Remote Sensing of Environment 283 (2022) 113309

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

VSDF: A variation-based spatiotemporal data fusion method

Chen Xu [a,b,c], Xiaoping Du [a,b,*], Zhenzhen Yan [a,b], Junjie Zhu [a,b], Shu Xu [d], Xiangtao Fan [a,b]

[a] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[b] International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China
[d] China Remote Sensing Satellite Ground Station, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

ARTICLE INFO

Edited by Jing M. Chen

Keywords:
Data fusion
Spatiotemporal fusion
Landsat
MODIS
Guided filter
Change detection

ABSTRACT

The fusion of spatiotemporal data provides the possibility to improve both the spatial and temporal resolution of remote sensing data. Nevertheless, the performance of current spatiotemporal data fusion methods is affected by several aspects, e.g., (1) retrieval of abrupt land cover changes, (2) recovery of detailed spatial information, and (3) the need to reduce side effects related to the performance differences between sensors. Concerning the above aspects, this study proposes the use of a Variation-based Spatiotemporal Data Fusion (VSDF) method. In VSDF, an abundant variation classification (AVC) is used to identify explicitly the land cover changes for spectral unmixing. In addition, feature-level fusion is introduced to strengthen the spatial structures, i.e., the edges and texture. Furthermore, a relative reliability index (RRI) is proposed to guide the prediction process to lower the uncertainty of the input datasets. With reference to the all-round performance assessment (APA) metrics, the performance of VSDF was compared with five popular methods, i.e., Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM), Flexible Spatiotemporal DAta Fusion (FSDAF), FSDAF 2.0, Reliable and Adaptive Spatiotemporal Data Fusion (RASDF), and Fit-FC (regression model Fitting, spatial Filtering and residual Compensation). The experimental results demonstrated that VSDF can realize a more accurate prediction of temporal land cover changes with better spatial detail than the benchmark methods. Consequently, VSDF has the potential to generate accurate high-spatiotemporal-resolution simulations for global remote sensing studies.

1. Introduction

The scale of remote sensing research and applications has expanded from the single time phase to multiple time phases, and from discrete regions to the global scale (Guo et al., 2021). However, high-resolution earth observation data with both spatial and temporal continuity is limited due to there being insufficient remote sensing sensors and unstable observation conditions (Claverie et al., 2018; Wulder et al., 2015). Sensors with a fine spatial resolution of 10–30 m (such as Sentinel and Landsat series satellites) typically have long revisit periods of around 15 days. In addition, due to the influence of clouds and cloud shadows, it is difficult to ensure the availability of data in a specified area and time, which severely limits the application of remote sensing data. Hence, there is an urgent need to improve the temporal resolution of remote sensing data to support large-scale remote sensing research and applications, e.g., land cover change monitoring, crop yield, and urban dynamics mapping (Hansen and Loveland, 2012; Kang and Özdoğan, 2019; Zhao et al., 2020). The fusion of spatiotemporal data provides a practical approach to improving spatiotemporal resolution, whereby the fusion simulates missing high-resolution datasets by blending high-spatial and high-temporal remote sensing datasets (Zhu et al., 2018). Spatiotemporal data fusion can obtain simulations of unavailable remote sensing data at a low cost. In recent years, spatiotemporal data fusion has been used in large-scale remote sensing research and applications (Li and Long, 2020; Liu et al., 2021; Liu et al., 2019).

In the past decade, various novel spatiotemporal data fusion methods have been developed, and they can generally be categorized into two groups: learning-based and non-learning-based methods (Zhou et al., 2021; Zhu et al., 2018). Learning-based methods, e.g., dictionary-pair learning and deep learning, implicitly construct the relationships between the fused results and the input datasets by learning the training datasets (Cai et al., 2022). For instance, SParse-representation-based SpatioTemporal reflectance Fusion Model (SPSTFM) estimates the fine images by establishing a transforming model for the change between

* Corresponding author at: Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China.
E-mail address: duxp@aircas.ac.cn (X. Du).

https://doi.org/10.1016/j.rse.2022.113309
Received 24 June 2022; Received in revised form 27 September 2022; Accepted 8 October 2022
Available online 22 October 2022

two coarse and fine image pairs (Huang and Song, 2012). The GAN-based SpatioTemporal Fusion Model (GAN-STFM) predicts a fine image using a conditional generative adversarial network and switchable normalization technique, which minimizes the input datasets for the predictions (Tan et al., 2022). Learning-based methods have become popular in recent years due to their reliable prediction accuracy and processing efficiency. However, most prediction models, especially deep learning-based models, must be pre-trained with additional datasets, which is time-consuming and restricts their use in large-scale applications (Ao et al., 2022).

Non-learning-based methods explicitly build the relationships between the input and fused images by exploiting remote sensing laws and assumptions, i.e., the linear mixing model, Tobler's first law of geography (Tobler, 1970), and temporal dependence (Zhu et al., 2018). Most existing non-learning-based methods follow two basic approaches, i.e., unmixing and the weight function (Gevaert and García-Haro, 2015; Hilker et al., 2009). The multisensor multiresolution technique (MMT) predicts the fine pixels by unmixing the coarse pixels based on linear spectral mixing theory, and this is regarded as the first method to introduce unmixing to spatiotemporal data fusion (Zhukov et al., 1999). The weight function approach predicts the fine pixels by averaging the neighborhood information within a window, first introduced in STARFM (Gao et al., 2006). Additionally, some other methods estimate the spatiotemporal data fusion with Bayesian estimation theory (Li et al., 2013; Shen et al., 2016). Most advanced non-learning-based spatiotemporal data fusion models are based on mixed models of multiple basic methods (Zhu et al., 2016). For example, Fit-FC constructs a linear regression model by considering spectral information from coarse images at T_1 and T_2, and refines the prediction by conducting spatial filtering along with the weight function (Wang and Atkinson, 2018). Fit-FC performs well in fusing rapid temporal change with considerable spectral accuracy. FSDAF is a well-known mixed spatiotemporal data fusion method that takes the abrupt land cover changes into account (Zhu et al., 2016). FSDAF 2.0 is an optimized version of FSDAF, which improves the predictions of spatial detail and areas of land cover changes by further evaluating the pixels that have changed after unmixing (Guo et al., 2020). In addition, RASDF introduces an adaptive local unmixing model that can retrieve strong temporal changes before filtering and distributing the residuals (Shi et al., 2022).

Currently, spatiotemporal data fusion is still facing several challenges, and in this research, we focus on the following three aspects: (1) Retrieval of abrupt land cover changes. The abrupt changes (e.g., land cover changes caused by floods or human activities) are more difficult to estimate compared to synchronized changes (e.g., the growth of vegetation and the change in soil moisture). This is because these changes can be totally different from similar neighboring pixels, and there is limited information with which to estimate abrupt changes if the size of the changes is comparable to the coarse pixels (Zhu et al., 2018). Early spatiotemporal data fusion methods (e.g., STARFM) did not consider abrupt land cover changes, and it remains challenging to estimate and recover abrupt changes. (2) Recovering detailed spatial structures. Given that there is not sufficient fine-resolution information at the time of the prediction, it is difficult to recover precisely the spatial structures, e.g., small shapes, texture, and edges. Some methods introduce neighboring information and optimize the prediction by pixel-level fusion methods, e.g., FSDAF and Fit-FC. However, pixel-level fusion methods will smooth these structures producing “blur” artifacts (Zhang, 2010). (3) Reducing the side effects related to performance differences between sensors. Typically, there are geometric and spectral differences between coarse and fine images due to the use of different sensors, the sun elevation, etc., which will give rise to uncertainty in the prediction (Zhou et al., 2021). Some spatiotemporal data fusion models pre-process input datasets before the fusion process to reduce systematically such differences and reduce the side effects. However, errors in the input datasets are inevitable, and the magnitude of the differences needs to be further considered during the prediction to determine intelligently the fusion strategy to lower the possible introduction of errors.

To cope with the aforementioned problems, a novel spatiotemporal data fusion model named the Variation-based Spatiotemporal Data Fusion (VSDF) model is proposed. By considering systematically the temporal changes, introducing feature-level information, and exploiting the reliability of coarse images, VSDF can accurately capture abrupt land cover changes in heterogeneous areas and estimate the detailed spatial structure. The proposed method requires only minimal input datasets, that is, one fine image and two coarse images. To evaluate the proposed method, experiments were conducted with 31 Landsat/MODIS datasets and the performance of VSDF was compared with five other popular models, i.e., STARFM, FSDAF, FSDAF 2.0, RASDF, and Fit-FC. The remainder of the paper is organized as follows: Section 2 describes the principles of VSDF. Section 3 introduces the experimental datasets and the design of the experiments, and Section 4 presents the results. Finally, a discussion and conclusions are given in Section 4 and Section 5, respectively.

2. Method

VSDF fuses one pair of coarse/fine images at T_1 (C_1 and F_1) and one coarse image at T_2 (C_2) to estimate the fine image at T_2 (F_2). There are four main steps, that is, evaluating the input data, unmixing and estimating the preliminary prediction, distributing the residuals, and edge fusion, as shown in Fig. 1 Also, the main abbreviations and symbols mentioned in this section are listed in Table 1.

2.1. Evaluating the relative reliability index (RRI) of input datasets

According to the spectral unmixing theory, the value of a coarse pixel should be equal to the average value of the fine pixels. Given that the coarse and fine images are from two different satellite sensors, there are errors due to the systematic differences between the sensors and capturing conditions (e.g., resolution, solar geometry, spectral band range, and capturing angle). These errors are known collectively as the spectral error, which may be used to describe the errors in the coarse images. Considering that only one set of fine/coarse images at T_1 is available, we can describe the spectral error between the fine/coarse datasets with the root mean squared error (RMSE):

$$\mathrm{RMSE}_{C_1,F_1^C} = \frac{1}{B}\sum_{b=0}^{B}\sqrt{\frac{\sum_{n=1}^{N}(C_{1,n,b}-F_{1,n,b}^C)^2}{N}} \tag{1}$$

where C_1 is the coarse image at T_1; F_1^C is the fine image at T_1 (F_1) upscaled to the coarse resolution; B is the band number of the input datasets; N is the pixel number in the coarse images; $C_{1,n,b}$ and $F_{1,n,b}^C$ are the b band value of pixel n of C_1 and F_1^C, respectively.

$\mathrm{RMSE}_{C_1,F_1^C}$ can describe quantitatively the reliability of the input datasets. For example, linear spectral mixing theory assumes that the value of the coarse pixel equals the average value of all the fine pixels inside, which means that the $\mathrm{RMSE}_{C_1,F_1^C}$ should be zero. However, we can describe the difference in coarse images between T_1 and T_2 in the same way, that is:

$$\mathrm{RMSE}_{C_1,C_2} = \frac{1}{B}\sum_{b=0}^{B}\sqrt{\frac{\sum_{n=1}^{N}(C_{1,n,b}-C_{2,n,b})^2}{N}} \tag{2}$$

where C_2 is the coarse image at T_2; and $C_{2,n,b}$ is the b band value of the n pixel of C_2.

RMSE_{C_1,C_2} describes the intensity of spectral change between T_1 and T_2. The more significant change occurs, the higher RMSE_{C_1,C_2} is. Finally, the relative reliability index (RRI) of the input datasets can be defined as:

$$\mathrm{RRI} = \frac{\mathrm{RMSE}_{C_1,C_2}}{\mathrm{RMSE}_{C_1,F_1^C}} \tag{3}$$

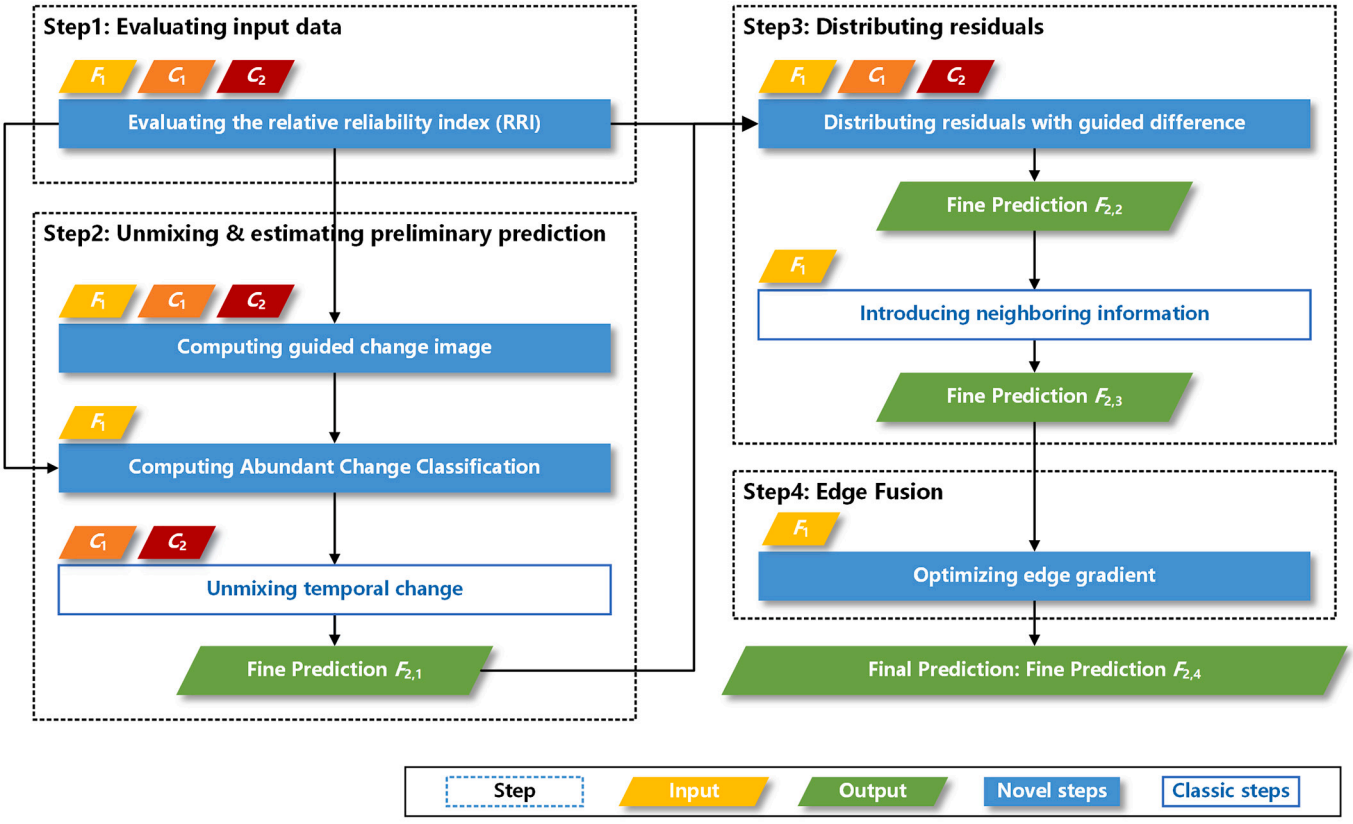

Fig. 1. The flowchart of the VSDF method.

Table 1
Definitions of main abbreviations and symbols.

Variable	Description
T_1	Time with an available fine image and coarse image pair
T_2	Time expected to predict the fine image
C_1/C_2	Coarse image (MODIS) at T_1/T_2
F_1/F_2	Fine image (Landsat) at T_1/T_2
$\Delta C/\Delta F$	Difference of coarse/fine images between T_1 and T_2
$F_{2,1}/F_{2,1}/F_{2,3}$	Intermediate predictions of VSDF at T_2
$F_{2,4}$	Final prediction of VSDF at T_2
F^C	Fine image upscaled to the coarse resolution
C^F	Coarse image downscaled to the fine resolution
RRI	Relative reliability index (Eq. (3))
RMSE	Root mean squared error
VC	Variation classification
AVC	Abundant variation classification
GF	Guided filter (He et al., 2010, 2013)
APA	All-round performance assessment (Zhu et al., 2022)

RRI represents the relative possibility of estimating accurately the variation from the input fine/coarse datasets. To be specific, $RMSE_{C_1, F_1^C}$ describes the reliability of the coarse images. Coarse images are the only information source that may be used for detecting temporal changes and the accuracy of the coarse images affects the ability to capture accurately the temporal variation. However, $RMSE_{C_1, C_2}$ represents the changing intensity during the fusion. Thus, the lower the RRI is, the more error there is in comparison with the variation, which means there is less possibility of estimating the variation accurately. Consequently, the information from the coarse images should be considered less, while more attention should be given to evaluating the fine images. For instance, for datasets of a few days apart, there are few temporal changes and the error of the coarse images ($RMSE_{C_1, F_1^C}$) could be larger than the variation ($RMSE_{C_1, C_2}$), where the RRI might be <1. Under such circumstances, the fine image at T_1 (F_1) should be emphasized more during the prediction compared with the coarse images (C_1 and C_2). In VSDF, the RRI guides the fusion process and sets the key parameters by determining which information should receive more attention.

2.2. Unmixing and estimating the preliminary prediction

It is assumed that all pixels of the same class exhibit a similar change in reflectance. Following the methodology of the unmixing method, VSFD obtains the preliminary prediction of F_2. For a given band b, the simulated value of a fine pixel at (x_i, y_j) can be given by:

$$F_2(x_i, y_j, b) = F_1(x_i, y_j, b) + \Delta F(c, b) + \varepsilon \quad (4)$$

where c is the class of the pixel at (x_i, y_j); $\Delta F(c, b)$ indicates the change in the value of class c in band b; ε is the residual for the local difference between $\Delta F(x_i, y_j, b)$ and $\Delta F(c, b)$.

2.2.1. Variation classification (VC)

The class c determines the change in value $\Delta F(c, b)$ for the fine pixel at (x_i, y_j), and which can be estimated by unmixing ΔC. Previous research estimates the class as land cover types at T_1 by classifying F_1. However, if a land cover change occurs, $\Delta F(x_i, y_j, b)$ will no longer be equal to $\Delta F(c, b)$ because the class of the fine pixel at T_2 is not the same class as at T_1. More specifically, the $\Delta F(x_i, y_j, b)$ should not be similar to the changes of other fine pixels with the same class at T_1. In this regard, we define the variation classification (VC), as shown in Fig. 2. Any pixel classified as the same class of VC should be with the same land cover at both T_1 and T_2, e.g., pixel 2 and pixel 6. Under such circumstances, the $\Delta F(c, b)$ for each class of VC can be considered approximately the same.

2.2.2. Guided change image

The fine-resolution land cover at T_1 can be estimated with F_1, but there is no information to indicate the fine-resolution land cover at T_2. The critical problem is identifying the temporal changes while only the coarse image is available at T_2. Some methods predict the temporal

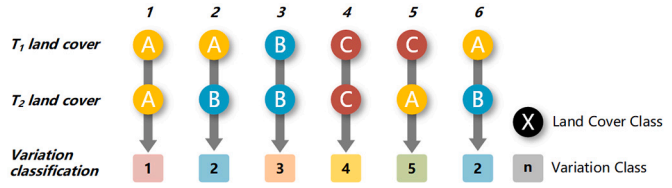

Fig. 2. Variation classification. Any pixel with the same land cover class at both T_1 and T_2 are regarded as the same class of variation classification.

variation at fine-resolution by downscaling the coarse-resolution images with thin-plate spline (TPS) interpolation, which might induce over-smoothed edges and the loss of spatial details (Guo et al., 2020; Wang and Atkinson, 2018; Zhu et al., 2016). To solve this problem, the guided change image with the Guided Filter (GF) is implemented in the present method to introduce the local spatial details from F_1 to guide the downscaling of ΔC. The GF is an efficient filter algorithm that can retain the edge and texture information from the guided image (He et al., 2010, 2013). The GF assumes that the filter acts as a local linear model between the guided fine image and the filtering output. Thus, the filtered fine-resolution image can be estimated under the guidance of the fine image at T_1. Next, the process of estimating the guided change images based on the use of the GF is introduced. A pixel within the local window $\omega_{x_i,\,y_j,\,r}$ at the center of (x_i,y_j) is given by:

$$\Delta F_{guided}(n,b,\omega_{x_i,y_j,r}) = a_{x_i,y_j,b,r}F_1(n,b) + b_{x_i,y_j,b,r} \tag{5}$$

where r is the radius of the local window; n is the index of the pixel in the local window; $a_{x_i,\,y_j,\,b,\,r}$ and $b_{x_i,\,y_j,\,b,\,r}$ are the linear coefficients computed with the pixels in the local window, which is given by:

$$a_{x_i,y_j,b,r} = \frac{\frac{1}{N_\omega}\sum_{n=1}^{N_\omega}(F_1(n,\mathrm{b}) - \mu_\omega)\Delta C^F(n,b)}{\sigma_\omega^2 + \epsilon} \tag{6}$$

$$b_{x_i,y_j,b,r} = \frac{1}{N_\omega}\sum_{n=1}^{N_\omega}\Delta C(x_i,y_j) - a_{x_i,y_j,r}\mu_\omega \tag{7}$$

where N_ω is the number of pixels in the local window $\omega_{x_i,\,y_j,\,r}$; ΔC^F is the difference in the coarse images between T_1 and T_2 downscaled to fine resolution and is regarded as the filtered input of the GF; μ_ω and σ_ω^2 are the mean and variance of F_1 in local window $\omega_{x_i,\,y_j,\,r}$; ϵ is a regularization parameter preset by the user. The detailed derivation of the GF is elaborated in (He et al., 2010, 2013).

Then, the final $\Delta F_{guided}(x_i,y_j,b)$ is the average of $\Delta F_{guided}(n,b,\omega_{x_i,\,y_j,\,r})$ in all local windows. Overall, Eqs. (5)–(7) can be summarized as:

$$\Delta F_{guided} = GuidedFilter(F_1,\Delta C^F) \tag{8}$$

where ΔF_{guided} is the guided change image; ΔC^F is the difference of the coarse images between T_1 and T_2 downscaled to coarse resolution; *GuidedFilter* is the process of the GF.

Eq. (5) which is linear ensures that the guided change image ΔF_{guided} retains the texture from the fine image at T_1 because $\nabla(\Delta F_{guided}) = a\,\nabla F_1$, which means that the guided change image approaches the spatial details with the given fine image. Moreover, ΔF_{guided} is derived from the ΔC^F, which means that the guided change image retains the spectral information with the coarse images. Compared with interpolation implemented in other research (e.g., TPS or bicubic interpolation), the guided change image reserves more detailed information on features such as edges and texture. However, the guided change image cannot describe accurately the quantitative variation during fusion. First, the guided change image is still over-smoothed compared to the ground truth change image. In addition, the filtering input ΔC^F and the guided image F_1 cannot strictly meet the assumption of the local linear model required by the GF. More importantly, the shape of edges can change from T_1 to T_2 in case of an abrupt land cover change. As a result, there might be unreal gradient errors. Hence, the guided change image cannot be used to estimate the VC. To cope with this problem, the abundant variation classification (AVC) is proposed, which is not sensitive to the accuracy of the guided change images at a quantitative level.

2.2.3. Abundant variation classification (AVC)

Assuming that the guided change image is unreliable, it is impossible to estimate accurately the temporal change classification, hence VSDF uses the abundant variation classification (AVC) instead. AVC is the joining of the classification of land cover at T_1 and the classification with the guided change image, whereby we classify the pixels into as many classes as possible. As shown in Fig. 3, the pixels of the same class from VC might be further estimated as different classes of AVC. For AVC, the error caused by over smoothing and assigning the incorrect edge value might induce more classes instead of mixed classes. As a result, $\Delta F(c,b)$ will be coordinated with the AVC variation of the pixel. Although there will be some $\Delta F(c_n,b)$ which equal other $\Delta F(c_m,b)$, this does not, in theory, affect the accuracy.

VSDF estimates the AVC by implementing the well-known unsupervised clustering algorithm K-Means clustering (Hartigan and Wong, 1979), feeding the combined array of the fine image at T_1 and the guided change image. K-Means can be replaced by other clustering algorithms, e.g., fuzzy C-means (FCM) (Bezdek et al., 1984), but image segmentation (e.g., SLIC (Achanta et al., 2012)) is not recommended here, given that it considers the spatial information during clustering and because pixel-based clustering can concentrate on the spectral information and avoid the spatial-smoothed classification results.

The cluster number, or the number of AVC classes, is a crucial parameter for unsupervised clustering. Compared with the land cover classification, there should be more clusters for AVC than the number of land cover types. The number of clusters should be closely related to the intensity of the temporal change besides the land cover at T_1. The cluster number is set empirically in many fusion methods by the users, e.g., LMGM (Rao et al., 2015) and UBDF (Zurita-Milla et al., 2008). Additionally, some studies have determined the number of clusters based on the fine image T_1, e.g., (Peng et al., 2022). However, information from a single time does not contain the information on land cover changes. To solve this problem, the number of clusters is determined using a semi-empirical equation with the RRI:

$$n_{\mathrm{AVC}} = (3 - \frac{1}{\mathrm{RRI}}) \times 2n_F \tag{9}$$

where n_F is the value of the empirical number; 5 is adopted in the present work.

As mentioned above, the higher the RRI is, the more variation there is in the information that can be obtained, and the more the possibility to estimate the temporal changes. In contrast, a low RRI means that sufficient reliable information is not available for estimating the AVC compared with the uncertainty of the input datasets, and fewer numbers of the cluster are considered for classification.

2.2.4. Unmixing temporal change

According to the linear spectral mixing theory, the temporal change of a coarse pixel at (x_I,y_J) can be expressed as:

$$\Delta C(x_I,y_J,b) = \sum_{c=1}^{n_{\mathrm{AVC}}}\mu(x_I,y_J,c) \times \Delta F_{AVC}(c,b) \tag{10}$$

where n_{AVC} is the number of abundant variation classes in the coarse pixel at (x_I,y_J); $\mu(x_I,y_J,c)$ is the number of fine pixels of class c within the coarse pixel; b is the spectral band.

There is an equation for each coarse pixel as in Eq. (10), and we can estimate each $\Delta F_{AVC}(c,b)$ for all classes by the least square method. Subsequently, a preliminary prediction $F_{2,\,1}$ using Eq. (4) can be obtained as:

$$F_{2,1}(x_i,y_j,b) = F_1(x_i,y_j,b) + \Delta F_{AVC}(c,b) + \varepsilon \tag{11}$$

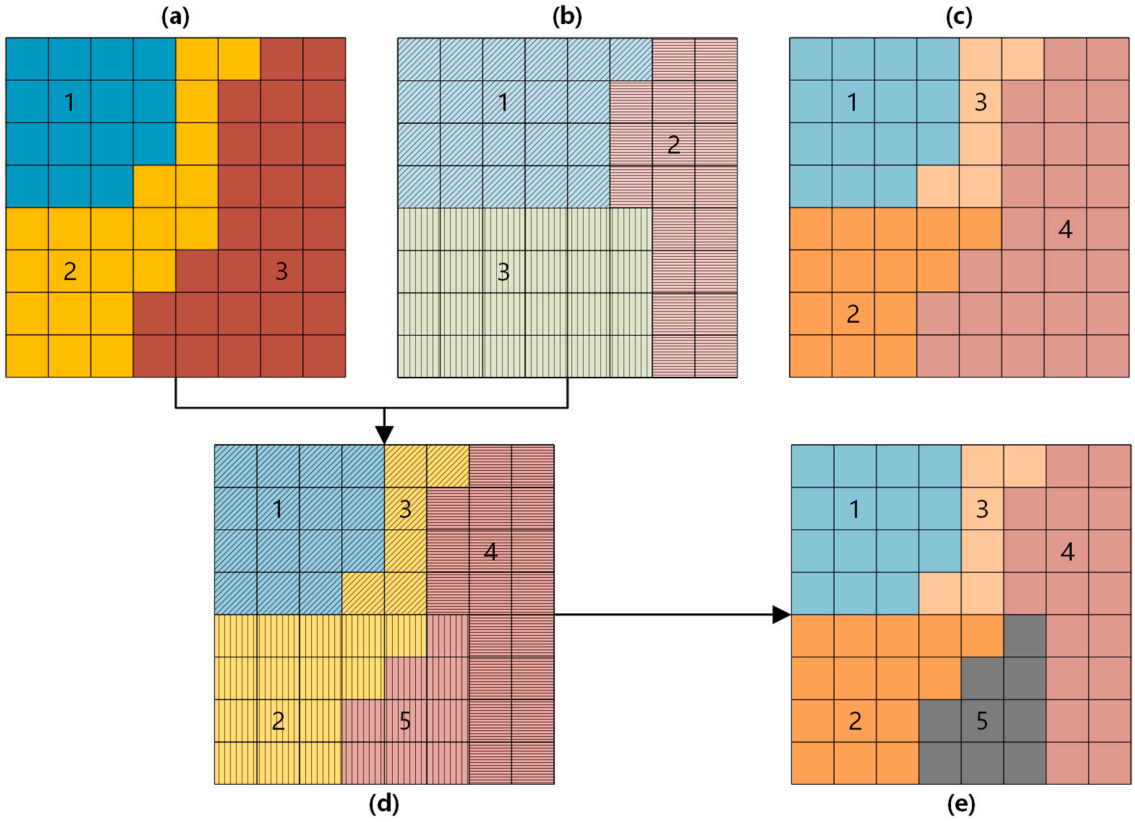

Fig. 3. Diagram of AVC at fine resolution: (a) land cover classification at T_1; (b) classification with guided change image; (c) real VC (target classification); (d) and (e) joint result of (a) and (b) (AVC). Compared with (a) and (b), (e) contains more classes: there are three classes of land cover at T_1 in (a), and during the fusion, there are four classes of temporal changes in (c), where the pixels of class 2 in (a) have reverted to class 2 and class 3 in (c). Although the edges of the guided change image are over-smoothed, the detailed information can be compensated with land cover classification at T_1. For example, the information for class 3 in (e) cannot be decided in (b), but it may be recognized with the information from (a). However, this process might induce over-classification. For instance, class 3 in (b) is recognized as class 2 and class 5 in (e). Class 2 in (e) corresponds precisely to class 2 in the real VC (c), whereas class 5 is an additional class.

2.3. Distributing the residuals

The residuals (ε in Eq. (11)) of the preliminary prediction F_2^1 are caused mainly by (1) the difference between $\Delta F_{AVC}(c,b)$ and the real $\Delta F(x_i,y_j,b)$; (2) the errors of AVC. The residuals between $\Delta F_{AVC}(c,b)$ and the real $\Delta F(x_i,y_j,b)$ will cause the loss of detailed texture by averaging the temporal change of the pixels from the same AVC. However, the errors of AVC will cause artifacts in the predictions of $F_{2,1}$. To reduce the aforementioned errors, VSDF distributes the residuals and introduces the neighboring information in this step.

2.3.1. Distributing the residuals with the guided difference

Given that only the coarse image at the prediction date is available, the residuals can be obtained at coarse resolution as:

$$C_{residual} = C_2 - F_{2,1}^C \tag{12}$$

where $F_{2,1}$ is the preliminary prediction from Eq. (11); $F_{2,1}^C$ is the upscale of $F_{2,1}$ at coarse resolution.

Some studies distribute the residuals by interpolation to detect the temporal change. However, in the present method, the temporal change is considered during the unmixing. Moreover, we are more interested in recovering the texture and removing the block effects. Considering that detailed spatial information is not available at T_2, it is reasonable to introduce the information from T_1. VSDF migrates the texture and edges from the fine image at T_1 to guide the distribution of the residuals. To be specific, we downscale the residuals into fine-resolution by the GF to preserve the image structure from the fine image at T_1 as:

$$F_{2,2} = F_{2,1} + GuidedFilter(F_1, C_{residual}) \tag{13}$$

The distribution of the residuals can be processed several times. Theoretically, if the coarse image at T_1 (C_2) is strictly accurate, the distribution can always approach the prediction and give a stable and optimal result. However, there are avoidable errors in the coarse images, and the prediction will be misled. Under such circumstances, the introduction of information from the coarse images (residuals) should be restricted. To balance the errors and make full use of the residuals, the loop times of Eq. (13) are determined according to the RRI:

$$t_{loop} = \begin{cases} n_{max}\left(1 - \dfrac{1}{\mathrm{RRI}}\right)^2, & \mathrm{RRI} \geq 1 \\ 0, & \mathrm{RRI} < 1 \end{cases} \tag{14}$$

When the RRI < 1, the error of the coarse image is larger than the temporal change and it is considered that $C_{residual}$ cannot provide positive information, and distribution of the residuals should not be performed. In contrast, when the error is ignored compared with the temporal change, the loop times are set as n_{max}, because it is found that the output will tend to be stable when the loop times are greater than a specific number. In present experiments, the recommended n_{max} is 5.

2.3.2. Introducing neighboring information

The AVC will produce abnormal artifacts, which leads to unreasonable unmixing results in some areas. In most cases, such pixels are sporadic and can be fixed by introducing neighboring information. The introduction of neighboring information originated from the weight function proposed in STARFM. In this study, the same strategy as in FSDAF is applied, the elaboration of which can be referred to (Zhu et al., 2016). In that work, n_s similar pixels were selected within the window centered at (x_i, y_j) according to the spectral similarity. Different from FSDAF, the selection of similar pixels is not limited to the pixels in the same class as the central pixel. Then, the central pixel is replaced by the weighted average of all similar pixels in the window. The weights are evaluated by the spatial distance between similar pixels and the central pixel. Accordingly, the abnormal pixels can be fixed by the neighboring pixels, and the optimized result is given by:

$$F_{2,3}(x_i, y_j, b) = F_1(x_i, y_j, b) + \sum_{k=1}^{n_s} w_k \times \Delta F_{2,2}(x_k, y_k, b) \quad (15)$$

where n_s is the number of similar pixels in the window centered at (x_i, y_j); $\Delta F_{2,\,2}$ is the difference of the band value between $F_{2,\,2}$ and F_1; w_k is the weight of the value of (x_k, y_k), which is related to the distance between (x_k, y_k, b) and (x_i, y_j, b).

2.4. Edge fusion

Although introducing the neighboring information can improve the spectral accuracy, it will also smooth the texture and edges. Given that there is no available information at T_2 to indicate the edges, it is reasonable to transfer the edges from the fine image at T_1 to the prediction. We can obtain the F_2^{Edge} by filtering the prediction image under the guidance of the fine image at T_2:

$$F_2^{Edge} = F_1 + GuidedFilter(F_1, \Delta F_{2,3}) \quad (16)$$

where $\Delta F_{2,\,3}$ is the difference between $F_{2,\,3}$ and F_1.

The GF can effectively transfer the presence of the edges from the guidance image. However, it should be noted that the gradient is not transferable when the local linear assumption is not satisfied. To avoid this problem, VSDF adopts feature-level fusion to preserve the texture information. According to the spectral mixing theory, the pixels on the edges are probably mixed types of adjacent pixels, which are coordinated with the local linearity. Hence, it is reasonable to replace the pixels at the edges and texture with the guided prediction. The edges and significant texture are first extracted as features by the Canny edge detection algorithm, a robust edge detection algorithm (Canny, 1986). Then, VSDF only optimizes $F_{2,\,3}$ with the pixel from F_2^{Edge} if the pixel is recognized as an edge by the Canny edge detection algorithm such that the final prediction of VSDF at T_2 is obtained as:

$$F_{2,4}(x_i, y_j) = \begin{cases} F_2^{Edge}(x_i, y_j), & (x_i, y_j)\ is\ edge \\ F_{2,3}(x_i, y_j), & (x_i, y_j)\ is\ not\ edge \end{cases} \quad (17)$$

3. Test experiment

3.1. Datasets

Experiments were performed using the Moderate Resolution Imaging Spectroradiometer (MODIS) and the Landsat datasets. MODIS provides daily global observations, but the spatial resolution of MODIS is >200 m. In contrast, the Landsat series satellite data has a resolution of 30 m, but there is a long revisit cycle of around 16 days (Masek et al., 2020). Specifically, Landsat datasets were adopted as the fine images, and six bands of Landsat 5 TM (bands 1, 2, 3, 4, 5, and 6) were stacked and resampled to 25 m spatial resolution before prediction. Six bands (bands 3, 4, 1, 2, 6, and 7) of MOD09GA images, which provided the daily 500 m surface reflectance data for the globe, were used as the coarse images. The MOD09GA datasets were re-projected from sinusoidal projection to the Universal Transverse Mercator (UTM) project before prediction (Emelyanova et al., 2013).

Two sites were used to evaluate the performance of VSDF. The first site was the Coleambally Irrigation Area (CIA), located in the southern part of New South Wales, Australia. The CIA site is heterogeneous and covered by croplands and woodlands. The second site was the Gwydir Catchment (GWY) in the northern part of New South Wales, Australia, which is more homogeneous than the CIA. First, we elaborate on the process of VSDF with an inundated area from GWY acquired on November 26, 2004 (T_1) and December 12, 2004 (T_2). As shown in Fig. 4, the Landsat and simulated-MODIS data, which were upscaled from the Landsat images, were used. Additional datasets were also used to analyze the all-round performance of the proposed method. Two regions in the CIA and three regions in the GWY were selected for evaluation, and each region contained 800*800 pixels, as shown in Fig. 5. To evaluate the performance for different intensities of temporal change, four pairs of Landsat/MODIS images were selected, for which one pair was used as the base pair, and the other three pairs were expected to be fused. To be specific, the base image pair of the CIA was acquired on October 8, 2001, and the three pairs to be fused were acquired on October 17, 2001, November 25, 2001, and January 5, 2002, respectively; the base image pair of the GWY was acquired on April 16, 2004, and the three pairs to be fused were acquired on May 2, 2004, July 5, 2004, and August 22, 2004, respectively. For convenience, the three pairs with different time intervals were called close, medium, and far time interval datasets. Apart from the real Landsat/MODIS pairs, tests were also implemented with the real Landsat images and the simulated MODIS image pairs, which were upscaled from the corresponding Landsat images. The Landsat/simulated-MODIS pairs have been applied widely in research to evaluate performance because they avoid errors in the input datasets and satisfy strictly the basic assumption of spatiotemporal data fusion. Hence, the evaluation is only related to the robustness of the algorithms instead of any other uncertainties. Overall, 31 datasets (1 + 5 regions * 2 types of MODIS * 3 temporal intervals) were used to evaluate thoroughly the performance of VSDF.

3.2. Experimental design

For comparison purposes, experiments with the datasets were conducted using VSDF and five other widely-used spatiotemporal data fusion methods, i.e., STARFM, FSDAF, FSDAF 2.0, Fit-FC, and RASDF. The parameters for each of the methods were selected carefully for the CIA and GWY sites as recommended (Zhou et al., 2021) for the optimal predictions. In particular, the parameters for RASDF were set to the default because the default parameters have already been optimized for the CIA and GWY landscapes. For VSDF, only similar pixel numbers for neighboring pixels and the moving window size for searching similar neighboring pixels were comparable with the other algorithms, and both of the parameters were not optimized in order to focus on the comparison with the novel VSDF process. Full details of the parameters selected are given in Table 2. All experiments were carried out on the ScienceEarth Platform (Xu et al., 2022a, 2022b).

To evaluate comprehensively the performance of the methods, the all-round performance assessment (APA) proposed by Zhu et al. (2022) was adopted to evaluate the fusion results with respect to the global spectral accuracy, the pixel spectral accuracy, the edge, and texture. The APA metric includes four indexes, i.e., the root-mean-square error (RMSE), the average difference (AD), Robert's edge (EDGE), and the local binary patterns (LBP) (Table 3). For each index, the optimal value is 0.

4. Results

4.1. Elaborating the process of VSDF with the inundated GWY region

In this test, the Landsat/simulated MODIS datasets of Fig. 4 were used to elaborate on the process of VSDF with visual and statistical comparisons. The detailed prediction and APA performance metrics for each (prediction) method, are presented in Fig. 6 and Table 4, respectively.

VSDF estimated the preliminary prediction $F_{2,\ 1}$ (Fig. 6(b)) by unmixing the temporal difference referring to AVC. In general, it was found that the preliminary prediction recovered the land cover changes caused by flooding. Fig. 7 shows the guided change image which was used to estimate AVC. It is clear that the guided change image (Fig. 7(c)) captured the inundated area with better spatial detail compared to the original difference at coarse resolution (Fig. 7(a)). Fig. 8 demonstrates different classifications by K-Means. VSDF used the combination of F_1 and the guided change image (ΔF_{guided}) as input, which successfully clustered the pixels according to their temporal change features (Fig. 8 (a)). In comparison, classification with F_1 failed to distinguish between the inundated and non-inundated areas (Fig. 8(b, c)). Then, VSDF distributed the local residuals in each coarse-pixel and produced the $F_{2,\ 2}$ (Fig. 6(c)). VSDF guided the distribution of the residuals with the gradient of F_1, as shown in Fig. 9. Compared with the original residuals at coarse resolution (Fig. 9(a)), the guided-residuals (Fig. 9(b)) contained more spatial information referring to the reference (Fig. 9(c)); moreover, the process sharpened the texture and caused significant augmentation of the standard local binary pattern (LBP) in $F_{2,\ 2}$ (Table 4). Subsequently, by introducing the neighboring information, VSDF optimized the small artifacts caused by the errors in AVC (e.g., marked in red in Fig. 6(2)(d)) and produced $F_{2,\ 3}$. In addition, introducing the neighboring information induced smoothness in the edges and the texture, and both EDGE and LBP decreased in $F_{2,\ 3}$. Finally, the edges were recovered from F_1 in the last step and this caused an increase of EDGE. For instance, the unreal transition in $F_{2,\ 3}$ was optimized in $F_{2,\ 4}$, e.g., the area marked in red in Fig. 6(2)(e).

The fusion predictions by the proposed VSDF and the five other spatiotemporal data fusion methods are shown in Fig. 10. It was found that all methods estimated successfully the general boundary of the inundated area. The zoomed region marked in yellow in Fig. 10(b) is represented in Fig. 11. Unfortunately, all predictions failed to capture precisely part of the detailed texture in the inundated areas (e.g., the rectangular area marked in red in Fig. 11(b)), which was caused by the absence of fine-resolution information of the texture at the prediction time. In addition, it is apparent that the edges and texture of FSDAF and FSDAF 2.0 were over-smoothed, e.g., the regions marked in red in Fig. 11(e, f). There were clear spectral errors in the predictions of RASDF, Fit-FC, and STARFM, which are marked in red in Fig. 11(d, g, h).

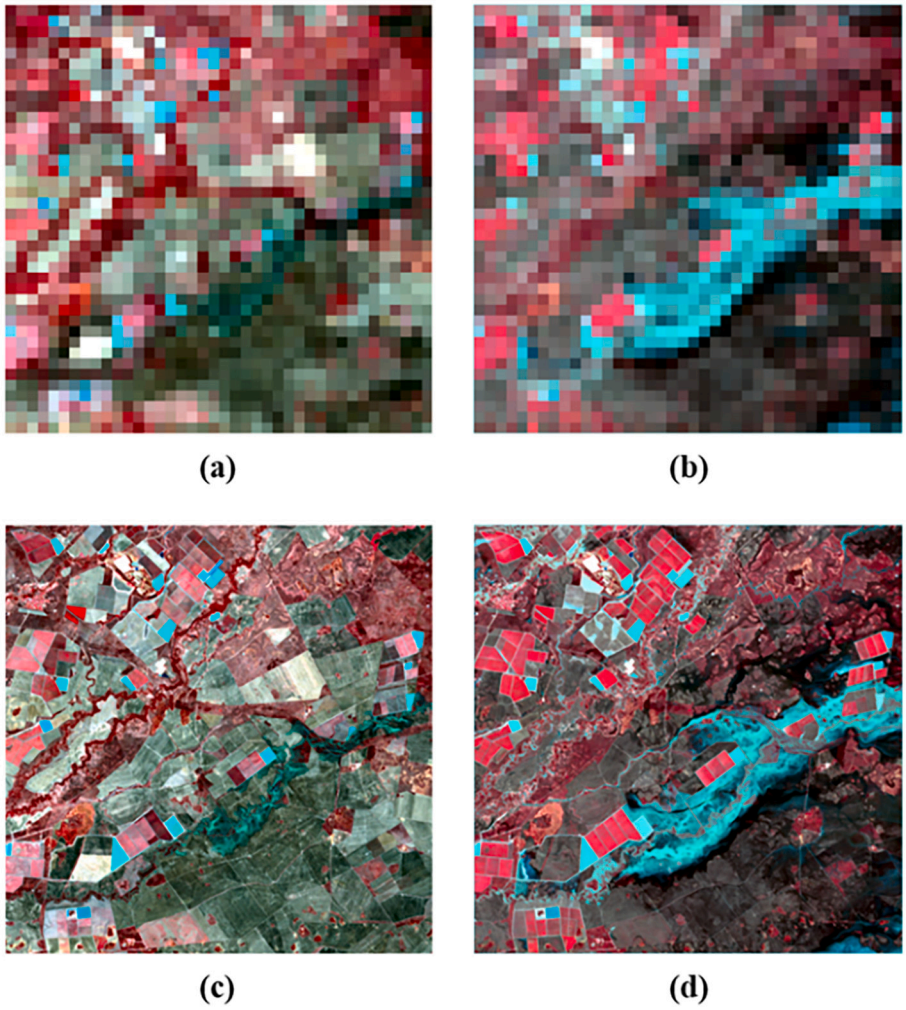

Fig. 4. Test region in GWY with abrupt land cover type changes: (a) simulated MODIS image at T_1; (b) simulated MODIS image at T_2; (c) Landsat image at T_1; (d) Landsat image at T_2. It should be noted that, in this study, images are displayed with the composite false colors of the NIR, red, and green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

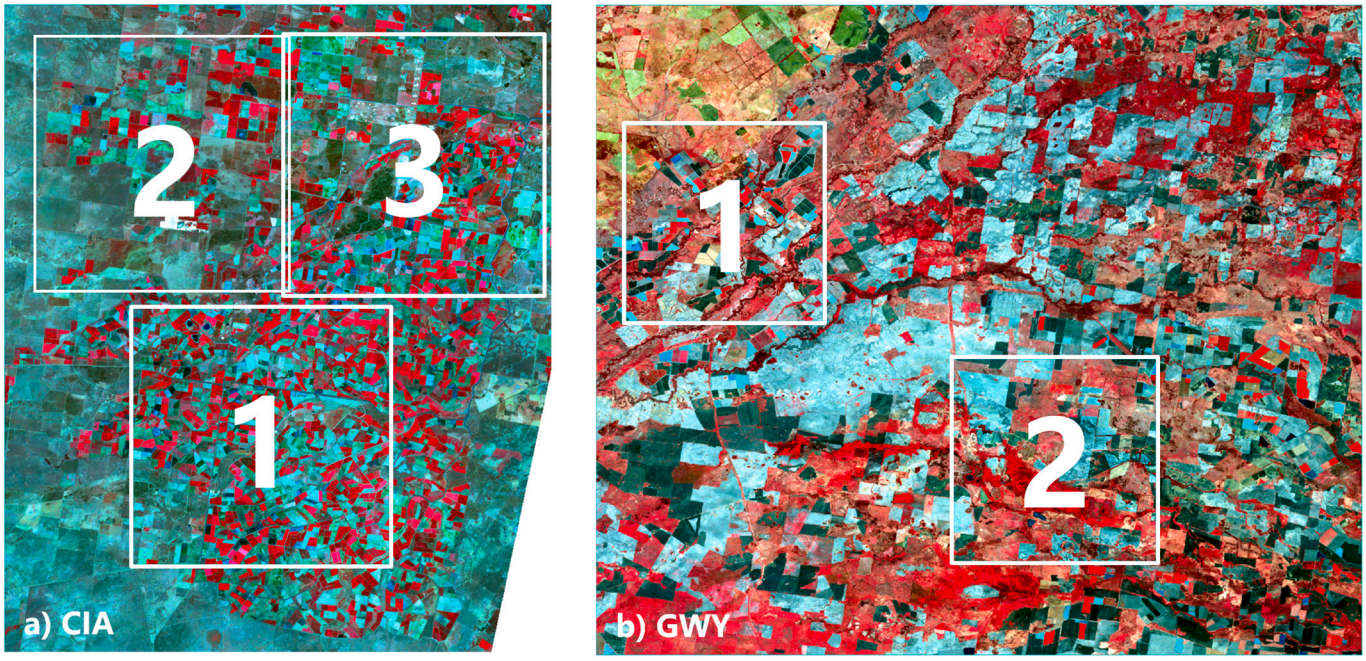

Fig. 5. The input test data of the CIA and GWY. Three areas were selected from the CIA, and two areas were selected from the GWY.

Table 2
Key parameters for the spatiotemporal data fusion methods where n_c is the number of clusters for classification, w is the moving window size, n_s is the number of similar neighboring pixels and w_s is the window size for searching similar neighboring pixels.

	n_c		w	n_s	w_s
	CIA	GWY			
Fit-FC	N/A	N/A	3*3	30	31
FSDAF	6	5	N/A	30	31
STARFM	6	5	N/A	N/A	31
FSDAF2.0	6	5	N/A	N/A	N/A
RASDF	4–7	4–7	N/A	N/A	N/A
VSDF	N/A	N/A	N/A	30	31

Table 3
The composites and meaning of the all-round performance assessment (APA) metrics.

Metric	Meaning	Range	Optimal value	Positive value	Negative value
Root-mean-square error (RMSE)	pixel spectral accuracy	[0,1]	0	spectral errors	/
Average difference (AD)	global spectral accuracy	[−1,1]	0	over estimated	under estimated
Robert's edge (EDGE)	edge	[−1,1]	0	over sharpened	over smoothed
Local binary patterns (LBP)	texture	[−1,1]	0	over sharpened	over smoothed

In comparison, VSDF estimated successfully the spectral changes and recovered the edges. However, there was extra texture detail caused by the errors of AVC (marked in red in Fig. 11(c)). Table 5 lists the APA performance metrics for STARFM, FSDAF, FSDAF 2.0, Fit-FC, RASDF, and VSDF, and it may be seen that the accuracy of the proposed VSDF was superior to the other methods, especially with respect to the spectral accuracy (RMSE) and the texture (LBP).

4.2. Test with heterogeneous landscape (CIA)

The predictions with the simulated-MODIS images for three heterogeneous CIA landscape regions are presented in Fig. 12. The heterogeneous CIA landscape is characterized by the presence of small patches of farmland, the size of which are comparable to a coarse pixel of MODIS. Based on a visual comparison, all predictions fused successfully the heterogeneous landscape and the patches. Parts of the edges and the texture were over-smoothed in the prediction results for FSDAF, Fit-FC, and STARFM, especially when the time interval was large and the land cover changes were intensive. In contrast, VSDF and RASDF captured the edges of the patches more clearly. VSDF adopted the AVC to distinguish the types of land cover changes, which means VSDF could estimate effectively the changes of the small patches in the heterogeneous landscape. The zoomed area marked in yellow in Fig. 12 (b) is represented by Fig. 13. Apart from the red area marked in Fig. 13(b), which was not estimated clearly by all methods due to the loss of fine-resolution information, the proposed VSDF recovered the approximate estimation for most small patches. In comparison, all other methods failed to estimate the specific patches marked in red in Fig. 13.

The predictions based on feeding real MODIS images are presented in Fig. 14. Real MODIS images contain unexpected uncertainties (e.g., geometric and spectral errors). Such uncertainties contradict the basic assumption of spectral unmixing, which induce errors in the prediction. It was found that all predictions were more ambiguous than the predictions based on feeding simulated MODIS images, which is evident at the edges of the small patches. When the size of the registration errors in the MODIS images was comparable with the scale of the small patches, the spectral variation overflowed into neighboring patches and induced blurry edges and spectral errors in the patches. As shown in Fig. 14, VSDF decreased such side effects by controlling the consideration of the MODIS information via the RRI. When the errors of the MODIS images were significant, VSDF reduced the clustering number for AVC and lowered the considerations of the residuals.

The all-band average abs-APA performance metrics for VSDF, RASDF, FSDFA 2.0, FSDAF, Fit-FC, and STARFM are presented in Fig. 15, where the performance feeding real MODIS data is in blue and that feeding simulated MODIS data is in orange. For the convenience of visual comparison, the abs-APA performance metrics, and the absolute

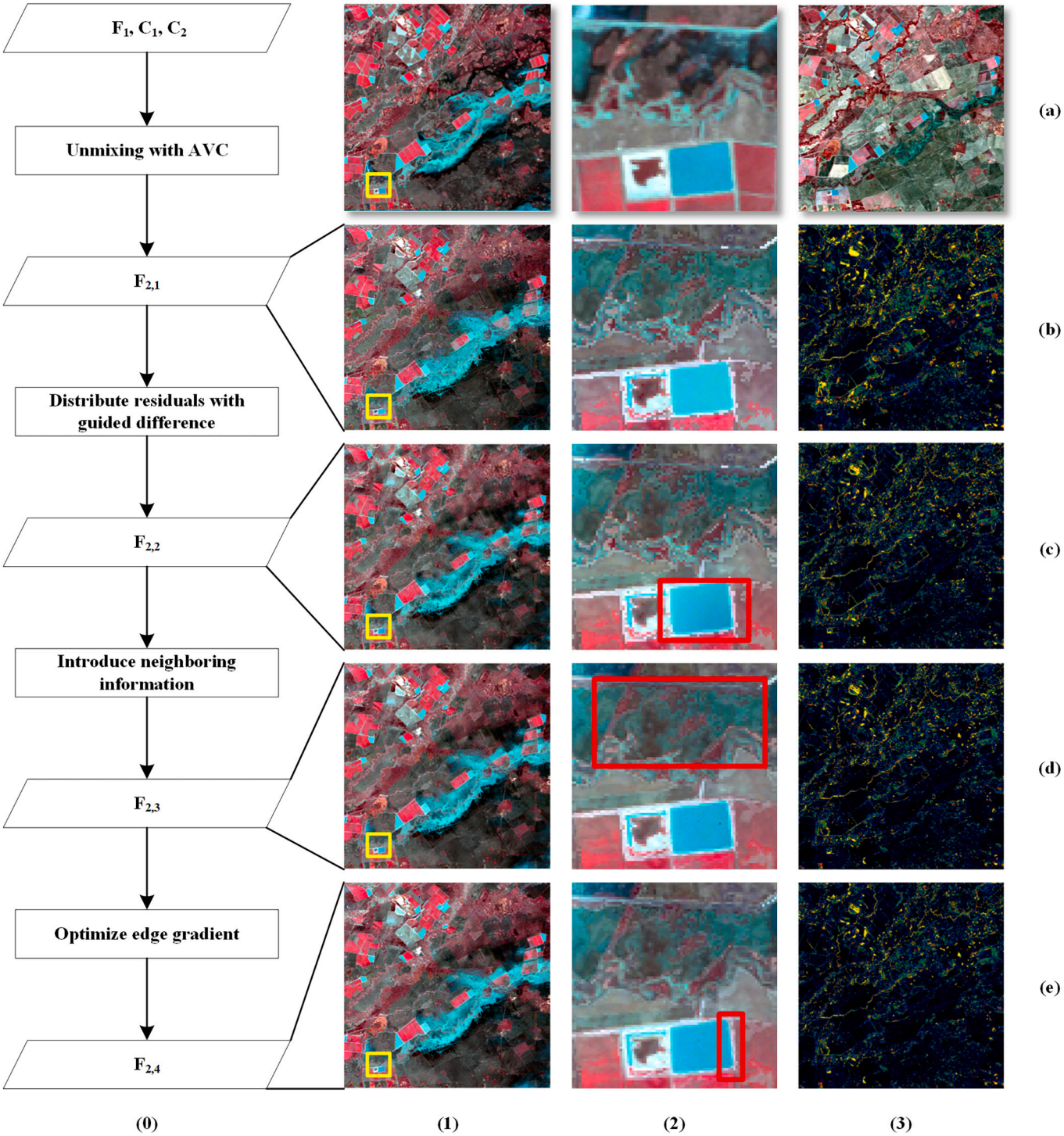

Fig. 6. The visual process of VSDF. Row: (a) ground truth; (b) $F_{2,\ 1}$; (c) $F_{2,\ 2}$; (d) $F_{2,\ 3}$; (e) $F_{2,\ 4}$. Column: (0) simplified process of VSDF represented by four intermediate predictions; (1) false colour composites of predictions; (2) zoomed predictions for the sub-area marked in yellow; (3) the differences between the prediction and the ground truth. To be specific, image (3)(a) is the ground truth of F_1.

value of the origin APA performance metrics are presented, and where a lower bar height means better performance. All the spatiotemporal data fusion methods produced nearly unbiased predictions with acceptable AD values (abs-AD <0.003). Statistically, VSDF performed better than the five other spatiotemporal data fusion methods for the heterogeneous CIA landscape. VSDF obtained the lowest RMSE, abs-EDGE, and abs-LBP in most of the CIA regions with both real and simulated MODIS. It can be seen that Fit-FC realized comparable RMSE to VSDF in several cases. However, Fit-FC performed badly in recovering the edges and texture, as reflected in the high abs-EDGE and abs-LBP values. Such results for Fit-FC were consistent with those in independent studies (Liu et al., 2019a; Zhou et al., 2021). In addition, it can be seen that VSDF estimated successfully the prediction with low AD when the land cover changes were intensive (longer time intervals).

Table 4
All-round performance assessment (APA) metrics (AD, RMSE, EDGE, and LBP) for $F_{2,1}$, $F_{2,2}$, $F_{2,3}$, and $F_{2,4}$ of VSDF. Only the mean value of all bands is shown. Values <0.001 were replaced by 0.

	$F_{2,1}$	$F_{2,2}$	$F_{2,3}$	$F_{2,4}$
AD	0	0	0	0
RMSE	0.034647	0.03109	0.028907	0.027959
EDGE	−0.37574	−0.36889	−0.41887	−0.39213
LBP	0.012963	0.037231	0.009977	0.024289

4.3. Test with homogenous landscape (GWY)

The GWY is homogenous and is characterized by large parcels and complex texture. Hence, the main difficulties are in recovering the texture and estimating the land cover change. The predictions for the GWY regions feeding simulated-MODIS data are presented in Fig. 16. From the visual comparison, it may be seen that all spatiotemporal data fusion methods generally recovered the edges and spectral changes except Fit-FC and STARFM. For the predictions of close intervals, visual

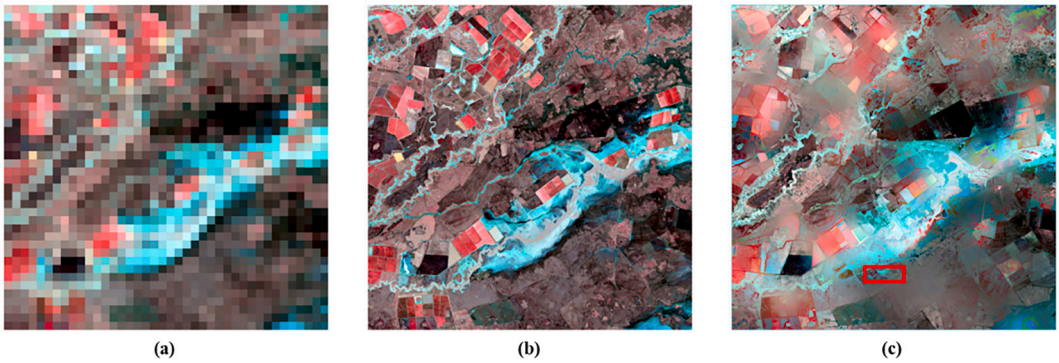

Fig. 7. The temporal change: (a) difference of C_1 and C_2; (b) difference of F_1 and F_2; (c) guided change image (ΔF_{guided} given by Eq. (8)).

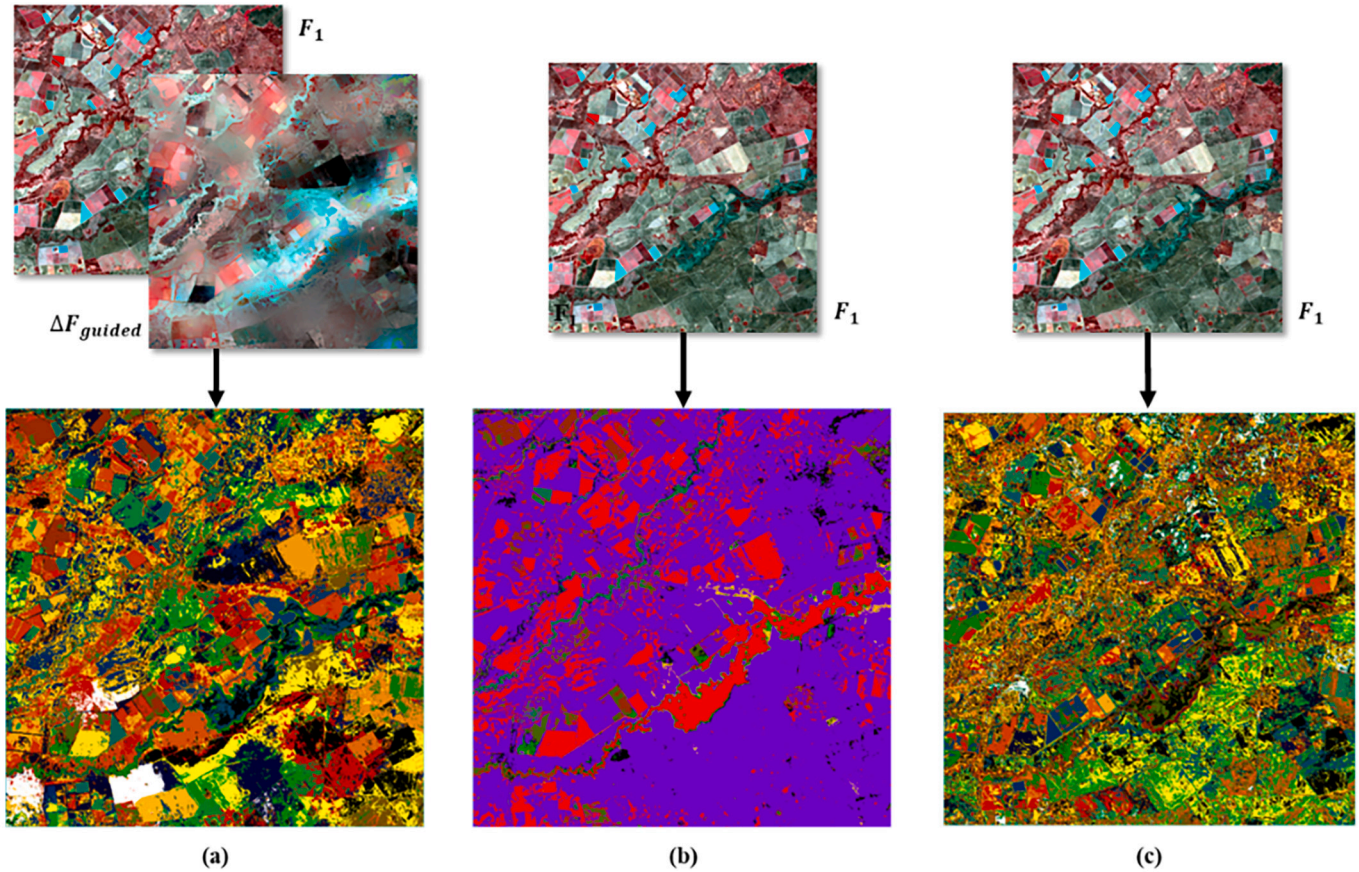

Fig. 8. Classification: (a) 7 K-Means classification of AVC with 45 clusters; (b) the K-Means classification of F_1 with 5 clusters; (c) the K-Means classification of F_1 with 45 clusters.

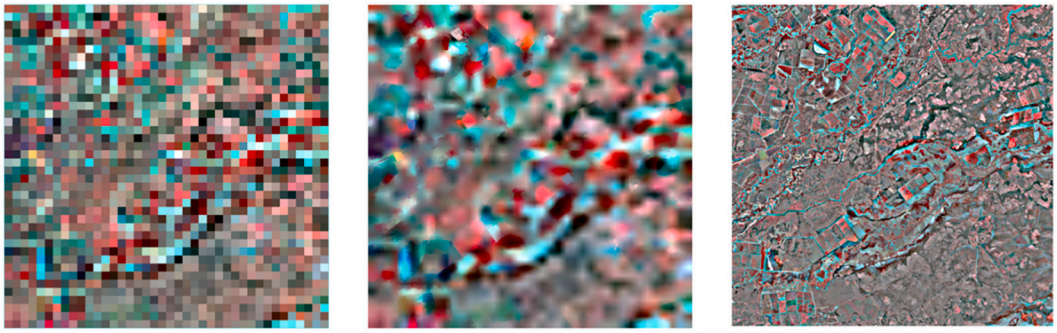

Fig. 9. Residuals of $F_{2,\ 1}$: (a) residuals of $F_{2,\ 1}$ (difference between $F^{C}_{2,\ 1}$ and C_2 at coarse-resolution); (b) filtered residuals of $F_{2,\ 1}$ under the guidance of F_1; (c) ground truth of the residuals (difference between $F_{2,\ 1}$ and F_2 at fine-resolution).

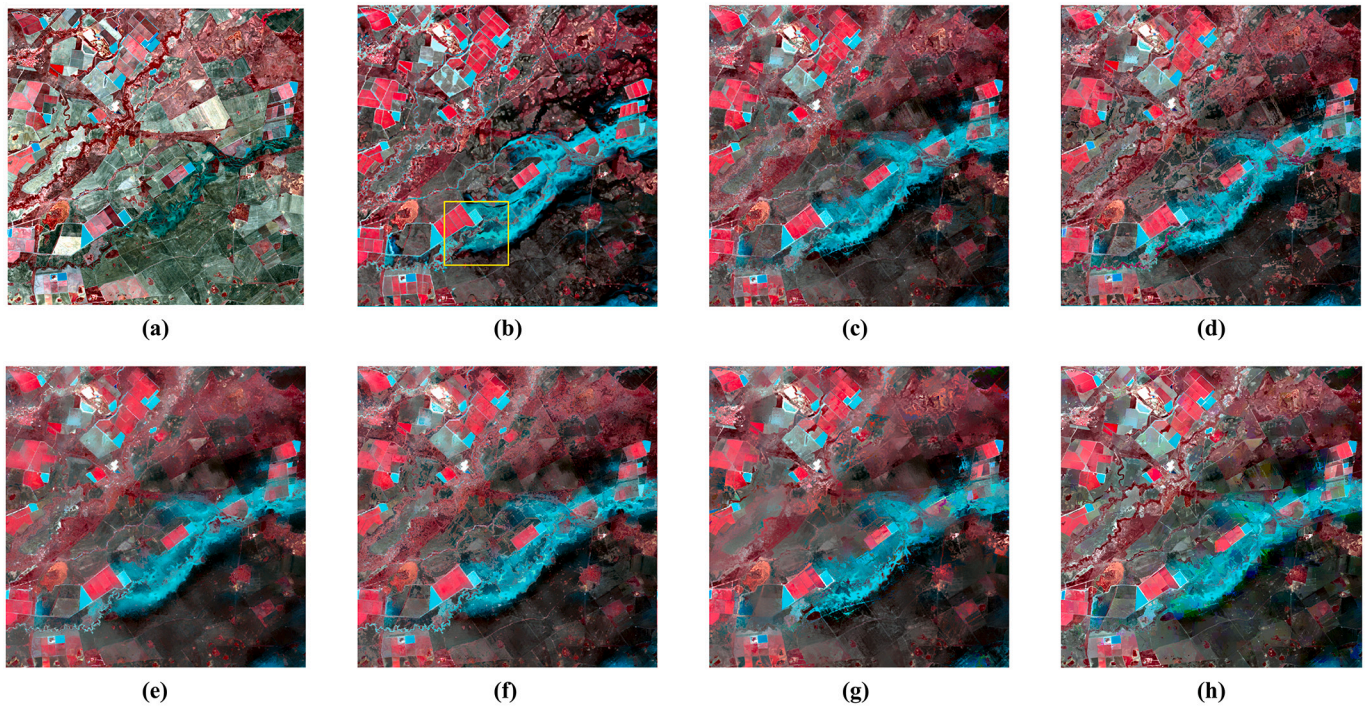

Fig. 10. Predictions of the region with land cover change: (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), and STARFM (h).

differences between the predicted images and the reference images were rare. The zoomed area marked in yellow in Fig. 16(b) to highlight the details of the predictions are represented by Fig. 17. It can be seen that the texture for the GWY sites was more complicated and irregular than those for the CIA sites, and there were apparent spectral errors in most predicted images. Part of the texture in the predictions, especially the results for Fit-FC, were over-smoothed because there was, in theory, no information for estimating the texture at fine resolution. However, the predictions of VSDF captured accurately the irregular spectral changes for small irregular land cover changes, which benefitted from the AVC. VSDF retrieved better texture by transferring explicitly the texture and edges from the available fine image at T_1.

Comparing the predictions feeding real-MODIS data, as shown in Fig. 18, it can be discerned that registration errors caused the blurred edges, as evidenced in the fusion predictions by FSDAF, Fit-FC, and STARFM. With reference to the zoomed area in Fig. 19, there were apparent errors (marked in white) in all the predictions which were fed the misregistered real MODIS images. However, VSDF preserved the edges of the parcels more clearly compared to the other methods.

The statistical evaluation data for VSDF, RASDF, FSDAF 2.0, FSDAF, Fit-FC, and STARFM are presented in Fig. 20. All the methods generated acceptable unbiased predictions based on the feeding of the real MODIS data (abs-AD <0.005) and where ignorable unbiased predictions were obtained feeding the simulated MODIS data (abs-AD <0.0001). Clearly, the RMSE of the predictions for fusion by VSDF was lower than those for most of the other methods, which means that the spectral accuracy of VSDF was superior and confirms the considerable potential for VSDF to retrieve the spectral changes. In addition, the fused images of VSDF had smaller abs-EDGE and abs-LBP values in most cases, suggesting that VSDF preserved the edges and textural information in the GWY landscape.

5. Discussion

5.1. All-round performance of VSDF

The APA performance metrics for each intermediate prediction of VSDF are presented in Table 6. After distributing the guided residuals

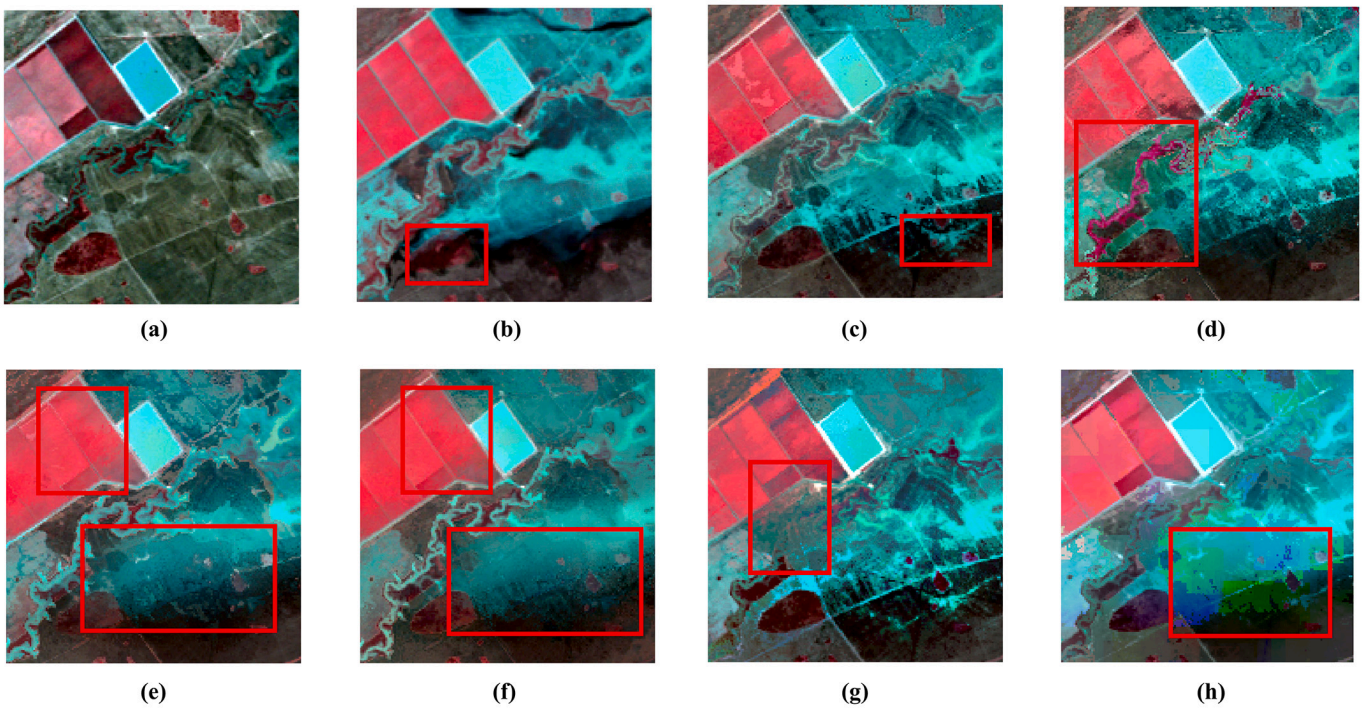

Fig. 11. Zoomed area marked in yellow in Fig. 10(b): (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), and STARFM (h). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
All-round performance assessment (APA) metrics (AD, RMSE, EDGE, and LBP) of VSDF, RASDF, FSDAF 2.0, FSDAF, Fit-FC, RASDF, and STARFM. Only the mean value of all bands is shown. The best results are marked in bold and any value <0.0001 is marked as 0.

	VSDF	RASDF	FSDAF2.0	FSDAF	Fit-FC	STARFM
AD	**0**	**0**	**0**	**0**	**0**	0.00126
RMSE	**0.02796**	0.03089	0.02816	0.03010	0.03114	0.02979
EDGE	−0.39213	**−0.36125**	−0.48706	−0.46141	−0.53324	0.45882
LBP	**0.02429**	0.05327	−0.09644	−0.04955	−0.15690	−0.23585

from $F_{2,\ 1}$ to $F_{2,\ 2}$, it can be seen that the RMSE decreased, but the spatial accuracy was not affected, which means that VSDF successfully prevented the induction of a blurry prediction during the distribution of the residuals. Introducing neighboring information from $F_{2,\ 2}$ to $F_{2,\ 3}$ affected the spatial accuracy as the abs-EDGE increased a lot. This was because introducing neighboring information averaged the difference of similar pixels, which caused the blurry edges. However, the introduction of neighboring information repaired the minor artifacts caused by the errors in AVC and thus improved the evaluation of texture (abs-LBP) and spectral accuracy (RMSE). In the final step, introducing the edges improved the estimation of the edges, and the abs-EDGE decreased while other performance metrics remained nearly unchanged. Overall, it is clear that in each step of VSDF, the RMSE decreased steadily.

In Table 7, it can be seen that apart from Fit-FC, the errors in the global spectrum may be ignored (abs-AD <0.002), and FSDAF delivered the best performance of global spectral accuracy for spectral accuracy among all the fusion methods, followed by VSDF. It can be seen that VSDF had the lowest RMSE compared with the other methods, and the improvement in accuracy was at least 7%. Furthermore, it can be seen that the spectral performances of RASDF and FSDAF 2.0 were close to each other given that they both had similar abs-AD and abs-EDGE values. In terms of spatial accuracy, VSDF obtained the best abs-EDGE and abs-LBP values, indicating that VSDF captured the edges and texture more accurately than the other methods. Overall, the performance of VSDF was superior to the benchmark methods in terms of the experimental results.

5.2. Retrieving abrupt land cover changes with AVC

As mentioned previously, it is difficult to predict abrupt changes. Fit-FC captures the temporal changes by establishing the regression model between the fine image at T_1 and the fine image to be fused. The regression model can capture effectively the prominent changes and changes that are evident, but ambiguous edges will be introduced, which is evident as shown in Fig. 13(g) and Fig. 17(g). FSDAF and FSDAF 2.0 estimate the abrupt temporal changes by interpolating the coarse-resolution residuals, and introduce the abrupt temporal changes by distributing the interpolated residuals at fine resolution. Interpolation of the residuals will also cause ambiguous results, which is evident around the abrupt changes, as demonstrated in areas marked in red in Fig. 11(e, f). RASDF improves the predictions by implementing additional local unmixing. RASDF can thus preserve the structure and details where there are no abrupt changes. Nevertheless, RASDF does not consider systematically the land cover changes and regards such changes as exceptions that need to be repaired. Hence, RASDF failed to recognize definitively some small parcels in the experiments, e.g., areas marked in Fig. 11(d), Fig. 13(d), and Fig. 19(d). Apart from the compared methods, an enhanced FSDAF that incorporates sub-pixel class fraction change information (SFSDAF) considers the temporal changes by introducing land cover class fraction changes. However, SFSDAF downscales the sub-pixel land cover class fraction change from coarse resolution to fine resolution by interpolation, and this will induce blurring effects in the abrupt change area (Li et al., 2020).

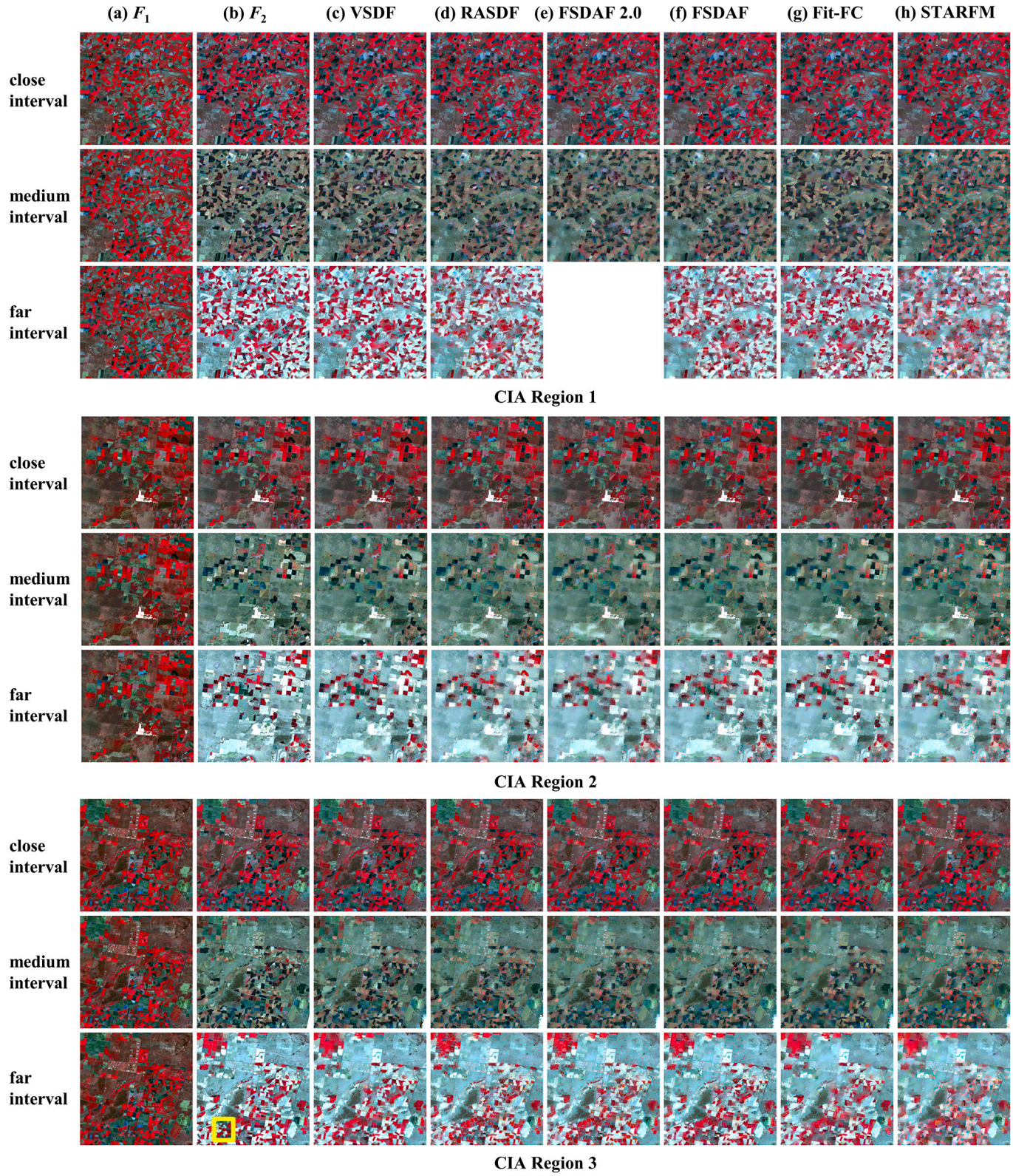

Fig. 12. Predictions feeding Landsat/simulated-MODIS image pairs for three CIA regions: (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), and STARFM (h). For each region, three tests with different time intervals were carried out. FSDAF 2.0 failed to implement the prediction for the far interval image pairs in region 1.

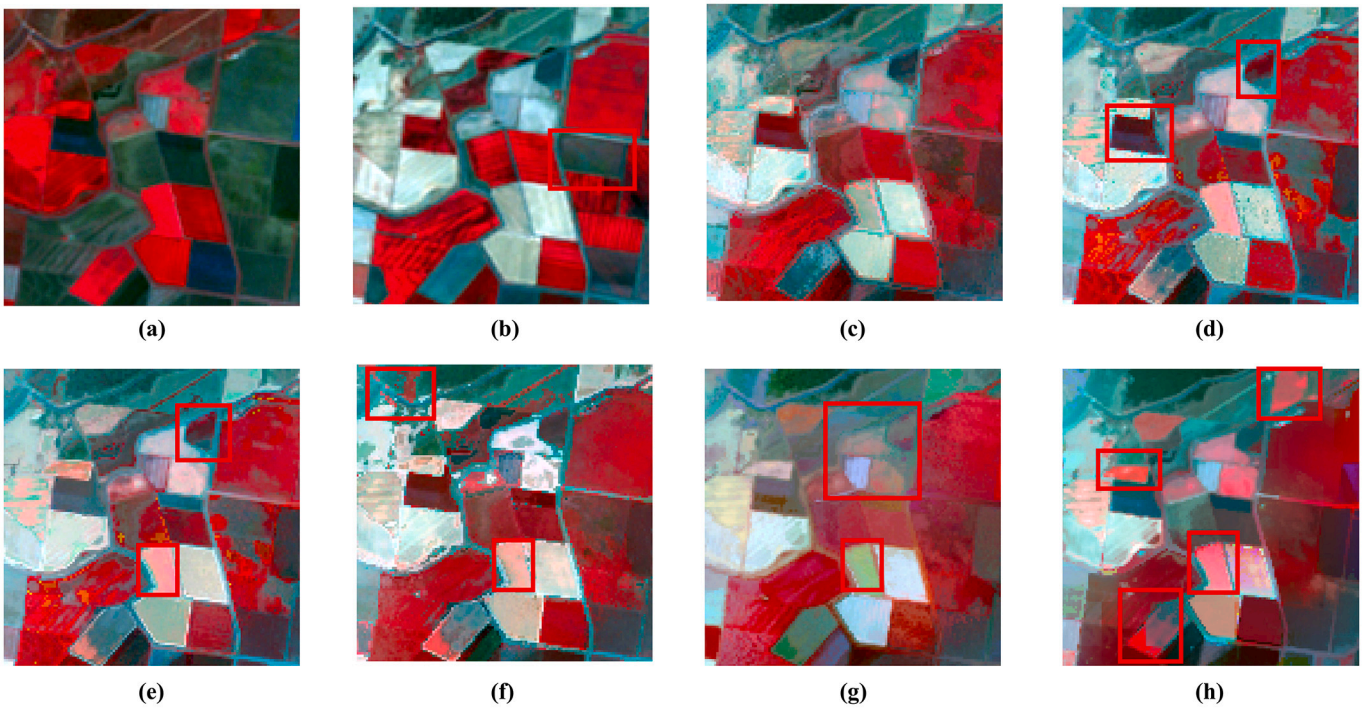

Fig. 13. The zoomed area marked in yellow in Fig. 12(b): (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), and STARFM (h). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

VSDF adopts a novel method to retrieve systematically abrupt land cover changes, and this distinguishes it from any other non-learning-based spatiotemporal data fusion method. VSDF recognizes qualitatively the types of change of each fine pixel with AVC and estimates quantitatively the abrupt land cover changes by unmixing. To evaluate the fine-resolution temporal changes, VSDF implements the GF to downscale the temporal changes from coarse resolution to fine resolution under the guidance of F_1. The GF has been used in Fit-FC to construct the regression model, but the filter output is not strictly a local linear transform of the guidance of F_1, which means that the GF cannot estimate quantitatively the gradient around the edges and the texture. As a result, there might be unreal predictions in the results of Fit-FC. However, the proposed AVC is not sensitive to accuracy at a quantitative level, thus the AVC can be used to identify qualitatively the temporal changes with detailed spatial structures for classification purposes. Then, VSDF obtains the quantitative results through global unmixing. VSDF captured accurately the temporal changes of small parcels (e.g., the zoomed areas in Fig. 13, Fig. 17, and Fig. 19). With reference to the evaluation of RMSE, as in Fig. 15 and Fig. 20, in most tests, VSDF estimated the spectrum and the temporal change more accurately than RASDF, FSDAF 2.0, FSDAF, Fit-FC, and STARFM.

To better demonstrate the effect of AVC in VSDF and compare its advantages with ordinary land cover classification adopted by other methods, we utilized the classification feeding F_1 instead of AVC in VSDF (Table 8). At the same time, to eliminate the possible influence caused by different numbers of clusters, the cluster numbers were set the same as the AVC. The results showed that the introduction of AVC improved the spectral accuracy of the prediction affording better spatial structures. However, the introduction of AVC also decreased the accuracy of the prediction of the edges, which might be induced by the blur during downscaling. Such a loss of accuracy for the edges is relatively acceptable given that the performance was better than the benchmark methods (Table 7). The results show that VSDF benefits from employing AVC.

5.3. Optimizing the prediction with the relative reliability index

VSDF introduces the reliability of the input datasets to guide the prediction. The uncertainty caused by differences in the characteristics and the performance of different sensors might induce unstable prediction results. Current fusion models have considered such uncertainties by pre-processing, which can lower the side effects during fusion (Gao et al., 2006; Zhu et al., 2010). Additionally, such uncertainties have been used to guide the usage of input information according to its reliability (Shi et al., 2022). Analogously, the use of the RRI is proposed in order to present the ratio of the intensity of the temporal changes and the reliability of the input datasets, which is adopted to guide the prediction process of VSDF. The RRI is a "relative" index that indicates the relative possibility of retrieving the temporal changes. On the one hand, the RRI will increase with increase of information that can be obtained from the coarse image pair to evaluate the temporal changes. On the other hand, the RRI will decrease with increase of the errors in the input datasets because errors increase the uncertainty of the input information in estimating the temporal changes. In present research, the RRI determines the number of clusters of AVC and the number of loops needed to distribute the residuals (Table 9). For instance, the RRI for three tests of CIA regions with close intervals was lower than 1, which means the errors in the MODIS images were more significant than the differences in the temporal changes. Under such circumstances, it is supposed to be impossible to estimate accurately the temporal changes, or the temporal changes are not obvious. As a result, fewer cluster numbers of AVC were set (19, 15, and 18) and there was no need to introduce the residuals from the coarse images (loop number = 0).

As expressed in Eq. (9), the cluster number is determined by the RRI, but the n_F is set empirically. To discuss the effect of the cluster number and the RRI, we further explored the sensitivity of VSDF with different n_F, which is shown in Fig. 21. It can be found that the spectral accuracy

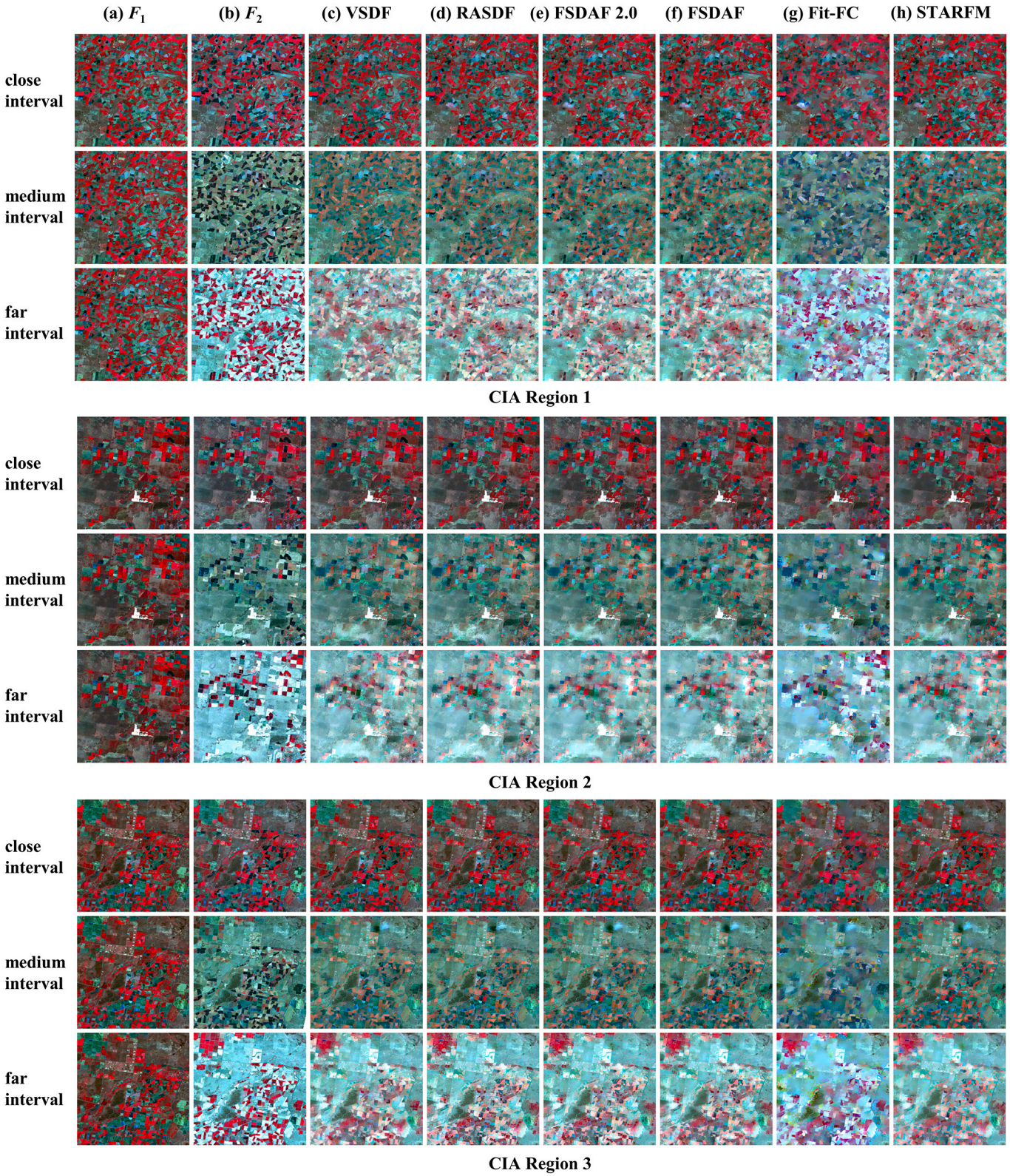

Fig. 14. Predictions feeding Landsat/real-MODIS image pairs for three CIA regions shown in Fig. 5: (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), and STARFM (h). For each region, three tests with different time intervals were carried out.

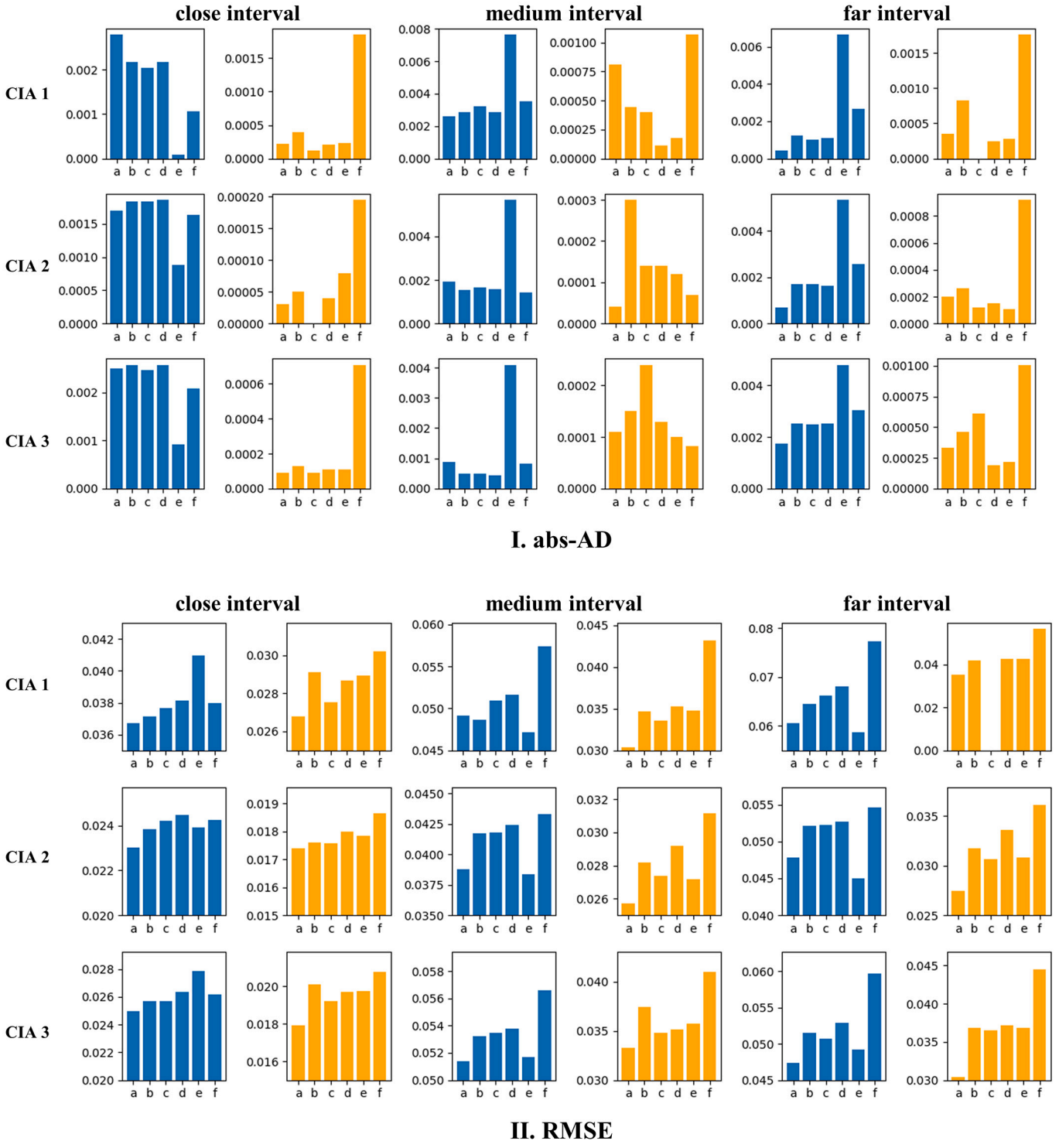

Fig. 15. Absolute all-round performance assessment (APA) metrics for CIA predictions: I: abs-AD; II: RMSE; III: abs-EDGE; IV: abs-LBP. The a, b, c, e, and f refer to the performance for VSDF, RASDF, FSDFA 2.0, FSDAF, Fit-FC, and STARFM, respectively. A lower bar height represents better performance. The full metrics are given in the Appendix.

(EDGE and abs-LBP) of the final predictions ($F_{2,\ 4}$) changed little with the variance of n_F. In contrast, the spectral accuracy (abs-AD and RMSE) was affected by n_F. On the one hand, more cluster numbers (larger n_F) yielded a better AD accuracy. On the other hand, the average RMSE decreased with the increase of n_F at first, and the best performance was reached when $n_F = 5$. It should be noted that the average abs-AD value was <0.002, which is less important than the RMSE. Hence, we recommend the use of $n_F = 5$ according to the RMSE. Overall, although the cluster number affected the performance of VSDF, it is evident that the performance of VSDF ($F_{2,\ 4}$) was always better than the other methods except for AD, which indicates that the optimal performance of VSDF compared to the benchmark methods was not affected by the impact of different cluster numbers.

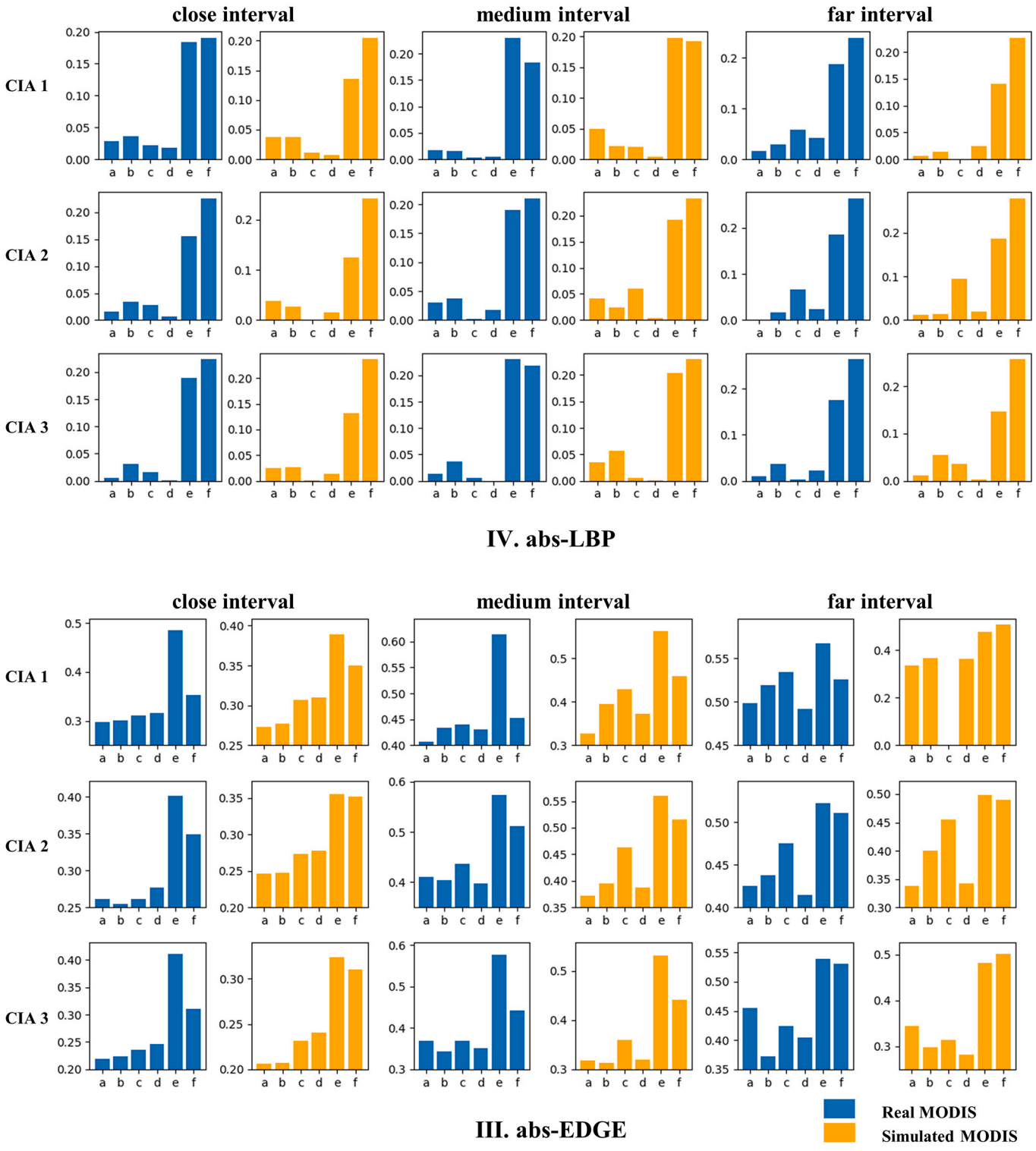

Fig. 15. (*continued*).

5.4. Limitations and further improvements

The results of the experiments demonstrate that the VSDF can predict abrupt land cover changes with satisfactory accuracy. Nevertheless, there are some known limitations in the method which need to be addressed. VSDF introduces the GF to transform the detailed spatial structure from F_1 to the predicted fine image, which is used to evaluate AVC and distribute the residuals. Although it can be very effective for part of the abrupt land cover changes, the GF cannot capture the abrupt change if the texture has changed. This is because the guided temporal image cannot estimate the texture at T_2 under the guidance of the fine image at T_1. The approach might introduce false spatial details if the spatial structure changes from T_1 to T_2, as in the patch marked in red in Fig. 11(c). Unfortunately, it is presumed this is inevitable with the given input datasets because there is no available fine-resolution information to indicate such detailed abrupt changes at T_2. In subsequent research, the real spatial structure could be extracted from supplementary fine images at T_2 (e.g., multisource images, panchromatic images) to avoid

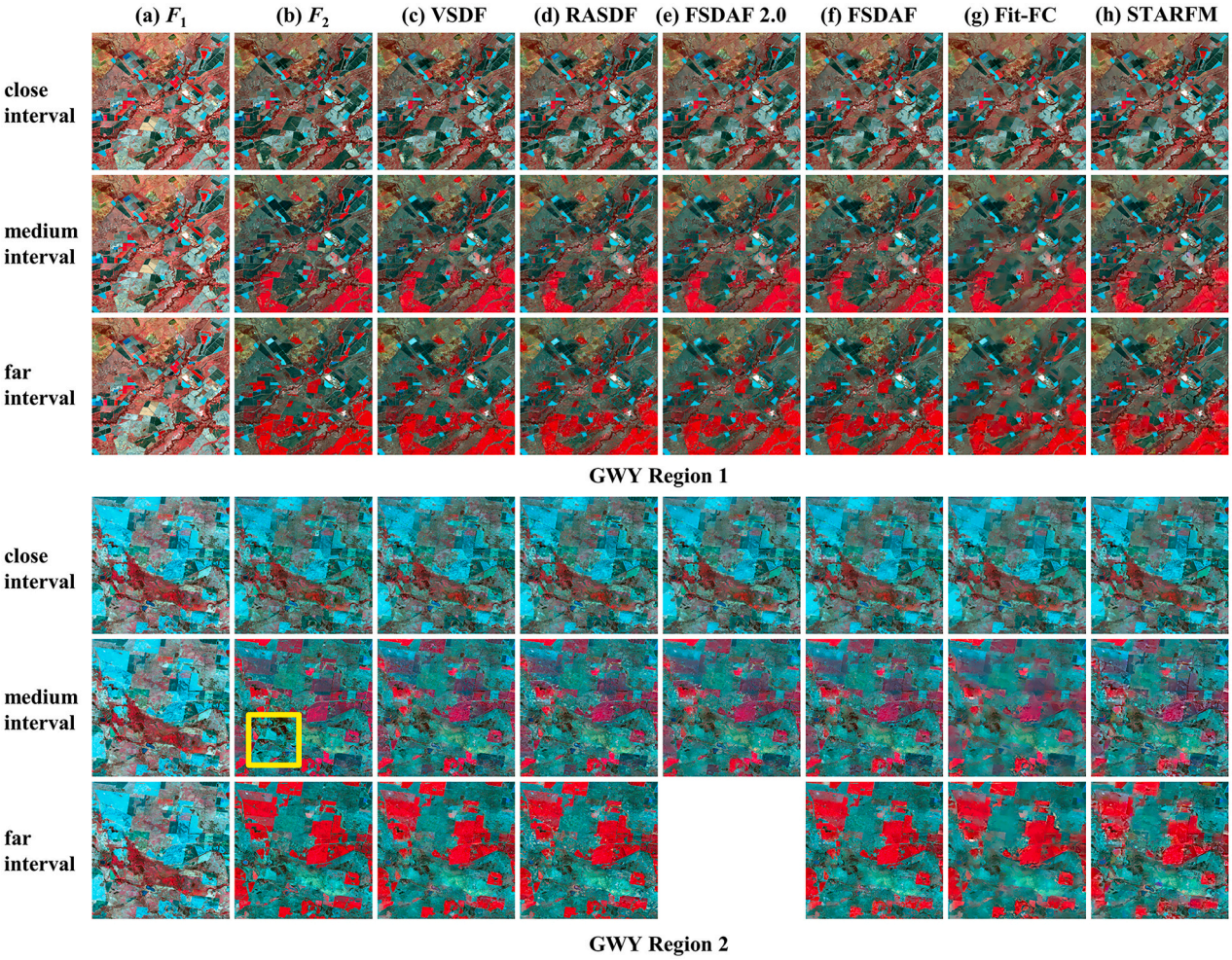

Fig. 16. Predictions feeding Landsat/simulated-MODIS image pairs for two GWY regions: (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), and STARFM (h). For each region, three tests with different time intervals were carried out. FSDAF 2.0 failed to implement the prediction for the far interval image pairs in region 2.

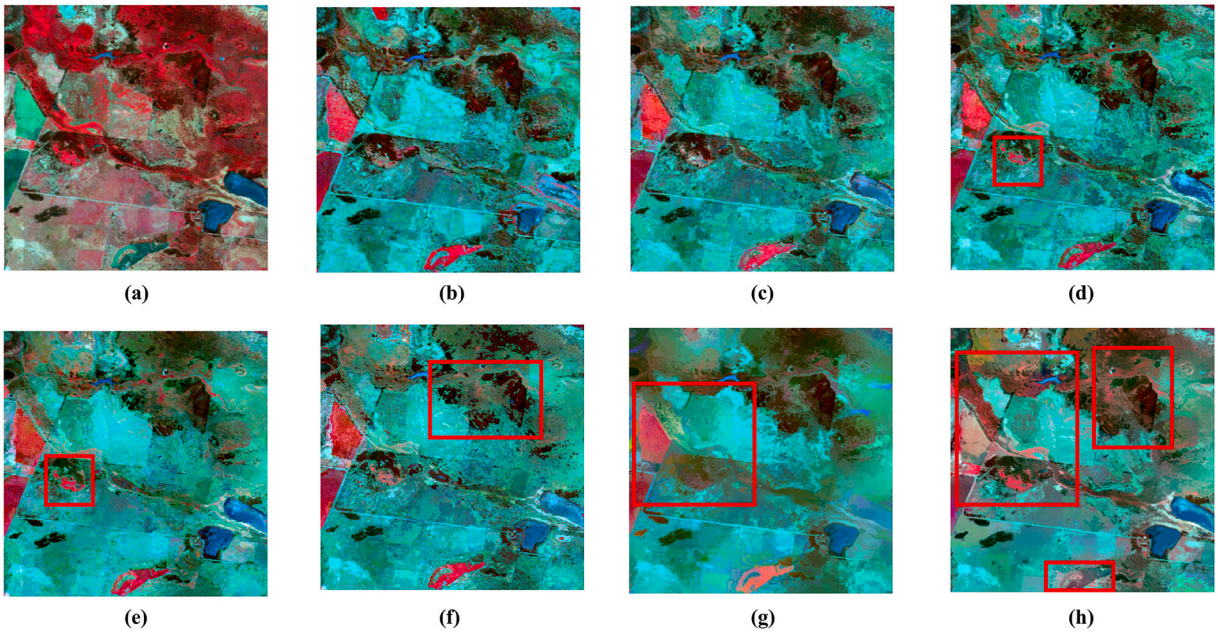

Fig. 17. The zoomed area marked in yellow in Fig. 16(b): (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), STARFM (h). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

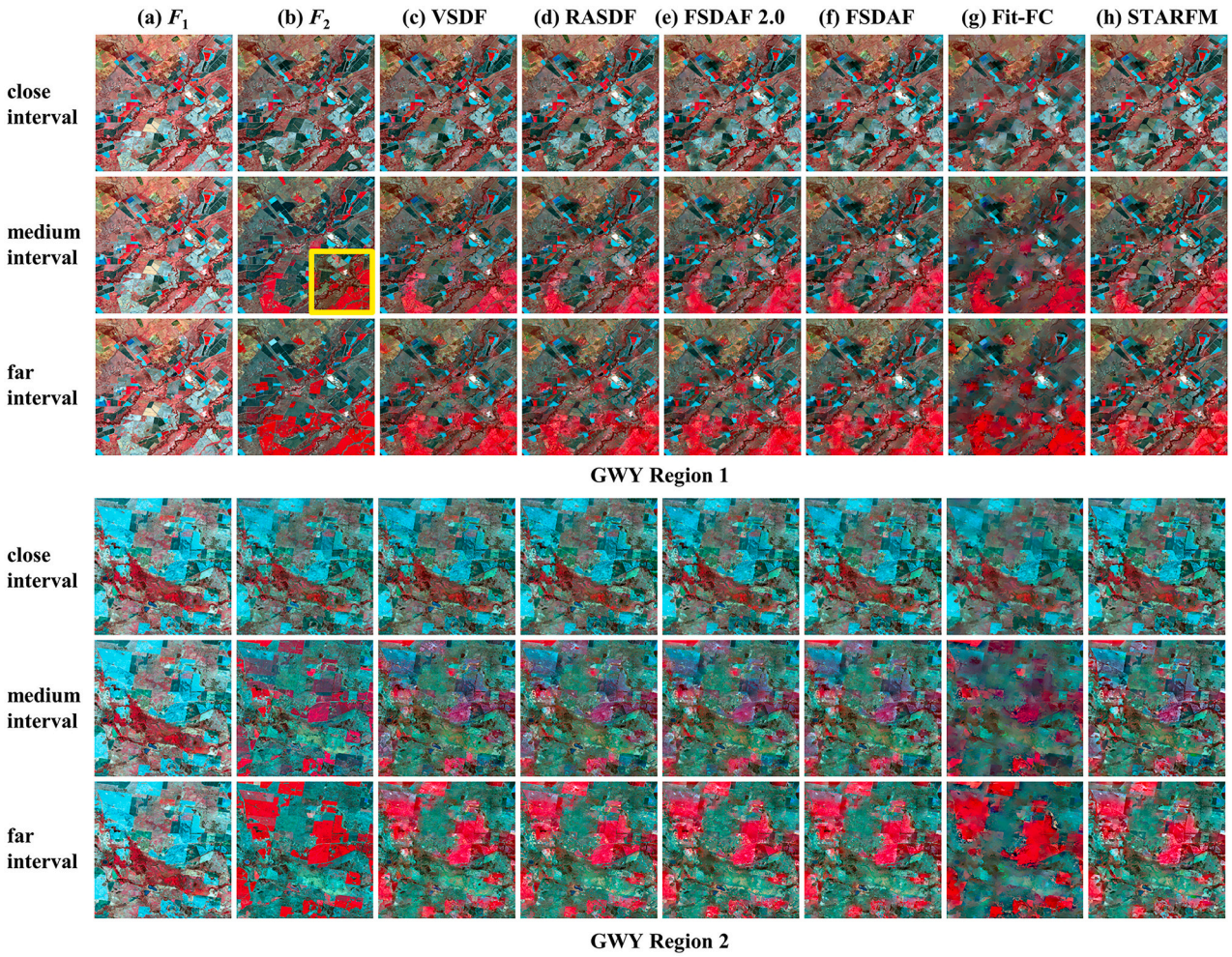

Fig. 18. Predictions feeding the Landsat/real-MODIS image pairs for two GWY regions: (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), and STARFM (h). For each region, three tests with different time intervals were carried out.

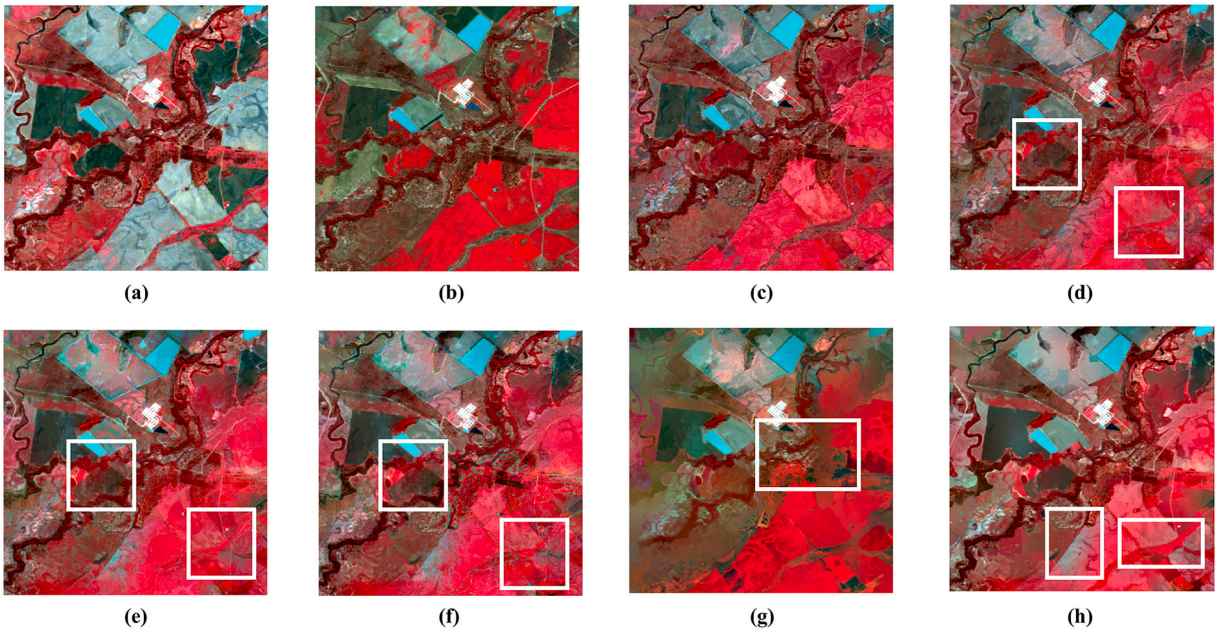

Fig. 19. The zoomed area marked in yellow in Fig. 18(b): (a) F_1; (b) F_2 (ground truth of the prediction); (c)-(h) predictions of VSDF (c), RASDF (d), FSDAF 2.0 (e), FSDAF (f), Fit-FC (g), STARFM (h). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

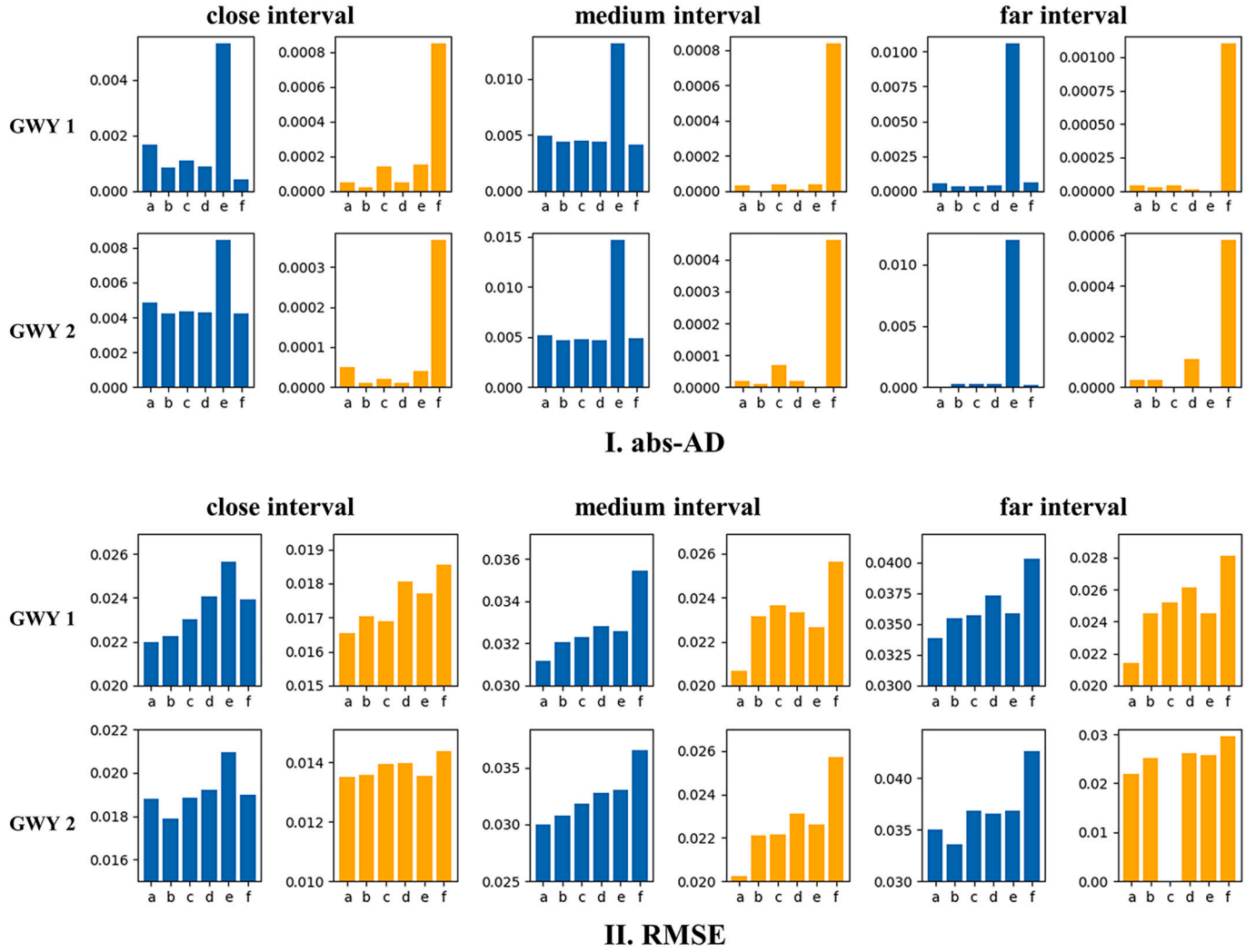

Fig. 20. All-round performance assessment (APA) metrics for GWY predictions: I: abs-AD; II: RMSE; III: abs-EDGE; IV: abs-LBP. The a, b, c, e, and f refer to the performance metrics for VSDF, RASDF, FSDFA 2.0, FSDAF, Fit-FC, and STARFM, respectively. The prediction for GWY region 2 in the far time interval is unavailable. The full metrics are given in the Appendix.

false spatial structures. Additionally, VSDF uses the minimum input (two coarse images and one fine image) for the prediction, which means that the time-series information is not used for prediction. As a result, VSDF can be robust for the prediction with a near time interval, but the performance might decrease when predicting a long time interval, which restricts the potential of VSDF in fine resolution image predictions in a time series. Therefore, further research efforts are needed to introduce multiple pairs of coarse/fine images for a stable prediction for long time intervals and time-series applications, just like the enhanced spatial and temporal adaptive reflectance fusion model (ESTARFM) (Zhu et al., 2010). In particular, the estimation of AVC and land cover changes can be optimized by feeding multiple fine images and exploiting the patterns of time-series information. Moreover, VSDF was tested over heterogeneous/homogeneous landscapes with MODIS/Landsat image pairs. Though the results show the robust performance of VSDF on both surfaces, more tests are expected to further explore the practicality of VSDF for different landscapes (e.g., complex terrain) and the use of remote sensing data acquired by other satellite sensors (e.g., Sentinel, GaoFen). Finally, it is proposed to use the RRI to indicate the relative possibility of recovering the temporal change and guiding the prediction process, e.g., the cluster number of classification in Eq. (9). However, the use of the RRI in VSDF currently is empirical or semi-empirical. In the future, it is planned to exploit the RRI to guide the prediction process.

6. Conclusions

This study has proposed a novel spatiotemporal data fusion model named VSDF. VSDF only needs one pair of reference coarse/fine images and one coarse image at the prediction time. In VSDF, AVC is used to capture the temporal changes, and this classifies explicitly the fine pixels according to their land cover types and temporal change types. Furthermore, a novel index, the RRI, is proposed to characterize the input datasets and guide the processing of VSDF. Finally, feature-level fusion is introduced to recover the edges of the prediction. Experiments were performed for VSDF with Landsat and MODIS/simulated MODIS images, and the performance was compared with five spatiotemporal data fusion methods, i.e., STARFM, FSDAF, FSDAF 2.0, Fit-FC, and RASDF. The quantitative evaluations showed that VSDF gave better performance than the benchmark methods. Specifically, compared with RASDF, FSDAF 2.0, FSDAF, Fit-FC, and STARFM, the improvement in spectral accuracy (RMSE) reached 7%, 7%, 10%, 7%, and 20%, respectively. In summary, VSDF can effectively capture the temporal land cover changes and recover the detailed structure in heterogeneous and homogeneous landscapes.

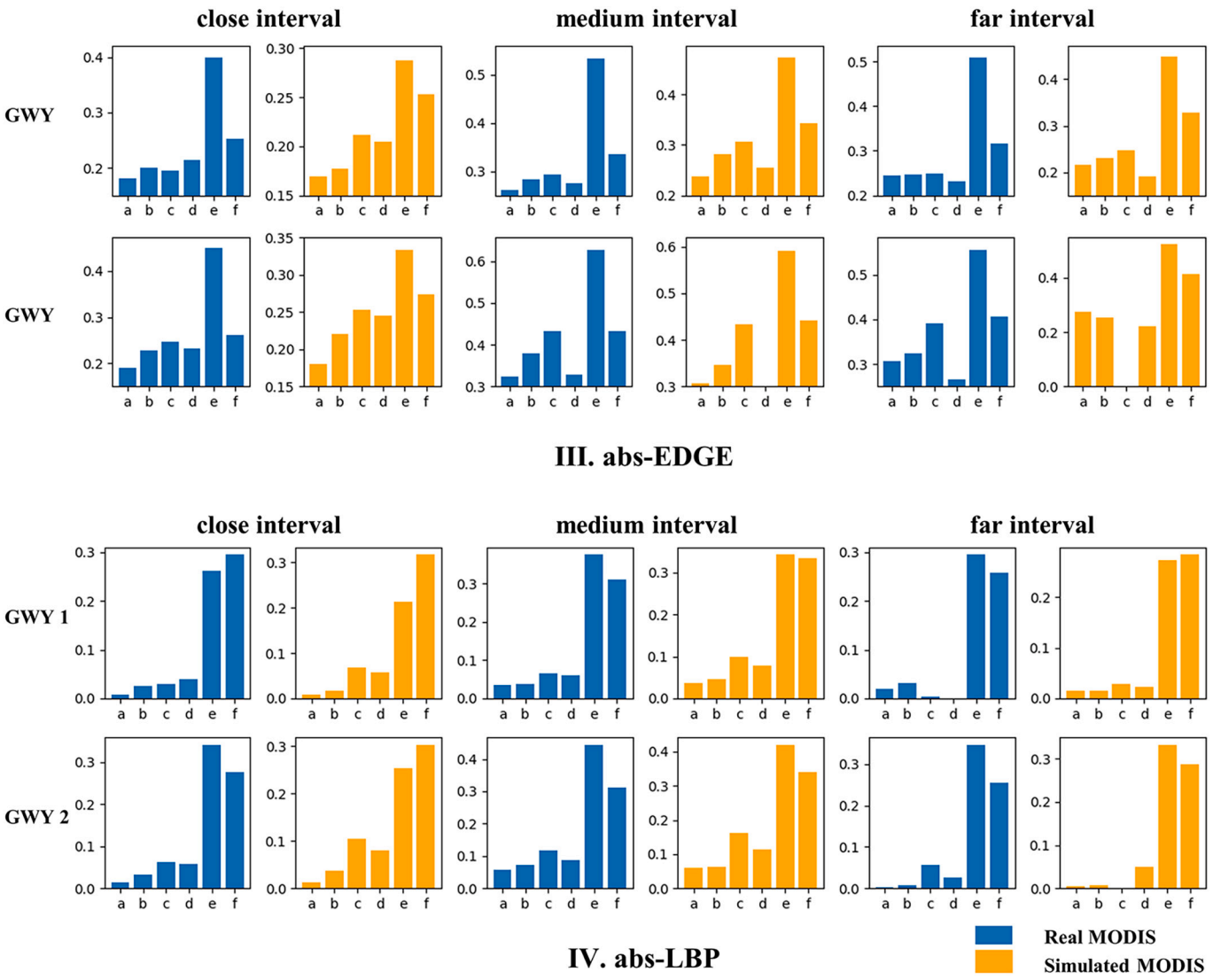

Fig. 20. (*continued*).

Table 6
Average absolute all-round performance assessment (APA) metrics for four intermediate predictions of VSDF (Average abs-AD = average value of absolute AD of 31 tests, Average RMSE = average value of RMSE of 31 tests, Average abs-EDGE = average value of absolute EDGE of 31 tests, Average abs-LBP = average value of absolute LBP of 31 tests). The best results are marked in bold, and the detailed metrics are given in the Appendix.

	$F_{2,\ 1}$	$F_{2,\ 2}$	$F_{2,\ 3}$	$F_{2,\ 4}$
Average abs-AD	**0.001046**	0.001076	0.001110	0.001123
Average RMSE	0.034514	0.03252	0.030278	**0.030236**
Average abs-EDGE	0.274867	**0.271764**	0.318256	0.302958
Average abs-LBP	0.026778	0.026008	**0.022328**	0.022380

Table 7
Average absolute all-round performance assessment (APA) metrics for VSDF, RASDF, FSDAF 2.0, FSDAF, Fit-FC, and STARFM, and VSDF performance improvement of VSDF compared to other algorithms (Average abs-AD = average value of absolute AD of 31 tests, Average RMSE = average value of RMSE of 31 tests, Average abs-EDGE = average value of absolute EDGE of 31 tests, Average abs-LBP = average value of absolute LBP of 31 tests). The best results are marked in bold, and the detailed APA performance metrics are given in the Appendix.

abs-APA performance	VSDF	RASDF	FSDAF 2.0	FSDAF	Fit-FC	STARFM
Average abs-AD	0.001123	0.001131	0.001180	**0.001073**	0.003286	0.001495
Average RMSE	**0.030236**	0.032398	0.032372	0.033341	0.032258	0.036413
Average abs-EDGE	**0.302958**	0.313582	0.347045	0.311166	0.488135	0.400776
Average abs-LBP	**0.022380**	0.03208	0.045828	0.030625	0.233526	0.255779
Improvement of VSDF compared to other algorithms		RASDF	FSDAF 2.0	FSDAF	Fit-FC	STARFM
Average abs-AD		1%	5%	−4%	193%	33%
Average RMSE		7%	7%	10%	7%	20%
Average abs-EDGE		4%	15%	3%	61%	33%
Average abs-LBP		43%	105%	37%	943%	1043%

Table 8
Average absolute all-round performance assessment (APA) metrics of 31 tests with or without AVC (Average abs-AD = average value of absolute AD from 31 tests, Average RMSE = average value of RMSE from 31 tests, Average abs-EDGE = average value of absolute EDGE from 31 tests, Average abs-LBP = average value of absolute LBP from 31 tests). The best results are marked in bold.

	Average abs-AD	Average RMSE	Average abs-EDGE	Average abs-LBP
VSDF with AVC	**0.001123**	**0.030236**	0.302958	**0.022380**
VSDF without AVC	0.001257	0.031850	**0.277731**	0.034960

Table 9
The RRI, cluster number of AVC ($n_F = 5$) and loop number for the tests with MODIS/real Landsat images.

	RRI					Cluster number of AVC					Loop number				
	CIA			GWY		CIA			GWY		CIA			GWY	
Region	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
Close interval	0.93	0.71	0.90	1.14	1.37	19	15	18	21	22	0	0	0	0	0
Medium interval	1.49	2.17	2.17	1.79	2.44	23	25	25	24	25	0	1	1	0	1
Far interval	2.70	3.57	3.85	2.04	3.13	26	27	27	25	26	1	2	2	1	2

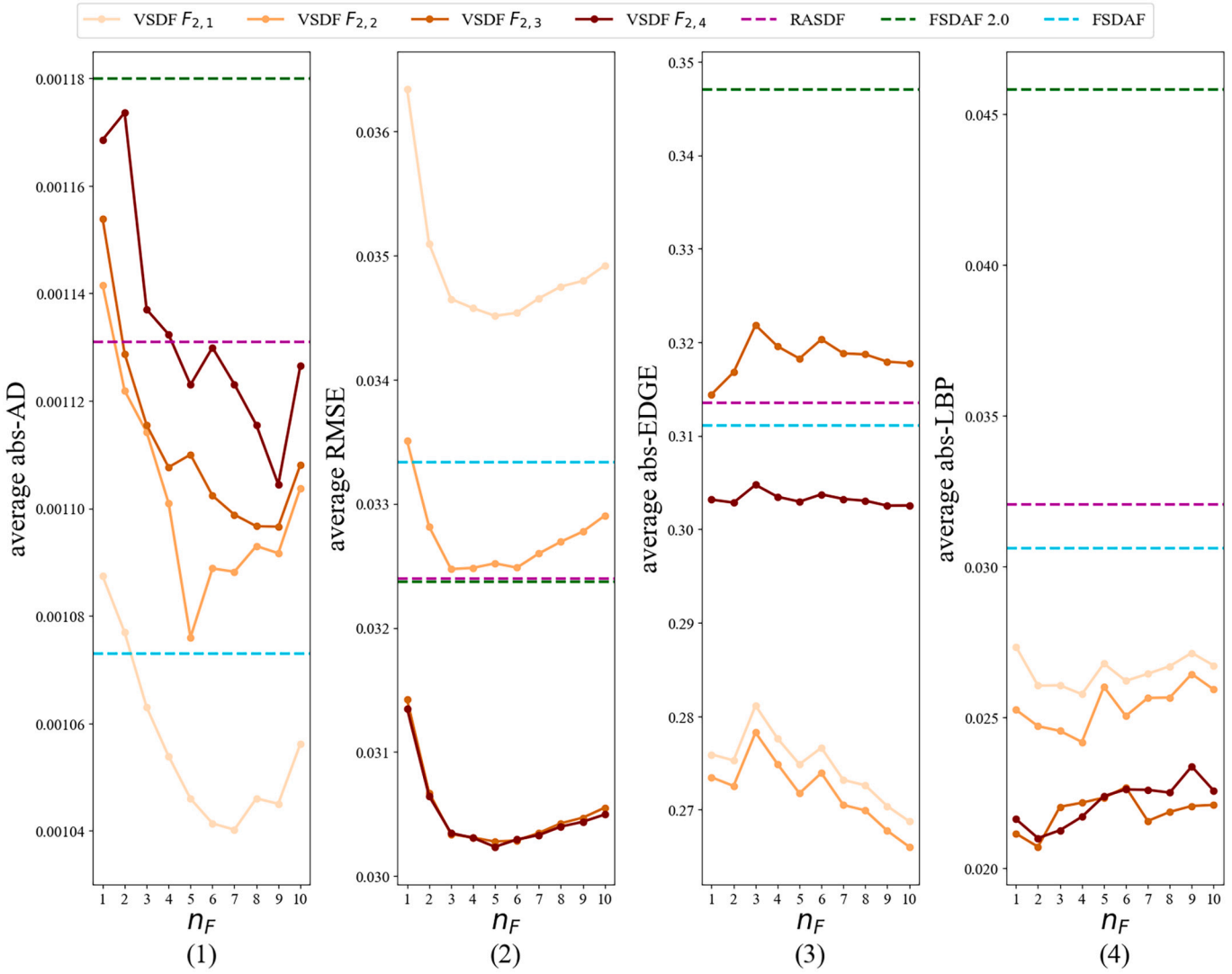

Fig. 21. Average absolute all-round performance assessment (APA) metrics of VSDF with different n_F: (1) average abs-AD of 31 tests, (2) average RMSE of 31 tests, (3) average abs-EDGE of 31 tests, and (4) average abs-LBP of 31 tests. The performance of RASDF, FSDAF 2.0, and FSDAF are added for comparison.

Code availability

The Python script of VSDF is available at https://github.com/ChenXuAxel/VSDF, or available upon request to the corresponding author.

CRediT authorship contribution statement

Chen Xu: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Software. **Xiaoping Du:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Funding acquisition. **Zhenzhen Yan:** Writing – review & editing, Funding acquisition. **Junjie Zhu:** Writing – review & editing, Funding acquisition. **Shu Xu:** Formal analysis, Writing – review & editing. **Xiangtao Fan:** Conceptualization, Funding acquisition, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Project title: CASEarth (XDA19080103 and XDA19080101)), the Innovation Drive Development Special Project of Guangxi (GuikeAA20302022), and the National Natural Science Foundation of China (41974108).

Appendix A. Appendix

Table A1
All-round performance assessment (APA) metrics for experiments in Section 4.2 and Section 4.3: AD.

Prediction feeding Landsat/simulated MODIS pairs															
Type	**CIA**									**GWY**					
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**	
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**
$F_{2,1}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$F_{2,2}$	−0.00001	0.00000	0.00001	−0.00001	0.00001	0.00000	0.00000	0.00002	0.00000	0.00001	0.00000	0.00000	0.00000	−0.00001	0.00000
$F_{2,3}$	−0.00011	0.00002	0.00001	−0.00010	0.00015	0.00015	0.00021	0.00026	0.00034	−0.00001	0.00006	0.00005	0.00012	0.00008	0.00016
$F_{2,4}$ (VSDF)	−0.00022	−0.00003	−0.00009	−0.00081	0.00004	−0.00011	−0.00035	0.00020	0.00033	0.00005	0.00005	−0.00003	0.00002	−0.00004	0.00003
RASDF	0.00040	0.00005	0.00013	0.00044	0.00030	0.00015	0.00083	0.00026	0.00046	0.00002	0.00001	0.00000	0.00001	0.00003	0.00003
FSDAF 2.0	0.00012	0.00000	0.00009	0.00040	0.00014	0.00024	N.A.	0.00012	0.00061	−0.00014	−0.00002	0.00004	0.00007	0.00004	N.A.
FSDAF	0.00021	0.00004	0.00011	0.00011	0.00014	0.00013	0.00024	0.00015	0.00019	0.00005	0.00001	−0.00001	0.00002	0.00001	0.00011
Fit-FC	0.00023	0.00008	0.00011	0.00018	0.00012	0.00010	0.00028	0.00011	0.00022	0.00015	0.00004	0.00004	0.00000	0.00000	0.00000
STARFM	0.00185	0.00020	0.00071	0.00107	−0.00007	0.00008	−0.00176	−0.00092	−0.00101	0.00085	0.00037	0.00084	0.00046	0.00110	0.00058
Prediction feeding Landsat/real MODIS pairs															
Type	**CIA**									**GWY**					
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**	
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**
$F_{2,1}$	−0.00264	−0.00175	−0.00250	0.00299	0.00169	0.00074	−0.00033	−0.00065	−0.00173	−0.00168	−0.00491	−0.00510	−0.00521	−0.00034	−0.00014
$F_{2,2}$	−0.00264	−0.00175	−0.00250	0.00299	0.00179	0.00090	−0.00027	−0.00081	−0.00183	−0.00168	−0.00491	−0.00510	−0.00530	−0.00061	−0.00016
$F_{2,3}$	−0.00265	−0.00174	−0.00245	0.00302	0.00188	0.00098	−0.00020	−0.00071	−0.00170	−0.00165	−0.00483	−0.00497	−0.00515	−0.00052	0.00002
$F_{2,4}$ (VSDF)	−0.00279	−0.00170	−0.00250	0.00263	0.00191	0.00089	−0.00043	−0.00068	−0.00173	−0.00165	−0.00484	−0.00490	−0.00518	−0.00054	−0.00001
RASDF	−0.00218	−0.00184	−0.00257	0.00287	0.00156	0.00049	−0.00122	−0.00170	−0.00251	−0.00084	−0.00423	−0.00441	−0.00469	0.00035	0.00027
FSDAF 2.0	−0.00204	−0.00184	−0.00246	0.00324	0.00166	0.00050	−0.00100	−0.00170	−0.00250	−0.00111	−0.00433	−0.00447	−0.00474	0.00033	0.00023
FSDAF	−0.00218	−0.00186	−0.00257	0.00288	0.00160	0.00045	−0.00110	−0.00164	−0.00252	−0.00088	−0.00430	−0.00437	−0.00470	0.00042	0.00028
Fit-FC	0.00009	−0.00088	−0.00091	0.00764	0.00568	0.00409	0.00664	0.00533	0.00479	−0.00530	−0.00844	−0.01315	−0.01466	−0.01057	−0.01201
STARFM	−0.00106	−0.00163	−0.00207	0.00352	0.00144	0.00083	−0.00266	−0.00257	−0.00302	−0.00041	−0.00425	−0.00410	−0.00491	0.00060	−0.00015

Table A2
All-round performance assessment (APA) metrics for experiments in Section 4.2 and Section 4.3: RMSE.

Prediction feeding Landsat/simulated MODIS pairs															
Type	**CIA**									**GWY**					
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**	
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**
$F_{2,1}$	0.03280	0.02118	0.02236	0.03984	0.03273	0.04294	0.04598	0.03763	0.04025	0.02009	0.01707	0.02678	0.02605	0.02837	0.02899
$F_{2,2}$	0.02972	0.01918	0.01940	0.03449	0.02831	0.03624	0.04001	0.03205	0.03481	0.01795	0.01540	0.02362	0.02332	0.02491	0.02555
$F_{2,3}$	0.02752	0.01770	0.01819	0.03053	0.02601	0.03388	0.03508	0.02793	0.03070	0.01666	0.01364	0.02064	0.01997	0.02130	0.02148
$F_{2,4}$ (VSDF)	0.02679	0.01740	0.01792	0.03035	0.02570	0.03327	0.03507	0.02743	0.03036	0.01655	0.01351	0.02068	0.02023	0.02141	0.02198
RASDF	0.02914	0.01760	0.02010	0.03466	0.02818	0.03742	0.04174	0.03177	0.03680	0.01702	0.01359	0.02314	0.02212	0.02453	0.02520
FSDAF 2.0	0.02753	0.01756	0.01923	0.03357	0.02738	0.03483	N.A.	0.03069	0.03655	0.01688	0.01394	0.02365	0.02215	0.02522	N.A.
FSDAF	0.02869	0.01799	0.01968	0.03523	0.02916	0.03516	0.04280	0.03363	0.03714	0.01805	0.01398	0.02335	0.02312	0.02611	0.02626
Fit-FC	0.02895	0.01783	0.01972	0.03476	0.02715	0.03569	0.04250	0.03084	0.03684	0.01770	0.01353	0.02267	0.02260	0.02452	0.02579
STARFM	0.03022	0.01864	0.02079	0.04314	0.03116	0.04095	0.05666	0.03610	0.04452	0.01854	0.01438	0.02563	0.02570	0.02809	0.02954

(*continued on next page*)

Table A2 (*continued*)

Prediction feeding Landsat/simulated MODIS pairs															
Type	**CIA**									**GWY**					
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**	
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**
Prediction feeding Landsat/real MODIS pairs															
Type	**CIA**									**GWY**					
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**	
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**
$F_{2,1}$	0.03767	0.02337	0.02543	0.05006	0.04080	0.05289	0.06358	0.04893	0.04954	0.02262	0.01968	0.03195	0.03261	0.03632	0.03675
$F_{2,2}$	0.03767	0.02337	0.02543	0.05006	0.04042	0.05235	0.06238	0.05072	0.04932	0.02262	0.01968	0.03195	0.03215	0.03594	0.03801
$F_{2,3}$	0.03681	0.02312	0.02506	0.04889	0.03879	0.05127	0.06014	0.04790	0.04701	0.02201	0.01890	0.03086	0.02964	0.03356	0.03452
$F_{2,4}$ (VSDF)	0.03671	0.02304	0.02498	0.04916	0.03874	0.05141	0.06066	0.04783	0.04735	0.02198	0.01880	0.03119	0.02998	0.03386	0.03503
RASDF	0.03712	0.02386	0.02571	0.04866	0.04172	0.05324	0.06449	0.05211	0.05145	0.02225	0.01791	0.03208	0.03080	0.03543	0.03362
FSDAF 2.0	0.03768	0.02421	0.02570	0.05093	0.04181	0.05348	0.06621	0.05218	0.05070	0.02302	0.01885	0.03228	0.03180	0.03569	0.03689
FSDAF	0.03814	0.02449	0.02633	0.05166	0.04240	0.05378	0.06802	0.05266	0.05283	0.02407	0.01923	0.03283	0.03281	0.03733	0.03653
Fit-FC	0.04099	0.02393	0.02783	0.04715	0.03837	0.05167	0.05873	0.04500	0.04915	0.02564	0.02096	0.03257	0.03305	0.03587	0.03685
STARFM	0.03796	0.02424	0.02616	0.05736	0.04333	0.05657	0.07729	0.05461	0.05965	0.02392	0.01897	0.03543	0.03654	0.04033	0.04260

Table A3
All-round performance assessment (APA) metrics for experiments in Section 4.2 and Section 4.3: EDGE.

Prediction feeding Landsat/simulated MODIS pairs															
Type	**CIA**									**GWY**					
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**	
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**
$F_{2,1}$	−0.24591	−0.23198	−0.20082	−0.2711	−0.3385	−0.28815	−0.23388	−0.25982	−0.26693	−0.17894	−0.19695	−0.2328	−0.27243	−0.20124	−0.21816
$F_{2,2}$	−0.24104	−0.22794	−0.19745	−0.26126	−0.3297	−0.27972	−0.22379	−0.25109	−0.25821	−0.17703	−0.19441	−0.22938	−0.26771	−0.19622	−0.21373
$F_{2,3}$	−0.27066	−0.24923	−0.20683	−0.32869	−0.3905	−0.33017	−0.27573	−0.32651	−0.31293	−0.18916	−0.21293	−0.27297	−0.3414	−0.25376	−0.30431
$F_{2,4}$ (VSDF)	−0.27276	−0.24621	−0.20643	−0.32707	−0.37157	−0.31772	−0.33689	−0.3374	−0.34346	−0.16885	−0.17963	−0.23849	−0.30728	−0.21517	−0.2764
RASDF	−0.27675	−0.24767	−0.20724	−0.39466	−0.39432	−0.31272	−0.3652	−0.39984	−0.29867	−0.17716	−0.21993	−0.28178	−0.34567	−0.23098	−0.25406
FSDAF 2.0	−0.30708	−0.2728	−0.2311	−0.42847	−0.46309	−0.35913	N.A.	−0.4551	−0.31321	−0.21174	−0.25314	−0.3069	−0.43416	−0.24596	N.A.
FSDAF	−0.31009	−0.27821	−0.24084	−0.37267	−0.38686	−0.31949	−0.36412	−0.34231	−0.28148	−0.20428	−0.24461	−0.25493	−0.30082	−0.19	−0.22206
Fit-FC	−0.3891	−0.3557	−0.32347	−0.56231	−0.56032	−0.53029	−0.47599	−0.49856	−0.48155	−0.28774	−0.33369	−0.47271	−0.59105	−0.44896	−0.52292
STARFM	−0.35026	−0.35192	−0.31036	−0.45969	−0.51470	−0.44136	−0.50621	−0.48924	−0.50003	−0.25320	−0.27336	−0.34307	−0.44206	−0.32758	−0.41285
Prediction feeding Landsat/real MODIS pairs															
Type	**CIA**									**GWY**					
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**	
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**
$F_{2,1}$	−0.2887	−0.25587	−0.21368	−0.36548	−0.37619	−0.34362	−0.41155	−0.37823	−0.40079	−0.19741	−0.21639	−0.27032	−0.29691	−0.234	−0.25839
$F_{2,2}$	−0.2887	−0.25587	−0.21368	−0.36548	−0.37522	−0.3451	−0.41225	−0.3734	−0.40098	−0.19741	−0.21639	−0.27032	−0.2956	−0.2408	−0.25592
$F_{2,3}$	−0.313	−0.2661	−0.22444	−0.4336	−0.42411	−0.39359	−0.503	−0.42966	−0.46059	−0.20321	−0.22995	−0.3009	−0.36198	−0.29472	−0.34244
$F_{2,4}$ (VSDF)	−0.29837	−0.26087	−0.21955	−0.40644	−0.41068	−0.36928	−0.49835	−0.42548	−0.45534	−0.18144	−0.19116	−0.26108	−0.32368	−0.24549	−0.30703
RASDF	−0.30038	−0.25458	−0.22441	−0.43449	−0.40478	−0.34324	−0.5197	−0.43724	−0.37158	−0.2006	−0.22895	−0.2834	−0.38041	−0.24575	−0.32362
FSDAF 2.0	−0.3107	−0.26125	−0.23599	−0.44046	−0.43584	−0.36984	−0.53419	−0.4756	−0.42403	−0.19557	−0.24618	−0.29277	−0.43152	−0.24952	−0.39191
FSDAF	−0.31527	−0.27716	−0.24594	−0.42983	−0.39679	−0.35039	−0.49171	−0.41431	−0.40474	−0.21403	−0.23224	−0.27569	−0.32766	−0.23054	−0.26564
Fit-FC	−0.48451	−0.40138	−0.41082	−0.61288	−0.57302	−0.57687	−0.56729	−0.52264	−0.53954	−0.40056	−0.45045	−0.533	−0.62627	−0.50931	−0.55605
STARFM	−0.35282	−0.34854	−0.31068	−0.45302	−0.51192	−0.44240	−0.52583	−0.51074	−0.53030	−0.25179	−0.26062	−0.33606	−0.43295	−0.31517	−0.40653

Table A4
All-round performance assessment (APA) metrics for experiments in Section 4.2 and Section 4.3: LBP.

Prediction feeding Landsat/simulated MODIS pairs															
Type	**CIA**									**GWY**					
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**	
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**
$F_{2,1}$	0.04083	0.03786	0.02374	0.04440	0.03002	0.02887	−0.01829	−0.01703	−0.01493	0.02889	0.03556	0.00228	−0.00307	0.05693	0.04950
$F_{2,2}$	0.05433	0.04450	0.03376	0.06790	0.04971	0.05140	0.01161	0.01031	0.01528	0.00688	0.01538	−0.02118	−0.02565	0.03256	0.02800
$F_{2,3}$	0.03135	0.03492	0.02228	0.03437	0.03222	0.02465	−0.01063	0.00218	0.00114	−0.01640	−0.02357	−0.05314	−0.07754	−0.00489	−0.02417
$F_{2,4}$ (VSDF)	0.03737	0.03762	0.02541	0.04970	0.04179	0.03497	0.00637	0.01186	0.01202	−0.00842	−0.01224	−0.03756	−0.06061	0.01440	−0.00429
RASDF	0.03803	0.02597	0.02600	0.02161	0.02382	0.05683	0.01323	−0.01332	0.05403	−0.01694	−0.03704	−0.04627	−0.06405	0.01515	0.00652
FSDAF 2.0	0.01112	0.00057	0.00163	−0.02016	−0.06059	0.00536	N.A.	−0.09375	0.03551	−0.06718	−0.10533	−0.09835	−0.16294	−0.02919	N.A.
FSDAF	0.00665	−0.01471	−0.01426	0.00444	0.00329	0.00060	−0.02459	−0.01912	0.00248	−0.05741	−0.08096	−0.07905	−0.11256	−0.02239	−0.04987
Fit-FC	−0.13487	−0.12501	−0.13232	−0.19791	−0.19246	−0.20362	−0.14138	−0.18615	−0.14658	−0.21239	−0.25367	−0.34161	−0.41982	−0.27257	−0.33227
STARFM	−0.20521	−0.24263	−0.23667	−0.19253	−0.23284	−0.22976	−0.22699	−0.27842	−0.25729	−0.31654	−0.30334	−0.33467	−0.34116	−0.28399	−0.28628
Prediction feeding Landsat/real MODIS pairs															

(*continued on next page*)

Table A4 (*continued*)

Prediction feeding Landsat/simulated MODIS pairs																
Type	**CIA**									**GWY**						
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**		
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**	
Type	**CIA**									**GWY**						
Time	**close interval**			**medium interval**			**far interval**			**close interval**		**medium interval**		**far interval**		
Region	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**3**	**1**	**3**	**1**	**3**	
$F_{2,1}$	0.03667	0.01929	0.00901	0.02490	0.02012	0.00871	−0.02885	−0.03229	−0.03273	0.02427	0.02405	0.00283	−0.00872	0.05985	0.05265	
$F_{2,2}$	0.03667	0.01929	0.00901	0.02490	0.03261	0.01940	−0.00759	−0.00326	−0.00414	0.02427	0.02405	0.00283	−0.03075	0.03230	0.02949	
$F_{2,3}$	0.02399	0.01421	0.00282	0.00276	0.02300	0.00312	−0.03115	−0.00540	−0.02050	−0.00736	−0.02234	−0.04111	−0.07345	0.00620	−0.01131	
$F_{2,4}$ **(VSDF)**	0.02840	0.01585	0.00558	0.01641	0.03026	0.01373	−0.01581	0.00119	−0.00994	−0.00730	−0.01374	−0.03445	−0.05900	0.01947	0.00372	
RASDF	0.03554	0.03428	0.03092	0.01622	0.03747	0.03751	−0.02890	−0.01718	0.03674	−0.02610	−0.03255	−0.03826	−0.07270	0.03098	−0.00703	
FSDAF 2.0	0.02258	0.02759	0.01618	−0.00334	−0.00246	0.00518	−0.05762	−0.06692	−0.00289	−0.02988	−0.06340	−0.06568	−0.11702	0.00444	−0.05571	
FSDAF	0.01823	0.00697	−0.00086	−0.00435	0.01698	−0.00023	−0.04210	−0.02406	−0.02292	−0.03877	−0.05733	−0.06005	−0.08756	0.00067	−0.02635	
Fit-FC	−0.18430	−0.15613	−0.18858	−0.23040	−0.18981	−0.23062	−0.18660	−0.18599	−0.17588	−0.26107	−0.34098	−0.37443	−0.44400	−0.29432	−0.34667	
STARFM	−0.19096	−0.22594	−0.22355	−0.18254	−0.21019	−0.21906	−0.23951	−0.26387	−0.26392	−0.29420	−0.27624	−0.31023	−0.31267	−0.25694	−0.25517	

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2274–2282. https://doi.org/10.1109/TPAMI.2012.120.

Ao, Z., Sun, Y., Pan, X., Xin, Q., 2022. Deep learning-based spatiotemporal data fusion using a patch-to-pixel mapping strategy and model comparisons. IEEE Trans. Geosci. Remote Sens. 60, 1–18. https://doi.org/10.1109/TGRS.2022.3154406.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. 10, 191–203. https://doi.org/10.1016/0098-3004(84)90020-7.

Cai, J., Huang, B., Fung, T., 2022. Progressive spatiotemporal image fusion with deep neural networks. Int. J. Appl. Earth Obs. Geoinf. 108, 102745 https://doi.org/10.1016/j.jag.2022.102745.

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8, 679–698. https://doi.org/10.1109/TPAMI.1986.4767851.

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.C., Skakun, S.V., Justice, C., 2018. The harmonized landsat and Sentinel-2 surface reflectance data set. Remote Sens. Environ. 219, 145–161. https://doi.org/10.1016/j.rse.2018.09.002.

Emelyanova, I.V., McVicar, T.R., Van Niel, T.G., Li, L.T., van Dijk, A.I.J.M., 2013. Assessing the accuracy of blending landsat-MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: a framework for algorithm selection. Remote Sens. Environ. 133, 193–209. https://doi.org/10.1016/j.rse.2013.02.007.

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 Trans. Geosci. Remote Sens. 44, 2207–2218. https://doi.org/10.1109/TGRS.2006.872081.

Gevaert, C.M., García-Haro, F.J., 2015. A comparison of STARFM and an unmixing-based algorithm for landsat and MODIS data fusion. Remote Sens. Environ. 156, 34–44. https://doi.org/10.1016/j.rse.2014.09.012.

Guo, D., Shi, W., Hao, M., Zhu, X., 2020. FSDAF 2.0: improving the performance of retrieving land cover changes and preserving spatial details. Remote Sens. Environ. 248, 111973 https://doi.org/10.1016/j.rse.2020.111973.

Guo, H., Chen, F., Sun, Z., Liu, J., Liang, D., 2021. Big Earth Data: a practice of sustainability science to achieve the sustainable development goals. Sci. Bull. 66, 1050–1053. https://doi.org/10.1016/j.scib.2021.01.012.

Hansen, M.C., Loveland, T.R., 2012. A review of large area monitoring of land cover change using Landsat data. Remote Sens. Environ. 122, 66–74. https://doi.org/10.1016/j.rse.2011.08.024.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a K-means clustering algorithm. Appl. Stat. 28, 100. https://doi.org/10.2307/2346830.

He, K., Sun, J., Tang, X., 2013. Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1397–1409. https://doi.org/10.1109/TPAMI.2012.213.

He, K., Sun, J., Tang, X., 2010. Guided image filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision – ECCV 2010. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 1–14.

Hilker, T., Wulder, M.A., Coops, N.C., Linke, J., McDermid, G., Masek, J.G., Gao, F., White, J.C., 2009. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on landsat and MODIS. Remote Sens. Environ. 113, 1613–1627. https://doi.org/10.1016/j.rse.2009.03.007.

Huang, B., Song, H., 2012. Spatiotemporal reflectance fusion via sparse representation. IEEE Trans. Geosci. Remote Sens. 50, 3707–3716. https://doi.org/10.1109/TGRS.2012.2186638.

Kang, Y., Özdoğan, M., 2019. Field-level crop yield mapping with landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163. https://doi.org/10.1016/j.rse.2019.04.005.

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 Sens. Environ. 135, 52–63. https://doi.org/10.1016/j.rse.2013.03.021.

Li, X., Foody, G.M., Boyd, D.S., Ge, Y., Zhang, Y., Du, Y., Ling, F., 2020. SFSDAF: an enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion. Remote Sens. Environ. 237, 111537 https://doi.org/10.1016/j.rse.2019.111537.

Li, X., Long, D., 2020. An improvement in accuracy and spatiotemporal continuity of the MODIS precipitable water vapor product based on a data fusion approach. Remote Sens. Environ. 248, 111966 https://doi.org/10.1016/j.rse.2020.111966.

Liu, H., Gong, P., Wang, J., Wang, X., Ning, G., Xu, B., 2021. Production of global daily seamless data cubes and quantification of global land cover change from 1985 to 2020 - iMap world 1.0. Remote Sens. Environ. 258, 112364 https://doi.org/10.1016/j.rse.2021.112364.

Liu, Maolin, Ke, Y., Yin, Q., Chen, X., Im, J., 2019. Comparison of five spatio-temporal satellite image fusion models over landscapes with various spatial heterogeneity and temporal variation. Remote Sens. 11 https://doi.org/10.3390/rs11222612.

Liu, Meng, 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 Sens. Environ. 227, 74–89. https://doi.org/10.1016/j.rse.2019.03.012.

Masek, J.G., Wulder, M.A., Markham, B., McCorkel, J., Crawford, C.J., Storey, J., Jenstrom, D.T., 2020. Landsat 9: empowering open science and applications through continuity. Remote Sens. Environ. 248 https://doi.org/10.1016/j.rse.2020.111968.

Peng, K., Wang, Q., Tang, Y., Tong, X., Atkinson, P.M., 2022. Geographically weighted spatial unmixing for spatiotemporal fusion. IEEE Trans. Geosci. Remote Sens. 60 https://doi.org/10.1109/TGRS.2021.3115136.

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

Shen, H., Meng, X., Zhang, L., 2016. An integrated framework for the spatio-temporal-spectral fusion of remote sensing images. IEEE Trans. Geosci. Remote Sens. 54, 7135–7148. https://doi.org/10.1109/TGRS.2016.2596290.

Shi, W., Guo, D., Zhang, H., 2022. A reliable and adaptive spatiotemporal data fusion method for blending multi-spatiotemporal-resolution satellite images. Remote Sens. Environ. 268, 112770 https://doi.org/10.1016/j.rse.2021.112770.

Tan, Z., Gao, M., Li, X., Jiang, L., 2022. A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network. IEEE Trans. Geosci. Remote Sens. 60 https://doi.org/10.1109/TGRS.2021.3050551.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234. https://doi.org/10.2307/143141.

Wang, Q., Atkinson, P.M., 2018. Spatio-temporal fusion for daily Sentinel-2 images. Remote Sens. Environ. 204, 31–42. https://doi.org/10.1016/j.rse.2017.10.046.

Wulder, M.A., Hilker, T., White, J.C., Coops, N.C., Masek, J.G., Pflugmacher, D., Crevier, Y., 2015. Virtual constellations for global terrestrial monitoring. Remote Sens. Environ. 170, 62–76. https://doi.org/10.1016/j.rse.2015.09.001.

Xu, C., Du, X., Fan, X., Yan, Z., Kang, X., Zhu, J., Hu, Z., 2022a. A modular remote sensing big data framework. IEEE Trans. Geosci. Remote Sens. 60, 1–11. https://doi.org/10.1109/TGRS.2021.3100601.

Xu, C., Du, X., Jian, H., Dong, Y., Qin, W., Mu, H., Yan, Z., Zhu, J., Fan, X., 2022b. Analyzing large-scale data cubes with user-defined algorithms: a cloud-native approach. Int. J. Appl. Earth Obs. Geoinf. 109, 102784 https://doi.org/10.1016/j.jag.2022.102784.

Zhang, J., 2010. Multi-source remote sensing data fusion: status and trends. Int. J. Image Data Fusion 1, 5–24. https://doi.org/10.1080/19479830903561035.

Zhao, M., Zhou, Y., Li, X., Cheng, W., Zhou, C., Ma, T., Li, M., Huang, K., 2020. Mapping urban dynamics (1992–2018) in Southeast Asia using consistent nighttime light data from DMSP and VIIRS. Remote Sens. Environ. 248, 111980 https://doi.org/10.1016/j.rse.2020.111980.

Zhou, J., Chen, J., Chen, X., Zhu, X., Qiu, Y., Song, H., Rao, Y., Zhang, C., Cao, X., Cui, X., 2021. Sensitivity of six typical spatiotemporal fusion methods to different influential

factors: a comparative study for a normalized difference vegetation index time series reconstruction. Remote Sens. Environ. 252, 112130 https://doi.org/10.1016/j.rse.2020.112130.

Zhu, X., Cai, F., Tian, J., Williams, T.K.A., 2018. Spatiotemporal fusion of multisource remote sensing data: literature survey, taxonomy, principles, applications, and future directions. Remote Sens. 10 https://doi.org/10.3390/rs10040527.

Zhu, X., Chen, J., Gao, F., Chen, X., Masek, J.G., 2010. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. Remote Sens. Environ. 114, 2610–2623. https://doi.org/10.1016/j.rse.2010.05.032.

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 Sens. Environ. 172, 165–177. https://doi.org/10.1016/j.rse.2015.11.016.

Zhu, X., Zhan, W., Zhou, J., Chen, X., Liang, Z., Xu, S., Chen, J., 2022. A novel framework to assess all-round performances of spatiotemporal fusion models. Remote Sens. Environ. 274, 113002 https://doi.org/10.1016/j.rse.2022.113002.

Zhukov, B., Oertel, D., Lanzl, F., Reinhackel, G., 1999. Unmixing-based multisensor multiresolution image fusion. IEEE Trans. Geosci. Remote Sens. 37, 1212–1226. https://doi.org/10.1109/36.763276.

Zurita-Milla, R., Clevers, J.G.P.W., Schaepman, M.E., 2008. Unmixing-based landsat TM and MERIS FR data fusion. IEEE Geosci. Remote Sens. Lett. 5, 453–457. https://doi.org/10.1109/LGRS.2008.919685.