MST-Net: A General Deep Learning Model for Thick Cloud Removal from Optical Images

Lanxing Wang, Qunming Wang, Xiaohua Tong, and Peter M. Atkinson

***Abstract*—Temporally neighboring homologous images are crucial to provide auxiliary information for thick cloud removal. Due to the inherent satellite revisit period and frequent cloud obscuration, there is often a significant time interval between the target cloudy images and neighboring cloud-free homologous images, leading to potential land surface condition changes. Moreover, multi-temporal cloudy images that may contain valuable complementary information in the non-cloudy regions, are often neglected in practice. This paper focused on thick cloud removal from Landsat 8 OLI images. We proposed to fuse the temporally more frequent Sentinel-2 MSI images and also cloudy multi-temporal images consisting of Sentinel-2 MSI and Landsat 8 OLI time-series. Acquired by a sensor different from Landsat 8 OLI, Sentinel-2 MSI images exhibit great similarities in data characteristics. To fully exploit the spatio-temporal-spectral information in Multi-Source and multi-Temporal auxiliary images, we proposed a novel deep Network called MST-Net. MST-Net was validated using 12 simulated and two real cloudy Landsat 8 OLI images. The results show that the MST-Net can produce more satisfactory predictions than five benchmark methods. Both the images acquired by a different sensor and homogeneous multi-temporal cloudy images are beneficial. Under different sizes of clouds, the MST-Net produces consistently the most accurate predictions. Furthermore, due to the fusion of all bands simultaneously in the temporally closest Sentinel-2 MSI images, the MST-Net is less affected by thin cloud occlusion errors. Overall, the MST-Net shows great potential for cloud removal from optical images produced by a wide range of sensors and, more generally, filling gaps in various global scale products.**

***Index Terms*—Cloud removal, gap filling, deep learning, Landsat 8, Sentinel-2**

I. INTRODUCTION

Cloud, and cloud shadow, contamination is a common issue in optical remote sensing images [1], [2], [3], [4]. At the global scale, the cloud coverage exceeds 60% at any time [5]. Cloud removal (also known as gap filling or reconstruction) is crucial for using cloud-contaminated remote sensing images effectively. In general, clouds are divided into thick and thin clouds. For thin clouds, part of the information under them can be used for reconstruction [6], [7], [8], [9], [10]. However, thick clouds can cause complete loss of land cover information (i.e., spatial gaps), making their recovery more challenging. In this paper, we focus on thick cloud removal, and all clouds mentioned hereafter refer to thick clouds. To clarify, this research is based on known cloud masks and does not focus on cloud detection process [11].

Up to now, a wide range of cloud removal methods have been developed. Existing methods can be roughly divided into four categories, namely, spatial-based methods [12], [13], temporal-based methods [14], [15], spatio-temporal-based methods [16], [17], [18] and machine learning-based methods [19], [20], [21]. First, spatial-based methods are performed based on the assumption that there exists a link between spatially adjacent pixels. Specifically, spatial-based methods fill in the gaps by using the non-cloud data of the cloudy image itself. For example, Maalouf et al. [22] used the geometric flow curves derived from the Bandelet transform of the non-cloudy area to guide prediction of the missing areas. Spatial-based methods typically assume that for cloud-contaminated regions, the spatially closer non-cloud information is more relevant. Thus, the reliability of gap filling will be greatly reduced when the cloud coverage is large. Therefore, spatial-based methods are more suitable for removing small clouds and reconstruction of homogeneous regions. Second, temporal-based methods use cloud-free images acquired at times close to the image of interest to provide auxiliary information [23], [24], [25]. Generally, the relation between the auxiliary images and cloudy images is fitted through the non-cloudy data between them. Based on the fitted relation, the final prediction is produced using the effective data in the auxiliary image corresponding to the target cloudy region. Lorenzi et al. [26] developed a compressive sensing-based solution that assumes a similar spatial structure between the cloudy and auxiliary images. Lin et al. [27] cloned information from cloud-free patches to corresponding cloud patches based on the temporal correlation of multi-temporal images. Considering possible changes in land cover, temporal-based methods tend to select auxiliary images that are temporally closest to the target cloudy image. Third, spatio-temporal-based methods combine the advantages of the previous two methods by fully exploiting the effective spatio-temporal information of the cloudy and auxiliary images. The modified neighborhood similar pixel interpolator (i.e., MNSPI) approach [28] is a classical method of this type. The prediction is a linear combination of spatial prediction (spatial interpolation of spectra of neighboring similar pixels) and temporal prediction (spatial interpolation of temporal changes of the neighboring similar pixels). Similarly, Chen et al. [29] proposed a spatially and temporally weighted regression method for Landsat image cloud removal. Tang et al. [18] reconstructed Landsat images with large gaps using spectral-temporal metrics computed from one-year of Landsat images. Fourth, machine learning-based cloud removal methods have been developed rapidly recently, due to the continuous development of computing power and successive enrichment of datasets.

This work was supported by the National Natural Science Foundation of China under Grants 42222108, 42221002 and 42171345. (*Corresponding author: Q. Wang*.)

L. Wang, Q. Wang, and X. Tong are with the College of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road, Shanghai 200092, China (e-mail: wqm11111@126.com).

P.M. Atkinson is with the Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YR, UK and also with Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

Amongst these methods, deep learning methods with strong fitting ability for complex nonlinear relationships have received increasing attention [30], [31]. Malek et al. [32] proposed an autoencoder neural network (AE) for cloud removal. Zhang et al. [33] proposed a convolutional neural network (CNN) that uses spatio-temporal-spectral information for gap filling, namely, STS. Moreover, based on a loss function considering both global consistency and local particularity, Zhang et al. [34] adopted the weighted aggregation and progressive iteration to reconstruct multi-temporal cloudy images (i.e., PSTCR). Sebastianelli et al. [35] employed a generative adversarial network (GAN) to convert SAR data directly into optical images (i.e., PLFM), and integrate this intermediate result with time-series images to generate the final prediction.

Due to the complete loss of information under thick clouds, proper selection and use of temporal auxiliary data is important for gap filling, as done in existing temporal- and spatio-temporal-based methods. Normally, auxiliary images temporally closer to the target cloudy image are a preferable choice, as shorter time intervals generally correspond to fewer and smaller land cover changes [36]. Generally, homological images have been used to provide auxiliary information for gap filling, because of the same wavelength settings, spatial resolution and coordinate system. However, cloud occlusion is spatially extensive and long-term, and the revisit period of satellites is relatively long. Therefore, the use of homologous auxiliary data faces the problem of large time interval from the target cloudy images, reducing the value of the auxiliary data. In this case, it may be worthwhile to explore the use of multi-source auxiliary images with finer temporal resolution, which can provide auxiliary data affected less by land cover changes.

In this paper, we focused on cloud removal of Landsat 8 OLI images [37], [38], [39], [40] with a coarse temporal resolution of 16 days. As a source of potential heterogeneous auxiliary data, Sentinel-2 MSI images [41], [42], [43], [44], [45], [46] have a temporal resolution of 5 days. The frequency is three times finer compared with Landsat 8 OLI images, making it more possible to provide images temporally closer to the target cloudy Landsat 8 OLI images. More importantly, Sentinel-2 MSI images have the same projected coordinates (when acquired in the same region) and similar spectral settings as Landsat 8 OLI images [47], [48], [49]. For example, the Sentinel-2 MSI blue, green, red, vegetation red edge (i.e., VRE, the 8A band), SWIR 1 and SWIR 2 bands have almost the same central wavelength settings as those of the Landsat 8 OLI blue, green, red, NIR, SWIR 1 and SWIR 2 bands, respectively. Moreover, the spatial resolution of Sentinel-2 MSI images is 10 m for blue, green and red bands and 20 m for VRE, SWIR 1 and SWIR 2 bands, which is comparable to (or even finer than) that of Landsat images. Therefore, Sentinel-2 MSI images hold significant promise for gap filling of Landsat 8 OLI cloudy images. In this paper, we proposed to use multi-source Landsat 8 OLI and Sentinel-2 MSI images for cloud removal of Landsat 8 OLI images.

It is worth noting that most existing methods often use the temporally closest cloud-free images as auxiliary data. That is, the multi-temporal images contaminated by cloud are abandoned directly, even if they contain partial (particularly a large number of) effective data and are temporally closer to the target cloudy images than the auxiliary cloud-free image. Actually, the partial effective data in the multi-temporal images can be of great value in the cloud removal task, as they may provide important auxiliary information for the target cloud areas. In this paper, we also considered the use of multi-temporal cloudy images as auxiliary data. In fact, the clouds in the multi-temporal images are often spatially staggered because of their mobility. The key issue is how to fully exploit the complementary effective information in the remaining non-cloudy regions to facilitate more accurate cloud removal.

In this paper, we proposed to use the multi-source and multi-temporal images jointly as auxiliary data for cloud removal. Compared to traditional cloud removal models, deep learning-based methods can effectively explore information from large amounts of training data due to their powerful learning and nonlinear fitting abilities. For cloud removal, however, few existing deep learning methods consider the fusion of multi-source images and partially cloud-contaminated multi-temporal images. The key to fusion of these data for cloud removal is to fully exploit the rich spatio-temporal-spectral information in the auxiliary data. To this end, we proposed a deep learning-based method to fuse Multi-Source and multi-Temporal images, namely, MST-Net. To avoid underfitting and facilitate the training process, the use of multi-source and multi-temporal images is decomposed into two stages. Accordingly, the MST-Net consists of two networks: MS-Net and MT-Net. Specifically, MS-Net exploits the spatio-spectral information from Multi-Source images, while MT-Net exploits the spatio-temporal information from Multi-Temporal images. Considering that the temporally closest auxiliary image tends to contain the most valuable auxiliary information, MS-Net uses the temporally more frequent Sentinel-2 MSI data to preliminarily reconstruct the target cloudy image. Then, MT-Net further integrates the spatio-temporal information from time-series images composed of multi-temporal cloudy Sentinel-2 and Landsat images. The contributions of this paper are summarized as follows:

1) We proposed to use the temporally closer (compared with homologous Landsat 8 OLI auxiliary images) Sentinel-2 MSI images for cloud removal of Landsat 8 OLI images, taking full advantage of the similarities between these two types of images.
2) We explored the temporally close cloudy time-series (composed of cloudy Sentinel-2 and Landsat images) to make full use of the valuable information in the non-cloudy regions of the multi-temporal images.
3) We proposed the MST-Net to effectively exploit the spatio-temporal-spectral information from the auxiliary multi-source and multi-temporal images for cloud removal, by designing a two-stage network composed of MS-Net and MT-Net.

The remainder of the paper is structured as follows. The proposed MST-Net is detailed in Section II. In Section III, we conducted experiments based on simulated and real clouds to demonstrate the effectiveness of the proposed MST-Net. Section IV further discusses the effectiveness of MST-Net, its potential capabilities, and limitations. Section V concludes the paper.

II. METHODS

A. Overview of the proposed MST-Net

Considering their spectral similarity, the bands of interests in this paper are the Landsat 8 OLI blue, green, red, NIR, SWIR 1 and SWIR 2 bands, corresponding to the Sentinel-2 MSI blue, green, red, vegetation red edge (i.e., VRE, the 8A band), SWIR 1 and SWIR 2 bands, respectively. The proposed MST-Net consist of MS-Net and MT-Net, which are designed specifically to effectively extract spatio-temporal-spectral information from auxiliary multi-source and multi-temporal images through a process of decomposition and gradual integration. First, it is widely accepted that the auxiliary image temporally closest to the target cloudy image usually contains the most valuable information, which deserves to be exploited with priority. Considering this, MS-Net fuses the six bands of the temporally closest Sentinel-2 MSI image simultaneously to preliminarily reconstruct the target cloudy Landsat 8 OLI image. MS-Net is composed mainly of fully connected layers. Since each neuron is connected to all the neurons in the previous layer, the fully connected layer is able to capture effectively the relevant information in the input data, thus, fully exploiting the multi-spectral information in the temporally closest auxiliary image. Based on the output of MS-Net, MT-Net further integrates spatio-temporal information from the non-cloud data of the multi-temporal Sentinel-2 and Landsat images for final prediction. MT-Net is a deep CNN where each neuron is only convolved with a small local region of the input data, which enables it to efficiently learn the local correlations of the original data. Note that MS-Net and MT-Net are trained separately, with the former reconstructing all bands of the target cloudy image simultaneously and the latter reconstructing the cloudy image band by band. In summary, the MST-Net combines the strengths of MS-Net and MT-Net to efficiently exploit the spatio-temporal-spectral information in the auxiliary multi-source and multi-temporal images. The overall framework of the proposed MST-Net is illustrated in Fig. 1, in which band bn denotes the Sentinel-2 MSI or Landsat 8 OLI band with the closest central wavelength setting.

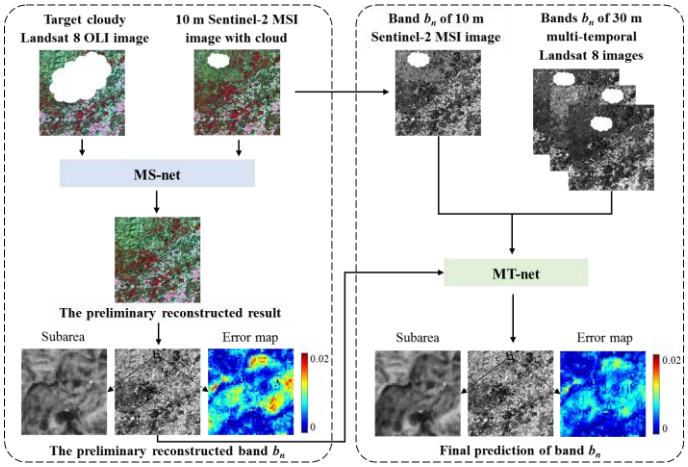


Fig. 1. Flowchart of the proposed MST-Net (prediction of a single band bn as an example).

It is worth noting that before being fed into MS-Net and MT-Net, the original 20 m Sentinel-2 bands (i.e., VRE, SWIR 1 and SWIR 2) need to be downscaled to 10 m (e.g., by fusion with 10 m observed bands). Amongst the various fusion methods available for downscaling Sentinel-2 data [50], [51], [52], we have selected the area-to-point regression kriging method (ATPRK) [53] due to its simplicity and demonstrated accuracy.

B. Fusion of temporally closest multi-source data (MS-Net)

MS-Net is designed to fuse spatio-spectral information from the temporally closest multi-source images (i.e., the auxiliary Sentinel-2 MSI and target Landsat 8 OLI cloudy images) for preliminary reconstruction. MS-Net consists of one CNN layer and five fully connected layers, three of which are connected to the activation function ReLU, as shown in Fig. 2. In this paper, Sentinel-2 MSI images were used in MS-Net, with a spatial resolution of 10 m (note that the 20 m original VRE, SWIR 1 and SWIR 2 bands of Sentinel-2 were downscaled to 10 m by ATPRK, as mentioned in Section II-A).

ATPRK is first used to downscale the three 20 m Sentinel-2 MSI bands to 10 m by fusing with the three 10 m bands. Then, the produced six 10 m bands are fed into MS-Net. In MS-Net, the prediction (i.e., output of the network) is a single Landsat pixel. Thus, for MS-Net, the unit input patch is six bands of 3×3 Sentinel-2 pixels. Based on the identified cloud masks, the common non-cloudy regions between the auxiliary and target cloudy images are used to construct a loss function based on the mean absolute error to guide the model training:

$$Loss_1 = \| (\mathbf{1} - \mathbf{M}_u) \odot \left(f_1(\mathbf{K}_1) - \mathbf{C} \right) \|_1 \quad (1)$$

where $Loss_1$ is the overall loss of the MS-Net training, and $\| \|_1$ represents the 1-norm. $\mathbf{C}$ and $\mathbf{K}_1$ represent the cloudy and temporally closest auxiliary images, respectively. $\mathbf{M}_u$ is the cloud mask union of $\mathbf{C}$ and $\mathbf{K}_1$ stored as a 0-1 matrix, where 1 represents the cloud pixel, and $\mathbf{1}$ represents an all-1 matrix with the same size of $\mathbf{M}_u$. $\odot$ represents the point multiplication operation between two matrices, and f_1 is the MS-Net model. Finally, the effective data in auxiliary images corresponding to the cloud regions in the target cloudy images are fed into the trained MS-Net to obtain a preliminary prediction of the target cloudy image. In this paper, the final training epoch (set to 60 epochs in this paper) is determined based on the loss curve. The initial learning rate of MS-Net was set to 1, and then it was set to decay to 95% of the last value every 10 epochs. MS-Net reconstructs all target cloudy bands simultaneously.

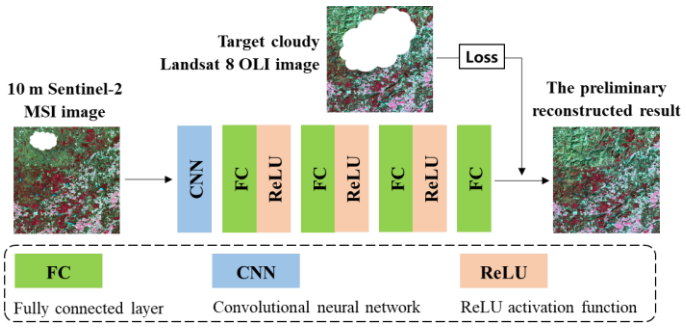


Fig. 2. Structure of the proposed MS-Net module.

C. Fusion of multi-temporal data (MT-Net)

Based on the preliminary reconstruction results obtained in Section II-B, MT-Net further fuses the multi-temporal cloudy

images to make the final prediction. The multi-temporal images used in this paper include 10 m Sentinel-2 MSI bands and three 30 m Landsat 8 OLI images. As shown in Fig. 3, MT-Net consists of 15 CNN layers where all, but three input- and one output-CNN layers are connected with a ReLU activation function. For each image, the unit input of MT-Net is a single band with $W \times W$ pixels. To strengthen the learning of the target cloud region while maintaining the consistency between the reconstructed and non-cloudy regions, a loss function [34] is designed for MT-Net, as shown in Eq. (2):

$$Loss_2 = Loss_{\text{global}} + \lambda \cdot Loss_{\text{target}} \quad (2)$$

where $Loss_{\text{global}}$ is the loss term considering global consistency, $Loss_{\text{target}}$ is the loss term measuring the accuracy of local details, and λ is a weight (set to 0.15 in this paper). The specific expressions of $Loss_{\text{global}}$ and $Loss_{\text{target}}$ are shown in Eqs. (3) and (4), respectively:

$$Loss_{\text{global}} = \| f_2(\mathbf{P}(b_m), \mathbf{K}_1(b_m), ..., \mathbf{K}_n(b_m)) - \mathbf{C}(b_m) \|_F \quad (3)$$

$$Loss_{\text{target}} = \| \mathbf{M} \odot (f_2(\mathbf{P}(b_m), \mathbf{K}_1(b_m), ..., \mathbf{K}_n(b_m)) - \mathbf{C}(b_m)) \|_F \quad (4)$$

where $\| \|_F$ represents the Frobenius norm computing the square root sum of all elements within a matrix. f_2 represents the MT-Net model. $\mathbf{K}_1(b_m)$, …, $\mathbf{K}_n(b_m)$ are the multi-temporal images in band b_m from $\mathbf{K}_1$ to $\mathbf{K}_n$, whose acquisition date moves away from the target cloudy image gradually. $\mathbf{M}$ is the cloud mask of the target cloudy image. n is the number of multi-temporal images and in this paper n was set to 4. $\mathbf{P}(b_m)$ represents the MS-Net output in band b_m, that is, $\mathbf{P} = f_1(\mathbf{K}_1)$. For each target cloudy band, the corresponding MT-Net is trained for 100 epochs (decided through the loss curve, and the training-validation ratio in this paper is about 9:1), with the initial learning rate set to 0.001 and decayed to 80% of the previous value every 20 epochs. Furthermore, considering the learning efficiency and accuracy, W was set to 40 in this paper. Using the MT-Net trained for each band, all bands are reconstructed in turn.

D. Implementation of the proposed MST-Net

Implementation of the full MST-Net consists of the following steps:

1) Training:
   - 1.1) The three 20 m Sentinel-2 MSI bands are downscaled to 10 m by fusion of the three 10 m bands using ATPRK.
   - 1.2) The six 10 m Sentinel-2 MSI bands obtained in Step 1.1) are fed into MS-Net for model training. The non-cloud, effective information in the target cloudy image is used to constrain the MS-Net output through the loss function defined in Eq. (1).
   - 1.3) The effective data from the 10 m Sentinel-2 MSI image in Step 1.1) (corresponding to the target cloud region) are input into the trained MS-Net in Step 1.2) to generate a preliminary reconstructed image.
   - 1.4) Both the same band of preliminary reconstructed image in Step 1.3) and multi-temporal images are fed into MT-Net, and the model training is guided by the loss function in Eq. (2). This step is repeated for each band in turn.
2) Predicting:
   - 2.1) Steps 1.1)-1.3) are performed based on target cloudy and Sentinel-2 MSI images in the testing data to obtain a preliminary reconstructed image.
   - 2.2) Each single band from the preliminary reconstructed image in Step 2.1) and corresponding multi-temporal images are input into the MT-Net trained in Step 1.4). The procedure is repeated for each band in turn to produce the final cloud removal result.

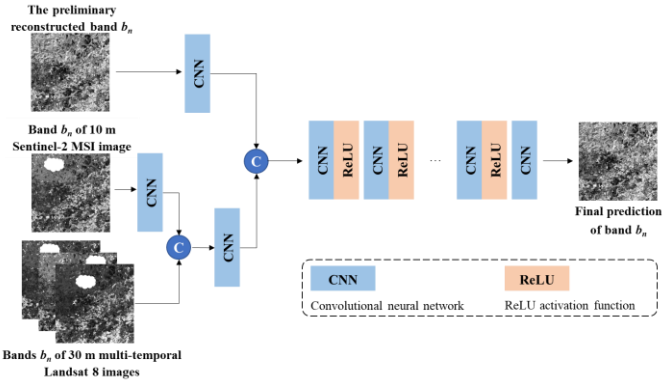


Fig. 3. Structure of the proposed MT-Net module (prediction of a single band as an example).

III. Experiments

A. Data and experimental design

In this section, the performance of the proposed MST-Net method was examined with simulated (Sections III-B to G) and real (Section III-H) clouds using the Collection 2 Landsat analysis ready surface reflectance data [54]. For simulated cloud experiments, we used the quality control files of real cloudy images (obtained from USGS) to generate known cloud masks (0-1 matrices) to simulate areas contaminated by thick clouds. These known cloud masks are employed to enable the network to recognize non-cloud data (corresponding to 0-values on the mask) during model training. The simulated cloud pixels (not involved in the training process) were originally observed and can be used to conduct a comprehensive quantitative and qualitative evaluation of the reconstruction results. Specifically, the clouds were simulated in cloud-free images by applying cloud masks of real cloudy images acquired on a different date. The experimental datasets for cloud simulation consist of two parts, one for training and the other for testing, both of which are originally cloud-free images. Due to the particularity of the auxiliary image used in this paper (multi-source and multi-temporal), the existing dataset is inadequate. Therefore, we collected data at a global scale for model training and testing, as illustrated in Fig. 4, where each dot represents the geographic location of a target cloudy image and four time-series auxiliary images (including one Sentinel-2 MSI and three Landsat 8 OLI images). Specifically, deep learning-based (or machine learning-based) models were trained using this global dataset from 13 regions (red dots in Fig. 4) and tested in 14 regions worldwide (triangles in Fig. 4; yellow represents simulated cloudy regions and blue represents real cloudy regions) in this paper. Four test regions (i.e., Regions 1-4) were used for the central experiments and to display the results (Sections III-B to H). Real cloud removal experiments were conducted in two

regions (blue triangles in Fig. 4), as detailed in Section III-I. Fig. 5 shows the original cloud-free images used for simulating the target cloudy images in Regions 1-12, all with a spatial size of 1000×1000 Landsat pixels. Fig. 6 shows the time-series auxiliary images (composed of Sentinel-2 MSI and Landsat 8 OLI) with simulated clouds in Regions 1-4.

B. Comparison with benchmark methods

In the experiments, we compared the proposed MST-Net with several benchmark methods, including three deep learning-based methods (i.e., PLFM, PSTCR and STS), one machine learning-based method (i.e., AE), and one classical non-deep learning-based method (i.e., MNSPI). Amongst them, PLFM, MST-Net and PSTCR use time-series data for training and predicting, while other methods use only the auxiliary image temporally closest to the target cloudy image (that is, the Sentinel-2 MSI image) for model training. The simulated cloud removal experiments were carried out in 12 regions. Considering the space limitation, the results of four of them (Regions 1-4) were evaluated in various respects. Fig. 7 displays the reconstruction results of the six methods in Regions 1-4. Two subareas of the predictions in Fig. 7 (Subareas 1 and 2 are marked in red and blue in the first line) are shown in Fig. 8 to facilitate clearer comparison between the six methods.

It can be seen from Fig. 8 that land cover details are blurred in the STS predictions. Moreover, the STS prediction presents noticeable hue abnormalities in the first subarea of Region 4. The AE and MNSPI predictions exhibit noticeable noise, particularly in the second subarea of Region 1 where both predictions show evident distortions in the blue objects. Moreover, there are tonal deviations in the AE and MNSPI predictions. For example, in the first subarea of Region 4, the shadows of the dark green objects are missed. Compared with STS, AE and MNSPI, PSTCR reconstructs the land cover texture more accurately, but there is still color distortion. For example, dark blue objects in the second subarea of Region 1 are predicted as light blue by PSTCR, while both subareas of Region 3 are predicted as whitened overall, which differs greatly from the reference. Similarly, the PLFM predictions also exhibit some tonal anomalies in different regions. For example, the color of the red objects in two subareas of Region 4 appears notably brighter than the reference. In contrast, the proposed MST-Net restores the texture and tone of land cover more satisfactorily, with the predictions closest to the reference amongst all methods. The reason is that for MST-Net, the mining of effective information in the temporally closest auxiliary images, coupled with its comprehensive utilization of multi-source time-series data, enhances its adaptability and enables the achievement of optimal results across diverse regions.

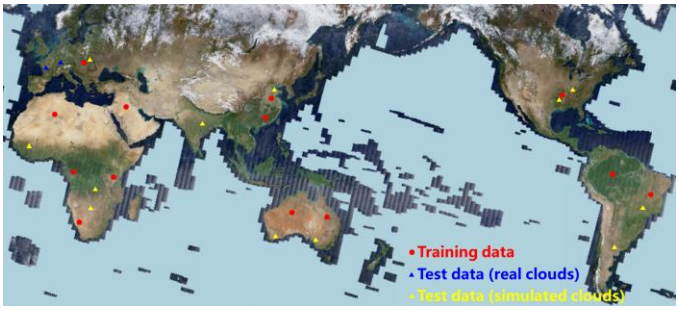


Fig. 4. Global distribution of training and testing data (downloaded from the Copernicus Browser).

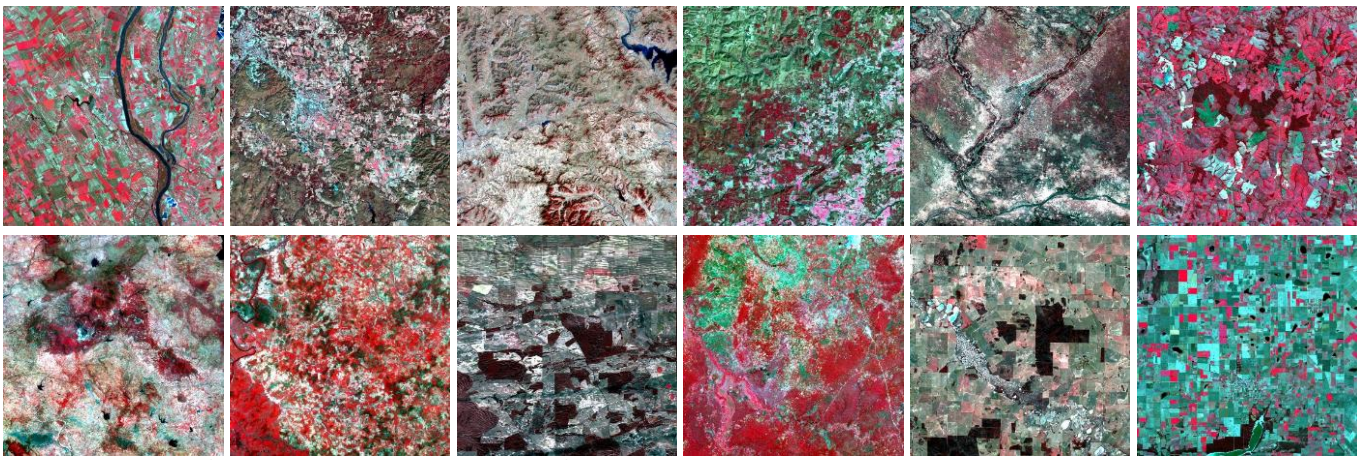


Fig. 5. The original cloud-free Landsat images (each with a spatial of 1000×1000 pixels; bands NIR, red and green as RGB) used to simulate the target cloudy image in Regions 1-12 (from left to right and top to bottom).

To further compare the predictions of the six cloud removal methods, the error maps of two subareas in Regions 3 and 4 are shown in Fig. 9 (showing the blue and NIR bands as a demonstration), where the 0-value pixels are set to black. From Fig. 9, it can be seen that the overall error of the MST-Net prediction is smaller than that of the other five methods. Specifically, the number of pixels with large errors (shown as blue-green and red) in the MST-Net predictions in all bands is the least, and the MST-Net prediction contains more 0-value pixels, especially in the subarea of Region 4. Moreover, Fig. 9 clearly indicates that the errors of all methods in the NIR band surpass that in the blue band, which can be attributed to the distinctive characteristics of the NIR band. Specifically, the NIR band exhibits greater sensitivity to subtle variations in vegetation, and its temporal variability may be more intricate, posing a greater challenge for accurate reconstruction compared to the visible light bands. Fig. 10 shows the scatterplots of the predictions in Region 2. It can be seen that the MST-Net predictions present the most tightly clustered scatter and the greatest consistency with the y = x line. In contrast, the other

five predictions present scatters that are more diffuse than that of the MST-Net predictions. Moreover, there are obvious outliers in the scatterplots of the MNSPI predictions. Generally, the scatterplots show that the MST-Net can reconstruct cloudy bands most accurately.

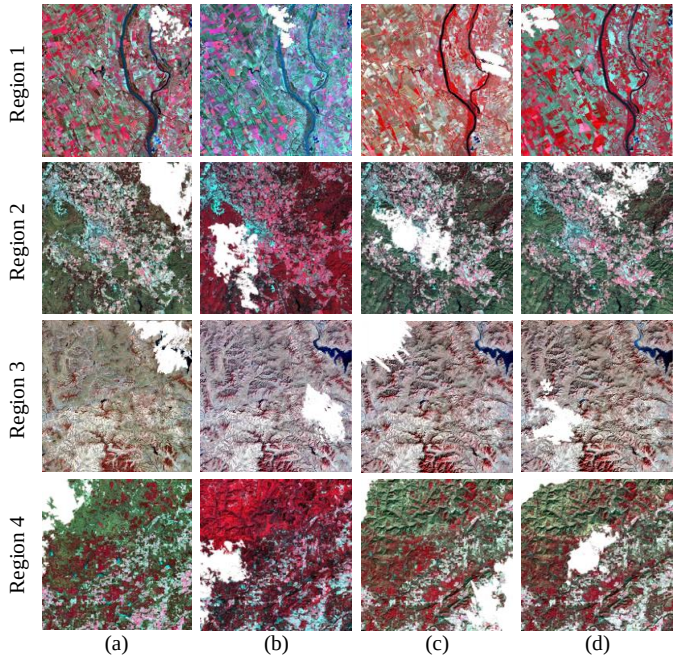


Fig. 6. The simulated partially cloud-contaminated time-series auxiliary images in Regions 1-4 (bands NIR, red, and green as RGB). From left to right, the acquisition date gradually moves away from the simulated target cloudy image. (a) Sentinel-2 MSI image. (b)-(d) are Landsat 8 OLI images.

For quantitative evaluation, the root mean square error (RMSE), universal image quality index (UIQI), correlation coefficient (CC) and spectral angle mapper (SAM) were used for all 12 regions, as shown in Table 1. It is evident that the MST-Net predictions generally yield superior RMSEs, CCs and UIQIs. For example, in Region 2, the average RMSE of the MST-Net prediction is 0.0027, 0.0083, 0.0055, 0.0049 and 0.0016 smaller than that of the PSTCR, PLFM, STS, AE and MNSPI predictions, respectively, while the average CC is 0.0393, 0.0661, 0.0831, 0.0454 and 0.0307 larger and the average UIQI is 0.0424, 0.0717, 0.0788, 0.0503 and 0.0335 larger. Furthermore, the performances of the six methods vary across different test regions. For example, in Regions 6 and 8, the prediction accuracies of the methods are relatively small, possibly due to obvious temporal changes in the auxiliary images in these regions. Amongst them, however, the MST-Net predictions still present the greatest accuracies, followed by the PSTCR predictions. Moreover, in Region 11, where there are small land cover changes between the target cloudy image and the temporally closest auxiliary image (i.e., Sentinel-2 MSI image), each method can achieve more accurate predictions with relatively small difference in accuracies. In this case, MST-Net and MNSPI perform best, followed by PSTCR, PLFM and STS. Overall, these results indicate that MST-Net generally produces more accurate predictions with stable performance and presents superior generalization ability when applied to globally sampled test datasets.

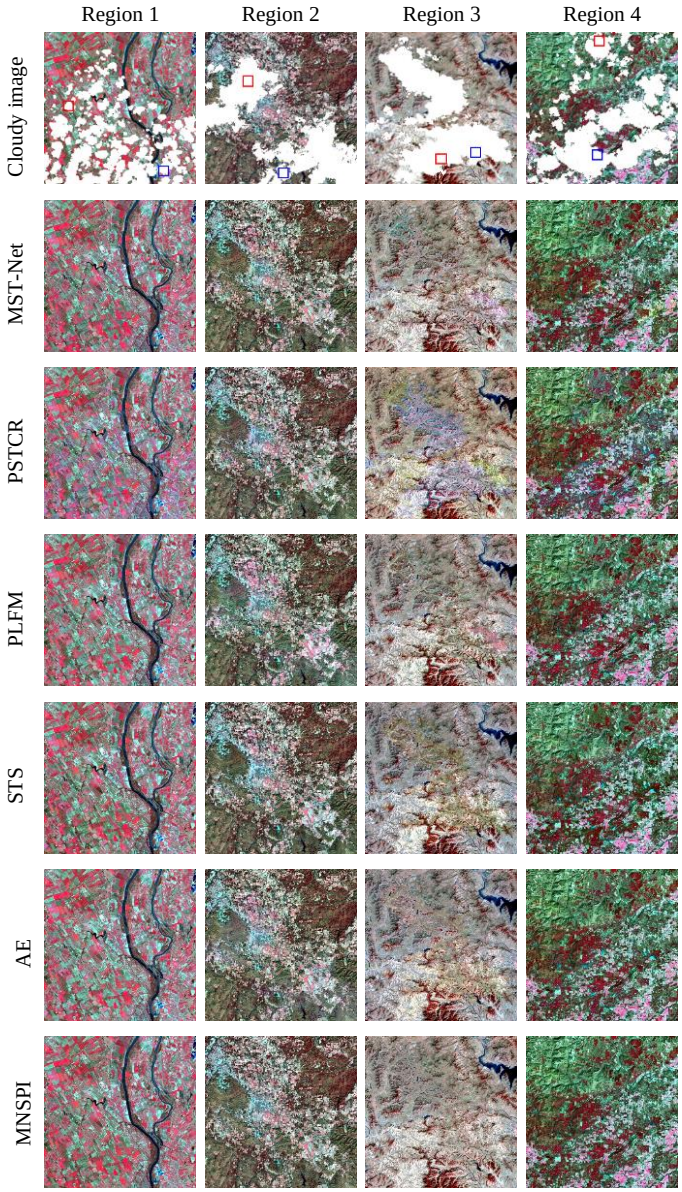


Fig. 7. Cloud removal results of the six methods in Regions 1-4 (bands NIR, red, and green as RGB). The red and blue boxes in the first row indicate the locations of the enlarged Subareas 1 and 2 in Fig. 8, respectively.

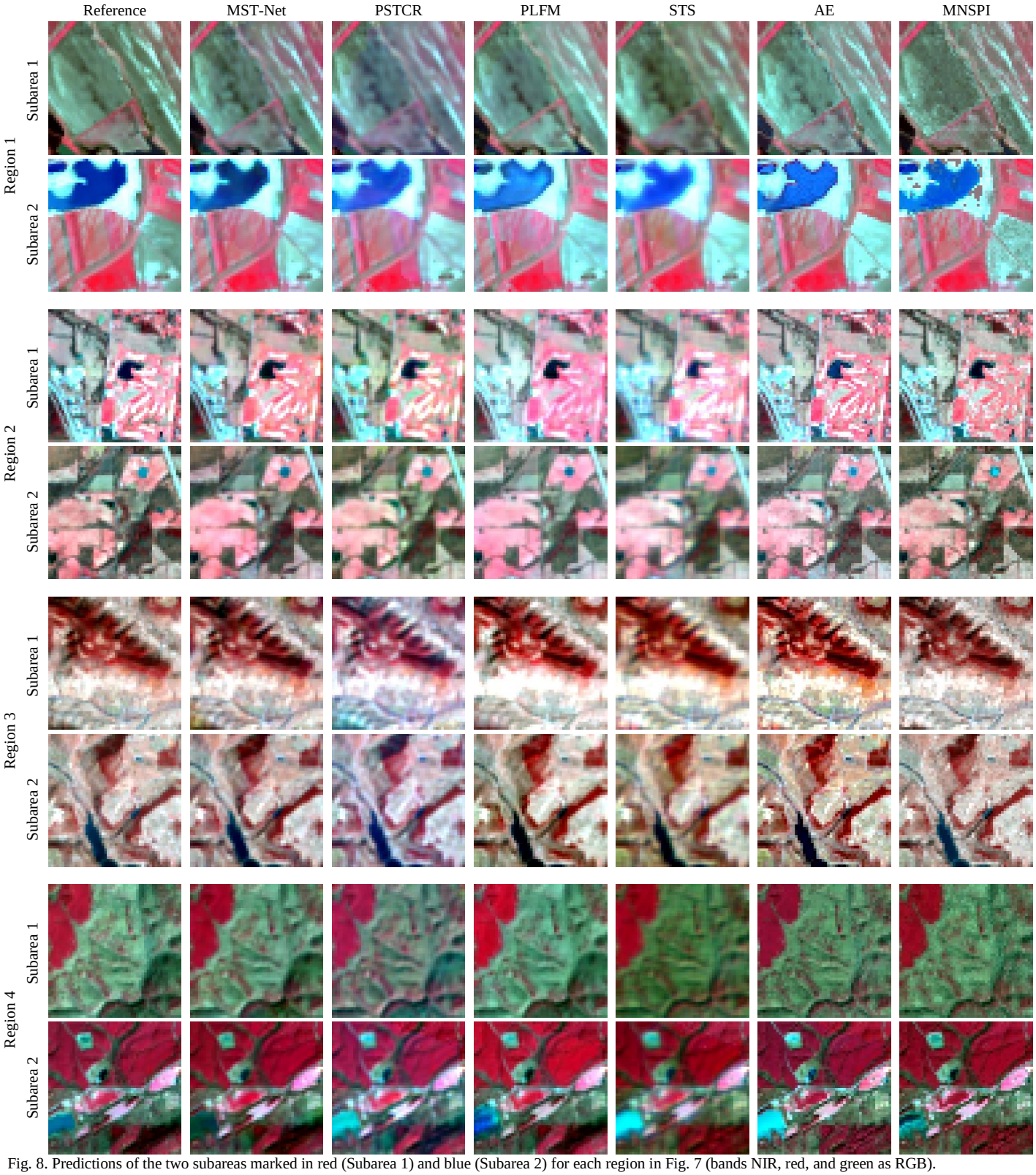


Fig. 8. Predictions of the two subareas marked in red (Subarea 1) and blue (Subarea 2) for each region in Fig. 7 (bands NIR, red, and green as RGB).

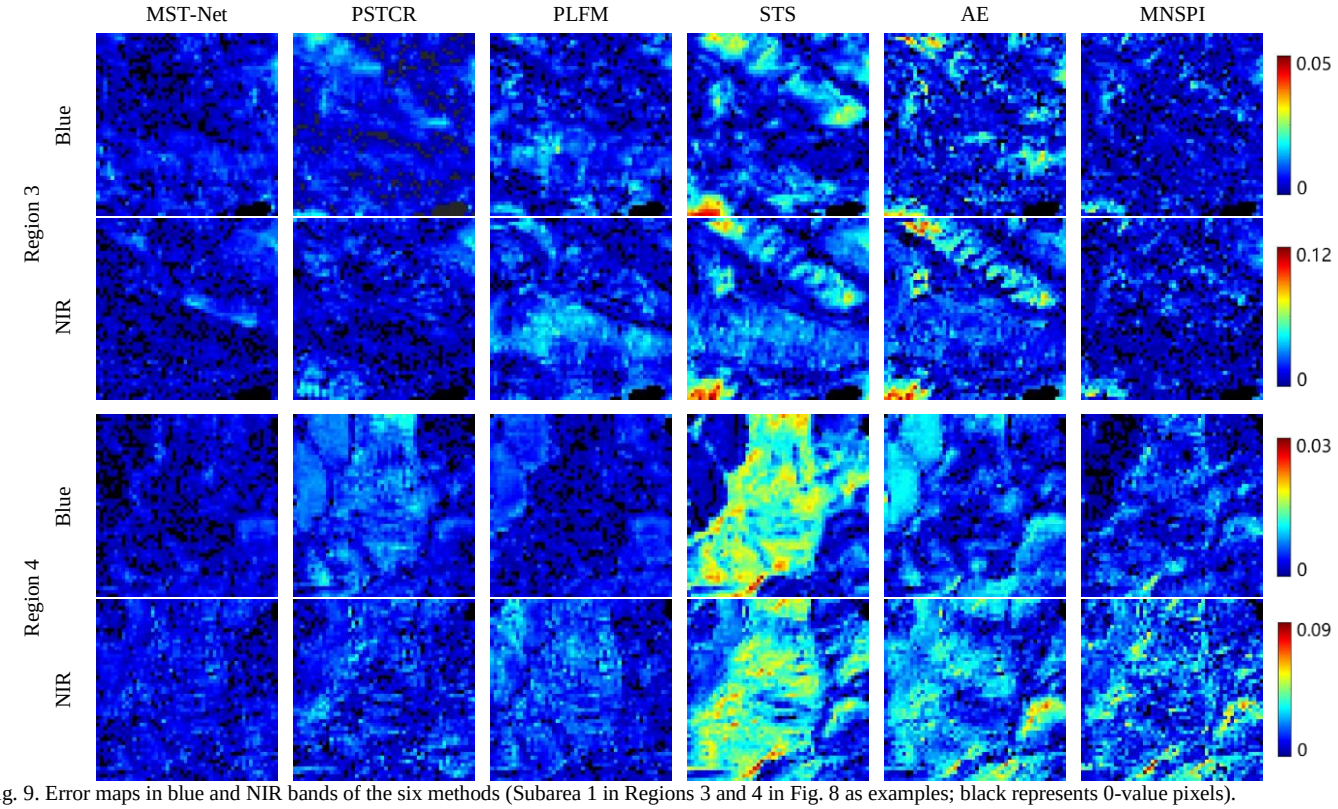


Fig. 9. Error maps in blue and NIR bands of the six methods (Subarea 1 in Regions 3 and 4 in Fig. 8 as examples; black represents 0-value pixels).

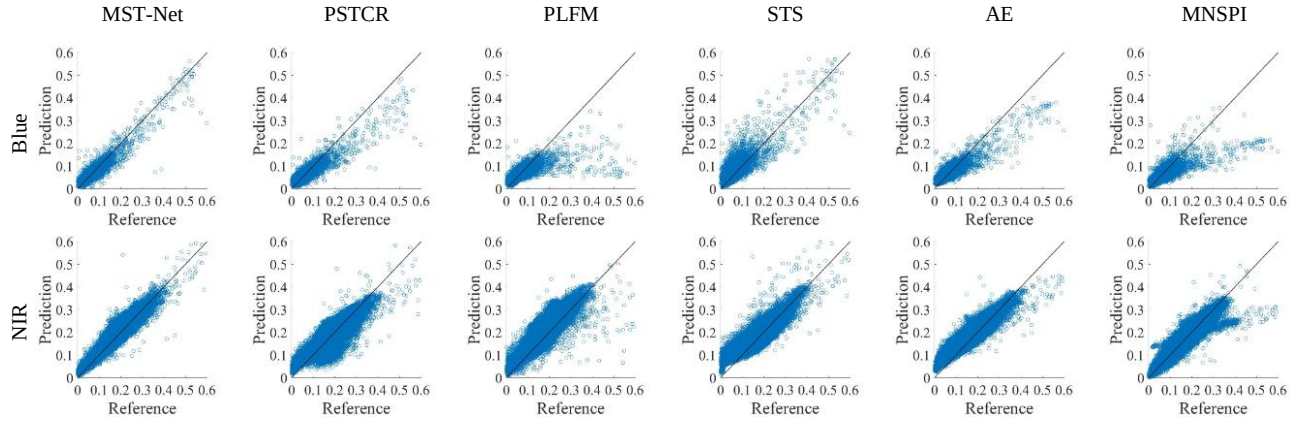


Fig. 10. Scatterplots between the predicted and actual reflectances in the blue and NIR bands for the six methods in Region 2.

Table 1 Accuracies (RMSEs, CCs and UIQIs are the averages of the six bands) of six cloud removal methods in Regions 1-12 (the values in bold are the most accurate results in each case).

	RMSE	CC	UIQI	SAM		RMSE	CC	UIQI	SAM		RMSE	CC	UIQI	SAM
	Region 1					Region 2					Region 3			
MST-Net	**0.0183**	**0.9186**	**0.9093**	**0.0829**		**0.0135**	**0.9188**	**0.9119**	0.0613		**0.0117**	**0.9428**	**0.9412**	0.0479
PSTCR	0.0222	0.8814	0.8381	0.1104		0.0162	0.8795	0.8696	0.0742		0.0143	0.9298	0.9205	0.0752
PLFM	0.0215	0.8724	0.8630	0.0958		0.0218	0.8526	0.8402	0.0721		0.0143	0.9255	0.9194	0.0499
STS	0.0217	0.8778	0.8728	0.1104		0.0190	0.8357	0.8331	0.0777		0.0223	0.8118	0.8063	0.0577
AE	0.0229	0.8668	0.8539	0.1223		0.0185	0.8734	0.8616	0.0677		0.0223	0.8096	0.8028	0.0566
MNSPI	0.0220	0.8467	0.8385	0.0884		0.0152	0.8881	0.8785	**0.0501**		0.0146	0.9019	0.8959	**0.0310**
	Region 4					Region 5					Region 6			
MST-Net	**0.0097**	**0.9525**	**0.9496**	0.0518		**0.0117**	**0.9748**	**0.9737**	0.0315		**0.0222**	**0.8526**	**0.8409**	**0.0869**
PSTCR	0.0125	0.9292	0.9246	0.0708		0.0150	0.9623	0.9594	0.0547		0.0238	0.8329	0.8096	0.1003
PLFM	0.0113	0.9430	0.9403	0.0564		0.0222	0.8915	0.8717	0.0500		0.0286	0.7927	0.7850	0.1154
STS	0.0164	0.8857	0.8798	0.0684		0.0152	0.9457	0.9432	0.0409		0.0256	0.8220	0.8207	0.0976
AE	0.0141	0.8934	0.8859	0.0545		0.0178	0.9491	0.9204	0.0344		0.0251	0.8051	0.8008	0.0971
MNSPI	0.0123	0.9280	0.9264	**0.0450**		0.0120	0.9650	0.9644	**0.0288**		0.0237	0.8253	0.8201	0.0925
	Region 7					Region 8					Region 9			
MST-Net	**0.0135**	**0.9184**	**0.9127**	0.0503		**0.0095**	**0.8541**	**0.8308**	**0.0544**		**0.0152**	**0.9144**	**0.9015**	0.0707
PSTCR	0.0152	0.9044	0.8984	0.0596		0.0113	0.8365	0.7804	0.0676		0.0174	0.8979	0.8828	0.0890

PLFM	0.0289	0.8567	0.8072	0.0965		0.0232	0.7461	0.7117	0.0970		0.0213	0.8344	0.8175	0.0853
STS	0.0436	0.8517	0.7322	0.1281		0.0142	0.7866	0.7816	0.0794		0.0169	0.8863	0.8842	0.0869
AE	0.0324	0.8890	0.7532	0.1168		0.0147	0.7682	0.7400	0.0832		0.0204	0.8540	0.8419	0.0950
MNSPI	0.0140	0.9027	0.9016	**0.0461**		0.0107	0.8100	0.7878	0.0565		0.0156	0.9019	0.8972	**0.0659**
	Region 10					Region 11					Region 12			
MST-Net	0.0114	**0.9399**	**0.9371**	0.0259		**0.0218**	**0.9596**	**0.9577**	0.0420		**0.0183**	**0.9502**	**0.9480**	0.0472
PSTCR	0.0130	0.9231	0.9194	0.0515		0.0254	0.9478	0.9448	0.0709		0.0195	0.9460	0.9410	0.0632
PLFM	0.0248	0.9101	0.8549	0.0951		0.0291	0.9488	0.9386	0.0586		0.0375	0.8231	0.8139	0.0819
STS	0.0356	0.9210	0.8642	0.0839		0.0244	0.9452	0.9401	0.0494		0.0251	0.9385	0.9278	0.0602
AE	0.0293	0.9058	0.8454	0.0945		0.0255	0.9345	0.9304	0.0512		0.0262	0.9128	0.8818	0.0774
MNSPI	**0.0096**	0.9372	0.9323	**0.0229**		0.0223	0.9547	0.9542	**0.0384**		0.0185	0.9441	0.9417	**0.0448**

C. Validation of the use of temporally closer images (Sentinel-2 MSI images)

In Section III-B, the effectiveness of the proposed MST-Net was demonstrated based on globally sampled training data. This section aims to examine the effectiveness of using Sentinel-2 MSI images. Specifically, the predictions of MST-Net using different auxiliary images are compared, including MST-Net (S): using multi-temporal data consisting of three Landsat 8 OLI and one Sentinel-2 MSI images (the same as in Section III-B), and MST-Net (L): using four Landsat 8 OLI images, which are temporally further than the Sentinel-2 MSI images used in Section III-B. The results, displayed in Fig. 11, show that compared with MST-Net (L), the predictions of MST-Net (S) are visually closer to the references. Moreover, Fig. 12 shows that the accuracies of the MST-Net prediction are obviously greater when Sentinel-2 MSI images are used. The results demonstrate that the Sentinel-2 MSI images play an important role in cloud removal. That is, when the available homologous auxiliary images are temporally further, the use of Sentinel-2 MSI images with shorter time intervals can facilitate more accurate prediction.

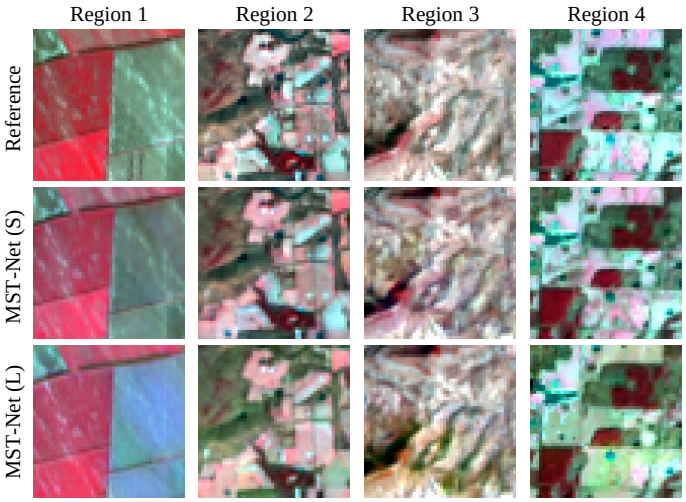


Fig. 11. Subareas of the MST-Net predictions in Regions 1-4 with different auxiliary images (bands NIR, red, and green as RGB). MST-Net (S): using time-series consisting of Landsat 8 OLI and Sentinel-2 MSI images; MST-Net (L): using only Landsat 8 OLI time-series data.

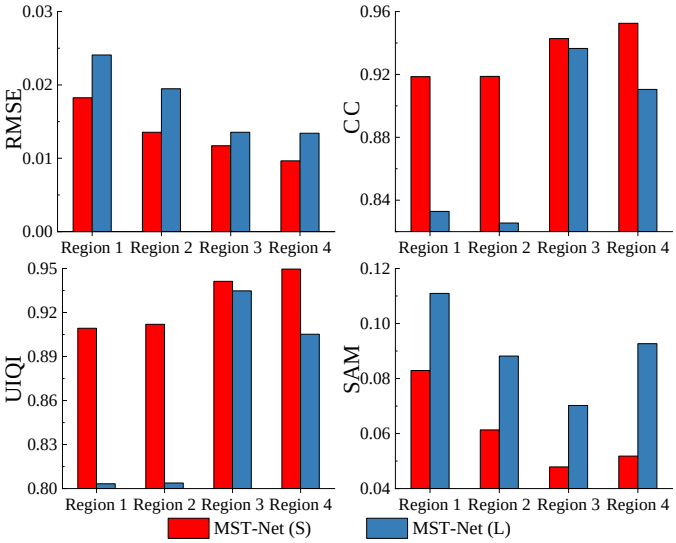


Fig. 12. Accuracies (averages of the six bands) of the MST-Net predictions in Regions 1-4 with different time-series auxiliary images. MST-Net (S): using time-series data consisting of Landsat 8 OLI and Sentinel-2 MSI images; MST-Net (L): using only Landsat 8 OLI time-series data.

D. Validation of the use of temporally closer but cloudy images

Table 2 Time-series data used for validation of temporally closer but cloudy images (data from one sampling point of the training data as an example).

Time-series (1)		Time-series (2)	
Number	Acquisition date	Number	Acquisition date
L1	2022.9.20	L3	2021.12.6
L2	2021.12.22	L4	2021.11.4
L3	2021.12.6	L5	2020.11.17
L4	2021.11.4	L6	2020.10.16

In this section, based on globally sampled training data, a comparative experiment is designed to validate the effectiveness of using temporally closer, but cloudy images. Based on MST-Net, we compared the predictions produced using different Landsat 8 OLI time-series images: (1) four partially cloud-contaminated Landsat 8 OLI images, and (2) the last two images of the time-series data in (1) plus two cloud-free images that are temporally further from the target cloudy image. The models trained through two different time-series auxiliary images were applied to the test data to obtain the predictions of two different versions. Details of the time-series data used are shown in Table 2 (data from one sampling point of the training data as an example). The average accuracies of the MST-Net predictions based on the two sets of auxiliary time-series (i.e., (1) and (2)) for the four regions are shown in Fig. 13. It can be seen that the MST-Net can produce more accurate predictions when using auxiliary time-series (1). This suggests that the temporally closer but cloudy multi-temporal images can provide more

valuable auxiliary information than temporally further, cloud-free images.

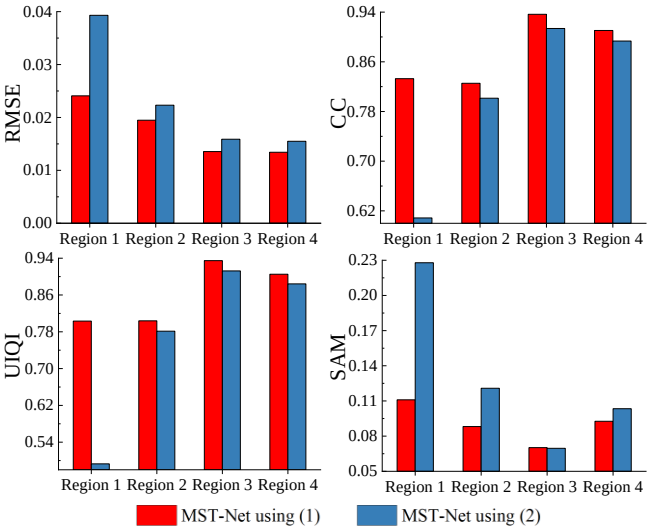


Fig. 13. Accuracies (averages of the six bands) of the MST-Net predictions using different auxiliary images: (1) Four partially cloud-contaminated Landsat 8 OLI images; (2) the last two images in (1), as well as two cloud-free images that are temporally further from the target cloudy image.

E. Ablation experiments

The proposed MST-Net consists of two networks, that is, MS-Net and MT-Net. MS-Net aims to fully mine the spatio-spectral information in the temporally closest auxiliary image, and MT-Net aims to further integrate the spatio-temporal information from auxiliary multi-temporal data for final prediction. This section aims to examine the effectiveness of combining MS-Net and MT-Net based on the data of Regions 1-4 in Section III-B. Specifically, MS-Net was trained using the common non-cloud data between the temporally closest auxiliary image (that is, Sentinel-2 MSI image with partial clouds) and the target cloudy images, and produced the prediction using the valid information in auxiliary images corresponding to the target cloud region. For MT-Net, the 10 m Sentinel-2 MSI image was degraded to 30 m to fill the cloudy image directly, as the preliminary prediction (one of the inputs of MT-Net). Then, the 10 m Sentinel-2 MSI image was fused with the other three Landsat OLI images for the final prediction. Fig. 14 shows the predictions of MS-Net, MT-Net and MST-Net. It is found that the MT-Net predictions have relatively noticeable color distortion, while the MS-Net predictions are more accurate in color, but present noise. This indicates that MS-Net can effectively exploit the multi-spectral information in the temporally closest auxiliary image and reconstruct the cloud regions more satisfactorily spectrally. However, MS-Net is trained pixel-by-pixel without considering the spatial relations between adjacent pixels, resulting in relative distortions in texture. In contrast, the MST-Net combines the advantages of MS-Net and MT-Net to obtain predictions that are visually closer to the references. Table 3 lists the average accuracies of all six bands for MS-Net, MT-Net and MST-Net. Amongst them, the MST-Net achieves the greatest accuracies in all regions. For example, in Region 3, the average CC of MST-Net is 0.0658 and 0.0290 larger than that of MS-Net and MT-Net, respectively.

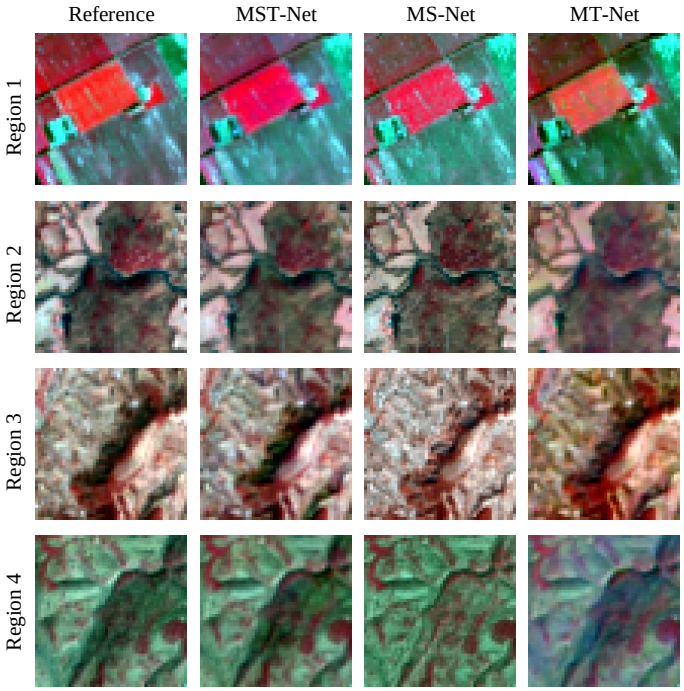


Fig. 14. Predictions of MST-Net, MS-Net and MT-Net in Regions 1-4 (one subarea was selected for each region; bands NIR, red and green as RGB).

F. Influence of cloud size

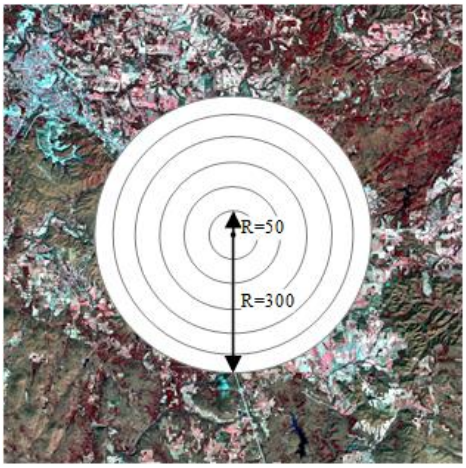


Fig. 15. Schematic diagram of clouds with different sizes (radius varies from 50 to 300 pixels).

This section aims to examine the performances of the six methods in removing clouds with different sizes. The original cloud-free data in four regions in Section III-B were covered with simulated clouds which share the same center, but different radii (ranging from 50 to 300 pixels), as shown in Fig. 15. The accuracies of the cloud removal results under different radii are shown in Fig. 16. The results indicate that the proposed MST-Net consistently produces the most accurate predictions. Specifically, the accuracies of most predictions in Regions 1 and 3 show a decreasing trend. In Regions 2 and 4, the results vary relatively smoothly, except for the results for a cloud radius of 50 pixels, which have the lowest accuracies. For Region 2, the reason for this phenomenon is that the vegetation in the 50 pixel radius area changed seasonally with a complex pattern. The reason for the result for Region 4 is that abrupt land cover

changes (different water color) occurred in the 50 pixel radius area. Due to the large proportion of land cover changes in the 50 pixel radius cloud regions, the overall accuracies of the predictions are relatively small when focusing on this small-sized region.

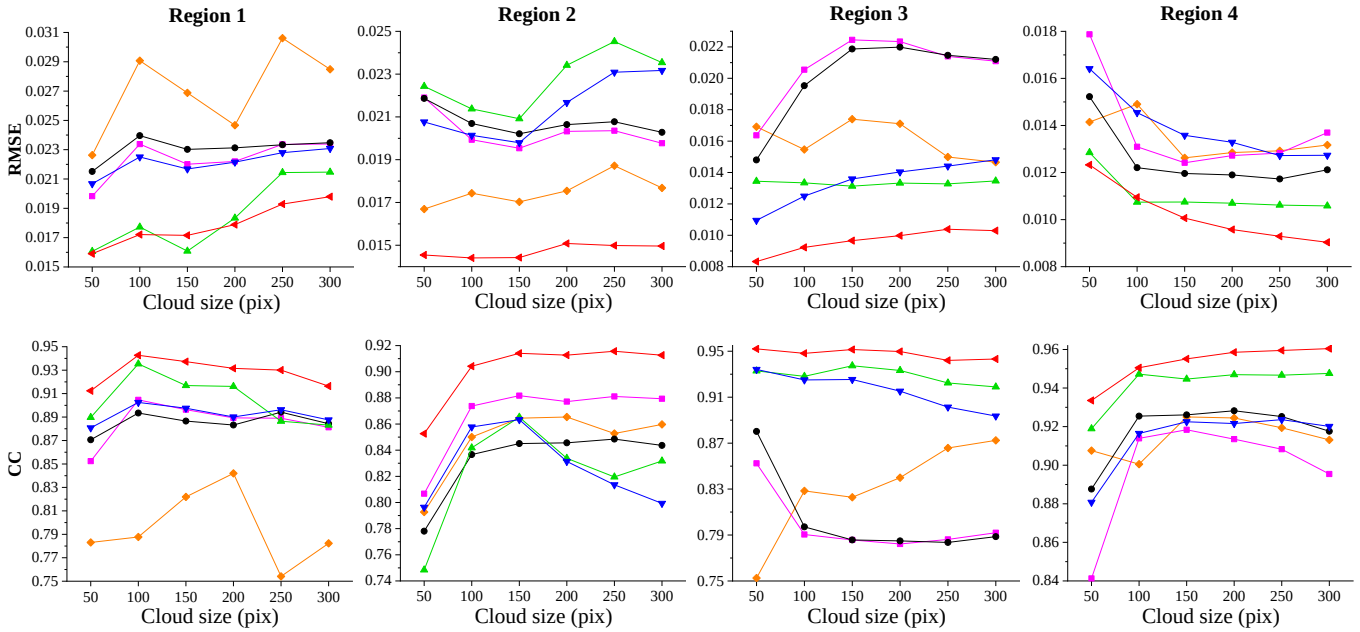
G. Influence of thin clouds

In the previous experiments in this paper, known cloud masks were used to simulate the cloud-contaminated areas, which means that the cloud masks are 100% accurate. However, in practice, thick and thin clouds often appear simultaneously, and existing cloud detection methods often suffer from a certain degree of omission error in identifying thin clouds. Considering this, we designed simulation experiments to evaluate the effectiveness of each method in the presence of thin cloud omission. Specifically, based on the data of Regions 1 and 2 in Section III-B, the cirrus band was used to simulate a certain range of thin clouds and add them to the target cloudy image and the auxiliary time-series images [55]. Fig. 17 shows the distribution of simulated thick clouds (same as in Section III-B) and thin clouds.

Fig. 18 shows the accuracies (average of six bands) of the six methods with, and without, thin cloud omission. It can be seen that the proposed MST-Net can always produce more accurate predictions than the other five methods, either with or without thin cloud omission. Moreover, it can be seen that the accuracies of the PSTCR and MNSPI predictions decrease remarkably with thin cloud omission in Regions 1 and 2, respectively. For PSTCR, the significant decrease in accuracies may be due to the global linear histogram match pretreatment of the corresponding bands between the auxiliary and cloudy images, which propagates the local thin cloud omission error to the entire auxiliary image, thereby greatly affecting the accuracies of the PSTCR predictions. For MNSPI, the effective adjacent information of thick clouds, which includes similar pixels involved directly in the calculation, is crucial for the final prediction. Consequently, thin clouds surrounding thick clouds exert significant adverse effects on the MNSPI predictions. The PLFM prediction is less affected by thin cloud omission in Region 2, but its RMSE is significantly increased in Region 1. This could be attributed to its sequential processing of each auxiliary image from distant to near temporally, and the spatial coverage of thin cloud in Region 1 is larger than in Region 2, resulting in a more significant negative effect on the final prediction. Compared to the other methods, the accuracies of the MST-Net predictions decrease the least with the influence of thin clouds, indicating that it is least affected by thin clouds. The reason is that MST-Net (the MS-Net module) handle multiple bands simultaneously, and the fusion of multi-spectral data enables the longer wavelength (e.g., SWIR) bands (less affected by the thin cloud contamination) to correct the thin cloud errors in the short wavelength bands to a certain extent. This reflects the advantage of integrating multi-spectral information in MS-Net. To sum up, the MST-Net achieves more stable performance than the benchmark methods with thin cloud omission.

Table 3 Accuracies (averages of the six bands) of MST-Net, MS-Net and MT-Net in Regions 1-4.

	RMSE			CC			UIQI			SAM		
	MST-Net	MS-Net	MT-Net	MST-Net	MS-Net	MT-Net	MST-Net	MS-Net	MT-Net	MST-Net	MS-Net	MT-Net
Region 1	**0.0182**	0.0194	0.0240	**0.9186**	0.9048	0.8720	**0.9092**	0.8974	0.8673	**0.0829**	0.0880	0.1172
Region 2	**0.0135**	0.0139	0.0180	**0.9188**	0.9090	0.8820	**0.9119**	0.9030	0.8793	**0.0613**	0.0524	0.0692
Region 3	**0.0117**	0.0163	0.0169	**0.9428**	0.8770	0.9138	**0.9412**	0.8675	0.9076	0.0479	**0.0336**	0.0852
Region 4	**0.0097**	0.0128	0.0121	**0.9525**	0.9164	0.9271	**0.9496**	0.9126	0.9247	**0.0518**	0.0564	0.0650



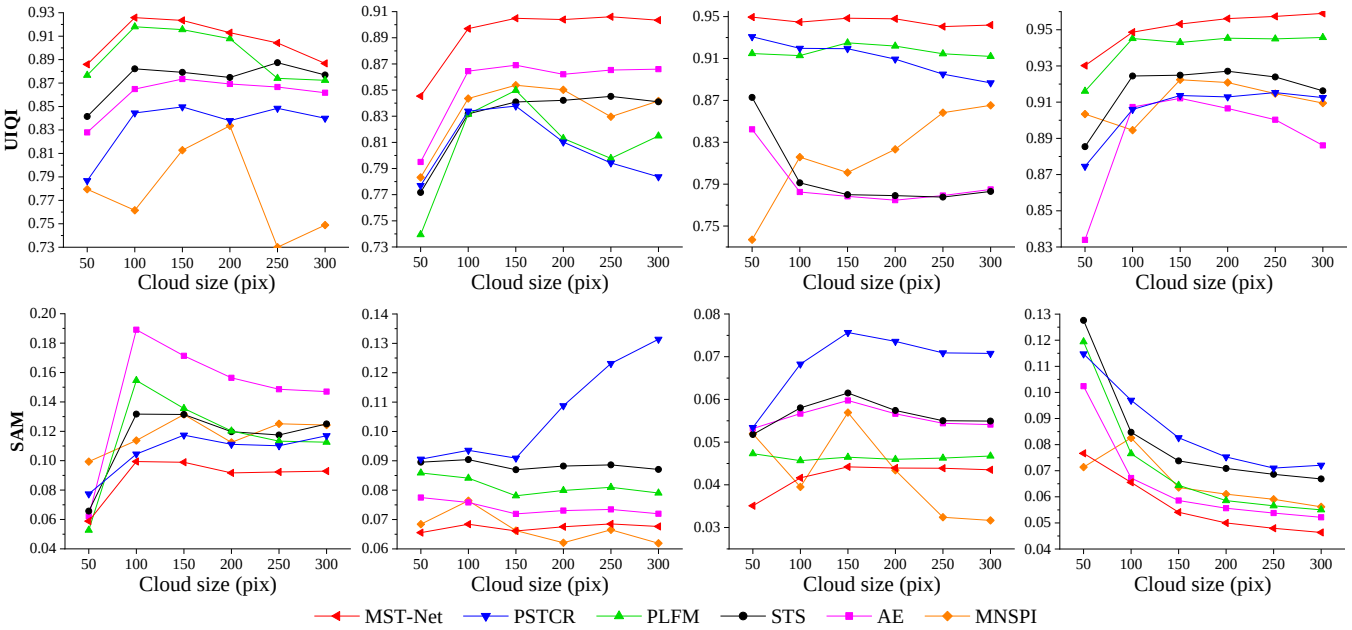


Fig. 16. Accuracies (averages of the six bands) of the six cloud removal methods under clouds with different sizes (e.g., different radii in this experiment).

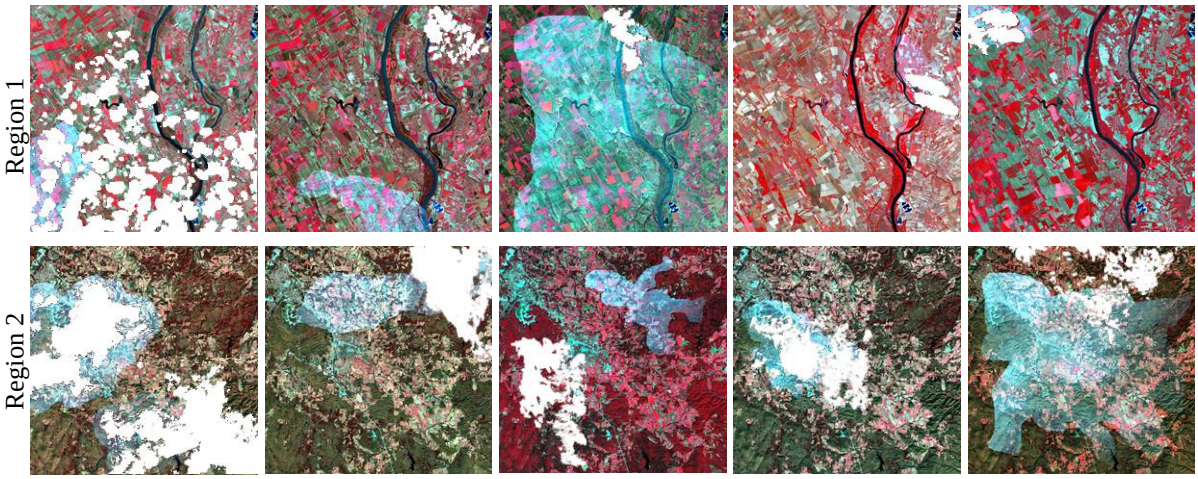


Fig. 17. The simulated thick (same as in Section III-B) and thin clouds in Regions 1-2 (bands NIR, red, and green as RGB). For each column, the first is the target cloudy image, and the later four are the auxiliary multi-temporal images.

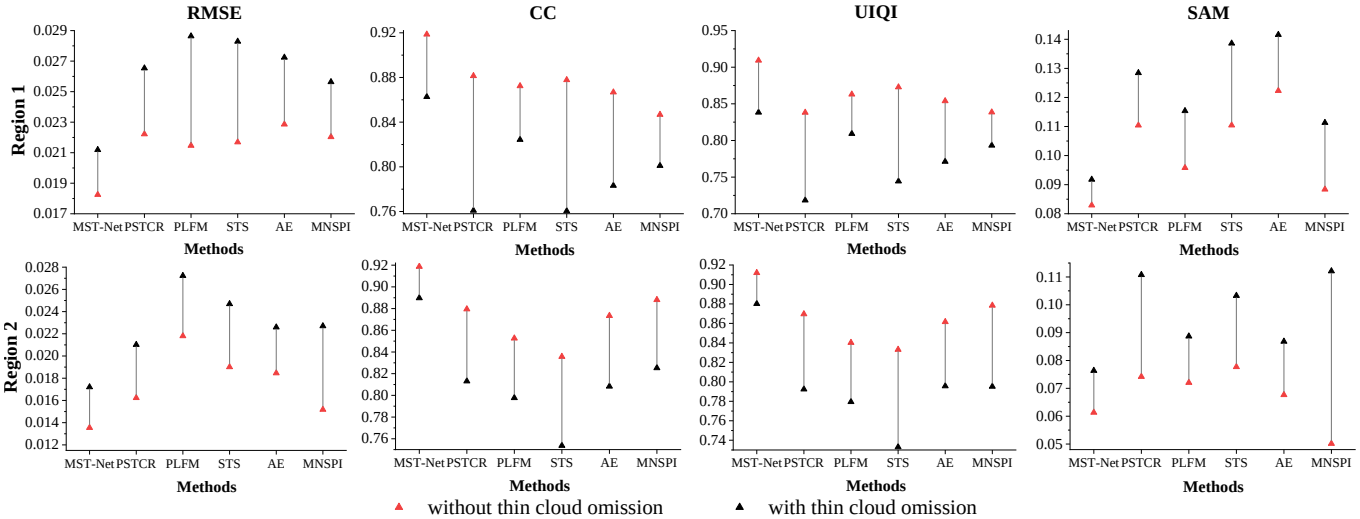


Fig. 18. Accuracies (averages of the six bands) of the six cloud removal methods with, and without, thin cloud omission.

H. Application to land cover mapping based on the filled data

The reconstruction of thick cloud occlusion information is crucial for downstream applications. To further illustrate the practical significance of removing thick clouds, we performed a land cover mapping [56] experiment using the predictions of the six methods. Specifically, we applied the k-means clustering algorithm to conduct unsupervised classification of the results in Fig. 7, setting the number of land cover types in all four regions as three, according to their characteristics. Table 4 presents the overall accuracies (OA) of the classification results for filled area, demonstrating that the MST-Net predictions lead to more accurate land cover mapping. This indicates that the MST-Net results have potential to facilitate smooth development of subsequent application-oriented research.

Table 4 Classification accuracies (OA) of the results in Fig. 7 (values in bold represent the most accurate result in each case).

	MST-Net	PSTCR	PLFM	STS	AE	MNSPI
Region 1	**0.9029**	0.8812	0.8872	0.8691	0.8715	0.8785
Region 2	**0.8132**	0.7548	0.7503	0.7105	0.7341	0.7744
Region 3	**0.8531**	0.8208	0.8306	0.7421	0.7337	0.7884
Region 4	**0.8982**	0.8531	0.8747	0.7996	0.8215	0.8357

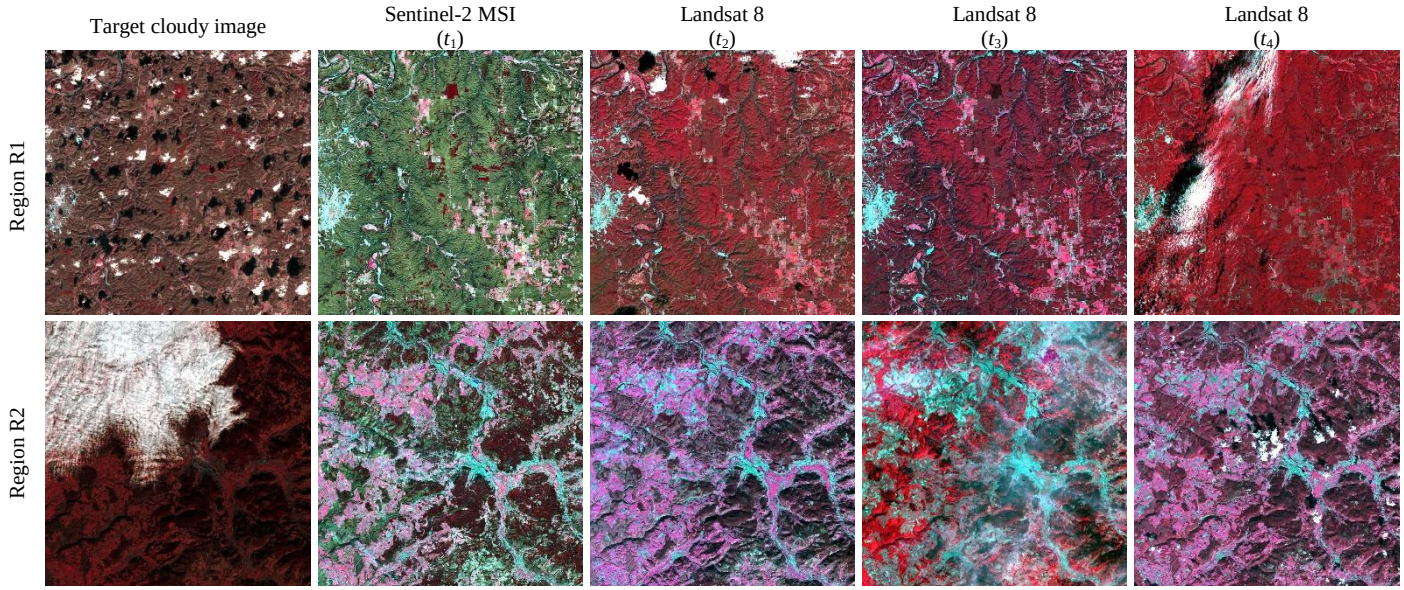


Fig. 19. Real cloudy images (bands NIR, red, and green as RGB)

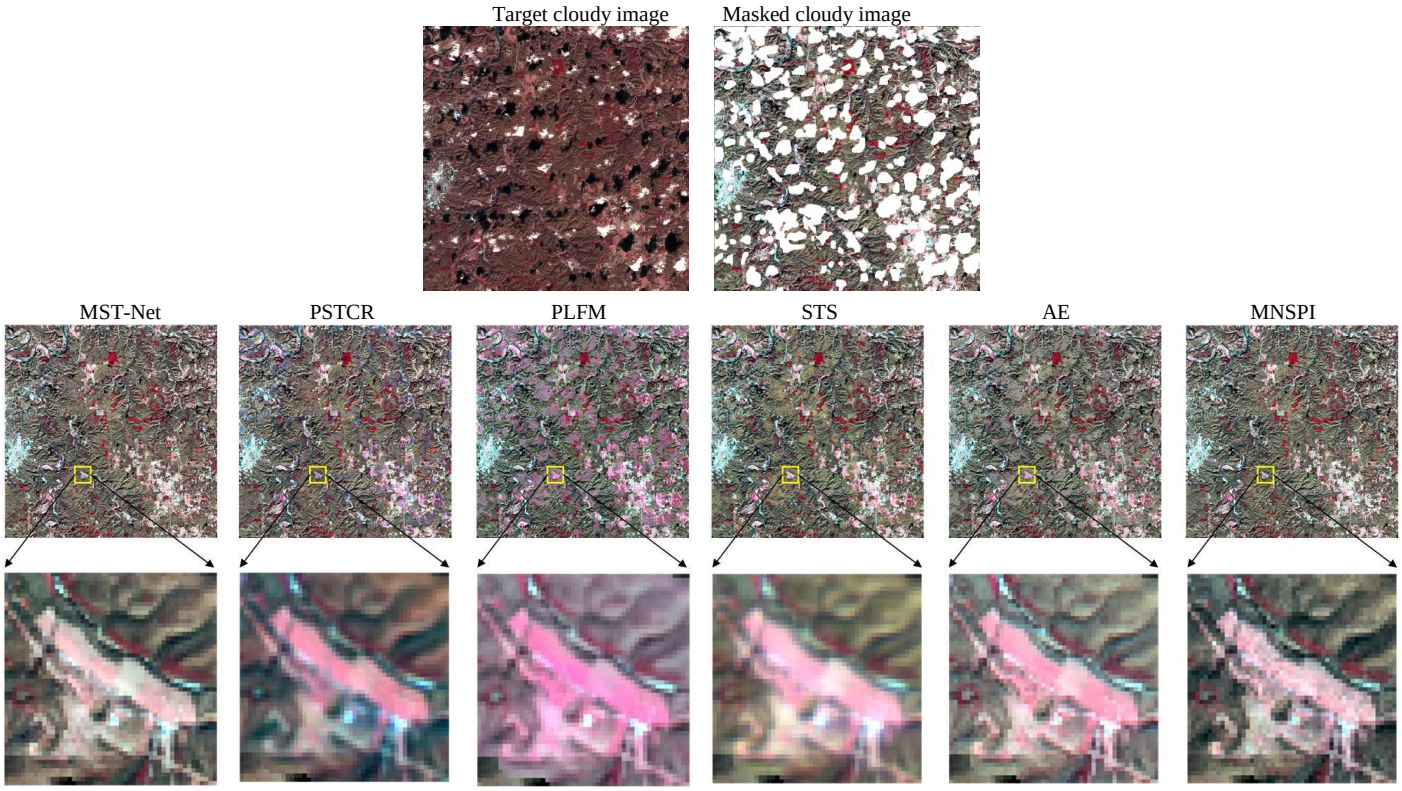


Fig. 20. Cloud removal results of the six methods for the real cloudy image in Region R1 (bands NIR, red and green as RGB).

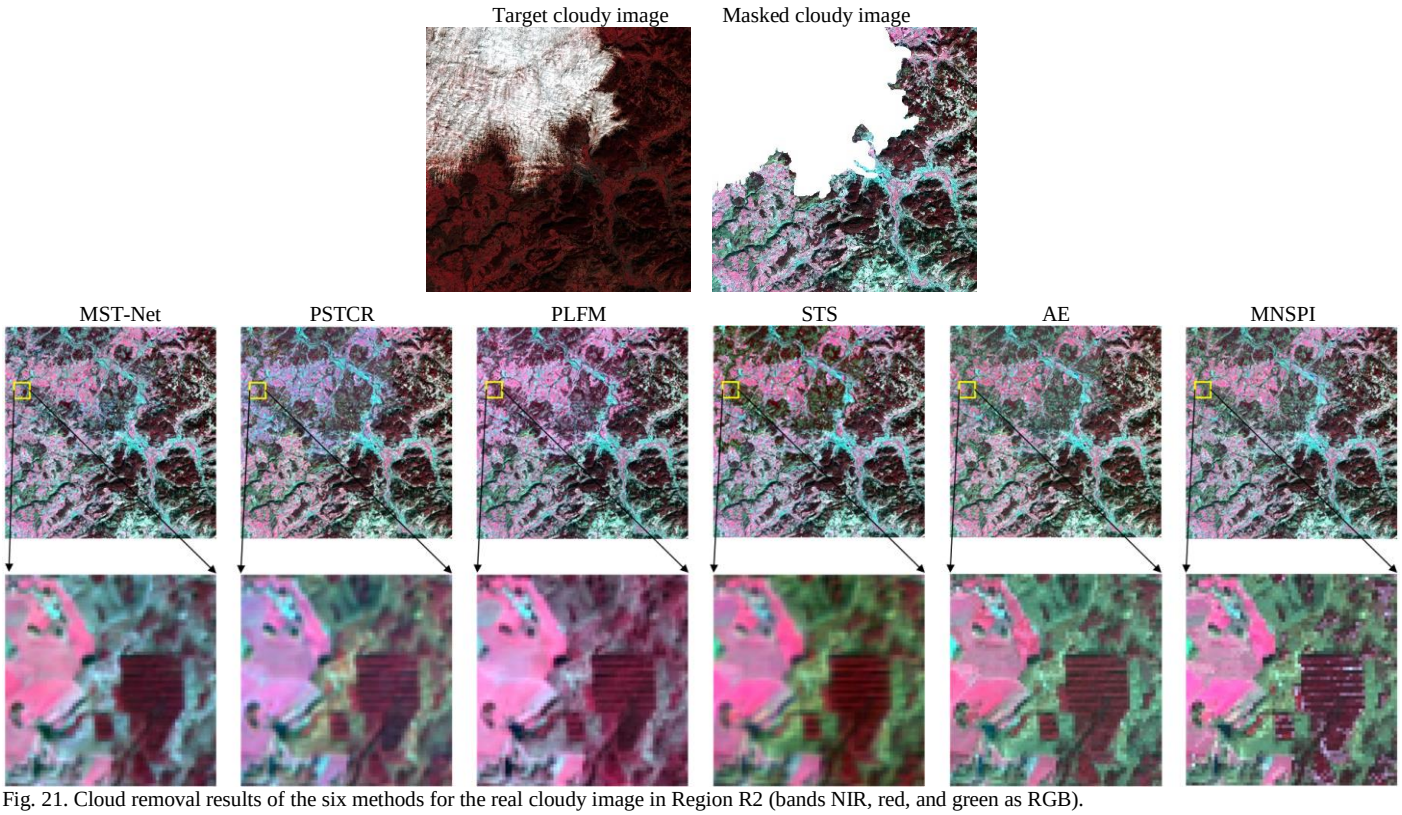


Fig. 21. Cloud removal results of the six methods for the real cloudy image in Region R2 (bands NIR, red, and green as RGB).

I. Experiments on real clouds

In this section, we examined the performance of the proposed MST-Net for removing real clouds. The experiment is based on real cloudy images from two different regions (all covering 1000×1000 Landsat pixels), with some of the temporally neighboring images contaminated by clouds, as shown in Fig. 19. Figs. 20 and 21 show the real cloud removal results of the six methods in the two regions, where the target real cloudy images are shown in the first line (left is the original cloudy image, and right is the cloudy image with cloud masked for clearer observation). Specifically, for the results of Region R1 in Fig. 20, the MST-Net prediction presents a great consistency, with no seams between the reconstructed and original non-cloudy areas. In contrast, the PLFM and AE predictions show clear seams. Similarly, for the results of Region 2 in Fig. 21, the MST-Net produces a more satisfactory prediction in both color and texture, while there is significant noise in the MNSPI prediction. The experimental results indicate that MST-Net outperforms the benchmark methods in real-world scenarios, demonstrating its enhanced generalization capability. That is, the proposed MST-Net has great potential for practical applications.

IV. Discussion

A. Generalization ability of the MST-Net

The proposed MST-Net fuses Sentinel-2 MSI images to reconstruct Landsat 8 OLI cloudy images, and its effectiveness was validated in Section III. In this section, we examined the performance of the MST-Net based on Landsat 8 OLI images only, that is, the Sentinel-2 MSI images used in Section III-B were replaced by Landsat 8 OLI images acquired further from the target cloudy image than the three other Landsat 8 OLI images in the time-series. The masks used are the same as in Section III-B. Fig. 22 shows the accuracies of the MST-Net and the five benchmark methods for the four regions. It is found that the MST-Net consistently produces the greatest accuracies. Taking Region 4 as an example, the average CC of the MST-Net prediction is 0.1926, 0.0032, 0.0462, 0.1076 and 0.0197 larger than that of PSTCR, PLFM, STS, AE and MNSPI predictions, respectively. The results indicate that even when only Landsat 8 OLI auxiliary images are used, the MST-Net can still achieve more accurate predictions than the benchmark methods, revealing the advantages of the designed network.

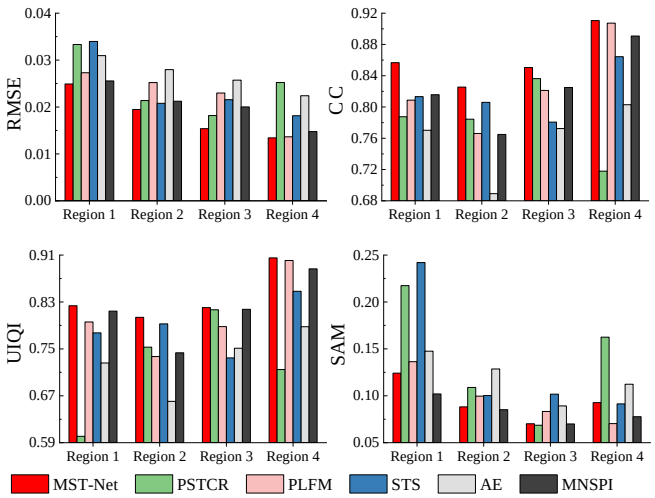


Fig. 22. Accuracies (averages of the six bands) of the six methods using completely Landsat 8 OLI time-series images in Regions 1-4.

B. Advantages of the proposed MST-Net

The advantages of MST-Net are summarized in three parts. First, from the perspective of data utilization, the MST-Net can fuse multi-source data (i.e., Sentinel-2 images) for cloud removal of Landsat images, which are rarely explored in existing studies. Moreover, the MST-Net can take full advantage of the multi-temporal images containing clouds with staggered spatial positions, which are always abandoned directly in previous methods. Thus, MST-Net exhibits greater adaptability to frequent instances of cloud occlusion encountered in practical cases. Second, from the perspective of model construction, considering the vast amount of information contained in the multi-source and multi-temporal data, the MST-Net incorporates their spatio-temporal-spectral information effectively via the designed two-stage structure composed of MS-Net and MT-Net. Thus, the model design is significantly simplified while ensuring fitting the relationship between the auxiliary and target cloudy images. For MS-Net, the mechanism of using multi-spectral bands simultaneously enables comprehensive exploitation of longer wavelength (e.g., SWIR) band information and, thus, adaptation to scenarios involving thin cloud omission errors. In practical applications, thin and thick cloud always exist simultaneously. Thus, the MS-Net part endows the MST-Net with stronger generalization capabilities in handling various clouds. Third, from the perspective of practical applications, considering the generalization ability of MST-Net demonstrated in handling both homologous (Section IV-A) and multi-source (Section III-B) images, it is believed that the MST-Net holds great promise for cloud removal of other datasets. However, it is worth investigating how to properly adjust the network structure so that the MST-Net can fully mine the effective information in different auxiliary data.

C. Fusion of multi-modal data

The task of gap filling relies heavily on the information in the auxiliary time-series corresponding to the cloud regions in the target cloudy image. The land cover changes between multi-temporal images have always been a challenging problem for gap filling, especially for reconstruction of the small-sized, new land cover classes under the target clouds. In the experiments analyzing the influence of cloud size in Section III-F, an abrupt change to the small-sized water body in Region 3 occurred under cloud within radius 50 pixels, resulting in a significant decrease in accuracies of the six methods when only the small region was considered for accuracy evaluation. This suggests that the role of optical images as auxiliary data is limited in this case. As a type of data with a different mode, SAR data [35], [57], [58], [59] can penetrate clouds to obtain land cover information under them, which may provide valuable auxiliary information for cloud removal, enhancing the performance of cloud removal. However, there is always obvious noise in SAR images [60], [61]. Moreover, there is a significant difference in the signals between SAR and optical remote sensing images. Thus, it is crucial to account for these issues in fusion of multi-modal data. Generative adversarial networks (GANs) have strong learning ability and adaptability, which can harmonize the difference in information between multi-modal data, showing great potential for using optical and SAR data jointly in cloud removal. However, GAN needs to maintain balance in the adversarial training between the generator and discriminator. Therefore, how to construct the network structure of the discriminator and generator and design a reasonable loss function accordingly is a great challenge, which deserves further study.

D. Future research

The proposed MST-Net reconstructs the entire cloud-contaminated area for each region using a consistent network structure and training process. However, it is important to note that the complexity of different local areas may vary in practice. In regions with strong spatial homogeneity, where land covers are relatively uniform and spatial texture is simple, accurate reconstruction results can be achieved using shallow networks. Conversely, for heterogeneous regions with rich spatial details, they may require deeper networks to characterize the temporal variation reliably, and to achieve accurate reconstruction. In future research, indicators such as the semi-variogram function could be utilized to evaluate the complexity of spatial texture and temporal variation (e.g., cross semi-variance). These indicators can then guide the selection of appropriate network structures (e.g., skipping layers based on a complexity threshold) or can be incorporated into the loss function to determine network weight updates. Moreover, at the application level, it is also worthwhile to generalize the MST-Net to gap filling of multi-modal remote sensing data, such as land surface temperature [62], [63], [64] and surface soil moisture [65], [66], [67]. Furthermore, as heterogeneous data sources, Sentinel-2 MSI and Landsat 8 OLI images inevitably exhibit differences in atmospheric correction methods, spectral response functions, and lighting conditions. The proposed MST-Net achieves spectral feature conversion between Sentinel-2 MSI and Landsat 8 OLI images through network training. In the future, spectral conversion methods based on invariant ground objects can be considered to obtain spectrally coordinated images that retain more ground object information at the Sentinel-2 acquisition time, such as to fill the gaps in Landsat series data more accurately.

Note that the tests presented in this paper were derived from the MST-Net trained using simulated cloudy images. In real-world applications, regions that are persistently covered by clouds may suffer from a lack of sufficient available training data, which can lead to decreased accuracy in the MST-Net predictions. The richness and diversity of the training data are crucial for deep learning models operating on global datasets, as these factors directly influence the generalization ability of the learning model. The targeted data enhancement strategies may provide effective solutions. Commonly employed techniques, such as rotation and random flipping, could be utilized for this purpose. Furthermore, future research could explore the development of GAN-based data generation methods, leveraging a well-trained GAN to synthesize images with specific characteristics that can enhance model training.

V. Conclusion

Considering the large temporal interval between effective homologous auxiliary images and the target cloudy image, as well as the waste of valuable complementary information in multi-temporal cloudy images, in this paper, we proposed to

comprehensively utilize multi-source image with finer temporal resolution and multi-temporal images with moving clouds for thick cloud removal. Specifically, for cloud removal of Landsat 8 OLI images, we proposed to fuse Sentinel-2 images with great similarities in data characteristics, but finer temporal resolution. Moreover, we exploited comprehensively the remaining effective information in the multi-temporal cloudy Sentinel-2 MSI and Landsat 8 OLI images. To fully exploit the spatio-temporal-spectral information in the auxiliary multi-source and multi-temporal images, we proposed the MST-Net consisting of two stages. Specifically, it initially explores the spatio-spectral information in the temporally closest Sentinel-2 image via the MS-Net module, and then integrates spatio-temporal information in the multi-temporal cloudy images with the MT-Net module. Experiments were carried out with the aim of cloud removal from Landsat 8 OLI images in six regions. The conclusions are summarized as follows:

1) MST-Net can produce visually continuous predictions without obvious seams and artifacts, and it is more accurate than five benchmark methods.
2) Utilization of the temporally closest, heterogonous source images (i.e., Sentinel-2 MSI images) can lead to more accurate predictions. For example, the increase in average CC of all bands is over 0.05 after using Sentinel-2 MSI images.
3) Compared with cloud-free images that are temporally further from the target cloudy images, temporally closer, but cloudy images can facilitate more accurate prediction. For example, in Regions 2 and 3, the average CC values are 0.10 and 0.05 larger when using temporally closer, but cloudy images.
4) Both multi-source and multi-temporal images are important in cloud removal, and the performance can be further enhanced by effectively integrating both parts (i.e., the proposed MST-Net).
5) Under different cloud sizes, the MST-Net can consistently produce more accurate predictions compared to the five benchmark methods.
6) The performance of MST-Net is least affected by thin cloud omission errors.

REFERENCES

[1] P. Ebel, V. Garnot, M. Schmitt, J. Wegner, and X. Zhu, "UnCRtainTS: uncertainty quantification for cloud removal in optical satellite time series," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[2] C. Stucker, V. Garnot, and K. Schindler, "U-TILISE: a sequence-to-sequence model for cloud removal in optical satellite time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[3] X. Zou, K. Li, J. Xing, P. Tao, and Y. Cui, "PMAA: a progressive multi-scale attention autoencoder model for high-performance cloud removal from multi-temporal satellite imagery," *European Conference on Artificial Intelligence (ECAI)*, 2023.

[4] M. Ruß wurm and M. Körner, "Self-attention for raw optical satellite time series classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 421–435, 2020.

[5] M.King, S. Platnick, W. Menzel, S. Ackerman, and P. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua Satellites," *Transactions on Geoscience and Remote Sensing*, vol. 51 no. 7, pp. 3826–3852, 2013.

[6] Ji, C.Y., "Haze reduction from the visible bands of LANDSAT TM and ETM+ images over a shallow water reef environment," *Remote Sensing of Environment*, vol. 112, no. 4, pp. 1773–1783. 2008.

[7] X. Ma, Q. Wang, X. Tong, and P. M. Atkinson, "A deep learning model for incorporating temporal information in haze removal," *Remote Sensing of Environment*, vol. 274, pp. 113012, 2022.

[8] M. Xu, F. Deng, S. Jia, X. A. Jia, and J. Plaza, "Attention mechanism-based generative adversarial networks for cloud removal in Landsat images," *Remote Sensing of Environment*, vol. 271, pp. 112902, 2022.

[9] J. Yeom, J. Roujean, K. Han, K. Lee, and H. Kim, "Thin cloud detection over land using background surface reflectance based on the BRDF model applied to Geostationary Ocean Color Imager (GOCI) satellite data sets," *Remote Sensing of Environment*, vol. 239, pp. 111610, 2020.

[10] Y. Zhang, B. Guindon, and J. Cihlar, "An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat image," *Remote Sensing of Environment*, vol. 82, no. 2–3, pp. 173–187, 2002.

[11] S. Skakun, J. Wevers, C. Brockmann, G. Doxani, M. Aleksandrov, M. Batič, D. Frantz, F. Gascon, L. Gómez-Chova, O. Hagolle, D. López-Puigdollers, J. Louis, M. Lubej, G. MateoGarcía, J. Osman, D. Peressutti, B. Pflug, J. Puc, R. Richter, J. Roger, P. Scaramuzza, E. Vermote, N. Vesel, A. Zupanc, and L. Žust, "Cloud Mask Intercomparison eXercise (CMIX): an evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2," *Remote Sensing of Environment*, vol. 274, pp. 112990, 2022.

[12] Q. Cheng, H. Shen, L. Zhang, and Z. Peng, "Missing information reconstruction for single remote sensing images using structure-preserving global optimization," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1163–1167, 2017.

[13] C. Guillemot and O. L. Meur, "Image inpainting: overview and recent advances," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 127–144, 2014.

[14] G. Gao, and Y. Gu, "Multitemporal Landsat missing data recovery based on tempo-spectral angle model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3656–3668, 2017.

[15] C. Lin, K. Lai, Z. Chen, and J. Chen, "Patch-based information reconstruction of cloud-contaminated multitemporal image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 163–174, 2014.

[16] R. Cao, Y. Chen, J. Chen, X. Zhu, and M. Shen, "Thick cloud removal in Landsat images based on autoregression of Landsat time-series data," *Remote Sensing of Environment*, vol. 249, pp. 112001, 2020.

[17] Q. Cheng, H. Shen, L. Zhang, Q. Yuan, and C. Zeng, "Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 92, pp. 54–68, 2014.

[18] Z. Tang, H. Adhikari, P. Pellikka, and J. Heiskanen, "A method for predicting large-area missing observations in Landsat time series using spectral-temporal metrics," *International Journal of Applied Earth Observation and Geoinformation*, vol. 99, pp. 102319, 2021.

[19] Y. Chen, L. Yang, X. Yang, R. Fan, M. Bilal, and Q. Li, "Thick clouds removal from multitemporal ZY-3 Satellite images using deep learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 143–152, 2020.

[20] J. Gao, Q. Yuan, J. Li, and X. Su, "Unsupervised missing information reconstruction for single remote sensing image with Deep Code Regression," *International Journal of Applied Earth Observations and Geoinformation*, vol. 105, pp. 102599, 2021.

[21] J. Zheng, X. Liu, and X. Wang, "Single image cloud removal using U-net and generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6371–6385, 2021.

[22] A. Maalouf, P. Carre, B. Augereau, C. Fernandez-Maloigne, "A bandelet-based inpainting technique for clouds removal from remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2363–2371, 2009.

[23] L. Wang and Q. Wang, "Fast spatial-spectral random forests for thick cloud removal of hyperspectral images," International Journal of Applied Earth Observation and Geoinformation, vol. 112, pp. 102916, 2022.

[24] Q. Wang, L. Wang, X. Zhu, Y. Ge, X. Tong, and P. M. Atkinson, "Remote sensing image gap filling based on spatial-spectral random forests," *Science of Remote Sensing*, vol. 5, pp. 100048, 2022.

[25] Z. Zhu, C. E. Woodcock, C. Holden, and Z. Yang, "Generating synthetic Landsat images based on all available Landsat data: Predicting Landsat

surface reflectance at any given time," *Remote Sensing of Environment*, vol. 162, pp. 67–83, 2015.

[26] L. Lorenzi, F. Melgani, and G. Mercier, "Missing-area reconstruction in multispectral images under a compressive sensing perspective," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 3998–4008, 2013.

[27] C. Lin, P. Tsai, K. Lai, and J. Chen, "Cloud removal from multitemporal satellite images using information cloning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 232–241, 2013.

[28] X. Zhu, F. Gao, D. Liu, and J. Chen, "A modified neighborhood similar pixel interpolator approach for removing thick clouds in Landsat images," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 3, pp. 521–525, 2012.

[29] B. Chen, B. Huang, L. Chen, and B. Xu, "Spatially and temporally weighted regression: a novel method to produce continuous cloud-free Landsat imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 27–37, 2017.

[30] F. Lyu, S. Wang, S. Han, C. Catlett, and S. Wang, "An integrated cyberGIS and machine learning framework for fine-scale prediction of Urban Heat Island using satellite remote sensing and urban sensor network data," *Urban Informatics*, vol. 1, no. 6, pp. 1–15, 2022.

[31] Y. Song, Y. Xu, B. Chen, Q. He, Y. Tu, F. Wang, and J. Cai, "Dynamic population mapping with AutoGluon," *Urban Informatics*, vol. 1, no. 13, pp. 1–13, 2022.

[32] S. Malek, F. Melgani, Y. Bazi, and N. Alajlan, "Reconstructing cloud-contaminated multispectral images with contextualized autoencoder neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2270–2282, 2018.

[33] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4274–4288, 2018.

[34] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang, "Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 148–160, 2020.

[35] A. Sebastianelli, E. Puglisi, M. P. Del Rosso, J. Mifdal, A. Nowakowski, P. P. Mathieu, F. Pirri, and S. L. Ullo, "PLFM: pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[36] Z. Zhu, S. Qiu, and S. Ye, "Remote sensing of land change: a multifaceted perspective," *Remote Sensing of Environment*, vol. 282, pp. 113266, 2022.

[37] S. Kabir, N. Pahlevan, R. E. O'Shea, and B. B. Barnes, "Leveraging Landsat-8/-9 underfly observations to evaluate consistency in reflectance products over aquatic environments," *Remote Sensing of Environment*, vol. 296, pp. 113755, 2023.

[38] V. S. Martins, D. P. Roy, H. Huang, L. Boschetti, H. K. Zhang, and L. Yan, "Deep learning high resolution burned area mapping by transfer learning from Landsat-8 to PlanetScope," *Remote Sensing of Environment*, vol. 280, pp. 113203, 2022.

[39] J. Michel, O. Hagolle, S. J. Hook, J. Roujean, and P. Gamet, "Quantifying thermal infra-Red directional anisotropy using master and Landsat-8 simultaneous acquisitions," *Remote Sensing of Environment*, vol. 297, pp. 113765, 2023.

[40] Y. Zhai, D. P. Roy, V. S. Martins, H. K. Zhang, L. Yan, and Z. Li, "Conterminous United States Landsat-8 top of atmosphere and surface reflectance tasseled cap transformation coefficients," *Remote Sensing of Environment*, vol. 274, pp. 112992, 2022.

[41] M. Drusch, U. D. Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012.

[42] X. Liu, J. Frey, C. Munteanu, N. Still, and B. Koch, "Mapping tree species diversity in temperate montane forests using Sentinel-1 and Sentinel-2 imagery and topography data," *Remote Sensing of Environment*, vol. 292, pp. 113576, 2023.

[43] A. Radman, M. Mahdianpari, and D. J. Varon, F. Mohammadimanesh, "S2MetNet: a novel dataset and deep learning benchmark for methane point source quantification using Sentinel-2 satellite imagery," *Remote Sensing of Environment*, vol. 295, pp. 113708, 2023.

[44] B. Slagter, J. Reiche, D. Marcos, A. Mullissa, E. Lossou, M. Peña-Claros, and M. Herold, "Monitoring direct drivers of small-scale tropical forest disturbance in near real-time with Sentinel-1 and -2 data," *Remote Sensing of Environment*, vol. 295, pp. 113655, 2023.

[45] M. K. Vanderhoof, L. Alexander, J. Christensen, K. Solvik, P. Nieuwlandt, and M. Sagehorn, "High-frequency time series comparison of Sentinel-1 and Sentinel-2 satellites for mapping open and vegetated water across the United States (2017–2021)," *Remote Sensing of Environment*, vol. 288, pp. 113498, 2023.

[46] M. Wang, D. Mao, Y. Wang, X. Xiao, H. Xiang, K. Feng, L. Luo, M. Jia, K. Song, and Z. Wang, "Wetland mapping in East Asia by two-stage object-based Random Forest and hierarchical decision tree algorithms on Sentinel-1/2 images," *Remote Sensing of Environment*, vol. 297, pp. 113793, 2023.

[47] G. Doxani, E. F. Vermote, J. Roger, S. Skakun, F. Gascon, A. Collison, L. D. Keukelaere, C. Desjardins, D. Frantz, O. Hagolle, M. Kim, J. Louis, F. Pacifici, B. Pflug, H. Poilvé, D. Ramon, R. Richter, and F. Yin, "Atmospheric Correction Inter-comparison eXercise, ACIX-II Land: an assessment of atmospheric correction processors for Landsat 8 and Sentinel-2 over land," *Remote Sensing of Environment*, vol. 285, pp. 113412, 2023.

[48] R. Shang and Z. Zhu, "Harmonizing Landsat 8 and Sentinel-2: a time-series-based reflectance adjustment approach," *Remote Sensing of Environment*, vol. 235, pp. 111439, 2019.

[49] R. Shang, Z. Zhu, J. Zhang, S. Qiu, Z. Yang, T. Li, and X. Yang, "Near-real-time monitoring of land disturbance with harmonized Landsats 7-8 and Sentinel-2 data," *Remote Sensing of Environment*, vol. 278, pp. 113073, 2022.

[50] H. Duba-Sullivan, E. J. Reid, S. Voisin, C. A. Bouman, G. T. Buzzard, "ResSR: A residual approach to super-resolving multispectral images," *Image and Video Processing*, arXiv:2408.13225.

[51] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, K. Schindler, "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.

[52] C. Lanaras, J. Bioucas-Dias, E. Baltsavias and K. Schindler, "Super-Resolution of Multispectral Multiresolution Images from a Single Sensor," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, pp. 1505-1513, 2017.

[53] Q. Wang, W. Shi, Z. Li, and P.M., Atkinson, "Fusion of Sentinel-2 images," *Remote Sensing of Environment*, vol. 187, pp. 241–252, 2016.

[54] C. J. Crawford, D. P. Roy, S. Arab, C. Barnes, E. Vermote, G. Hulley, A. Gerace, M. Choate, C. Engebretson, E. Micijevic, G. Schmidt, C. Anderson, M. Anderson, M. Bouchard, B. Cook, R. Dittmeier, D. Howard, C. Jenkerson, M. Kim, T. Kleyians, T. Maiersperger, C. Mueller, C. Neigh, L. Owen, B. Page, N. Pahlevan, R. Rengarajan, J. Roger, K. Sayler, P. Scaramuzza, S. Skakun, L. Yan, H. K. Zhang, Z. and Zhu, S. Zahn, "The 50-year Landsat collection 2 archive," *Science of Remote Sensing*, vol. 8, pp. 100103, 2023.

[55] J. Guo, J. Yang, H. Yue, H. Tan, C. Hou and K. Li, "RSDehazeNet: Dehazing network with channel refinement for multispectral remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2535–2549, 2021.

[56] F. Xu, S. Heremans, and B. Somers, "Urban land cover mapping with Sentinel-2: a spectro-spatio-temporal analysis. Urban land cover mapping with Sentinel-2: a spectro-spatio-temporal analysis," *Urban Informatics*, vol. 1, no. 8, pp. 1–18, 2022.

[57] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imager," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5866–5878, 2021.

[58] R. Eckardt, C. Berger, C. Thiel, and C. Schmullius, "Removal of optically thick clouds from multi-spectral satellite images using multi-frequency SAR data," *Remote Sensing*, vol. 5, no. 6, pp. 2973–3006, 2013.

[59] A. Meraner, P. Ebel, X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020.

[60] E. Dalsasso, L. Denis, and F. Tupin, "SAR2SAR: A semi-supervised despeckling algorithm for SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4321–4329, 2021.

[61] D. Ienco, R. Interdonato, R. Gaetano, and D. Minh, "Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 11–22, 2019.

[62] J. Ma, H. Shen, P. Wu, J. Wu, M. Gao, and C. Meng, "Generating gapless land surface temperature with a high spatio-temporal resolution by fusing multi-source satellite-observed and model-simulated data," *Remote Sensing of Environment*, vol. 278, pp. 113083, 2022.
[63] Y. Yu, L. J. Renzullo, T. R. McVicar, B. P. Malone, and S. Tian, "Generating daily 100 m resolution land surface temperature estimates continentally using an unbiased spatiotemporal fusion approach," *Remote Sensing of Environment*, vol. 297, pp. 113784, 2023.
[64] X. Zhu, S. Duan, Z. Li, P. Wu, H. Wu, W. Zhao, and Y. Qian, "Reconstruction of land surface temperature under cloudy conditions from Landsat 8 data using annual temperature cycle model," *Remote Sensing of Environment*, vol. 281, pp. 113261, 2022.
[65] Y. Kim, H. Park, J. S. Kimball, A. Colliander, and M. F. McCabe, "Global estimates of daily evapotranspiration using SMAP surface and root-zone soil moisture," *Remote Sensing of Environment*, vol. 298, pp. 113803, 2023.
[66] T. Schmidt, M. Schrön, Z. Li, T. Francke, S. Zacharias, A. Hildebrandt, and J. Peng, "Comprehensive quality assessment of satellite- and model-based soil moisture products against the COSMOS network in Germany," *Remote Sensing of Environment*, vol. 301, pp. 113930, 2024.
[67] L. Zhu, J. Dai, Y. Liu, S. Yuan, T. Qin, and J. P. Walker, "A cross-resolution transfer learning approach for soil moisture retrieval from Sentinel-1 using limited training samples," *Remote Sensing of Environment*, vol. 301, pp. 113944, 2024.

Lanxing Wang received the B.S. degree from Changsha University of Science & Technology, Changsha, China, in 2019. She is currently working toward the M.S. degree at Tongji University, Shanghai, China. Her research interests include remote sensing data reconstruction.

Qunming Wang received the Ph.D. degree from the Hong Kong Polytechnic University, Hong Kong, in 2015.

He is currently a Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. He was a Lecturer (Assistant Professor) with Lancaster Environment Centre, Lancaster University, Lancaster, U.K., from 2017 to 2018. His 3-year Ph.D. study was supported by the hypercompetitive Hong Kong Ph.D. Fellowship and his Ph.D. thesis was awarded as the Outstanding Thesis in the Faculty. He has authored or coauthored over 90 peer-reviewed articles in international journals such as *Remote Sensing of Environment*, *IEEE Transactions on Geoscience and Remote Sensing*, and *ISPRS Journal of Photogrammetry and Remote Sensing*. His research interests include remote sensing, image processing, and geostatistics.

Dr. Wang serves as Associate Editor for *Science of Remote Sensing* (sister journal of *Remote Sensing of Environment*) and *Photogrammetric Engineering & Remote Sensing*, and was Associate Editor for *Computers and Geosciences* (2017−2020).

Xiaohua Tong received the Ph.D. degree in traffic engineering from Tongji University, Shanghai, China, in 1999.

He is currently a Professor with the College of Surveying and Geoinformatics, Tongji University. He is also an Academician at the Chinese Academy of Engineering. He was a Research Fellow with Hong Kong Polytechnic University, Hong Kong, in 2006, and a Visiting Scholar with the University of California, Santa Barbara, CA, USA, between 2008 and 2009. His research interests include remote sensing, geographic information system, uncertainty and spatial data quality, and image processing for high-resolution and hyperspectral images.

Peter M. Atkinson received the Ph.D. degree from the University of Sheffield (NERC CASE award with Rothamsted Experimental Station) in 1990. More recently, he received the MBA degree from the University of Southampton in 2012.

He is currently Distinguished Professor of Spatial Data Science and Dean of the Faculty of Science and Technology at Lancaster University, UK. He was previously Professor of Geography at the University Southampton, where he is currently Visiting Professor. He is also Visiting Professor at the Chinese Academy of Sciences, Beijing. He previously held the Belle van Zuylen Chair at Utrecht University, the Netherlands, is a recipient of the Peter Burrough Award of the International Spatial Accuracy Research Association and is a Fellow of the Learned Society of Wales. The main focus of his research is in remote sensing, geographical information science and spatial (and space-time) statistics applied to a range of environmental science and socio-economic problems. He has published over 300 peer-reviewed articles in international scientific journals and around 50 refereed book chapters. He has also edited nine journal special issues and eight books.

Professor Atkinson is Editor-in-Chief of *Science of Remote Sensing*, a sister journal of *Remote Sensing of Environment*. He also sits on the editorial boards of several further journals including *Geographical Analysis*, *Spatial Statistics*, *International Journal of Applied Earth Observation and Geoinformation*, and *Environmental Informatics*.