Enhancing Spatio-Temporal Fusion of MODIS and Landsat Data by Incorporating 250 m MODIS Data

Qunming Wang, Yihang Zhang, Alex O. Onojeghuo, Xiaolin Zhu, Peter M. Atkinson

***Abstract*—Spatio-temporal fusion of MODIS and Landsat data aims to produce new data that have simultaneously the Landsat spatial resolution and MODIS temporal resolution. It is an ill-posed problem involving large uncertainty, especially for reproduction of abrupt changes and heterogeneous landscapes. In this paper, we proposed to incorporate the freely available 250 m MODIS images into spatio-temporal fusion to increase prediction accuracy. The 250 m MODIS bands 1 and 2 are fused with 500 m MODIS bands 3 to 7 using the advanced area-to-point regression kriging (ATPRK) approach. Based on a standard spatio-temporal fusion approach, the interim 250 m fused MODIS data are then downscaled to 30 m with the aid of the available 30 m Landsat data on temporally close days. The 250 m data can provide more information for the abrupt changes and heterogeneous landscapes than the original 500 m MODIS data, thus, increasing the accuracy of spatio-temporal fusion predictions. The effectiveness of the proposed scheme was demonstrated using two datasets.**

***Index Terms*—Downscaling, spatio-temporal fusion, image fusion, geostatistics, MODIS, Landsat.**

I. INTRODUCTION

Landsat and MODIS data have been used widely for global monitoring, due to their large swath, free availability and regular revisit capabilities. The MODIS sensor can revisit the same area on a daily basis, which is of great use for timely monitoring of rapid changes on the Earth's surface, such as vegetation phenology [1], [2] and land-cover/land-use change [3]. However, the spatial resolution of MODIS data (ranging from 250 m to 1000 m) is often too coarse to provide the information desired which may exist at a finer spatial scale than the sensor resolution. The Landsat sensor can provide images at a much finer spatial resolution of 30 m, but it can only revisit the same area every 16 days. Furthermore, in most cases, the acquired Landsat data for the specific areas can be contaminated by cloud and shadow, meaning that obtaining one clean Landsat image per month, in many cases, would be considered a good outcome.

Spatio-temporal fusion involves blending fine temporal resolution, but coarse spatial resolution (e.g., 500 m MODIS) data with fine spatial resolution, but coarse temporal resolution (e.g., 30 m Landsat) data to create data at both fine temporal and spatial resolutions [4]-[6]. It is carried out based on the availability of at least one coarse-fine spatial resolution image pair (e.g., MODIS-Landsat image pair) acquired on approximately the same day or at least one fine spatial resolution image that is temporally close to the prediction day. Generally, three types of spatio-temporal fusion approaches can be identified: image-pair-based, spatial unmixing-based and hybrid methods.

The spatial and temporal adaptive reflectance fusion model (STARFM) proposed by Gao *et al*. [7] is a typical image-pair-based approach, and one of the earliest spatio-temporal fusion approaches. With at least one coarse-fine image pair on temporally close days, STARFM calculates the fine spatial resolution reflectance on the prediction day as a linearly weighted combination of the coarse temporal changes (i.e., the reflectance difference between the available coarse images on different days) added to the available fine spatial resolution reflectance. Zhu *et al*. [8] proposed an enhanced STARFM (ESTARFM) method to enhance the performance of STARFM for heterogeneous landscapes. Different from STARFM where the temporal changes of all classes within a coarse pixel are assumed to be uniform (i.e., each coarse pixel is assumed to be pure), ESTARFM considers the temporal changes of each class separately, based on the hypothesis that the change rate of a class is stable during the period of interest. STARFM was also extended with a version termed spatial temporal adaptive algorithm for mapping reflectance change (STAARCH) to produce a dense stack of spatially coincident MODIS images for mapping forest disturbance [9]. Recently, the sparse representation was applied to spatio-temporal fusion. Based on two known coarse-fine image pairs, Huang and Song [10] developed the sparse-representation-based spatio-temporal reflectance fusion model (SPSTFM) to characterize the relationship between the coarse-fine temporal changes. The trained dictionary was used to predict the unknown fine spatial resolution reflectance according to the known coarse temporal changes. In their later work, sparse representation was further extended to the case where only one image pair is available [11]. Specifically, the relationship between the coarse-fine reflectance is characterized directly by a dictionary and a two-layer strategy is used (an intervening spatial resolution is involved) to cope with the large spatial resolution difference between MODIS and Landsat.

The spatial unmixing approaches are usually performed using a fine spatial resolution thematic map, which can be obtained by standard hard classification of the available fine spatial resolution data [12]-[16] or from other sources, such as an aerial image [17], or land-use database [18]. Spatial unmixing is different from the well-known spectral unmixing. The latter

Manuscript received XXX. *(Corresponding author: Q. Wang.)*

Q. Wang is with the Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK (e-mail: wqm11111@126.com).

A. O. Onojeghuo is with the Department of Surveying and Geoinformatics, Nnamdi Azikiwe University, Anambra state, PMB 5025, Awka, Nigeria.

Y. Zhang is with Institute of Geodesy and Geophysics, Chinese Academy of Sciences, Wuhan 430077, China and University of Chinese Academy of Sciences, Beijing 100049, China.

X. Zhu is with The Hong Kong Polytechnic University, Hong Kong, and also with Wuhan University, Wuhan 430072, China.

P.M. Atkinson is with the Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YR, UK; School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, BT7 1NN, Northern Ireland, UK; and also with Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

aims to estimate the proportions of each class within the coarse pixel and the class endmembers (spectra) are pre-determined (by either endmember extraction or referring to supervised information), while the former aims to estimate the class endmembers (for each band) within each coarse pixel and the class proportions are known (calculated from the fine spatial resolution thematic map by upscaling). Spatial unmixing is performed based on the strong assumption that no land-cover/land-use changes occur during the period, and thus, the class proportions are constant for each coarse image. Using a single fine spatial resolution land-use database LGN5, Zurita-Milla *et al*. [18] produced 30 m Landsat-like images from temporally dense, 300 m Medium Resolution Imaging Spectrometer (MERIS) images to monitor vegetation seasonal dynamics, and also investigated the optimal window size and number of classes in spatial unmixing [12]. Amorós-López *et al*. [13], [14] added a new regularization term to the cost function of the spatial unmixing to avoid large deviations of the estimated endmembers from the pre-defined endmembers extracted from the coarse data. Based on the availability of one coarse-fine image pair, Wu *et al*. [15] and Gevaert *et al*.'s method [16] first estimates changes in class endmember spectra from the time of the image pair to prediction and then adds the changes to the known fine spatial resolution reflectance.

The abovementioned two types of approaches were also combined and some hybrid methods were developed. Xu *et al*. [19] proposed a modified regularized spatial unmixing method, in which the regularization term is constructed using the fine spatial resolution endmembers that are extracted from the pre-STARFM predictions, rather than the coarse endmembers of the original coarse data in [13], [14]. Xie *et al*. [20] applied spatial unmixing to decompose the coarse data and STARFM was performed on the unmixing-based predictions. Zhu *et al*. [21] proposed a flexible fusion method that first estimates changes in class endmember spectra during the period of interest and then uses additional neighborhood information, as was done in STARFM, to achieve robust prediction.

Amongst the spatio-temporal fusion approaches, STARFM is one of the most widely used methods for blending MODIS and Landsat data and has been applied in various domains (e.g., forest monitoring, crop monitoring [4] and land surface temperature monitoring [22]), appreciating its simple implementation. Different from ESTARFM that requires at least two coarse-fine image pairs, STARFM can be implemented using only one image pair. Compared to spatial unmixing methods, STARFM requires less strict assumptions about the land-cover/land-use changes during the study period. Similarly to other spatio-temporal fusion approaches, when downscaling MODIS bands 1 to 7 (except band 5, as the wavelength of this band does not match any band of Landsat), STARFM downscales the 500 m bands directly to 30 m, which involves a large zoom factor of 16. Spatio-temporal fusion of MODIS and Landsat data is essentially an ill-posed problem. Such a large zoom factor brings great challenges in downscaling, especially for restoration of temporal changes (e.g., abrupt changes) and spatially heterogeneous landscapes. For abrupt changes, the information is not adequately represented in the observed Landsat data. For example, in the observed Landsat data, the corresponding area may be dominated by a large pure patch (e.g., bare soil), but may be broken down into several smaller patches of very different classes (e.g., vegetation, water, and impervious surface) on the prediction day. In spatio-temporal fusion, the restoration of abrupt changes relies mainly on the pre-interpolation process (e.g., bicubic interpolation in this paper), and the observed Landsat data cannot provide much helpful information. A straightforward solution to this issue would be to seek as many temporally close Landsat data as possible to provide related land-cover boundary information for the abrupt changes, but this may be challenging due to cloud contamination. The cloud contamination problem is actually a key motivation for spatio-temporal fusion.

It is worth noting that MODIS sensors also produce 250 m images in the red (band 1) and near-infrared (band 2) bands. The 250 m images are free and, thus, there is little data cost to use them. The 250 m images provide valuable information at an intermediate spatial resolution and are expected to be able to decrease the uncertainty in the conventional 500 m to 30 m downscaling process (from 250 m to 30 m, the downscaling process involves a much smaller zoom factor of 8).

In this paper, based on the popular STARFM approach, we take full advantage of the freely available 250 m information provided by MODIS sensors to produce more reliable 30 m Landsat-like data. The 250 m data are only available for bands 1 and 2 and bands 3 to 7 need to be downscaled to 250 m. Specifically, the 250 m MODIS bands 1 and 2 are fused with 500 m bands 3 to 7 to produce the interim 250 m bands 1 to 7. The MODIS bands 1 and 2 and bands 3 to 7 have different spectral range and a highly reliable image fusion approach is required for this issue. In our previous research, area-to-point regression kriging (ATPRK) was proposed for fusion of 500 m MODIS bands 3 to 7 and 250 m bands 1 and 2 [23]. ATPRK is an accurate approach for reproduction of spatial detail in bands 3 to 7 and is superior to some state-of-the-art benchmark methods. It is easy to implement and can perfectly preserve the 500 m spectral properties. It accounts explicitly for the size of support, spatial correlation, and the point spread function of the sensor. Based on the free 250 m MODIS data and the appealing advantages of ATPRK (both theoretically and experimentally), there exists a clear opportunity to enhance the spatio-temporal fusion of MODIS and Landsat data, which is developed in this paper.

The remainder of this paper is organized into four sections. Section II introduces briefly the principles of STARFM and the image fusion approach used to incorporate the 250 m MODIS data in spatio-temporal fusion. The experimental results for two datasets are provided in Section III, including validation. Section IV further discusses the proposed scheme of incorporating 250 m MODIS data, followed by a conclusion in Section V.

II. Methods

In this paper, the advanced ATPRK approach is used to fuse 250 m bands with 500 m bands of MODIS to produce the interim 250 m MODIS data. Such a fusion process is conducted for all MODIS data during the period of interest. With the 250 m MODIS and 30 m Landsat image pairs on temporally close days, STARFM is performed on the 250 m fused MODIS data on the prediction date to produce the 30 m Landsat-like data. The principles of ATPRK and STARFM are introduced briefly below.

A. ATPRK

ATPRK was also shown to outperform 13 current state-of-the-art benchmark methods in pan-sharpening in our previous work [24]. Based on its advantages and encouraging performance, ATPRK is adopted for image fusion of MODIS data (fusing 500 m bands with 250 m bands).

ATPRK is a two-step approach including regression modelling and area-to-point kriging (ATPK)-based residual downscaling. In the first step, the fine spatial resolution (250 m) prediction for each observed 500 m coarse band is a linear transformation of the fine spatial resolution band (e.g., 250 m PAN-like band). For MODIS image fusion, a 250 m PAN-like band needs to be selected for each 500 m band based on the correlation coefficient [23]. The coefficients are determined by the regression model built between the observed 500 m band and the corresponding PAN-like 250 m band (the 250 m band needs to be upscaled to 500 m). In the second step, the residuals (at a spatial resolution of 500 m) between the regression prediction and the original coarse data are calculated. The 500 m residuals are then downscaled to 250 m using ATPK, and the estimated 250 m residuals are added back to the 250 m prediction of the first step to produce the final 250 m result. Further details on ATPRK can be found in our previous work [23], [24].

In ATPRK, the second step (i.e., ATPK-based residual downscaling) aims to perfectly preserve the original 500 m data. It is an iterative process that increases the computational cost of ATPRK. In this paper, for fast fusion of the 500 m and 250 m bands, an approximate version of ATPRK is considered. In spatio-temporal fusion, there exists an unavoidable reflectance inconsistency between the images acquired by different sensors. Thus, the 500 m MODIS data do not necessarily need to be perfectly, but rather need to be sufficiently, preserved in the 250 m predictions before being downscaled to the 30 m Landsat resolution. On this basis, a simple and fast bicubic interpolation method is applied in the second step of ATPRK instead. The approximate version with bicubic interpolation in the second step can satisfactorily preserve the 500 m MODIS data (always with a CC between the 250 m predictions and 500 m original data close to 1), and more importantly, greatly expedite the fusion process.

B. STARFM

For a given time t_0, the Landsat-like image $L(x, y, t_0)$ is estimated as

$$L(x_{w/2}, y_{w/2}, t_0) = \sum_{i=1}^{w}\sum_{j=1}^{w}\sum_{k=1}^{n} W_{ijk} \times \left(M(x_i, y_j, t_0) + L(x_i, y_j, t_k) - M(x_i, y_j, t_k)\right) \quad (1)$$

where w is the size of a spatially neighboring window (in units of Landsat pixels), x and y are coordinates of the pixels, n is the number of temporally neighboring images, M means MODIS data, L means Landsat data, and $L(x, y, t_k)$ and $M(x, y, t_k)$ are an image pair at time t_k. W is a weight for neighboring pixels and it is a function of three factors: the spectral difference between the MODIS and Landsat reflectance of the image pair, temporal difference between the MODIS reflectance on different days, and the spatial distance between Landsat pixels. Further details on the approach can be found in [7]. In this paper, w was set to 31, and 20 spectrally similar pixels were considered in each moving window.

In the proposed scheme, the MODIS data in (1) are 250 m data produced using ATPRK, rather than the 500 m data in the original STARFM. Fig. 1 illustrates the flowchart of incorporating 250 m MODIS information in spatio-temporal fusion, where the example involves two pairs of MODIS-Landsat images at t_1 and t_3 are available for prediction of the 30 m Landsat 8 image at t_2. Note that the wavelength of Landsat 8 bands 2-7 (i.e., blue, green, red, NIR, SWR1 and SWR2 bands) corresponds to MODIS bands 3, 4, 1, 2, 6 and 7, respectively.

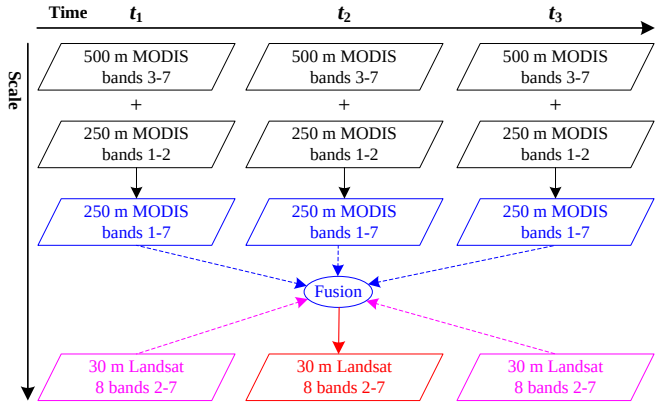

Fig. 1 A flowchart illustrating the proposed scheme of incorparating 250 m MODIS images in spatio-temporal fusion of MODIS and Landsat 8 data.

In (1), for compatibility in image size, the MODIS images need be interpolated to the Landsat spatial resolution (30 m) beforehand. This process can be achieved using the simple bicubic interpolation. Using the original 500 m MODIS images, the interpolation involves a zoom factor of 16, but the factor greatly reduces to 8 when the interim 250 m MODIS images are used as inputs in the proposed scheme. The much smaller zoom factor is, thus, able to decrease the uncertainty in downscaling.

The proposed method was tested using two datasets; one covers a tropical forest area in Amazon, while the other covers an irrigation area in Coleambally, New South Wales, Australia. The results are reported in the following section.

III. Experiments

A. Experiment on the Amazon forest dataset

In this experiment, two Landsat 5 and MODIS surface reflectance image pairs covering a tropical rainforest in the Amazon were used. They were acquired on 13 June 2001 and 21 July 2003. MODIS data from the daily surface reflectance products MOD09GA (500 m) and MOD09GQ (250 m) were used. To illustrate the performance of the proposed method in restoring abrupt changes, two sites experiencing noticeable deforestation were selected for testing. The two sites cover areas of 10 km by 10 km (site 1) and 20 km by 20 km (site 2), respectively. Figs. 2(a) and 2(b) show the two MODIS images and Figs. 2(i) and 2(j) show the two Landsat 5 images for site 1, while Figs. 3(a) and 3(b) show the two MODIS images and Figs. 3(c) and 3(d) show the two Landsat 5 images for site 2. The task of spatio-temporal fusion in this experiment was to predict the 30 m Landsat 5 image (including blue, green, red, NIR, SWR1 and SWR2 bands) on 21 July 2003, where the MODIS images on

the two days and the 30 m Landsat image on 13 June 2001 were used as inputs, and the true 30 m Landsat image on 21 July 2003 was used as reference in accuracy assessment.

The 500 m bands were first fused with the 250 m bands using ATPRK and the 250 m fused results for site 1 are shown in Figs. 2(c) and 2(d). It is clear that the 250 m fused images present more detailed information than the original 500 m images and clearer boundaries can be observed. Using bicubic interpolation, both 500 m and 250 m MODIS images were then downscaled to 30 m, which was used as the inputs of STARFM, as shown in Figs. 2(e)-2(h). Again, the interpolation results produced from the 250 m images are visually more informative than those from the 500 m data. Based on STARFM, the 30 m Landsat predictions of the two different schemes (i.e., using the original 500 m and 250 fused MODIS images) are shown in Figs. 2(k) and 2(l). The results for site 2 are displayed in Fig. 3.

The two blue patches representing abrupt changes (i.e., deforestation) from 13 June 2001 and 21 July 2003 in Fig. 2 were used for comparison of the two different schemes. Obviously, using the 250 m images, the two patches were more satisfactorily restored than those from the 500 m images where the patches appear as circle-like shapes. This is because the boundaries can be characterized by more pixels in the 250 m image in Fig. 2(d) than in the 500 m image in Fig. 2(b). When interpolated to 30 m, the 250 m neighboring pixels can provide more support than the 500 m neighboring pixels. As shown in Fig. 2(h), the 30 m interpolation results of the two patches are more accurate than those in Fig. 2(f). Based on the more reliable 30 m interpolation results in Figs. 2(g) and 2(h), a more accurate spatio-temporal fusion result was produced in Fig. 2(l). The advantages of the proposed scheme of using 250 m images can be also be observed clearly in the result for site 2, where abrupt changes in Fig. 3(f) are more accurately restored than in Fig. 3(e).

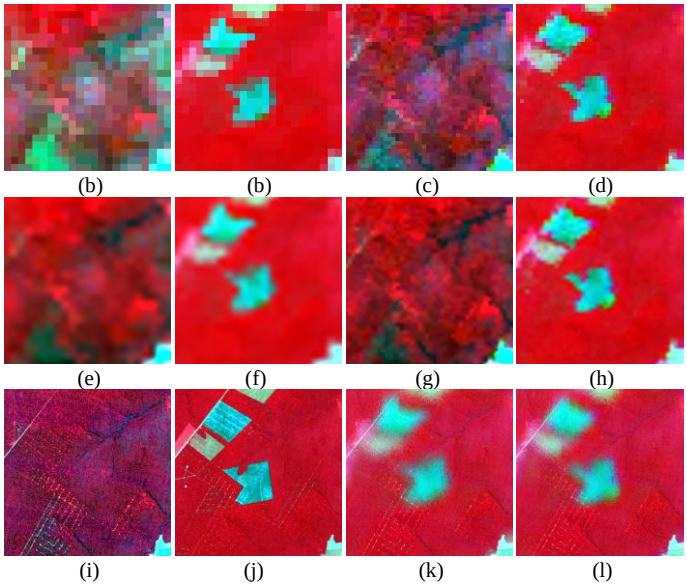

Fig. 2 The 30 m Landsat results for site 1 (10 km by 10 km) of the Amazon forest dataset (green, red and NIR bands as RGB). (a) and (b) are 500 m MODIS images on 13 June 2001 and 21 July 2003. (c) and (d) are 250 m fused MODIS images on 13 June 2001 and 21 July 2003. (e) and (f) are the 30 m bicubic interpolation results of (a) and (b). (g) and (h) are the 30 m bicubic interpolation results of (c) and (d). (i) and (j) are the 30 m Landsat 5 images on 13 June 2001 and 21 July 2003. (k) and (l) are the 30 m STARFM-derived Landsat images on 21 July 2003 using the original 500 m MODIS images and the 250 m fused MODIS images (with the aid of the MODIS-Landsat image pair on 13 June 2001), respectively.

Quantitative assessment of the 30 m results for the two sites is exhibited in Table 1. Three indices are used, including the root mean square error (RMSE), correlation coefficient (CC), and universal image quality index (UIQI) [25]. The ideal values for RMSE, CC and UIQI are 0, 1 and 1, respectively. For all six bands, the 250 m fused images lead to a smaller RMSE and larger CC and UIQI. It should be noted that the accuracy gains of using 250 m images for the red and NIR bands are larger than for the other four bands. For example, for site 2, the CCs of the red and NIR band increase by 0.033 and 0.045, but increase by around 0.01, 0.02, 0.01 and 0.01 for the blue, green, SWR1 and SWR2 bands. The reason is that the red and NIR bands are available at 250 m in the MODIS products, but the other four 250 m bands were produced by downscaling where uncertainties exist. Regarding the mean of the CC, the gains of using the 250 m images are 0.015 and 0.02 for sites 1 and 2, respectively.

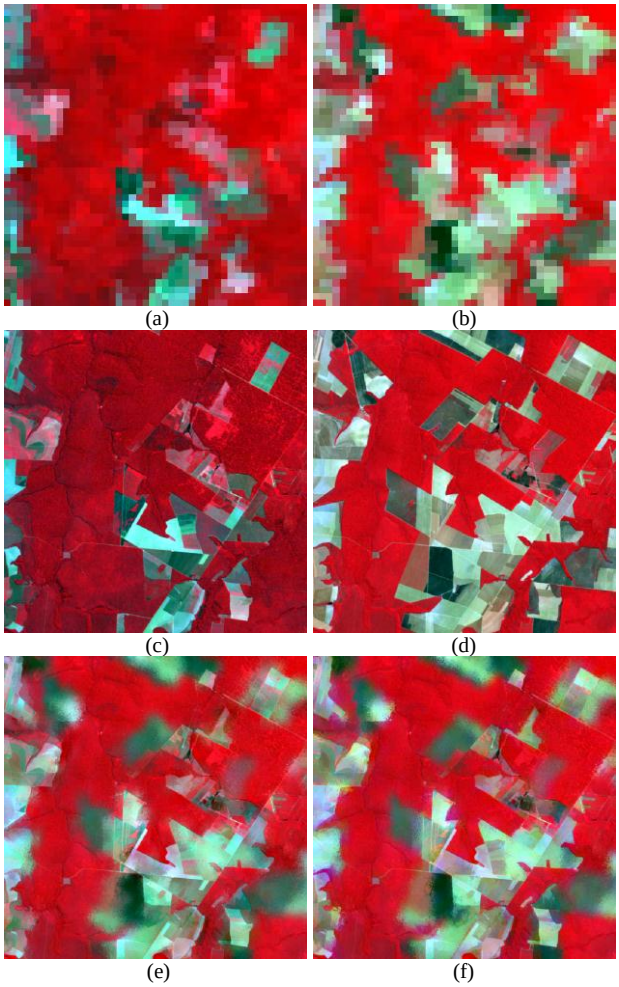

Fig. 3 The 30 m Landsat results for site 2 (20 km by 20 km) of the Amazon forest dataset (green, red and NIR bands as RGB). (a) and (b) are 500 m MODIS images on 13 June 2001 and 21 July 2003. (c) and (d) are the 30 m Landsat images on 13 June 2001 and 21 July 2003. (e) and (f) are the 30 m STARFM-derived Landsat 5 images on 21 July 2003 using the original 500 m MODIS images and the 250 m fused MODIS images (with the aid of the MODIS-Landsat image pair on 13 June 2001), respectively.

Table 1 Quantitative assessment for the Amazon forest dataset based on the STARFM method

	Bands	Site 1		Site 2	
		500 m MODIS	250 m MODIS	500 m MODIS	250 m MODIS
RMSE	Blue	0.0075	**0.0074**	0.0097	**0.0094**
	Green	0.0085	**0.0083**	0.0142	**0.0135**
	Red	0.0123	**0.0113**	0.0230	**0.0190**
	NIR	0.0280	**0.0264**	0.0315	**0.0282**
	SWIR1	0.0212	**0.0208**	0.0404	**0.0384**
	SWIR2	0.0228	**0.0223**	0.0451	**0.0437**
	Mean	0.0167	**0.0161**	0.0273	**0.0253**
CC	Blue	0.7424	**0.7509**	0.8468	**0.8578**
	Green	0.7499	**0.7601**	0.8728	**0.8881**
	Red	0.8293	**0.8562**	0.9038	**0.9362**
	NIR	0.8107	**0.8362**	0.7992	**0.8442**
	SWIR1	0.8316	**0.8401**	0.9069	**0.9163**
	SWIR2	0.8556	**0.8636**	0.8819	**0.8908**
	Mean	0.8032	**0.8178**	0.8686	**0.8889**
UIQI	Blue	0.7037	**0.7147**	0.8408	**0.8525**
	Green	0.7080	**0.7222**	0.8461	**0.8640**
	Red	0.8215	**0.8517**	0.8994	**0.9334**
	NIR	0.8101	**0.8360**	0.7979	**0.8439**
	SWIR1	0.8245	**0.8346**	0.9021	**0.9128**
	SWIR2	0.8506	**0.8597**	0.8791	**0.8883**
	Mean	0.7864	**0.8031**	0.8609	**0.8825**

B. Experiment on the Coleambally dataset

Three pairs of Landsat 8 and MODIS surface reflectance images covering a 28 km by 28 km area in Coleambally, Australia were used in this experiment. They were acquired on 6 July 2013, 14 August 2013 and 8 September 2013, respectively. For the MODIS data, the eight-day composite surface reflectance products MOD09A1 (500 m) and MOD09Q1 (250 m) were used. The three pairs of images are shown in Fig. 4. We predicted the 30 m Landsat image from the MODIS image on 14 August 2013, based on the availability of the two MODIS-Landsat image pairs on 6 July 2013 and 8 September 2013. The ESTARFM method was also performed as a benchmark method in this experiment. The true 30 m Landsat image on 14 August 2013 was used as a reference.

Fig. 5 shows the 30 m Landsat predictions on 14 August 2013 based on the two methods (STARFM and ESTARFM) coupled with two schemes (using 500 m and 250 m MODIS images). The results for two 1.5 km by 1.5 km heterogeneous sub-areas (marked in yellow in Fig. 5(a)) are shown in Fig. 6 to facilitate visual comparison. Using the 250 m fused images, the 30 m results produced for all two sub-areas are much closer to the references than those produced using the original 500 m images. For example, in sub-area S1, for both STARFM and ESTARSM, some green pixels were incorrectly predicted as blue pixels using the 500 images, but were restored adequately using the 250 m images. The main reason for the phenomenon is that the size of the patches of interest is smaller than one 500 m pixel. They cannot be reproduced accurately when bicubic interpolation is performed on the 500 m pixels (where support from the neighboring 500 m pixels is limited), and 250 m pixel-based interpolation is more reliable as more neighboring 250 m pixels are available to support the interpolation. In addition, amongst all predictions, STARFM with the 250 m image produces the result closest to the reference.

Table 2 lists the RMSE, CC and UIQI for the entire study area. The quantitative assessment also supports the findings of visual inspection. First, the 250 m images can produce fusion results with a smaller RMSE and larger CC and UIQI than the original 500 m images. More precisely, the gains in mean CC are 0.008 and 0.015 for STARFM and ESTARSM, respectively. For STARFM, the increase in UIQI for the blue, green, red, NIR, SWR1 and SWR2 bands is 0.010, 0.005, 0.024, 0.029, 0.010 and 0.006, respectively. Similarly to the results in Table 1, the accuracy gains for the red and NIR bands are larger than for the other four bands. Second, the accuracy of STARFM is greater than that of ESTARFM in both schemes. As mentioned in [26], for areas where temporal variance is dominant (as in the study area here), STARFM is more suitable. ESTARFM is more suitable for areas dominated by spatial variance.

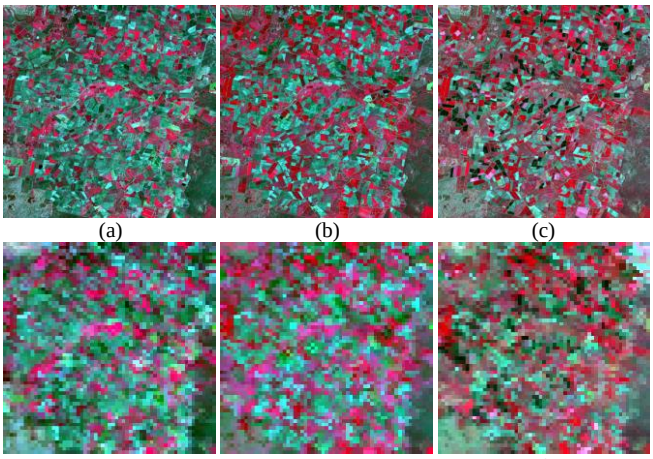

Fig. 4 The 30 m Landsat 8 and 500 m MODIS images for the Coleambally dataset (green, red and NIR bands as RGB). (a), (b) and (c) are 30 m Landsat 8 images on 6 July 2013, 14 August 2013 and 8 September 2013. (d)-(f) are the corresponding MODIS images for (a)-(c).

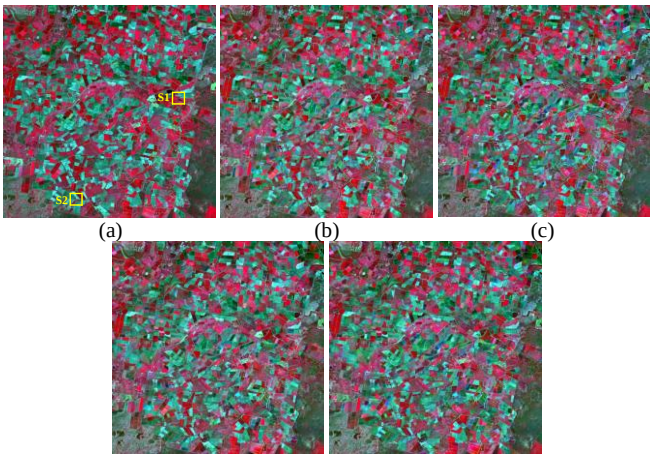

Fig. 5 The 30 m Landsat 8 results for the Coleambally dataset (green, red and NIR bands as RGB). (a) is the 30 m true Landsat 8 image on 14 August 2013 (the marked yellow sub-areas are used for analysis in Fig. 6). (b) and (c) are the 30 m ESTARFM-derived Landsat 8 images on 14 August 2013 using the original 500 m MODIS images and the 250 m fused MODIS images (with the aid of the MODIS-Landsat image pairs on 6 July 2013 and 8 September 2013), respectively. (d) and (e) are the 30 m STARFM-derived Landsat 8 images on 14 August 2013 using the original 500 m MODIS images and the 250 m fused MODIS images, respectively.

Table 2 Quantitative assessment for the Coleambally dataset

	Bands	ESTARFM		STARFM	
		500 m MODIS	250 m MODIS	500 m MODIS	250 m MODIS
RMSE	Blue	0.0101	**0.0099**	0.0085	**0.0083**
	Green	0.0113	**0.0111**	0.0104	**0.0103**
	Red	0.0189	**0.0177**	0.0173	**0.0160**

	NIR	0.0479	**0.0415**	0.0519	**0.0459**
	SWIR1	0.0483	**0.0476**	0.0383	**0.0375**
	SWIR2	0.0425	**0.0420**	0.0337	**0.0331**
	Mean	0.0298	**0.0283**	0.0267	**0.0252**
CC	Blue	0.8011	**0.8043**	0.8217	**0.8315**
	Green	0.7670	**0.7692**	0.7857	**0.7901**
	Red	0.8075	**0.8327**	0.8285	**0.8532**
	NIR	0.8923	**0.9197**	0.8883	**0.9162**
	SWIR1	**0.6851**	0.6824	0.7236	**0.7339**
	SWIR2	**0.7939**	0.7880	0.8023	**0.8098**

	Mean	0.7911	**0.7994**	0.8084	**0.8225**
UIQI	Blue	0.7878	**0.7929**	0.8152	**0.8249**
	Green	0.7659	**0.7682**	0.7855	**0.7898**
	Red	0.8026	**0.8264**	0.8245	**0.8483**
	NIR	0.8909	**0.9189**	0.8817	**0.9101**
	SWIR1	**0.6768**	0.6748	0.7209	**0.7306**
	SWIR2	**0.7644**	0.7618	0.7904	**0.7967**
	Mean	0.7814	**0.7905**	0.8030	**0.8167**

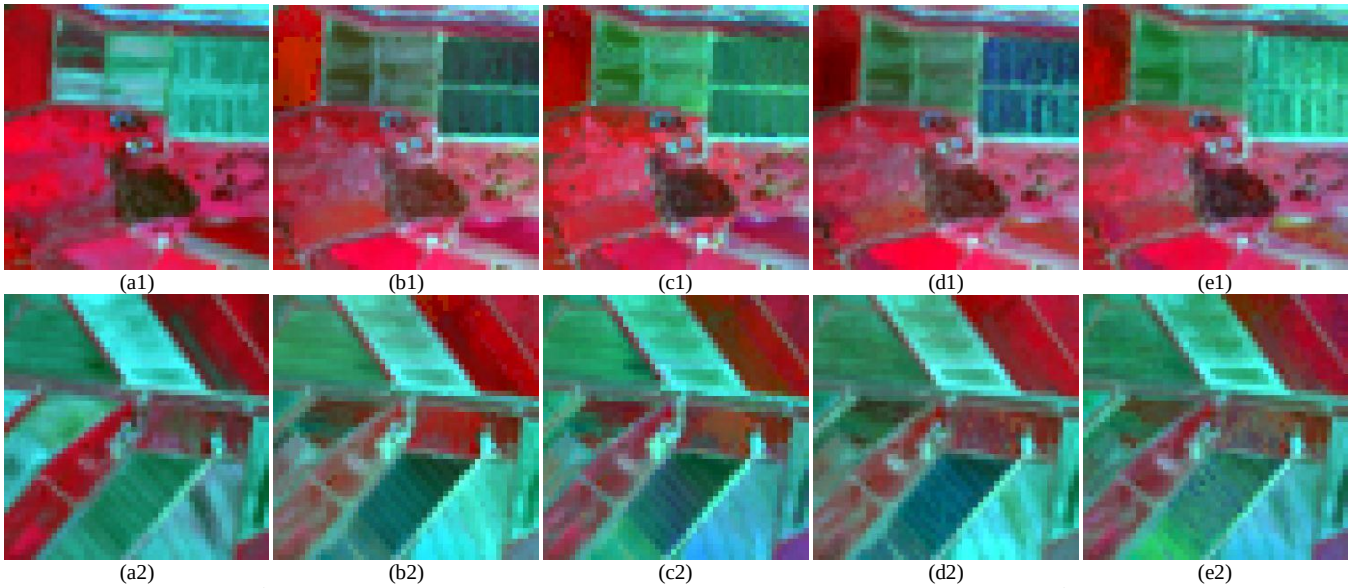

Fig. 6 The 30 m Landsat 8 results for the two heterogeneous sub-areas marked in yellow in Fig. 5(a). Lines 1 and 2 are the results for sub-areas S1 and S2 marked in Fig. 5(a). (a)-(e) have the same meanings as in Fig. 5.

IV. Discussion

In this paper, the freely available 250 m MODIS bands 1 and 2 were used as additional information for enhancement of spatio-temporal fusion of MODIS and Landsat data. The experiments in Section III show that the proposed scheme of incorporating 250 m MODIS images is able to reproduce more abrupt temporal changes (e.g., the two blue patches in Fig. 2(j)) and heterogeneous landscapes (Fig .6). Abrupt changes can be captured more adequately using the 250 m images on the prediction day than using the 500 m images alone, thus, providing more support in bicubic interpolation and post-spatio-temporal fusion. The proposed solution to enhance spatio-temporal fusion of MODIS and Landsat data has several advantages.

1) The 250 m data has little data cost as such data are totally free. Moreover, the 250 m data are acquired at exactly the same time as the 500 m data, and are available for any 500 m MODIS images that need to be downscaled.
2) The ATPRK-based image fusion approach is accurate [23], [24], easy to implement and can be automated and built into the well-known STARFM software straightforwardly. The approximate version of ATPRK greatly expedites the process of fusion of 500 m and 250 m MODIS images (the computational costs for the two datasets in the experiments are only several seconds).

These advantages facilitate technically the incorporation of 250 m data and enhancement of spatio-temporal fusion. Thus, the proposed scheme has great value in practice.

By incorporating the 250 m information, the spatio-temporal fusion of MODIS and Landsat data is actually divided into a two-step fusion process: from 500 m to 250 m and from 250 m to 30 m. In the traditional spatio-temporal fusion process, this first step (i.e., from 500 m to 250 m) is accomplished mainly by simple interpolation without any additional information. In the proposed scheme, information from the 250 m bands 1 and 2 is borrowed. The adopted image fusion approach (i.e., ATPRK in this paper) is more reliable than bicubic interpolation, which can undoubtedly increase the accuracy of this first step in the traditional spatio-temporal fusion process. However, it should be noted that the advanced ATPRK approach has uncertainties. For example, the regression model built from the coarse images might not be universal for fine spatial resolution images. Such uncertainties in incorporating the 250 m images motivate the development of more reliable image fusion approaches for further enhancement of spatio-temporal fusion of MODIS and Landsat data in future research.

In this paper, the popular STARFM method was considered as the fundamental spatio-temporal fusion approach, as it does not require strong assumptions of a stable land-cover/land-use distribution and can be performed using only one MODIS-Landsat image pair (rather than at least two pairs in ESTARFM). It would be interesting to consider incorporating 250 m images using other spatio-temporal fusion approaches, such as spatial unmixing-based methods, particularly for areas experiencing no (or very few) land-cover/land-use changes. Compared to another image-pair-based approach ESTARFM, STARFM is more suitable for cases where temporal variance

was dominant [26]. In the experiment on the Coleambally dataset, the temporal variance was dominant for the study area and, thus, STARFM produced greater accuracy. It would be worthwhile to consider using 250 m data for ESTARFM in cases where spatial variance was dominant.

The scheme of incorporating 250 m MODIS data was investigated for enhancing the fusion of 500 m MODIS and 30 m Landsat data in this paper. MODIS data can also be fused with other fine spatial resolution (but coarse temporal resolution) data, such as SPOT and the new Sentinel-2 data [27], to produce daily, fine spatial resolution data for global monitoring. In these new spatio-temporal fusion problems, the proposed scheme of incorporating 250 m MODIS data would also have great value in increasing the accuracy.

The 250 m data incorporation (i.e., ATPRK) and spatio-temporal fusion (i.e., STARFM) are performed separately in the proposed methodology. The uncertainty from the first step can be propagated directly to the second step. It is not clear how such uncertainty will affect the final predicted 30 m Landsat data. In future research, it would be interesting to develop a one-step spatio-temporal fusion approach to integrate the 500 m, 250 m and 30 m data in a single process, where one might expect to be able to gain greater control over the propagation of uncertainty (i.e., the uncertainty from processing the three different datasets can be controlled jointly).

The Landsat sensors (i.e., Landsat 7 and 8) can provide a 15 m panchromatic (PAN) band. The 15 m PAN band can be fused with the 30 m multispectral bands to produce 15 m pan-sharpened Landsat images [28]. The 15 m information is helpful for mapping of small objects, such as residential buildings and roads in urban areas. It would be of great interest to downscale the MODIS data further to 15 m based on the availability of the 15 m pan-sharpened Landsat images on temporally close days. As the target fine spatial resolution increases, however, uncertainty always increases as well. On the one hand, the 15 m available Landsat images on temporally close days were not observed in real data, but produced by pan-sharpening, a process which unavoidably introduces uncertainty. On the other hand, compared to the 30 m target resolution, 15 m means that a larger number of sub-pixel reflectance values need to be predicted and a larger solution space is involved. Thus, the reliability of the spatio-temporal fusion process may decrease. This issue provides a promising avenue for future research.

V. Conclusion

In this paper, to reduce the uncertainty inherent in the spatio-temporal fusion of 500 m MODIS and 30 m Landsat data, the freely available 250 m images acquired by MODIS were incorporated into the fusion process. The advanced ATPRK approach was used to fuse the 250 m bands with the 500 m bands to produce the interim 250 m seven-band MODIS data. The popular STARFM approach was then applied to fuse the 250 m MODIS and 30 m Landsat data. The ATPRK-based sharpening step makes full use of the 250 m observed MODIS data and yields a reliable interim product for input to the post-spatio-temporal fusion step. The 250 m fused images present more detailed spatial information than the original 500 m images (see Figs. 2(a)-2(d)), which is beneficial for restoration of abrupt changes and heterogeneous landscapes in spatio-temporal fusion. Experiments on two datasets show that the proposed scheme of incorporating 250 m MODIS images can produce more accurate spatio-temporal fusion results at 30 m resolution. It is especially advantageous in restoring abrupt temporal changes and heterogeneous landscapes.

The 250 m data are free and ATPRK can be automated and readily built into the well-known STARFM software. Therefore, the proposed scheme of incorporating 250 m data in spatio-temporal fusion has great potential application value and should see widespread adoption.

References

[1] S. Ganguly, M. A. Friedl, B. Tan, X. Zhang, and M. Verma. “Land surface phenology from MODIS: Characterization of the Collection 5 global land cover dynamics product,” *Remote Sensing of Environment*, vol. 114, no. 8, pp. 1805–1816, 2010.

[2] X. Y. Zhang, M. A. Friedl, C. B. Schaaf, A. H. Strahler, J. C. F. Hodges, F. Gao, B. C. Reed, and A. Huete. “Monitoring vegetation phenology using MODIS,” *Remote Sensing of Environment*, vol. 84, no. 3, pp. 471–475, 2003.

[3] M. C. Hansen, R. S. DeFries, J. R. G. Townshend, and R. Sohlberg. “Global land cover classification at the 1km spatial resolution using a classification tree approach,” *International Journal of Remote Sensing*, vol. 21, pp. 1331–1364, 2000.

[4] F. Gao, T. Hilker, X. Zhu, M. Anderson, J. G. Masek, P. Wang, and Y. Yang. “Fusing Landsat and MODIS data for vegetation monitoring,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, pp. 47–60, 2015.

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

[6] B. Chen, B. Huang, and Bing Xu. “Comparison of Spatiotemporal Fusion Models: A Review,” *Remote Sensing*, pp. 1798–1835, 2015.

[7] F. Gao, J. Masek, M. Schwaller, and F. Hall, “On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2207–2218, 2006.

[8] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek. “An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions,” *Remote Sensing of Environment*, vol. 114, pp. 2610–2623, 2010.

[9] T. Hilker, M. A. Wulder, N. C. Coops, J. Linke, J. McDermid, J. G. Masek, F. Gao, and J. C. White. “A new data fusion model for high spatial- and temporal-resolution mapping of forest based on Landsat and MODIS,” *Remote Sensing of Environment*, vol. 113, pp. 1613–1627, 2009.

[10] B. Huang and H. Song. “Spatiotemporal reflectance fusion via sparse representation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, pp. 3707–3716, 2012.

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

[12] R. Zurita-Milla, J. G. P. W. Clevers, and M. E. Schaepman. “Unmixing-based Landsat TM and MERIS FR data fusion,” *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 453–457, 2008.

[13] J. Amorós-López, L. Gómez-Chova, L. Alonso, L Guanter, R. Zurita-Milla, J. Moreno, and G. Camps-Valls. “Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 23, pp. 132–141, 2013.

[14] J. Amorós-López, L. Gómez-Chova, L. Alonso, L. Guanter, J. Moreno, and G. Camps-Valls. “Regularized multiresolution spatial unmixing for ENVISAT/MERIS and Landsat/TM image fusion,” *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 5, pp. 844–848. 2011.

[15] M. Wu, C. Wang, and L. Wang. “Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model,” *Journal of Applied Remote Sensing*, vol. 6, 2012.

[16] C. M. Gevaert and F. J. García-Haro. “A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion,” *Remote Sensing of Environment*, vol. 156, pp. 34–44, 2015.

[17] Y. T. Mustafa, V. A. Tolpekin, and A. Stein. “Improvement of spatio-temporal growth estimates in heterogeneous forests using Gaussian

bayesian networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4980–4991, 2014.

[18] R. Zurita-Milla, G. Kaiser, J. G. P. W. Clevers, W. Schneider, and M. E. Schaepman. "Downscaling time series of MERIS full resolution data to monitor vegetation seasonal dynamics," *Remote Sensing of Environment*, vol. 113, 1874–1885, 2009.

[19] Y. Xu, B. Huang, Y. Xu, K. Cao, C. Guo, and D. Meng. "Spatial and temporal image fusion via regularized spatial unmixing," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 6, pp. 1362–1366, 2015.

[20] D. Xie, J. Zhang, X. Zhu, Y. Pan, H. Liu, Z. Yuan, and Y. Yun. "An improved STARFM with help of an unmixing-based method to generate high spatial and temporal resolution remote sensing data in complex heterogeneous regions," *Sensors*, vol. 16, 2016.

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

[22] H. Shen, L. Huang, L. Zhang, P. Wu, C. Zeng. "Long-term and fine-scale satellite monitoring of the urban heat island effect by the fusion of multi-temporal and multi-sensor remote sensed data: A 26-year case study of the city of Wuhan in China," *Remote Sensing of Environment*, vol. 172, pp. 109–125, 2016.

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

[24] Q. Wang, W. Shi, and P. M. Atkinson, "Area-to-point regression kriging for pan-sharpening," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 151–165, 2016.

[25] Z. Wang and A. C. Bovik. "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, 2002.

[26] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. J. M. van Dijk. "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*, vol. 133, pp. 193–209, 2013.

[27] M. Drusch, *et al*. "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012.

[28] B. Chen, B. Huang, and B. Xu. "Fine land cover classification using daily synthetic Landsat-like images at 15-m resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 12, pp. 2359–2363, 2015.