M^2-STF: Integration of Multi-modal Data for Spatio-temporal Fusion

Qunming Wang, Aijing Li, Yijie Tang, Peter M. Atkinson

***Abstract*—Spatio-temporal fusion is a general technique used to blend fine spatial resolution and fine temporal resolution remote sensing data from multiple sensors, to generate time-series data with both fine spatial and temporal resolutions. It has received increasing attention in recent years. Drastic changes in land surface, however, pose great challenges for spatio-temporal fusion. To address this issue, this paper proposed a spatio-temporal fusion method which integrates multi-modal data (M^2-STF), specifically SAR data with optical data. Considering the scenario of flooding (which causes drastic land surface changes) as an example, this study focused on spatio-temporal fusion based on Sentinel-2 MSI and Sentinel-3 OLCI data, and developed the M^2-STF method by integrating Sentinel-1 SAR data. For the changed area, M^2-STF integrates the Sentinel-2 image at the known time and the Sentinel-1 SAR image at the prediction time to obtain a more accurate fine spatial resolution classification map at the prediction time. Based on this map, a spatial unmixing model and spatial interpolation model were developed taking into account both homogeneity and heterogeneity characteristics, which were then combined into a homogeneity index. For the unchanged area, a new similar pixel selection strategy was constructed to exclude the influence of similar pixels from the changed area. In the experiments, three regions were selected for validation, and M^2-STF was compared with five typical spatio-temporal fusion methods. By integrating Sentinel-1 SAR data at the prediction time, the accuracy of spatio-temporal fusion was increased remarkably, especially when the land surface changes greatly from the known to the prediction times. Specifically, the M^2-STF method outperforms all five benchmark methods, by reducing root mean square error (RMSE) by at least 16%.**

***Index Terms*—Sentinel-1; Sentinel-2; Sentinel-3; spatio-temporal fusion; multi-modal.**

I. INTRODUCTION

With the increasing availability of remote sensing data at various spatial and temporal resolutions, unprecedented opportunities are arising for monitoring surface dynamics. However, due to the inherent trade-off between spatial resolution and revisit period, satellite sensor images from a single data source often fail to meet the requirements for fine spatial and temporal resolution monitoring. Spatio-temporal fusion is a technique to cope with this limitation. That is, by fusing fine spatial resolution, but temporally sparse images with temporally dense, but coarse spatial resolution images, images with both fine spatial and temporal resolutions can be created. In recent years, methods for spatio-temporal fusion have been developed rapidly [1-6] for a wide range of applications, including the generation of fine spatio-temporal resolution surface temperature [7-9], soil moisture [10], surface evapotranspiration [11] and vegetation indices [12, 13]. Existing spatio-temporal fusion methods can be divided mainly into four categories: spatial weighting-based, spatial unmixing-based, learning-based and hybrid methods.

The most typical spatial weighting-based method is the spatial and temporal adaptive reflectance fusion model (STARFM) [14], which is suitable for areas with no changes in land cover types and strong homogeneity. The enhanced STARFM (ESTARFM) [15] extends STARFM for areas with great heterogeneity. To overcome the challenge of strong seasonal variation between images at different times, Wang and Atkinson [16] proposed the Fit-FC model. Additionally, to reduce the uncertainty in estimation of the fine spatial resolution increment, the virtual image pair-based spatio-temporal fusion (VIPSTF) [17] method was proposed.

Spatial unmixing-based methods evolved mainly from the original multisensor multiresolution technique (MMT) [18]. MMT was performed based on the assumption that the distribution of land cover classes remains unchanged during the study period, and it utilized a classification map based on the known fine spatial resolution image to inversely calculate the class reflectance in the coarse spatial resolution image at the prediction time. For example, Zurita-Milla *et al*. [19] employed unmixing-based data fusion (UBDF) to generate images with Landsat TM spatial resolution and MERIS spectral resolution. The spatial temporal data fusion approach (STDFA) [20] calculated the fine spatial resolution temporal variation in reflectance for each class by unmixing coarse difference images (between the known and prediction images), and then added it to the known fine spatial resolution image to obtain the fusion results. Liu *et al*. [21] applied linear spectral analysis to the fine spatial resolution image to obtain a fine resolution abundance map at the known time, which was used to predict the class reflectance in the coarse image at the prediction time. The final prediction was a linear combination of the class reflectance and corresponding fine spatial resolution abundance.

Hybrid methods integrate the advantages of the above two types of methods. The flexible spatiotemporal data fusion (FSDAF) [22] is a representative method of this type, which uses spatial unmixing to obtain the temporal change in reflectance of each land cover class, and maintains the coarse spatial resolution data at the prediction time through spatial weighting. FSDAF can cope with various challenging scenarios, especially when land surface changes occur. Based on this, several improved

Manuscript received 8 July 2025; revised 16 September 2025; accepted 30 October 2025. This research was supported by the National Natural Science Foundation of China under Grants 42222108 and 42171345. *(Corresponding author: Q. Wang.)*

Q. Wang, A. Li and Y. Tang 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; Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

FSDAF methods have been proposed, including IFSDAF [23], SFSDAF [24], FSDAF 2.0 [25], cuFSDAF [26] and cuFSDAF 2.0 [27]. Moreover, object-level-based spatio-temporal fusion methods also demonstrated satisfactory performance, including the object-based spatial and temporal vegetation index unmixing model [28], object-restricted strategy [29], object-based spatiotemporal fusion model [30], object-level hybrid spatiotemporal fusion method [31], and the object-based spatial unmixing model [32]. Additionally, Bayesian-based methods [33-36] have also been developed.

Learning-based methods focus primarily on modeling the complex relationship between the coarse and fine spatial resolution images. Early methods include Sparse representation-based spatio-temporal reflectance fusion model (SPSTFM) [37]. Recently, deep learning-based methods have received increasing attention. For example, the enhanced cross-paired wavelet based spatiotemporal fusion network (ECPW-STFN) [38] was proposed to separately train the high- and low-frequency components of images to better extract features of different levels. The methods based on generative adversarial network (GAN) have also been developed, such as the GAN-based spatiotemporal fusion model (GAN-STFM) [39] and the multilevel feature fusion with generative adversarial network (MLFF-GAN) [40]. These approaches do not estimate the data samples distribution directly, but model the latent distribution and generate new samples from it through adversarial learning. In addition, Transformer has shown promising results in time-series change feature extraction. It possesses a unique self-attention mechanism that effectively captures long-term dependencies while enabling more efficient parallel computing. For example, the spatial-temporal integration network (STINet) [41] was effective in fusing multiscale spatiotemporal dynamic features to reconstruct vegetation changes. The multistage remote sensing image spatio-temporal fusion network (MSFusion) [42] combined the advantages of Transformer and Convolutional Neural Network to adaptively fuse multi-scale features. The swin spatiotemporal fusion model (SwinSTFM) [43] constructed a hybrid-based model by integrating swin transformer and linear unmixing theories.

The essence of the existing spatio-temporal fusion methods is to estimate the fine spatial resolution incremental information in addition to the known fine spatial resolution image, with the former obtained mainly by downscaling the corresponding coarse spatial resolution increment. The estimation process, however, contains great uncertainty in the case of drastic land surface changes. The variation-based spatiotemporal data fusion method (VSDF) [44] was developed to detect different land surface changes by improving the unmixing model. The Agri-Fuse method [45] emphasized change information for agricultural scenarios with different phenological changes. These methods, however, mainly targeted attribute, and not geometric, changes of land cover, that is, scenarios where the boundaries of land cover remain unchanged. In reality, the shape or geometry of land cover objects usually changes over time, especially in sudden change scenarios. For example, after a sudden flood, the boundaries of both the water area and the surroundings can change drastically. In this case, it is difficult to accurately predict the changed water boundary at fine spatial resolution by relying only on the coarse-fine spatial resolution image pairs at the known time (i.e., before the flood event) and the coarse spatial resolution image at the prediction time.

To address the challenge of drastic land surface changes, and focusing specifically on the example of changes in surface water caused by flooding, a spatio-temporal fusion method integrating multi-modal data (M^2-STF) was proposed in this paper. It should be noted that existing spatio-temporal fusion missions focus mainly on fusing 30 m, 16-day Landsat images with 500 m, daily MODIS images, to create 30 m, daily time-series images. In this paper, Sentinel-2 MSI and Sentinel-3 OLCI were selected for spatio-temporal fusion. Sentinel-2 provides fine spatial resolution data at 10 m. Although its twin satellites can increase the temporal resolution to 5 days [46], the widespread cloud contamination can reduce the effective temporal resolution, especially in cloudy and rainy regions. The Sentinel-3 twin satellites can provide OLCI data with a temporal resolution of < 1.4 days, whereas the 300 m spatial resolution is too coarse for regional monitoring. By fusing Sentinel-2 MSI with Sentinel-3 OLCI data, 10 m time-series data with finer frequency can be obtained. Compared to the fusion of 30 m Landsat and 500 m MODIS data, the fusion of 10 m Sentinel-2 and 300 m Sentinel-3 data involves a much larger zoom factor of 30, which is almost twice for the fusion of 500 m MODIS to 30 m Landsat. This results in greater technical challenges, especially in the case of drastic land surface changes.

To deal with the challenges brought by drastic land surface changes, by M^2-STF, this paper innovatively introduces the use of Sentinel-1 SAR data into the spatio-temporal fusion of Sentinel-2 MSI and Sentinel-3 OLCI data. Sentinel-1 can penetrate clouds and fog, offering all-day and all-weather Earth observation. Sentinel-1 data are also freely available and the spatial resolution is consistent with Sentinel-2, providing effective representation of land cover at the prediction time. Additionally, for water areas, Sentinel-1 microwave signals are specularly reflected, resulting in lower signal intensity. Thus, the backscattering coefficient value of Sentinel-1 images in water areas is obviously lower than that of vegetation and other land cover. That is, Sentinel-1 SAR can effectively distinguish between water and other land cover types, bringing unique advantages for water extraction at fine spatial resolution [47-50].

The proposed M^2-STF method treats the changed area (i.e., the flood area) and the unchanged area separately. For the changed area, the fine spatial resolution water area boundary information obtained from Sentinel-1 is utilized to redefine the water mask at the prediction time, and based on this a new spatial unmixing method is constructed. For the unchanged area, a new similar pixel selection strategy was developed based on the water mask extracted from Sentinel-1, which effectively avoids the influence of similar pixels (identified by existing schemes) that have actually changed. The contributions of this paper are as follows:

1) Multi-modal data were integrated into the spatio-temporal fusion method to address the challenging issue of drastic land surface changes. Different from traditional spatio-temporal fusion methods based on the assumption that the land surface does not change (or change slowly), this paper focuses on drastic changes in land surface boundaries. Specifically, in the process of spatio-temporal fusion of Sentinel-2 MSI and Sentinel-3 OLCI data, Sentinel-1

SAR data were introduced to accurately estimate the change of land surface boundaries.

2) The M^2-STF model was proposed. This model generates a more accurate fine classification map by integrating the known fine spatial resolution classification map with the fine spatial resolution water mask at the prediction time. Based on this, a spatial unmixing model and a homogeneity index were established to estimate the fine spatial resolution reflectance in the changed area. For the unchanged area, a new similar pixel selection strategy was developed to eliminate the effect of the incorrectly identified similar pixels in traditional models.

The remainder of this paper is organized into four sections. Section II introduces the study data and the proposed M^2-STF approach. Section III presents the experiments involving validation on three datasets, and compares M^2-STF with five typical spatio-temporal fusion methods. Moreover, the influences of water mask and degrees of water changes on predictions are investigated. Section IV further discusses the experimental results and potential future research. Finally, Section V concludes the paper.

II. Methods

A. Data and study area

1) Data

In this paper, 10 m Sentinel-2 MSI and 300 m Sentinel-3 OLCI cloud-free images were acquired as the fine and coarse spatial resolution images, respectively. Additionally, during the fusion process, Sentinel-1 (10 m) data were introduced to obtain water boundary information at the prediction time. For Sentinel-2 MSI and Sentinel-3 OLCI, the wavebands of the two sensors are similar in the blue, green, red and near-infrared parts of the spectrum (i.e., bands 2, 3, 4, and 8a in Sentinel-2 and bands Oa4, Oa6, Oa8, and Oa17 in Sentinel-3). Thus, only these four bands were considered in the spatio-temporal fusion process. The data were downloaded from the ESA platform. Specifically, Sentinel-1 Level-1 GRD and Sentinel-2 L2A data were pre-processed in SNAP to obtain backscattering coefficients and surface reflectance, respectively. The Sentinel-1 pre-processing steps mainly include orbit correction, thermal noise removal, radiation calibration, speckle filtering and terrain correction. Notably, the Sentinel-2 band 8a (20 m) data were downscaled to 10 m by fusion with the 10 m bands using the area-to-point regression kriging (ATPRK) approach [51]. For Sentinel-3 OLCI Level-1 EFR data, geometric correction, radiometric calibration and atmospheric correction operations were performed in ENVI 5.6. Finally, using the Sentinel-2 image as reference, the Sentinel-1 and Sentinel-3 images were projected, georeferenced and cropped to the same area. Considering the difference between the Sentinel-2 MSI and Sentinel-3 OLCI sensor systems, in this paper, a linear model was thus, used to describe the relationship between the Sentinel-2 and -3 data at the known time, which was used to correct the Sentinel-3 image at the prediction time.

2) Study area

We selected three regions to validate the proposed M^2-STF in the experiments. The first region (12 km × 12 km) is located in the Hawkesbury River (150.8311 °E, 33.5883 °S) in Richmond, around Sydney (called Region 1 hereafter). The Hawkesbury River is one of the most important rivers in New South Wales, Australia, and one of the most important waterways in the Sydney region. It is surrounded by numerous historic towns and settlements. However, the river is prone to challenges such as frequent flooding. The second region (24 km × 24 km) is the Jamuna River (89.7862 °E, 25.7789 °N) in Bangladesh (called Region 2 hereafter). The Jamuna River is a major river of Bangladesh as well as one of the major tributaries of the Ganges River in India. It is a classic braided channel and its floodplain is an important agricultural cultivation area. The third region (9 km × 9 km) is located in the Koga Reservoir (11.3584 °E, 37.1697 °N) in Ethiopia (called Region 3 hereafter). The region is characterized by gradual expansion of dynamic watersheds, driven mainly by changes in rainfall. Ethiopia frequently experiences high-intensity rainfall and soil erosion is common in the Ethiopian highlands [52].

Table 1 Information on the data for the three study regions

Study area	Base date	Prediction date	CC
Hawkesbury River (Region 1)	February 18, 2022 (20220218)	March 10, 2022 (20220310)	0.6431
Jamuna River (Region 2)	March 20, 2022 (20220320)	October 21, 2022 (20221021)	0.4685
Koga Reservoir (Region 3)	May 21, 2023 (20230521)	June 30, 2023 (20230630)	0.6374
		October 03, 2023 (20231003)	0.3154

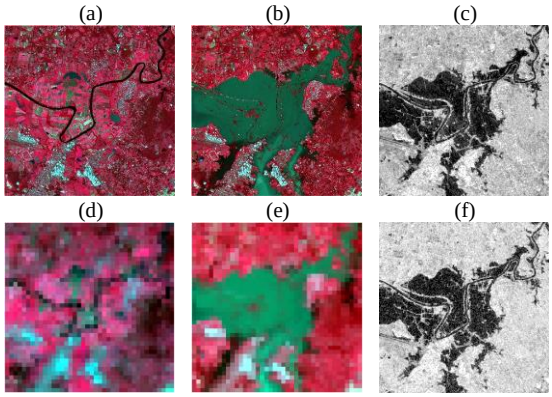


Fig. 1. The data used in Region 1. (a) and (b) are Sentinel-2 data on 20220218 (base date) and 20220310 (prediction date), respectively. (d)-(e) are the corresponding Sentinel-3 data. (c) and (f) are Sentinel-1 VV and VH on the prediction date, respectively.

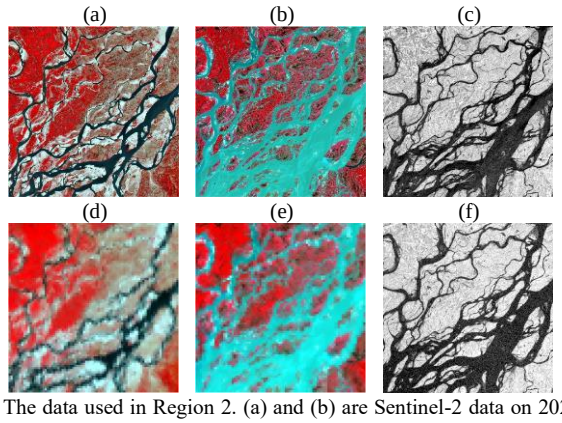


Fig. 2. The data used in Region 2. (a) and (b) are Sentinel-2 data on 20220320 (base date) and 20221021 (prediction date), respectively. (d)-(e) are the corresponding Sentinel-3 data. (c) and (f) are Sentinel-1 VV and VH on the prediction date, respectively.

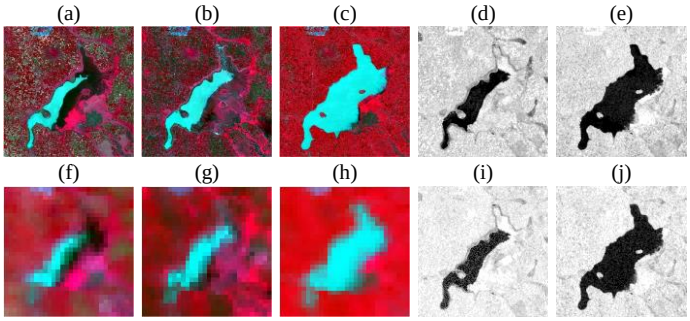


Fig. 3. The data used in Region 3. (a)-(c) are Sentinel-2 data on 20230521 (base date), 20230630 (first prediction date) and 20231003 (second prediction date), respectively. (f)-(h) are the corresponding Sentinel-3 data. (d)-(i) are Sentinel-1 VV and VH that are temporally closest to the first prediction date (i.e., 20230701), respectively. (e)-(j) are Sentinel-1 VV and VH on the second prediction date, respectively.

The detailed acquisition dates of the images for the three regions are shown in Table 1. The average correlation coefficient (CC) of the four bands between the images on the known and prediction dates in Region 1 is 0.6431. For Region 2, the corresponding CC is 0.4685. For Region 3, images on two dates need to be predicted, and the CC between the base and two prediction dates is 0.6374 and 0.3154. The data for the three regions are shown in Fig. 1-3. It can be seen that, in each region, heavy rainfall led to extensive flooding resulting in significant land surface changes. The Sentinel-1 data at the prediction time can effectively capture the distribution of water after flooding.

B. The proposed M²-STF approach

In the proposed method, the input data include a pair of coarse ($\mathbf{C}_{t1}$) - fine ($\mathbf{F}_{t1}$) spatial resolution images acquired at t_1 (base date), and a coarse image ($\mathbf{C}_{t2}$) and SAR image acquired at t_2 (prediction date). The target is to predict the fine spatial resolution image at t_2. M²-STF consists of three main steps: 1) Land cover classification at t_2. The classification map at t_2 was generated by combining the fine spatial resolution classification map at t_1 with the water mask extracted from the SAR data at t_2; 2) Prediction for the changed area. For the changed area, the fine spatial resolution reflectance at t_2 was predicted by the proposed spatial unmixing model and homogeneity index; 3) Prediction for the unchanged area. $\mathbf{F}_{t1}$ was used to determine similar pixels, based on the unchanged area identified by the classified map at t_2. The overall flowchart is shown in Fig. 4.

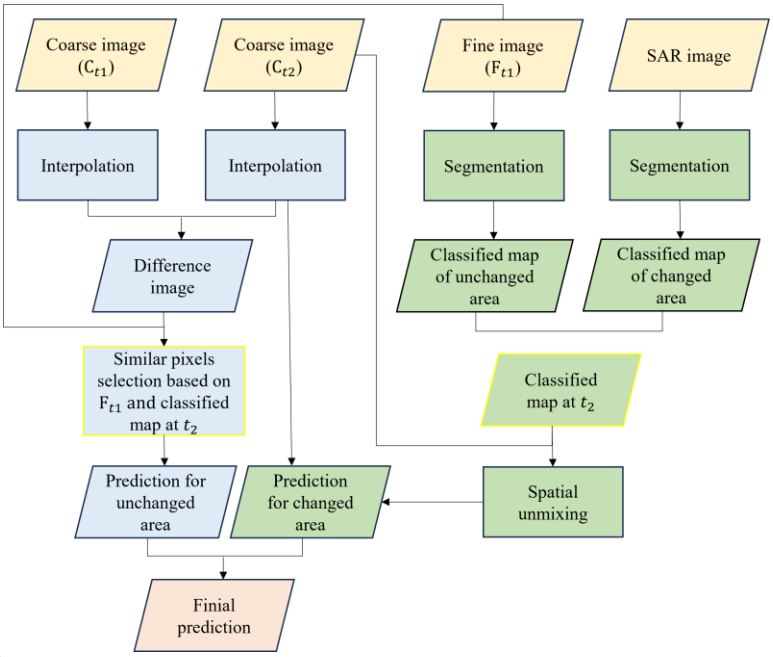


Fig. 4. Flowchart of the proposed M²-STF method.

1) Land cover classification at the prediction time

This paper investigated spatio-temporal fusion based on the scenario of drastic water area expansion (e.g., flooding). One of the keys lies in the accurate generation of the fine spatial resolution classification map at t_2. Existing strategies usually assume that the land surface remains unchanged and the classification map is obtained directly from the fine spatial resolution image at t_1. In contrast, this paper proposed a new method to obtain a more accurate land cover map at t_2 by integrating the fine spatial resolution Sentinel-1 data at t_2. The core of the process lies in the accurate extraction of the water mask (i.e., classification of the drastically changed area). The water area shows great spatial homogeneity in Sentinel-1 data, and its signal is obviously different from other land cover types. In this paper, the water area was extracted by a threshold segmentation method using Sentinel-1 Dual-Polarized Water Index (SDWI) defined as follows [53]:

$$\mathrm{SDWI} = \ln\left(10 \times \mathrm{VH} \times \mathrm{VV}\right) - 8 \quad (1)$$

The water area was defined as that with SDWI greater than 0. This threshold was determined after multiple trials and was identified as the optimal value for effectively distinguishing water bodies from non-water surfaces in our study areas. Additionally, it is noted that to reduce the impact of noise in the Sentinel-1 data, water area objects extracted with the size of a single pixel were excluded.

For the non-water area, it was assumed that no land surface changes occurred. Based on this, the land cover classification

result of this area can be obtained by classifying the fine spatial resolution image at t_1 into five categories. In this paper, the *K*-Means method was used to classify the Sentinel-2 image at t_1. The final fine spatial resolution classification map at t_2 was obtained by combining the classification results of the water and the non-water areas.

2) Prediction for the changed area

For the land surface changed area (i.e., water area), we fully mine the coarse spatial resolution information in the Sentinel-3 image at t_2 to obtain the reflectance at fine spatial resolution. Generally, the landscape can be divided into homogeneous and heterogeneous types. For homogeneous landscapes, the fine spatial resolution homogeneity information can be characterized by interpolating the original coarse resolution data (e.g., using bicubic interpolation). For heterogeneous landscapes, the fine resolution classification map can be utilized to capture the corresponding fine spatial resolution heterogeneity such that a spatial unmixing model can be used to predict reflectance. Specifically, different from existing strategies that use the classification map at t_1 directly, the proposed spatial unmixing model employs the fine spatial resolution classification map at t_2 (obtained in Section II-B 1)) to predict the class reflectance in a moving window:

$$\begin{bmatrix} C_{t2}(x_1,y_1,b) \\ \vdots \\ C_{t2}(x_i,y_i,b) \\ \vdots \\ C_{t2}(x_m,y_m,b) \end{bmatrix} = \begin{bmatrix} f_1(x_1,y_1) \cdots f_c(x_1,y_1) \cdots f_l(x_1,y_1) \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ f_1(x_i,y_i) \cdots f_c(x_i,y_i) \cdots f_l(x_i,y_i) \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ f_1(x_m,y_m) \cdots f_c(x_m,y_m) \cdots f_l(x_m,y_m) \end{bmatrix} \begin{bmatrix} F_{t2}^{\mathrm{U}}(1,b) \\ \vdots \\ F_{t2}^{\mathrm{U}}(c,b) \\ \vdots \\ F_{t2}^{\mathrm{U}}(l,b) \end{bmatrix} \tag{2}$$

where m is the number of coarse pixels in the moving window, l is the number of classes, $C_{t2}(x_i, y_i, b)$ is the reflectance of the coarse pixel at (x_i, y_i) in band b at t_2, $f_c(x_i, y_i)$ is the proportion of class c for the coarse pixel located at (x_i, y_i), and $F_{t2}^{\mathrm{U}}(c, b)$ is the predicted (by the Unmixing model) reflectance of class c at (x_i, y_i) in band b at t_2. For the changed area at t_2, the class reflectance of each fine pixel can be regarded as the expression of the corresponding heterogeneity information of the pixel. For the changed area at t_2, the reflectance of any fine pixel can be predicted by combining the above two types of results. Specifically, we proposed a homogeneity index (I_{ho}) to describe the weights for the two cases to allow combination:

$$\begin{aligned} F_{2\_\mathrm{ch}}(x_i,y_i,b) = {} & \left[1 - I_{\mathrm{ho}}(x_i,y_i,b)\right] \times F_{t2}^{\mathrm{U}}\left(c(x_i,y_i),b\right) + \\ & I_{\mathrm{ho}}(x_i,y_i,b) \times F_{t2}^{\mathrm{In}}(x_i,y_i,b) \end{aligned} \tag{3}$$

where $F_{2\_\mathrm{ch}}(x_i, y_i, b)$ is the predicted reflectance of the fine pixel at changed location (x_i, y_i) in band b at t_2. $F_{t2}^{\mathrm{U}}(c(x_i, y_i), b)$ is the corresponding result of spatial unmixing and $c(x_i, y_i)$ is the land cover class at t_2 for the fine pixel located at (x_i, y_i). $F_{t2}^{\mathrm{In}}(x_i, y_i, b)$ is the corresponding result of bicubic interpolation. I_{ho} was calculated based on the spatial homogeneity (denoted as I_{spatial}) and the spectral similarity (denoted as I_{spectral}) between the two fine spatial resolution results (i.e., the spatial unmixing result $\mathbf{F}_{t2}^{\mathrm{U}}$ and the interpolation result $\mathbf{F}_{t2}^{\mathrm{In}}$):

$$I_{\mathrm{ho}}(x_i,y_i,b) = I_{\mathrm{spectral}}(x_i,y_i,b) \times I_{\mathrm{spatial}}(x_i,y_i,b) \tag{4}$$

Theoretically, the higher the similarity between the spatial unmixing result $\mathbf{F}_{t2}^{\mathrm{U}}$ and the interpolation result $\mathbf{F}_{t2}^{\mathrm{In}}$, the stronger the homogeneity of the region. To simplify the calculation, it is assumed that the difference between the two images follows a Gaussian distribution, and it is considered that there is no spectral similarity when the difference exceeds three standard deviations from the mean difference. Its expression is as follows:

$$\begin{aligned} & I_{\mathrm{spectral}}(x_i,y_i,b) = 1 - \\ & \left| D_{\mathrm{UI}}(x_i,y_i,b) - \mathrm{mean}(D_{\mathrm{UI}}) \right| \Big/ \left(3 \times \mathrm{sd}(D_{\mathrm{UI}})\right) \end{aligned} \tag{5}$$

where D_{UI} is the difference value between the spatial unmixing result and the interpolation result. $\mathrm{mean}(D_{\mathrm{UI}})$ is the average difference in band b. $\mathrm{sd}(D_{\mathrm{UI}})$ is the standard deviation of the difference in band b. I_{spectral} ranges from 0 to 1, and larger values indicate greater spectral similarity.

The spatial homogeneity index I_{spatial} reflects the amount of smoothness in land cover. The higher the homogeneity of the image, the less spatial detail in the land cover and, consequently, the less that detailed information may be lost through interpolation. The spatial homogeneity of the fine spatial resolution image at t_2 is described as follows:

$$I_{\mathrm{spatial}}(x_i,y_i,b) = \sin\left[\left(\frac{1}{n}\sum_{p=1}^{n} L_p\right) \times \frac{\pi}{2}\right] \tag{6}$$

where n is the number of fine spatial resolution pixels in the moving window. Based on the fine spatial resolution classification map at t_2 (obtained in Section II-B 1)), L_p was set to 1 if the p-th fine pixel in the moving window has the same land cover type as the center pixel; otherwise, L_p=0. I_{spatial} ranges from 0 to 1, with larger values indicating stronger spatial homogeneity.

3) Prediction for the unchanged area

For the unchanged area, this paper proposed a new similar pixel selection strategy to avoid the influence of pixels that have actually changed. Specifically, different from the traditional strategy that searches similar pixels based on the spectral information of the fine resolution image at t_1, we excluded all pixels in the changed area and searched similar pixels only in the unchanged area according to the fine spatial resolution classification map at t_2. Specifically, in the moving window, we selected pixels that have the greatest spectral similarity with the center pixel in the unchanged area. After similar pixels are determined, the weight of each similar pixel is further calculated to determine its contribution to the central pixel:

$$\begin{aligned} F_{2\_\mathrm{uc}}(x_0,y_0,b) = {} & F_{t1}(x_i,y_i,b) + \\ & \sum_{i=1}^{w}\sum_{i=1}^{w} W_{ij} \times \left(C_{t2}(x_i,y_i,b) - C_{t1}(x_i,y_i,b)\right) \end{aligned} \tag{7}$$

where (x_0, y_0) is the position of the central (unchanged) pixel in the moving window, $F_{2\_\mathrm{uc}}(x_0, y_0, b)$ is the predicted reflectance of the fine pixel at unchanged location (x_0, y_0) in band b at t_2, $C_{t1}(x_i, y_i, b)$ is the reflectance of the coarse pixel at (x_i, y_i) in band b at t_1, and $F_{t1}(x_i, y_i, b)$ is the reflectance of the fine pixel at (x_i, y_i) in band b at t_1. w is the size of the moving window (in units of fine pixels). W_{ij} is the weight of neighboring similar pixels, which is calculated based on the spectral similarity and spatial distance between the pixels.

Combining the prediction $\mathbf{F}_{2\_\mathrm{ch}}$ for the changed area and $\mathbf{F}_{2\_\mathrm{uc}}$ for the unchanged area, the final prediction for the fine spatial resolution image at t_2 is obtained.

III. Experiments

The three datasets introduced in Section II-A were used for validation of the proposed M^2-STF method. In Section III-A, we compared the performance of the proposed M^2-STF approach with several typical spatio-temporal fusion methods. Note that deep learning-based spatio-temporal fusion methods were not considered for comparison, as they generally need at least two coarse-fine spatial resolution image pairs for training. Alternatively, the proposed method requires only one image pair and does not involve any training process. Thus, for fair comparison, the benchmark methods requiring only one image pair were considered. Specifically, the benchmark methods include two spatial weighting-based methods (STARFM and VIPSTF-SW), two spatial unmixing-based methods (UBDF and STDFA), and one hybrid method (FSDAF). In Section III-B, we examined the effectiveness of the proposed strategy of water mask incorporation on different spatio-temporal fusion methods. Section III-C investigates the impact of the range of water changes on various fusion methods. Section III-D examines the effect of several modules in the proposed M^2-STF approach. Five evaluation metrics were utilized in the experiments for accuracy assessment: CC, Universal Image Quality Index (UIQI), Root Mean Square Error (RMSE), Normalized Global Error (ERGAS) and Spectral Angle Mapper (SAM).

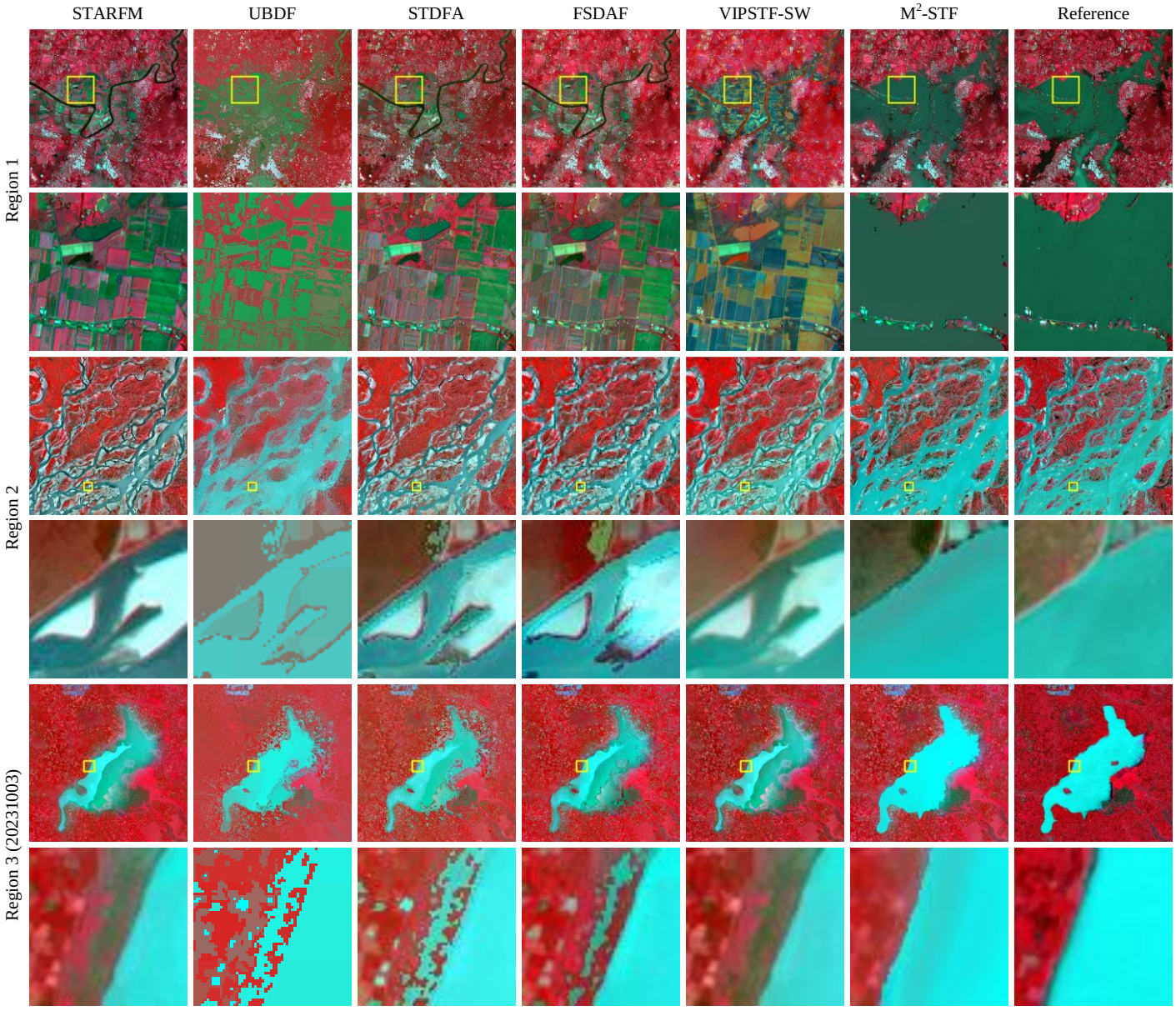


Fig. 5. Results of different methods for Regions 1-3 (NIR, red, and green bands as RGB).

A. Comparison with other methods

The results of different methods for the three datasets are shown in Fig. 5. From visual inspection, it is seen that M^2-STF produces the results closest to the reference image across all three regions. For example, in Region 2, where the water area varies greatly between the known and prediction times, there are notable errors in the predictions for the water area by the STARFM, STDFA, FSDAF and VIPSTF-SW methods. Compared with the four methods, the UBDF method produces more accurate prediction for the water area, as it relies heavily

on the coarse spatial resolution image at the prediction time for spatial unmixing. However, it can be seen from the sub-area that UBDF still fails to accurately predict the water boundary. In contrast, by integrating fine spatial resolution SAR data at the prediction time, M^2-STF not only produces the most accurate water boundary, but also the hue closest to the reference image. Similarly, in Regions 1 and 3 (20231003), checking the water boundary and land cover hue, M^2-STF also produces the most accurate results.

Table 2 Accuracy of different methods in Region 1

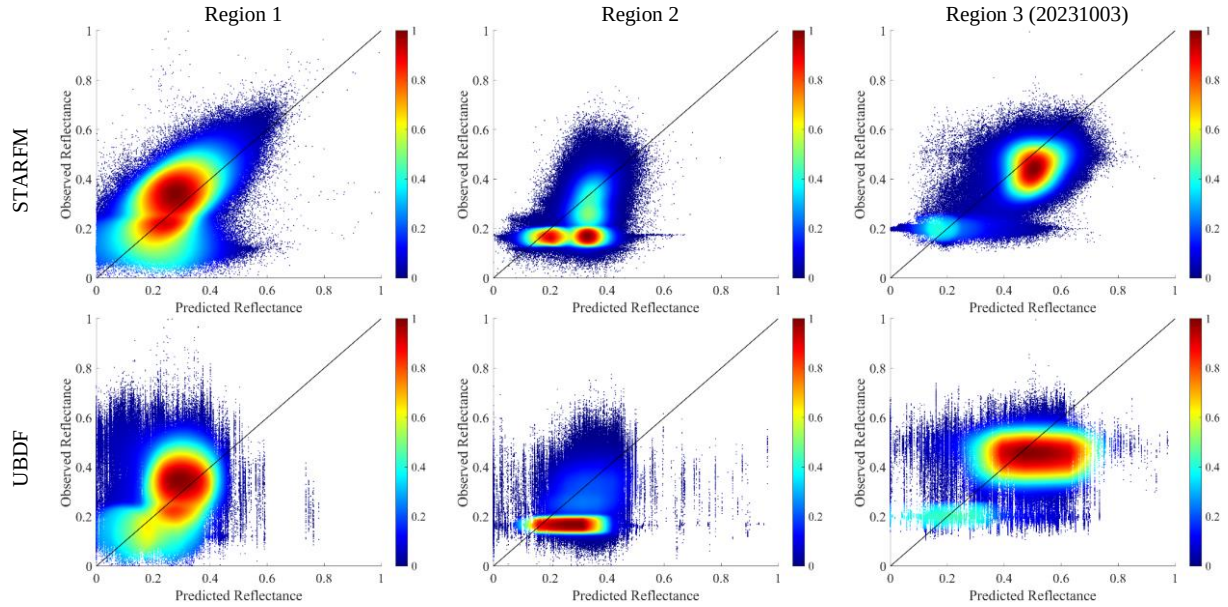
		STARFM	UBDF	STDFA	FSDAF	VIPSTF-SW	M^2-STF
CC	Blue	0.8685	0.6164	0.8651	0.8634	0.8629	**0.8886**
	Green	0.8219	0.5962	0.8202	0.8247	0.8219	**0.8651**
	Red	0.8043	0.5784	0.7694	0.8037	0.8236	**0.8619**
	NIR	0.6103	0.4127	0.4595	0.6150	0.4739	**0.8241**
	Average	0.7763	0.5509	0.7286	0.7767	0.7456	**0.8599**
UIQI	Blue	0.8378	0.4690	0.8340	0.8326	0.8101	**0.8554**
	Green	0.8061	0.4844	0.8037	0.8106	0.7716	**0.8458**
	Red	0.7988	0.4922	0.7637	0.7972	0.7835	**0.8522**
	NIR	0.6077	0.3908	0.4567	0.5907	0.4384	**0.7879**
	Average	0.7626	0.4591	0.7145	0.7578	0.7009	**0.8353**
RMSE	Blue	0.0352	0.0401	0.0353	0.0350	0.0353	**0.0346**
	Green	0.0331	0.0376	0.0332	0.0324	0.0326	**0.0316**
	Red	0.0295	0.0357	0.0313	0.0287	0.0278	**0.0253**
	NIR	0.1152	0.1269	0.1330	0.1053	0.1193	**0.0801**
	Average	0.0532	0.0601	0.0582	0.0504	0.0537	**0.0429**
ERGAS		0.9463	1.0662	1.0358	0.8964	0.9581	**0.7756**
SAM		0.2970	0.2717	0.3435	0.2433	0.2324	**0.1568**

Table 3 Accuracy of different methods in Region 2

		STARFM	UBDF	STDFA	FSDAF	VIPSTF-SW	M^2-STF
CC	Blue	0.6648	0.6394	0.6847	0.7184	0.7188	**0.7637**
	Green	0.6602	0.6372	0.6856	0.7163	0.7328	**0.7565**
	Red	0.6450	0.6506	0.6745	0.7080	**0.7694**	0.7522
	NIR	0.5189	0.4822	0.5092	0.5732	0.5737	**0.6842**
	Average	0.6222	0.6024	0.6385	0.6790	0.6987	**0.7391**
UIQI	Blue	0.5253	0.5694	0.5671	0.6096	0.6348	**0.6781**
	Green	0.5047	0.5610	0.5533	0.5926	0.6484	**0.6561**
	Red	0.5122	0.5907	0.5642	0.6038	**0.7099**	0.6683
	NIR	0.5005	0.4638	0.4779	0.5317	0.5383	**0.6108**
	Average	0.5107	0.5462	0.5406	0.5844	0.6328	**0.6533**
RMSE	Blue	0.0443	0.0371	0.0406	0.0382	0.0356	**0.0338**
	Green	0.0748	0.0688	0.0719	0.0702	0.0675	**0.0665**
	Red	0.0870	0.0787	0.0830	0.0814	**0.0757**	0.0767
	NIR	0.1117	0.1149	0.1102	0.1051	0.1050	**0.0940**
	Average	0.0795	0.0749	0.0765	0.0737	0.0710	**0.0678**
ERGAS		1.3330	1.2516	1.2816	1.2397	1.1887	**1.1462**
SAM		0.1788	0.2003	0.1886	0.1739	0.1749	**0.1615**

Table 4 Accuracy of different methods in Region 3 (20231003)

		STARFM	UBDF	STDFA	FSDAF	VIPSTF-SW	M^2-STF
CC	Blue	0.8214	0.6893	0.6833	0.7623	0.8109	**0.8840**
	Green	0.8499	0.7147	0.7357	0.7936	0.8220	**0.9104**
	Red	0.8727	0.7324	0.6947	0.7859	0.8516	**0.9225**
	NIR	0.7683	0.5011	0.5797	0.7056	0.7065	**0.8249**
	Average	0.8281	0.6594	0.6734	0.7619	0.7977	**0.8855**
UIQI	Blue	0.8129	0.6865	0.6761	0.7591	0.8035	**0.8648**
	Green	0.8445	0.7138	0.7338	0.7927	0.8196	**0.8994**
	Red	0.8716	0.7160	0.6689	0.7734	0.8483	**0.9204**
	NIR	0.7530	0.4406	0.5321	0.6861	0.6748	**0.8093**
	Average	0.8205	0.6392	0.6527	0.7528	0.7866	**0.8735**
RMSE	Blue	0.0226	0.0287	0.0306	0.0257	0.0231	**0.0194**
	Green	0.0288	0.0404	0.0397	0.0345	0.0314	**0.0238**
	Red	0.0444	0.0725	0.0814	0.0634	0.0493	**0.0353**
	NIR	0.0997	0.1748	0.1503	0.1130	0.1185	**0.0880**
	Average	0.0489	0.0791	0.0755	0.0591	0.0556	**0.0416**
ERGAS		0.6963	1.1133	1.1064	0.8714	0.7827	**0.5853**
SAM		0.1594	0.2350	0.2316	0.2085	0.1763	**0.1353**



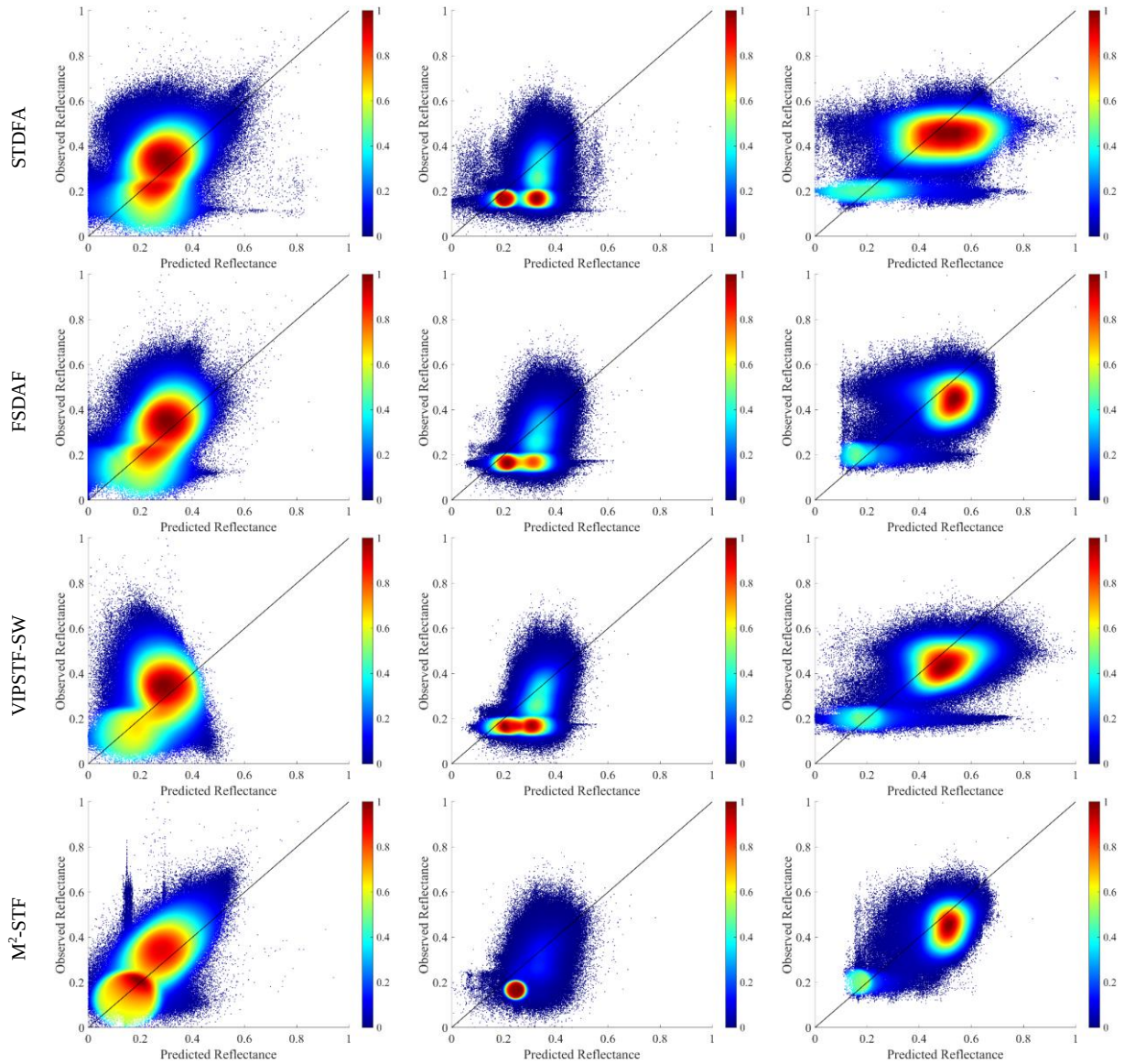


Fig. 6. Scatterplots of the actual and predicted values of the NIR band for Regions 1-3.

Tables 2-4 present the quantitative evaluation results of the different methods for Regions 1-3. Overall, the accuracy of the proposed M^2-STF method is greater than that of the other five methods, with the most significant improvement in the NIR band. Specifically, in Region 1, the average CC of M^2-STF is 0.8599, with increases of 0.0836, 0.3090, 0.1313, 0.0832 and 0.1143 compared to STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW, respectively. Amongst the four bands, the CCs of the five methods in the NIR band range from 0.4127 to 0.6150, which are much smaller than that of M^2-STF (0.8241). Additionally, the average RMSE and ERGAS of M^2-STF are 0.0429 and 0.7756, respectively, with decreases of 0.0075 and 0.1208 compared to the second most accurate method (i.e., FSDAF). For SAM, the value of M^2-STF is 0.1568, which is 0.1042, 0.1149, 0.1867, 0.0865 and 0.0756 smaller than that of STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW, respectively. In the braided river area of Region 2, M^2-STF still produces the greatest accuracy. For example, the mean UIQI and RMSE of M^2-STF are 0.6533 and 0.0678, respectively, which are 0.0205 larger and 0.0032 smaller than the second-ranked method (i.e., VIPSTF-SW). In Region 3, STARFM shows greater competitiveness, but M^2-STF still produces the largest CC and UIQI, and the smallest RMSE, ERGAS and SAM, with gains of 0.0574, 0.0530, 0.0073, 0.1110 and 0.0241 compared to STARFM. To show the difference between the results of different methods and the reference more intuitively, the scatterplots for the three regions are shown in Fig. 6 (taking the NIR band as an example). In the three regions, the points in the STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW scatterplots are obviously more dispersed, while the points in the M^2-STF scatterplot are more clustered and closer to the $y = x$ line.

B. Influence of the inclusion of different water masks at the prediction time

1) Extraction of water mask

We extracted water masks from the Sentinel-1 SAR data and Sentinel-3 optical data (after interpolation) at the prediction time, and the extracted masks were denoted as S1 mask and S3 mask, respectively. Specifically, the S3 mask was extracted using the NDWI method [54]. The extraction results using the S1 and S3

masks for the three regions are shown in Fig. 7. The water mask extracted by the Sentinel-2 data (S2 mask) was used as the reference for evaluation. From visual inspection, the S1 mask is much more similar to the S2 mask for all three regions, indicating that the Sentinel-1 data can produce relatively accurate water mask extraction results. In the S3 mask, due to the coarse spatial resolution of Sentinel-3, the boundary of the extracted water mask is rough, which differs greatly from the S2 mask. Furthermore, we conducted a quantitative evaluation of the accuracy of the S1 and S3 masks, taking Intersection over Union (IoU) and Overall Accuracy (OA) as evaluation indices, as shown in Table 5. Overall, the accuracy of the S1 masks in the three regions is obviously greater than that for the S3 mask. It is worth noting that the strategy of using the S1 mask was proposed in this paper, which has not been previously considered in the existing literature. To ensure a fair comparison with the proposed M^2-STF method, the S3 mask was further introduced into spatio-temporal fusion in Section III-B 2) to investigate the effects of applying the water mask at the prediction time on different spatio-temporal fusion methods.

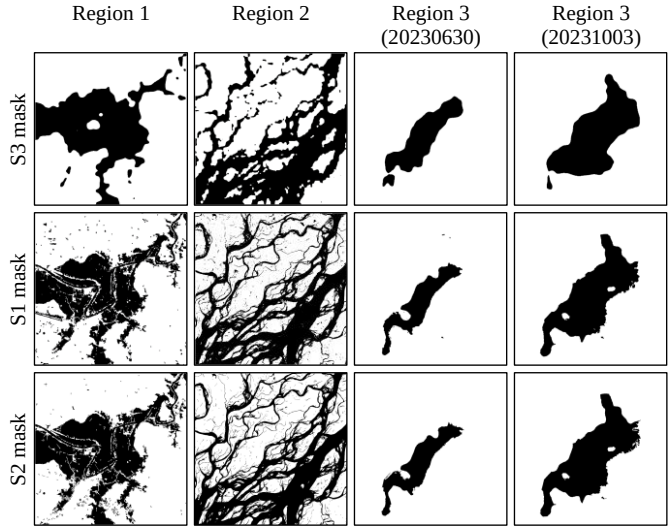


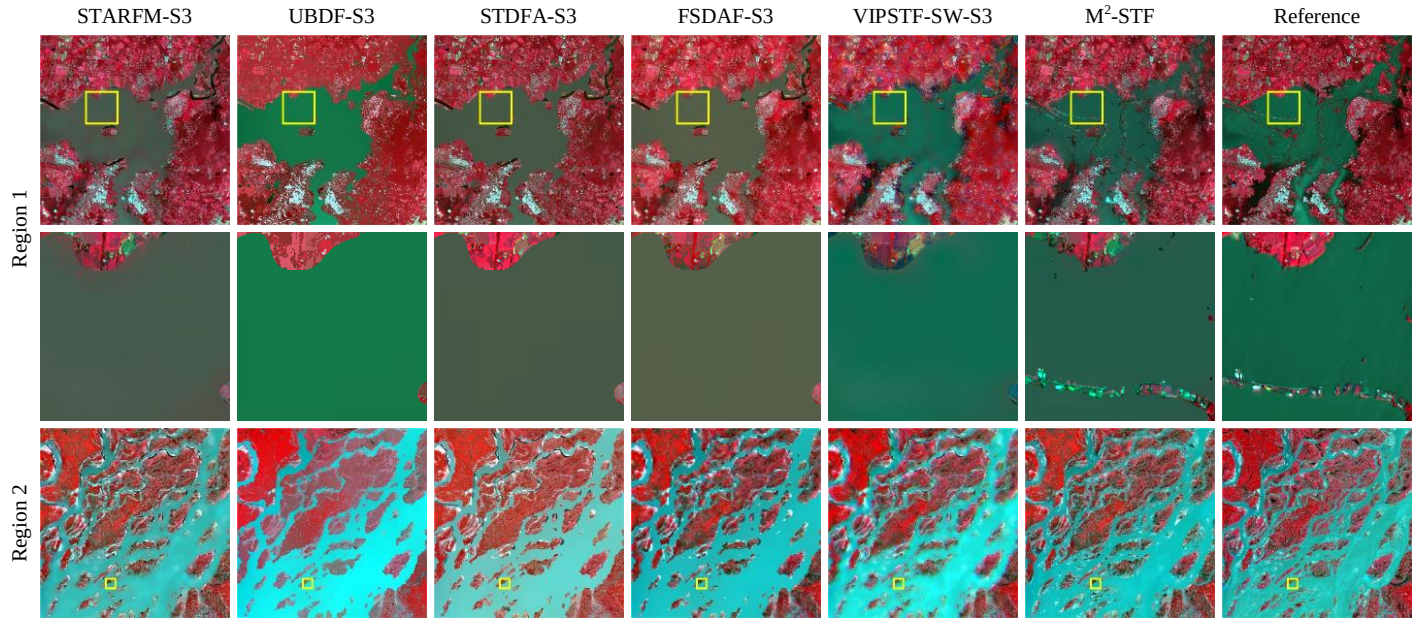
Fig. 7. Water masks extracted from different data for Regions 1-3.

Table 5 IoU and OA of different water masks for Regions 1-3

	IoU		OA	
	S3 mask	S1 mask	S3 mask	S1 mask
Region 1	0.6796	**0.8458**	0.8953	**0.9503**
Region 2	0.6345	**0.8536**	0.8177	**0.9361**
Region 3 (20230630)	0.6923	**0.9241**	0.9681	**0.9939**
Region 3 (20231003)	0.8054	**0.9623**	0.9591	**0.9928**

2) Results of inclusion of different water masks

To explore the applicability of the water mask for the various spatio-temporal fusion methods, we applied the S3 mask to the five methods: STARFM, UBDF, STDFA, FSDAF, and VIPSTF-SW. It is worth noting that these improvements were only applied to the changed area, and the original methods were retained for the unchanged area. Specifically, for the spatial weighted-based methods (STARFM and VIPSTF-SW), based on the fine spatial resolution information in the S3 mask, we transformed the problem of downscaling coarse spatial resolution increments to that of downscaling coarse spatial resolution data at the prediction time. For STDFA and FSDAF, we utilized the fine spatial resolution classification map at t_2, which was constructed by the S3 mask and classification map at t_1, to unmix the coarse spatial resolution data at the prediction time, instead of unmixing their increments in traditional versions. Fig. 8 shows the results of the S3-based version of each method. Compared with the original methods in Fig. 5, the S3 version of each method produces more accurate predictions in the water area. Furthermore, the M^2-STF prediction is still closer to the reference image in terms of water boundary and overall tone in the three regions. For example, in Region 3 (20231003), focusing on the sub-area, it can be found that the S3-based predictions of the water boundary are still biased due to the coarse spatial resolution (although spatially interpolated to 10 m) of the mask boundary, and the M^2-STF result is closer to the boundary in the reference image.



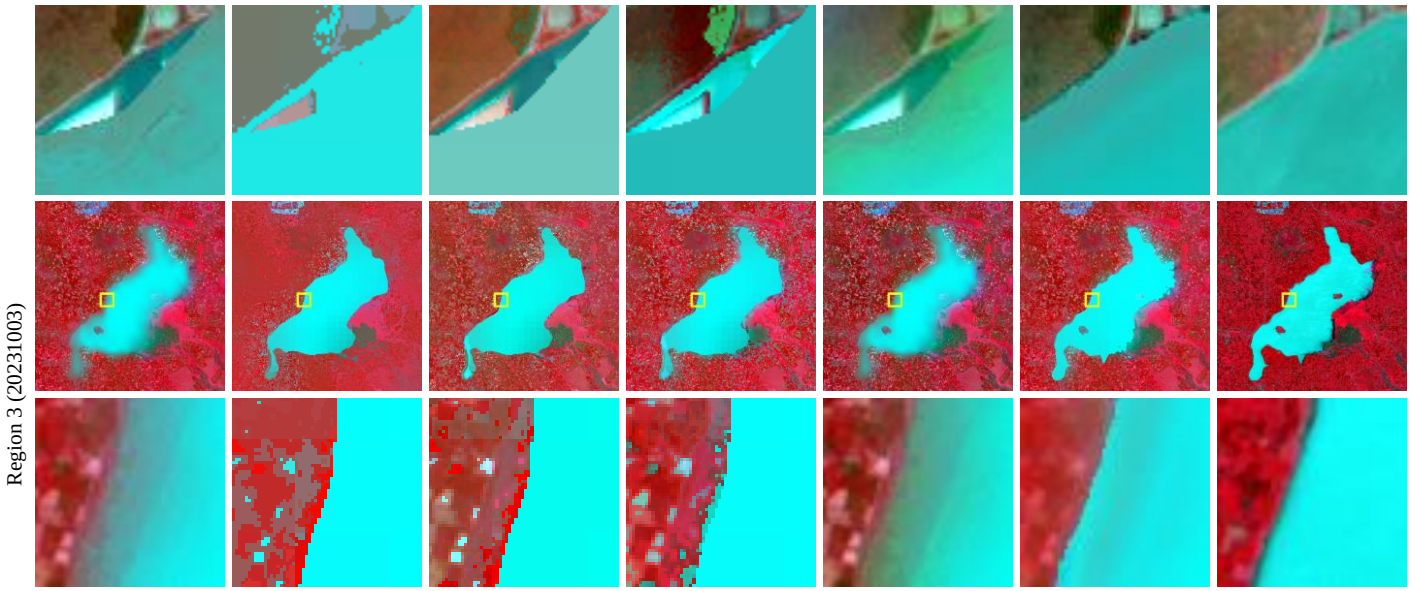


Fig. 8. Predictions for Regions 1-3 based on the S3 mask version of the different methods (NIR, red, and green bands as RGB).

Table 6 Comparison between the original and S3-based versions of the methods for Regions 1-3 (X-S3 indicates that the S3 mask was used in the X method; Bold represents the largest CC or smallest RMSE, while underline represents the next largest CC or next smallest RMSE)

	CC			RMSE		
	Region 1	Region 2	Region 3 (20231003)	Region 1	Region 2	Region 3 (20231003)
STARFM	0.7763	0.6222	0.8281	0.0532	0.0795	0.0489
STARFM-S3	<u>0.8299</u>	0.7061	<u>0.8332</u>	<u>0.0466</u>	0.0691	<u>0.0471</u>
UBDF	0.5509	0.6024	0.6594	0.0601	0.0749	0.0791
UBDF-S3	0.6499	0.7028	0.7738	0.0528	0.0685	0.0574
STDFA	0.7286	0.6385	0.6734	0.0582	0.0765	0.0755
STDFA-S3	0.7957	0.6901	0.7755	0.0514	0.0691	0.0583
FSDAF	0.7767	0.6790	0.7619	0.0504	0.0737	0.0591
FSDAF-S3	0.8159	0.7177	0.8124	0.0475	0.0680	0.0518
VIPSTF-SW	0.7456	0.6987	0.7977	0.0537	0.0710	0.0556
VIPSTF-SW-S3	0.7736	<u>0.7285</u>	0.8208	0.0523	**0.0669**	0.0535
M^2-STF	**0.8599**	**0.7391**	**0.8855**	**0.0429**	<u>0.0678</u>	**0.0416**

Table 6 shows the quantitative evaluation results of each method before and after the introduction of the S3 mask for the three regions (with X-S3 denoting the use of the S3 mask in the X method), using the average CC and RMSE of each band. Overall, in all three regions, the S3 version of all methods show varying degrees of accuracy increase compared to the original methods. For example, in Region 1, the S3 versions of the STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW methods increase the CCs by 0.0536, 0.0990, 0.0671, 0.0392 and 0.0280, respectively, while decreasing the RMSEs by 0.0066, 0.0073, 0.0068, 0.0029 and 0.0014, respectively. It is noted that STARFM-S3 and STDFA-S3 present the most obvious accuracy increases. In Region 2, the CCs for the S3 version of the five methods range from 0.6901 to 0.7285 and the RMSEs range from 0.0669 to 0.0691. Generally, the accuracy increased by FSDAF-S3 and VIPSTF-SW-S3 is relatively small. The most substantial increase is observed for UBDF-S3, which increases the CC by 0.1004 and reduced the RMSE by 0.0064 compared to its original version. In Region 3 (20231003), the S3 versions of the STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW methods increase the CCs by 0.0051, 0.1144, 0.1021, 0.0505 and 0.0231 compared to the original methods, respectively. In all three regions, M^2-STF still produces the most accurate results compared to the S3 versions of the five spatio-temporal fusion methods. For example, in Regions 1 and 3, compared to the second most accurate method, STARFM-S3, the CCs of M^2-STF are 0.0300 and 0.0523 larger, and the RMSEs are 0.0037 and 0.0055 smaller.

C. Influence of the range of water changes

To explore the advantages of M^2-STF over other methods under different extents of water changes, we selected data at two different prediction times in Region 3 for validation. Specifically, using the image on 20230521 as auxiliary data, the fine spatial resolution images on dates of 20230630 and 20231003 were predicted. As shown in Fig. 3, during this period, the water area continues to expand, and the expanding area from 20230521 to 20230630 is smaller than that from 20230521 to 20231003. That is, the longer the time interval, the larger the area of the surface water changes. Fig. 9 shows the average CC and RMSE gains in the water area of M^2-STF relative to the results of the five methods. Specifically, the CC of the M^2-STF result on the date of 20230630 is 0.1200, 0.1784, 0.1651, 0.1204 and 0.1276 larger than for STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW, respectively. On the date of 20231003, the corresponding CC gains are 0.1587, 0.2770, 0.2486, 0.1978 and 0.1770. That is, the gains on 20231003 are larger than those for 20230630, indicating that the M^2-STF approach demonstrates greater advantages over other methods when the water area undergoes more significant changes.

Additionally, we calculated the gains in the water area results of the S3 versions over the corresponding original methods, so as to further investigate the effect of using a water mask under different water area changes. As shown in Fig. 10, the red bars represent the gains of the S3 versions relative to the original methods on the date 20230630, while the blue bars represent the gains on the date 20231003. Checking the results, it is found that the blue bar is higher than the red bar for each method, indicating that the greater the change of water area, the more obvious the effect of including the water mask. For example, when using the S3 mask, the UBDF method presents the maximum average CC gain of 0.0876 on the date 20231003, which is also larger than

the gain of 0.0310 on 20230630. Similarly, the UBDF method on 20231003 produces the largest average RMSE gain of 0.0295, which is larger than the gain of 0.0138 on 20230630.

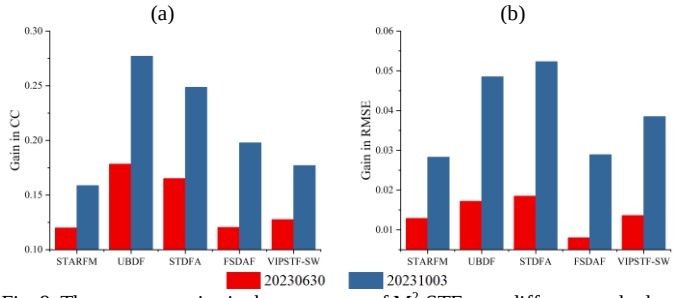


Fig. 9. The accuracy gains in the water area of M^2-STF over different methods on the two dates for Region 3. (a) Gain in CC. (b) Gain in RMSE.

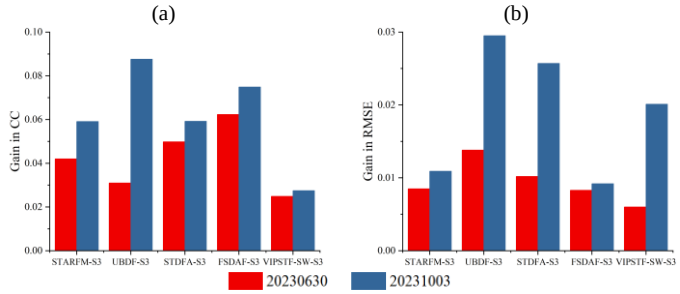


Fig. 10. The accuracy gains in the water area of different methods (over the original version without water mask) on the two dates for Region 3. (a) Gain in CC. (b) Gain in RMSE.

D. *Ablation analysis of the different modules in M^2-STF*

To investigate the influence of the proposed similar pixel selection and homogeneity index I_{ho} on the final results of M^2-STF, we compared four strategies: similar pixels in the changed area are not excluded and the homogeneity index is not introduced (denoted as NS-NI); similar pixels in the changed area are not excluded but the homogeneity index is introduced (denoted as NS-I); similar pixels in the changed area are excluded but the homogeneity index is not introduced (denoted as S-NI), and similar pixels in the changed area are excluded and the homogeneity index is introduced (denoted as S-I, that is, the strategy used in M^2-STF). We evaluated the accuracy of spatio-temporal fusion obtained using these four strategies, as shown in Fig. 11. Generally, the accuracy of NS-I is greater than that of NS-NI, and the accuracy of M^2-STF is greater than that of S-NI, indicating that the introduction of the homogeneity index I_{ho} can effectively increase the accuracy. Additionally, it can be found that the accuracy of S-NI is generally greater than that of NS-NI, and the accuracy of M^2-STF is generally greater than that of NS-I, indicating that the accuracy of M^2-STF can also be enhanced by excluding similar pixels in the changed area. Generally, by combining both the strategies of removing similar pixels from the changed area and using the homogeneity index I_{ho}, the greatest prediction accuracy can be obtained for all three regions.

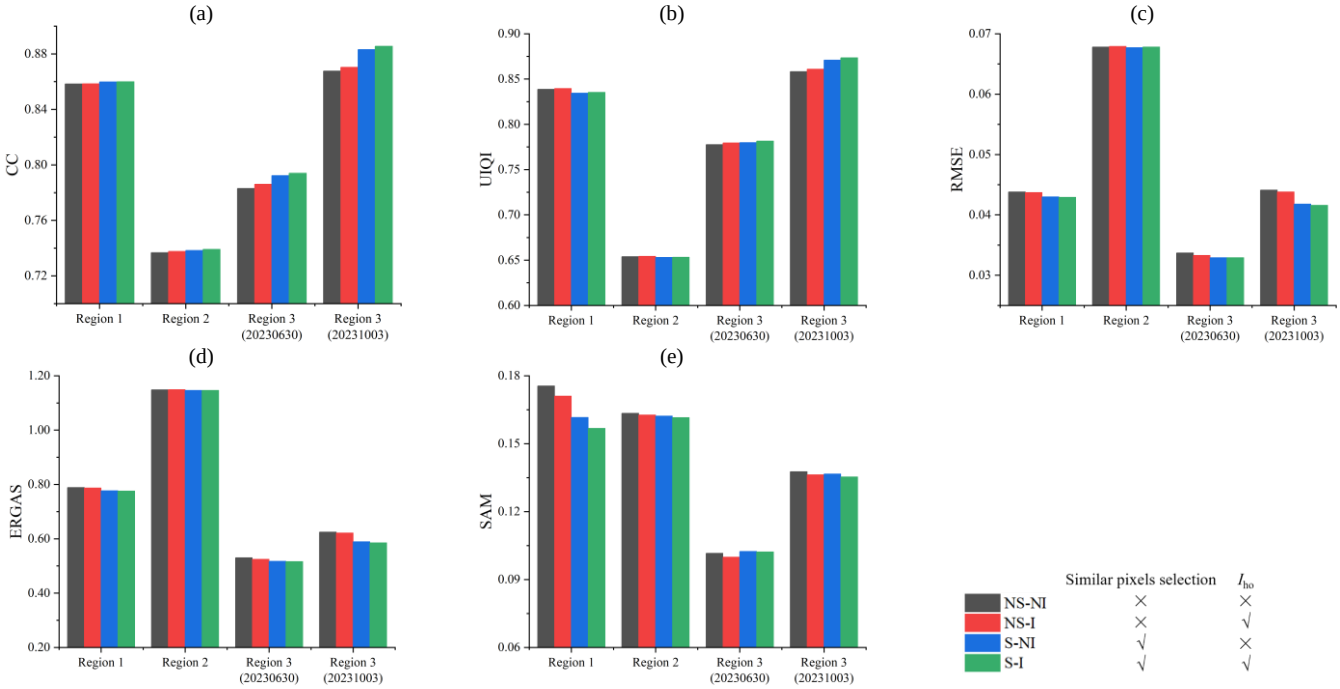


Fig. 11. Accuracy of M^2-STF with different strategies (considering whether to select similar pixels only from the unchanged area and whether to add I_{ho}). (a) CC. (b) UIQI. (c) RMSE. (d) ERGAS. (e) SAM.

IV. Discussion

A. *Uncertainty in 10 m water extraction*

In Section II-B 1), we employed the SDWI-based threshold segmentation method to extract water masks. Although thresholding may introduce additional uncertainties, its impact in this study is limited, as the water masks derived from Sentinel-1 are highly consistent with those from Sentinel-2. To quantify the potential uncertainty, we compared the fusion results generated by M^2-STF (using S1 mask) with M^2-STF-S2

(using S2 mask). As shown in Table 7, the CC of the fusion results obtained with M^2-STF-S2 is 0.0016 to 0.0185 larger than that of M^2-STF, while the RMSE decease ranges from 0.0001 to 0.0050. To further assess the thresholding approach, we also compared it with the Sentinel-1 Flood Finder (called S1FF hereafter), which utilizes a pixel-based random forests approach to detect water pixels within a scene. The corresponding water masks are illustrated in Fig. 12. Visually, the SDWI-based masks appear more accurate, as S1FF is prone to speckle noise and classification errors in heterogeneous land cover. As summarized in Table 7, M^2-STF and M^2-STF-S1FF yield very similar accuracies, whereas the threshold-based approach is computationally simpler. Therefore, given its comparable performance and reduced complexity, the threshold segmentation method was considered in this study.

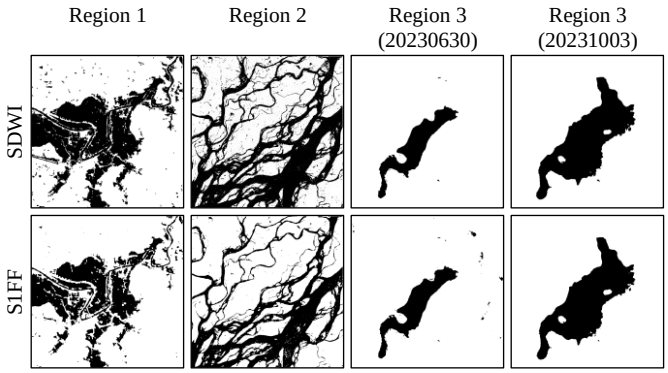


Fig. 12. Water masks extracted from different methods for Regions 1-3.

Table 7 Comparison between different mask versions of M^2-STF for Regions 1-3

	CC				RMSE			
	Region 1	Region 2	Region 3 (20230630)	Region 3 (20231003)	Region 1	Region 2	Region 3 (20230630)	Region 3 (20231003)
M^2-STF	0.8599	0.7391	0.8029	0.8855	0.0429	0.0678	0.0322	0.0416
M^2-STF-S1FF	0.8523	0.7301	0.7957	0.8817	0.0438	0.0701	0.0327	0.0420
M^2-STF-S2	0.8719	0.7576	0.8045	0.8928	0.0417	0.0628	0.0321	0.0405

B. Influence of water mask

In Section III-B 2), we incorporated the water mask extracted from Sentinel-3 into two spatial weighting-based methods (STARFM and VIPSTF-SW), two spatial unmixing-based methods (UBDF and STDFA), and a hybrid method (FSDAF). The experimental results demonstrate that the introduction of the S3 mask can effectively increase the accuracy under drastic changes. In Section III-B 1), the results show that the S1 mask provides more accurate results than the S3 mask. Thus, we replaced the coarse S3 water mask with the fine S1 mask, and introduced it into the five spatio-temporal fusion methods to explore whether the fusion accuracy can be further increased. The results of the S1 mask versions of all five methods are shown in Fig. 13 (Region 3 (20231003) is used as an example). Generally, compared with the S3 versions in Section III-B 2), the S1 versions exhibit greater visual performance, especially in the prediction of the water boundary, which is closer to the reference data. Table 8 provides a detailed comparison of the accuracy of the S1 versions across each band. Similarly, the accuracy of the S1 version is greater than for the S3 version of each method (see Table 4 for details). Furthermore, when comparing the S1 version of each method with the M^2-STF approach, it is found that M^2-STF presents the greatest accuracy. Additionally, it should be emphasized that this paper adopts the classical threshold method to extract the water mask, as it does not focus on the accuracy of water mask. In future research, efforts can be made to explore more accurate water extraction methods to further enhance M^2-STF.

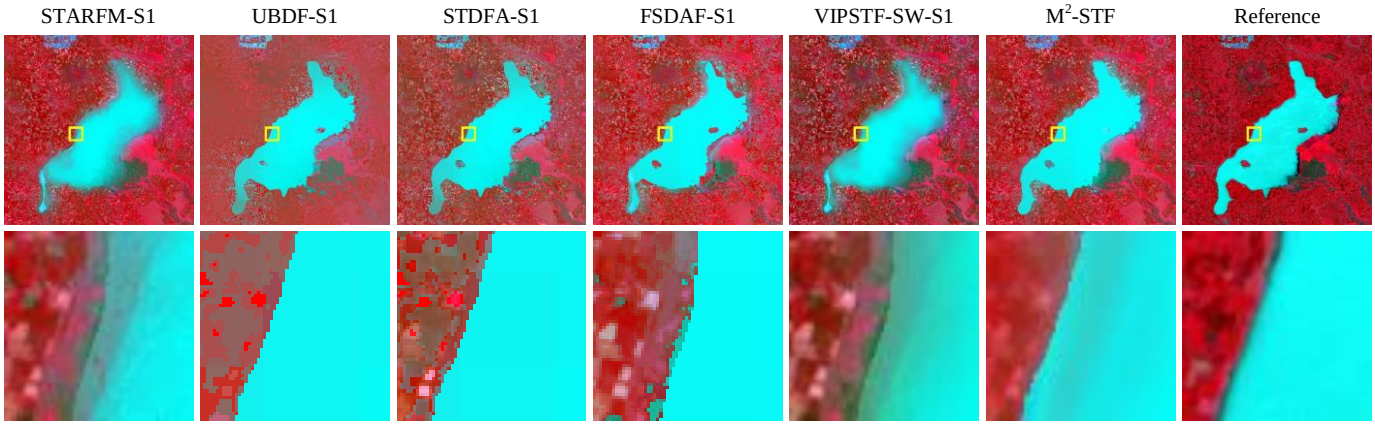


Fig. 13. Predictions for S1 version of different methods (NIR, red, and green bands as RGB) (with Region 3 (20231003) as an example).

Table 8 Accuracy for S1 version of different methods (with Region 3 (20231003) as an example)

		STARFM-S1	UBDF-S1	STDFA-S1	FSDAF-S1	VIPSTF-SW-S1	M^2-STF
CC	Blue	0.8365	0.7950	0.7850	0.8306	0.8257	**0.8840**
	Green	0.8690	0.8323	0.8513	0.8665	0.8466	**0.9104**
	Red	0.8827	0.8750	0.8293	0.8679	0.8676	**0.9225**
	NIR	0.7864	0.7318	0.7536	0.7632	0.7474	**0.8249**

	Average	0.8437	0.8085	0.8048	0.8320	0.8218	**0.8855**
UIQI	Blue	0.8251	0.7906	0.7814	0.8239	0.8164	**0.8648**
	Green	0.8600	0.8301	0.8496	0.8629	0.8418	**0.8994**
	Red	0.8820	0.8676	0.8149	0.8624	0.8652	**0.9204**
	NIR	0.7717	0.7013	0.7206	0.7445	0.7192	**0.8093**
	Average	0.8347	0.7974	0.7916	0.8234	0.8107	**0.8735**
RMSE	Blue	0.0219	0.0238	0.0252	0.0224	0.0224	**0.0194**
	Green	0.0271	0.0307	0.0291	0.0274	0.0291	**0.0238**
	Red	0.0424	0.0471	0.0576	0.0476	0.0461	**0.0353**
	NIR	0.0955	0.1142	0.1102	0.1025	0.1085	**0.0880**
	Average	0.0467	0.0539	0.0555	0.0500	0.0515	**0.0416**
ERGAS		0.6654	0.7600	0.8091	0.7149	0.7279	**0.5853**
SAM		0.1525	0.1665	0.1648	0.1726	0.1599	**0.1353**

C. The performance of M^2-STF in different regions

In this paper, three datasets with different scenarios were used for validation. It can be noted that the performance of M^2-STF varies across the different regions. Specifically, the prediction accuracy of M^2-STF in Region 3 (20231003) is greater than that of Regions 1 and 2. This can be attributed to the spatial pattern of land cover in each region. Compared with Regions 1 and 2, Region 3 (20231003) covers relatively homogeneous land cover types before and after water expansion. In contrast, Region 1 covers diverse land cover types before the flood, including cities, rivers and farmland, presenting strong heterogeneity. Similarly, the heterogeneity of the unchanged area in Region 2 is also strong, and the braided drainage patterns in the changed area are complicated. From the perspective of water extraction, unlike the large-scale water expansion in Regions 1 and 3 (20231003), the extraction of the slender water planform in the braided water system in Region 2 tends to be intermittent. Additionally, Sentinel-1 data are affected by various noise, which further contributes to the lowest accuracy of the results in Region 2 amongst the three regions. Judging from the results of M^2-STF in the three regions, the restoration of spectral differences within each class is still challenging. Although the issue of intra-class spectral variation has been considered in some regions, it has not been completely solved [55]. This indicates that further optimization of the model is still required in future to better deal with scenarios with complex intra-class spectral variation.

D. The applicability of M^2-STF

In this paper, Sentinel-2 MSI and Sentinel-3 OLCI were selected as input data for the M^2-STF approach, and Sentinel-1 data were used to provide auxiliary fine spatial resolution information at the prediction time. The main reason of using Sentinel-2 and -3 for the fusion example in this paper is that the Sentinel data (including Sentinel-1, -2 and -3) are produced by the Copernicus program, ensuring harmonized calibration between the series. Generally, M^2-STF provides a general model for enhancing spatio-temporal fusion of remote sensing data in scenarios experiencing drastic land surface changes, and has the potential to be extended to other remote sensing data fusion tasks. For example, in the case of Landsat and MODIS data commonly used for spatio-temporal fusion, Sentinel-1 (or other SAR images) acquired closer to the prediction time can also be incorporated to enhance prediction accuracy. Additionally, this paper discusses the application of M^2-STF in the scenario of water expansion. In the changed area, M^2-STF introduces the water mask extracted from Sentinel-1 at the prediction time to enhance the prediction accuracy. This strategy is simple and easy to implement, especially since the availability of free Sentinel data lowers the barrier to entry for research. Once we acquire coarse resolution data and SAR data for the prediction time, we can obtain fine resolution optical data. Governments and disaster management agencies will benefit from this, as high-frequency, multi-source data fusion can provide more timely information on water body expansion, helping governments to quickly develop evacuation or disaster prevention strategies.

E. The limitations of M^2-STF

In this study, the proposed method requires Sentinel-1 data to be temporally very close to the prediction date (ideally on the same day). It is acknowledged that in regions where Sentinel-1 data have relatively coarse temporal resolution, the temporal gaps with Sentinel-2 observations may affect model performance. Moreover, SAR data are prone to speckle noise, and threshold-based segmentation may introduce additional uncertainties. It should be noted that in this work, water masks serve only as auxiliary inputs to facilitate spatio-temporal fusion, which is the core of M^2-STF. Although the experimental results demonstrate that the thresholding approach is sufficient for water extraction in this study, more advanced techniques could be developed in future work to enhance robustness. Furthermore, it would be interesting to explore the direct integration of SAR data into spatio-temporal fusion. For example, Ye *et al*. [56] proposed a novel optical-SAR fusion framework, termed visual saliency features fusion, which operates by extracting and balancing significant complementary features from the two modalities. Additionally, the multi-CNN sequence-to-sequence model [57] was employed to formulate correlations between optical and SAR time series for crop dynamic monitoring. However, direct multi-modal fusion still faces challenges related to modality differences, speckle noise, and geometric distortions. Future work will aim at building on these advances by incorporating deep learning to achieve more generalized and robust fusion strategies.

F. Multi-modal data fusion

This paper integrates SAR data to extract the changed area. In addition to Sentinel-1 data, other SAR datasets such as Radarsat-2, ALOS-2 and TerraSAR-X also have potential for similar applications. Furthermore, in addition to combining SAR data with optical data for spatio-temporal fusion tasks, in future studies, we can also consider other multi-modal data for fusion, such as LiDAR data. LiDAR can provide 3-D spatial geometry

and compensate for optical data for more accurate land cover and land use (LCLU) classification. Additionally, the fusion of remote sensing data and geospatial 'big data' (e.g., point of interest (POI) or street view imagery) can also be considered. In a previous study, Mantsis *et al.* [58] used snow-related instant social media data, together with Sentinel-1 data to achieve snow depth estimation. Lu *et al.* [59] extracted information from remote sensing imagery and POI data for urban functional zone mapping. Wang [60] fused multi-source social media data with street view images to characterize the quality of urban space. Moreover, there is also potential for integrating remote sensing data with ground-based observation and dynamic models for fusion. This practice could provide extra information, such as spatial distribution and temporal change rules.

V. Conclusion

In spatio-temporal fusion tasks, the greater the land surface changes between the known and prediction times, the greater the uncertainty of the predictions. Focusing on the water expansion scenario due to flood inundation, the M^2-STF method was proposed to address the challenging problem of drastic land surface changes. Specifically, in spatio-temporal fusion of Sentinel-2 MSI and Sentinel-3 OLCI, Sentinel-1 SAR data were incorporated to provide valuable fine spatial resolution land cover distribution information at the prediction time. We conducted experiments on three datasets and compared M^2-STF with five typical spatio-temporal fusion methods, including STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW. The main conclusions are as follows:

1) Compared with STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW, M^2-STF can estimate the boundary of changed land surface more accurately and produce greater accuracy. For example, in Region 1, the average CC of M^2-STF is 0.8599, which is 0.0836, 0.3090, 0.1313, 0.0832, and 0.1143 larger than for STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW, respectively. The average RMSE of M^2-STF is 0.0429, which is 0.0103, 0.0172, 0.0153, 0.0075 and 0.0108 smaller than those of STARFM, UBDF, STDFA, FSDAF and VIPSTF-SW, respectively.
2) The inclusion of a fine spatial resolution water mask can effectively increase the accuracy of changed boundary prediction. The increase in accuracy is more obvious when the extent of the water changes from the known to the prediction times is larger.
3) Compared with the water mask (S3 mask) extracted from Sentinel-3 OLCI data, the water mask (S1 mask) extracted from Sentinel-1 SAR data is more accurate. For each spatio-temporal fusion method, the inclusion of the S1 mask results in greater accuracy.

References

[1] B. Chen, B. Huang, and B. Xu, "Comparison of Spatiotemporal Fusion Models: A Review," *Remote Sensing*, vol. 7, no. 2, pp. 1798–1835, Feb. 2015.

[2] H. K. Zhang, B. Huang, M. Zhang, K. Cao, and L. Yu, "A generalization of spatial and temporal fusion methods for remotely sensed surface parameters," *International Journal of Remote Sensing*, vol. 36, no. 17, pp. 4411–4445, Sep. 2015.

[3] X. Zhu, F. Cai, J. Tian, and T. Williams, "Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions," *Remote Sensing*, vol. 10, no. 4, p. 527, Mar. 2018.

[4] M. Belgiu and A. Stein, "Spatiotemporal Image Fusion in Remote Sensing," *Remote Sensing*, vol. 11, no. 7, p. 818, Apr. 2019.

[5] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data: an overview and new benchmark," *Sci. China Inf. Sci.*, vol. 63, no. 4, p. 140301, Apr. 2020.

[6] Q. Wang, Y. Tang, Y. Ge, H. Xie, X. Tong, and P. M. Atkinson, "A comprehensive review of spatial-temporal-spectral information reconstruction techniques," *Science of Remote Sensing*, vol. 8, p. 100102, Dec. 2023.

[7] 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, p. 113083, Sep. 2022.

[8] 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, p. 113784, Nov. 2023.

[9] Q. Wang, Y. Tang, X. Tong, and P. M. Atkinson, "Filling gaps in cloudy Landsat LST product by spatial-temporal fusion of multi-scale data," *Remote Sensing of Environment*, vol. 306, p. 114142, May 2024.

[10] H. Yang, Q. Wang, W. Zhao, X. Tong, and P. M. Atkinson, "Reconstruction of a Global 9 km, 8-Day SMAP Surface Soil Moisture Dataset during 2015–2020 by Spatiotemporal Fusion," *J Remote Sens*, vol. 2022, p. 2022/9871246, Jan. 2022.

[11] S. Wang *et al.*, "A classification-based spatiotemporal adaptive fusion model for the evaluation of remotely sensed evapotranspiration in heterogeneous irrigated agricultural area," *Remote Sensing of Environment*, vol. 273, p. 112962, May 2022.

[12] M. Zhang *et al.*, "A spatio-temporal fusion strategy for improving the estimation accuracy of the aboveground biomass in grassland based on GF-1 and MODIS," *Ecological Indicators*, vol. 157, p. 111276, Dec. 2023.

[13] Y. Gao *et al.*, "FARM: A fully automated rice mapping framework combining Sentinel-1 SAR and Sentinel-2 multi-temporal imagery," *Computers and Electronics in Agriculture*, vol. 213, p. 108262, Oct. 2023.

[14] Feng Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance," *IEEE Trans. Geosci. Remote Sensing*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006.

[15] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sensing of Environment*, vol. 114, no. 11, pp. 2610–2623, Nov. 2010.

[16] Q. Wang and P. M. Atkinson, "Spatio-temporal fusion for daily Sentinel-2 images," *Remote Sensing of Environment*, vol. 204, pp. 31–42, Jan. 2018.

[17] Q. Wang, Y. Tang, X. Tong, and P. M. Atkinson, "Virtual image pair-based spatio-temporal fusion," *Remote Sensing of Environment*, vol. 249, p. 112009, Nov. 2020.

[18] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhackel, "Unmixing-based multisensor multiresolution image fusion," *IEEE Trans. Geosci. Remote Sensing*, vol. 37, no. 3, pp. 1212–1226, May 1999.

[19] R. Zurita-Milla, J. Clevers, and M. E. Schaepman, "Unmixing-Based Landsat TM and MERIS FR Data Fusion," *IEEE Geosci. Remote Sensing Lett.*, vol. 5, no. 3, pp. 453–457, Jul. 2008.

[20] M. Wu, Z. Niu, C. Wang, C. Wu, and L. Wang, "Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model," *J. Appl. Remote Sens*, vol. 6, no. 1, p. 063507, Mar. 2012.

[21] W. Liu, Y. Zeng, S. Li, and W. Huang, "Spectral unmixing based spatiotemporal downscaling fusion approach," *International Journal of Applied Earth Observation and Geoinformation*, vol. 88, p. 102054, Jun. 2020.

[22] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sensing of Environment*, vol. 172, pp. 165–177, Jan. 2016.

[23] M. Liu *et al.*, "An Improved Flexible Spatiotemporal DAta Fusion (IFSDAF) method for producing high spatiotemporal resolution normalized difference vegetation index time series," *Remote Sensing of Environment*, vol. 227, pp. 74–89, Jun. 2019.

[24] X. Li *et al.*, "SFSDAF: An enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion," *Remote Sensing of Environment*, vol. 237, p. 111537, Feb. 2020.

[25] D. Guo, W. Shi, M. Hao, and X. Zhu, "FSDAF 2.0: Improving the performance of retrieving land cover changes and preserving spatial details," *Remote Sensing of Environment*, vol. 248, p. 111973, Oct. 2020.

[26] H. Gao *et al.*, "cuFSDAF: An Enhanced Flexible Spatiotemporal Data Fusion Algorithm Parallelized Using Graphics Processing Units," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–16, 2022.

[27] D. Guo, W. Shi, F. Qian, S. Wang, and C. Cai, "Monitoring the spatiotemporal change of Dongting Lake wetland by integrating Landsat and MODIS images, from 2001 to 2020," *Ecological Informatics*, vol. 72, p. 101848, Dec. 2022.

[28] M. Lu, J. Chen, H. Tang, Y. Rao, P. Yang, and W. Wu, "Land cover change detection by integrating object-based data blending model of Landsat and MODIS," *Remote Sensing of Environment*, vol. 184, pp. 374–386, Oct. 2016.

[29] H. Guan, Y. Su, T. Hu, J. Chen, and Q. Guo, "An Object-Based Strategy for Improving the Accuracy of Spatiotemporal Satellite Imagery Fusion for Vegetation-Mapping Applications," *Remote Sensing*, vol. 11, no. 24, p. 2927, Dec. 2019.

[30] H. Zhang, Y. Sun, W. Shi, D. Guo, and N. Zheng, "An object-based spatiotemporal fusion model for remote sensing images," *European Journal of Remote Sensing*, vol. 54, no. 1, pp. 86–101, Jan. 2021.

[31] D. Guo and W. Shi, "Object-Level Hybrid Spatiotemporal Fusion: Reaching a Better Tradeoff Among Spectral Accuracy, Spatial Accuracy, and Efficiency," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 16, pp. 8007–8021, 2023.

[32] H. Guo, D. Ye, H. Xu, and L. Bruzzone, "OBSUM: An object-based spatial unmixing model for spatiotemporal fusion of remote sensing images," *Remote Sensing of Environment*, vol. 304, p. 114046, Apr. 2024.

[33] K. Nagarajan, C. Krekeler, K. C. Slatton, and W. D. Graham, "A Scalable Approach to Fusing Spatiotemporal Data to Estimate Streamflow via a Bayesian Network," *IEEE Trans. Geosci. Remote Sensing*, vol. 48, no. 10, pp. 3720–3732, Oct. 2010.

[34] L. Liao, J. Song, J. Wang, Z. Xiao, and J. Wang, "Bayesian Method for Building Frequent Landsat-Like NDVI Datasets by Integrating MODIS and Landsat NDVI," *Remote Sensing*, vol. 8, no. 6, p. 452, May 2016.

[35] J. Xue, Y. Leung, and T. Fung, "A Bayesian Data Fusion Approach to Spatio-Temporal Fusion of Remotely Sensed Images," *Remote Sensing*, vol. 9, no. 12, p. 1310, Dec. 2017.

[36] Y. Zhang *et al.*, "Updating Landsat-based forest cover maps with MODIS images using multiscale spectral-spatial-temporal superresolution mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 63, pp. 129–142, Dec. 2017.

[37] B. Huang and H. Song, "Spatiotemporal Reflectance Fusion via Sparse Representation," *IEEE Trans. Geosci. Remote Sensing*, vol. 50, no. 10, pp. 3707–3716, Oct. 2012.

[38] X. Zhang, S. Li, Z. Tan, and X. Li, "Enhanced wavelet based spatiotemporal fusion networks using cross-paired remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 211, pp. 281–297, May 2024.

[39] Z. Tan, M. Gao, X. Li, and L. Jiang, "A Flexible Reference-Insensitive Spatiotemporal Fusion Model for Remote Sensing Images Using Conditional Generative Adversarial Network," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–13, 2022.

[40] B. Song *et al.*, "MLFF-GAN: A Multilevel Feature Fusion With GAN for Spatiotemporal Remote Sensing Images," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–16, 2022.

[41] S. Wang and F. Fan, "STINet: Vegetation Changes Reconstruction Through a Transformer-Based Spatiotemporal Fusion Approach in Remote Sensing," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–16, 2024.

[42] G. Yang *et al.*, "MSFusion: Multistage for Remote Sensing Image Spatiotemporal Fusion Based on Texture Transformer and Convolutional Neural Network," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 15, pp. 4653–4666, 2022.

[43] G. Chen, P. Jiao, Q. Hu, L. Xiao, and Z. Ye, "SwinSTFM: Remote Sensing Spatiotemporal Fusion Using Swin Transformer," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–18, 2022.

[44] C. Xu, X. Du, Z. Yan, J. Zhu, S. Xu, and X. Fan, "VSDF: A variation-based spatiotemporal data fusion method," *Remote Sensing of Environment*, vol. 283, p. 113309, Dec. 2022.

[45] Z. Gu, J. Chen, Y. Chen, Y. Qiu, X. Zhu, and X. Chen, "Agri-Fuse: A novel spatiotemporal fusion method designed for agricultural scenarios with diverse phenological changes," *Remote Sensing of Environment*, vol. 299, p. 113874, Dec. 2023.

[46] F. Xu, S. Heremans, and B. Somers, "Urban land cover mapping with Sentinel-2: a spectro-spatio-temporal analysis," *Urban Info*, vol. 1, no. 1, p. 8, Oct. 2022.

[47] Y. Li, Z. Niu, Z. Xu, and X. Yan, "Construction of High Spatial-Temporal Water Body Dataset in China Based on Sentinel-1 Archives and GEE," *Remote Sensing*, vol. 12, no. 15, p. 2413, Jul. 2020.

[48] Z. Dong, G. Wang, S. O. Y. Amankwah, X. Wei, Y. Hu, and A. Feng, "Monitoring the summer flooding in the Poyang Lake area of China in 2020 based on Sentinel-1 data and multiple convolutional neural networks," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102400, Oct. 2021.

[49] D. Marzi and P. Gamba, "Inland Water Body Mapping Using Multitemporal Sentinel-1 SAR Data," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 14, pp. 11789–11799, 2021.

[50] M. Tazmul Islam and Q. Meng, "An exploratory study of Sentinel-1 SAR for rapid urban flood mapping on Google Earth Engine," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 103002, Sep. 2022.

[51] Q. Wang, W. Shi, P. M. Atkinson, and Y. Zhao, "Downscaling MODIS images with area-to-point regression kriging," *Remote Sensing of Environment*, vol. 166, pp. 191–204, Sep. 2015.

[52] G. T. Ayele *et al.*, "Sediment Yield and Reservoir Sedimentation in Highly Dynamic Watersheds: The Case of Koga Reservoir, Ethiopia," *Water*, vol. 13, no. 23, p. 3374, Nov. 2021.

[53] S. Jia, D. Xue, C. Li, J. Zheng, and W. Li, "Study on new method for water area information extraction based on Sentinel-1 data," Yangtze River, vol. 50, no. 2, pp. 213–217, 2019.

[54] S. K. McFeeters, "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432, May 1996.

[55] Q. Wang, K. Peng, Y. Tang, X. Tong, and P. M. Atkinson, "Blocks-removed spatial unmixing for downscaling MODIS images," *Remote Sensing of Environment*, vol. 256, p. 112325, Apr. 2021.

[56] Y. Ye, J. Zhang, L. Zhou, J. Li, X. Ren, and J. Fan, "Optical and SAR Image Fusion Based on Complementary Feature Decomposition and Visual Saliency Features," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–15, 2024.

[57] W. Zhao, Y. Qu, J. Chen, and Z. Yuan, "Deeply synergistic optical and SAR time series for crop dynamic monitoring," *Remote Sensing of Environment*, vol. 247, p. 111952, Sep. 2020.

[58] D. F. Mantsis *et al.*, "Multimodal Fusion of Sentinel 1 Images and Social Media Data for Snow Depth Estimation," *IEEE Geosci. Remote Sensing Lett.*, vol. 19, pp. 1–5, 2022.

[59] W. Lu, C. Tao, H. Li, J. Qi, and Y. Li, "A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data," *Remote Sensing of Environment*, vol. 270, p. 112830, Mar. 2022.

[60] Y. Wang, "Fusing multi-source social media data and street view imagery to inform urban space quality: a study of user perceptions at Kampong Glam and Haji Lane," *Urban Info*, vol. 3, no. 1, p. 21, Jun. 2024.