Remote Sensing of Environment 172 (2016) 165–177

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

A flexible spatiotemporal method for fusing satellite images with different resolutions

Xiaolin Zhu [a,*], Eileen H. Helmer [b], Feng Gao [c], Desheng Liu [d], Jin Chen [e], Michael A. Lefsky [a]

[a] Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, CO, 80523, USA
[b] International Institute of Tropical Forestry, USDA Forest Service, Río Piedras, PR, 00926-0113, USA
[c] Hydrology and Remote Sensing Laboratory, USDA Agricultural Research Service (ARS), Beltsville, MD 20705, USA
[d] Department of Geography, The Ohio State University, Columbus, OH 43210, USA
[e] State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 6 April 2015
Received in revised form 31 August 2015
Accepted 13 November 2015
Available online xxxx

Keywords:
Spatiotemporal
Data fusion
Landsat
MODIS
Reflectance
Time-series

ABSTRACT

Studies of land surface dynamics in heterogeneous landscapes often require remote sensing data with high acquisition frequency and high spatial resolution. However, no single sensor meets this requirement. This study presents a new spatiotemporal data fusion method, the Flexible Spatiotemporal DAta Fusion (FSDAF) method, to generate synthesized frequent high spatial resolution images through blending two types of data, i.e., frequent coarse spatial resolution data, such as that from MODIS, and less frequent high spatial resolution data such as that from Landsat. The proposed method is based on spectral unmixing analysis and a thin plate spline interpolator. Compared with existing spatiotemporal data fusion methods, it has the following strengths: (1) it needs minimum input data; (2) it is suitable for heterogeneous landscapes; and (3) it can predict both gradual change and land cover type change. Simulated data and real satellite images were used to test the performance of the proposed method. Its performance was compared with two very popular methods, the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) and an unmixing-based data fusion (UBDF) method. Results show that the new method creates more accurate fused images and keeps more spatial detail than STARFM and UBDF. More importantly, it closely captures reflectance changes caused by land cover conversions, which is a big issue with current spatiotemporal data fusion methods. Because the proposed method uses simple principles and needs only one fine-resolution image as input, it has the potential to increase the availability of high-resolution time-series data that can support studies of rapid land surface dynamics.

1. Introduction

Dense time-series data composited by satellite images with frequent coverage are important sources for studying land surface dynamics, such as for monitoring vegetation phenology (Shen, Tang, Chen, Zhu, & Zheng, 2011), mapping shrub encroachment into grassland (Zhou, Chen, Chen, Cao, & Zhu, 2013), detecting land cover and land use change (Yang & Lo, 2002), and estimating agriculture intensity (Galford et al., 2008). In heterogeneous areas, these studies also require dense time-series data with high spatial resolution so that land surface dynamics can be characterized at fine scales related to human activities, e.g., from a few meters to tens of meters. However, no single satellite sensor currently provides global coverage of dense time-series data with this fine of a spatial resolution due to the tradeoff between pixel size and swath width as well as cloud contamination (Gevaert & García-Haro, 2015).

For applications requiring imagery from the past several decades, there are two types of satellite images, one with frequent coverage of every 1–2 days, but coarse spatial resolution of 250 m to 1 km, such as imagery from the MODerate resolution Imaging Spectroradiometer (MODIS) images (hereafter referred to as "coarse-resolution" images), and the other with fine spatial resolution of 10–30 m, but a long revisit cycle of ~16 days, such as Landsat images (hereafter, "fine-resolution" images). In the last decade, spatiotemporal data fusion methods have been developed to blend these two types of satellite images to generate synthesized data with both high spatial resolution and frequent coverage (Fu, Chen, Wang, Zhu, & Hilker, 2013; Gao, Masek, Schwaller, & Hall, 2006; Gevaert & García-Haro, 2015; Hilker, Wulder, Coops, Linke, et al., 2009; Huang & Zhang, 2014; Song & Huang, 2013; Wu, Wang, & Wang, 2012; Zhu, Chen, Gao, Chen, & Masek, 2010; Zurita-Milla, Clevers, & Schaepman, 2008). These synthesized data can support the investigation of land surface dynamics in heterogeneous landscapes (Hilker, Wulder, Coops, Seitz, et al., 2009; Senf, Leitão, Pflugmacher, van der

* Corresponding author at: Department of Ecosystem Science and Sustainability, Colorado State University, NESB 108, 1499 Campus Delivery, Fort Collins, CO 80523-1499, USA.
E-mail address: zhuxiaolin55@gmail.com (X. Zhu).

http://dx.doi.org/10.1016/j.rse.2015.11.016
0034-4257/

Linden, & Hostert, 2015; Walker, de Beurs, Wynne, & Gao, 2012; Watts, Powell, Lawrence, & Hilker, 2011).

Existing spatiotemporal data fusion methods can be categorized into three groups: weighted function based, unmixing based, and dictionary-pair learning based (Table 1). All of these methods need one or more observed pairs of coarse- and fine-resolution images for training and a coarse-resolution image at prediction date as input data. The output of these methods is a synthetic fine-resolution image at prediction date. Intrinsically, all spatiotemporal data fusion methods use spatial information from the input fine-resolution images and temporal information from the coarse-resolution images.

Among the weighted function based methods, the spatial and temporal adaptive reflectance fusion model (STARFM) is the one developed first (Gao et al., 2006). STARFM assumes that changes of reflectance are consistent and comparable at coarse and fine resolutions if pixels in coarse-resolution images (hereafter referred to as "coarse pixels") are "pure" pixels, in that one coarse pixel only includes one land cover type. In this case, changes derived from coarse pixels can be directly added to pixels in fine-resolution images (hereafter referred to as "fine pixels") to get the prediction. However, this ideal situation cannot be satisfied when coarse pixels are mixed, having a mixture of different land cover types. Therefore, STARFM predicts pixels with a function that gives a higher weight to purer coarse pixels based on information from neighboring fine pixels. STARFM was later modified and improved for more complex situations, resulting in the spatial temporal adaptive algorithm for mapping reflectance change (STAARCH), which improves STARFM's performance when land cover type change and disturbance exist (Hilker, Wulder, Coops, Linke, et al., 2009), and the Enhanced STARFM (ESTARFM), which improves STARFM's accuracy in heterogeneous areas (Zhu et al., 2010).

Among the unmixing based methods, the multisensor multiresolution technique (MMT) proposed by Zhukov et al. (1999) is perhaps the first one to fuse images acquired at different times and with different resolutions. MMT has four steps to predict a fine-resolution image: (1) classify the input fine-resolution data to define endmembers at coarse resolution; (2) compute endmember fractions of each coarse pixel; (3) unmix the coarse pixels at the prediction date within a moving window; (4) assign unmixed reflectance to fine pixels (Zhukov et al., 1999). In recent years, MMT has been modified by several studies to improve its accuracy. Zurita-Milla et al. (2008) introduced constraints into the linear unmixing process to ensure that the solved reflectance values were positive and within an appropriate range. Wu et al. (2012) estimated reflectance change through unmixing endmember reflectance at both input and prediction date and then added the estimated change back to the base fine-resolution image to get the prediction. Amorós-López et al. (2013) modified the cost function to prevent the solved endmember reflectance from being greatly different from a predefined endmember reflectance. Gevaert and García-Haro (2015) directly unmixed the change of coarse pixels to estimate the change of endmembers and applied Bayesian theory to constrain the estimation.

Compared with weighted function and unmixing based methods, dictionary-pair learning based spatiotemporal data fusion methods are relatively new. Dictionary-pair learning based algorithms establish correspondences between fine- and coarse-resolution images based on their structural similarity, which can be used to capture the main features, including land cover type changes, in the predictions. The Sparse-representation-based SpatioTemporal reflectance Fusion Model (SPSTFM) is perhaps the first to bring dictionary-pair learning techniques from natural image superresolution to spatiotemporal data fusion (Huang & Song, 2012). SPSTFM establishes a correspondence between the change of two fine-resolution images and two coarse-resolution images through dictionary-pair learning, and then the trained dictionary is applied to predict a high-resolution image at the prediction date. Following SPSTFM, Song and Huang (2013) developed another dictionary-pair learning based fusion method which uses only one pair of fine- and coarse-resolution images. This method trains a dictionary pair on the input fine- and coarse-resolution image pair, and then downscales the coarse-resolution image at the prediction date by a sparse coding technique. Due to the large scale difference between MODIS and Landsat, this method is implemented in a two-layer framework, i.e., it first predicts an image with a middle-resolution between the fine- and coarse-resolution and then predicts the fine-resolution image based on the middle-resolution image (Song & Huang, 2013).

Studies have demonstrated that all of the above spatiotemporal data fusion methods from the three groups can improve the spatial and temporal resolution of satellite images for specific applications (Gao et al., 2006; Song & Huang, 2013; Zhu et al., 2010; Zurita-Milla et al., 2008). However, they face challenges in heterogeneous regions with abrupt land cover type changes. Most weighted function based methods assume no land cover type changes between input and prediction date (Fu et al., 2013; Gao et al., 2006; Weng, Fu, & Gao, 2014; Zhu et al., 2010). As a result, they can successfully predict pixels with changes in attributes like vegetation phenology or soil moisture, because these changes are strongly related to the changes in similar pixels selected from the input imagery. However, current methods are not effective for predicting spectral changes that are sudden or not observed in input imagery, in that the changes are not predictable from pixels that were similar in the input date. These changes include urbanization, deforestation/reforestation, wildfires, floods and land cover transitions caused by other forces. Song and Huang (2013) applied STARFM in an urbanized area and found that it failed to recover pixels with land cover type changes. Emelyanova, McVicar, Van Niel, Li, and van Dijk (2013) conducted a comprehensive study to investigate the performance of STARFM and ESTARFM in two landscapes with contrasting spatial and temporal dynamics. Their results demonstrate that the performance of data fusion methods is strongly associated with land cover spatial and temporal variance. ESTARFM is better than STARFM in heterogeneous landscapes, but it is even worse than STARFM for predicting abrupt changes in land cover types (Emelyanova et al., 2013). STAARCH can handle land cover changes or disturbances if they are predictable

Table 1
Summary of main spatiotemporal data fusion methods: W = weighted function based, U = unmixing based, and D = dictionary-pair learning based.

Name of method	Category	Input requirement*	Reference
STARFM	W	One or more pairs	Gao et al. (2006)
STAARCH	W	Two fine images and a time-series of coarse images	Hilker et al. (2009a)
ESTARFM	W	Two pairs	Zhu et al. (2010)
MMT	U	One fine image	Zhukov, Oertel, Lanzl, and Reinhäckel (1999)
Constrained unmixing	U	One fine image	Zurita-Milla et al. (2008)
STDFA	U	Two or more pairs	Wu et al. (2012)
Spatial unmixing	U	A time-series of fine images	Amorós-López et al. (2013)
STRUM	U	One pair	Gevaert and García-Haro (2015)
SPSTFM	D	Two pairs	Huang and Song (2012)
One-pair learning	D	One pair	Song and Huang (2013)

* All methods need one coarse image at prediction date as input, so the only other required input data are listed in the table. "One pair" means one image with finer spatial resolution and one image with coarser spatial resolution acquired on the same or nearly the same date as the finer resolution image.

from one of the input Landsat images (i.e. recorded as spectrally different from surrounding areas due to earlier disturbance, for example), but it requires two Landsat images, one from before and one from after the change (Hilker, Wulder, Coops, Linke, et al., 2009). Unmixing based fusion methods also require that no land cover type change occurs between the input and prediction dates (Amorós-López et al., 2013; Wu et al., 2012; Zhukov et al., 1999; Zurita-Milla et al., 2008). In addition, to capture the spatial variability of each class, these unmixing based methods solve fine-resolution reflectance locally using a moving window which may produce unrealistic results because of co-linearity problem and noises contained in both fine- and coarse-resolution data (Gevaert & García-Haro, 2015). Dictionary-pair learning methods only use statistical relationships between fine- and coarse-resolution images rather than any physical properties of remote sensing signals (Huang & Song, 2012; Song & Huang, 2013). Although they can better predict pixels with land cover type changes, they do not accurately maintain the shape of objects, especially when the scale difference between fine and coarse-resolution images is large (see Fig. 3 in Song & Huang, 2013).

Besides the above-mentioned intrinsic problems of the methods in each group, most spatiotemporal data fusion methods suffer another limitation: they require two or more fine-resolution images as input data (Table 1). For example, both ESTARFM and STAARCH (as mentioned above) need at least two pairs of fine and coarse-resolution images to improve the performance of STARFM (Hilker, Wulder, Coops, Linke, et al., 2009; Zhu et al., 2010). Huang and Zhang (2014) also need two Landsat images to improve the performance of unmixing-based method in a situation where land cover type has changed. However, in many regions, it is not easy to collect two pairs of available images within a reasonable period (e.g., a season or a year) because of cloud contamination, sub-optimal acquisition schedule, data archive access restrictions or other reasons (Ju & Roy, 2008; Senf et al., 2015). The scan-line corrector failure on Landsat 7 in May of 2003 led to regular data gaps in all scenes collected since then. In addition, Landsat 5 data are sparse in regions with no local receiver, and these places include many of the places with persistent cloud cover. This problem is more serious when these spatiotemporal data fusion methods are applied to regions with rapid seasonal changes or frequent disturbances, such as cropland (Watts et al., 2011). The data record is poor in many ecologically important areas on the planet (Yu, Shi, & Gong, 2015), and we sought a method that would require only one cloud-free observation at the spatial resolution of Landsat.

To overcome the above-mentioned limitation of current spatiotemporal data fusion methods, we propose a Flexible Spatiotemporal DAta Fusion (FSDAF) model in this paper. Its goal is to more accurately predict fine-resolution images in heterogeneous areas by capturing both gradual and abrupt land cover type changes and it requires minimal input data. In particular, it requires only one image with fine spatial resolution. We test FSDAF with simulated images and real Landsat images and then compare it with other methods that require only one fine resolution image as input: we compare it with STARFM and an unmixing based method. In the rest of the paper, we will introduce steps of FSDAF in Section 2, describe the test experiments and results in Sections 3 and 4, and discuss the strengths and limitations of FSDAF in the last section.

2. Methodology

2.1. Notations and definitions

Before describing the details of FSDAF, some notations and definitions are given here for convenience.

m the number of fine pixels (also named as subpixels) within one coarse pixel;
(x_i, y_i) coordinate index of the ith pixel;
i index of a coarse pixel;
j index of a fine pixel within one coarse pixel, $j = 1,\ldots, m$;
$C_1(x_i, y_i, b)$ and $C_2(x_i, y_i, b)$ band b value of coarse pixel (e.g., MODIS) at location (x_i, y_i) observed at t_1 and t_2 respectively;
$F_1(x_{ij}, y_{ij}, b)$ and $F_2(x_{ij}, y_{ij}, b)$ band b value of the jth fine pixel (e.g., Landsat) within the coarse pixel at location (x_i, y_i) observed at t_1 and t_2 respectively;
$f_c(x_i, y_i)$ the fraction of class c of the (x_i, y_i) coarse pixel;
$\Delta C(x_i, y_i, b)$ change of band b value of the (x_i, y_i) coarse pixel between t_1 and t_2;
$\Delta F(c, b)$ change of band b value of class c at fine resolution between t_1 and t_2.

2.2. FSDAF

In FSDAF, the input data include one pair of coarse- and fine-resolution images acquired at t_1 and one coarse-resolution image at t_2 (Fig. 1). The output is a predicted fine-resolution image at t_2. Before the implementation of FSDAF, both coarse- and fine-resolution images should be calibrated to the same physical quantity, such as top-of-atmosphere reflectance or surface reflectance, and they need to be co-registered. The co-registration process can be done in the following steps: re-projecting coordinates of coarse-resolution image to that of the fine-resolution image if they are different, resampling coarse-resolution image to fine resolution by nearest neighbor algorithm, geo-referencing one image to another one by selecting control points or maximizing correlation between the two images, and then cropping them to cover the same area (Emelyanova et al., 2013; Gevaert & García-Haro, 2015). To further reduce the difference between coarse- and fine-resolution data caused by sensor configuration and processing chains, a radiometric normalization can be applied by assuming a linear relationship between both data sets (Gao, Masek, Wolfe, & Huang, 2010; Gevaert & García-Haro, 2015). FSDAF includes six main steps: (1) classify the fine-resolution image at t_1; (2) estimate the temporal change of each class in the coarse-resolution image from t_1 to t_2; (3) predict the fine-resolution image at t_2 using the class-level temporal change and calculate residuals at each coarse pixel; (4) predict the fine-resolution image from the coarse image at t_2 with a Thin Plate Spline (TPS) interpolator; (5) distribute residuals based on TPS prediction; and (6) get the final prediction of the fine-resolution image using information in neighborhood. Detailed descriptions of each step in FSDAF are given below.

To intuitively illustrate procedures of the proposed FSDAF method, we use simulated images to show the outputs of intermediate steps. The two pairs of simulated images include both gradual change (e.g., phenology) and land cover type change. Fig. 2 shows the two pairs of simulated images with only one band. Fig. 2(a) and (b) are two Landsat-like images at t_1 and t_2 (30 m resolution, size 480 × 480 pixels), while Fig. 2(c) and (d) are their corresponding MODIS-like images (480 m resolution) which were aggregated from Fig. 2(a) and

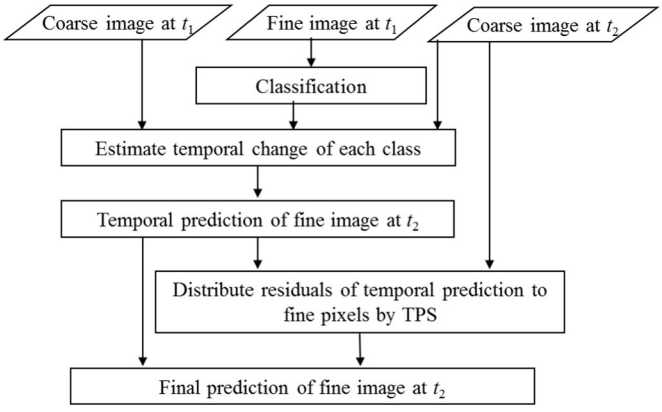

Fig. 1. Flowchart of the proposed spatiotemporal data fusion method.

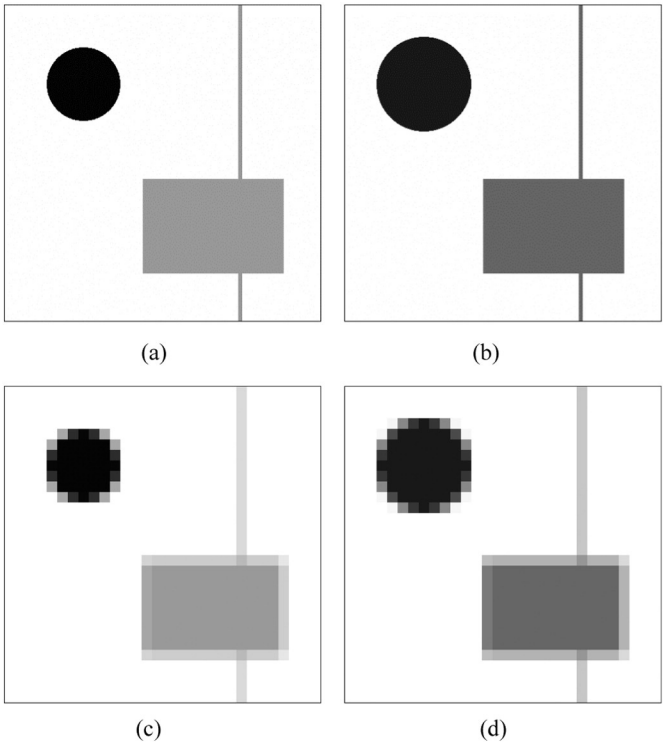

Fig. 2. Simulated Landsat-like images at t_1 (a) and t_2 (b) and their corresponding MODIS-like images (c) and (d).

(b) by averaging values of all fine pixels inside a coarse pixel. There are three objects simulated: a circle, a rectangle, and a line. From t_1 to t_2, the circle increases its radius from 56 pixels to 72 pixels, and its reflectance value increases from 0.01 to 0.05. The rectangle and the line remain the same size but change their values from 0.3 to 0.2. The background has a constant value of 0.5. Random noise of less than ± 0.001 was also added to the simulated Landsat-like images.

2.2.1. Classify fine-resolution image at t_1

To get the fraction of each class within one coarse pixel, the fine-resolution image at t_1 is classified by either supervised or unsupervised algorithms using all image bands. The selection of a classification algorithm depends on the specific application and data availability. If the ground reference data are available, supervised classifiers such as support vector machines and maximum likelihood classifiers can be used. Otherwise, unsupervised classifiers can automatically divide the fine-resolution image into several spectral classes. In this study, an unsupervised classifier, i.e., ISODATA, is used to classify the input fine-resolution image to make FSDAF automatic. Users need to set minimum and maximum number of classes in ISODATA, which can be determined by users' prior knowledge of the study area or visual inspection of the input fine-resolution image. ISODATA outputs optimal classification results through merging and splitting classes according to the distribution of pixel values in feature space (Ball & Hall, 1965). The simulated fine image at t_1 (Fig. 2(a)) was finally classified into three spectral classes: the circle object is class 1, the background is class 2, and the rectangle and the line are class 3.

After classification of the fine-resolution image at t_1, we can calculate the class fractions within a coarse pixel through counting the number of fine pixels of each class:

$$f_c(x_i, y_i) = N_c(x_i, y_i)/m\,, \tag{1}$$

where N_c (x_i, y_i) is the number of fine pixels belonging to class c within the coarse pixel at (x_i, y_i).

2.2.2. Estimate the temporal change of each class

For band b, the temporal change of the coarse pixel at (x_i, y_i) is:

$$\Delta C(x_i, y_i, b) = C_2(x_i, y_i, b) - C_1(x_i, y_i, b)\,. \tag{2}$$

According to spectral linear mixing theory, the temporal change of a coarse pixel is the weighted sum of the temporal change of all classes within it:

$$\Delta C(x_i, y_i, b) = \sum_{c=1}^{l} f_c(x_i, y_i) \times \Delta F(c, b). \tag{3}$$

where l is the number of classes. Eq. (3) is valid only when no land cover type change happens between t_1 and t_2. Theoretically, in order to solve for $\Delta F(c, b)$, $c = 1,\ldots, l$, we need at least l equations. Assuming that the temporal change of each class is the same among all coarse pixels, we can select n $(n > l)$ coarse pixels to compose a system of linear mixture equations:

$$\begin{bmatrix} \Delta C(x_1, y_1, b) \\ \vdots \\ \Delta C(x_i, y_i, b) \\ \vdots \\ \Delta C(x_n, y_n, b) \end{bmatrix} = \begin{bmatrix} f_1(x_1, y_1) & f_2(x_1, y_1) & \cdots & f_l(x_1, y_1) \\ \vdots & \vdots & & \vdots \\ f_1(x_i, y_i) & f_2(x_i, y_i) & \cdots & f_l(x_i, y_i) \\ \vdots & \vdots & & \vdots \\ f_1(x_n, y_n) & f_2(x_n, y_n) & \cdots & f_l(x_n, y_n) \end{bmatrix} \begin{bmatrix} \Delta F(1, b) \\ \vdots \\ \Delta F(c, b) \\ \vdots \\ \Delta F(l, b) \end{bmatrix} \tag{4}$$

$\Delta F(c, b)$, $c = 1,\ldots, l$, can be solved through the inversion of Eq. (4) by computing a least squares best fit solution. However, there are two factors which will affect the accuracy of inversion: collinearity and land cover type change. First, the collinearity problem happens when the fractions of one class in these selected coarse pixels have a linear relationship with fractions of any other classes. To avoid this situation, for each class k coarse pixels are selected that have the highest fraction of a given class, i.e., k purest coarse pixels of that class. Second, for these k purest coarse pixels, if some of them have land cover type change, the temporal change ΔC of these coarse pixels would be outliers assuming that land cover type change happens in a relatively small portion of the whole image and pixels with the largest changes are relatively rare. Accordingly, of the k purest coarse pixels of each class, the ones with ΔC outside of the range of 0.1–0.9 quantiles (or a narrower range, e.g., 0.25–0.75, if land cover type change is large through inspecting the two coarse-resolution images) are excluded. After the above two-step selection, a total of n coarse pixels are used to compose Eq. (4). For the simulated images in Fig. (2), the solved change ΔF of the three classes are 0.03998, −0.00002, and −0.09998, while the true change values are 0.04, 0.00, −0.10 respectively, indicating that the change values at fine resolution have been accurately estimated from the linear equation system.

2.2.3. Predict fine-resolution image and residuals from temporal changes

The temporal change of each class can be assigned to relevant fine pixels without considering the within-class variability. If land cover types do not change between t_1 and t_2, adding the temporal change to values of fine pixels observed at t_1 can obtain the prediction of values of fine pixels at t_2:

$$F_2^{TP}\left(x_{ij}, y_{ij}, b\right) = F_1\left(x_{ij}, y_{ij}, b\right) + \Delta F(c, b) \text{ if } \left(x_{ij}, y_{ij}\right) \text{ belongs to class } c. \tag{5}$$

where $F_2^{TP}(x_{ij}, y_{ij}, b)$ is referred to as the temporal prediction because it only uses the temporal change information between input and prediction dates rather than any spatial information, such as spatial

dependence. Fig. 3(a) shows the temporal production of simulated Landsat-like image at t_2. Compared with the true image in Fig. 2(b), we can see that temporal prediction has accurately estimated the value of pixels without land cover type change but it failed to predict the expanded circle object.

For each coarse pixel, its value is equal to the sum of values of all fine pixels inside it and a bias factor ξ which is the system difference between two sensors caused by differences in bandwidth and solar geometry (Gao et al., 2006). This system difference can be considered constant between t_1 and t_2, so the values of coarse pixels at t_1 and t_2 can be written as:

$$C_1(x_i, y_i, b) = \frac{1}{m}\sum_{j=1}^{m} F_1\left(x_{ij}, y_{ij}, b\right) + \xi \quad , \tag{6}$$

$$C_2(x_i, y_i, b) = \frac{1}{m}\sum_{j=1}^{m} F_2\left(x_{ij}, y_{ij}, b\right) + \xi \quad . \tag{7}$$

We already have a temporal prediction of the fine-resolution image at t_2, but it is not a very accurate prediction where land cover type change has occurred and large within-class variability exists. We introduce a residual term R between the true values and temporal prediction of fine pixels:

$$\frac{1}{m}\sum_{j=1}^{m} F_2\left(x_{ij}, y_{ij}, b\right) = \frac{1}{m}\sum_{j=1}^{m} F_2^{TP}\left(x_{ij}, y_{ij}, b\right) + R(x_i, y_i, b) \, . \tag{8}$$

From Eqs. (6)–(8), we can derive:

$$R(x_i, y_i, b) = \Delta C(x_i, y_i, b) - \frac{1}{m}\left[\sum_{j=1}^{m} F_2^{TP}\left(x_{ij}, y_{ij}, b\right) - \sum_{j=1}^{m} F_1\left(x_{ij}, y_{ij}, b\right)\right] \, . \tag{9}$$

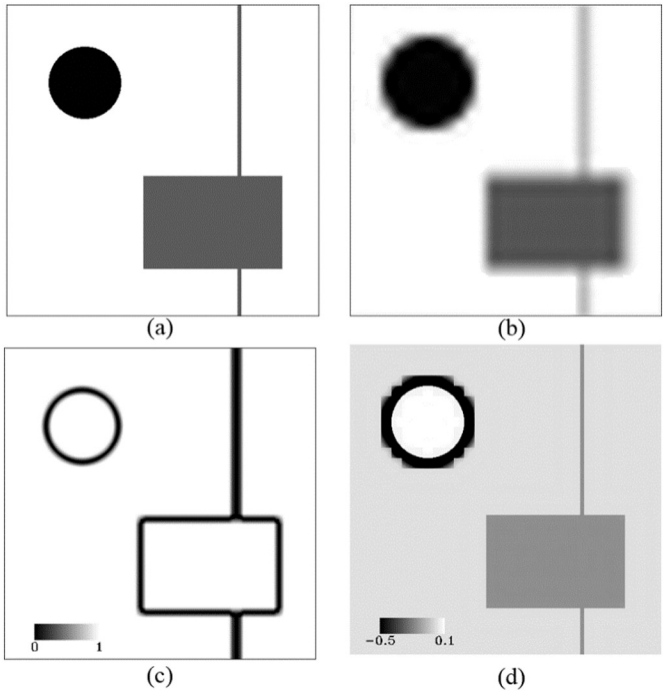

Fig. 3. Temporal prediction (Eq. (5)) (a) and spatial prediction (Eq. (13)) (b) of simulated Landsat-like image at t_2, the homogeneity index image (Eq. (16)) (c) and estimation of total change (Eq. (20)) (d).

From Eq. (8), we can see that distributing residual $R(x_i, y_i, b)$ to fine pixels within a coarse pixel is a key step to improve the accuracy of temporal prediction of fine pixel value at t_2.

2.2.4. Get TPS interpolation for guiding residual distribution

As described above, residuals of the temporal prediction mainly come from land cover type change and within-class variability. However, since the fine-resolution image is unknown at t_2, all true information about land cover type change and within-class variability is contained in the coarse-resolution image at t_2. Therefore, downscaling the coarse-resolution image at t_2 to fine resolution can get another prediction of the fine-resolution image at t_2, which further helps to distribute the residuals from the temporal prediction. Since this prediction only uses spatial dependence among the coarse pixels at t_2, rather than any information at t_1, we refer to this prediction as the spatial prediction.

In this method, we adopt a thin plate spline (TPS) method to downscale the coarse-resolution image at t_2. TPS is a spatial interpolation technique for point data based on spatial dependence (Dubrule, 1984). The value of each coarse pixel is attributed to the location at the center to get a regular point data set. TPS first fits a spatial dependent function using known point data through minimizing an energy function. Given N known points, the basic TPS function for band b is defined as:

$$f_{TPS-b}(x, y) = a_0 + a_1 x + a_2 y + \frac{1}{2}\sum_{i=1}^{N} b_i r_i^2 \, \log r_i^2, \tag{10}$$

with the constraints:

$$\sum_{i=1}^{N} b_i = \sum_{i=1}^{N} b_i x_i = \sum_{i=1}^{N} b_i y_i = 0 \quad , \tag{11}$$

where $r_i^2 = (x - x_i)^2 + (y - y_i)^2$. The coefficients in Eq. (10) are optimized by minimizing:

$$E_{TPS-b} = \sum_{i=1}^{N} \left\| C_2(x_i, y_i, b) - f_{TPS-b}(x_i, y_i) \right\|^2 \quad . \tag{12}$$

After optimizing the parameters in the TPS function, it is then used to predict the values of each fine pixel:

$$F_2^{SP}\left(x_{ij}, y_{ij}, b\right) = f_{TPS-b}\left(x_{ij}, y_{ij}\right) \quad . \tag{13}$$

Since TPS prediction only uses spatial dependence of the coarse pixels, it produces a smooth result. In other words, TPS prediction captures the spatial patterns shown in the coarse image but cannot retrieve all spatial details. The strength of TPS prediction is that it maintains the land cover type change signals and local variability in the result. This is the very limitation of the temporal prediction. Compared with the temporal prediction (Fig. 3(a)), we can see that spatial prediction of the simulated Landsat-like image at t_2 (Fig. 3(b)) better captures the expanded circle object, but it has larger errors in the small or narrow objects (i.e., the line in the simulated image) and boundaries between two classes.

2.2.5. Distribute residuals to fine pixels

As mentioned before, the distribution of the residuals from the temporal prediction to individual fine pixels inside each coarse pixel is the key step to improving the accuracy of the temporal prediction. Existing downscaling approaches distribute residuals to subpixels equally (Chen, Li, Chen, Rao, & Yamaguchi, 2014), or weighted by the initial estimate of each subpixel (Liu & Zhu, 2012). These simple strategies ensure that the re-aggregated fused fine-resolution image exactly matches the original coarse-resolution image, but they may not give help to improve

accuracy of individual subpixels, because they do not consider the real sources of where the residuals come from. Errors of temporal prediction (Eq. 9) are mainly caused by land cover type change and within-class variability across the image. Therefore, this study designs a new weighted function to distribute more residuals to the subpixels with larger errors.

In the case of the homogenous landscape, we can assume that the TPS spatial prediction best represents true values of the fine pixels at t_2, and the error of the temporal prediction can be estimated as:

$$E_{ho}\left(x_{ij}, y_{ij}, b\right) = F_2^{SP}\left(x_{ij}, y_{ij}, b\right) - F_2^{TP}\left(x_{ij}, y_{ij}, b\right) . \quad (14)$$

However, the error estimated from Eq. (14) is not valid for fine pixels in heterogeneous landscapes, or at edges between two land cover types, because TPS prediction smoothes these edges in space. Where the landscape is heterogeneous, or at land cover edges, assuming that all fine pixels within a coarse pixel with equal error is reasonable if we have no other information available:

$$E_{he}\left(x_{ij}, y_{ij}, b\right) = R(x_i, y_i, b) . \quad (15)$$

To integrate the two cases into one weighted function to guide the residual distribution, here we introduce a homogeneity index:

$$HI\left(x_{ij}, y_{ij}\right) = \left(\sum_{k=1}^{m} I_k\right) \Big/ m , \quad (16)$$

where $I_k = 1$ when the kth fine pixels within a moving window (its size is one coarse pixel) with the same land cover type as the central fine pixel (x_{ij}, y_{ij}) being considered, otherwise $I_k = 0$. HI ranges from 0 to 1, and larger values indicate a more homogenous landscape (see Fig. 3(c)). The weight for combining the two cases through HI is:

$$CW\left(x_{ij}, y_{ij}, b\right) = E_{ho}\left(x_{ij}, y_{ij}, b\right) \times HI\left(x_{ij}, y_{ij}\right) + E_{he}\left(x_{ij}, y_{ij}, b\right) \times \left[1 - HI\left(x_{ij}, y_{ij}\right)\right] \quad (17)$$

The weight is then normalized as:

$$W\left(x_{ij}, y_{ij}, b\right) = CW\left(x_{ij}, y_{ij}, b\right) \Big/ \sum_{j=1}^{m} CW\left(x_{ij}, y_{ij}, b\right) . \quad (18)$$

Then, the residual distributed to jth fine pixel is:

$$r\left(x_{ij}, y_{ij}, b\right) = m \times R(x_i, y_i, b) \times W\left(x_{ij}, y_{ij}, b\right) . \quad (19)$$

Summing the distributed residual and the temporal change, we can obtain the prediction of the total change of a fine pixel between t_1 and t_2 (see Fig. 3(d)):

$$\Delta F\left(x_{ij}, y_{ij}, b\right) = r\left(x_{ij}, y_{ij}, b\right) + \Delta F(c, b) \text{ if } \left(x_{ij}, y_{ij}\right) \text{ belongs to class } c. \quad (20)$$

2.2.6. Obtain a robust prediction of fine image using neighborhood

Theoretically, adding the total change term obtained in Eq. (20) to the value of fine pixel at t_1 can get the final prediction at t_2. However, this prediction is on a pixel-by-pixel basis, which inevitably has many uncertainties caused by errors in previous steps and noise contained in all input images. In addition, the distribution of residuals is implemented within each coarse pixel, which leads to block effects as shown in Fig. 3(d). STARFM and ESTARFM both use additional neighborhood information to reduce the uncertainties in final predictions and while mitigating block effects (Gao et al., 2006; Zhu et al., 2010). In this study, we employ a similar strategy as STARFM and ESTARFM to get a more robust prediction of fine pixel values at t_2. First, in the fine image at t_1, for a target fine pixel (x_{ij}, y_{ij}), we select n fine pixels (named as similar pixels including the target pixel itself) of the same class and with the smallest spectral difference from the target fine pixel within its neighborhood (Fig. 4). The spectral difference between kth fine pixel and the target pixel is defined as:

$$S_k = \sum_{b=1}^{B} \left[\left| F_1(x_k, y_k, b) - F_1(x_{ij}, y_{ij}, b) \right| / F_1\left(x_{ij}, y_{ij}, b\right) \right] . \quad (21)$$

For the number of similar pixels n, there is a tradeoff between accuracy of the fused results and computing time. Through trial-and-error experiments, we found that the accuracy of fused results will be stable with $n > 20$, so we recommend selecting 20 similar pixels in practice.

Second, the weight of each similar pixel is determined by the spatial distance between similar pixels and the target pixel. The spatial distance of kth similar pixel D_k is a relative distance defined in ESTARFM (Zhu et al., 2010):

$$D_k = 1 + \sqrt{\left(x_k - x_{ij}\right)^2 + \left(y_k - y_{ij}\right)^2} \Big/ (w/2) , \quad (22)$$

where w is the size of neighborhood, which is determined by the homogeneity of the study area and commonly the size of one to three coarse pixels. Using a larger size in more heterogonous areas ensures that enough similar pixels are selected (Zhu et al., 2010). D_k is a relative distance ranging from 1 to $1+\sqrt{2}$. Assuming that similar pixels that are further away contribute less to estimate target pixel, the weight for the kth similar pixel is calculated as:

$$w_k = (1/D_k) \Big/ \sum_{k=1}^{n} (1/D_k) . \quad (23)$$

Change information of all similar pixels is summed by weight to get the total change value of the target pixel. Adding this final estimate of total change to the initial observation at t_1 yields the final prediction of the target pixel value at t_2:

$$\hat{F}_2\left(x_{ij}, y_{ij}, b\right) = F_1\left(x_{ij}, y_{ij}, b\right) + \sum_{k-1}^{n} w_k \times \Delta F(x_k, y_k, b) . \quad (24)$$

3. Testing experiment

3.1. Study area and data

FSDAF was tested by both simulated data in Fig. 2 and real satellite images. For the real satellite images, fine-resolution images are Landsat images, while corresponding coarse-resolution images are simulated MODIS-like images aggregated from the original Landsat images. In this study, we used simulated MODIS-like images rather than real MODIS images for algorithm tests because the accuracy of spatiotemporal data fusion algorithms are affected by the radiometric and geometric

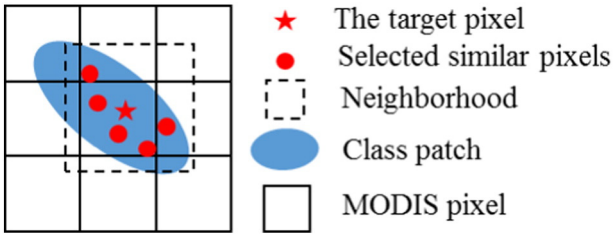

Fig. 4. Diagram of similar pixels selected in a neighborhood of the target pixel.

inconsistencies between two sensors (Gevaert & García-Haro, 2015). Using simulated coarse-resolution images can eliminate the interference of these confounding factors so that we can directly compare the performance of different methods given that the difference in accuracy is only caused by different methods themselves. This strategy was used in recent studies to assess the performance of spatiotemporal data fusion methods (Gevaert & García-Haro, 2015; Wu et al., 2012). Application of the proposed method to real coarse-resolution images and assessment of the influence of radiometric and geometric inconsistencies on the accuracy are beyond the objectives of this study, but they will be explored in our future studies.

Landsat images were provided by Emelyanova et al. (2013) and have been atmospherically corrected. These Landsat images cover two study sites with contrasting spatial and temporal dynamics, i.e., one with heterogeneous landscape and another with land cover type change. These two sites have been used to evaluate different spatiotemporal data fusion methods, including STARFM and ESTARFM, in a previous study (Emelyanova et al., 2013), and the authors make these data sets freely available to remote sensing community. These data sets can be used as benchmark for comparing or testing spatiotemporal data fusion methods. For example, these data were used to test a newly developed method for blending MODIS NDVI time-series and Landsat images (Rao, Zhu, Chen, & Wang, 2015).

In the first site with heterogeneous landscape, we used two cloud-free Landsat 7 ETM+ images (excluding the 6th thermal band and 8th panchromatic band) covering an area of 20 km × 20 km in southern New South Wales, Australia (145.0675°E, 34.0034°S). The false color composite of Landsat images and their corresponding aggregated MODIS-like images were shown in Fig. 5. The two Landsat images (Path/Row 93/84) were acquired on November 25, 2001 (Fig. 5(a)) and January 12, 2002 (Fig. 5(b)) respectively. The major land cover types in this area are irrigated rice cropland, dryland agriculture, and woodlands. Rice croplands are often irrigated in October–November (Emelyanova et al., 2013). In Fig. 5, we can see that there are many small parcels of cropland in this site. These cropland parcels range in size from one to several MODIS pixels, have irregular shapes and are spatially scattered. Comparing the two Landsat images, it is clear that irrigated cropland (i.e., darker pixels in Fig. 5(a)) has larger reflectance changes than surrounding dryland agriculture or woodlands. In this experiment, we used the pair of images on November 25, 2001 (Fig. 5(a) and (c)) and the MODIS-like image on January 12, 2002 (Fig. 5(c)) to predict the Landsat image in Fig. 5(b). In the process of FSDAF, Fig. 5(a) was classified into three spectral classes by ISODATA method.

The second site with land cover type change is located in northern New South Wales, Australia (149.2815°E, 29.0855°S). This site covers an area of 20 km × 20 km and is relatively homogenous, with large parcels of croplands and natural vegetation (Fig. 6). Two Landsat images were acquired on November 26, 2004 and December 12, 2004 (Path/Row 91/80). A large flood occurred in December 2004. From the Landsat image of December 12, 2004 (Fig. 6(b)), we can see a large inundated area. The flood event caused land cover type change to water in some pixels from Fig. 6(a) to (b). In this experiment, the pair of Landsat and MODIS-like images of November 26, 2004 (Fig. 6(a) and (c)) and the MODIS-like image of December 12, 2004 (Fig. 6(d)) were used to predict the Landsat image of December 12, 2004 (Fig. 6(b)). In the process of FSDAF, Fig. 6(a) was classified into four spectral classes by ISODATA method.

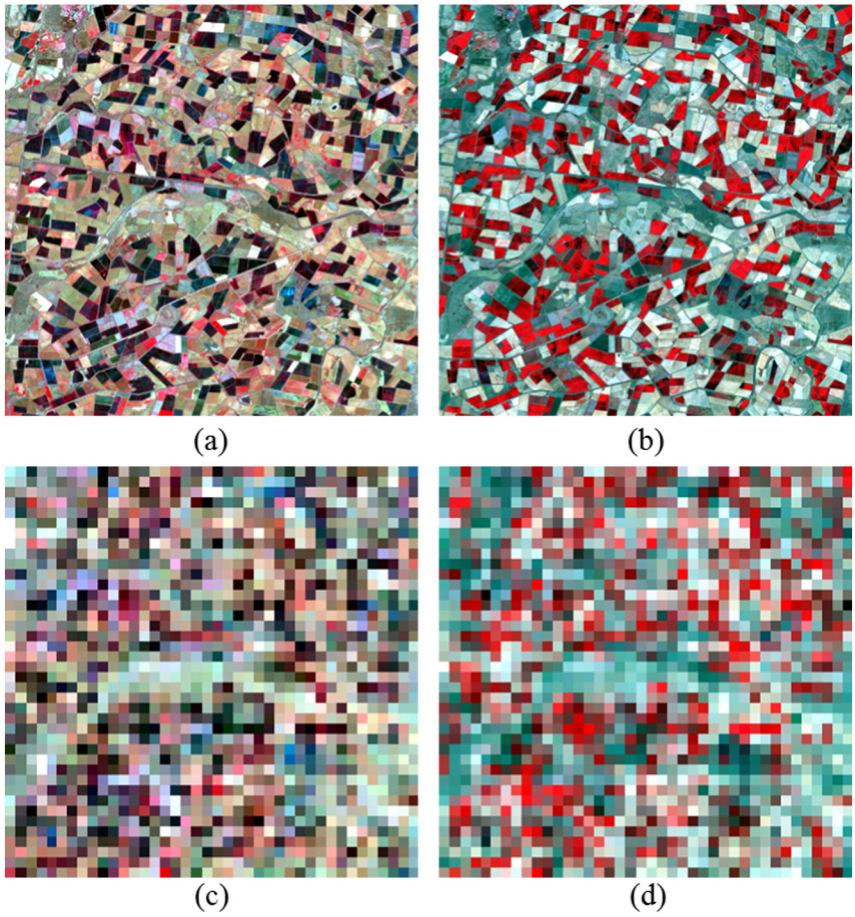

Fig. 5. Test data in a heterogeneous landscape: Landsat images (800 × 800 pixels) acquired on (a) November 25, 2001 and (b) January 12, 2002, (c) and (d) are 500 m MODIS-like images aggregated from (a) and (b). All images use NIR-red-green as RGB, and MODIS-like images are resampled to have same image size as the Landsat images.

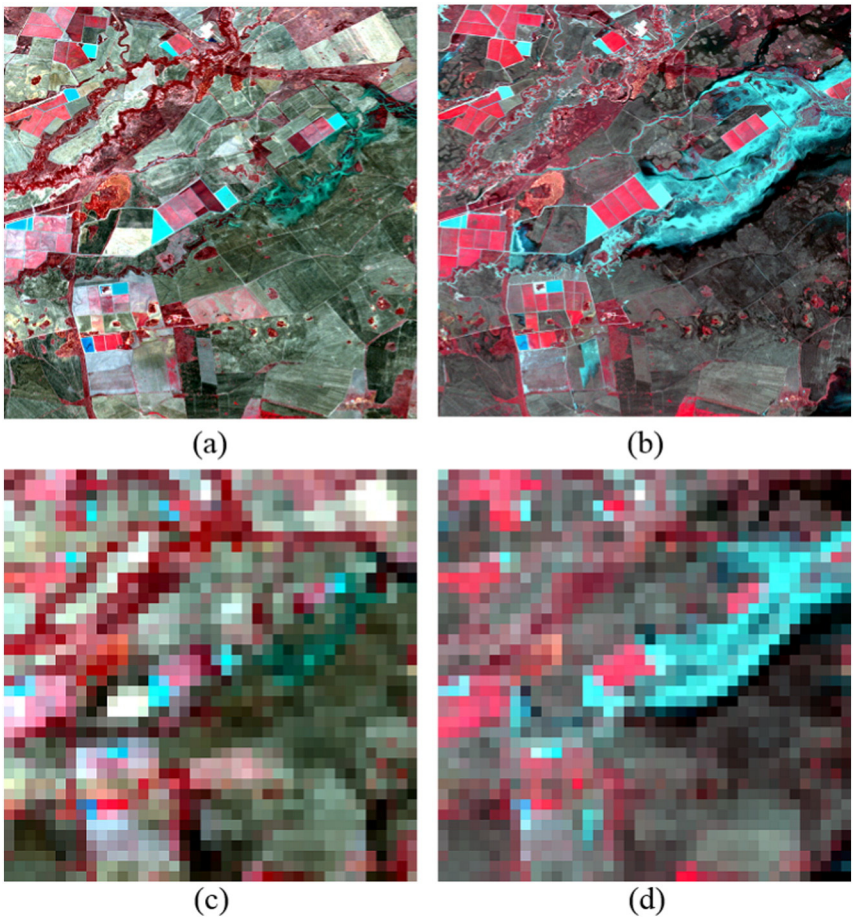

Fig. 6. Test data in area with land cover type change: Landsat images (800 × 800 pixels) acquired on (a) November 26, 2004 and (b) December 12, 2004, (c) and (d) are 500 m MODIS-like images aggregated from (a) and (b). All images use NIR-red-green as RGB and MODIS-like images are resampled to have the same image size as the Landsat images.

3.2. Comparison and evaluation

The performance of FSDAF was also compared with the STAFRM algorithm (Gao et al., 2006) and an unmixing-based data fusion (UBDF) algorithm (Zurita-Milla et al., 2008), because both algorithms have been widely used, and each of them only needs one-pair of fine- and coarse-resolution images. The fine-resolution images predicted by all methods were compared with the true images quantitatively and visually. Several indices were calculated to represent different aspects of accuracy. Root mean square error (*RMSE*) was used to gauge the difference between the predicted reflectance and the actual reflectance. Correlation coefficient r was used to show the linear relationship between predicted and actual reflectance. Average difference (*AD*) between predicted and true images was used to represent the overall bias of predictions. Positive *AD* indicates that the fused image generally overestimates the actual values, while negative *AD* means underestimation. Besides the above quantitative assessment, a visual assessment index, structure similarity (*SSIM*) (Wang, Bovik, Sheikh, & Simoncelli, 2004), was also used to evaluate the similarity of the overall structure between the true and predicted images:

$$SSIM = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X + \sigma_Y + C_2)}, \quad (25)$$

where μ_X and μ_Y are means, σ_X and σ_Y are variance of true and predicted images, σ_{XY} is the covariance of the two images, C_1 and C_2 are two small constants to avoid unstable results when the denominator of Eq. (25) is very close to zero. A *SSIM* value closer to 1 indicates more similarity between the two images. To better demonstrate the effectiveness of data fusion methods, these four accuracy indices were also calculated between the actual fine-resolution image at prediction time and the input fine-resolution image. Indices from two actual fine-resolution images were used as baseline to evaluate whether data fusion methods can add correct temporal information to the input fine-resolution image.

Images produced by spatiotemporal data fusion methods have various applications (Emelyanova et al., 2013). Land cover classification is one important application of these fused images. To evaluate whether or not FSDAF can benefit the further applications, we classified the original Landsat image and all predicted images of the second site experienced a large flood in December 2004 (Fig. 6) to get land cover maps. To exclude effects from other factors, we applied the same classifier, i.e., support vector machine (SVM), and the same set of training data to all images. These images were classified into vegetation, low-vegetation, inundated land, and water. The classification map of original Landsat image was used as reference map to quantitatively assess the agreement between it and other three classification maps of predicted images by error matrix (Liu, Frazier, & Kumar, 2007). Overall accuracy (*oa*) and *kappa* coefficient derived from error matrix were reported to evaluate the agreement at map level. For the category level, the average value of user's and producer's accuracy (*aup*) was used to assess the agreement of each class (Liu et al., 2007).

4. Results

4.1. Test with simulated data

Through visually comparing the predicted results via the three methods (Fig. 7), we can see that all of them can maintain spatial details

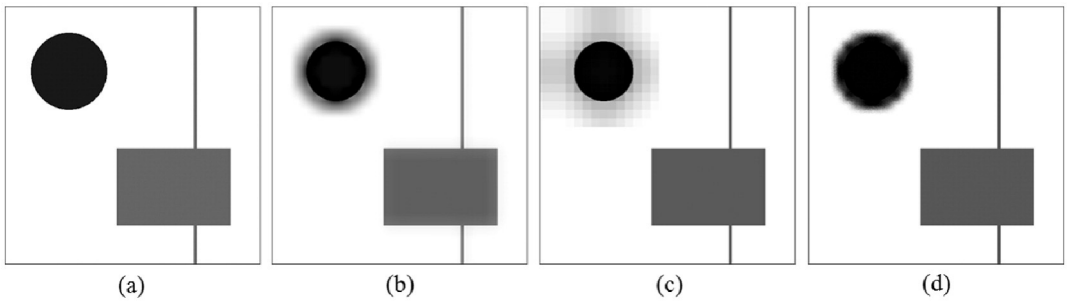

Fig. 7. Comparisons between the original and the predicted Landsat-like images for the simulated data set: (a) original simulated Landsat-like image at t_2, (b) predicted image by STARFM, (c) predicted image by UBDF, and (d) predicted image by the proposed FSDAF method.

Table 2
Accuracy assessment of three data fusion methods applied to the simulated dataset in Fig. 2. The units are reflectance (RMSE = Root Mean Square Error, r = correlation coefficient, AD = average difference from true reflectance, SSIM = structural similarity).

Method	*RMSE*	*r*	*AD*	*SSIM*
Input fine image	0.0846	0.836	0.0251	0.8129
STARFM	0.0405	0.9617	0.0020	0.9592
UBDF	0.0583	0.9187	0.0003	0.9131
FSDAF	0.0256	0.9841	0.0001	0.9843

where no land cover type change happens. For the expanded circle object, however, UBDF produced a blurred zone around the circle (Fig. 7(c)). STARFM produced a less blurry zone than UBDF, but the circular object is still very different from the true image. In contrast, the circle object predicted by FSDAF is much more similar in shape to the true image than the object predicted by other two methods, suggesting that FSDAF is able to produce a satisfactory simulation of the shape of objects that have undergone land cover type change. Quantitative comparisons show that predicted images by all data fusion methods have smaller *RMSE* and *AD*, and higher *r* and *SSIM* than those computed between the input fine-resolution image and the actual fine-resolution image at prediction time (Table 2), which implies that all data fusion methods can more or less gain temporal information from coarse-resolution images to adjust the values of fine pixels between t_1 and t_2. Among the three methods, FSDAF has the smallest errors and highest similarity to the true image. STARFM performed worse than FSDAF but better than UBDF. The prediction of both FSDAF and UBDF is nearly unbiased, while STARFM overestimated the true values.

4.2. Test with satellite images in heterogeneous landscape

Fig. 8 presents the Landsat-like image on January 12, 2002 predicted by the three methods. A zoom-in area was also used to highlight the difference between predicted images and the actual image (Fig. 9). From the visual comparison, the images that the three methods predict are generally similar to the original Landsat image in Fig. 8(a), suggesting that all methods are able to capture the general temporal change in croplands from November 25, 2001 to January 12, 2002. However, the predicted image of FSDAF is more similar to the original image than are the images predicted by STARFM and UBDF in regards to spatial details, which can be seen from the zoom-in images in Fig. 9. Particularly, comparing zoom-in area of the two original Landsat images, we can see that there is a parcel of cropland changed from vegetation to non-vegetation. For this small parcel, both STARFM and UBDF cannot accurately predict its pixel values. In addition, a block effect can be seen in the result of UBDF. In contrast, FSDAF is better at preserving the shapes of small objects. Comparing the quantitative indices calculated using the input Landsat image from November 25, 2001 with the fused results

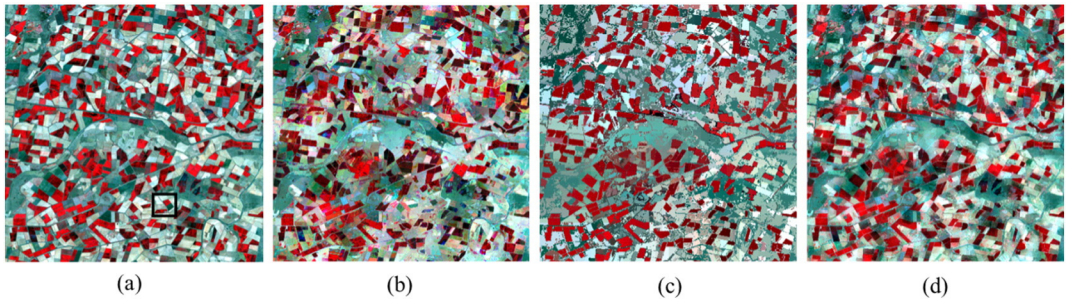

Fig. 8. Original Landsat image of January 12, 2002 (a) and its predicted images by STARFM (b), UBDF (c), and FSDAF (d).

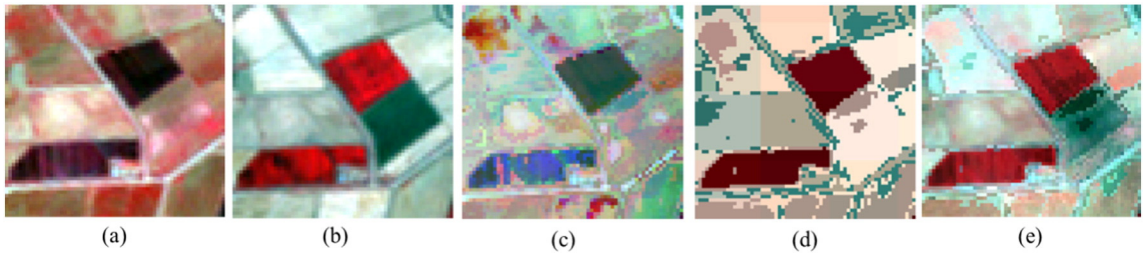

Fig. 9. Zoom in scenes of area marked in Fig. 8(a): original Landsat image of November 25, 2001 (a), original Landsat image of January 12, 2002 (b), and predicted images by STARFM (c), UBDF (d), and FSDAF (e).

Table 3
Accuracy assessment of three data fusion methods applied to the heterogeneous study site (Fig. 8). The units are reflectance (RMSE = Root Mean Square Error, r = correlation coefficient, AD = average difference from true reflectance, SSIM = structural similarity).

	Landsat 11/25/2001				STARFM				UBDF				FSDAF			
	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM
band1	0.028	0.596	−0.016	0.473	0.018	0.771	0.000	0.738	0.019	0.781	0.000	0.780	0.014	0.872	0.000	0.867
band2	0.045	0.603	−0.028	0.447	0.028	0.775	0.000	0.741	0.030	0.756	0.000	0.755	0.022	0.872	0.000	0.865
band3	0.068	0.728	−0.038	0.555	0.045	0.825	0.000	0.806	0.052	0.778	0.001	0.778	0.034	0.900	0.000	0.894
band4	0.131	0.093	−0.092	0.087	0.061	0.506	−0.001	0.482	0.065	0.416	0.000	0.392	0.045	0.743	0.000	0.718
band5	0.092	0.798	−0.065	0.769	0.052	0.860	0.000	0.859	0.061	0.814	0.001	0.814	0.044	0.901	0.000	0.899
band7	0.062	0.803	−0.036	0.784	0.041	0.865	0.000	0.863	0.046	0.838	0.001	0.838	0.035	0.903	0.000	0.902

(Table 3), we can see that all three methods have successfully added certain temporal change information to the input Landsat image to get the prediction on January 12, 2002. For all 6 bands, the fused results of FSDAF have smaller *RMSE* and higher *r* and *SSIM* than STARFM and UBDF (Table 3), suggesting that FSDAF predictions are more accurate than those of STARFM and UBDF. Among all bands, the 4th near infrared (NIR) band has the largest difference in accuracy between FSDAF and other two methods (*RMSE* 0.045 vs. 0.061 and 0.065, *r* 0.743 vs. 0.506 and 0.416, *SSIM* 0.718 vs. 0.482 and 0.392). The scatter plots of actual vs. predicted NIR band by the three methods also confirmed that values predicted by FSDAF are closer to the actual values than the other two methods (Fig. 10). The two Landsat images were acquired within the early growing season of crops, so the NIR band experienced larger reflectance change than other bands. Compared with STARFM and UBDF, the large improvement in predicting the NIR band by FSDAF indicates that it is more capable of capturing large temporal change between input and prediction dates. For the overall prediction bias, all methods can obtain nearly unbiased results for each band ($|AD|<0.001$).

4.3. Test with satellite images experiencing land cover type change

Fig. 11 presents Landsat-like images predicted by the three methods, the original Landsat image of December 12, 2004, and zoom-in scenes of the region inundated by floods to show more details. It is apparent that the Landsat-like image predicted by FSDAF seems to be closer to the true image than are those predicted by STARFM and UBDF. The zoom-in images show that FSDAF successfully captures reflectance changes in pixels that have been inundated. In contrast, STARFM predicted a

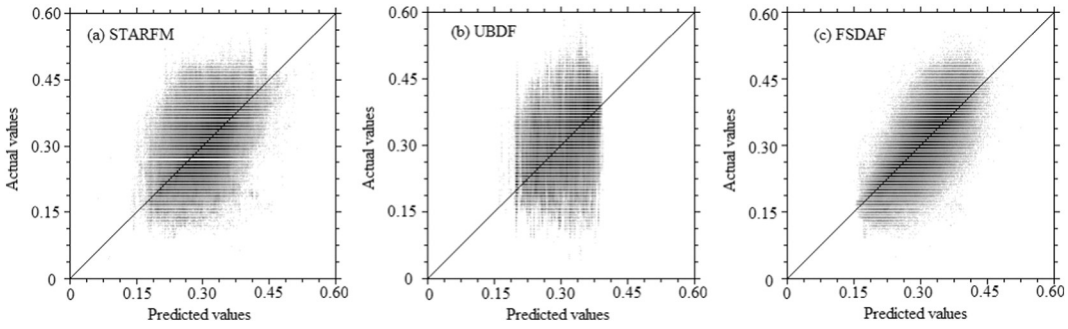

Fig. 10. Scatter plots of the actual and predicted values for NIR band (darker color indicates a higher density of points, and the line is 1:1 line). Panels (a)–(c) are the scatter plots of Fig. 8(b)–(d), respectively.

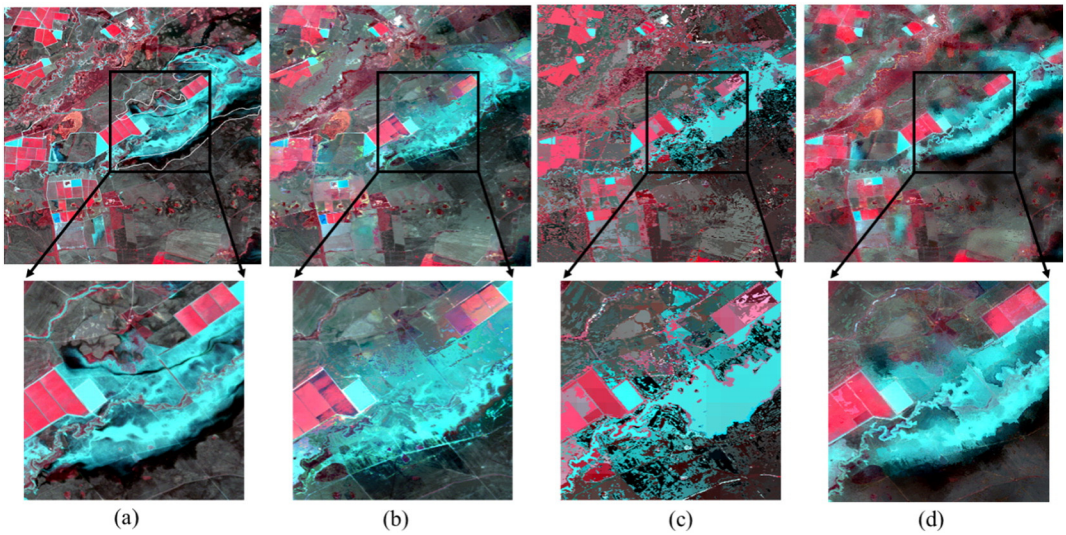

Fig. 11. Original Landsat image of December 12, 2004 (a) and its predicted images by STARFM (b), UBDF (c), and FSDAF (d). White line in (a) delineates the boundaries of inundated area. The lower row images are zoom-in scenes of area marked in the upper row images.

Table 4
Accuracy assessment of three data fusion methods applied to the study site with land cover type change (Fig. 11). The units are reflectance (RMSE = Root Mean Square Error, r = correlation coefficient, AD = average difference from true reflectance, SSIM = structural similarity).

	Landsat 11/26/2004				STARFM				UBDF				FSDAF			
	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM
band1	0.030	0.420	0.023	0.394	0.011	0.816	0.000	0.803	0.018	0.506	0.000	0.505	0.010	0.855	0.000	0.848
band2	0.040	0.395	0.030	0.368	0.016	0.812	0.000	0.800	0.025	0.552	0.000	0.551	0.013	0.865	0.000	0.857
band3	0.053	0.378	0.041	0.350	0.019	0.792	0.000	0.778	0.030	0.524	0.000	0.523	0.016	0.852	0.000	0.843
band4	0.072	0.601	0.057	0.534	0.026	0.875	0.000	0.868	0.045	0.682	0.000	0.679	0.022	0.917	0.000	0.910
band5	0.147	0.426	0.124	0.346	0.045	0.841	0.000	0.828	0.071	0.613	0.001	0.610	0.040	0.881	0.001	0.872
band7	0.121	0.441	0.106	0.348	0.033	0.839	0.000	0.827	0.052	0.607	0.001	0.604	0.030	0.874	0.001	0.864

more "blurry" image with less clear boundaries for the inundated area and UBDF predicted an image with very large errors in the inundated area. The quantitative indices calculated from fused results and from the input Landsat image on November 26, 2004 demonstrate that all data fusion methods have captured certain temporal change information between the input and prediction images (Table 4). For all 6 bands, FSDAF provided the most accurate predictions with the smallest *RMSE* and highest *r* and *SSIM*. STARFM has lower accuracy than FSDAF in terms of quantitative assessment but it performed much better than UBDF. Scatter plots of NIR bands shown in Fig. 12 also suggest that values predicted by FSDAF are closer to actual values than the other two methods. For the overall bias of the prediction, the small *AD* values reveal that all three methods obtained nearly unbiased results. Table 4 shows quantitative assessment for the whole image. To better compare their performance for predicting pixels at the boundaries of the inundated area, we also calculated accuracy indices of all methods only using pixels within the region marked in Fig. 11(a) (Table 5). The improvement in accuracy with FSDAF is most obvious for the marked boundaries of inundated area. The 5th band is the one with most change when flooded. For this band, the *SSIM* values of FSDAF, STARFM, and UBDF are 0.787, 0.563, and 0.303 respectively, suggesting that FSDAF is more powerful for predicting pixels located in a complex area, i.e., a transitional area of change vs. non-change, which is a condition that challenges existing spatiotemporal data fusion methods.

Table 6 indices of agreement between the classifications of the predicted images (Fig. 11(b)–(d)) and classification of the original Landsat image (Fig. 11(a)) using the same training data. Higher values of *oa* and *kappa* suggest a higher similarity between the classification map of a given predicted image and the original image. Classification of the FSDAF-predicted image has the largest values *oa* and *kappa*, followed by FSTARFM, and UBDF has lowest indices of agreement. For individual classes, FSDAF also has higher agreement than STARFM and UBDF for all four of the classes. In particular, inundated land and water mapped in image predicted by FSDAF has much higher agreement than the other two methods, suggesting that FSDAF can better retrieve pixels that have undergone land cover type change during the flooding event.

5. Discussion and conclusions

To increase our ability to monitor rapid land surface dynamics in heterogeneous areas, spatiotemporal data fusion methods have been developed to blend satellite images with different spatial and temporal resolutions. However, previous methods have difficulties predicting pixel values at fine resolution in heterogeneous areas where land cover type change happens during the period between the input and prediction dates when a fine spatial resolution image is only available before the change. To overcome this limitation, this study proposed a new spatiotemporal data fusion method, FSDAF, to blend temporally sparse fine-resolution images with temporally dense coarse-resolution images. FSDAF integrates ideas from unmixing based methods, spatial interpolation, and STARFM into one framework. FSDAF was tested here in a simple simulated scenario and two real landscapes and compared with two popular spatiotemporal data fusion method, STARFM and the unmixing-based data fusion method that also can use only one fine resolution image as input. All results demonstrate that FSDAF can achieve higher accuracy, keep more spatial details, and better retrieve land cover type changes in the predicted fine-resolution images. The better results obtained by FSDAF can be attributed to the strengths described next.

First, the temporal change of endmembers solved in FSDAF is more robust than other unmixing based methods because of the following strategies: 1) FSDAF solves the temporal change globally, avoiding missing the small objects. Existing unmixing based methods often discard endmembers with small fractional abundance (e.g., less than 0.1) in a moving window to avoid unrealistic estimates (Gevaert & García-Haro, 2015), which results in missing temporal change for small objects. It is common for endmembers with small fractional abundance in one coarse pixel to be more abundant in other coarse pixels. In our method, for each endmember, we seek a certain number of coarse pixels which contain each endmember from the whole image, so we can predict the changes for small objects. 2) FSDAF uses the purest coarse pixels to compose the linear equation system, minimizing the impact of collinearity among coarse pixels on the solution. Existing unmixing methods

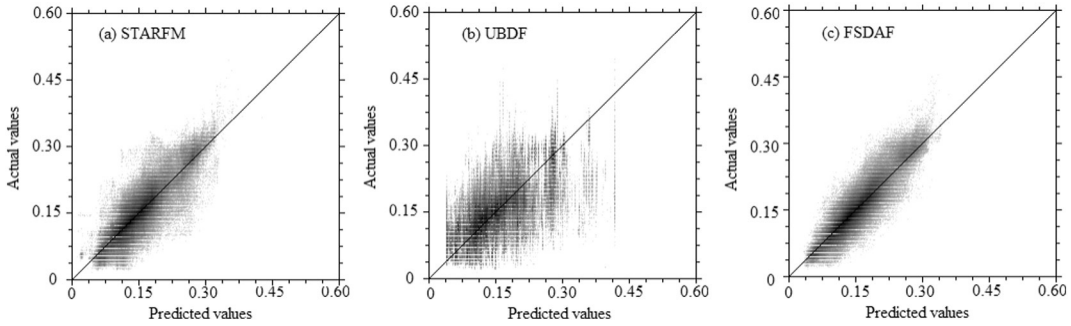

Fig. 12. Scatter plots of the actual and predicted values for NIR band (darker color indicates a higher density of points, and the line is 1:1 line). Panels (a)–(c) are the scatter plots of Fig. 11(b)–(d), respectively.

Table 5
Accuracy assessment of three data fusion methods only using pixels within the boundaries of inundated area marked in Fig. 11(a). The units are reflectance (RMSE = Root Mean Square Error, r = correlation coefficient, AD = average difference from true reflectance, SSIM = structural similarity).

	Landsat 11/26/2004				STARFM				UBDF				FSDAF			
	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM	RMSE	r	AD	SSIM
band1	0.007	0.434	0.024	0.328	0.004	0.662	0.004	0.596	0.006	0.367	0.005	0.366	0.003	0.872	0.000	0.857
band2	0.010	0.363	0.031	0.259	0.006	0.622	0.005	0.572	0.008	0.384	0.007	0.382	0.005	0.875	0.000	0.859
band3	0.013	0.303	0.043	0.225	0.007	0.589	0.006	0.544	0.010	0.347	0.008	0.345	0.006	0.873	0.000	0.858
band4	0.022	0.532	0.091	0.371	0.007	0.693	0.008	0.659	0.012	0.497	0.014	0.474	0.005	0.872	0.002	0.854
band5	0.049	0.222	0.208	0.095	0.013	0.598	0.014	0.563	0.019	0.315	0.022	0.303	0.010	0.806	0.005	0.787
band7	0.037	0.291	0.156	0.125	0.008	0.638	0.009	0.600	0.013	0.321	0.015	0.310	0.007	0.804	0.003	0.779

use coarse pixels within a moving window, which will usually be spatially autocorrelated, causing collinearity in the equation system. 3) FSDAF refines the selected purest coarse pixels to avoid the effects of land cover type change on temporal change estimation. Existing unmixing based methods use endmember fractions obtained directly from the input fine-resolution image, assuming no land cover type change between the input image and the prediction date. However, this assumption may be not valid in areas with frequent disturbance or land cover type change, or a long time interval between input and prediction dates. If no land cover type change happens, the temporal change of selected purest coarse pixels for each endmember should be similar to each other. In addition, pixels with land cover type change are often rare compared with the no-change pixels in an image. Therefore, it is reasonable to use quantiles to exclude coarse pixels with probable land cover type change based on their relative abundance.

Second, local variability of temporal change caused by land cover conversions or within-class differences is modeled well through the distribution of residuals. STARFM and unmixing based methods estimate reflectance change within a moving window to account for within-class differences. However, this moving-window strategy leads to inaccurate estimates in highly mixed landscapes in STARFM and unrealistic solutions for pixels with high correlation in unmixing based methods. In addition, these methods do not consider land cover type change. The proposed method decomposes the total change of each fine pixel into global change and local change. The local change, including both within-class differences and land cover type change, is estimated through distribution of the residuals from the temporal prediction. Because the coarse-resolution image is the only available information showing the situation at the prediction date, it is used to guide the distribution of residuals. Assuming both within-class differences and land cover type change have spatial dependence, FSDAF applies the TPS method to downscale the coarse-resolution image at the prediction date to fine resolution. The downscaled image can help us to judge which pixels have land cover type change or within-class variance so that we can better distribute residuals.

Third, FSDAF predicts images with good spatial continuity through bringing in neighborhood information. This strategy has been used in STARFM and STARFM-like methods but never in unmixing based methods. Current unmixing based methods are implemented one coarse pixel at a time. Although they use neighboring coarse pixels in a moving window to solve spectral values or temporal change of endmembers, the neighboring coarse pixels are not involved in predicting fine pixels within the coarse pixel at the center of the moving window. In other words, fine pixels with the same class in one coarse pixel use the same solved values which are more or less different from the neighboring coarse pixels. When the solved values between two neighboring coarse pixels are significantly different from each other, there inevitably exists discontinuity between fine pixels crossing the boundaries of two neighboring coarse pixels. This discontinuity causes block effects, i.e., the visible trace of coarse pixels. Like these unmixing based method, the distribution of residuals in FSDAF is also done for each individual coarse pixel. If we use the result after residual distribution as the final prediction, it may also have block effect. Considering that closer same-class fine pixels should have similar temporal change patterns, for each fine pixel, FSDAF uses the weighted average of its surrounding fine pixels to obtain its total change. This step ultimately improves the spatial continuity of predicted fine-resolution images.

Last, FSDAF has comparable efficiency with STARFM and unmixing based methods, even though it seems to have more steps. Through checking the computing time of each step in FSDAF, we found that the last step, i.e., final prediction using information in neighborhood, is the most time-consuming one. This is because other steps are implemented in coarse pixels while the last step is done for each fine pixel. Actually, the last step is like the procedure of STARFM. However, FSDAF only uses 20 selected similar pixels while STARFM uses much more similar pixels. In other words, FSDAF only needs to perform the neighborhood calculation for a small portion of the similar pixels used in STARFM. As a result, FSDAF needs nearly equal time as STARFM even though it has more steps. For unmixing based methods, most computing time is used to invert the linear equation system to obtain endmember values at fine resolution for each coarse pixel. In contrast, FSDAF only needs to invert the linear equation system one time, because it uses global change values of endmembers. Therefore, FSDAF only needs about one third more processing time than unmixing based methods despite its additional steps.

Although FSDAF can predict both gradual reflectance change and land cover type change between the input and prediction dates, it cannot capture tiny changes in land cover type, for example if only a few fine pixels experienced land cover type change and the change is invisible in the coarse-resolution image. For the tiny change, a possible solution is bringing another available fine-resolution image acquired after the change into the process. With a later image, change detection between the two fine-resolution images could identify fine pixels with land cover type change, and then values could be predicted for these pixels. However, this solution has challenges in cloudy regions where it is hard to acquire another cloud-free fine-solution image, and it also does not work if the change was short-lived, as might happen, for example, with small flood events.

Like STARFM and unmixing based methods, FSDAF is not only designed for fusing reflectance data of different sensors, it can also be directly applied to any products derived from reflectance data if these products are linearly additive in space, such as some simple spectral indices, leaf area index, and fraction of photosynthetically active radiation. A current study suggests that spatiotemporal data fusion methods directly applied to vegetation indices could obtain higher accuracy than blending reflectance of individual bands and then calculating indices, because of less error propagation (Jarihani et al., 2014). Some other products, such as NDVI and LST, are intrinsically nonlinearly additive,

Table 6
Indices of agreement between land cover classification of each predicted image in Fig. 11(b)–(d) as compared with the actual Landsat image in Fig. 11(a).

	oa	*kappa*	*aup*			
			Vegetation	Low-vegetation	Inundated	Water
FSTARFM	0.783	0.66	0.790	0.844	0.770	0.588
UBDF	0.650	0.47	0.649	0.752	0.595	0.356
FSDAF	0.835	0.74	0.815	0.875	0.820	0.689

but assuming them to be linearly additive would probably introduce only very small errors, such that these products with different spatial and temporal resolutions can also be fused by the proposed method. In addition, similar to STARFM, FSDAF could be extended to use two pairs of fine and coarse-resolution images as input data. In this situation, FSDAF can be applied to each pair of input to get two separate predictions. Then, a temporal weight can be used to combine the two predictions (Gao et al., 2006; Zhu et al., 2010). It may get more accurate and robust results if the second pair can provide complementary information. The code of FSDAF is available by sending request to authors. In conclusion, the proposed FSDAF method needs minimal input data and is able to capture both gradual change and land cover type change. It is an important supplement to the family of spatiotemporal data fusion methods to support studies of land surface dynamics which require satellite images with high frequency and high spatial resolution. The potential applications of synthetic high-spatial resolution data include monitoring forest disturbance and phenology, mapping land cover and land use change, real-time disaster detection, and tracking crop progress and condition.

Acknowledgments

This study was supported by the USDA Forest Service International Institute of Tropical Forestry (cooperative agreement 13-CA-11120101-029) and the Southern Research Station Forest Inventory and Analysis Program, and it was conducted in cooperation with the University of Puerto Rico and the Rocky Mountain Research Station.

References

Amorós-López, J., Gómez-Chova, L., Alonso, L., Guanter, L., Zurita-Milla, R., Moreno, J., & Camps-Valls, G. (2013). Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring. *International Journal of Applied Earth Observation and Geoinformation*, *23*, 132–141. http://dx.doi.org/10.1016/j.jag.2012.12.004.

Ball, G. H., & Hall, D. J. (1965). ISODATA, a novel method of data analysis and pattern classification. *Proceedings of the IEEE* (Retrieved from http://www.stormingmedia.us/61/6169/0616996.pdf).

Chen, X., Li, W., Chen, J., Rao, Y., & Yamaguchi, Y. (2014). A combination of TsHARP and thin plate spline interpolation for spatial sharpening of thermal imagery. *Remote Sensing*, *6*, 2845–2863. http://dx.doi.org/10.3390/rs6042845.

Dubrule, O. (1984). Comparing splines and kriging. *Computers & Geosciences*, *10*, 327–338. http://dx.doi.org/10.1016/0098-3004(84)90030-X.

Emelyanova, I. V., McVicar, T. R., Van Niel, T. G., Li, L. T., & van Dijk, A. I. J. M. (2013). Assessing the accuracy of blending Landsat–MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection. *Remote Sensing of Environment*, *133*, 193–209. http://dx.doi.org/10.1016/j.rse.2013.02.007.

Fu, D., Chen, B., Wang, J., Zhu, X., & Hilker, T. (2013). An improved image fusion approach based on enhanced spatial and temporal the adaptive reflectance fusion model. *Remote Sensing*, *5*(12), 6346–6360.

Galford, G. L., Mustard, J. F., Melillo, J., Gendrin, A., Cerri, C. C., & Cerri, C. E. P. (2008). Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil. *Remote Sensing of Environment*, *112*, 576–587. http://dx.doi.org/10.1016/j.rse.2007.05.017.

Gao, F., Masek, J., Schwaller, M., & Hall, F. (2006). On the blending of the Landsat and MODIS surface reflectance : Predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(8), 2207–2218.

Gao, F., Masek, J., Wolfe, R., & Huang, C. (2010). Building consistent medium resolution satellite data set using moderate resolution imaging spectroradiometer products as reference. *Journal of Applied Remote Sensing*, *4*, 043526. http://dx.doi.org/10.1117/1.3430002.

Gevaert, C. M., & García-Haro, F. J. (2015). A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. *Remote Sensing of Environment*, *156*, 34–44. http://dx.doi.org/10.1016/j.rse.2014.09.012.

Hilker, T., Wulder, M. A., Coops, N. C., Linke, J., McDermid, G., Masek, J. G., ... White, J. C. (2009a). A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. *Remote Sensing of Environment*, *113*, 1613–1627. http://dx.doi.org/10.1016/j.rse.2009.03.007.

Hilker, T., Wulder, M. A., Coops, N. C., Seitz, N., White, J. C., Gao, F., ... Stenhouse, G. (2009b). Generation of dense time series synthetic Landsat data through data blending with MODIS using a spatial and temporal adaptive reflectance fusion model. *Remote Sensing of Environment*, *113*, 1988–1999. http://dx.doi.org/10.1016/j.rse.2009.05.011.

Huang, B., & Song, H. (2012). *Spatiotemporal reflectance fusion via sparse representation50(10). (pp. 3707–3716), 3707–3716.*

Huang, B., & Zhang, H. (2014). Spatio-temporal reflectance fusion via unmixing: Accounting for both phenological and land-cover changes. *International Journal of Remote Sensing*, 1–21 (September 2014). (doi:10.1080/01431161.2014.951097).

Jarihani, A. A., McVicar, T. R., Van Niel, T. G., Emelyanova, I. V., Callow, J. N., & Johansen, K. (2014). Blending Landsat and MODIS data to generate multispectral indices: A comparison of "index-then-blend" and "blend-then-index" approaches. *Remote Sensing*, *6*(10), 9213–9238. http://dx.doi.org/10.3390/rs6109213.

Ju, J., & Roy, D. P. (2008). The availability of cloud-free Landsat ETM + data over the conterminous United States and globally. *Remote Sensing of Environment*, *112*, 1196–1211. http://dx.doi.org/10.1016/j.rse.2007.08.011.

Liu, D., & Zhu, X. (2012). An enhanced physical method for downscaling thermal infrared radiance. *IEEE Geoscience and Remote Sensing Letters*, *9*(4), 690–694.

Liu, C., Frazier, P., & Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, *107*, 606–616. http://dx.doi.org/10.1016/j.rse.2006.10.010.

Rao, Y., Zhu, X., Chen, J., & Wang, J. (2015). An improved method for producing high spatial-resolution NDVI time series datasets with multi-temporal MODIS NDVI data and Landsat TM/ETM + images. *Remote Sensing*, *7*, 7865–7891. http://dx.doi.org/10.3390/rs70607865.

Senf, C., Leitão, J. P., Pflugmacher, D., van der Linden, S., & Hostert, P. (2015). Mapping land cover in complex Mediterranean landscapes using Landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery. *Remote Sensing of Environment*, *156*, 527–536. http://dx.doi.org/10.1016/j.rse.2014.10.018.

Shen, M., Tang, Y., Chen, J., Zhu, X., & Zheng, Y. (2011). Influences of temperature and precipitation before the growing season on spring phenology in grasslands of the central and eastern Qinghai–Tibetan Plateau. *Agricultural and Forest Meteorology*, *151*, 1711–1722. http://dx.doi.org/10.1016/j.agrformet.2011.07.003.

Song, H., & Huang, B. (2013). Spatiotemporal satellite image fusion through one-pair image learning. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(4), 1883–1896.

Walker, J. J., de Beurs, K. M., Wynne, R. H., & Gao, F. (2012). Evaluation of Landsat and MODIS data fusion products for analysis of dryland forest phenology. *Remote Sensing of Environment*, *117*, 381–393. http://dx.doi.org/10.1016/j.rse.2011.10.014.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*, 600–612. http://dx.doi.org/10.1109/TIP.2003.819861.

Watts, J. D., Powell, S. L., Lawrence, R. L., & Hilker, T. (2011). Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery. *Remote Sensing of Environment*, *115*, 66–75. http://dx.doi.org/10.1016/j.rse.2010.08.005.

Weng, Q., Fu, P., & Gao, F. (2014). Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data. *Remote Sensing of Environment*, *145*, 55–67. http://dx.doi.org/10.1016/j.rse.2014.02.003.

Wu, M., Wang, C., & Wang, L. (2012). Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model. *Journal of Applied Remote Sensing*, *6*. http://dx.doi.org/10.1117/1.JRS.6.063507.

Yang, X., & Lo, C. P. (2002). Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area. *International Journal of Remote Sensing*, *23*, 1775–1798. http://dx.doi.org/10.1080/01431160110075802.

Yu, L., Shi, Y., & Gong, P. (2015). Land cover mapping and data availability in critical terrestrial ecoregions: A global perspective with Landsat thematic mapper and enhanced thematic mapper plus data. *Biological Conservation*, *190*, 34–42. http://dx.doi.org/10.1016/j.biocon.2015.05.009.

Zhou, Y., Chen, J., Chen, X. H., Cao, X., & Zhu, X. L. (2013). Two important indicators with potential to identify Caragana microphylla in xilin gol grassland from temporal MODIS data. *Ecological Indicators*, *34*, 520–527. http://dx.doi.org/10.1016/j.ecolind.2013.06.014.

Zhu, X., Chen, J., Gao, F., Chen, X., & Masek, J. G. (2010). An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sensing of Environment*, *114*(11), 2610–2623. http://dx.doi.org/10.1016/j.rse.2010.05.032.

Zhukov, B., Oertel, D., Lanzl, F., & Reinhäckel, G. (1999). Unmixing-based multisensor multiresolution image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, *37*, 1212–1226. http://dx.doi.org/10.1109/36.763276.

Zurita-Milla, R., Clevers, J. G. P. W., & Schaepman, M. E. (2008). Unmixing-based landsat TM and MERIS FR data fusion. *IEEE Geoscience and Remote Sensing Letters*, *5*, 453–457. http://dx.doi.org/10.1109/LGRS.2008.919685.