A new cross-fusion method to automatically determine the optimal input image pairs for NDVI spatiotemporal data fusion

Yang Chen, Ruyin Cao, Jin Chen, Xiaolin Zhu, Ji Zhou, Guangpeng Wang, Miaogen Shen,

Xuehong Chen, Wei Yang

Abstract—**Spatiotemporal data fusion is a methodology to generate images with both high spatial and temporal resolution. Most spatiotemporal data fusion methods generate the fused image at a prediction date based on pairs of input images from other dates. The performance of spatiotemporal data fusion is greatly affected by the selection of the input image pair. There are two criteria for selecting the input image pair: the "similarity" criterion, in which the image at the base date should be as similar as possible to that at the prediction date, and the "consistency" criterion, in which the coarse and fine images at the base date should be consistent in terms of their radiometric characteristics and imaging geometry. Unfortunately, the "consistency" criterion has not been quantitatively considered by previous selection strategies. We thus develop a novel method (called "cross-fusion") to address the issue of the determination of the base image pair. The new method first chooses several candidate input image pairs according to the "similarity" criterion, and then takes the "consistency" criterion into account by employing all of the candidate input image pairs to implement spatiotemporal data fusion between them. We applied the new method to MODIS-Landsat NDVI data fusion. The results show that the cross-fusion method performs better than four other selection strategies, with lower average absolute difference values and higher correlation coefficients in various vegetated regions including a deciduous forest in Northeast China, an evergreen forest in South China, a cropland in North China Plain and a grassland in the Tibetan Plateau. We simulated scenarios for the inconsistence between MODIS and Landsat data and found that the simulated inconsistence is successfully quantified by the new method. In addition, the cross-fusion method is less affected by cloud omission errors. The fused NDVI time-series data generated by the new method tracked various vegetation growth trajectories better than previous selection strategies. We expect that the cross-fusion method can advance practical applications of spatiotemporal data fusion technology.**

Index Terms—**Landsat NDVI, MODIS-Landsat, NDVI time series, Spatiotemporal fusion, VIIRS NDVI**

I. INTRODUCTION

NORMALIZED Difference Vegetation Index (NDVI) data describe the vegetation greenness of land surfaces [1]. NDVI time-series data have been widely used for investigating various processes of terrestrial ecosystems, such as vegetation productivity [2], vegetation phenology [3]-[4], forest fires [5], and land cover classification [6]. Currently, NDVI time-series products are provided by a number of satellite sensors. However, these NDVI products usually have relatively coarse spatial resolution, such as the MODIS NDVI (250m-0.05°), SPOT VGT NDVI (1km), and the AVHRR GIMMS NDVI (8km). The coarse spatial resolution, ranging from hundreds of meters to several kilometers, is an obvious constraint that greatly limits their application to geographically heterogeneous areas [7]-[8]. NDVI time-series data with higher spatial resolution are thus necessary.

Due to the trade-offs between spatial and temporal resolutions, a single satellite sensor provides data with either high temporal frequency or high spatial resolution [9]-[10]. As such, spatiotemporal fusion technology has been proposed to simulate high spatiotemporal NDVI time series by blending the high-frequency but low spatial-resolution images (e.g., MODIS, referred to as the coarse resolution image) with high spatial-resolution but low-frequency images (e.g., Landsat, referred to as the fine resolution image). During the past decade, more than 50 spatiotemporal fusion algorithms have

*Corresponding author: R. Cao (email: cao.ruyin@uestc.edu.cn)

Y. Chen, R. Cao and J. Zhou are with the School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China

J. Chen and X.H. Chen are with the State Key Laboratory of Earth Surface Processes and Resource Ecology, Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China.

X.L. Zhu is with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University.

G.P. Wang and M.G. Shen are with the Key Laboratory of Alpine Ecology, Institute of Tibetan Plateau Research, CAS Center for Excellence in Tibetan Plateau Earth Sciences, Chinese Academy of Sciences, 16 Lincui Road, Beijing 100101, China.

W. Yang is with the Center for Environmental Remote Sensing, Chiba University, Chiba 263-8522, Japan.

been developed (Zhu et al. [11]). These fusion algorithms can be roughly grouped into four types. The first is unmixing-based, in which the values of fine pixels are predicted by using the spectral linear unmixing technique (e.g., LAC-GAS by Maselli and Rembold [12]; STDFA by Wu [13]; MMT by Zhukov et al. [14]). The second is weight-function-based, in which the values of fine pixels are estimated from the combined information of all input images by using weight functions (e.g., STNLFFM by Cheng et al. [15]; STARFM by Gao et al. [16]; STAARCH by Hilker et al. [17]; semi-physical fusion approach by Roy et al. [18]; ESTARFM by Zhu et al. [19]). The third is machine-learning-based, in which the relationship between the coarse and fine image pairs is described by machine learning (SPSTFM by Huang and Song [20]; BME by Li et al. [21]; EBSPTM by Wu et al. [22]). The fourth type comprises hybrid methods, in which two or more of the above technologies are integrated (e.g., STRUM by Gevaert and García-Haro [23]; FSDAF by Zhu et al. [24]).

Spatiotemporal data fusion algorithms normally employ spatial information from the fine images at base dates to assist in the production of the fusion image at the prediction date. Therefore, at least one pair of fine and coarse images at a base date is required for spatiotemporal fusion. For example, one base image pair is necessary for STARFM, STRUM, and FSDAF, and two base image pairs are required by STDFA and ESTARFM. It has been recognized that in addition to the fusion algorithms, the accuracy of spatiotemporal data fusion also strongly depends on the selection of the base image pair [25]-[26]. For better spatiotemporal data fusion performance, there are two criteria in regards to the selection of the base image pair. Taking MODIS-Landsat fusion and one-pair base images as an example, the first criterion to be considered is that the image at the base date should be as similar as possible to the one at the prediction date (called the "similarity" criterion; Fig. 1). To meet this requirement, two automatic strategies have been adopted in previous studies: the "nearest date" (ND) strategy, in which the base image pair is determined to be acquired at the closest date to the prediction date, and the "highest correlation" (HC) strategy, in which the base date is determined when the correlation coefficient of the MODIS images between the base and prediction dates is the highest. The second criterion for selecting the base image pair is that the MODIS and Landsat images at the base date should be consistent in terms of their radiometric characteristics and imaging geometry (called the "consistency" criterion; Fig. 1). Inconsistency between MODIS and Landsat images, a major problem that decreases the accuracy of data fusion, may be caused by many factors, including differences in the spectral response function and viewing angles (large viewing angle for MODIS vs. near nadir view for Landsat) and geolocation accuracy [18][27]. Figure 1 graphically illustrates the "similarity" and "consistency" criteria.

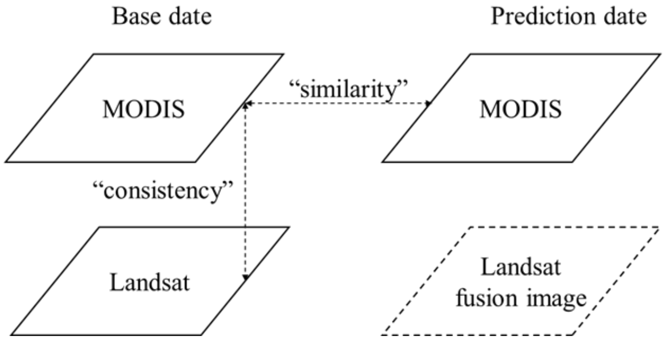


Fig. 1. A graphical illustration showing the two criteria for the selection of the input base image pair (i.e., "similarity" and "consistency").

Unfortunately, no method exists to automatically determine the base image pair by taking both "similarity" and "consistency" criteria into consideration. For example, Zhu et al. [24] adopted the ND strategy in the FSDAF spatiotemporal fusion algorithm that they developed. Wang et al. [25] proposed an operational data fusion framework in which a user can choose between the ND and HC strategies. Liu et al. [26] suggested to use multiple base image pairs to generate several fused images first, and then produce the final prediction by the weighted sum of these fused images. However, the weights were estimated by considering the difference in MODIS scale NDVI only (see Eq. 10 in [26]). Such treatment cannot account for the inconsistency between MODIS and Landsat images. In actuality, most previous studies employed only the "similarity" criterion, because MODIS images at the base and prediction dates are available for the determination of similarity. It is more difficult to consider the "consistency" criterion because consistency is hard to quantify; thus, several studies just provided some recommendations for this criterion [27]-[29]. For example, they suggested that compared with the MODIS directional reflectance product (MOD09GA), the bidirectional reflectance distribution function (BRDF) adjusted reflectance products (e.g., MCD43A4) might be better for MODIS-Landsat data fusion because the viewing angular differences between MODIS and Landsat can be corrected to a large extent [27]-[29]. Some methods were also developed to further normalize Landsat data to nadir BRDF-adjusted data [30]-[31]. Even after BRDF corrections, however, many other factors may also lead to inconsistency, such as the registration accuracy between MODIS and Landsat [25]. Therefore, it is important to develop a method to automatically determine the optimal base image pair, which is a crucial step for the practical applications of spatiotemporal data fusion technology.

In this study, we developed a novel method (called "cross-fusion") to address the issue of the determination of the base image pair. The new method can quantitatively take both "similarity" and "consistency" criteria into account. The new method operates by first choosing a number of base image pairs as candidates according to the "similarity" criterion. Considering the interannual vegetation growth cycle, these candidate image pairs are selected from multi-year images. The new method then employs all the candidate image pairs to implement spatiotemporal data fusion between them (referred to as "cross-fusion"). Performance of cross-fusion can be quantitatively evaluated because all of the true Landsat images at the candidate dates are available. Logically, we expect cross-

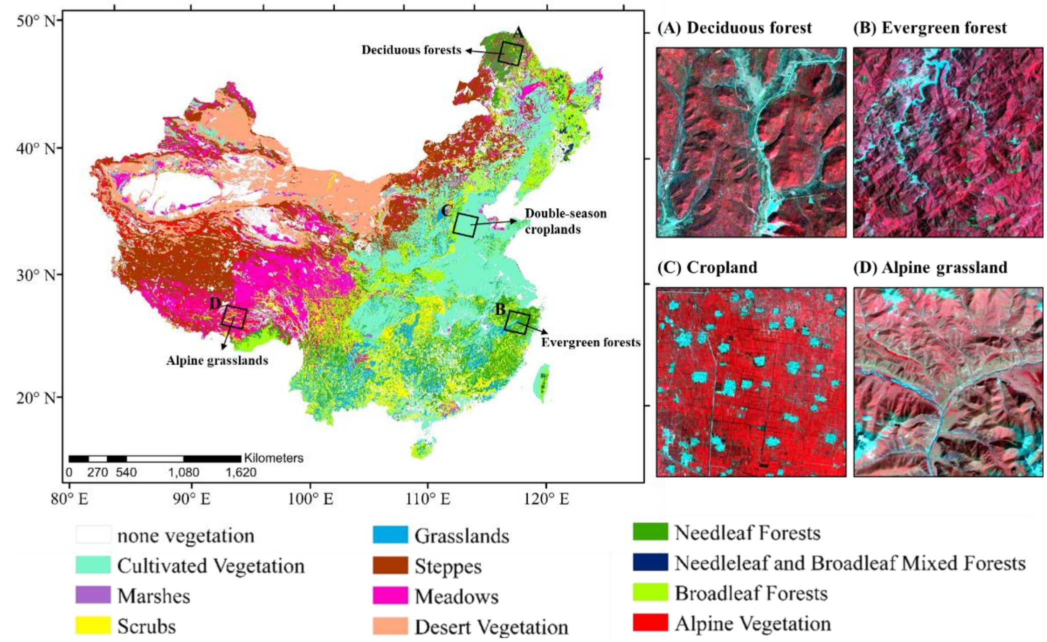


Fig. 2. Left: The spatial distribution of various vegetation types in China provided by the Editorial Board of the Vegetation Map of China CAS (2001) [38]. The four boxes A-D indicate the locations of the testing areas. Right: the Landsat images of the four testing regions (Standard false color composite).

fusion to perform well according to the "similarity" criterion. We assume that larger fusion errors can occur when using an "inconsistent" candidate image pair to implement data fusion at other candidate dates. In such a way, the cross-fusion method can further consider the "consistency" criterion. We compared the cross-fusion method with other existing selection strategies and found that the fused NDVI image from the cross-fusion method was more accurate (smaller errors) and robust. The new method provides an operational and automatic way to perform spatiotemporal data fusion in practical applications.

II. MATERIALS AND METHODS

A. MODIS and Landsat NDVI data and preprocessing

We collected MODIS and Landsat images in four testing areas covered by different vegetation types, including deciduous forests, evergreen forests, double-season croplands, and alpine grasslands (regions A, B, C and D in Fig. 2, respectively). We downloaded the BRDF-adjusted reflectance product (MCD43A4) for 2001-2016 from the website of the United States Geological Survey (USGS: https://lpdaac.usgs.gov/products/mcd43a4v006/). MCD43A4 data are available daily with a spatial resolution of 500 m. For Landsat data, we collected all available Landsat images (i.e., Landsat 5, 7, and 8) from 2001 to 2016 from the platform of Google Earth Engine. Subset areas (12×12 km2) in each of the regions (Fig. 2) were used for our experiments. Digital Numbers in Landsat images were radiometrically calibrated and atmospherically corrected by the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) [32]. We used the daily MCD43A4 product to generate 8-day composite reflectance data by averaging all cloud-free reflectance data during each 16-day time period centered at the composition date (correspondence with Dr. Zhuosen Wang from the MODIS BRDF team). There were still some missing data in the 8-day composite time-series reflectance. We therefore filled these missing data by the linear interpolation and smoothed the time-series data with the Savitzky-Golay filter [33]-[34].

To implement spatiotemporal data fusion, both MODIS and Landsat images were co-registered, and MODIS NDVI images were resampled by the nearest neighbor method to match the Landsat spatial resolution. Cloud contamination in each Landsat image was automatically detected by the Fmask (function of mask) method [35]-[36]. We followed Xie et al. [27] to define that only the12×12 km^2 subset images with cloud cover below 1% were used. As a result, we screened out in total 171, 137, 90 and 155 Landsat images for the areas of deciduous forests, evergreen forests, double-season croplands, and alpine grasslands, respectively.

There are two strategies to generate NDVI fused images. One is to blend the reflectance images and then calculate NDVI from the fused images (i.e., Blend-then-Index), and the other is to first calculate NDVI and then blend the coarse and fine NDVI images (i.e., Index-then-Blend). We adopted the Index-then-Blend strategy in this study because of its better performance, as suggested by Jarihani *et al*. [37]

B. Developing the cross-fusion method

There are two steps in the cross-fusion method. In the first step, we employ the "similarity" criterion to choose a certain number of candidate base image pairs. In the second step, we address the "consistency" issue by employing all the candidate image pairs for the cross-fusion process. Fig. 3 shows the flowchart of the new method.

1) Step 1: Determining the candidate base image pairs

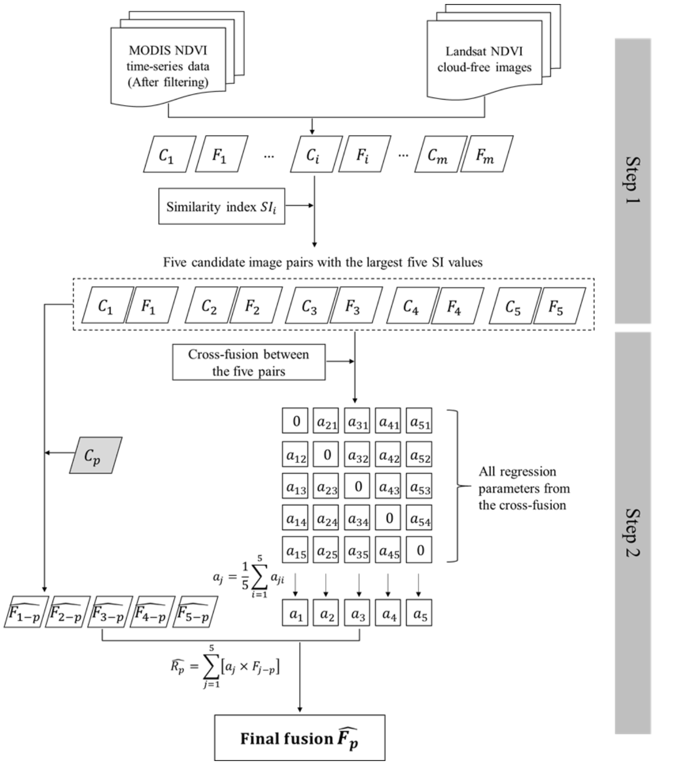

MODIS NDVI time-series data (After filtering)
Landsat NDVI cloud-free images
C_1 F_1 ... C_i F_i ... C_m F_m
Similarity index SI_i
Five candidate image pairs with the largest five SI values
C_1 F_1 C_2 F_2 C_3 F_3 C_4 F_4 C_5 F_5
Cross-fusion between the five pairs
0 a_{21} a_{31} a_{41} a_{51}
a_{12} 0 a_{32} a_{42} a_{52}
a_{13} a_{23} 0 a_{43} a_{53}
a_{14} a_{24} a_{34} 0 a_{54}
a_{15} a_{25} a_{35} a_{45} 0
All regression parameters from the cross-fusion
C_p
$a_j = \frac{1}{5}\sum_{i=1}^{5} a_{ji}$
a_1 a_2 a_3 a_4 a_5
$\widehat{F_{1-p}}$ $\widehat{F_{2-p}}$ $\widehat{F_{3-p}}$ $\widehat{F_{4-p}}$ $\widehat{F_{5-p}}$
$\widehat{R_p} = \sum_{j=1}^{5}[a_j \times F_{j-p}]$
Final fusion $\widehat{F_p}$
Step 1
Step 2

Fig. 3. Flowchart of the cross-fusion method. *Ci* and *Fi* represent the coarse (MODIS) and fine (Landsat) images, respectively. *Cp* is the coarse image at the prediction date.

To quantify the similarity between the MODIS image at the prediction date R_p and the one at a given date R_i, we calculated the reflectance difference $diff(R_{p\_i})$ and linear correlation coefficient $cor(R_{p\_i})$ between the two MODIS images as follows:

$$diff(R_{p\_i}) = \frac{1}{n}\sum_{j=1}^{n}\left|R_p(j) - R_i(j)\right|$$
$$cor(R_{p\_i}) = \frac{Cov(\{R_p(j)\},\{R_i(j)\})}{\sqrt{Var(\{R_p(j)\})}\sqrt{Var(\{R_i(j)\})}},$$
$$j = 1, \dots, n \qquad (1)$$

where n indicates the number of MODIS pixels in the subset image, and the functions *Var*() and *Cov*() represent estimations of the variance and covariance, respectively. Assuming that the total number of image pairs during multiple years is *m* (i.e., *i* = 1,…,m), we combined $diff(R_{p\_i})$ and $cor(R_{p\_i})$ to calculate the similarity index of the *i*th image (SI_i) as follows:

$$SI_i = \frac{(1-diff(R_{p\_i}))}{\sum_{i=1}^{m}(1-diff(R_{p\_i}))} \times \frac{cor(R_{p\_i})}{\sum_{i=1}^{m}cor(R_{p\_i})} \qquad (2)$$

Eq. (2) suggests that a smaller reflectance difference and a higher correlation coefficient lead to a larger *SI* value. Therefore, the candidate image pairs are determined to be those that have the largest five *SI* values. We also tested more or fewer images pairs and found that five is the best choice considering the balance between accuracy and computing time (see the discussion section for details).

2) Step 2: Cross-fusion among the candidate base image pairs

Assuming that the five candidate image pairs (coarse image, fine image) are denoted as (C_1, F_1), (C_2, F_2), (C_3, F_3), (C_4, F_4) and (C_5, F_5), we perform cross-fusion using these image pairs. For example, we can use (C_2, F_2), (C_3, F_3), (C_4, F_4), and (C_5, F_5) to predict F_1. For the reflectance at a given pixel (x, y) in the image F_1 (i.e., F_1(x, y)), the predictions of F_1(x, y) from spatiotemporal data fusion are $Fusion(F_2(x,y) \rightarrow F_1(x,y))$, $Fusion(F_3(x,y) \rightarrow F_1(x,y))$, $Fusion(F_4(x,y) \rightarrow F_1(x,y))$, and $Fusion(F_5(x,y) \rightarrow F_1(x,y))$. We assume that the final prediction of $F_1(x, y)$ (denoted as $\widehat{F_1}(x,y)$) can be calculated as

$$\begin{aligned}\widehat{F_1}(x,y) = {} & a_{21} \times Fusion(F_2(x,y) \rightarrow F_1(x,y)) + \\ & a_{31} \times Fusion(F_3(x,y) \rightarrow F_1(x,y)) + \\ & a_{41} \times Fusion(F_4(x,y) \rightarrow F_1(x,y)) + \\ & a_{51} \times Fusion(F_5(x,y) \rightarrow F_1(x,y)) \\ \text{s.t.} \quad & a_{21} + a_{31} + a_{41} + a_{51} = 1.0 \quad \text{and} \\ & 0 \le (a_{21},\ a_{31},\ a_{41},\ a_{51}) \le 1.0 \end{aligned} \qquad (3)$$

To estimate the four regression parameters in Eq. (3), we employ all pixels in a 5 × 5 local window with (x, y) as the central pixel and minimize the following objective function. We chose the local window size of 5 × 5 to generate enough equations to solve the four regression parameters for all the pixels including the pixels in the edge and corner of the image (see Fig. S1 in the supplementary materials).

$$\underset{(a_{21},\ a_{31}\ ,\ a_{41},\ a_{51})}{\mathbf{arg\,min}} \quad \sum_{i=1}^{25}(F_1(x_i,y_i) - \widehat{F_1}(x_i,y_i))^2 \qquad (4)$$

The four regression parameters are constrained between 0 and 1 and their sum is 1. Therefore, the regression parameter for a prediction item in Eq. (3) (e.g., a_{21}) could be small if this prediction item (e.g., $Fusion(F_2(x,y) \rightarrow F_1(x,y))$) deviates much from the true value. The larger prediction error at (x, y) can be explained by the inconsistency between MODIS and Landsat because all of the candidate image pairs are chosen according to the "similarity" criterion. In the same way, we can estimate all regression parameters for the cross-fusion, expressed as:

$$\begin{matrix} 0 & a_{21} & a_{31} & a_{41} & a_{51} \\ a_{12} & 0 & a_{32} & a_{42} & a_{52} \\ a_{13} & a_{23} & 0 & a_{43} & a_{53} \\ a_{14} & a_{24} & a_{34} & 0 & a_{54} \\ a_{15} & a_{25} & a_{35} & a_{45} & 0 \end{matrix} \qquad (5)$$

According to Eq. (3), the sum of each row of Eq. (5) is 1.0 (e.g., $a_{21} + a_{31} + a_{41} + a_{51}$ = 1.0). The values in each column of Eq. (5) represent the regression parameters for using a candidate image pair to generate other candidate image pairs. For example, a_{12}, a_{13}, a_{14}, and a_{15} in the first column are the parameters when using (C_1, F_1) to produce the pixel (x, y) in (C_2, F_2), (C_3, F_3), (C_4, F_4) and (C_5, F_5), respectively. As mentioned above, a smaller parameter for a given image pair generally indicates a larger prediction error which is likely caused by inconsistency between MODIS and Landsat data for this image pair. Therefore, we use the average value of each column to represent the contribution of each image pair to the final prediction of $R_p(x,y)$ ($\widehat{R_p}(x,y)$), expressed as:

$$\widehat{R_p}(x,y) = \sum_{j=1}^{5}[a_{j\_} \times Fusion\left(F_j(x,y) \rightarrow F_p(x,y)\right)]$$
$$\text{with } a_{j\_} = \frac{1}{5}\sum_{i=1}^{5} a_{ji} \qquad (6)$$

where $Fusion\left(F_j(x,y) \rightarrow F_p(x,y)\right)$ is to use the *j*th candidate image pair to produce the pixel (x, y) in the fusion image at the prediction date. Using Eqs. (3-6) and the moving window, we can perform spatiotemporal fusion to predict all pixels. In actuality, the cross-fusion uses not one but five base image pairs, and different candidate image pairs may contribute differently to predict different pixels.

To test the cross-fusion, we chose the FSDAF algorithm as the spatiotemporal data fusion algorithm (i.e., using FSDAF in Eqs. (3) and (6)) because of two reasons. First, the cross-fusion method is only applicable for spatiotemporal data fusion algorithms that require one input image pair, such as FSDAF [24]. Therefore, those multi-input-pair fusion algorithms (e.g., ESTARFM [19]) were not considered. Second, The FSDAF fusion algorithm was found to performs well in various scenarios, even in some challenging cases such as heterogeneous landscapes and abrupt changes of land cover types [11]. For more details about the FSDAF algorithm, please refer to Zhu *et al*. [24]. In the following experiments, we compared the performance of FSDAF when using the base image pair determined by the cross-fusion and other methods (see section 3.1).

III. EXPERIMENTAL DESIGN

A. Experiment I: Quantitative assessments at random prediction dates

We performed quantitative evaluations when using different selection strategies to determine the base image pair for FSDAF. We considered the strategies *ND* (nearest date), *HC* (highest correlation), *Diff* (smallest differences; Eq. 1), *SI* (largest similarity index; Eq. 2), and the cross-fusion method. In this experiment, we randomly selected 10% of the NDVI images in each testing area and used these images as the truth. The quantitative evaluations were performed by comparing the truth image with the fused image. Owing to a failure of the scan-line corrector (referred to as SLC-off), there are missing strips in Landsat 7 ETM+ images after May 2003. These SLC-off images were not chosen. As a result, 12, 14, 11, and 14 Landsat images were selected in the testing areas of deciduous forest, cropland, evergreen forest, and alpine grassland, respectively. Table 1 showed the detailed information for all the selected images. Two accuracy evaluation indices were used: the average absolute difference ($AAD = \left|NDVI_{fusion} - NDVI_{true}\right|$) and the correlation coefficient between the fused and true Landsat images. For the quantitative evaluations in this experiments and other experiments below, AAD and the Correlation Coefficient were computed for only the clear pixels.

B. Experiment II: Simulated the inconsistence between MODIS and Landsat

Compared with previous base-image pairs selection methods, the most significant improvement of the new method is that it can really quantify the inconsistence between the coarse and fine images at the base date. We thus performed a simulation experiment to illustrate this point. To be exact, we simulated the scenarios for the inconsistence between MODIS and Landsat data, in which each pixel in the Landsat base image is multiplied by a random number between 0.8 and 1.2. This simulation experiment was performed on the same images as those used in Experiment I (i.e., all images in Table 1). Because the cross-fusion method chose five candidate image pairs, we assume that the inconsistence occurs in a different number of image pairs (varying from 1 to 5). Specifically, at each predication date, we randomly chose 1, 2, 3, 4, and 5 image pairs from the five candidate image pairs, respectively, and simulated noise on these selected images. We tested the performance of the cross-fusion method under these simulated scenarios.

C. Experiment III: Effect of cloud omission errors on the cross-fusion method

Only a cloud-free Landsat image can be used as the base image. Currently, cloud pixels in Landsat images are identified by the Fmask method [35]-[36]. However, there are cloud detection errors in Landsat images [39]-[40]. In this experiment, we assume that one of the five candidate image pairs have a small amount of clouds. This experiment was also performed on all the randomly selected images (i.e., all images in Table 1). We investigate whether and to what extent the cross-fusion method is affected by cloud omission errors.

D. Experiment IV: Generating the fused NDVI time-series data

In this experiment, we investigate the performance of the cross-fusion method to generate the fused NDVI time-series data. Because there are 46 MODIS images in a year (8-d temporal resolution), we employed the new method to produce 92 fused NDVI images in two consecutive years. The years were selected randomly for each testing region. For comparisons, we also generated the fused NDVI time series by using the ND, HC, Diff, and SI strategies, respectively.

E. Experiment V: Visual inspection for block effect of the fused image

We choose candidate image pairs based on the subset MODIS images (12 × 12 km^2). Therefore, different candidate image pairs may be chosen for a different subset area when applying the cross-fusion method to a larger area. In this experiment, we test the performance of the cross-fusion method in a 108 × 108 km2 region (i.e., 9 × 9 subset image), and investigate whether there is a block effect in the Landsat fused image. We performed this experiment for one image per vegetation type. The predication dates are 2014/105 (year/day of year) for deciduous forest, 2005/145 for cropland, 2010/144 for evergreen forest, and 2004/254 for grassland.

Table 1. The imaging dates (Year/Day) and the sensors for the randomly selected images for comparisons. L5 TM: Landsat 5 TM; L7 ETM+: Landsat 7 ETM+; L8 OLI: Landsat 8 OLI.

Deciduous Forest		**Evergreen Forest**	
Year/Day	**Sensor**	**Year/Day**	**Sensor**
2001/157	L7 ETM+	2001/327	L5 TM
2004/014	L5 TM	2002/314	L5 TM
2004/062	L5 TM	2003/269	L5 TM
2006/323	L5 TM	2003/349	L5 TM
2007/158	L5 TM	2004/208	L5 TM
2008/209	L5 TM	2006/213	L5 TM

2010/326	L5 TM	2008/003	L5 TM
2011/313	L5 TM	2008/187	L5 TM
2013/302	L8 OLI	2010/144	L5 TM
2014/241	L8 OLI	2011/139	L5 TM
2015/028	L8 OLI	2013/296	L8 OLI
2016/143	L8 OLI		

Cropland		Grassland	
Year/Day	Sensor	Year/Day	Sensor
2001/067	L7 ETM+	2003/115	L5 TM
2001/187	L5 TM	2003/339	L5 TM
2004/100	L5 TM	2004/326	L5 TM
2005/094	L5 TM	2005/312	L5 TM
2005/142	L5 TM	2005/320	L5 TM
2006/073	L5 TM	2006/299	L5 TM
2006/161	L5 TM	2007/118	L5 TM
2008/207	L5 TM	2007/318	L5 TM
2009/353	L5 TM	2007/334	L5 TM
2010/140	L5 TM	2010/326	L5 TM
2014/359	L8 OLI	2011/033	L5 TM
2015/138	L8 OLI	2012/140	L5 TM
2015/282	L8 OLI	2013/342	L8 OLI
2016/013	L8 OLI	2013/350	L8 OLI

IV. RESULTS

A. Results for fusion at random prediction dates (Experiment I)

Figure 4 shows the average values of both the *AAD* and the correlation coefficient with different selection strategies for different vegetation types (for the detailed results of each Landsat image, please refer to the Table. S1 in the supplementary materials). We found that the cross-fusion method achieved the lowest *AAD* values and the highest correlation coefficient compared with other selection strategies, suggesting better and stable performance of the new method (Fig. 4). The *SI* strategy seemed to be the second best, except that *SI* had a larger *AAD* value than the *Diff* strategy for deciduous forest. This result indicates that combining *HC* and *Diff* in SI (Eq. 2) is effective. The *ND* strategy performed the worst, especially for cropland, which may be explained as follows. Although the closest date was chosen, there may still have been a large difference in land cover between the base and prediction dates if there was temporally continuous cloud contamination or the land cover changed quickly, such as for multi-season cropland.

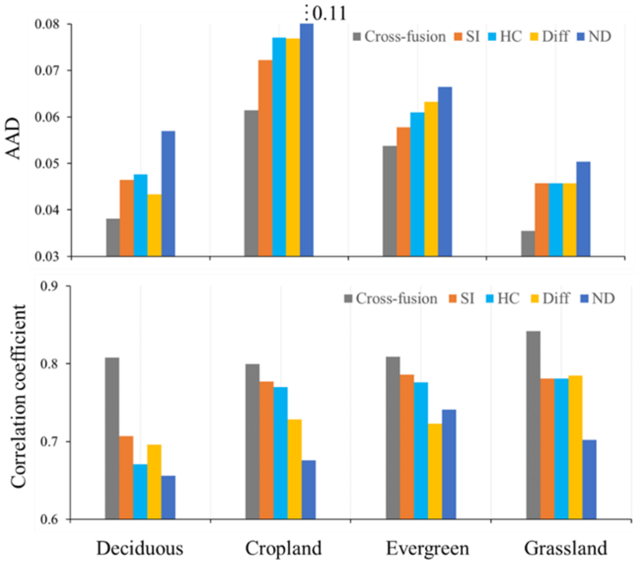


Fig. 4. Quantitative evaluation results (*AAD* and correlation coefficient) for data fusion with different selection strategies for different vegetation types. Note that: the *AAD* for the *ND* strategy and cropland is 0.11, which is beyond the range of the y-axis.

The new method implements cross-fusion using the five candidate image pairs. To investigate the effectiveness of such treatment, we compared the spatiotemporal data fusion accuracy when using the cross-fusion method to determine the base image pair or using each of the five candidates as the base image pair (referred to as SI_1, SI_2, SI_3, SI_4, and SI_5, respectively). The results showed that the cross-fusion method had the lowest *AAD* values and the highest correlation coefficient (Fig. 5). Interestingly, a candidate image pair with a higher similarity index (Eq. 2) did not necessarily achieve better performance. For example, SI_1 (highest similarity index) had lower *AAD* values than the other four candidate image pairs for deciduous and evergreen forests but not for cropland and grassland (Fig. 5). This result may be due to the inconsistency between the MODIS and Landsat base images and the inconsistency between the time-series of MODIS and Landsat images due to heterogeneous landscape. This investigation further highlights the importance of considering the “inconsistence” criterion in spatiotemporal data fusion.

Table 2 shows detailed comparisons at each prediction date for deciduous forest. It can be seen that the cross-fusion method performs better (lowest *AAD* value) than any one of the five candidate image pairs for 7 out of 12 prediction dates (i.e., 7/12). The percentages were 10/14, 5/11, and 10/14 for cropland, evergreen forest, and grassland (see Table S2 in the supplementary materials). The explanation for these results is that the five candidate image pairs used by the cross-fusion method contributed differently to different pixels; thus, the cross-fusion could achieve higher accuracy than any of the candidate image pairs at most of the prediction dates. We also noted that the cross-fusion is not the best at some prediction dates. However, the new method is still reliable in practical applications, which can be explained as: for those dates at which the cross-fusion did not achieve the lowest *AAD* values, it is almost impossible for the users to select the best candidate image pair (i.e., the lowest *AAD* value) because the best pair seemed to randomly occur in one of the five candidate image pairs (Table S2). However, the *AAD* values of the cross-fusion tends to be closer to, albeit somewhat larger than, the *AAD* values of the best candidate image pair. For example, for the 11 prediction dates of the evergreen forest, the cross-fusion performed the best for 5 dates and achieved the second lowest *AAD* values for other 6 prediction dates. On average, the cross-fusion method achieved the highest accuracy for all the vegetation types (Fig. 5).

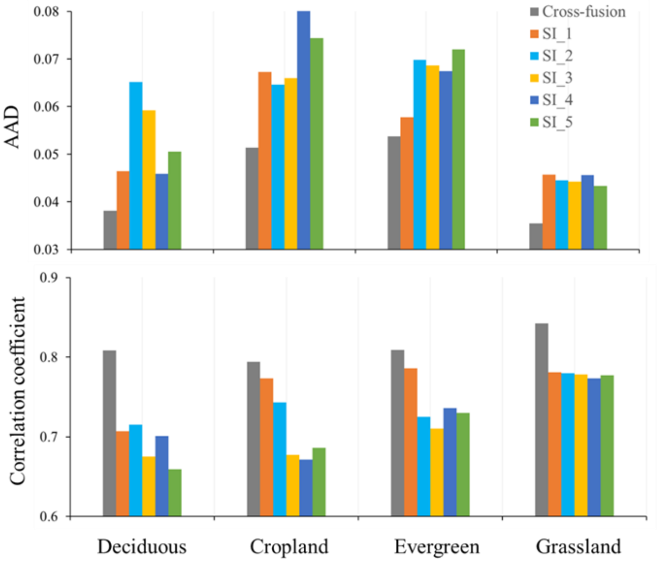


Fig. 5. Quantitative comparisons (*AAD* and correlation coefficient) between the cross-fusion method and the five candidate image pairs (denoted as SI_1, SI_2, SI_3, SI_4, and SI_5, respectively) for different vegetation types. Noted: results are the averaged *AAD* values and correlation coefficients for all the randomly selected images.

Table 2. Performance of spatiotemporal data fusion for both the cross-fusion method and the five candidate image pairs (denoted as *SI*_1, *SI*_2, *SI*_3, *SI*_4, and *SI*_5, respectively) at each prediction date for deciduous forest. We show (*AAD*/correlation coefficient) in the table. The highest *AAD* value at each predication date is in bold

Year/Day	**Cross-fusion**	***SI*_1**	***SI*_2**	***SI*_3**	***SI*_4**	***SI*_5**
2001/157	0.0688 /0.923	**0.0481** **/0.958**	0.0599 /0.777	0.0685 /0.915	0.0975 /0.857	0.093 /0.780
2004/014	**0.0399** **/0.785**	0.0444 /0.423	0.0462 /0.735	0.0403 /0.738	0.0432 /0.647	0.0648 /0.784
2004/062	**0.0149** **/0.694**	0.0215 /0.633	0.0178 /0.411	0.0189 /0.503	0.0192 /0.591	0.0169 /0.483
2006/323	0.0797 /0.770	0.0669 /0.594	0.0998 /0.704	0.1406 /0.415	**0.0423** **/0.573**	0.0816 /0.563
2007/158	**0.0297** **/0.934**	0.0834 /0.914	0.0371 /0.940	0.0622 /0.698	0.0416 /0.807	0.0459 /0.762
2008/209	**0.0181** **/0.853**	0.0376 /0.734	0.0353 /0.824	0.0537 /0.811	0.0556 /0.829	0.0466 /0.813
2010/326	0.0333 /0.751	0.041 /0.290	0.2565 /0.742	0.0431 /0.657	**0.0325** **/0.657**	0.0383 /0.638
2011/313	0.0518 /0.665	0.0662 /0.562	0.0721 /0.580	**0.0457** **/0.542**	0.0535 /0.664	0.0642 /0.137
2013/302	**0.0346** **/0.676**	0.0366 /0.741	0.0436 /0.556	0.095 /0.406	0.0457 /0.384	0.0395 /0.688
2014/241	**0.0246** **/0.799**	0.0287 /0.823	0.0472 /0.594	0.0329 /0.707	0.0273 /0.738	0.0431 /0.654
2015/028	**0.0156** **/0.916**	0.023 /0.939	0.026 /0.815	0.0232 /0.835	0.0247 /0.787	0.0232 /0.822
2016/143	0.0458 /0.933	0.0598 /0.88	**0.0415** **/0.903**	0.0866 /0.872	0.0674 /0.883	0.0493 /0.787
Average	**0.0381** **/0.808**	**0.0464** **/0.707**	**0.0652** **/0.715**	**0.0592** **/0.675**	**0.0459** **/0.701**	**0.0505** **/0.659**

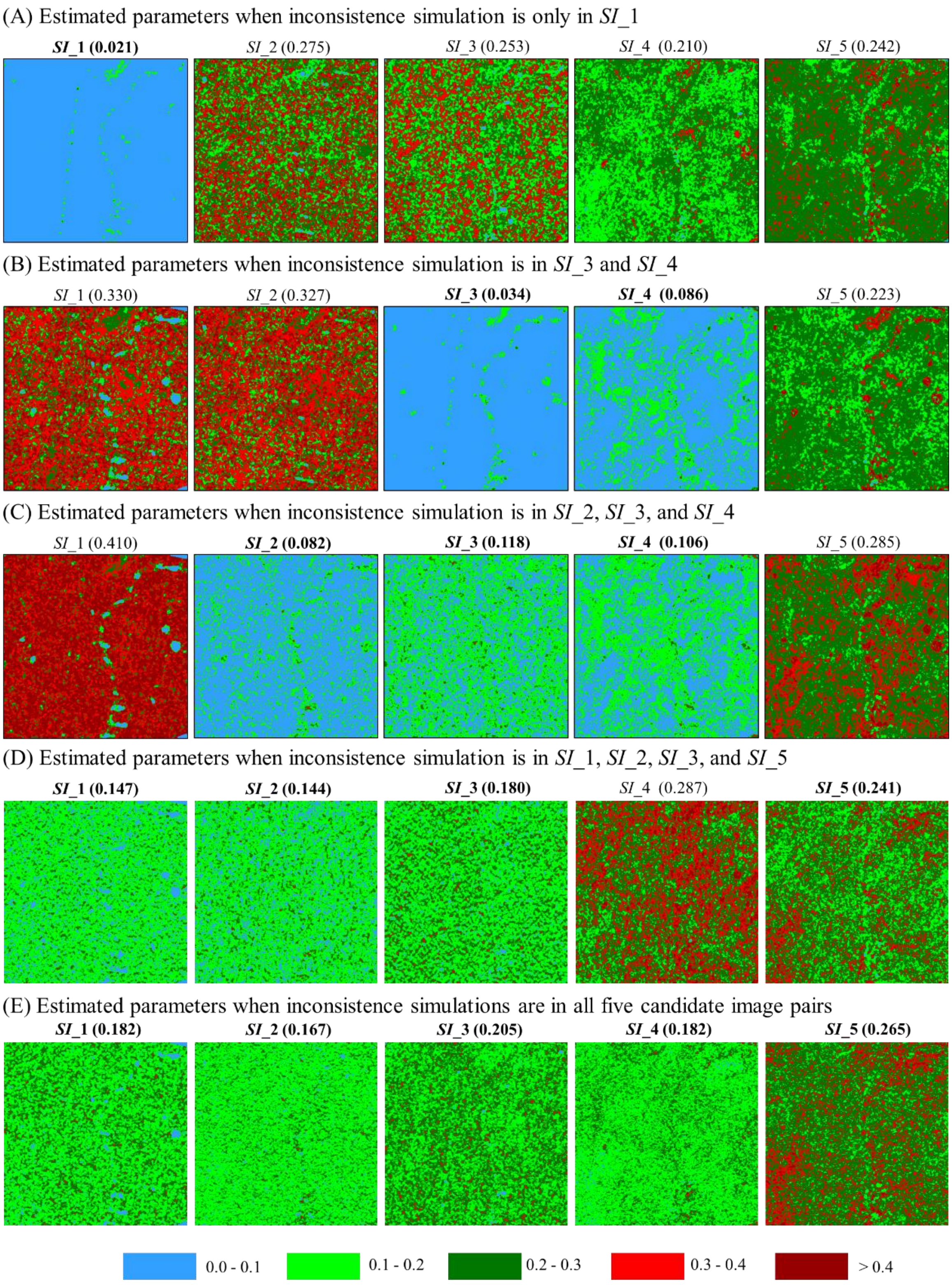
B. Results for the simulated inconsistence between MODIS and Landsat (Experiment II)

We simulated inconsistence between MODIS and Landsat data by multiplying each pixel in the Landsat image by a random number within the range 0.8-1.2. Due to page limitations, here we showed the results for one prediction date per vegetation type in Table 3. For the results at all the prediction dates, please refer to Table. S3 in the supplementary materials. We considered different numbers of candidate image pairs with the simulated inconsistence (see the second column in Table 3). As expected, inconsistence between MODIS and Landsat data greatly reduces the accuracy of spatiotemporal data fusion (see the bold numbers in Table 3). For example, the *AAD* value increased from 0.0376 to 0.0818 when simulating inconsistence for SI_1 of deciduous forest (compare Table 2 with Table 3). These results confirm that the "inconsistence" issue should be addressed for spatiotemporal data fusion.

We found that the cross-fusion method was less affected by these simulated inconsistences (Table 3). More importantly, the cross-fusion method performed better than any one of the five candidate image pairs, even for the case in which all of the five candidate image pairs were inconsistent. To further understand the performance of the cross-fusion method, we show the regression parameters (Eq. 6) for the five candidate image pairs of deciduous forest at the prediction date (year/day of year: 2008/209) (Fig. 6). We expect a small parameter for the image pair with simulated inconsistence, which suggests small contribution for the fusion from this image pair. The results confirmed that the parameters for the image pairs with simulated inconsistence were substantially smaller than the parameters for other candidate image pairs (e.g., 0.02 for SI_1 vs. 0.2-0.3 for the other four in Fig. 6A). This investigation suggests that inconsistence between MODIS and Landsat data is truly accounted for by the new method. For the extreme case of inconsistence existing in all of the five candidate image pairs, we found that the parameters of the five candidate image pairs were around 0.2 and their differences were relatively small (Fig. 6E). In summary, the cross-fusion method provides an automatic way to determine the base image pair by considering both "similarity" and "consistency" criteria.

Table 3. Performance of spatiotemporal data fusion for both the cross-fusion method and the five candidate image pairs (denoted as *SI*_1, *SI*_2, *SI*_3, *SI*_4, and *SI*_5, respectively) for the simulated scenarios of "inconsistence" between MODIS and Landsat. We simulated different numbers of image pairs with inconsistence (see the second column) and the simulated image pairs are expressed in bold. We show (*AAD* / correlation coefficient) in the table

Types /Day	**Pairs**	**·Cross -fusion**	***SI*_1**	***SI*_2**	***SI*_3**	***SI*_4**	***SI*_5**
Deciduous (2008/209)	1	0.0189 /0.848	**0.0818 /0.472**	0.0353 /0.824	0.0537 /0.811	0.0556 /0.829	0.0466 /0.813
	2	0.0200 /0.847	0.0376 /0.734	0.0353 /0.824	**0.0710 /0.599**	**0.0681 /0.581**	0.0466 /0.813
	3	0.0218 /0.833	0.0376 /0.734	**0.0728 /0.536**	**0.0681 /0.608**	**0.0677 /0.581**	0.0466 /0.813
	4	0.0296 /0.783	**0.0819 /0.470**	**0.0738 /0.533**	**0.0701 /0.603**	0.0556 /0.829	**0.0570 /0.642**
	5	0.0319 /0.762	**0.0821 /0.465**	**0.0745 /0.531**	**0.0687 /0.609**	**0.0698 /0.572**	**0.0558 /0.652**
Cropland (2006/073)	1	0.0188 /0.786	**0.0348 /0.658**	0.0223 /0.700	0.0595 /0.615	0.0227 /0.644	0.0219 /0.612
	2	0.0187 /0.786	**0.0347 /0.659**	0.0223 /0.700	**0.0596 /0.584**	0.0227 /0.644	0.0219 /0.612
	3	0.0197 /0.783	0.0306 /0.719	**0.0264 /0.644**	0.0595 /0.615	**0.0275 /0.582**	**0.0249 /0.553**
	4	0.0194 /0.776	**0.0350 /0.656**	**0.0263 /0.643**	**0.0596 /0.589**	**0.0274 /0.581**	0.0219 /0.612
	5	0.0198 /0.772	**0.0352 /0.653**	**0.0262 /0.644**	**0.0596 /0.582**	**0.0273 /0.583**	**0.0249 /0.554**
Evergreen (2013/296)	1	0.0263 /0.965	**0.0678 /0.866**	0.0581 /0.798	0.0345 /0.922	0.0635 /0.723	0.0357 /0.924
	2	0.0232 /0.965	0.0355 /0.947	**0.0861 /0.596**	0.0345 /0.922	**0.0929 /0.499**	0.0357 /0.924
	3	0.0278 /0.953	**0.0681 /0.866**	0.0581 /0.798	**0.0812 /0.733**	**0.0940 /0.491**	0.0357 /0.924
	4	0.0289 /0.951	**0.0686 /0.863**	**0.0865 /0.596**	0.0345 /0.922	**0.0934 /0.496**	**0.0641 /0.838**
	5	0.0361 /0.919	**0.0685 /0.864**	**0.0864 /0.596**	**0.0807 /0.734**	**0.0925 /0.500**	**0.0638 /0.838**
Grassland (2013/342)	1	0.0148 /0.981	**0.0407 /0.868**	0.0220 /0.950	0.0299 /0.916	0.0276 /0.930	0.0236 /0.946
	2	0.0155 /0.980	0.0366 /0.891	**0.0296 /0.934**	0.0299 /0.916	0.0276 /0.930	**0.0327 /0.926**
	3	0.0163 /0.978	0.0366 /0.891	**0.0295 /0.935**	**0.0373 /0.894**	0.0276 /0.930	**0.0326 /0.894**
	4	0.0166 /0.978	**0.0407 /0.867**	**0.0296 /0.934**	0.0299 /0.916	**0.0346 /0.916**	**0.0327 /0.924**
	5	0.0178 /0.976	**0.0407 /0.867**	**0.0296 /0.934**	**0.0374 /0.893**	**0.0346 /0.916**	**0.0326 /0.925**



(A) Estimated parameters when inconsistence simulation is only in *SI*_1

(B) Estimated parameters when inconsistence simulation is in *SI*_3 and *SI*_4

(C) Estimated parameters when inconsistence simulation is in *SI*_2, *SI*_3, and *SI*_4

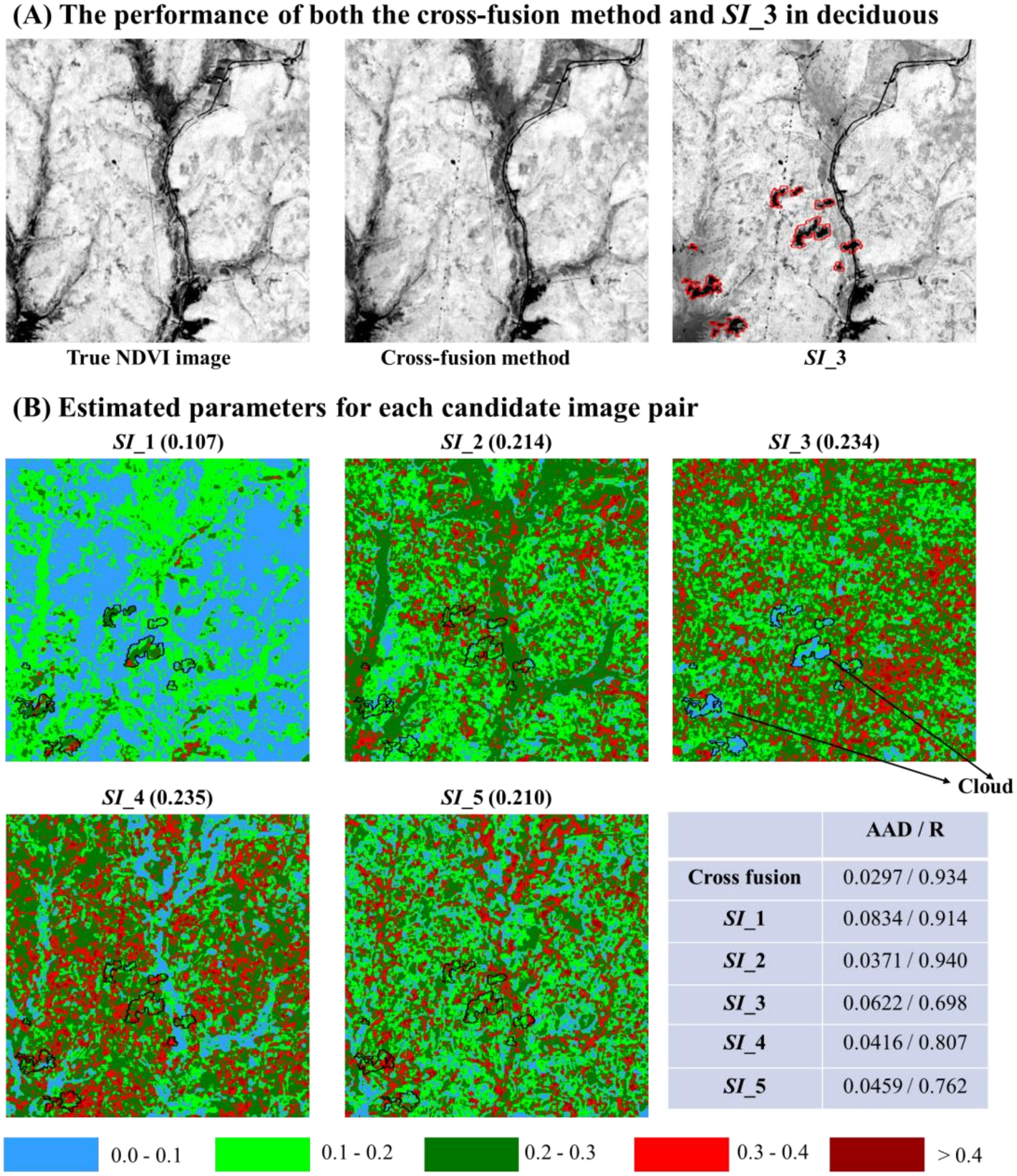
(D) Estimated parameters when inconsistence simulation is in *SI*_1, *SI*_2, *SI*_3, and *SI*_5

(E) Estimated parameters when inconsistence simulations are in all five candidate image pairs

Fig. 6. Estimated parameters (Eq. 6) for the five candidate image pairs for deciduous forest at the prediction date 2008/209 in the five simulated scenarios (A-E). The average value of each image is shown in parentheses above each image.

C. Results for the effect of cloud omission errors on the cross-fusion method (Experiment III)

We investigated the effect of cloud omission errors on the cross-fusion method by using a cloud-contaminated Landsat image as one of the five candidate base image pairs. Here, we showed the result for the prediction date (year/day of year: 2007/158) for the deciduous forest as an example. We found that compared with the fused image from the cloud-contaminated base image pair (i.e., SI_3), the fused image from the cross-fusion method was less affected by the cloud contamination (Fig. 7A). The pixels with cloud contamination have obvious fusion errors in SI_3 (see the red polygons in Fig. 7A), whereas these pixels can be successfully predicted by the new method. This result is because clouds in the Landsat image led to inconsistency between Landsat and MODIS data, and this inconsistency in local area can also be quantified by the new method. To verify this explanation, we further investigated the regression parameters (Eq. 6) for the five candidate image pairs (Fig. 7B). Interestingly, cloud-contaminated pixels have much smaller values than other pixels in SI_3 (see the black polygons in SI_3 of Fig. 7B), which suggests that inconsistency between Landsat and MODIS can be determined by the cross-fusion method at the pixel scale. Similar results for other prediction dates and vegetation types can be found in Table. S4 and Fig. S2 in the supplementary materials. It is almost impossible when cloud omission exists in the same location in all five candidate Landsat base images. Therefore, the cross-fusion method is resistant to noise such as cloud omission errors.



	AAD / R
Cross fusion	0.0297 / 0.934
***SI*_1**	0.0834 / 0.914
***SI*_2**	0.0371 / 0.940
***SI*_3**	0.0622 / 0.698
***SI*_4**	0.0416 / 0.807
***SI*_5**	0.0459 / 0.762

Fig. 7. (A) The true NDVI image for the date (2007/158) for deciduous forest, and the fused NDVI images using both the cross-fusion method and the *SI* strategy (i.e., *SI*_3). We highlighted the pixels with cloud omission errors with red polygons in *SI*_3. (B) The estimated parameters (Eq. 6) for the five candidate image pairs used by the cross-fusion method. The average values for each image are shown in the bracket above each image. Note that: The corresponding pixels with cloud omission errors are highlighted by black polygons. The last panel shows the performance of data fusion for all candidate image pairs.

D. Results for the generation of the fused NDVI time-series data (Experiment IV)

We calculated the averaged NDVI values for all pixels in each fused image and showed the time series of the average values in Figure 8. We found that the fused NDVI time series generated by cross-fusion were visually much better than those generated by other selection strategies. The cross-fusion method produced smoother NDVI time-series data. We further compared the fused NDVI time series with the existing true NDVI values at some dates by calculating the deviations between them (i.e., deviation = |fused NDVI – true NDVI|). Results showed that the averaged deviations were the smallest for the cross-fusion method in all the four testing regions. For example, for the area covered by deciduous forest, the averaged deviations are 0.029, 0.040, 0.053, 0.056, and 0.045 for cross-fusion, SI, HC, Diff, and ND strategies, respectively (Fig. 8).

E. Results of visual inspection for the block effect in the fused image (Experiment IV)

We performed the cross-fusion method in the large area for one image per vegetation type. Because the size of the image is 108 × 108 km^2, there are a total of 81 subset images (12 × 12 km^2 for each subset). Results showed that there is no block effect in the fused image for 2005/145 (year/day of year) for cropland (Fig. 9). Similar results were also found for other vegetation types (see Fig. S3 in the supplementary materials).

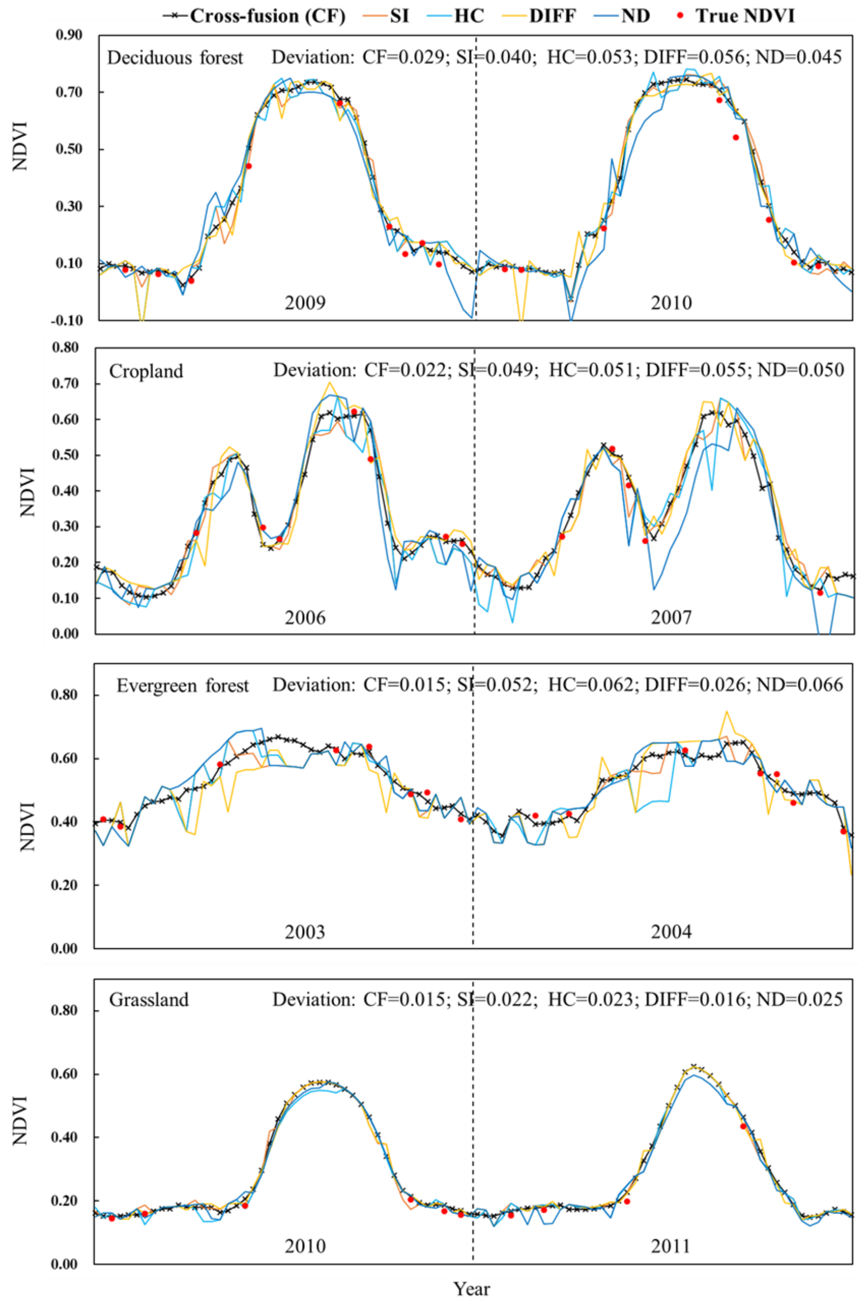


Fig. 8. The time series of the NDVI values averaged over all pixels in each fused image. The deviation values were calculated between the fused NDVI time series and the existing true NDVI values at some dates.

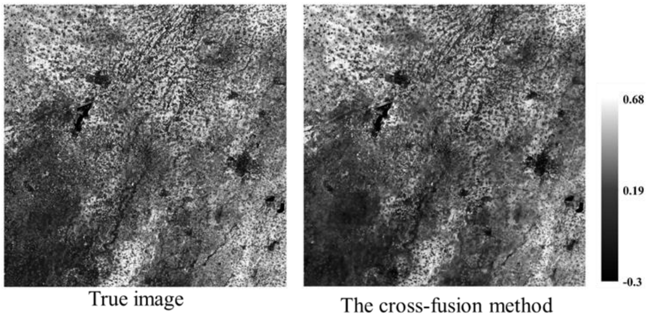


Fig. 9. The true image and the fused image by the cross-fusion method for cropland for 2005/145 (year/day of year). The size of the image is 108 × 108 km^2

V. DISCUSSION AND CONCLUSIONS

A. Performances of the cross-fusion method

The performance of spatiotemporal data fusion is mainly affected by two factors. One factor is to model the reflectance changes between the base and prediction dates, and the other is the selection of base image pairs. Most previous efforts to improve data fusion accuracy have focused on the first factor, and little effort has been made in exploring how to determine the optimal base image pair. A few studies compared some selection strategies such as the nearest date or the highest similarity [25][27]. However, these strategies consider only the "similarity" criterion (Fig. 1). As far as we know, the cross-fusion method may be the first to quantitatively take both "similarity" and "consistency" criteria into account for the selection of the optimal base image pair.

The cross-fusion method improves the accuracy of spatiotemporal data fusion. Our experiments showed that the cross-fusion method performs better than other selection strategies (i.e., *ND*, *HC*, *Diff* and *SI*) in various regions covered by different vegetation types (Fig. 4). Two reasons may explain the better performance of the new method. First, multi-year Landsat image data are employed for the selection of candidate image pairs. Wang *et al*. [25] indicated that spatiotemporal data fusion performs poorly when the base image pair is chosen from a season different from that of the prediction date. Their results are confirmed by our study. We found that using the *ND* strategy had the lowest fusion accuracy (Fig. 4), which may be because in some cases the base and prediction dates were from different seasons. These investigations suggest that it is better choice to choose the base image pairs from multi-year Landsat images instead of using single-year data. Second, the cross-fusion method considers the "consistency" criteria. Our simulation experiments confirmed that simulated inconsistence between Landsat and MODIS images can be well accounted for by the new method (see Table 3 and Fig. 6). This may also explain why the cross-fusion method outperforms any of the five candidate base image pairs in terms of the average *AAD* values (Fig. 5). It is worth noting that using multi-year Landsat data may involve different Landsat sensors (e.g., Landsat 5 or 8), which may lead to different levels of consistency between MODIS and Landsat data [27]. However, it is unnecessary to choose the five candidate image pairs from the same Landsat sensor, because clear Landsat images may be rarely available in some cloudy areas. More importantly, the cross-fusion method is developed to address the issue of inconsistency between the coarse and fine images of the input image pair.

Spatiotemporal data fusion that uses a single base image pair may suffer from some uncertainties. For example, we showed that cloud omission errors have less of an effect with the new method, but fusion accuracy greatly decreases when using a single Landsat base image with cloud omission (Fig. 7). This result is because cloud omission errors in Landsat images can be regarded as a type of inconsistence between Landsat and MODIS data in the local areas, and this inconsistence can also be considered in the process of cross-fusion. For stable performance of spatiotemporal data fusion, it is therefore essential to employ multiple base image pairs. In the cross-fusion method, we determined empirically to use five candidate base image pairs. Here, we further tested the performance of the new method by using different numbers of candidate base image pairs. Results show that the *AAD* values initially decrease with increasing candidate numbers and then vary little when the number is above five for all the four vegetation types (Fig. 10). According to the results of the current experiment, the use of five candidate base image pairs is acceptable, considering the balance between accuracy and computing time.

We determined that a Landsat subset image (12×12 km^2) with cloud contamination below 1% can be used by the new method. To produce an NDVI fusion image for a large area (e.g., one Landsat scene), we recommend running subset images one by one and then combining the fused subset images into a mosaic. In such a way, we can make full use of all the Landsat scenes that are partially cloudy. Because the new method works at the pixel scale, we found no block effect when combining the fused subset images (Fig.9). However, we cannot test the new method in all cases. In case of block effect, the smoothing process based on similarity of pixel can be further applied to the fused image, as suggested by Zhu et al. [24].

B. Uncertainties in the cross-fusion method

We recognize that some issues regarding the application of the cross-fusion method may need to clarify. First, spatiotemporal fusion algorithms are far from perfect and the fusion algorithms will also introduce errors. Different fusion algorithms have different fusion errors. We tested the cross-fusion method by using FSDAF to perform fusion. How about the performance of cross-fusion when using some other spatiotemporal data fusion algorithms? To address the concern, we further tested cross-fusion by using the pioneering algorithm STARFM [16]. We used the same data as in Table 1 and investigated the experiment in Fig. 5 again. Results showed that fusion accuracy was also improved by cross-fusion when using STARFM to perform data fusion (Fig. S4 in the supplementary materials), further suggesting the robustness of the cross-fusion method.

Second, cross-fusion used the five candidate image pairs that have the largest five SI values. Because the total number of image pairs (i.e., *m* in Eq. 2) is different in different regions, the SI values may vary a lot in different regions. Therefore, it is impossible to determine a SI threshold above which the image pairs can be used as candidates. However, one concern may be whether cross-fusion is still effective if the five candidate image pairs are not so similar to the MODIS image at the prediction date (i.e., SI is not so high). To test it, we investigated the experiment in Fig. 5 again but using the five candidates that have the median SI values. More specifically, we sorted all the candidates according to SI and selected the image-pairs with median SI values from the sorted image-pairs sequence. As we expected, compared with the fusion accuracy for the five candidate image pairs with the five largest SI (Fig. 5), the fusion accuracy greatly decreased for the candidates with median SI (see SI_1-SI_5 in Fig. S5 in the supplementary). For example, the AAD value of SI_1 increased from 0.046 (Fig. 5) to about 0.13 for deciduous forest (Fig. S5), suggesting the necessity of considering the "similarity" criterion for the selection of base image pair. Under this scenario (i.e., median SI), however, the cross-fusion method still further improved the fusion accuracy (Fig. S5). This additional investigation highlights that the new method can be applied to the scenario in which all the candidate image pairs are not so similar to the MODIS image at the prediction data.

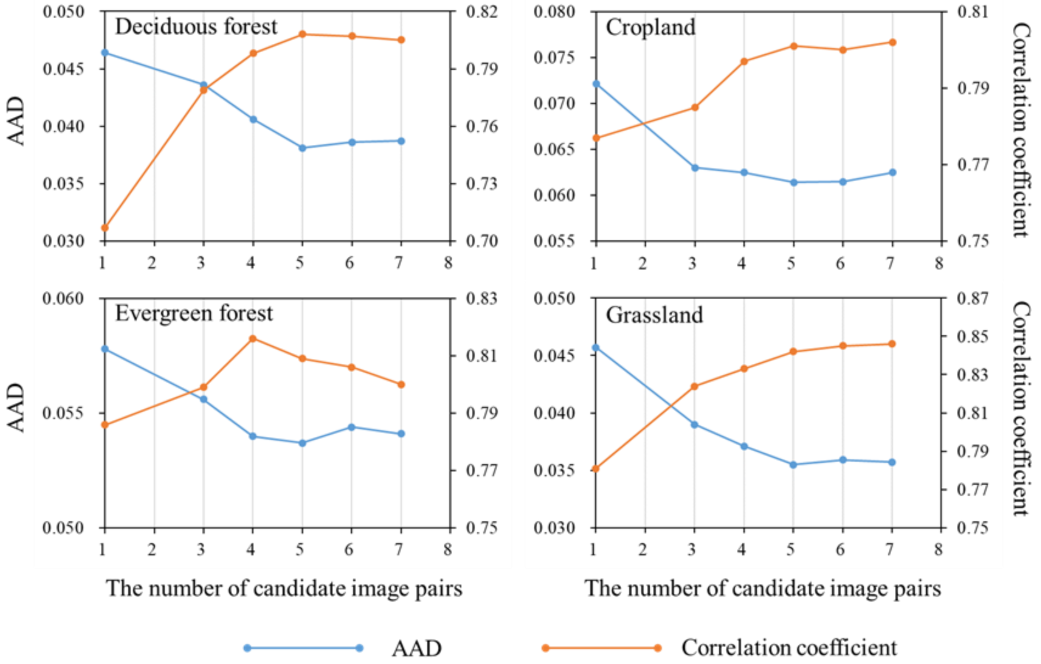


Fig. 10. The performance of the cross-fusion method when using different numbers of candidate base image pairs. Each point in the figure is the average of AAD values or correlation coefficients at all the random prediction dates.

Third, in spatiotemporal data fusion, MODIS images are required to be resampled to match the Landsat spatial resolution. In our experiment, the nearest neighbor resample method was adopted. Here, we further investigated how the different resample methods can affect the fused images. We first used the bilinear interpolation method to resample MODIS images and then performed the experiment in Fig. 5 again. Results showed that the fusion accuracy was comparable between two resample methods, and more importantly, cross-fusion method was still effective when the bilinear interpolation method was used (Fig. S6 in the supplementary materials).

Fourth, the cross-fusion method combined the five fused images that were generated by the five candidate image pairs (according to Eq. 6). We noted that the fusion method IFSDAF also generated the final prediction by the weighted sum of several different fused images [26]. The weights in IFSDAF were estimated by using the difference in MODIS scale NDVI between the base date and the prediction date (see Eq. 10 in [26]). Here, we further performed the experiment in Fig. 5 but using the combination method of IFSDAF (referred to as IFSDAF_combination) to combine the five fused images (i.e., SI_1 - SI_5). Results showed that IFSDAF_combination improved the fusion accuracy by combining the five fused images; however, cross-fusion performed better than IFSDAF_combination (Fig. S7 in the supplementary materials). Theoretically, the IFSDAF_combination method has two obvious limitations. First, because the weights in IFSDAF_combination are completely based on MODIS NDVI differences, IFSDAF_combination cannot address the issue of inconsistency between coarse and fine images (see Fig. S8 in the supplementary materials). Second, there is obvious block effect in the fused image by using IFSDAF_combination because the NDVI difference is calculated at the MODIS scale (see Fig. S9 in the supplementary materials). As a result, a smoothing postprocessing is necessary which may bring a certain uncertainty.

Fifth, in this study the cross-fusion method was tested for MODIS-Landsat NDVI fusion. One may wonder whether cross-fusion can be extended to reflectance images? Theoretically, cross-fusion can work because the fusion error of reflectance can be quantified during the process of cross fusion. It is worth noting that many methods have been developed to reconstruct high-quality MODIS NDVI time-series data, but few efforts have been made for the reconstruction of MODIS reflectance data [37]. As a result, the selected candidate image pairs may be not so similar to the

MODIS reflectance image at the prediction date due to the remaining noise in MODIS reflectance data. Our experiment suggests that under these scenarios (i.e., SI is not so high), the cross-fusion method is still effective (Fig. S5).

Sixth, some limitations in the cross-fusion method should be mentioned. The cross-fusion method is only applicable for one-input-pair fusion algorithms such as STARFM [16], FSDAF [24] and IFSDAF [26]. These one-input-pair fusion algorithms have been widely used and their fusion accuracy has been gradually improved [41]-[42]. Nevertheless, it is also necessary to address the issue of the selection of base image pairs for multi-input-pair fusion algorithms such as ESTARFM [19], which will be considered in our future research. Another limitation is that cross-fusion takes more computing time. For a 108 × 108 km^2 region, it takes about 15 hours to complete the spatiotemporal data fusion on a personal computer (CPU: Inter Core i7-8700). In fact, estimating the regression parameters with the least square method (Eq. 3) is very fast. Thus, the computing time of the cross-fusion method can be greatly reduced by faster spatiotemporal data fusion algorithms. The GPU version of the STARFM algorithm (https://github.com/HPSCIL/cuSTARFM) speeds up the estimation by a factor of 342. We hope that GPU versions of more spatiotemporal data fusion algorithms can be developed, which will popularize the cross-fusion method.

C. *A short summary*

For spatiotemporal data fusion, at least one pair of fine and coarse images at a base date is required for most fusion algorithms to produce the fusion image at the prediction date. It has been recognized that the performances of spatiotemporal data fusion were greatly affected by using different input base image pairs, but the selection of the input image pair was not well addressed by previous studies. We thus developed a new cross-fusion method for the determination of the input image pair. The new method considers both the "similarity" criterion (i.e., the coarse images at the base and prediction dates should be similar) and in particular the "consistency" criterion (the coarse and fine images at the base date should be consistent). We tested the cross-fusion method by using MODIS-Landsat NDVI fusion in the testing regions covered by different vegetation types (deciduous forest, evergreen forests, cropland, and alpine grassland). The experimental results showed that compared with four other selection strategies, the cross-fusion method performed better and achieved smaller fusion error. We simulated scenarios for the inconsistence between MODIS and Landsat data and found the new method successfully quantified the inconsistence even the local inconsistence (e.g., cloud omission in the Landsat image). Furthermore, the fused NDVI time-series data generated by the new method tracked various vegetation growth trajectories better than previous selection strategies. The cross-fusion method provides an effective way to determine the input image pair and improves the practical application of spatiotemporal fusion technology.

ACKNOWLEDGMENT

The authors would like to thank the Landsat team to make the Landsat data freely available.

REFERENCES

[1] J.W. Rouse, Jr. Haas, and R.H. Schell *et al*, "Monitoring vegetation systems in the Great Plains," *NASA Special Publication.*, vol. 1, pp. 309-317, 1974.

[2] R.B. Myneni, S. Hoffman, and Y. Knyazikhin *et al*, "Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data," *Remote Sens. Environ*., vol. 83, no. 1-2, pp. 214-231, Nov. 2002.

[3] R.Y. Cao, M.G. Shen, J. Zhou, and J. Chen, "Modeling vegetation green-up dates across the Tibetan Plateau by including both seasonal and daily temperature and precipitation," *Agric. For. Meteorol*., vol. 249, pp. 176-186, Feb. 2018.

[4] N. Delbart, T. Le Toan, and L. Kergoat *et al*, "Remote sensing of spring phenology in boreal regions: A free of snow-effect method using NOAA-AVHRR and SPOT-VGT data (1982-2004)," *Remote Sens. Environ*., vol. 101, no. 1, pp. 52-62, Mar. 2006.

[5] E. Chuvieco, D. Cocero, and D. Riaño *et al*, "Combining NDVI and surface temperature for the estimation of live fuel moisture content in forest fire danger rating," *Remote Sens. Environ*., vol. 92, no. 3, pp. 322-331, Aug. 2004.

[6] T.R. Loveland, B.C. Reed, and J.F. Brown *et al*, "Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data," *Int. J. Remote Sens*., vol. 21, no. 6-7, pp. 1303-1365, Nov. 2000.

[7] Y. Rao, X. Zhu, and J. Chen *et al*, "An Improved Method for Producing High Spatial-Resolution NDVI Time Series Datasets with Multi-Temporal MODIS NDVI Data and Landsat TM/ETM+ Images," *Remote Sens.*, vol. 7, no. 6, pp. 7865-7891, Jun. 2015.

[8] T. Hwang, C. Song, and P.V. Bolstad *et al*, "Downscaling real-time vegetation dynamics by fusing multi-temporal MODIS and Landsat NDVI in topographically complex terrain," *Remote Sens. Environ*., vol. 115, no. 10, pp. 2499-2512, Oct. 2011.

[9] I.V. Emelyanova, T.R. McVicar, and T.G. Van Niel *et al*, "Assessing the accuracy of blending Landsat–MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sens. Environ*., vol. 133, pp. 193-209, Jun. 2013.

[10] H.K. Zhang, B. Huang, and M. Zhang *et al*, "A generalization of spatial and temporal fusion methods for remotely sensed surface parameters," *Int. J. Remote Sens*., vol. 36, no. 17, pp. 4411-4445, Sep. 2015.

[11] X.L. Zhu, F.Y. Cai, and J.Q. Tian *et al*, "Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions," *Remote Sens*., vol. 10, pp. 527, Mar. 2018.

[12] F. Maselli, and F. Rembold, "Integration of LAC and GAC NDVI data to improve vegetation monitoring in semi-arid environments," *Int. J. Remote Sens*., vol. 23, no. 12, pp. 2475-2488, Nov. 2002.

[13] M.Q. Wu, Z. Niu, C.Y. Wang, C.Y. Wu, and L. W, "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, pp. 63507.1-63507.13, 2012.

[14] B. Zhukov, D. Oertel, and F. Lanzl *et al*, "Unmixing-based multisensor multiresolution image fusion," *IEEE Trans. Geosci. Remote Sens*., vol. 37, no. 3, pp. 1212-1226, May. 1999.

[15] Q. Cheng, H. Liu, and H. Shen *et al*, "A Spatial and Temporal Nonlocal Filter-Based Data Fusion Method," *IEEE Trans. Geosci. Remote Sens*., vol. 55, no. 8, pp. 4476-4488, May. 2017.

[16] F. Gao, J. Masek, and M. Schwaller *et al*, "On the Blending of the Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance," *IEEE Trans. Geosci. Remote Sens*., vol. 44, no. 8, pp. 2207-2218, Aug. 2006.

[17] T. Hilker, M.A. Wulder, and N.C. Coops *et al*, "A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS," *Remote Sens. Environ*., vol. 113, no. 8, pp. 1613-1627, Aug. 2009.

[18] D.P. Roy, J. Ju, and P. Lewis *et al*, "Multi-temporal MODIS-Landsat data fusion for relative radiometric normalization, gap filling, and prediction of Landsat data," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 3112-3130, Jun. 2008.

[19] X.L. Zhu, J. Chen, and F. Gao *et al*, "An enhanced spatial and temporal

adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sens. Environ*., vol. 114, no. 11, pp. 2610-2623, Nov. 2010.

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

[21] A. Li, Y. Bo, and Y. Zhu *et al*, "Blending multi-resolution satellite sea surface temperature (SST) products using Bayesian maximum entropy method," *Remote Sens. Environ*., vol. 135, pp. 52-63, Aug. 2013.

[22] B. Wu, B. Huang, and L. Zhang, "An Error-Bound-Regularized Sparse Coding for Spatiotemporal Reflectance Fusion," *IEEE Trans. Geosci. Remote Sens*., vol. 53, no. 12, pp. 6791-6803, Jul. 2015.

[23] C.M. Gevaert, and F.J. García-Haro, "A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion," *Remote Sens. Environ*., vol. 156, pp. 34-44, Jan. 2015.

[24] X.L. Zhu, E.H. Helmer, and F. Gao *et al*, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sens. Environ*., vol. 172, pp. 165-177, Jan. 2016.

[25] P. Wang, F. Gao, and J.G. Masek, "Operational data fusion framework for building frequent landsat-like imagery," *IEEE Trans. Geosci. Remote Sens*., vol. 52, no. 11, pp. 7353-7365, Nov. 2014.

[26] M. Liu, W. Yang, and X.L. Zhu et al, "An Improved Flexible Spatiotemporal DAta Fusion (IFSDAF) method for producing high spatiotemporal resolution normalized difference vegetation index time series," Remote Sens. Environ., vol. 227, pp. 74-89, Jun. 2019.

[27] D. Xie, F. Gao, L. Sun, and M. Anderson, "Improving Spatial-Temporal Data Fusion by Choosing Optimal Input Image Pairs," *Remote Sens*., vol. 10, no. 7, pp. 1142, Jul. 2018.

[28] J.J. Walker, K.M. De Beurs, R.H. Wynne and F. Gao, "Evaluation of Landsat and MODIS data fusion products for analysis of dryland forest phenology," *Remote Sens. Environ*., vol. 117, pp. 381-393, Feb. 2012.

[29] F. Gao, M.C. Anderson, and X. Zhang *et al*, "Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery," *Remote Sens. Environ*., vol. 188, pp. 9-25, Jan. 2017.

[30] F. Gao, T. He, J.G. Masek, Y. Shuai, C.B. Schaaf and Z. Wang, "Angular Effects and Correction for Medium Resolution Sensors to Support Crop Monitoring," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., Vol. 7, pp. 4480-4489, Aug. 2014.

[31] D.P. Roy, H.K. Zhang, J.C. Ju, J.L. Gomez-Dans, P.E. Lewis, and C.B. Schaaf et al, "A general method to normalize Landsat reflectance data to nadir BRDF adjusted reflectance," Remote Sens. Environ., vol. 176, pp. 255–271, Apr. 2016.

[32] J.G. Masek, E.F. Vermote, and N.E. Saleous *et al*, "A Landsat surface reflectance dataset for North America 1990-2000," *IEEE Geosci. Remote Sens. Lett*., vol. 3, no. 1, pp. 68-72, Jan. 2006.

[33] J. Chen, P. Jönsson, and M. Tamura *et al*, "A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter," *Remote Sens. Environ*., vol. 91, no. 3-4, pp. 332-344, Jun. 2004.

[34] R.Y. Cao, Y. Chen, and M.G. Shen *et al*, "A simple method to improve the quality of NDVI time-series data by integrating spatiotemporal information with the Savitzky-Golay filter," *Remote Sens. Environ*., vol. 217, pp. 244-257, Nov. 2018.

[35] Z. Zhu, and C.E. Woodcock, "Object-based cloud and cloud shadow detection in Landsat imagery," *Remote Sens. Environ*., vol. 118, pp. 83-94, Mar. 2012.

[36] Z. Zhu, S. Wang, and C.E. Woodcock, "Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images," *Remote Sens. Environ*., vol. 159, pp. 269-277, Mar. 2015.

[37] A. Jarihani, T. McVicar, and T. Van Niel *et al*, "Blending Landsat and MODIS Data to Generate Multispectral Indices: A Comparison of "Index-then-Blend" and "Blend-then-Index" Approaches," *Remote Sens*., vol. 6, no. 10, pp. 9213-9238, Sep. 2014.

[38] *Editorial Board of Vegetation Map of China Cas (2001) 1:1000, 000 Vegetation Atlas of China*, Science Press, Beijing, China, 2001, pp. 434

[39] S. Foga, P.L. Scaramuzza, and S. Guo *et al*, "Cloud detection algorithm comparison and validation for operational Landsat data products," *Remote Sens. Environ*., vol. 194, pp. 379-390, Jun. 2017.

[40] X.L. Zhu, and E.H. Helmer, "An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions," *Remote Sens. Environ*., vol. 214, pp. 135-153, Sep. 2018.

[41] Y. Luo, K. Guan, and J. Peng *et al*, "STAIR: A generic and fully-automated method to fuse multiple sources of optical satellite data to generate a high-resolution, daily and cloud-/gap-free surface reflectance product," *Remote Sens. Environ*., vol. 214, pp. 87-99, Sep. 2018.

[42] H. Song, and B. Huang, "Spatiotemporal Satellite Image Fusion through One-Pair Image Learning," IEEE Trans. Geosci. Remote Sens., vol. 51, no. 4, pp. 1883-1896, Oct. 2013.