Virtual image pair-based spatio-temporal fusion

Qunming Wang [a], Yijie Tang [a], Xiaohua Tong [a, *], Peter M. Atkinson [b, c]

[a] College of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road, Shanghai 200092, China

[b] Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YR, UK

[c] Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK

*Corresponding author. E-mail: xhtong@tongji.edu.cn

Abstract: Spatio-temporal fusion is a technique used to produce images with both fine spatial and temporal resolution. Generally, the principle of existing spatio-temporal fusion methods can be characterized by a unified framework of prediction based on two parts: (i) the known fine spatial resolution images (e.g., Landsat images), and (ii) the fine spatial resolution increment predicted from the available coarse spatial resolution increment (i.e., a downscaling process), that is, the difference between the coarse spatial resolution images (e.g., MODIS images) acquired at the known and prediction times. Owing to seasonal changes and land cover changes, there always exist large differences between images acquired at different times, resulting in a large increment and, further, great uncertainty in downscaling. In this paper, a virtual image pair-based spatio-temporal fusion (VIPSTF) approach was proposed to deal with this problem. VIPSTF is based on the concept of a virtual image pair (VIP), which is produced based on the available, known MODIS-Landsat image pairs. We demonstrate theoretically that compared to the known image pairs, the VIP is closer to the data at the prediction time. The VIP can capture more fine spatial resolution information directly from known images and reduce the challenge in downscaling. VIPSTF is a flexible framework suitable for existing spatial weighting- and spatial unmixing-based methods, and two versions VIPSTF-SW and VIPSTF-SU are, thus, developed. Experimental results on a heterogeneous site and a site experiencing land cover type changes show that both spatial weighting- and spatial unmixing-based methods can be enhanced by VIPSTF, and the

advantage is particularly noticeable when the observed image pairs are temporally far from the prediction time. Moreover, VIPSTF is free of the need for image pair selection and robust to the use of multiple image pairs. VIPSTF is also computationally faster than the original methods when using multiple image pairs. The concept of VIP provides a new insight to enhance spatio-temporal fusion by making fuller use of the observed image pairs and reducing the uncertainty of estimating the fine spatial resolution increment.

Keywords: Virtual image pair (VIP), Spatio-temporal fusion, Downscaling, Time-series images.

1. Introduction

Remote sensing satellite sensor data for the globe have been applied in many areas, such as land cover change monitoring (Dyer, 2012), vegetation monitoring (Shen et al., 2011) and ecological evaluation (Pisek et al., 2015). Among the satellite sensors, the Landsat series (e.g., Thematic Mapper (TM), Enhanced Thematic Mapper (ETM+), Operational Land Imager (OLI)) and the Terra/Aqua MODerate resolution Imaging Spectroradiometer (MODIS) are perhaps the most commonly used due to their regular revisit capabilities, wide swath and free availability. Normally, there is a trade-off between spatial and temporal resolutions. The Landsat sensors can acquire images at a fine spatial resolution of 30 m, but they have a revisit period of up to 16 days. Moreover, due to cloud contamination, the effective temporal resolution is much coarser (e.g., only a few useable Landsat images are available per year). On the contrary, MODIS can acquire images for the same scene at least once per day, but the images are at a coarse spatial resolution of 500 m. To meet the demand of timely, fine spatial resolution monitoring, spatio-temporal fusion methods have been developed to blend the available temporally sparse fine spatial resolution images and temporally dense coarse spatial resolution images to create time-series with both fine spatial and temporal resolutions (Belgiu and Stein, 2019; Chen et al., 2015; Gao et al., 2015; Zhang et al., 2015; Zhu et al., 2018). Generally, three main categories of spatio-temporal fusion methods can be identified: spatial weighting-based, spatial unmixing-based and hybrid methods.

The spatial and temporal adaptive reflectance fusion model (STARFM) (Gao et al., 2006) is one of the earliest and the most commonly applied spatial weighting-based methods. STARFM predicts the reflectance of fine spatial resolution pixels based on a linear weighting of the reflectances of spatially surrounding similar pixels. The similar pixels in the neighborhood are selected according to their spectral similarity with the center pixel. STARFM is more effective for homogeneous landscapes and areas with stable land cover during the period of interest. The spatial temporal adaptive algorithm for mapping reflectance change (STAARCH) increased the accuracy of spatio-temporal fusion for areas experiencing land cover change (i.e., forest disturbance) by introducing a disturbance factor to quantify the reflectance change in Landsat images (Hilker et al., 2009). To increase the accuracy for heterogeneous regions, an enhanced spatial and temporal adaptive reflectance fusion model (ESTARFM) was proposed by introducing a conversion coefficient to characterize the linear relationship between the changes in MODIS and Landsat reflectances (Zhu et al., 2010). ESTARFM was advantageous for reproducing small and linear targets. Wang and Atkinson (2018) introduced a Fit-FC method to deal with strong seasonal changes in spatio-temporal fusion. These spatial weighting-based methods have been applied widely to predict land surface temperature (LST) (Huang et al., 2013; Shen et al., 2016; Weng et al., 2014; Wu et al., 2015), leaf area index (Houborg et al., 2016; Zhang et al., 2014), and normalized difference vegetation index (NDVI) (Meng et al., 2013; Tewes et al., 2015) at both fine spatial and temporal resolutions.

Spatial unmixing-based methods are generally performed based on a coarse image at the prediction time and a land cover classification map produced from the known fine spatial resolution data (e.g., multispectral images at the target fine spatial resolution (Amorós-López et al., 2013; Gevaert et al., 2015; Zurita-Milla et al., 2008), and aerial image (Mustafa et al., 2014) or land-use database (Zurita-Milla et al., 2009) at the finer spatial resolution). Based on the assumption that the land cover does not change during a given period, the fine spatial resolution land cover map at known time is upscaled to characterize the coarse proportions of land cover classes at the prediction time. The representative reflectance of each land cover class within a coarse pixel can be predicted inversely from the coarse proportions and observed coarse reflectance. The multisensor multiresolution technique (MMT) proposed by Zhukov et al. (1999) is one of the first spatial unmixing-based methods. MMT assigns the predicted land cover class reflectance directly to a fine spatial resolution pixel according to its corresponding class. Busetto et al. (2008) considered both spatial and spectral differences for weighting the contributions of neighboring coarse pixels in the spatial unmixing model. To avoid large deviations of the predicted reflectance of each class, Amorós-López et al. (2013) introduced a new regularization term to the objective function in the spatial unmixing model, where the difference between the class reflectances at target fine and observed coarse spatial resolutions is minimized. The spatial-temporal data fusion approach (STDFA) calculated the temporal change in reflectance for each class by unmixing the coarse difference images. The predicted temporal change at fine spatial resolution is then added to the known fine spatial resolution image (Wu et al., 2012). Gevaert and García-Haro (2015) applied a Bayesian solution to constrain the fine spatial resolution reflectance in the unmixing model.

Hybrid methods combining the mechanisms of the above two categories of methods have also been developed. The Flexible Spatiotemporal DAta Fusion (FSDAF) method estimates the temporal change of each class by spatially unmixing the coarse difference images, and then distributing the residuals estimated from thin plate spline (TPS) interpolation based on spatial weighting of neighboring similar pixels (Zhu et al., 2016). Liu et al. (2019) proposed an improved FSDAF (IFSDAF) for producing NDVI time-series with both fine spatial and temporal resolutions. Instead of distributing the residuals entirely based on the TPS interpolation result (i.e., space-dependent increment), IFSDAF also considers temporally-dependent increment by spatial unmixing. To enhance the performance for restoration of land cover change, an enhanced FSDAF that incorporates sub-pixel class fraction change information (SFSDAF) was proposed by Li et al. (2020). SFSDAF accounts for the changes in class reflectance and proportions jointly in the spatial unmixing model. Xu et al. (2015) performed spatial weighting based on STARFM before spatial unmixing, where the STARFM prediction is used to construct a regularization term to avoid large deviations of predicted class reflectances.

Apart from the methods mentioned above, Bayesian-based methods (Li et al., 2013) and learning-based methods (Das and Ghosh, 2016; Huang and Song, 2012; Liu et al., 2016) have also been developed.

Although the specific mechanisms of the spatio-temporal fusion methods vary, the methods can be summarized by a unified framework

$$\hat{\mathbf{L}}(\text{t\_predict}) = \mathbf{L}(\text{t\_known}) + \Delta\mathbf{L} \tag{1}$$

$$\Delta\mathbf{L} = f(\Delta\mathbf{M}). \tag{2}$$

Eq. (1) indicates that the prediction of the Landsat image at the prediction time is divided into two parts; the known Landsat image $\mathbf{L}(\text{t\_known})$ and the unknown Landsat level increment $\Delta\mathbf{L}$ (Liu et al., 2019). Note that multiple known Landsat images (i.e., multiple MODIS-Landsat image pairs are available) can also be included in the term $\mathbf{L}(\text{t\_known})$, which is then a combination of the multiple Landsat images correspondingly. The first part makes use of available fine spatial resolution information directly, while the second part predicts fine spatial resolution information from the available coarse spatial resolution data. As seen from Eq. (2), the estimation of $\Delta\mathbf{L}$ depends on MODIS level increment $\Delta\mathbf{M}$, which is the difference between the MODIS images at the known and prediction times. Obviously, the estimation of $\Delta\mathbf{L}$ is the most pivotal issue: this involves downscaling, the quality of which exerts a direct influence on the accuracy of prediction. The function f (i.e., the downscaling operator) differs according to the specific spatio-temporal fusion method. For spatial weighting-based methods, f is usually a linear weighting function (Gao et al., 2006; Zhu et al., 2010), while for spatial unmixing-based methods, f is a linear unmixing model (Amorós-López et al., 2013; Zhukov et al., 1999). No matter which method is adopted, a smaller increment $\Delta\mathbf{M}$ will definitely decrease the uncertainty in estimating $\Delta\mathbf{L}$. To reduce the error produced by estimation of $\Delta\mathbf{L}$ and produce a greater accuracy for spatio-temporal fusion, it is important to minimize $\Delta\mathbf{M}$. One possible solution is to acquire MODIS-Landsat image pairs as temporally close to the prediction time as possible. Due to cloud and shadow contamination, however, the number of available high-quality Landsat images is always limited (Ju and Roy, 2008). Thus, it can be challenging to acquire image pairs that are sufficiently close to the prediction time; that is, it is always difficult to decrease $\Delta\mathbf{M}$ just from the perspective of using data.

Alternatively, another possible solution to reduce $\Delta\mathbf{M}$ is to perform transformations to the known MODIS images based on an identified model. As acknowledged widely, there exists a corresponding relationship between the Landsat and MODIS images acquired at the same time. Suppose the zoom factor between the MODIS and Landsat images is s such that the reflectance of each MODIS pixel can be regarded as the average of the reflectance of s^2 Landsat pixels covering the same area. Preserving this relationship, the transformation applied to known Landsat images can be linked to that of the MODIS images. Inspired by this, in this paper we introduced the concept of the virtual image pair (VIP), that is, the synthesization of a MODIS-Landsat image pair closer to that at the prediction time (i.e., with a smaller $\Delta\mathbf{M}$) than the original observed MODIS-Landsat image pairs. When the VIP is adopted, the input of the function f in Eq. (2) will become smaller, thus, reducing the burden of estimating $\Delta\mathbf{L}$. Actually, in this case, the final prediction is dependent on the new 'known' Landsat image (i.e., the virtual Landsat image) to a larger extent than existing methods, which is closer to the Landsat image to be predicted and can capture more fine spatial resolution information directly from the observed Landsat images.

In this paper, based on the concept of VIP, a VIP-based spatio-temporal fusion (VIPSTF) approach is proposed. VIPSTF produces the VIP based on the observed MODIS-Landsat image pairs that may have a considerable temporal distance to the prediction time. The new MODIS level increment is downscaled by the function f in Eq. (2) to predict the new Landsat level increment. As mentioned above, f varies when different methods are used. For the proposed VIPSTF approach, both spatial weighting- and spatial unmxing-based methods can be incorporated into it. Specifically, the popular STARFM (Gao et al., 2006) and STDFA (Wu et al., 2012) methods are adopted to characterize the function f in VIPSTF in this paper. VIPSTF can reduce the difference between MODIS images at the known and prediction times effectively, reducing the burden in estimation of the Landsat level increment and finally leading to greater prediction accuracy.

The remainder of this paper is organized into four sections. In Section 2, the relation between the MODIS and Landsat images in the VIP is first deduced in Section 2.1. Section 2.2 introduces the method to produce the VIP and demonstrates mathematically its validity in reducing $\Delta\mathbf{M}$. Furthermore, the proposed VIPSTF approach including both spatial weighting and spatial unmixing-based versions is introduced explicitly in Section 2.3. Section 3 presents the experimental results of VIPSTF and compares it with other spatio-temporal fusion methods. Section 4 discusses the main findings and the problems to be investigated further. Section 5 concludes the paper.

2. Methods

Similarly to most of existing spatio-temporal fusion methods, the proposed method is performed for each band separately. In this paper, for simplicity of mathematical expression, the principle is illustrated based on a single band of Landsat and MODIS images. The implementation can be applied to each band similarly.

2.1. Relation between Landsat and MODIS images in the virtual image pair (VIP)

In this paper, the VIP is proposed to decrease the difference between images acquired at the known time and prediction time, and further, to increase the accuracy of spatio-temporal fusion. The VIP is generated by combining the original known time-series images through a certain mathematical transformation. Suppose that we have N known MODIS-Landsat image pairs acquired at $t_1,\ldots, t_N$. The Landsat images are denoted as $\mathbf{L}_1,\ldots, \mathbf{L}_N$, while the MODIS images are denoted as $\mathbf{M}_1,\ldots, \mathbf{M}_N$. The functions g_1 and g_2 are applied to Landsat and MODIS time-series images to produce the VIP

$$\mathbf{L}_{\mathrm{VIP}} = g_1(\mathbf{L}_1,\ldots,\mathbf{L}_N) \tag{3}$$

$$\mathbf{M}_{\mathrm{VIP}} = g_2(\mathbf{M}_1, \ldots, \mathbf{M}_N) \tag{4}$$

where $\mathbf{L}_{\mathrm{VIP}}$ and $\mathbf{M}_{\mathrm{VIP}}$ are the virtual Landsat image and virtual MODIS image, respectively.

Suppose the zoom factor between the Landsat and MODIS images is s. The value (i.e., reflectance in this paper) of each MODIS pixel can generally be treated as the average of every s^2 Landsat pixel covering the same area at the same time (Li et al., 2020; Zhu et al., 2010). Based on this assumption, an intrinsic relation can be built between the corresponding Landsat and MODIS pixels for any MODIS-Landsat image pair

$$M(x_0, y_0) = \frac{1}{s^2} \sum_{i=1}^{s^2} L(x_{0i}, y_{0i}). \tag{5}$$

In Eq.(5), $M(x_0, y_0)$ is the value of the MODIS pixel located at (x_0, y_0), and $L(x_{0i}, y_{0i})$ is the value of the i th pixel of the s^2 Landsat pixels covering the same area as $M(x_0, y_0)$.

No matter which method is adopted to determine the two functions g_1 and g_2, it is always important to ensure consistency between the Landsat and MODIS images defined in Eq. (5). Accordingly, the corresponding pixels in $\mathbf{L}_{\mathrm{VIP}}$ and $\mathbf{M}_{\mathrm{VIP}}$ should satisfy the relationship as well, and the two functions can also be connected correspondingly. Specifically, according to Eqs. (3) and (5), we can simply characterize $\mathbf{M}_{\mathrm{VIP}}$ using g_1

$$M_{\mathrm{VIP}}(x_0, y_0) = \frac{1}{s^2} \sum_{i=1}^{s^2} L_{\mathrm{VIP}}(x_{0i}, y_{0i}) = \frac{1}{s^2} \sum_{i=1}^{s^2} g_1\left[L_1(x_{0i}, y_{0i}), \ldots, L_N(x_{0i}, y_{0i})\right]. \tag{6}$$

Suppose g_1 is a linear transformation function, the fixed coefficient $1/s^2$ can be applied to each Landsat pixel directly, that is, Eq. (6) can be rewritten as

$$\begin{aligned} M_{\mathrm{VIP}}(x_0, y_0) &= g_1\left[\frac{1}{s^2} \sum_{i=1}^{s^2} L_1(x_{0i}, y_{0i}), \ldots, \frac{1}{s^2} \sum_{i=1}^{s^2} L_N(x_{0i}, y_{0i})\right]. \\ &= g_1\left[M_1(x_0, y_0), \ldots, M_N(x_0, y_0)\right] \end{aligned} \tag{7}$$

When each pixel in the virtual MODIS image undergoes the same transformation in Eq. (7), the whole MODIS image can be represented as follows

$$\mathbf{M}_{\mathrm{VIP}} = g_1(\mathbf{M}_1, \ldots, \mathbf{M}_N)\ . \tag{8}$$

Comparing Eq. (8) with Eq. (4), it is clear that the function g_2 is the same as g_1. That is, the transformation applied to the MODIS time-series is consistent with that for the Landsat time-series. Note that such consistency exists based on the assumption of a linear transformation.

2.2. Production of the VIP

2.2.1 The specific form of the VIP

As mentioned in Section 2.1, the linear transformation is a feasible solution to produce the VIP and can relate the virtual Landsat and MODIS images effectively. Specifically, the transformation applied to the Landsat time-series to produce $\mathbf{L}_{\mathrm{VIP}}$ can be expressed explicitly as

$$\mathbf{L}_{\mathrm{VIP}} = g_1(\mathbf{L}_1, \ldots, \mathbf{L}_N) = \sum_{k=1}^{N} a_k \mathbf{L}_k + b \tag{9}$$

where a_k is the transformation coefficient for the kth image in the Landsat time-series and b is a constant. According to the consistency in linear transformation demonstrated above, the virtual MODIS image $\mathbf{M}_{\mathrm{VIP}}$ can be expressed similarly

$$\mathbf{M}_{\mathrm{VIP}} = g_1(\mathbf{M}_1, \ldots, \mathbf{M}_N) = \sum_{k=1}^{N} a_k \mathbf{M}_k + b\ . \tag{10}$$

In the linear transformation function, different coefficient sets (i.e., composed of a_k and b) will result in different VIPs. It is critical to develop a reliable scheme to estimate the coefficients appropriately. In this paper, the coefficient set is estimated based on the linear regression model fitted between the MODIS data at the known and prediction times

$$\mathbf{M}_p = \sum_{k=1}^{N} a_k \mathbf{M}_k + b + \mathbf{r}\ . \tag{11}$$

In Eq. (11), $\mathbf{r}$ is the residual image, and $\mathbf{M}_k$ and $\mathbf{M}_p$ are the kth known MODIS image and the MODIS at the prediction time, respectively. The coefficients a_k and b are obtained using the least squares method.

2.2.2 The rationale of the specific form

As the ultimate purpose of any definition of VIP is to reduce $\Delta\mathbf{M}$ (i.e., the virtual MODIS image needs to be closer to the MODIS image at the prediction time), the coefficient set should follow the key rule that the new $\Delta\mathbf{M}'$ between the virtual MODIS image and the MODIS image at the prediction time should be smaller than the original $\Delta\mathbf{M}$. To evaluate whether the coefficient set estimated by the regression model satisfies the rule, we need to quantify $\Delta\mathbf{M}$ and $\Delta\mathbf{M}'$ beforehand. The root mean square error (RMSE) is one of the most widely used indices to measure the statistical difference in the pixel values (i.e., reflectance in this paper) between two images. It is used to quantify $\Delta\mathbf{M}$ and $\Delta\mathbf{M}'$ in this paper. RMSE is defined as

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left[U(x_i,y_i)-V(x_i,y_i)\right]^2}=\sqrt{E[(\mathbf{U}-\mathbf{V})^2]} \tag{12}$$

where $\mathbf{U}$ and $\mathbf{V}$ represent two images composed of m pixels. Mathematically, the RMSE between two images equals the square root of the expectation of the square of the difference image $\mathbf{U}-\mathbf{V}$. Therefore, we can calculate the expectation of the square of $\Delta\mathbf{M}$ and $\Delta\mathbf{M}'$ (i.e., $E(\Delta\mathbf{M}^2)$ and $E(\Delta\mathbf{M}'^2)$) instead for their comparison.

For spatio-temporal fusion using *multiple* image pairs, the original $\Delta\mathbf{M}$ cannot be expressed simply as the difference between MODIS images. According to the general framework of spatio-temporal fusion summarized in the Introduction, prediction using multiple image pairs can be written as

$$\begin{aligned}\hat{\mathbf{L}}_p&=\sum_{i=1}^{N}w_i\left[\mathbf{L}_i+f(\mathbf{M}_p-\mathbf{M}_i)\right]\\&=\sum_{i=1}^{N}w_i\mathbf{L}_i+\sum_{i=1}^{N}w_i f(\mathbf{M}_p-\mathbf{M}_i)\end{aligned} \tag{13}$$

where w_i is the weight for the i th prediction and satisfies $\sum_{i=1}^{N} w_i = 1$. In Eq. (13), the prediction is divided into two parts. The first part $\sum_{i=1}^{N} w_i \mathbf{L}_i$ is known, while the second part, the weighted sum of $f(\mathbf{M}_p - \mathbf{M}_i)$, can be regarded as the increment term produced by multiple image pairs. The function f differs according to the used spatio-temporal fusion method, and usually a linear model can be adopted for its characterization (e.g., the linear weighting function in the spatial weighting-based methods and the linear unmixing model for spatial unmixing-based methods). In this case, the second part can be altered as

$$\begin{aligned}\sum_{i=1}^{N} w_i f(\mathbf{M}_p - \mathbf{M}_i) &= f\left[\sum_{i=1}^{N} w_i (\mathbf{M}_p - \mathbf{M}_i)\right]. \\ &= f(\Delta\mathbf{M})\end{aligned} \tag{14}$$

That is, $\Delta\mathbf{M}$ can be expressed as $\sum_{i=1}^{N} w_i (\mathbf{M}_p - \mathbf{M}_i)$ for fusion using multiple image pairs.

When the VIP is used, based on Eqs. (10) and (11), $\Delta\mathbf{M}'$ can be expressed as

$$\Delta\mathbf{M}' = \mathbf{M}_p - \mathbf{M}_{\mathrm{VIP}}. \tag{15}$$

To compare $E(\Delta\mathbf{M}^2)$ and $E(\Delta\mathbf{M}'^2)$, they are transformed individually, as presented in Appendix A. After derivation, $E(\Delta\mathbf{M}^2)$ and $E(\Delta\mathbf{M}'^2)$ can be expressed as

$$E(\Delta\mathbf{M}^2) = Var(\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i} \mathbf{M}_k) + Var(\mathbf{r}) + E^2\left[\sum_{i=1}^{N} w_i (\mathbf{M}_p - \mathbf{M}_i)\right] \tag{16}$$

$$E(\Delta\mathbf{M}'^2) = Var(\mathbf{r}). \tag{17}$$

Comparing Eq. (16) with Eq. (17), we can conclude that $E(\Delta\mathbf{M}'^2)$ is obviously smaller than $E(\Delta\mathbf{M}^2)$, suggesting that the produced VIP is closer to the data at the prediction time than that for conventional spatio-temporal fusion model. Furthermore, by setting the weight w_i for the i th known MODIS image in Eq. (16) as 1 (i.e., only the i th MODIS-Landsat image pair is used for fusion), we have

$$E(\Delta\mathbf{M}_i^{\,2}) = Var(\sum_{k=1}^{N} a_{k_i} \mathbf{M}_k) + Var(\mathbf{r}) + E^2(\mathbf{M}_p - \mathbf{M}_i). \tag{18}$$

It is clear that $E(\Delta\mathbf{M}_i{}^2)$ is still larger than $E(\Delta\mathbf{M}'^2)$. This means the VIP is closer to the data at the prediction time than *any* known image pair, thus, capturing more fine spatial resolution information directly from the known images. Therefore, it is feasible to use the regression model to estimate the coefficient set and produce the VIP.

2.3. VIP-based spatio-temporal fusion (VIPSTF)

According to the general framework in Eq. (13), the prediction of the Landsat image includes two parts: the linear superposition of known Landsat images and the increment computed by applying a function f to $\Delta\mathbf{M}$. When the VIP is introduced for spatio-temporal fusion, the framework in Eq. (13) is replaced by the proposed VIPSTF model as follows

$$\begin{aligned}\hat{\mathbf{L}}_p &= \mathbf{L}_{\mathrm{VIP}} + \Delta\mathbf{L}' \\ &= \mathbf{L}_{\mathrm{VIP}} + f(\Delta\mathbf{M}') \\ &= \mathbf{L}_{\mathrm{VIP}} + f(\mathbf{M}_p - \mathbf{M}_{\mathrm{VIP}})\end{aligned} . \quad (19)$$

The VIPSTF prediction is a combination of the produced $\mathbf{L}_{\mathrm{VIP}}$ and the Landsat level increment $\Delta\mathbf{L}'$. The increment $\Delta\mathbf{L}'$ is predicted by applying the function f to the MODIS level increment $\Delta\mathbf{M}'$.

As mentioned in the Introduction, there are two main types of methods to characterize f: one is spatial weighting (SW)-based and the other is spatial unmixing (SU)-based. In this paper, the popular STARFM and STDFA methods are considered as representative choices for SW and SU, respectively. We name the corresponding VIPSTF-based versions as VIPSTF-SW and VIPSTF-SU. The flowchart of the proposed VIPSTF approach (including both VIPSTF-SW and VIPSTF-SU versions) is shown in Fig. 1.

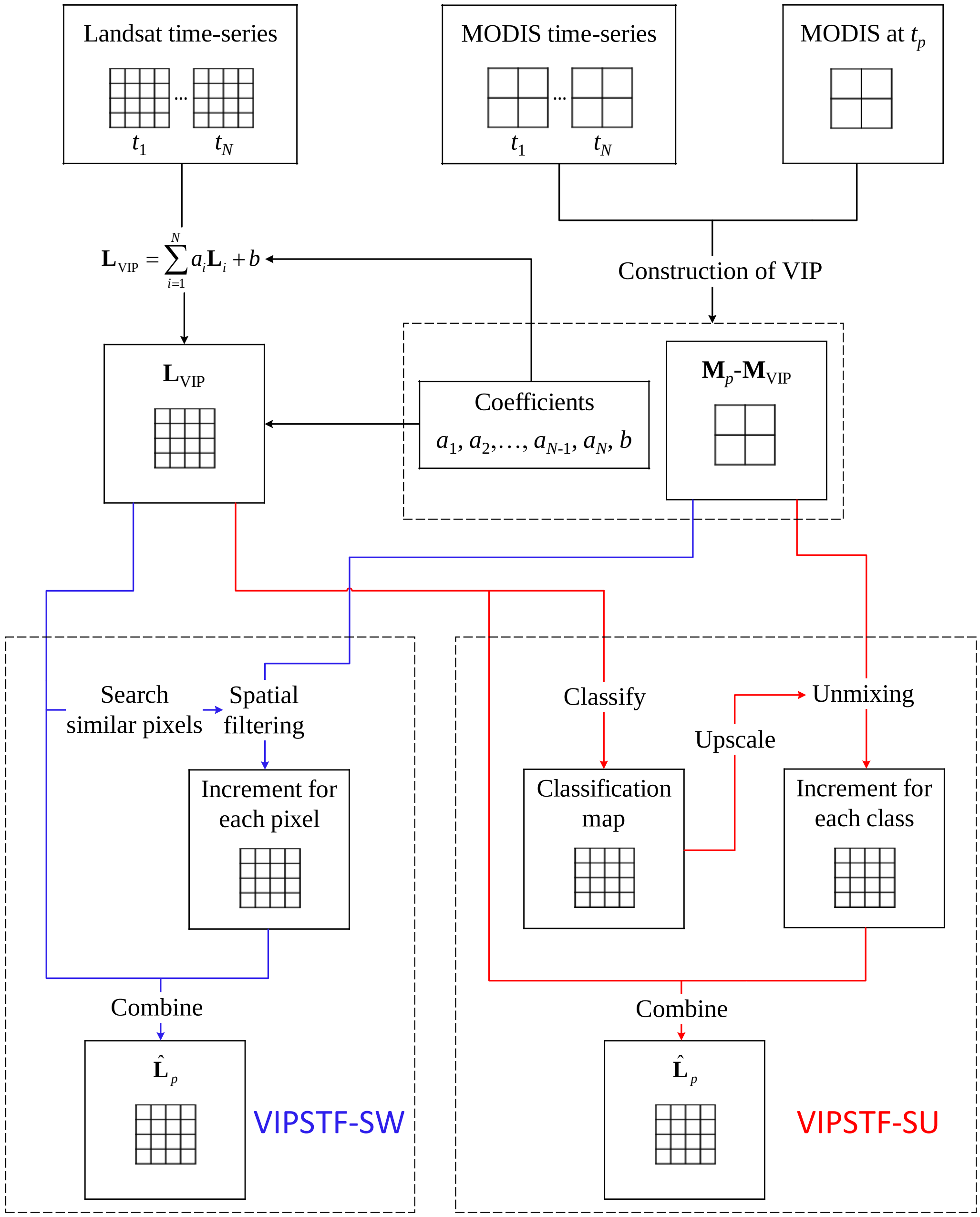


Fig. 1. Flowchart of VIPSTF, where both spatial weighting (SW)- and spatial unmixing (SU)-based solutions (i.e., VIPSTF-SW and VIPSTF-SU) are illustrated.

2.3.1 Spatial weighting-based VIPSTF (VIPSTF-SW)

In the proposed VIPSTF-SW method, a spatial weighting strategy is applied to predict the Landsat level increment $\Delta\mathbf{L}'$ from the MODIS level increment $\Delta\mathbf{M}'$, as shown in Eq. (20)

$$\Delta L'(x_0, y_0) = \sum_{i=1}^{n_s} \lambda_i \Delta M'(x_i, y_i) \tag{20}$$

where (x_i, y_i) is the spatial location of the similar pixels surrounding the pixel centered at (x_0, y_0), n_s is the number of similar neighboring pixels and λ_i is a weight assigned according to the distance between the center and similar pixels. Note that to match the spatial resolution of Landsat increment $\Delta\mathbf{L}'$, the MODIS increment $\Delta\mathbf{M}'$ needs to be interpolated (e.g., by bicubic interpolation) to the Landsat spatial resolution in advance. The similar pixels are searched according to the spectral difference between the center pixel and neighboring pixels in the virtual Landsat image $\mathbf{L}_{\mathrm{VIP}}$: the first n_s pixels with the smallest spectral difference are chosen as similar pixels in each local window. Eq. (20) means that the increment for the center Landsat pixel is determined as a linear combination of $\Delta\mathbf{M}'$ of neighboring similar pixels. As seen in Eq. (19), by combining the prediction in Eq. (20) with the virtual Landsat image $\mathbf{L}_{\mathrm{VIP}}$, the final prediction of VIPSTF-SW is obtained.

The main difference between the spatial weighting strategy in VIPSTF-SW and the conventional strategy in STARFM lies in two aspects. First, in VIPSTF-SW, the difference (i.e., $\Delta\mathbf{M}'$) between the MODIS image at the prediction time and the virtual MODIS image is used as the basis for spatial weighting. This is distinguished from STARFM where $\Delta\mathbf{M}$ is larger, as demonstrated in Section 2.2. Second, in VIPSTF-SW, the similar pixels for each center pixel are searched based on the single image $\mathbf{L}_{\mathrm{VIP}}$, rather than all known Landsat images in STARFM where the search is performed for each Landsat image in turn. Among the Landsat time-series images, some images are temporally far from the prediction time, which will decrease the validity of the selection of spectrally similar neighboring pixels. Therefore, the virtual Landsat image $\mathbf{L}_{\mathrm{VIP}}$, which combines Landsat time-series images with adaptive coefficients, is more appropriate for searching similar neighboring pixels.

2.3.2 Spatial unmixing-based VIPSTF (VIPSTF-SU)

In the proposed VIPSTF-SU method, land cover classification is performed on the virtual Landsat image $\mathbf{L}_{\text{VIP}}$ to acquire the fine spatial resolution land cover map. The map is upscaled to the MODIS spatial resolution to produce the coarse proportions for each land cover class. Based on the assumption that the distribution of land cover does not change during the period of interest, the coarse proportions at different times are the same. Thus, the proportion of each class for each MODIS pixel derived from the classification map of $\mathbf{L}_{\text{VIP}}$ is applied to unmix $\Delta\mathbf{M}'$ to produce the increment at the Landsat level. By solving the following linear SU model, the increment for each class can be obtained

$$\begin{bmatrix} p_1(x_1,y_1) & \cdots & p_c(x_1,y_1) & \cdots & p_C(x_1,y_1) \\ \cdots & & \cdots & & \cdots \\ p_1(x_i,y_i) & \cdots & p_c(x_i,y_i) & \cdots & p_C(x_i,y_i) \\ \cdots & & \cdots & & \cdots \\ p_1(x_{n_w},y_{n_w}) & \cdots & p_c(x_{n_w},y_{n_w}) & \cdots & p_C(x_{n_w},y_{n_w}) \end{bmatrix} \begin{bmatrix} \Delta L(1) \\ \cdots \\ \Delta L(c) \\ \cdots \\ \Delta L(C) \end{bmatrix} = \begin{bmatrix} \Delta M'(x_1,y_1) \\ \cdots \\ \Delta M'(x_i,y_i) \\ \cdots \\ \Delta M'(x_{n_w},y_{n_w}) \end{bmatrix}. \tag{21}$$

In Eq. (21), C is the number of classes, n_w is the number of coarse MODIS pixels in the moving window, $\Delta M'(x,y)$ is the MODIS level increment $\Delta\mathbf{M}'$ of the coarse MODIS pixel located at (x,y) in the moving window, $p_c(x,y)$ is the coarse proportion of class c for the coarse MODIS pixel located at (x,y), and $\Delta L(c)$ is the increment for the cth class. For each Landsat pixel, its increment $\Delta\mathbf{L}'$ is determined as

$$\Delta L'(x_0,y_0) = \Delta L\big(c(x_0,y_0)\big) \tag{22}$$

where $c(x_0,y_0)$ is the land cover class of the Landsat pixel located at (x_0,y_0) (determined by the classification map of $\mathbf{L}_{\text{VIP}}$). The final VIPSTF-SU prediction of a Landsat pixel can be obtained by combining the increment in Eq. (22) with the corresponding pixel in $\mathbf{L}_{\text{VIP}}$.

Similarly, the SU model in the proposed VIPSTF-SU method differs from the original SU-based model (i.e., STDFA) in two aspects. First, $\Delta\mathbf{M}'$ is used as the basis for unmixing, rather than $\Delta\mathbf{M}$ in STDFA. Second, in VIPSTF-SU, the single image $\mathbf{L}_{\text{VIP}}$ is used to produce the land cover map, rather than the composed Landsat image whose features are stacked by all known Landsat images.

3. Experiments

3.1. Data and experimental setup

For validation of the proposed VIPSTF approach, MODIS and Landsat time-series images for two sites were used in our experiments. The first site is located in southern New South Wales, Australia (145.0675°E, 34.0034°S) (called Site 1 hereafter) and presents a heterogeneous landscape, while the second site is located in southern New South Wales, Australia (145.0675°E, 34.0034°S) (called Site 2 hereafter) with great land cover change caused by flood inundation. In Site 1, we used Landsat 7 ETM+ time-series from 7 October 2001 to 3 May 2002 and the corresponding 15 MODIS Terra MOD09GA Collection 5 images acquired on almost the same days. In Site 2, 11 pairs of Landsat and MODIS images from 16 April 2004 to 14 February 2005 were used. For both sites the spatial extent is 20 km by 20 km. The detailed acquisition dates of the images are presented in Table 1. Chronologically, we numbered the Landsat images of Site 1 as L1 to L15, and the corresponding MODIS images as M1 to M15. A similar numbering system was applied to Site 2. Partial Landsat and MODIS data for Sites 1 and 2 are shown in Figs. 2 and 3, respectively. It is noted that Site 2 is defined as the site with land cover change. Except for visual inspection (e.g., the flood inundation), the correlation coefficient (CC) between images acquired on different dates for Site 2 is much smaller than that for Site 1, even for two images acquired close in time (e.g., the CC between L8 and L9 for Site 1 is 0.7312, while the CC between L8 and L9 for Site 2 is only 0.3963).

Table 1 Acquisition dates of the MODIS-Landsat data of the two sites

Site 1		Site 2	
Image ID	Date	Image ID	Date
M1-L1	2001.10.07	M1-L1	2004.04.16
M2-L2	2001.10.16	M2-L2	2004.05.02
M3-L3	2001.11.01	M3-L3	2004.07.05
M4-L4	2001.11.08	M4-L4	2004.08.06
M5-L5	2001.11.24	M5-L5	2004.08.22
M6-L6	2001.12.03	M6-L6	2004.10.25
M7-L7	2002.01.04	M7-L7	2004.11.26

M8-L8	2002.02.12	M8-L8	2004.12.12
M9-L9	2002.03.09	M9-L9	2005.01.13
M10-L10	2002.03.16	M10-L10	2005.01.29
M11-L11	2002.04.02	M11-L11	2005.02.14
M12-L12	2002.04.10		
M13-L13	2002.04.17		
M14-L14	2002.04.26		
M15-L15	2002.05.03		

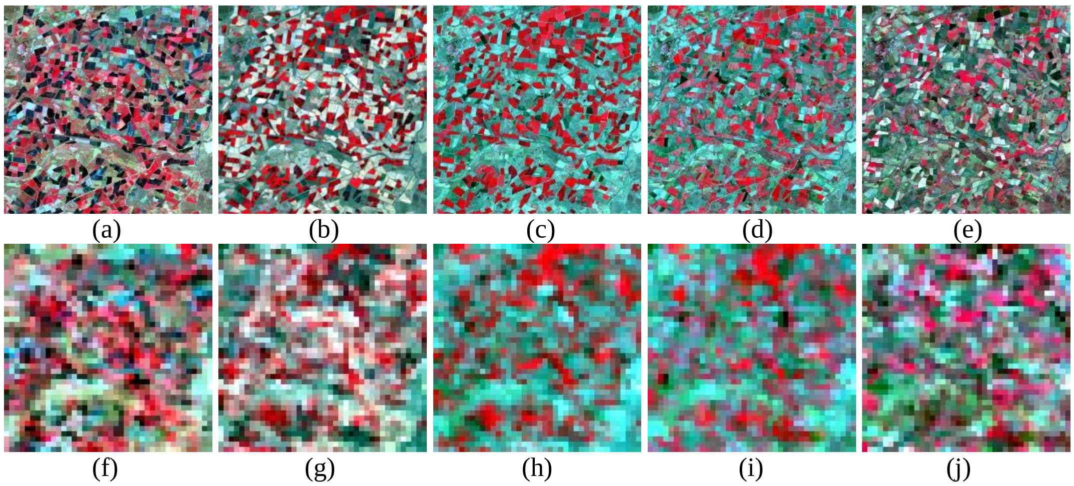


Fig. 2. Partial data of Site 1. (a) L4. (b) L7. (c) L8. (d) L9. (e) L13. (f)-(j) are corresponding MODIS data.

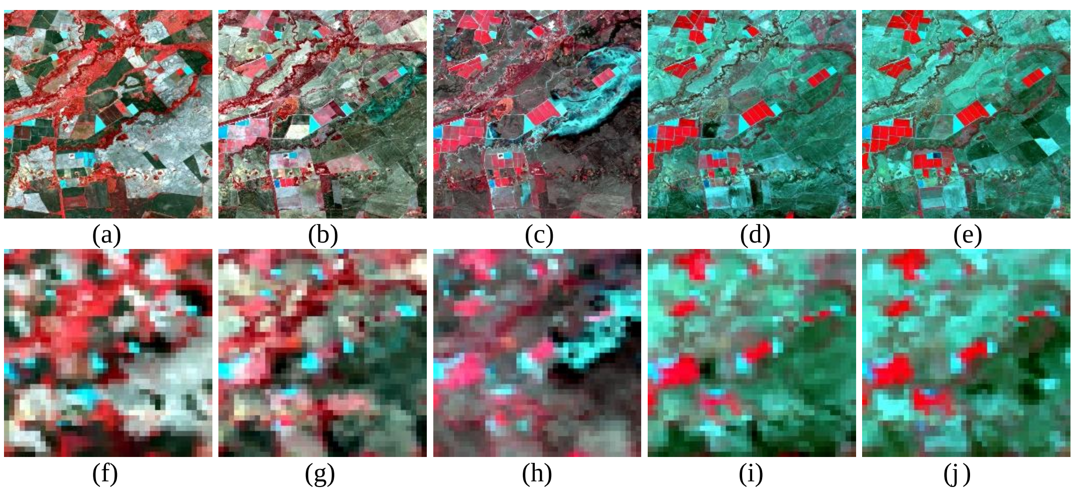


Fig. 3. Partial data of Site 2. (a) L2. (b) L7. (c) L8. (d) L9. (e) L11. (f)-(j) are corresponding MODIS data.

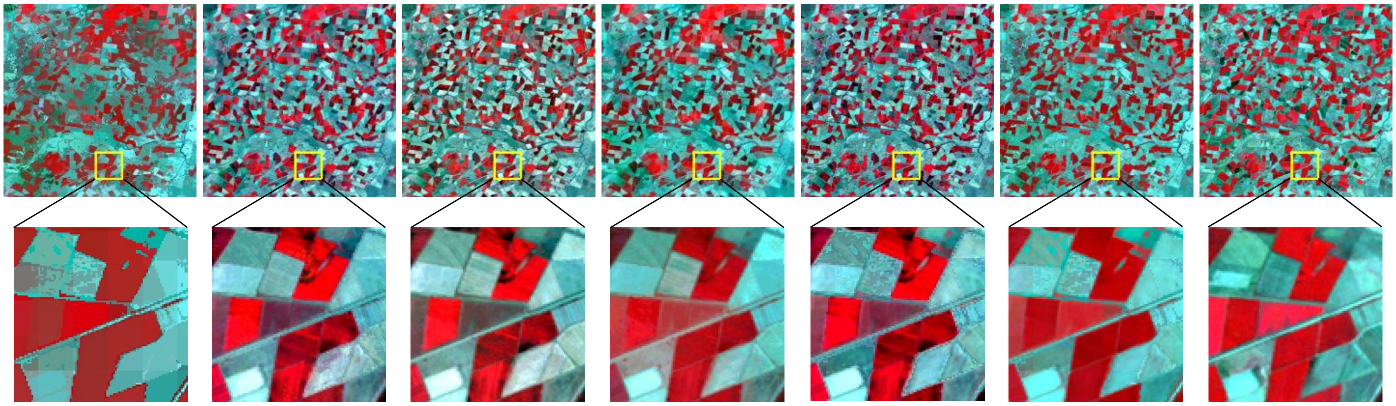
Sections 3.2 and 3.3 provide the results for Site 1 (the heterogeneous site) and Site 2 (the site with land cover change), respectively. For Site 1, spatio-temporal fusion was performed to predict the Landsat image on 12 February 2002 (i.e., L8), based on one MODIS-Landsat image pair (Section 3.2.1) and multiple image pairs

(Section 3.2.2). For Site 2, the prediction date is 12 December 2004, and the results based on one image pair are provided. The proposed VIPSTF approach (including both VIPSTF-SW and VIPSTF-SU versions) is compared with STARFM (Gao et al., 2006), STDFA (Wu et al., 2012), the unmixing-based data fusion (UBDF) algorithm (Zurita-Milla et al., 2008) and Flexible Spatiotemporal DAta Fusion (FSDAF) algorithm (Zhu et al., 2016). For STDFA and VIPSTF-SU, the images were classified into five classes with *k*-means-based unsupervised classification, and for STARFM and VIPSTF-SW, 20 similar pixels were selected within each local window.

3.2. Test for the heterogeneous site (Site 1)

3.2.1 Prediction by one image pair

Among the 15 MODIS-Landsat image pairs of Site 1, we chose one MODIS-Landsat image pair from L1 to L15 (except L8) as the known images, in turn, along with the MODIS image at the prediction time as input. That is, the spatio-temporal fusion methods predict L8 with 14 different inputs. The predictions of the six methods when using M7-L7 as the input image pair are exhibited in Fig. 4 for visual comparison. Obviously, vegetation in the reference image presents as vibrant red. However, the predictions of the vegetation for FSDAF, STARFM and STDFA have a noticeably different color. When the VIP is used in fusion by VIPSTF-SW and VIPSTF-SU, the predictions are visually closer to the reference compared to the original STARFM and STDFA methods as well as FSDAF. Although the color in the UBDF prediction resembles that in the reference image, the method fails to reproduce the intra-class change (i.e., a reflectance value is assigned to the pixels of the same class within the coarse pixel) and also the blocky artifacts is noticeable.



(a) (b) (c) (d) (e) (f) (g)

Fig. 4. Results of different spatio-temporal fusion methods for Site 1 (M7-L7 as known image pair) (NIR, red, and green bands as RGB). (a) UBDF. (b) FSDAF. (c) STARFM. (d) VIPSTF-SW. (e) STDFA. (f) VIPSTF-SU. (g) Reference.

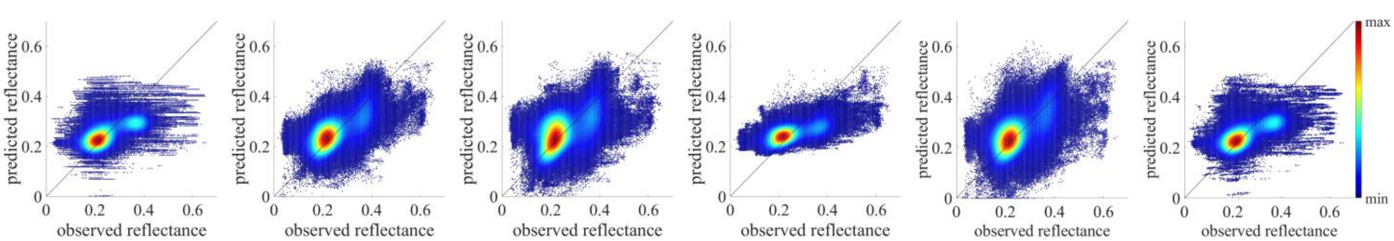
Quantitative evaluation was conducted using the RMSE and CC, as listed in Table 2. The UBDF and FSDAF methods produce mean CCs of around 0.7220 and 0.8314, respectively. For VIPSTF-SW, the mean CC is 0.8345, with an increase of 0.0392 compared to STARFM. For VIPSTF-SU, the mean CC is 0.0174 larger than for STDFA. STARFM and STDFA produced mean RMSEs of 0.0454 and 0.0453, respectively. For VIPSTF-SW and VIPSTF-SU, the corresponding mean RMSEs decrease by 0.0090 and 0.0060, respectively. Among all six methods, VIPSTF-SW produces the greatest accuracy, with the largest CC of 0.8435 and the smallest RMSE of 0.0321. The scatter plots in Fig. 5 reveal the difference between the actual Landsat image and the predictions, where the NIR band is used as an example. Clearly, the points in STARFM and STDFA present greater dispersion. In VIPSTF-SW and VIPSTF-SU predictions, the points are more aggregated and closer to the $y=x$ line.

Fig. 6 shows the RMSEs and CCs of the six methods based on the use of different image pairs (i.e., M1-L1 to M7-L7 and M9-L9 to M15-L15, 14 cases in all). The accuracy increases closer to the prediction time and decreases away from the prediction time, with the predictions using the Landsat images temporally closest to M8-L8 having the greatest accuracy. Checking the results for each method, FSDAF is found to be a competitive method that produces smaller RMSEs and larger CCs than UBDF, STARFM and STDFA in most cases. Moreover, the proposed VIPSTF-SW and VIPSTF-SU methods produce smaller RMSEs and larger

CCs than original STARFM and STDFA, and the two VIPSTF-based methods are also more accurate than FSDAF and UBDF. Interestingly, when different image pairs are used, the performances of VIPSTF-SW and VIPSTF-SU are more robust than the original STARFM and STDFA as well as FSDAF. More specifically, when temporally further image pairs are used, the gain in accuracy for VIPSTF is more obvious. As a result, the difference between VIPSTF and the original STARFM and STDFA methods varies greatly according to the used image pairs. For example, when using M7-L7, the CCs of STARFM and VIPSTF-SW are 0.8043 and 0.8435, respectively, with a difference of 0.0392, but the difference increases to 0.2552 when using M3-L3. Similarly, the difference between VIPSTF-SU and STDFA is 0.0174 when using M7-L7 but up to 0.1716 when using M3-L3.

Table 2 Accuracies of different spatio-temporal fusion methods for Site 1 (M7-L7 as known image pair)

		Ideal	UBDF	FSDAF	STARFM	VIPSTF-SW	STDFA	VIPSTF-SU
RMSE	Blue	0	0.0161	0.0148	0.0163	**0.0127**	0.0164	0.0134
	Green	0	0.0220	0.0199	0.0243	**0.0166**	0.0230	0.0175
	Red	0	0.0326	0.0311	0.0409	**0.0235**	0.0355	0.0251
	NIR	0	0.0684	**0.0664**	0.0788	0.0667	0.0753	0.0668
	SWR1	0	0.0601	0.0455	0.0500	**0.0400**	0.0513	0.0449
	SWR2	0	0.0513	0.0363	0.0365	**0.0332**	0.0404	0.0380
	Mean	0	0.0418	0.0357	0.0411	**0.0321**	0.0403	0.0343
CC	Blue	1	0.7260	0.8691	0.8643	**0.8732**	0.8470	0.8532
	Green	1	0.7223	0.8452	0.8251	**0.8506**	0.8134	0.8303
	Red	1	0.7619	0.8668	0.8562	**0.8818**	0.8484	0.8653
	NIR	1	0.5788	0.6272	0.4899	**0.6496**	0.5531	0.6073
	SWR1	1	0.7652	0.8768	0.8784	**0.8906**	0.8542	0.8632
	SWR2	1	0.7778	0.9036	0.9122	**0.9151**	0.8881	0.8894
	Mean	1	0.7220	0.8314	0.8043	**0.8435**	0.8007	0.8181



(a) (b) (c) (d) (e) (f)

Fig. 5. Scatter plots of the actual and predicted values of the NIR band for Site 1 (M7-L7 as known image pair). (a) UBDF. (b) FSDAF. (c) STARFM. (d) VIPSTF-SW. (e) STDFA. (f) VIPSTF-SU.

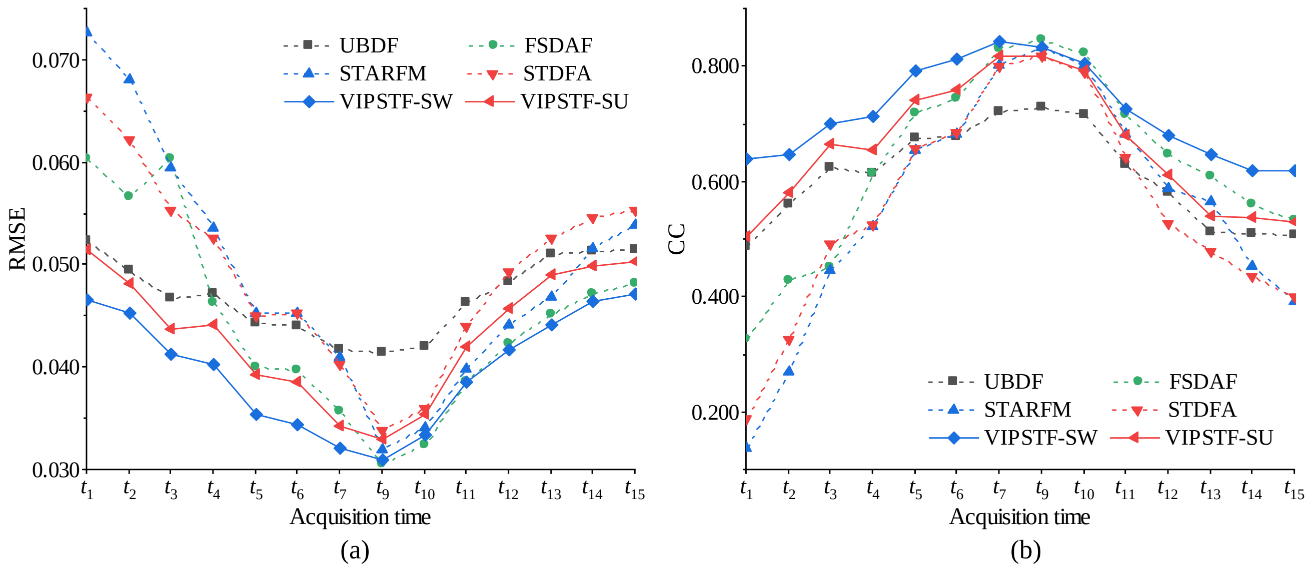


Fig. 6. The prediction accuracy based on different image pairs for Site 1. (a) RMSE. (b) CC.

3.2.2 Prediction by multiple image pairs

For prediction by multiple image pairs, we chose L8 as the Landsat image to predict and the temporally closest M7-L7 and M9-L9 image pairs were selected as the input. When using more image pairs for prediction, the selection of input spreads along both sides one-by-one. For the cases of using 2, 4, 6, 8, 10, 12 and 14 image pairs we compared STARFM, STDFA, VIPSTF-SW and VIPSTF-SU. Fig. 7 shows the sub-area for the predictions of the different methods using 2, 6, 10 and 14 image pairs. When two image pairs are used for prediction, the prediction of STARFM tends to be less accurate than the other three methods, as the prediction shows unexpected dark blocks. As the number of image pairs increases, the difference between the reference and the predictions of STARFM and STDFA enlarges, while the predictions of VIPSTF-SW and VIPSTF-SU are more accurate. It can be seen from the predictions using 14 image pairs that the restoration of the red and green patches in STARFM and STDFA is not as satisfactory as those for VIPSTF-SW and VIPSTF-SU, which are very close to the reference.

Fig. 8 shows the quantitative accuracy assessment of the predictions using multiple image pairs. The accuracy of the prediction by one image pair is also included for comparison. Obviously, no matter how the number of image pairs changes, VIPSTF always provides a more accurate prediction than the corresponding original method. Moreover, from using one to multiple image pairs for prediction, the CCs of VIPSTF increase greatly (e.g., by 0.1795 for STARFM and 0.1471 for STDFA). When using more than two image pairs, the prediction accuracy of VIPSTF increases slowly. More precisely, the CC of VIPSTF-SW is 0.8973 for two image pairs, and increases to 0.9032 for 14 image pairs. The increase of CC of VIPSTF-SW is about 0.0060 from using 2 to 14 image pairs. This is also the same case for VIPSTF-SU, where the corresponding increase in the CC is 0.0124. By contrast, the accuracies of STARFM and STDFA present an apparent fluctuation, and the main trend is that the accuracy can decrease as the number of image pairs increases to a large value. The CCs of STARFM and STDFA decrease by 0.0741 and 0.0667, respectively, when changing from using 6 to 12 image pairs.

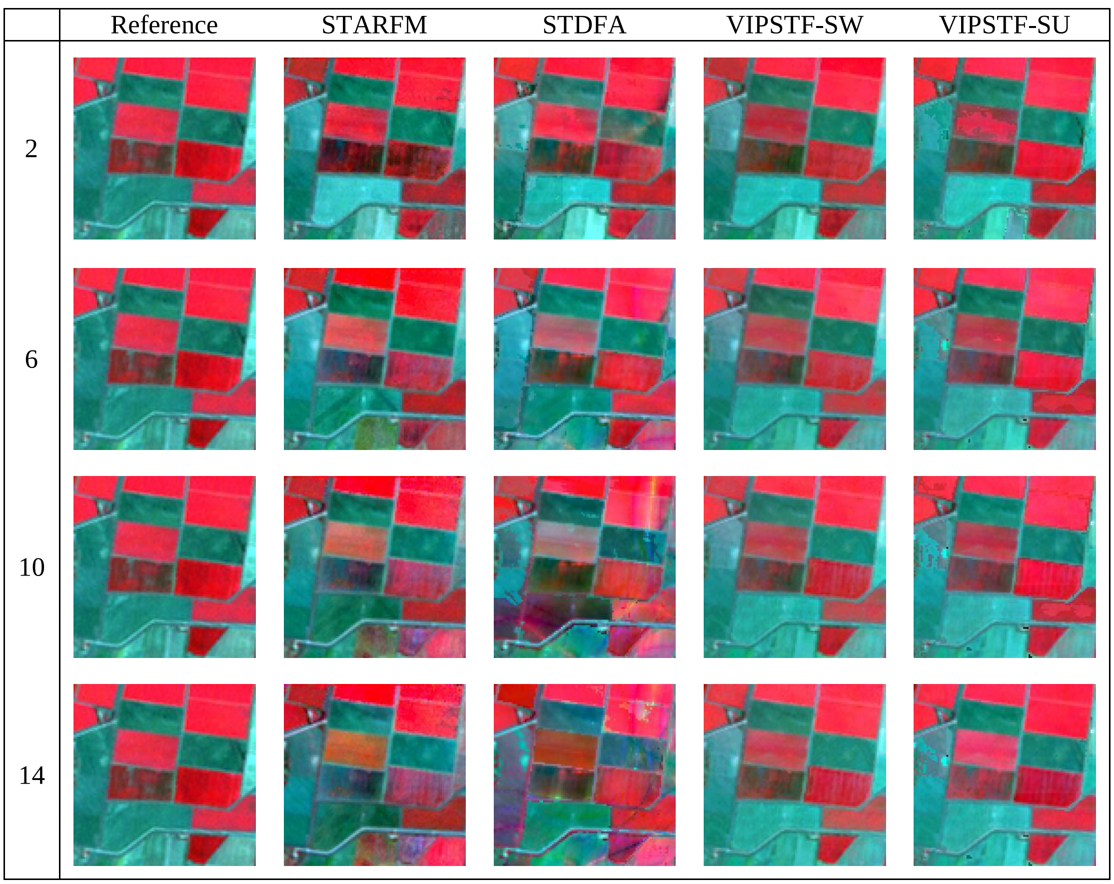


Fig. 7. The predictions based on different numbers of image pairs for Site 1.

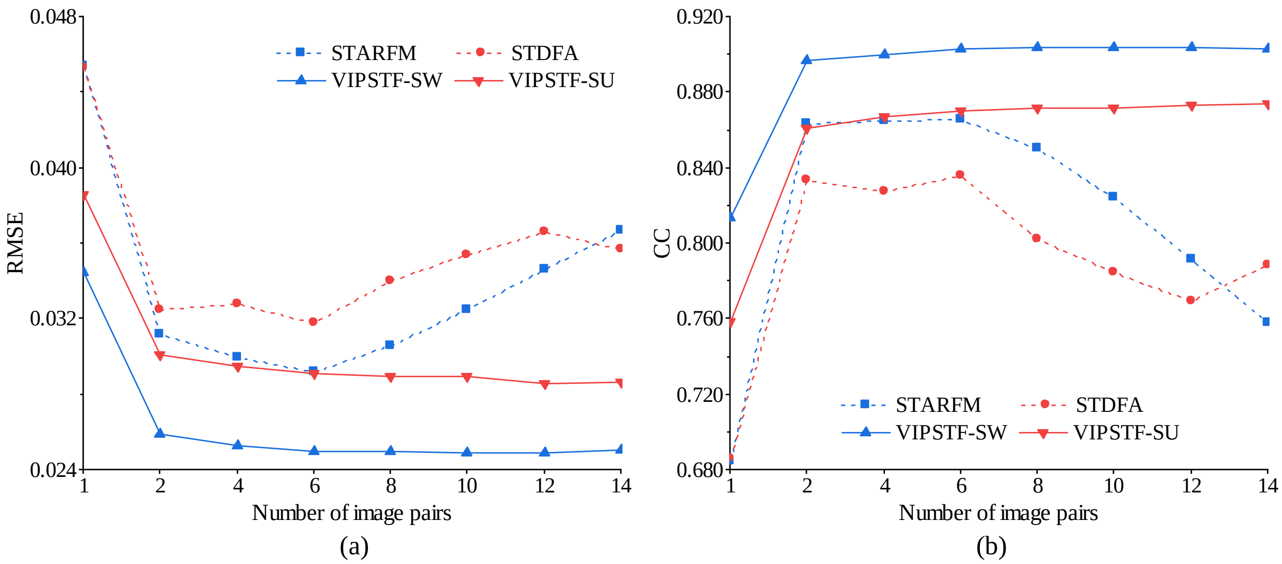
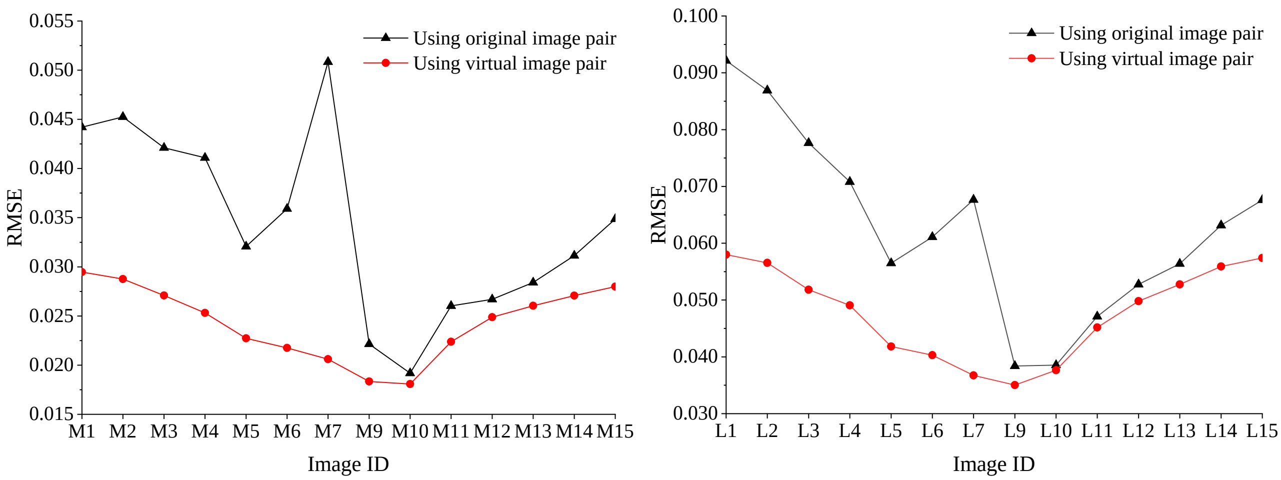


Fig. 8. The accuracy of prediction by multiple image pairs for Site 1. (a) RMSE. (b) CC.

3.2.3 Reduction in the difference between the images at the known and prediction times

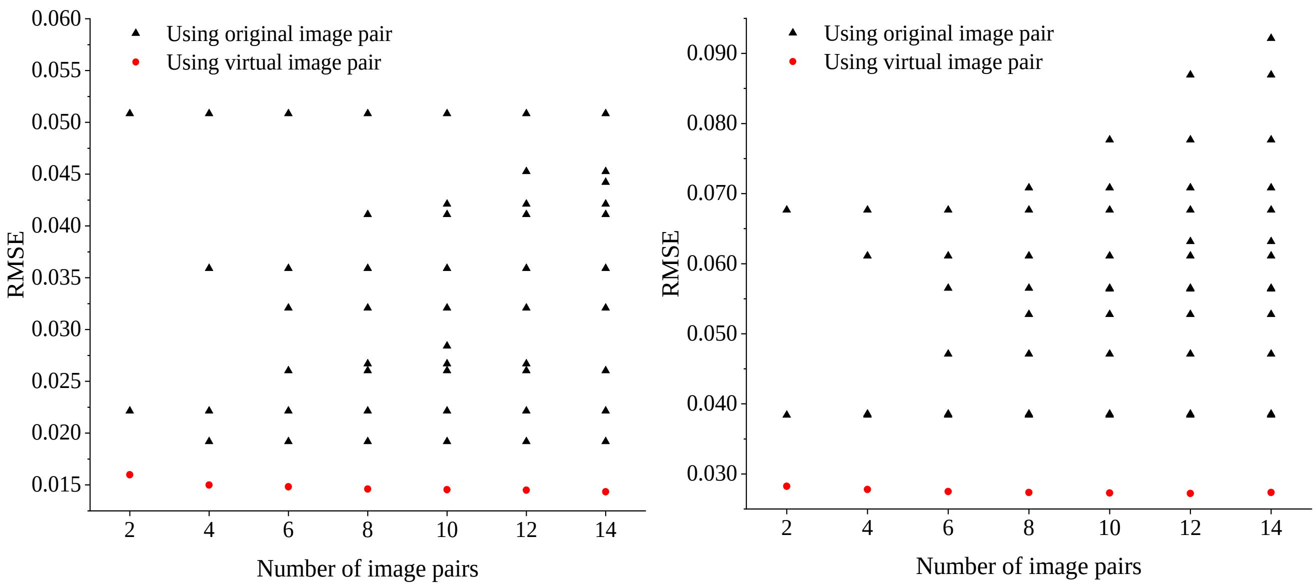
As demonstrated theoretically in Section 2.3, the square root of the expectation of $\Delta\mathbf{M}$, which equals the RMSE between the MODIS images at the known and prediction times, will decrease when using the VIP. Since the VIP includes both Landsat and MODIS images, we calculated the mean RMSEs between the Landsat images and also the mean RMSEs between the MODIS images when using the original image pair and the VIP for comparison. Fig. 9 displays the results for using one image pair (14 cases in all, as in Fig. 6). It can be noticed that the RMSEs between the MODIS images range from 0.0192 to 0.0508 when using the original image pair, and range from 0.0011 to 0.0302 when using the VIP. As for the Landsat images, the RMSEs range from 0.0384 to 0.0869 and 0.0350 to 0.0574 when the original image pair and the VIP are used, respectively. In each case, the RMSEs are obviously smaller when the VIP is used.

The corresponding results for multiple image pairs were also calculated, as shown in Fig. 10. The black triangles represent the mean RMSEs between the different known images (MODIS or Landsat images) and the image (MODIS or Landsat image) at the prediction time, while the red circles are the mean RMSEs between the virtual MODIS or Landsat image and the image (MODIS or Landsat image) at the prediction time. It is seen clearly that the red circle is always less than the black triangle for each prediction, indicating that the RMSE between the VIP and the image at the prediction is always smaller, which is consistent with Eq. (18). Therefore, the VIP can effectively reduce the difference between images at the known and prediction times (i.e., the increments at both the MODIS and Landsat levels).



(a) (b)

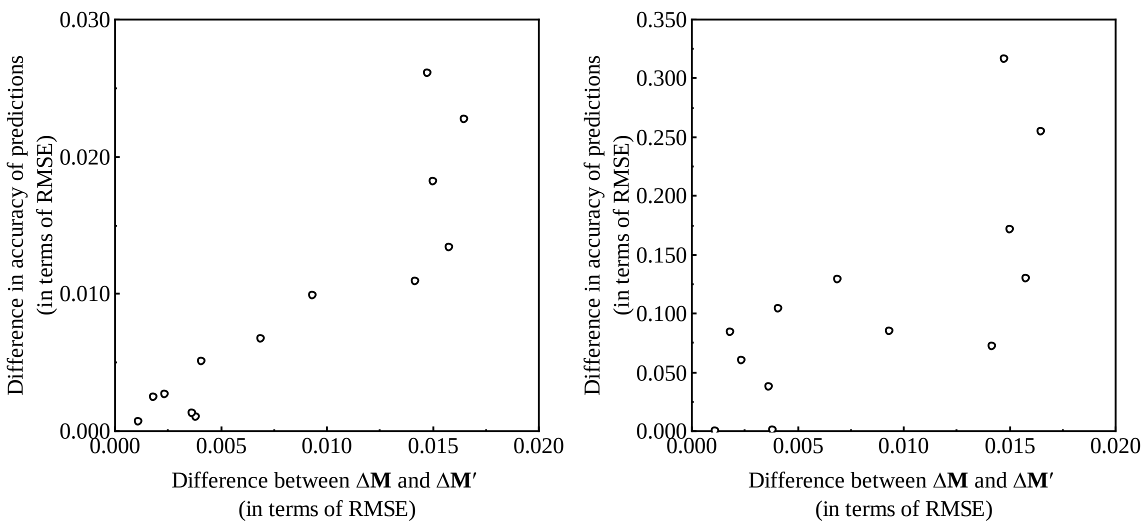
Fig. 9. The RMSE between images at the known and prediction times when using the original image pair and the VIP based on one image pair. (a) RMSE between MODIS images. (b) RMSE between Landsat images.



(a) (b)

Fig. 10. The RMSE between images at the known and prediction times when using the original image pair and the VIP based on multiple image pairs. (a) RMSE between MODIS images. (b) RMSE between Landsat images.

STARFM and STDFA use the original image pairs for prediction, which have a large MODIS level increment $\Delta\mathbf{M}$. In VIPSTF-SW and VIPSTF-SU, however, the virtual MODIS image with a smaller $\Delta\mathbf{M}'$ is used for prediction. To investigate how $\Delta\mathbf{M}$ can influence the prediction accuracy, we calculated the reduction in the increment (in terms of the difference between the mean RMSEs of $\Delta\mathbf{M}$ and $\Delta\mathbf{M}'$), and the corresponding increase in accuracy achieved by using VIPSTF (in terms of the difference between the prediction RMSEs of VIPSTF and the original methods). Fig. 11 shows the scatter plots for VIPSTF-SW and VIPSTF-SU. It can be seen that when the difference between $\Delta\mathbf{M}$ and $\Delta\mathbf{M}'$ increases, the difference between the prediction accuracy increases as well. That is, the increase in accuracy is larger when the reduction in the MODIS level increment $\Delta\mathbf{M}$ is larger.



(a) (b)

Fig. 11. Scatter plots of reduction in the MODIS level increment (in terms of the difference between $\Delta\mathbf{M}$ and $\Delta\mathbf{M}'$) and the corresponding increase of prediction accuracy (in terms of RMSE decrease) for Site 1. (a) STARFM and VIPSTF-SW. (b) STDFA and VIPSTF-SU.

3.2.4 Computational cost

The computational costs for STARFM, STDFA, VIPSTF-SW and VIPSTF-SU are shown in Fig. 12. It is obvious that the computational costs of STARFM and STDFA increases linearly when more image pairs are used, while those of VIPSTF-SW and VIPSTF-SU remain stable from using 1 to 14 image pairs. This is because both the spatial weighting procedure of STARFM and the spatial unmixing process of STDFA require time-consuming computation. When a new image pair is added, an additional time-consuming spatial weighting or spatial unmixing process is implemented. In VIPSTF, however, only a single VIP is constructed based on the simple linear transformation, and the time spent on producing the VIP is negligible. Moreover, the spatial weighting or spatial unmixing process is implemented only once, which saves computational cost significantly.

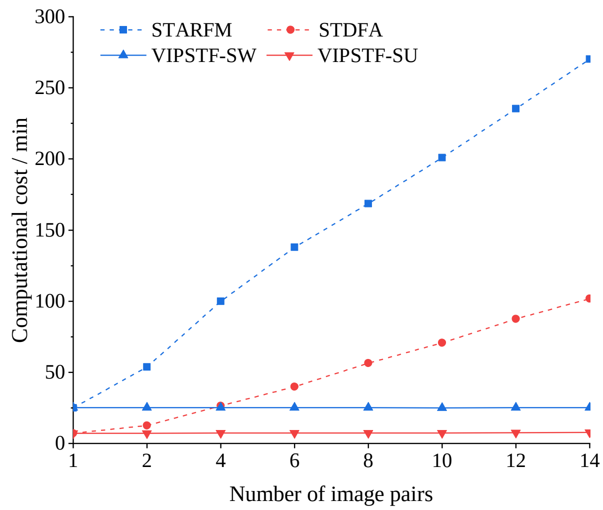
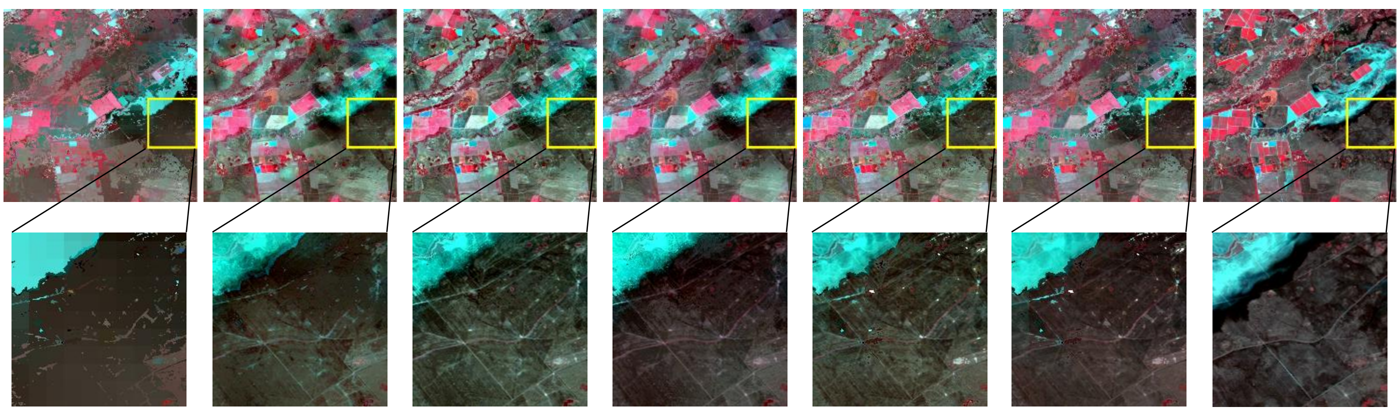


Fig. 12. Computational costs of the methods for Site 1.

3.3. Test for the site with land cover change (Site 2)

For the site with land cover change, we chose the image numbered L8 as the Landsat image to predict. The 10 Landsat images numbered L1 to L7 and L9 to L11 were selected as the inputs to prediction, respectively. The predictions produced using M7-L7 as input are shown in Fig. 13. Since the Landsat image to predict covers a large area inundated by floods which does not occur in the known Landsat images, large uncertainties exist in the predictions. From the visual comparison, all six methods can capture the flood information, but the boundary of the flood for each prediction varies noticeably. It is apparent that FSDAF, VIPSTF-SW and VIPSTF-SU can predict the boundary more accurately; see the black zone below the flood area. Furthermore, when comparing the sub-area, the predictions of VIPSTF-SW and VIPSTF-SU have a more similar color to the reference image than STARFM, STDFA and FSDAF. Table 3 lists the accuracy of the six methods when using M7-L7 as the image pair. Overall, UBDF produces the smallest mean CC of 0.5595, while VIPSTF-SW provides the largest mean CC of 0.7432. Compared to STARFM, the mean RMSE is decreased by 0.0048 and the mean CC is increased by 0.0324 using VIPSTF-SW. Similarly, when using VIPSTF-SU, the mean RMSE is decreased by 0.0022 and the mean CC is increased by 0.0101 compared to STDFA. FSDAF produces a more accurate prediction than UBDF, STDFA and STARFM, but is less accurate than VIPSTF-SW.



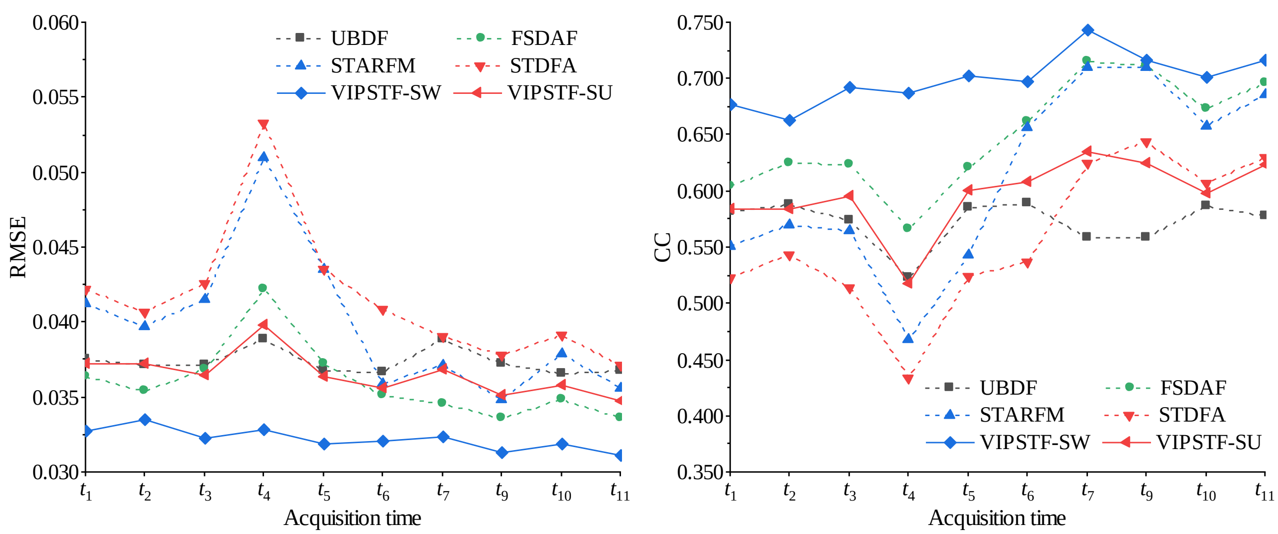
(a) (b) (c) (d) (e) (f) (g)

Fig. 13. Results of different methods for Site 2 (M7-L7 as known image pair). (a) UBDF. (b) FSDAF. (c) STARFM. (d) VIPSTF-SW. (e) STDFA. (f) VIPSTF-SU. (g) Reference.

Table 3 Accuracy of different spatio-temporal fusion methods for Site 2 (M7-L7 as known image pair)

		Ideal	UBDF	FSDAF	STARFM	VIPSTF-SW	STDFA	VIPSTF-SU
RMSE	Blue	0	0.0201	**0.0140**	0.0147	0.0143	0.0162	0.0162
	Green	0	0.0240	0.0201	0.0209	**0.0194**	0.0233	0.0222
	Red	0	0.0284	0.0242	0.0253	**0.0229**	0.0280	0.0265
	NIR	0	0.0462	0.0328	0.0325	**0.0315**	0.0401	0.0400
	SWR1	0	0.0633	0.0610	0.0681	**0.0584**	0.0674	0.0638
	SWR2	0	0.0512	0.0555	0.0614	**0.0481**	0.0593	0.0526
	Mean	0	0.0389	0.0346	0.0372	**0.0324**	0.0391	0.0369
CC	Blue	1	0.4774	0.6540	0.6396	**0.6949**	0.5597	0.5800
	Green	1	0.5265	0.6766	0.6586	**0.7026**	0.5700	0.5924
	Red	1	0.5011	0.6659	0.6466	**0.6952**	0.5554	0.5706
	NIR	1	0.6043	0.8317	0.8384	**0.8456**	0.7423	0.7351
	SWR1	1	0.6427	0.7494	0.7486	**0.7671**	0.6758	0.6800
	SWR2	1	0.6051	0.7168	0.7330	**0.7541**	0.6470	0.6525
	Mean	1	0.5595	0.7157	0.7108	**0.7432**	0.6250	0.6351

The prediction accuracies of the six methods based on the use of multiple image pairs are shown in Fig. 14. The prediction accuracies do not show an obvious trend as for Site 1, and the accuracies are smaller. The reason is that spatio-temporal fusion becomes more challenging when great land cover change exists. It is evident that either VIPSTF-SW or VIPSTF-SU produces greater accuracy than the original STARFM or STDFA. The CCs of VIPSTF-SW range from 0.6636 to 0.7432, while CCs of STARFM range from 0.4684 to 0.7108. As for VIPSTF-SU, the RMSEs are smaller than for STDFA, and the CCs are larger than for STDFA in most cases. In addition, the accuracy of FSDAF lies between that of STARFM and VIPSTF-SW, and the accuracy of UBDF fluctuates when using different image pairs.



(a) (b)

Fig. 14. The prediction accuracy based on different image pairs for Site 2. (a) RMSE. (b) CC.

4. Discussion

4.1. The impact of image pairs

In the experiments for the heterogeneous site, predictions using multiple image pairs were provided for different spatio-temporal fusion methods. From Fig. 8, we find that as the number of image pairs increases to a large value (e.g., larger than six), the accuracy increases slowly for VIPSTF-SW and VIPSTF-SU, but decreases obviously for STARFM and STDFA. For STARFM and STDFA, the final predictions are the weighted sum of separate predictions based on different image pairs. The weightings are mainly determined by the temporal difference between the known and prediction times in a local window. We calculated the absolute mean CCs of all six bands between the Landsat images at the known time (i.e., time of L1 to L15 except L8) and prediction time (i.e., time of L8), as shown in Fig. 15. The absolute CCs for the Landsat images of the eight image pairs are distributed between the two blue dotted lines in Fig. 15. It can be noted that when L4 and L12 were added for fusion, the absolute CCs decrease obviously on both sides, which corresponds to the dramatic decrease in the accuracy of STARFM and STDFA in Fig. 8. This means STARFM and STDFA are sensitive to the CC between the image at the known and prediction times, but the existing scheme of combining multiple image pairs cannot accurately account for this factor. As a result, the image pairs with small correlation (e.g., the CC between L2 and L8 is 0.0649) can affect greatly the final prediction accuracy. In contrast, for VIPSTF, when constructing the VIP, different coefficients were assigned to images at different known times, and the coefficients are closely related to the CC between the image at the known and prediction times. For clarification, the absolute coefficients $|a|$ of the green, red and NIR bands for L1 to L15 (except L8) in the case of using 14 image pairs are depicted in Fig. 16(a), while the relation with the CC (the red band is used as an example) is depicted in Fig. 16(b). In general, the lines of $|a|$ in Fig. 16(a) show a similar trend to that of the |CC| in Fig .15. Moreover, as seen from Fig. 16(b), $|a|$ is larger when |CC| is larger. This means the known image pairs with small correlation will be less informative in VIPSTF. Therefore, VIPSTF can assign $|a|$ to different known images adaptively according to its correlation with the image at the prediction time. In spatio-temporal fusion, several studies investigated how to determine the optimal input image pairs (Chen et al., 2020; Tang et al., 2020), such as using the CC between coarse observations or even the CC between the coarse and fine images in each image pair to find the optimal image pairs. However, this issue remains open. For the VIPSTF proposed in this paper, the adaptive assignment of weights to different image pairs is robust when using multiple image pairs, and more importantly, releases the requirement for image pair selection, which is a complicated task.

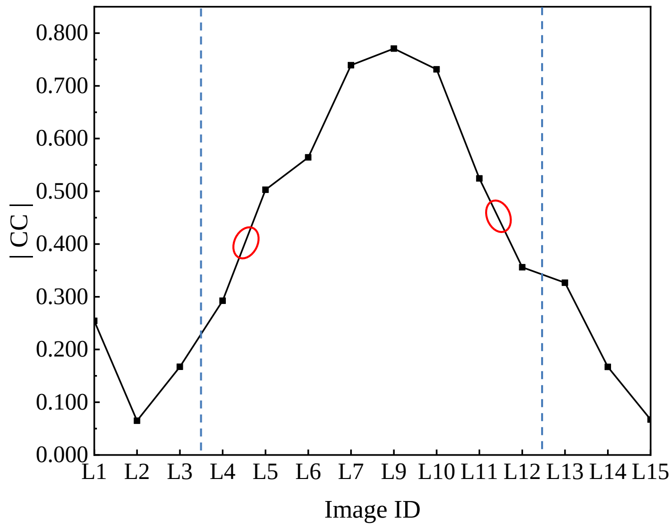


Fig. 15. The CC between Landsat images at the known and prediction times.

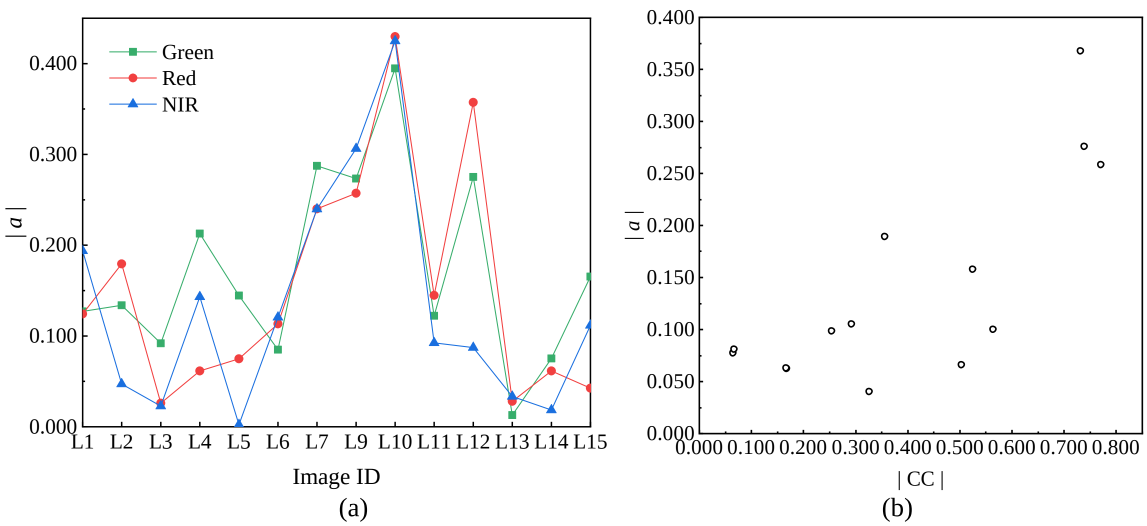


Fig. 16. Variation in the absolute regression coefficient |*a*|. (a) |*a*| of Landsat at different times (e.g., 14 images). (b) Scatter plot between |CC| and |*a*| for the Red band.

In practice, due to the influence of cloud contamination, it is difficult to acquire sufficient MODIS and Landsat time-series image pairs with reliable quality. Also, image pre-processing, including geometric registration between the MODIS and Landsat images, may require intensive effort. Intuitively, we expect the employment of more image pairs to be beneficial and to increase accuracy. According the experimental results, however, the inclusion of more image pairs does not necessarily benefit obviously VIPSTF if the number of image pairs is already large. Thus, there emerges an imbalance in the costs and benefits. To avoid futile efforts in acquiring the MODIS and Landsat data in practical applications, it is necessary to define an index based on the idea of cost-benefit ratio to guide the determination of the number of image pairs. It is expected that the optimal number may vary according to the study area.

4.2. The relation between the Landsat and MODIS images

In the proposed VIPSTF approach, it is assumed that the reflectance of each MODIS pixel is the average of the corresponding Landsat pixels covering the same area (Li et al., 2020; Zhu et al., 2010). However, there always exists inconsistency between MODIS and Landsat images, which produces a bias in the assumed relationship (Chen et al., 2020; Li et al., 2020; Xie et al., 2018). The reason for this phenomenon is that the acquisition conditions (e.g., atmospheric effects, Sun-sensor geometry, bidirectional reflectance distribution function (BRDF) effects, the response function, noise, etc.) vary for different sensors (Gao et al., 2014; Roy et al., 2016). For example, although Terra, Aqua and Landsat are all Sun-synchronous orbit satellites, their viewing angles are different. MODIS images are acquired at very large viewing angles, while Landsat images are acquired with near-nadir view. All these factors will cause an inevitable bias in the simple averaging model. The bias can also differ greatly for MODIS-Landsat pairs acquired in different spatial regions and at different times. Since the bias is difficult to characterize at the current stage, it is challenging to express the relationship between Landsat and MODIS in a perfectly accurate mathematical model. However, if any prior knowledge or auxiliary information is available, it can be used readily when constructing the relation between the Landsat and MODIS images for possible enhancement of the proposed VIPSTF approach.

4.3. Production of the VIP

This paper introduced the concept of the VIP to synthesize a MODIS-Landsat image pair closer to the prediction time. Theoretically, there should be opening solutions to produce the VIP. In this paper, it was determined specifically using a linear transformation model. See Eqs. (3) and (4), when constructing the VIP, we defined two functions, g_1 and g_2. Based on the assumption of linear transformation, g_1 and g_2 were defined as the linear weighted sum of MODIS and Landsat time-series images, as expressed in Eqs. (9) and (10). The rationale for the production of the VIP (i.e., the linear regression-based solution to determine the coefficients) was demonstrated mathematically. Experiments also validate that both the virtual MODIS and Landsat images are closer to that for the prediction time (see Figs. 9 and 10). Except for the linear transformation adopted in this paper, other transformation models such as nonlinear transformation may also be considered in future research. The application of these models may potentially lead to a more appropriate characterization of VIP and increase the fusion accuracy finally. Nevertheless, two points need to be emphasized when developing other transformation methods. First, the main objective of the production of the VIP is to reduce $\Delta\mathbf{M}$, that is, to produce a VIP closer to the prediction time. Second, the transformation should preserve the consistency between the MODIS and Landsat images, such as in Eq. (5). This means that the two functions g_1 and g_2 need to be connected in a certain way, either explicitly or intrinsically.

4.4. The applicability of VIPSTF

In the general framework of the existing spatio-temporal fusion methods in Eqs. (1) and (2), the function f is the most critical issue for prediction. For the SW and SU methods used in the proposed VIPSTF approach, f is a specific function that can be characterized explicitly by a mathematical expression. However, there also exists some other spatio-temporal fusion methods where f cannot be defined as an explicit function. For example, in some learning-based methods (e.g., sparse representation (Huang and Song, 2012; Song and

Huang, 2013; Zhao et al., 2018), support vector regression (Moosavi et al., 2015) and deep learning (Das and Ghosh, 2016; Song et al., 2018)), the processing of $\Delta\mathbf{M}$ is performed in a black box. In this paper, VIPSTF was demonstrated to be more accurate by applying the linear mechanism of SW and SU methods to process the new MODIS increment $\Delta\mathbf{M}'$ between the virtual MODIS image and the MODIS at the prediction time. Based on this encouraging performance, it is also worthwhile to investigate whether VIPSTF has the potential to be adopted to other spatio-temporal fusion methods (e.g., learning-based methods) where the function f cannot be expressed explicitly. For these methods, however, the combination with VIPSTF tends to be more complex, and the feasibility remains to be validated and developed. On the other hand, for some learning-based methods, at least two image pairs (one before and one after the prediction time) are required. The VIP produced in this paper is actually a single image pair. Thus, it would be interesting to construct multiple VIPs (e.g., one VIP before and one VIP after the prediction time) for these methods, or even extend the original learning-based methods to be applicable to only one image pair. This is part of our ongoing research.

4.5. Comparison between VIPSTF-SW and VIPSTF-SU

In this paper, two versions of VIPSTF were developed by extending existing SW and SU schemes for characterizing the function f. From the prediction by one image pair for the heterogeneous area in Section 3.2, the two types of methods have close performances and the difference in accuracy is small. For the area experiencing land cover changes in Section 3.3, however, the prediction of the SW methods have a greater accuracy than the SU methods in most cases; see the lines in Fig. 14(b). The reason is that there is a strong assumption in the SU-based methods: the proportions of land cover classes do not change during the time of interest. This assumption means the matrix of coarse proportions in Eq. (21) is fixed for any time, which makes the SU methods especially sensitive to land cover changes. In future research, it may be of great interest to develop more adaptive SU methods to account explicitly for land cover changes. For example, a bias term reflecting the degree of change in proportions could be included in the original coarse proportions to predict more reliable increments for each class. However, how to quantify the change degree would be a critical issue, which may require reliable change detection between coarse spatial resolution images. On the other hand, blocky artifacts always exist in the predictions of SU methods because the unmixing step is implemented in units of coarse pixels, so that the pixels belonging to the same class in a local window may have very different reflectances. The spatial filtering scheme used in the Fit-FC method proposed in our previous research (Wang and Atkinson, 2018) may be a plausible solution to remove them, but the prediction can sometimes be visually smooth. It is found that the use of coarse proportions upscaled from soft classification results of an available fine spatial resolution land cover map, rather than a fine hard classified map in spatial unmixing, can alleviate the blocky artifacts (Liu et al., 2020; Ma et al., 2018; Wang et al., 2020). The theoretical basis behind this needs to be investigated further. Therefore, it would also be interesting to seek solutions to reduce the blocky artifacts in SU-based methods including the proposed VIPSTF-SU method for further enhancement.

4.6. Comparison with solutions based on Landsat time-series

Some studies have been developed for predicting Landsat images based on the homologous Landsat time-series accumulated from other days (Hilker et al., 2009; Zhu et al., 2015; Zhu et al., 2018). For example, Zhu et al. (2015) synthesized Landsat images at any given time using all available Landsat data based on seasonal trend analysis. Zhu et al. (2018) filled the missing pixels due to SLC-off and cloud contamination to produce spatially complete Landsat data. These researches are different from the spatio-temporal fusion investigated in this paper. First, from the perspective of data, they are performed based on the availability of Landsat time-series, sometimes for a very long time (e.g., >30 years in Zhu et al. (2015)). Spatio-temporal fusion, however, is flexible to the number of available Landsat images and has a much lighter dependence on the number of data. That is, spatio-temporal fusion can also be performed using only one temporal neighboring Landsat image. Second, from the perspective of principles, spatio-temporal fusion actually focuses on the issue of downscaling, by taking full advantage of the coarse MODIS images and the fine Landsat images to predict the completely missing Landsat images on the same dates of MODIS images. The solutions based on long Landsat time-series account for seasonal trends and fit a model to characterize the reflectance at any time (Zhu et al., 2015). The gap-filling solution in Zhu et al. (2018) is performed using spatial and temporal interpolation, based on partly available Landsat data at the prediction time, rather than completely missing Landsat data at the prediction time as in spatio-temporal fusion. Given the common goal of predicting Landsat images, these two types of solutions can be potentially combined, which may be one breakthrough to enhance the performance of predicting missing Landsat data. Seasonal trends present the law of dynamic change of land cover at Landsat resolution at different times, while spatio-temporal fusion further exploits information from additional coarse MODIS images. This provides an interesting avenue for future research.

5. Conclusion

For spatio-temporal fusion, uncertainty exists mainly in the downscaling process of estimating the fine spatial resolution level increment (e.g., Landsat level increment) from the coarse level increment (e.g., MODIS level increment), which also means the difference between images of the known and prediction times. This paper proposed to construct a VIP which is closer to the data at the prediction time to capture more fine spatial resolution information directly from the known Landsat images, thus, reducing the burden of estimating the Landsat level increment. It was demonstrated theoretically that the VIP can reduce the MODIS level increment. Based on the concept of VIP, the VIPSTF approach was proposed. VIPSTF is a general approach suitable to both spatial weighting- and spatial unmixing-based methods. Accordingly, two versions of VIPSTF (i.e., VIPSTF-SW and VIPSTF-SU) were developed in this paper. Experiments were performed on two groups of datasets, and the proposed VIPSTF-based methods were compared to existing UBDF, FSDAF, STARFM and STDFA methods. The main findings are summarized as follows.

1) VIPSTF can enhance the performance of spatio-temporal fusion. The accuracies of both VIPSTF-SW and VIPSTF-SU are greater than the original STARFM and STDFA methods as well as the popular UBDF and FSDAF methods. For the prediction using M7-L7 as the known image pair for Site 1, the mean CC of VIPSTF-SW is 0.8435, which is 0.0392, 0.1215 and 0.0121 larger than for STARFM, UBDF and FSDAF, respectively. Also, the mean RMSE of VIPSTF-SU is 0.0060, 0.0075 and 0.0014 smaller than for STDFA, UBDF and FSDAF, respectively.
2) Both the virtual MODIS and Landsat images in the VIP are closer to the data at the prediction time than the original image pairs. The VIP can effectively reduce the increments at both the MODIS and Landsat levels. The advantage of VIPSTF is especially obvious when the reduction in the increment is large (i.e., the case where the original image pairs are temporally far from the prediction time).
3) VIPSTF is applicable to both heterogeneous sites and sites experiencing temporal land cover type changes.
4) For the prediction by multiple image pairs, as the number of image pairs increases, the prediction accuracies of STARFM and STDFA can decrease, but that of VIPSTF increases slowly or stays stable. This means that VIPSTF is robust to the use of different image pairs, which releases it from the complicated problem of image pair selection.
5) For the site with land cover changes, VIPSTF-SW is more accurate than VIPSTF-SU, and the latter is more sensitive to land cover changes. When using M7-L7 as the known image pair, the mean CC of VIPSTF-SW is 0.1081 larger than for VIPSTF-SU.
6) When using more image pairs, the computational cost of STARFM and STDFA increases noticeably, while VIPSTF always maintains a constant and smaller running time.

Acknowledgment

This work was supported by National Natural Science Foundation of China under Grant 41971297, Fundamental Research Funds for the Central Universities under Grant 02502150021 and Tongji University under Grant 02502350047.

Appendix A

As seen from Eq. (14), $\Delta\mathbf{M}$ can be expressed as $\sum_{i=1}^{N} w_i(\mathbf{M}_p - \mathbf{M}_i)$ when using multiple image pairs for fusion. Considering the relationship between the expectation and the variance, $E(\Delta\mathbf{M}^2)$ can be calculated as

$$\begin{aligned} E(\Delta\mathbf{M}^2) &= Var(\Delta\mathbf{M}) + E^2(\Delta\mathbf{M}) \\ &= Var\left[\sum_{i=1}^{N} w_i(\mathbf{M}_p - \mathbf{M}_i)\right] + E^2\left[\sum_{i=1}^{N} w_i(\mathbf{M}_p - \mathbf{M}_i)\right] \end{aligned} \tag{A1}$$

As for the variance term $Var\left[\sum_{i=1}^{N} w_i(\mathbf{M}_p - \mathbf{M}_i)\right]$, $\mathbf{M}_p$ can be represented by the transformation of $\mathbf{M}_k$ according to Eq. (11) (note that $\mathbf{M}_k$ and $\mathbf{M}_i$ do not refer to the same MODIS image). Thus, we have

$$\begin{aligned} Var(\Delta\mathbf{M}) &= Var\left[\sum_{i=1}^{N} w_i(\mathbf{M}_p - \mathbf{M}_i)\right] \\ &= Var\left[\sum_{i=1}^{N} w_i(\sum_{k=1}^{N} a_k\mathbf{M}_k + b + \mathbf{r} - \mathbf{M}_i)\right] \\ &= Var\left[\sum_{i=1}^{N} w_i(\sum_{k=1}^{N} a_{k_i}\mathbf{M}_k + b + \mathbf{r})\right] \\ &= Var(\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i}\mathbf{M}_k + \sum_{i=1}^{N} w_i b + \sum_{i=1}^{N} w_i\mathbf{r}) \\ &= Var(\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i}\mathbf{M}_k + \mathbf{r}) \end{aligned} . \tag{A2}$$

In Eq. (A2), $\mathbf{M}_i$ is merged with $\sum_{k=1}^{N} a_k\mathbf{M}_k$ by defining a new coefficient

$$a_{k_i} = \begin{cases} a_k - 1, & k = i \\ a_k, & k \neq i \end{cases}. \tag{A3}$$

Moreover, the term $\sum_{i=1}^{N} w_i b$ can be canceled in Eq. (A2) as both w_i and b are constant, and the term $\sum_{i=1}^{N} w_i \mathbf{r}$ is simplified as $\mathbf{r}$ since $\sum_{i=1}^{N} w_i = 1$.

Considering the expansion rule of the variance of the sum of two variables, Eq. (A2) can be rewritten as

$$\begin{aligned} Var(\Delta\mathbf{M}) &= Var(\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i} \mathbf{M}_k) + Var(\mathbf{r}) + 2Cov(\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i} \mathbf{M}_k, \mathbf{r}) \\ &= Var(\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i} \mathbf{M}_k) + Var(\mathbf{r}) + 2\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i} Cov(\mathbf{M}_k, \mathbf{r}) \end{aligned}. \tag{A4}$$

According to the relationship between the covariance and the expectation, $Cov(\mathbf{M}_k, \mathbf{r})$ can be transformed as

$$Cov(\mathbf{M}_k, \mathbf{r}) = E(\mathbf{M}_k \cdot \mathbf{r}) - E(\mathbf{M}_k)E(\mathbf{r}) \tag{A5}$$

where $\cdot$ means the inner product between two vectors.

For classical least squares-based linear regression modeling, there are two important properties. First, the expectation of the product of the independent variable and the residual is zero. Second, the expectation of the residual is zero (Draper and Smith, 2014)

$$\begin{aligned} E(\mathbf{M}_k \cdot \mathbf{r}) &= 0 \\ E(\mathbf{r}) &= 0 \end{aligned}. \tag{A6}$$

Therefore, Eq. (A5) equals to zero and Eq. (A4) can then be rewritten as

$$Var(\Delta\mathbf{M}) = Var(\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i} \mathbf{M}_k) + Var(\mathbf{r}). \tag{A7}$$

According to Eq. (A7), Eq. (A1) can be updated as

$$E(\Delta\mathbf{M}^2) = Var(\sum_{i=1}^{N} w_i \sum_{k=1}^{N} a_{k_i} \mathbf{M}_k) + Var(\mathbf{r}) + E^2\left[\sum_{i=1}^{N} w_i (\mathbf{M}_p - \mathbf{M}_i)\right]. \tag{A8}$$

When the VIP is used, based on Eqs. (10) and (11), $E(\Delta\mathbf{M}'^2)$ can be derived as

$$\begin{aligned} E(\Delta\mathbf{M}'^2) &= E\left[(\mathbf{M}_p - \mathbf{M}_{\mathrm{VIP}})^2\right] \\ &= E(\mathbf{r}^2) \\ &= Var(\mathbf{r}) + E^2(\mathbf{r}) \\ &= Var(\mathbf{r}) \end{aligned} . \tag{A9}$$

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.

Belgiu, M., Stein, A., 2019. Spatiotemporal image fusion in remote sensing. Remote Sensing 11(7), 818.

Busetto, L., Meroni, M., Colombo, R., 2008. Combining medium and coarse spatial resolution satellite data to improve the estimation of sub-pixel NDVI time series. Remote Sensing of Environment 112(1), 118–131.

Chen, B., Huang, B., Xu, B., 2015. Comparison of spatiotemporal fusion models: A review. Remote Sensing, 1798–1835.

Chen, Y., Cao, R., Chen, J., Zhu, X., Zhou, J., Wang, G., Shen, M., Chen, X., Yang, W., 2020. A new cross-fusion method to automatically determine the optimal input image pairs for NDVI spatiotemporal data fusion. IEEE Transactions on Geoscience and Remote Sensing.

Das, M., Ghosh, S. K., 2016. Deep-STEP: A deep learning approach for spatiotemporal prediction of remote sensing data. IEEE Geoscience and Remote Sensing Letters 13, 1984–1988.

Draper, N. R., Smith, H., 2014. Applied Regression Analysis, third ed. John Wiley and Sons, New York.

Dyer, C., 2012. Adaptability research of spatial and temporal remote sensing data fusion technology in crop monitoring. Remote Sensing Technology and Application 345(4), e4638.

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.; He, T.; Masek, J. G.; Shuai, Y.; Schaaf, C. B.; Wang, Z, 2014. Angular Effects and Correction for Medium Resolution Sensors to Support Crop Monitoring. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7, 4480–4489.

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

Gevaert, C. M., Javier Garcia-Haro, F., 2015. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. Remote Sensing of Environment 156, 34–44.

Houborg, R., McCabe, M. F., Gao, F., 2016. A spatio-temporal enhancement method for medium resolution LAI (STEM-LAI). International Journal of Applied Earth Observation and Geoinformation 47, 15–29.

Huang, B., Song, H., 2012. Spatiotemporal reflectance fusion via sparse representation. IEEE Transactions on Geoscience and Remote Sensing 50, 3707–3716.

Huang, B., Wang, J., Song, H., Fu, D., Wong, K., 2013. Generating high spatiotemporal resolution land surface temperature for urban heat island monitoring. IEEE Geoscience and Remote Sensing Letters 10(5), 1011–1015.

Huang, B., Zhang, H., Song, H., Wang, J., Song, C., 2013. Unified fusion of remote-sensing imagery: Generating simultaneously high-resolution synthetic spatial–temporal–spectral earth observations. Remote Sensing Letters 4, 561–569.

Hilker, T., Wulder, M. A., Coops, N. C., Seitz, N., White, J. C., Gao, F., Masek, J. G, Stenhouse, G., 2009. 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(9), 1988-1999.

Hilker, T., Wulder, M. A., 2009. 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(8), 1613–1627.

Ju, J., Roy, D. P., 2008. The availability of cloud-free Landsat ETM plus data over the conterminous United States and globally. Remote Sensing of Environment 112, 1196–1211.

Li, A., Bo, Y., Zhu, Y., Guo, P., Bi, J., He, Y., 2013. Blending multi-resolution satellite sea surface temperature (SST) products using Bayesian maximum entropy method. Remote Sensing of Environment 135, 52–63.

Li, J., Li, Y., He, L., Chen, J., Plaza, A., 2020. Spatio-temporal fusion for remote sensing data: an overview and new benchmark. SCIENCE CHINA Information Sciences 63(4), 140301.

Li, X., Foody, G. M., Boyd, D. S., Ge, Y., Zhang, Y., Du, Yun., Ling, F., 2020. SFSDAF: An enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion. Remote Sensing of Environment 237, 111537.

Li, Y., Li, J., Lin, H., Jin, C., Antonio, P., 2020. A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks. SCIENCE CHINA Information Sciences 63(4), 140302.

Liu, M., Yang, W., Zhu, X., Chen, J., Chen, X., Yang, L., Helmer, E. H., 2019. An Improved Flexible Spatiotemporal DAta Fusion (IFSDAF) method for producing high spatiotemporal resolution normalized difference vegetation index time series. Remote Sensing of Environment 227, 74–89.

Liu, W., Zeng, Y., Li, S., Huang, W., 2020. Spectral unmixing based spatiotemporal downscaling fusion approach. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 88, 102054.

Liu, X., Deng, C., Wang, S., Huang, G., Zhao, B., Lauren, P., 2016. Fast and accurate spatiotemporal fusion based upon extreme learning machine. IEEE Geoscience and Remote Sensing Letters 13, 2039–2043.

Ma, J., Zhang, W., Marinoni, A., Gao, L., Zhang, B., 2018. An improved spatial and temporal reflectance unmixing model to synthesize time series of landsat-like images. Remote Sensing 10, 1388.

Meng, J. H., Du, X., Wu, B. F., 2013. Generation of high spatial and temporal resolution NDVI and its application in crop biomass estimation. International Journal of Digital Earth 6, 203–218.

Moosavi, V., Talebi, A., Mokhtari, M. H., Shamsi, S. R. F., Niazi, Y., 2015. A wavelet-artificial intelligence fusion approach (WAIFA) for blending Landsat and MODIS surface temperature. Remote Sensing of Environment 169, 243–254.

Mustafa, Y. T., Tolpekin, V. A., Stein, A., 2014. Improvement of spatio-temporal growth estimates in heterogeneous forests using Gaussian bayesian networks. IEEE Transactions on Geoscience and Remote Sensing 52(8), 4980–4991.

Pisek, J., Lang, M., Kuusk, J., 2015. A note on suitable viewing configuration for retrieval of forest understory reflectance from multi-angle remote sensing data. Remote Sensing of Environment 156, 242–246.

Roy, D. P.; Zhang, H. K.; Ju, J.; Gomez-Dans, J. L.; Lewis, P. E.; Schaaf, C. B.; Sun, Q.; Li, J.; Huang, H.; Kovalskyy, V., 2016. A general method to normalize Landsat reflectance data to nadir BRDF adjusted reflectance. Remote Sensing of Environment 176, 255–271.

Shen, H., Huang, L., Zhang, L., Wu, P., Zeng, C., 2016. 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 172, 109–125.

Shen, M., Tang, Y., Chen, J., 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(12), 0–1722.

Song, H., Liu, Q., Wang, G., Hang, R., Huang, B. 2018. Spatiotemporal satellite image fusion using deep convolutional neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(3) 1-9.

Tang, Y., Wang, Q., Zhang, K., Atkinson, P. M., 2020. Quantifying the effect of registration error on spatio-temporal fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 487-503.

Tewes, A., Thonfeld, F., Schmidt, M., 2015. Using RapidEye and MODIS data fusion to monitor vegetation dynamics in semi-arid rangelands in South Africa. Remote Sensing 7, 6510–6534.

Wang, J., Schmitz, O., Lu, M., Karssenberg, D., 2020. Thermal unmixing based downscaling for fine resolution diurnal land surface temperature analysis. ISPRS Journal of Photogrammetry and Remote Sensing 161, 76–89.

Weng, Q., Peng, F., Feng, G., 2014. Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data. Remote Sensing of Environment 145(8), 55-67.

Wang, Q., Atkinson, P. M., 2018. Spatio-temporal fusion for daily Sentinel-2 images. Remote Sensing of Environment 204, 31–42.

Wu, M. et al., 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(13), 06357.

Wu, P. H., Shen, H. F., Zhang, L. P., 2015. Integrated fusion of multi-scale polar-orbiting and geostationary satellite observations for the mapping of high spatial and temporal resolution land surface temperature. Remote Sensing of Environment 156, 169–181.

Xie, D., Gao, F., Sun, L., Anderson, M., 2018. Improving spatial-temporal data fusion by choosing optimal input image pairs. Remote Sensing 10(7), 1142.

Xu, Y., Huang, B., Xu, Y., Cao, K., Guo, C., Meng, D., 2015. Spatial and temporal image fusion via regularized spatial unmixing. IEEE Geoscience and Remote Sensing Letters 12(6), 1362–1366.

Zhang, H. K., Chen, J. M., Huang, B., 2014. Reconstructing seasonal variation of Landsat vegetation index related to leaf area index by fusing with MODIS data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7, 950–960.

Zhang, H. K., Huang, B., Zhang, M., Cao, K., Yu, L., 2015. A generalization of spatial and temporal fusion methods for remotely sensed surface parameters. International Journal of Remote Sensing 36(17), 4411–4445.

Zhao, C., Gao, X., Emery, W. J., Wang, Y., Li, J., 2018. An integrated spatio-spectral-temporal sparse representation method for fusing remote-sensing images with different resolutions. IEEE Transactions on Geoscience and Remote Sensing 56(6), 1-13.

Zhu, X., Chen, J., Gao, F., 2010. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. Remote Sensing of Environment 114(11), 2610–2623.

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

Zhu, X., Cai, F., Tian, J., 2018. Spatiotemporal fusion of multisource remote sensing data literature survey, taxonomy, principles, applications, and future directions. Remote Sensing 10(4), 527.

Zhu, X., Helmer, E. H., Chen, J., Liu, D., 2018. An automatic system for reconstructing high-quality seasonal Landsat time-series. Boca Raton, USA.

Zhu, Z., Woodcock, C. E., Holden, C., Yang, Z., 2015. Generating synthetic Landsat images based on all available Landsat data: Predicting Landsat surface reflectance at any given time. Remote Sensing of Environment 162, 67-83.

Zhukov, B., Oertel, D., Lanzl, F., 1999. Unmixing-based multisensor multiresolution image fusion. IEEE Transactions on Geoscience and Remote Sensing 37(3), 1212–1226.

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(3), 453–457.

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

Fig. 1. Flowchart of VIPSTF, where both spatial weighting (SW)- and spatial unmixing (SU)-based solutions (i.e., VIPSTF-SW and VIPSTF-SU) are illustrated.

Fig. 2. Partial data of Site 1. (a) L4. (b) L7. (c) L8. (d) L9. (e) L13. (f)-(j) are corresponding MODIS data.

Fig. 3. Partial data of Site 2. (a) L2. (b) L7. (c) L8. (d) L9. (e) L11. (f)-(j) are corresponding MODIS data.

Fig. 4. Results of different spatio-temporal fusion methods for Site 1 (M7-L7 as known image pair) (NIR, red, and green bands as RGB). (a) UBDF. (b) FSDAF. (c) STARFM. (d) VIPSTF-SW. (e) STDFA. (f) VIPSTF-SU. (g) Reference.

Fig. 5. Scatter plots of the actual and predicted values of the NIR band for Site 1 (M7-L7 as known image pair). (a) UBDF. (b) FSDAF. (c) STARFM. (d) VIPSTF-SW. (e) STDFA. (f) VIPSTF-SU.

Fig. 6. The prediction accuracy based on different image pairs for Site 1. (a) RMSE. (b) CC.

Fig. 7. The predictions based on different numbers of image pairs for Site 1.

Fig. 8. The accuracy of prediction by multiple image pairs for Site 1. (a) RMSE. (b) CC.

Fig. 9. The RMSE between images at the known and prediction times when using the original image pair and the VIP based on one image pair. (a) RMSE between MODIS images. (b) RMSE between Landsat images.

Fig. 10. The RMSE between images at the known and prediction times when using the original image pair and the VIP based on multiple image pairs. (a) RMSE between MODIS images. (b) RMSE between Landsat images.

Fig. 11. Scatter plots of reduction in the MODIS level increment (in terms of the difference between $\Delta\mathbf{M}$ and $\Delta\mathbf{M}'$) and the corresponding increase of prediction accuracy (in terms of RMSE decrease) for Site 1. (a) STARFM and VIPSTF-SW. (b) STDFA and VIPSTF-SU.

Fig. 12. Computational costs of the methods for Site 1.

Fig. 13. Results of different methods for Site 2 (M7-L7 as known image pair). (a) UBDF. (b) FSDAF. (c) STARFM. (d) VIPSTF-SW. (e) STDFA. (f) VIPSTF-SU. (g) Reference.

Fig. 14. The prediction accuracy based on different image pairs for Site 2. (a) RMSE. (b) CC.

Fig. 15. The CC between Landsat images at the known and prediction times.

Fig. 16. Variation in the absolute regression coefficient $|a|$. (a) $|a|$ of Landsat at different times (e.g., 14 images). (b) Scatter plot between |CC| and $|a|$ for the Red band.