huang, "Sea ice sensing from GNSS-R data using convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1510–1514, Oct. 2018.

[13] M. Asgarimehr, I. Zhelavskaya, G. Foti, S. Reich, and J. Wickert, "A GNSS-R geophysical model function: Machine learning for wind speed retrievals," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1333–1337, Aug. 2019.

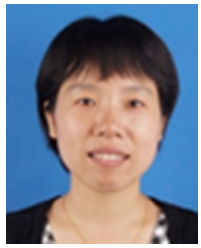
[14] X. Li *et al.*, "Analysis of coastal wind speed retrieval from CYGNSS mission using artificial neural network," *Remote Sens. Environ.*, vol. 260, 2021, Art. no. 112454.

[15] V. Senyurek *et al.*, "Machine learning-based CYGNSS soil moisture estimates over ISMN sites in CONUS," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1168.
[16] Q. Wang *et al.*, "A new GNSS-R altimetry algorithm based on machine learning fusion model and feature optimization to improve the precision of sea surface height retrieval," *Front. Earth Sci.*, vol. 9, 2021, Art. no. 758.
[17] H. Hotellings, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, pp. 417–441, 1933.
[18] R. G. Lloyd and G. R. Brereton, "Support vector machines for classification and regression," *Analyst*, vol. 135, no. 2, pp. 230–267, 2010.
[19] H. Tayara, K. G. Soo, and K. T. Chong, "Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network," *IEEE Access*, vol. 6, pp. 2220–2230, 2017.
[20] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, 2019.
[21] J. Tye, P. Jales, M. Unwin, and C. Underwood, "The first application of stare processing to retrieve mean square slope using the SGR-ReSI GNSS-R experiment on TDS-1," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4669–4677, Oct. 2017.
[22] G. Giangregorio, M. di Bisceglie, P. Addabbo, T. Beltramonte, S. D'Addio, and C. Galdi, "Stochastic modeling and simulation of delay–Doppler maps in GNSS-R over the ocean," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2056–2069, Apr. 2016.
[23] G. Piccioni, O. B. Andersen, and L. Stenseng, "Sentinel-3 for science workshop," in *Proc. Workshop*, Ed. L. Ouwehand, Venice, Italy, ESA SP-734, Jun. 2015.
[24] L. G. James and J. K. Stephen, "The application of reflected GPS signals to ocean remote sensing," *Remote Sens. Environ.*, vol. 73, pp. 175–187, 2000.
[25] J. Mashburn, P. Axelrad, S. T. Lowe, and K. M. Larson, "An assessment of the precision and accuracy of altimetry retrievals for a monterey bay GNSS-R experiment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 10, pp. 4660–4668, Oct. 2016.
[26] V. U. Zavorotny and A. G. Voronovich, "Scattering of GPS signals from the ocean with wind remote sensing application," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 2, pp. 951–964, Mar. 2000.
[27] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on Riemannian manifolds with Gaussian RBF kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2464–2477, Dec. 2015.
[28] S. Park and T. Suh, "Speculative backpropagation for CNN parallel training," *IEEE Access*, vol. 8, pp. 215365–215374, 2020.
[29] J. Hua *et al.*, "Proton exchange membrane fuel cell system diagnosis based on the multivariate statistical method," *Int. J. Hydrogen Energy*, vol. 36, pp. 9896–9905, 2011.
[30] W. Cao, X. Liu, and J. Ni, "Parameter optimization of support vector regression using henry gas solubility optimization algorithm," *IEEE Access*, vol. 8, pp. 88633–88642, 2020.

Yun Zhang received the Ph.D. degree in applied marine environmental studies from the Tokyo University of Maritime Science and Technology, Tokyo, Japan, in 2008.

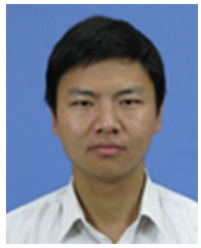
From 2011, he is a Professor with the College of Information and Technology, Shanghai Ocean University, Shanghai, China. His research interests include the study of navigation system reflection signal technique and its maritime application.

Shen Huang received the B.S. degree in computer science and technology from Nantong University, Nantong, China, in 2018. He is currently working toward the M.E. degree in computer technology from Shanghai Ocean University, Shanghai, China.



Yanling Han received the B.E. degree in mechanical design and manufacturing and the M.E. degree in mechanical automation from Sichuan University, Sichuan, China, and the Ph.D. degree in engineering and control theory from Shanghai University, Shanghai, China, in 1996, 1999, and 2005, respectively.

She is currently a Professor and with the Shanghai Ocean University, Shanghai, China. Her research interests include the study of ocean remote sensing, flexible system modelling, and deep learning.



Shuhu Yang received the Ph.D. degree in physics of physics from School of Physics, Nanjing University, Nanjing, China, in 2012.

Since September 2012, he has been the Lecturer with the College of Information Technology, Shanghai Ocean University, Shanghai, China. His research interests include evolution of the Antarctic ice sheet, hyperspectral remote sensing, and the use of navigational satellite reflections.

Zhonghua Hong received the Ph.D. degree in GIS from Tongji University, Shanghai, China, in 2014.

Since 2019, he has been an Associate Professor with the College of Information Technology, Shanghai Ocean University, Shanghai, China. His research interests include three-dimensional damage detection, coastal mapping, photogrammetry, GNSS-R, and deep learning.

Dehao Ma received the B.S. degree in electronic commerce major from Shandong University of Finance and Economics, Jinan, China, in 2017, and the M.E. degree in engineering in computer technology from Shanghai Ocean University, Shanghai, China, in 2021.

He has been engaged in GNSS-R research since 2018.

Wanting Meng received the B.S. degree in spatial information and digital technology and the M.S. degree in software engineering from Shang Ocean University, Shanghai, China, in 2013 and 2016, respectively.

She is currently with Shanghai Spaceflight Institute of TT&C and Telecommunication, Shanghai, China, where she is currently a Research Associate in the field of microwave radiometer calibration techniques and GNSS-R remote sensing.