

Online Fusion of Multi-resolution Multispectral Images with Weakly Supervised Temporal Dynamics

Haoqing Li^a, Bhavya Duvvuri^a, Ricardo Borsoi^b, Tales Imbiriba^a, Edward Beighley^a,
Deniz Erdoğan^a, Pau Closas^a

^a*Northeastern University, Boston, 02215, MA, USA*

^b*University of Lorraine, CNRS, CRAN, Nancy, F-54000, France*

Abstract

Real-time satellite imaging has a central role in monitoring, detecting and estimating the intensity of key natural phenomena such as floods, earthquakes, etc. One important constraint of satellite imaging is the trade-off between spatial/spectral resolution and their revisiting time, a consequence of design and physical constraints imposed by satellite orbit among other technical limitations. In this paper, we focus on fusing multi-temporal, multi-spectral images where data acquired from different instruments with different spatial resolutions is used. We leverage the spatial relationship between images at multiple modalities to generate high-resolution image sequences at higher revisiting rates. To achieve this goal, we formulate the fusion method as a recursive state estimation problem and study its performance in filtering and smoothing contexts. Furthermore, a calibration strategy is proposed to estimate the time-varying temporal dynamics of the image sequence using only a small amount of historical image data. Differently from the training process in traditional machine learning algorithms, which usually require large datasets and computation times, the parameters of the temporal dynamical model are calibrated based on an analytical expression that uses only two of the images in the historical dataset. A distributed version of the Bayesian filtering and smoothing strategies is also proposed to reduce its computational complexity. To evaluate the proposed methodology we consider a water mapping task where real data acquired by the Landsat and MODIS instruments are fused generating high spatial-temporal resolution image estimates. Our experiments show that the proposed methodology outperforms the competing methods in both estimation accuracy and water mapping tasks.

Keywords: Multimodal image fusion, Online Fusion, Bayesian Filtering, Water mapping, Super-resolution

1. Introduction

High spatial resolution satellite image data is a fundamental tool for remote sensing applications such as the monitoring of land cover changes [1, 2], deforestation [3, 4] or water mapping [5, 6] and water quality [7]. Moreover, to adequately deal with the variability

of such events over time it is important to have short time spans between different image acquisitions of the same scene (i.e., a high temporal resolution, or low revisit times). However, fundamental limitations of multiband imaging instruments and large sensor-to-target distances impose a trade-off between spatial and temporal resolutions of satellite image sequences.

This means that instruments providing high spatial resolution have long revisit times, while the converse holds for instruments with short revisit times. This can be illustrated, for instance, by considering Landsat 8 and MODIS instruments (with 30 and 250/500 meters spatial resolution, respectively). While MODIS is able to provide daily images at coarse resolution, Landsat-8 only revisits the same site once every 16 days [8].

Considering these limitations, many works proposed multimodal image fusion techniques to generate high (spatial, spectral or temporal) resolution remote sensing images. Multimodal image fusion aims to combine multiple observed images, each of which having high resolution in a given dimension – spatial, temporal, or spectral – to generate high resolution image sequences. Several instances of image fusion have been considered, some works aim to directly supply classification maps from multiple satellite image and surface elevation data at each time instant [9], integrating optical and radar data for time-series crop classification [10, 11], or fusing spatio-temporal optical and elevation data to obtain high-resolution land temperature maps [12].

In particular, classification or mapping tasks based on time-series remote sensing data is receiving increasing interest in the literature [13, 14, 10, 11]. Thus, to overcome the limitations of existing instruments, fusing images with different spectral and spatial resolutions has been extensively studied to generate images with high spatial and spectral resolutions, which are critical for accurately distinguishing different materials in a pixel [15, 16, 17]. Recently, an increasing interest has been observed in applying multimodal image fusion to generate image sequences with high spatial and temporal resolutions [18], with particular interest dedicated to fusing data from multiple satellites to obtain daily images with high (e.g., 30 m) resolution [19]. This has already had an important impact in applications such as the generation of daily snow cover maps [20] and the study of drought-induced tree mortality [21]. Existing spatiotemporal image fusion methods are usually divided in weighted fusion, unmixing-based, learning-based and Bayesian approaches [22]. There also exist hybrid techniques, which leverage ideas from more than one family of approaches.

Weighted fusion methods assume that the temporal changes occurring between two time instants are consistent between the high and low spatial resolution images for low resolution pixels which are composed of only a single material [23]. However, coarse resolution pixels are often mixtures of different materials. The predicted high resolution pixels are then computed as a weighted linear combination of the previous high resolution pixels and of the changes occurring at low resolution pixels in a given neighborhood [24, 25]. Different works have designed various weighting functions, which aim to select neighboring pixels that are homogeneous and spatially/spectrally similar to the pixel whose change is being predicted [24, 22, 26]. Other works have extended such framework account for sudden changes [27] or to use different weighting functions [25].

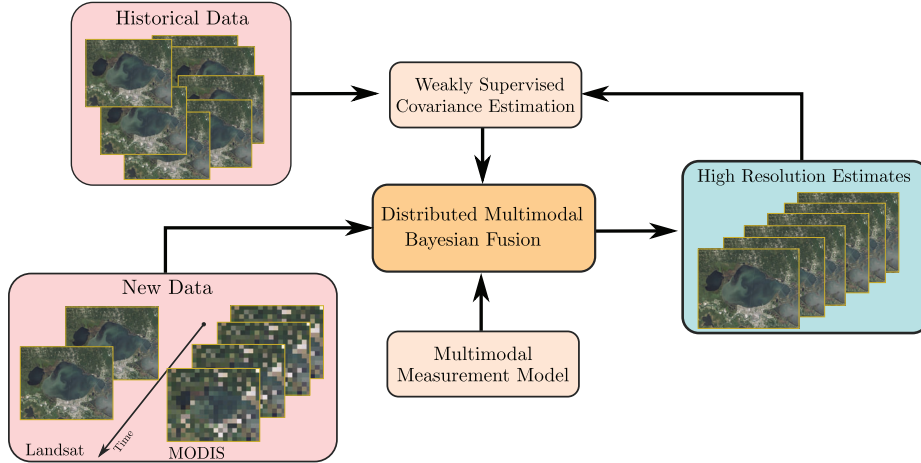


Figure 1: Overview of the proposed method. Multimodal (e.g., Landsat and MODIS over time) images time series are fused by the Distributed Multimodal Bayesian Fusion algorithm resulting in a high spatial-temporal resolution estimated sequence. Covariance estimates for the dynamical model are estimated through a weakly supervised strategy based on local high-resolution historical data. We highlight that the Bayesian fusion methodology employed here is agnostic to the multimodal measurement model making the strategy easily generalizable to different data scenarios.

Unmixing-based methods make use of the linear mixing model (LMM), which assumes that each pixel in the low resolution image can be represented as a convex combination of the reflectance of a small number of pure spectral signatures, called endmembers [28, 29]. The LMM has been used for multimodal image fusion by assuming the proportions of each material in a low resolution pixel to be stable/constant over time [30, 31, 32]. This way, spectral unmixing [28] is used to estimate the endmembers at different time instants from low resolution images, while using different strategies to mitigate the spectral variability of a single material [30, 33, 34]. However, abrupt abundance variations (originating from, e.g., land cover changes) are commonly found in multitemporal image streams [35, 36, 37, 38], which may negatively impact the performance of such methods and can be particularly challenging to address when occurring jointly with finer endmember variations [35]. Thus, special care is required when fusing images which are temporally distant from one another [39], motivating the development of strategies using, e.g., spatially adaptive quantification of the reliability of the input images to guide unmixing based image fusion strategies [40].

Learning-based approaches leverage training data and different machine learning algorithms in order to perform image fusion. Those approaches are varied, ranging from approaches such as dictionary learning [41], which are based on a sparse representation of image pixels and have a strong connection to the LMM, to convolutional neural networks [42], which are flexible function approximations which are typically used to learn a mapping from the low-resolution to high resolution data.

Bayesian methods are flexible alternatives to the previous approaches that take into account the uncertainty present both in the imaging model and in the estimated images. The

Bayesian framework is based on the definition of probabilistic models to describe the relationship between images of different spatial, spectral and temporal resolutions acquired by different instruments. This allows image fusion to be formulated as a maximum a posteriori estimation problem [43]. Although Bayesian methods usually consider Gaussian distributions for mathematical tractability, different variations have been proposed depending on how the image acquisition process is modelled and on how the mean and covariance matrices are estimated. This included assuming them diagonal [44], estimating image covariance matrices based on an initial estimate of the high resolution image [43], or based on the low resolution image pixels [45].

A recent work considered a Kalman filter-based approach to estimate a high resolution image sequence based on mixed resolution observations from the Landsat and MODIS instruments [46]. However, to define the model for the Kalman filter, two Landsat+MODIS image pairs at times t_0 and t_N are considered, as well as a time series of MODIS images at instants $t_k \in [t_0, t_N]$, making it unsuitable for online operation. Moreover, changes between each pair of images were assumed to be constant/uniform over predefined groups of high resolution image pixels, which can be restrictive (due to the large resolution difference between the measured images, the groups must contain many pixels in order to make the model well-posed). It also does not benefit from auxiliary information that could aid the estimation of the high resolution images. Another work used the Kalman filter to estimate normalized difference vegetation indices (NDVI) time series images from Landsat and MODIS observations, using an affine model for the dynamics of the states whose coefficients are selected based on the seasonality, and another affine model to relate the NDVI estimate obtained from MODIS and Landsat measurements [47]. The Kalman filter was also recently applied to estimate land surface temperature by fusing thermal infrared and microwave data [48].

In this paper, we propose a weakly supervised Kalman filter and smoother framework for spatio-temporal fusion of multispectral images. The proposed framework relies on explicit modeling assumptions about the image acquisition and temporal evolution processes, under which the proposed solution is statistically optimal. The Kalman filter-based methods can operate in a fully online setting, where high-resolution images are only available as past data. We also develop a smoother-based method to optimally exploit information contained in future high-resolution observed images when processing images in a time window. However, the quality of the reconstruction of Kalman filter and smoother strategies depend directly on the quality of the dynamical image evolution model. Thus, to overcome this limitation, a weakly supervised strategy is proposed to learn the temporal dynamics of the high-resolution images from a small amount of past data. More precisely, instead of considering the changes to be constant over areas comprising large amounts of image pixels, we propose an analytical calibration strategy to estimate a more informative time-varying dynamical image model by leveraging historical data. This allows for a better localization of changes in the high resolution image even in intervals where only coarse resolution observations (e.g., MODIS) are available. Moreover, to mitigate the high computational complexity of the Kalman filter and smoother, we propose a distributed implementation

by exploiting different independence assumptions about the high-resolution state space, allowing the proposed methods to be applied to large datasets and geographical areas. Figure 1 depicts the proposed methodology where high-resolution (spatially and temporally) estimates are generated by fusing different data modalities. We illustrate the application of the proposed framework by fusing images from the Landsat and MODIS instruments. Experimental results indicate that the proposed method can lead to considerable improvements compared to using a non-informative dynamical model and to widely used image fusion algorithms, both in image reconstruction and in downstream water classification and hydrograph estimation tasks. A software package containing an implementation of the proposed method and the image dataset is available at <https://github.com/HaoqingLi/Multi-resolution-Multispectral-image-fusion-based-weakly-supervised-constrained-Kalman-fusion>.

This paper is organized as follows. In Section 2, we present the paper notation and the proposed imaging model. Section 3 presents the Kalman filter and smoother approaches for multimodal image fusion. Section 5 contains simulation experiments that illustrate the performance of the proposed method. Finally, Section 6 concludes the paper.

2. Dynamical Imaging Model

2.1. Definitions and notation

Let us denote the ℓ -th band of the k -th acquired image reflectances from modality $m \in \Omega$ by $\mathbf{y}_{k,\ell}^m \in \mathbb{R}^{N_{m,\ell}}$, with $N_{m,\ell}$ pixels for each of the bands $\ell = 1, \dots, L_m$, and Ω denoting the set of image modalities. As a practical example, we consider $\Omega = \{\text{L}, \text{M}\}$ to contain the Landsat-8, and MODIS image modalities, without loss of generality. We also denote by Ω_H the highest resolution image modality, e.g., $\Omega_H = \{\text{L}\}$. We denote the corresponding high resolution latent reflectances by $\mathbf{S}_k \in \mathbb{R}^{N_H \times L_H}$, with N_H pixels and L_H bands, with $L_H \geq L_m$ and $N_H \geq N_{m,\ell}$, $\forall \ell, m$. Subindex $k \in \mathbb{N}_*$ denotes the acquisition time index. We also denote by $\text{vec}(\cdot)$, $\text{col}\{\cdot\}$, $\text{diag}\{\cdot\}$ and by $\text{blkdiag}\{\cdot\}$ the vectorization, vector stacking, diagonal and block diagonal matrix operators, respectively. The notation $\mathbf{x}_{a:b}$ for $a, b \in \mathbb{N}_*$ represents the set $\{\mathbf{x}_a, \mathbf{x}_{a+1}, \dots, \mathbf{x}_b\}$. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2.2. Measurement model

To formulate our measurement model we assume that the acquired image at time index k , for any imaging modality, is a spatially degraded and spectrally transformed version of the high resolution latent reflectance image $\mathbf{S}_k$. Following this assumption our measurement model for the m -th modality becomes:

$$\mathbf{y}_{k,\ell}^m = \mathcal{H}_\ell^m(\mathbf{S}_k) \mathbf{c}_\ell^m + \mathbf{r}_{k,\ell}^m, \quad \ell = 1, \dots, L_m, \quad (1)$$

where $\mathbf{c}_\ell^m \in \mathbb{R}^{L_H}$ denotes a spectral transformation vector, mapping all bands in $\mathbf{S}_k$ to the ℓ -th measured band at modality m ; $\mathcal{H}_\ell^m$ is a linear operator representing the band-wise spatial degradation, modeling blurring and downsampling effects of each high resolution band, and $\mathbf{r}_{k,\ell}^m$ represents the measurement noise. Note that, while we consider the spatial

resolution of the high resolution bands in $\mathbf{S}_k$ to be the same, different bands from the same modality can have different resolutions. We also assume the measurement noise to be Gaussian and uncorrelated among bands, that is, $\mathbf{r}_{k,\ell}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_\ell^m)$ with time-invariant covariance matrix given by $\mathbf{R}_\ell^m \in \mathbb{R}^{N_{m,\ell} \times N_{m,\ell}}$, and $\text{cov}(\mathbf{r}_{k,j}^m, \mathbf{r}_{k,\ell}^m) = \mathbf{0}$ for all $j \neq \ell$.

Note that satellite images may be corrupted by several effects, including dead pixels in the sensor, incorrect atmospheric compensation, and the presence of heavy cloud cover. Such pixels cannot be reliably used in the image fusion process as they may degrade the performance of the method. Directly addressing these effects using a statistical model would require the choice of a non-Gaussian distribution for the noise vector $\mathbf{r}_{k,\ell}^m$, which could make the computational complexity of the fusion procedure prohibitive. Thus, we consider a matrix $\mathbf{D}_k^m \in \mathbb{R}^{\tilde{N}_m \times N_m}$, which eliminates outlier pixels from the image, leading to the following transformed measurement model:

$$\tilde{\mathbf{y}}_{k,\ell}^m = \mathbf{D}_k^m \mathcal{H}_\ell^m(\mathbf{S}_k) \mathbf{c}_\ell^m + \tilde{\mathbf{r}}_{k,\ell}^m, \quad (2)$$

where $\tilde{\mathbf{y}}_{k,\ell}^m = \mathbf{D}_k^m \mathbf{y}_{k,\ell}^m$ and $\tilde{\mathbf{r}}_{k,\ell}^m = \mathbf{D}_k^m \mathbf{r}_{k,\ell}^m$ denotes the measured image band and the measurement noise in which the outlier values have been removed.

Using (2) and the properties of the vectorization operator, we can write this model equivalently as

$$\begin{aligned} \tilde{\mathbf{y}}_{k,\ell}^m &= [(\mathbf{c}_\ell^m)^\top \otimes \mathbf{D}_k^m] \text{vec}(\mathcal{H}_\ell^m(\mathbf{S}_k)) + \tilde{\mathbf{r}}_{k,\ell}^m \\ &= [(\mathbf{c}_\ell^m)^\top \otimes \mathbf{D}_k^m] \mathbf{H}_\ell^m \mathbf{s}_k + \tilde{\mathbf{r}}_{k,\ell}^m \end{aligned} \quad (3)$$

where $\otimes$ denotes the Kronecker product. The variable $\mathbf{s}_k \in \mathbb{R}^{L_H N_H}$ denotes a vector-ordering of the high-resolution image $\mathbf{S}_k$ which is obtained by grouping all pixels such that the bands of a single HR pixel are adjacent to each other, and the pixels that are contained within a single “lowest-resolution” pixel are also adjacent to each other, that is:

$$\mathbf{s}_k = \left[\begin{bmatrix} s_{k,1,\iota(1,1)} \\ \vdots \\ s_{k,L_H,\iota(1,1)} \\ s_{k,1,\iota(2,1)} \\ \vdots \\ s_{k,L_H,\iota(d,1)} \end{bmatrix}^\top, \dots, \begin{bmatrix} s_{k,1,\iota(1,N_{m',\ell'})} \\ \vdots \\ s_{k,L_H,\iota(1,N_{m',\ell'})} \\ s_{k,1,\iota(2,N_{m',\ell'})} \\ \vdots \\ s_{k,L_H,\iota(d,N_{m',\ell'})} \end{bmatrix}^\top \right]^\top, \quad (4)$$

where $s_{k,i,j}$ is the (i,j) -th position of $\mathbf{S}_k$, m' and ℓ' are the modality and spectral band with the lowest spatial resolution (i.e., for which $N_{m,\ell}$ is smallest), $d = N_H/N_{m',\ell'}$ is the number of HR pixels inside each low resolution pixel of band ℓ' and modality m' , and $\iota: \mathbb{N}_* \times \mathbb{N}_* \rightarrow \mathbb{N}_*$ is a function such that $\iota(i,j)$ returns the index (in $\mathbf{S}_k$) of the i -th HR pixel contained inside the j -th low resolution pixel (where $i \in \{1, \dots, d\}$) for modality m' and band ℓ' . $\mathbf{H}_\ell^m$ is a matrix form representation of the operator $\mathcal{H}_\ell^m$, such that $\text{vec}(\mathcal{H}_\ell^m(\mathbf{S}_k)) = \mathbf{H}_\ell^m \mathbf{s}_k$.

We can now represent all bands from each modality in the form of a single vector $\tilde{\mathbf{y}}_k^m \in \mathbb{R}^{\tilde{N}_m L_m}$ as

$$\tilde{\mathbf{y}}_k^m = \underbrace{\begin{pmatrix} [(\mathbf{c}_1^m)^\top \otimes \mathbf{D}_k^m] \mathbf{H}_1^m \\ \vdots \\ [(\mathbf{c}_{L_m}^m)^\top \otimes \mathbf{D}_k^m] \mathbf{H}_{L_m}^m \end{pmatrix}}_{\tilde{\mathbf{H}}_k^m} \mathbf{s}_k + \tilde{\mathbf{r}}_k^m, \quad (5)$$

where $\tilde{\mathbf{r}}_k^m \sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{R}}_k^m)$, and

$$\tilde{\mathbf{y}}_k^m = \text{col} \{ \tilde{\mathbf{y}}_{k,1}^m, \dots, \tilde{\mathbf{y}}_{k,L_m}^m \}, \quad (6)$$

$$\tilde{\mathbf{r}}_k^m = \text{col} \{ \tilde{\mathbf{r}}_{k,1}^m, \dots, \tilde{\mathbf{r}}_{k,L_m}^m \}, \quad (7)$$

$$\tilde{\mathbf{R}}_k^m = \text{blkdiag} \{ \mathbf{D}_k^m \mathbf{R}_1^m (\mathbf{D}_k^m)^\top, \dots, \mathbf{D}_k^m \mathbf{R}_{L_m}^m (\mathbf{D}_k^m)^\top \} \quad (8)$$

Note that at most time instants k , one or more of the modalities $m \in \Omega$ is not observed. In this case, we set the matrix $\mathbf{D}_k^m$ as an empty (zero-dimensional) matrix, which simplifies the problem and avoids introducing additional notation.

2.3. Dynamical evolution model

Defining reasonable dynamical models for image fusion requires detailed knowledge regarding the scene evolution over time, which is often unattainable. In this contribution, we aim at a complete data driven strategy assuming very little knowledge regarding the scene evolution except for past data coming from the imaging modalities being used. To match such lack of prior knowledge we consider a simple random-walk process to model the latent state dynamics as:

$$\mathbf{s}_{k+1} = \mathbf{F}_k \mathbf{s}_k + \mathbf{q}_k, \quad (9)$$

where $\mathbf{F}_k \in \mathbb{R}^{L_H N_H \times L_H N_H}$ is the state transition matrix, which is assumed to satisfy $\|\mathbf{F}_k\|_2 \leq 1$, and $\mathbf{q}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ with $\mathbf{Q}_k \in \mathbb{R}^{L_H N_H \times L_H N_H}$ being the state process noise covariance matrix. Note that the above model plays a crucial role in the estimation results, as it describes both the distribution of the changes occurring in the image at time k , as well as the marginal distribution of the states. This means that more sophisticated dynamics can be introduced in the problem through the appropriate design of the process noise covariance matrix $\mathbf{Q}_k$. Although expectation maximization (EM) can be used to estimate $\mathbf{Q}_k$ in time invariant models [49], the problem becomes extremely ill-posed in the time-varying setting. Another issue relates to the computational complexity of EM-based strategies requiring the solution of the Kalman filter and smoother systems multiple times, becoming unfeasible when dealing with large images. For these reasons, we propose an alternative route to estimate $\mathbf{Q}_k$.

2.4. A weakly supervised approach for estimating $\mathbf{Q}_k$

We consider $\mathbf{Q}_k(\mathcal{D}_k)$ as a function of the set $\mathcal{D}_k = \{\tilde{\mathbf{y}}_\ell^{m \in \Omega_H}\}_{\ell < k}$ of past high resolution images. The set $\mathcal{D}_k$ represents historical data and images currently being fused up the the time step k . Although many strategies could be leveraged to find suitable past time windows to account for more relevant covariance estimation and consider full covariance matrices, in this preliminary work we choose a simple route to validate this type of approach. For this, let $\mathbf{y}_{k-\tau}^{m \in \Omega_H}$ be the the most recently observed high resolution image¹. We compute $\mathbf{Q}_k$ by finding in our historical data the most similar image to $\mathbf{y}_{k-\tau}^{m \in \Omega_H}$ and then computing the pixelwise variance across the following $n \in \mathbb{N}_*$ images in our historical data. That is, we compute $\mathbf{Q}_k$ executing the following three steps for every time step k :

1. Identify the most similar state over $\mathcal{D}_k$, that is, the image that is most similar, according to a metric $\mathcal{L}$

$$\ell^* = \arg \min_{\ell \in \mathcal{I}_{\mathcal{D}_k}} \mathcal{L}(\mathbf{y}_{k-\tau}^{m \in \Omega_H}, [\mathcal{D}_k]_\ell), \quad (10)$$

with $[\mathcal{D}_k]_\ell$ being the ℓ -th image in the historical set $\mathcal{D}_k$, and $\mathcal{I}_{\mathcal{D}_k} \subseteq \mathbb{Z}$ is the set containing the time index of each image in $\mathcal{D}_k$.

2. select a time window $[\mathcal{D}_k]_{\ell^*:\ell^*+n}$.
3. compute the diagonal process noise covariance matrix, i.e., $\mathbf{Q}_k = \text{diag}\{q_{k,1}^2, \dots, q_{k,L_H N_H}^2\}$, as

$$q_{k,j}^2 = \max \left(\frac{\text{var}([\mathcal{D}_k]_{\ell^*:\ell^*+n}^{(j)})}{\Delta_{\mathcal{D}_k}^{\ell^*}}, \varepsilon^2 \right) \times \Delta_k, \quad (11)$$

where $[\mathcal{D}_k]_{\ell^*:\ell^*+n}^{(j)} = [\tilde{y}_{\ell^*,j}^{m \in \Omega_H}, \dots, \tilde{y}_{\ell^*+n,j}^{m \in \Omega_H}]$, $\varepsilon > 0$ is a small scalar allowing for changes on the scene that were unseen on the historical data window $[\mathcal{D}_k]_{\ell^*:\ell^*+n}$, Δ_k is the time interval (in days) between $\mathbf{y}_k^{m \in \Omega_H}$ and $\mathbf{y}_{k+1}^{m \in \Omega_H}$, and $\Delta_{\mathcal{D}_k}^{\ell^*}$ is the time interval (in days) between $[\mathcal{D}_k]_{\ell^*}$ and $[\mathcal{D}_k]_{\ell^*+n}$. As similarity metric we used the cosine similarity $\mathcal{L}(\mathbf{y}, \mathbf{z}) = \cos(\mathbf{y}, \mathbf{z})$.

3. Multimodal image fusion using a weakly supervised constrained Kalman filter

Considering models (5) and (9), the online multimodal image fusion problem can be formulated as the problem of computing the posterior distribution of the high resolution image given all previous measurements available, i.e.,

$$p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k}^m\}_{m \in \Omega}) = \mathcal{N}(\mathbf{s}_{k|k}, \mathbf{P}_{k|k}). \quad (12)$$

Due to the choice of a linear Gaussian model, this distribution is also Gaussian. Moreover, its mean vector $\mathbf{s}_{k|k}$ and covariance matrix $\mathbf{P}_{k|k}$ can be computed recursively using the standard Kalman filter with a prediction and update steps [50].

¹That is, $\tau \in \mathbb{Z}_+$ is the smallest integer such that a high resolution image was observed at time instant $k - \tau$.

More precisely, the prediction step of the Kalman filter computes the first and second order moments of $p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k-1}^m\}_{m \in \Omega})$ as:

$$\mathbf{s}_{k|k-1} = \mathbf{F}_{k-1} \mathbf{s}_{k-1|k-1} \quad (13)$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_{k-1} \mathbf{P}_{k-1|k-1} \mathbf{F}_{k-1}^\top + \mathbf{Q}_{k-1} \quad (14)$$

The update step computes then computes of (12). Note that the update step can be simplified and implemented separately for each data modality by using the Markov property of the model and the independence between noise vectors of different modalities:

$$\begin{aligned} p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k}^m\}_{m \in \Omega}) \\ \propto p(\{\tilde{\mathbf{y}}_k^m\}_{m \in \Omega} | \mathbf{s}_k) p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k-1}^m\}_{m \in \Omega}) \\ = p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k-1}^u\}_{u \in \Omega}) \prod_{m \in \Omega} p(\tilde{\mathbf{y}}_k^m | \mathbf{s}_k). \end{aligned} \quad (15)$$

By computing the first product in the right hand side as:

$$\begin{aligned} p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k-1}^u\}_{u \in \Omega}) p(\tilde{\mathbf{y}}_k^m | \mathbf{s}_k) \\ \propto p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k-1}^u\}_{u \in \Omega}, \tilde{\mathbf{y}}_k^m), \end{aligned} \quad (16)$$

which is an update step of the Kalman filter with image modality m to yield a new posterior in the r.h.s. of (16). This can be computed as:

$$\mathbf{v}_k^m = \tilde{\mathbf{y}}_k^m - \tilde{\mathbf{H}}_k^m \mathbf{s}_{k|k-1} \quad (17)$$

$$\mathbf{T}_k^m = \tilde{\mathbf{H}}_k^m \mathbf{P}_{k|k-1} (\tilde{\mathbf{H}}_k^m)^\top + \tilde{\mathbf{R}}_k^m \quad (18)$$

$$\mathbf{K}_k^m = \mathbf{P}_{k|k-1} (\tilde{\mathbf{H}}_k^m)^\top (\mathbf{T}_k^m)^{-1} \quad (19)$$

$$\mathbf{s}_{k|k} = \mathbf{s}_{k|k-1} + \mathbf{K}_k^m \mathbf{v}_k^m \quad (20)$$

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_k^m \mathbf{T}_k^m (\mathbf{K}_k^m)^\top \quad (21)$$

for $m \in \Omega$. By proceeding with the computation of the product in the r.h.s. of (15) recursively, the Kalman update can then be performed separately for each of the modalities observed at time instant k . Note that after the first modality is processed, the update equations above are used again for the subsequent modalities by setting $\mathbf{s}_{k+1|k}$ and $\mathbf{P}_{k+1|k}$ as equal to the posterior estimates from the previously processed modality.

3.1. The Linear Smoother

Given a window of K image samples, the Bayesian smoothing problem consists of computing the posterior distribution of the high resolution image given all available measurements available, i.e.,

$$p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:K}^m\}_{m \in \Omega}) = \mathcal{N}(\mathbf{s}_{k|K}, \mathbf{P}_{k|K}), \quad (22)$$

which is also a Gaussian. Just like in the filtering problem, the linear and Gaussian model allows this solution to be computed efficiently using the Rauch-Tung-Striebel (RTS) smoothing equations [50], which consist of a forward pass of the Kalman filter (as described before), followed by a backwards recursion that updates the previously computed mean and covariances matrices of the state with information from future time instants.

We note that the smoothing can also be performed efficiently for the case when multiple image modalities are available. Let us consider the Bayesian smoothing equations as defined in [51, 50], which is performed in two steps. Starting from the Kalman state estimate at time K , given by $p(\mathbf{s}_K | \{\tilde{\mathbf{y}}_{1:K}^m\}_{m \in \Omega})$, the smoothing distribution is computed recursively for $k = K - 1, \dots, 1$, according to the following relation:

$$p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:K}^m\}_{m \in \Omega}) = p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k}^m\}_{m \in \Omega}) \times \int \frac{p(\mathbf{s}_{k+1} | \mathbf{s}_k) p(\mathbf{s}_{k+1} | \{\tilde{\mathbf{y}}_{1:K}^m\}_{m \in \Omega})}{p(\mathbf{s}_{k+1} | \{\tilde{\mathbf{y}}_{1:k}^m\}_{m \in \Omega})} d\mathbf{s}_{k+1}, \quad (23)$$

where $p(\mathbf{s}_k | \{\tilde{\mathbf{y}}_{1:k}^m\}_{m \in \Omega}) = \mathcal{N}(\mathbf{s}_{k|k}, \mathbf{P}_{k|k})$ is the Kalman estimate of the state PDF at time k , $p(\mathbf{s}_{k+1} | \mathbf{s}_k)$ is the state transition PDF, computed according to (9), $p(\mathbf{s}_{k+1} | \{\tilde{\mathbf{y}}_{1:K}^m\}_{m \in \Omega}) = \mathcal{N}(\mathbf{s}_{k+1|K}, \mathbf{P}_{k+1|K})$ is the smoothing distribution obtained at the previous iteration, and $p(\mathbf{s}_{k+1} | \{\tilde{\mathbf{y}}_{1:k}^m\}_{m \in \Omega})$ is the predictive state distribution, which is computed exactly as in the prediction step of the Kalman filter.

In the linear and Gaussian case this translates into the following closed form solution [50], with

$$\mathbf{s}_{k+1|k} = \mathbf{F}_k \mathbf{s}_{k|k} \quad (24)$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}_k \mathbf{P}_{k|k} \mathbf{F}_k^\top + \mathbf{Q}_k \quad (25)$$

being used to compute the predictive state distribution, and

$$\mathbf{G}_k = \mathbf{P}_{k|k} \mathbf{F}_k^\top \mathbf{P}_{k+1|k}^{-1} \quad (26)$$

$$\mathbf{s}_{k|K} = \mathbf{s}_{k|k} + \mathbf{G}_k (\mathbf{s}_{k+1|K} - \mathbf{s}_{k+1|k}) \quad (27)$$

$$\mathbf{P}_{k|K} = \mathbf{P}_k + \mathbf{G}_k (\mathbf{P}_{k+1|K} - \mathbf{P}_{k+1|k}) \mathbf{G}_k^\top \quad (28)$$

to update the covariances. It should be noted that the mean and covariance $\mathbf{s}_{k|k}$ and $\mathbf{P}_{k|k}$ used in the Smoothing equations are the final result obtained from the Kalman update after processing all image modalities that were available at instant k .

Thus, while in the Kalman filtering the update equations must be computed sequentially at each time step w.r.t. the different image modalities, smoothing only needs only the final state estimates at each instant, no matter how many modalities are present.

3.2. Constraining the estimates

Although the Kalman filter provides closed-form solutions to the estimation of the high-resolution image sequence, it relies on a Gaussian assumption on the states and observations

which does not correspond to the physics of the problem. In fact, represented in reflectance values, each pixel and band of a high-resolution images $\mathbf{s}_k$ is actually constrained to an interval $s_{k,i,j} \in [0, s_{\max}]$, where $s_{\max}$ is the maximum reflectance values of the scene. Since this information can potentially improve the accuracy of the estimated states, we propose to incorporate this information by considering the linearly constrained Kalman filter [52], in which the final constrained state $\mathbf{s}_{k|k}^+$ is obtained as the solution to a constrained optimization problem:

$$\begin{aligned} \mathbf{s}_{k|k}^+ = \arg \min_{\mathbf{s}} \quad & (\mathbf{s} - \mathbf{s}_{k|k})^\top \mathbf{P}_{k|k}^{-1} (\mathbf{s} - \mathbf{s}_{k|k}) \\ \text{subject to } & \mathbf{s} \in [0, s_{\max}]^{N_H L_H} \end{aligned} \quad (29)$$

Problem (29) consists in a constrained quadratic program, which can be costly to solve due to the high dimensionality of the variables. Thus, we propose a simple solution consisting of truncating the result of the traditional Kalman update:

$$\mathbf{s}_{k|k}^+ = \max(\min(\mathbf{s}_{k|k}, s_{\max}), 0), \quad (30)$$

where functions $\max(\cdot, \cdot)$ and $\min(\cdot, \cdot)$ compute the elementwise maximum and minimum value between a vector and a scalar. Note that this truncation provides the exact solution when $\mathbf{P}_{k|k}$ is diagonal. The same truncation strategy was also applied to the results of the linear smoother $\mathbf{s}_{k|K}$. We generally observed that this gave good results in practice. $s_{\max}$ can be estimated as the maximum value of the observed images in a time window, or from the historical data.

4. A distributed implementation

A problem with the Kalman filter is the need to compute and store the state covariance matrix, $\mathbf{P}_{k|k}$. This incurs in storage and operations asymptotic complexity in the order of $\mathcal{O}(N_H^2 L_H^2)$ and $\mathcal{O}(N_H^3 L_H^3)$, respectively. This can make the method intractable for images with a large number of pixels. Thus, to reduce the complexity of the filter and of the smoother, we consider splitting the pixels in the estimated state $\mathbf{s}_k$ into multiple groups which are assumed to be statistically independent [53, 54, 55]. To this end, we divide the state space into G groups as:

$$\mathbf{s}_k = \text{vec}([\mathbf{s}_k^{(1)}, \dots, \mathbf{s}_k^{(G)}]), \quad (31)$$

where the variables within each block $\mathbf{s}_k^{(g)}$ are correlated, but different blocks $\mathbf{s}_k^{(g_1)}$ and $\mathbf{s}_k^{(g_2)}$ are assumed to be independent for $g_1 \neq g_2$. This leads to the following approximation for the predictive and posterior covariance matrices $\mathbf{P}_{k|k-1}$ and $\mathbf{P}_{k|k}$ as block diagonal matrices:

$$\mathbf{P}_{k|k-1} = \text{blkdiag}\{\mathbf{P}_{k|k-1}^{(1)}, \dots, \mathbf{P}_{k|k-1}^{(G)}\} \quad (32)$$

$$\mathbf{P}_{k|k} = \text{blkdiag}\{\mathbf{P}_{k|k}^{(1)}, \dots, \mathbf{P}_{k|k}^{(G)}\} \quad (33)$$

We consider different splitting possibilities, with different trade-offs between approximation accuracy with respect to the full-state-covariance Kalman filter and complexity:

- i) A fully diagonal model (with $G = N_H L_H$ blocks).
- ii) A block diagonal model where each block consists of all bands of one single high-resolution pixel (with $G = N_H$ blocks).
- iii) A block diagonal model, with blocks corresponding to the high-resolution pixels which reside inside a single MODIS pixel (with $G = N_H L_H / N_{\text{MODIS}}$ blocks).

Following [54], the Kalman equations for the prediction step (13)–(14) can be written for each block as:

$$\mathbf{s}_{k+1|k}^{(g)} = [\mathbf{F}_k]_{(g),:} \mathbf{s}_k \quad (34)$$

$$\mathbf{P}_{k+1|k}^{(g)} = [\mathbf{F}_k]_{(g),:} \mathbf{P}_k ([\mathbf{F}_k]_{(g),:})^\top + \mathbf{Q}_k^{(g)} \quad (35)$$

where $[\mathbf{F}_k]_{(g),:}$ means the matrix formed by taking from $\mathbf{F}_k$ the rows which correspond to the indices in the group of states g , and all columns. Matrices $\mathbf{Q}_k^{(g)}$ are defined as:

$$\mathbf{Q}_k = \text{blkdiag} \{ \mathbf{Q}_k^{(1)}, \dots, \mathbf{Q}_k^{(G)} \}. \quad (36)$$

Similarly, the Kalman update equations (17)–(21) are performed separately for each block of variables, and are given by:

$$\mathbf{s}_k^{(g)} = \mathbf{s}_{k|k-1}^{(g)} + \mathbf{K}_k^{(g)} \mathbf{v}_k^m \quad (37)$$

$$\mathbf{P}_k^{(g)} = \mathbf{P}_{k|k-1}^{(g)} - \mathbf{K}_k^{(g)} \mathbf{T}_k^m (\mathbf{K}_k^{(g)})^\top \quad (38)$$

with:

$$\mathbf{K}_k^{(g)} = \Sigma_{xy,k|k-1}^{(g)} (\mathbf{T}_k^m)^{-1} \quad (39)$$

$$\mathbf{v}_k^m = \tilde{\mathbf{y}}_k^m - \tilde{\mathbf{H}}_k^m \mathbf{s}_{k|k-1} \quad (40)$$

$$\mathbf{T}_k^m = \tilde{\mathbf{H}}_k^m \mathbf{P}_{k|k-1} (\tilde{\mathbf{H}}_k^m)^\top + \tilde{\mathbf{R}}_k^m \quad (41)$$

$$\begin{aligned} \Sigma_{xy,k|k-1}^{(g)} &= [\mathbf{P}_{k|k-1} (\tilde{\mathbf{H}}_k^m)^\top]_{(g),:} \\ &= [\mathbf{P}_{k|k-1}]_{(g),:} (\tilde{\mathbf{H}}_k^m)^\top \end{aligned} \quad (42)$$

where $[\mathbf{P}_{k|k-1}]_{(g),:}$ means the matrix formed by taking from $\mathbf{P}_{k|k-1}$ the rows which correspond to the indices in the group of states g , and all columns. Note that the block diagonal structure of $\mathbf{P}_{k|k-1}$ and $\mathbf{P}_{k|k}$ can be explored to perform the above operations efficiently, since these matrices are very sparse.

Following the same approach, the linear smoother can also be approximated in blockwise fashion as in [55], for the predictive equations (24)–(25):

$$\mathbf{s}_{k+1|k}^{(g)} = [\mathbf{F}_k]_{(g),:} \mathbf{s}_k \quad (43)$$

$$\mathbf{P}_{k+1|k}^{(g)} = [\mathbf{F}_k]_{(g),:} \mathbf{P}_k ([\mathbf{F}_k]_{(g),:})^\top + \mathbf{Q}_k^{(g)} \quad (44)$$

and for the smoothing equations (26)–(28):

$$\begin{aligned} \mathbf{G}_k^{(g)} &= [\mathbf{P}_k \mathbf{F}_k^\top]_{(g),(g)} (\mathbf{P}_{k+1|k}^{(g)})^{-1} \\ &= [\mathbf{P}_k]_{(g),(g)} ([\mathbf{F}_k]_{(g),(g)})^\top (\mathbf{P}_{k+1|k}^{(g)})^{-1} \end{aligned} \quad (45)$$

$$\mathbf{s}_{k|K}^{(g)} = \mathbf{s}_k^{(g)} + \mathbf{G}_k^{(g)} (\mathbf{s}_{k+1|K}^{(g)} - \mathbf{s}_{k+1|k}^{(g)}) \quad (46)$$

$$\mathbf{P}_{k|K}^{(g)} = \mathbf{P}_k^{(g)} + \mathbf{G}_k^{(g)} (\mathbf{P}_{k+1|K}^{(g)} - \mathbf{P}_{k+1|k}^{(g)}) (\mathbf{G}_k^{(g)})^\top \quad (47)$$

where

$$\mathbf{G}_k = \text{blkdiag} \{ \mathbf{G}_k^{(1)}, \dots, \mathbf{G}_k^{(G)} \}. \quad (48)$$

One last issue is that the innovation covariance matrix $\mathbf{T}_k^m$ can also be large for big images (e.g., Landsat measurements), as it has $(L_m \prod_{\ell=1}^{L_m} N_{m,\ell})^2$ elements. Fortunately, the model implicitly imposes a simple structure for this matrix. To show this, let us consider a permutation of the pixels $\mathbf{\Pi}_m$, such that $\mathbf{\Pi}_m \tilde{\mathbf{y}}_k^m$ reorders $\tilde{\mathbf{y}}_k^m$ by making different bands of each LR pixel contiguous:

$$\mathbf{\Pi}_m \tilde{\mathbf{y}}_k^m = \left[\begin{bmatrix} \tilde{y}_{k,1,1}^m \\ \vdots \\ \tilde{y}_{k,L_m,1}^m \end{bmatrix}^\top, \dots, \begin{bmatrix} \tilde{y}_{k,1,N_m}^m \\ \vdots \\ \tilde{y}_{k,L_m,N_m}^m \end{bmatrix}^\top \right]^\top, \quad (49)$$

where $\tilde{y}_{k,\ell,n}^m$ is the n -th pixel of the ℓ -th band of $\tilde{\mathbf{y}}_k$.

If we assume that $\mathbf{H}_\ell^m$ is a local filter, i.e., each pixel in the low-resolution image is generated according to a fixed linear combination of a *distinct* subset of HR pixels, this allows us to express the row-permuted version of $\tilde{\mathbf{H}}_k^m$ equivalently as:

$$\mathbf{\Pi}_m \tilde{\mathbf{H}}_k^m = \text{blkdiag} \{ \underbrace{\mathbf{H}, \mathbf{H}, \dots, \mathbf{H}}_{N_m \text{ times}} \}, \quad (50)$$

where matrix $\mathbf{H} \in \mathbb{R}^{L_m \times d^2 L_H}$ is given by:

$$\mathbf{H} = \mathbf{h}^m \otimes \mathbf{C}^m, \quad (51)$$

where $\mathbf{C}^m = [(\mathbf{c}_1^m)^\top, \dots, (\mathbf{c}_{L_m}^m)^\top]^\top$ is the spectral response function for all bands, $\mathbf{h}^m \in \mathbb{R}^{1 \times d}$ is the local spatial response filter, which defined how the HR pixels inside each LR pixels are combined, and d is the number of HR pixel in each LR pixel.

Using this permutation, the innovation covariance matrix can be written as:

$$\begin{aligned} \mathbf{\Pi}_m \mathbf{T}_k^m \mathbf{\Pi}_m^\top &= \mathbf{\Pi}_m \tilde{\mathbf{H}}_k^m \mathbf{P}_{k|k-1} (\tilde{\mathbf{H}}_k^m)^\top \mathbf{\Pi}_m^\top + \mathbf{\Pi}_m \tilde{\mathbf{R}}_k^m \mathbf{\Pi}_m^\top \\ &= \text{blkdiag} \{ \mathbf{H}, \dots, \mathbf{H} \} \mathbf{P}_{k|k-1} \text{blkdiag} \{ \mathbf{H}^\top, \dots, \mathbf{H}^\top \} \\ &\quad + \mathbf{\Pi}_m \tilde{\mathbf{R}}_k^m \mathbf{\Pi}_m^\top \\ &= \text{blkdiag} \{ \mathbf{H} \mathbf{P}_{k|k-1}^{(1)} \mathbf{H}^\top, \dots, \mathbf{H} \mathbf{P}_{k|k-1}^{(G)} \mathbf{H}^\top \} \\ &\quad + \mathbf{\Pi}_m \tilde{\mathbf{R}}_k^m \mathbf{\Pi}_m^\top. \end{aligned} \quad (52)$$

Algorithm 1: Weakly supervised online image fusion

Input : Measured multimodal images $\mathbf{y}_k^m$, for all time instants $k = 1, \dots, K$ and modalities m , historical datasets of high-resolution images $\mathcal{D}_k$, parameters $s_{\max}$.

Output: Estimated image sequence $\mathbf{s}_{k|K}$

```
1 Initialize  $\mathbf{P}_{0|0}$  and  $\mathbf{s}_{0|0}$ ;
2 // Filter ;
3 for  $k = 1, 2, \dots, K$  do
4   Compute innovation covariance matrix  $\mathbf{Q}_k$  using  $\mathbf{s}_{k-1}$  and  $\mathcal{D}_k$  according to Section 2.4 ;
5   Compute  $\mathbf{s}_{k|k-1}$  and  $\mathbf{P}_{k|k-1}$  using equations (31), (32), (34) and (35) ; // Prediction
6   Compute  $\mathbf{s}_{k|k}$  and  $\mathbf{P}_{k|k}$  using equation (33) and equations (37)–(42) ; // Update
7   Constrain  $\mathbf{s}_{k|k}$  using (30) ;
8 end
9 // Smoother ;
10 for  $k = K, K-1, \dots, 1$  do
11   Compute  $\mathbf{s}_{k+1|k}$  and  $\mathbf{P}_{k+1|k}$  using equations (31), (32), (43) and (44) ; // Prediction
12   Compute  $\mathbf{s}_{k|K}$  and  $\mathbf{P}_{k|K}$  using equations (45)–(47) and equation (48) ; // Backwards
    update
13 end
14 return Estimated images  $\mathbf{s}_{k|K}$ 
```

Thus, as long as the noise is independent among different pixels (i.e., $\tilde{\mathbf{R}}_k^m$ is block diagonal), it is possible to express the innovation covariance matrix in block diagonal form by adequately permuting the LR image pixels. This shows that each pixel from the lowest resolution image modality can be processed independently when $\mathbf{Q}_k$ and $\mathbf{P}_{0|0}$ also have a block diagonal structure. The proposed image fusion method is summarized in Algorithm 1.

5. Experiments

In this section, we use the proposed methodology to fuse Landsat and MODIS image over time. The Kalman filter and smoother are built under the three different assumptions for the state covariance matrices regarding the distributed implementation discussed in Section 4: *i*) diagonal state covariance (denoted by KF-D and SM-D); *ii*) block-diagonal state covariance with one block per Landsat multispectral pixel (denoted by KF-B and SM-B); and *iii*) block-diagonal with blocks for all Landsat multispectral pixels corresponding to the same coarse pixel in a MODIS image being correlated (denoted by KF-F and SM-F). A filter in which Landsat multispectral pixels corresponding to more than one coarse pixel in a MODIS image being all correlated could not be implemented due to computational and memory limitations.

Although in our experiments we consider only two modalities the proposed methodology admits multiple different modalities provided that enough computational power is available. As benchmark, we compare the performance of Kalman filter and smoother under all three assumptions to that of the *Enhanced Spatial and Temporal Adaptive ReFlectancefusion Model* (ESTARFM) algorithm [25], and the *Prediction Smooth Reflectance Fusion Model*

(PSRFM) algorithm [56, 57]. The ESTARFM algorithm requires two high-resolution (e.g., Landsat) images at the beginning of the image sequence, and can generate high-resolution reconstructions at later time instants based on MODIS measurements. Thus, it is a good candidate for comparison with the Kalman filtering based strategies, which also do not require future data. The PSRFM method, on the other hand, uses two high-resolution (e.g., Landsat) images (one at the beginning and one at the end of the sequence), and provides high-resolution reconstruction for the intermediate MODIS images. Thus, it consists in an adequate comparison to the smoother algorithms, which also require future high-resolution images. In the following, we describe the data and simulation setup, followed by the results and the discussions.

5.1. Study region

For the experiments, we consider two sites. The first is the Oroville dam (Figure 2, left panel), located on the Feather River, in the Sierra Nevada Foothills ($38^{\circ} 35.3'$ North and $122^{\circ} 27.8'$ W) is the tallest dam in USA and is major water storage facility in California State Water Project. The reservoir has a maximum storage capacity of 1.54×10^{11} ft³ or 4.36×10^9 m³, which fills during heavy rains or large spring snow melts and water is carefully released to prevent flooding in downstream areas, mainly to prevent large flooding in Butte County and area along the Feather River. The reservoir water storage change in between 07/03 and 09/21 of 2018 is as shown as the hydrograph curve in Figure 8. Another unique characteristic is that it has three power plants at this reservoir. The water released downstream is used to maintain the Feather and Sacramento Rivers and the San Francisco-San Joaquin delta. Lake Oroville is at an elevation of 935 feet (285 meters) above sea level. We focus at a particular area of the Oroville dam delimited by the red box in Figure 2.

The second site is the Elephant Butte reservoir (Figure 2, right panel), located in the southern part of the Rio Grande river, in New Mexico, USA ($33^{\circ} 19.4'$ N and $107^{\circ} 26.2'$ W). It is the largest reservoir in New Mexico, providing power and irrigation to southern New Mexico and Texas. Elephant Butte reservoir is at an elevation of 4,414 ft (1,345 meters), and has a surface area of 36,500 acres (14,800 ha).

Table 1: Spectral angle mapper between the estimated high-resolution image and the Landsat measurement for the Oroville Dam example (note that the Landsat images at dates 07/19, 08/20, and 09/05 were not supplied to the algorithms and only used for evaluation purposes). However, the Landsat image at 09/21 was available to all algorithms. Note that the spectral angle is not reported for PSRFM at 09/21. This is so since PSRFM uses the last pair (MODIS-Landsat) of images and directly sets its estimations at this dates to the ground-truth.

Method	KF-F	SM-F	KF-B	SM-B	KF-D	SM-D	ESTARFM	PSRFM
Image (07/19)	7.1240	10.8537	4.2356	6.1515	4.9304	5.9064	6.0810	6.8837
Image (08/20)	27.6343	26.2786	26.1229	25.1520	27.1928	26.1758	29.0892	27.7802
Image (09/05)	8.5741	6.0366	6.6246	3.6838	7.4482	4.4135	11.4553	6.0354
Image (09/21)	8.0588	3.6385	6.4042	0.5471	6.9754	0.6960	11.9584	—
Average	12.8478	11.7019	10.8468	8.8836	11.6367	9.2979	14.6460	10.1748

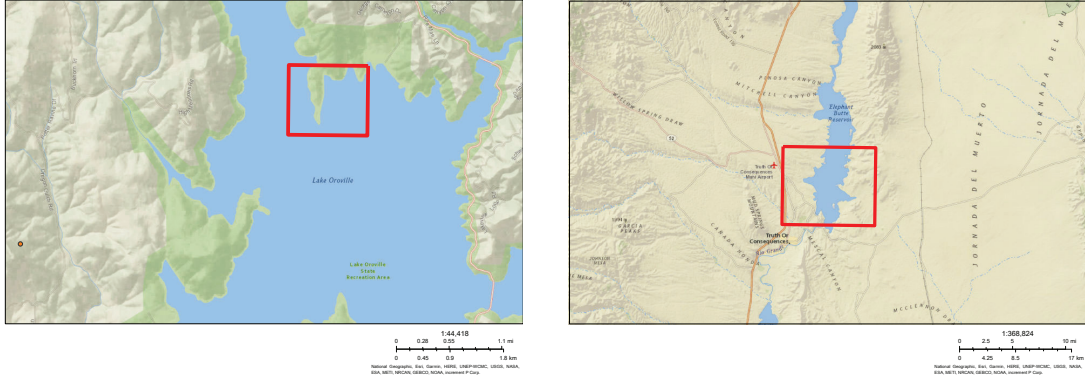


Figure 2: (Left) Oroville dam site. (Right) Elephant Butte site. The red boxes delimit the specific study areas used in our experiments.

Table 2: Percentage of misclassified pixels for the Oroville Dam example (the Landsat image at 09/21 was available to all algorithms). Note that the misclassification percentage is not reported for PSRFM at 09/21. This is so since PSRFM uses the last pair (MODIS-Landsat) of images and directly sets its estimations at this dates to the ground-truth.

Method	KF-F	SM-F	KF-B	SM-B	KF-D	SM-D	ESTARFM	PSRFM
Image (07/19)	9.5412	7.6360	6.4472	8.2914	6.1119	8.0171	5.4870	5.2431
Image (08/20)	14.9215	10.4405	7.8647	4.1000	7.2245	3.7799	18.2899	17.9851
Image (09/05)	13.4888	8.2152	9.6632	4.7859	9.4345	4.5877	22.7404	20.8962
Image (09/21)	11.7360	3.8409	9.3583	0.2439	9.2974	0.2591	26.3374	—
Average	12.4219	7.5332	8.3333	4.3553	8.0171	4.1610	18.2137	11.0311

5.2. Remote Sensed data

For our simulations with the Oroville Dam site, we collected MODIS and Landsat data acquired from the region marked with a red square on Figure 2, and on a interval ranging from 2018/07/03 to 2018/09/21. This interval was selected since the hydrograph analysis indicates high variation in the water level of the reservoir, see, the hydrograph curve in Figure 8. Such variation in the water levels result in large changes in the acquired images, exposing flooded areas. In this experiment we will focus on the red and near-infrared (NIR) bands since they are often used to distinguish water from other landcover elements in the image [58]. We also collected 5 Landsat data from 2017/08/01 to 2017/12/07 to serve as a past historical dataset $\mathcal{D}_k$.

The study region marked in the left panel of Figure 2 corresponds to Landsat and MODIS images with 81×81 and 9×9 pixels, respectively². After filtering for heavy cloud cover during the designated time periods, a set of 6 Landsat and 16 MODIS images were obtained. We used the first MODIS and Landsat images for initialization of all methods leading to 5 and 15 images used in the remaining fusion process.

²The Landsat images were also upsampled to a spatial resolution of 27.77 meters to make its resolution exactly 9 times that of MODIS.

From the set of 5 Landsat images of the Oroville Dam site that were available for testing, three of them were set aside and not processed by any of the the algorithms. These images were acquired at dates 07/19, 08/20 and 09/05, when MODIS observations were also available, and will be used in the form of a reference for the evaluation of the algorithms' capability of estimating the high resolution images at these dates solely from the low resolution MODIS measurements.

For the simulations with the Elephant Butte site, shown in the right panel of Figure 2, we aim to evaluate the performance of the algorithms when processing a larger geographical area, with an area of approximately $9km \times 9km$. The setup is similar to the Oroville Dam example. We focus on the red and near-infrared bands of the Landsat and MODIS instruments, and collect 47 Landsat images from 2014/01/16 to 2017/11/24 to serve as the past historical dataset $\mathcal{D}_k$.

The study region corresponds to Landsat and MODIS images with 324×324 and 36×36 pixels, respectively. After removing images with significant cloud cover, we obtained a set of 5 Landsat and 7 MODIS images to process. We used the first MODIS and Landsat image pair to initialize the algorithms, leading to 4 Landsat and 6 MODIS images to be used in the remaining fusion process. From the set of 4 Landsat images that were available for testing, 2 of them were set aside as ground truth to evaluate the algorithms. Theses images are acquired at dates 06/07 and 06/23. However, the MODIS measurements at those dates contained significant cloud cover, and had to be discarded. Therefore, we evaluate the performance of the algorithms through the estimation results obtained dates 06/14 and 06/27 (in which the MODIS observations were available).

5.3. Algorithm setup

We initialized the proposed Kalman filter and smoother using a high resolution Landsat observation as the state, i.e., $\mathbf{s}_{0|0} = \tilde{\mathbf{y}}_0^L$, and set $\mathbf{P}_{0|0} = 10^{-10}\mathbf{P}_0$. The structure of $\mathbf{P}_0$ varies with different assumptions: *i)* $\mathbf{P}_0 = \mathbf{I}$ if the state covariance is diagonal; *ii)* $\mathbf{P}_0 = \text{blkdiag}\{\mathbf{P}_{0,1}, \mathbf{P}_{0,2}, \dots, \mathbf{P}_{0,N_H}\}$, where $\mathbf{P}_{0,i} = \frac{1}{2}\mathbb{1} + \frac{1}{2}\mathbf{I}$, with $\mathbb{1}$ being an all ones matrix, if the state covariance matrix has a block-diagonal structure with one block per Landsat multispectral pixel; *iii)* $\mathbf{P}_0 = \text{blkdiag}\{\mathbf{P}_{0,1}, \mathbf{P}_{0,2}, \dots, \mathbf{P}_{0,\tilde{N}_m \times L_m}\}$, where $\mathbf{P}_{0,i} = \frac{1}{2}\mathbb{1} + \frac{1}{2}\mathbf{I}$ if the state covariance matrix has a block-diagonal structure with each block containing all Landsat multispectral pixels corresponding to the same coarse pixel in a MODIS image. Figure 7 shows an example of the final $\mathbf{P}_{k|k}$, $k = 13$, obtained with the KF under all the assumptions discussed in Section 4. The noise covariance matrices were set as $\mathbf{R}_\ell^L = 10^{-10}\mathbf{I}$ and $\mathbf{R}_\ell^M = 10^{-4}\mathbf{I}$, for all ℓ . The blurring and downsampling matrices were set as $\mathbf{H}_\ell^L = \mathbf{I}$ for Landsat, while for MODIS $\mathbf{H}_\ell^M$ consisted of a convolution by an uniform 9×9 filter, defined by $\mathbf{h} = \frac{1}{81}\mathbb{1}_{9 \times 9}$ (where $\mathbb{1}_{9 \times 9}$ is a 9×9 matrix of ones), followed by decimation by a factor of 9, which represents the degradation occurring at the sensor (see, e.g., [44]). We also set $\mathbf{F}_k = \mathbf{I}$ for all k . The vectors $\mathbf{c}_\ell^m$ contained a positive gain in the ℓ -th position which compensated for scaling differences between Landsat and MODIS sensors, and zeros elsewhere.

The matrices $\mathbf{D}_k^M$ were constructed based on the quality codes (i.e., the QA bits) released

by MODIS for each image pixel [59]. QA bits provides information regarding pixel quality and cloud cover for all pixels and all bands. In our experiments we dropped any pixel not classified as *corrected product produced at ideal quality* in the QA bits [59] by adding zeros at corresponding positions in $\mathbf{D}_k^M$. Matrices $\mathbf{Q}_k$ were computed following our data-driven strategy described in Section 2.4 where $\varepsilon^2 = 10^{-5}$ and $n = 1$.

The ESTARFM algorithm was parametrized as follows [25], $w = 14$ as half of the window size, the number of classes was set to 4, and the pixels range was set to $[0, 0.5]$. The PSRFM algorithm was parametrized as follows, `CLUSTER_METHOD` = KMEAN, and `CLUSTER_DATA` = `fine + coarse`. We highlight that all methods have access only to the first (07/03) and last (09/21) Landsat images, which allows the algorithms to produce estimates for the MODIS images observed from the second (07/09, $k = 2$) up to the last date (09/21, $k = 16$). However, PSRFM uses the the last pair (MODIS-Landsat) during its inference process. For this reason, error metrics computed for PSRFM on (09/21) should be disregarded as the estimate is directly the ground-truth (i.e., the Landsat image) and, thus, are not reported in the experimental results.

All algorithms are evaluated using three metrics, which are computed taking as reference the Landsat images, three of which are not observed by the algorithms. The first metric is the Spectral Angle Mapper (SAM), which attempts to measure the estimation accuracy directly:

$$\text{SAM}(\mathbf{S}, \hat{\mathbf{S}}) = \frac{1}{N_H} \sum_{r=1}^{N_H} \arccos \left(\frac{\mathbf{s}_r^\top \hat{\mathbf{s}}_r}{\|\mathbf{s}_r\| \|\hat{\mathbf{s}}_r\|} \right), \quad (53)$$

where $\mathbf{S}$ and $\hat{\mathbf{S}}$ denote the true and the estimated images, respectively. $\mathbf{s}_r$ and $\hat{\mathbf{s}}_r$ denote the r -th pixels of different bands in $\mathbf{S}$ and $\hat{\mathbf{S}}$, respectively. The two remaining metrics are related to downstream tasks of water classification and water level monitoring, which are performed on the reconstructed image sequence.

We evaluate the direct benefit of the different fusion strategies in classifying water pixels from the estimated images. To classify water pixels we resorted to a KNN classifier whose centroids of water and non-water 2-band pixels were computed using K-Means algorithm. Finally, we evaluate the performance of the algorithms for hydrograph estimation by plotting the proportion of pixels in the image classified as water over time against the true hydrograph for the period, for all algorithms.

5.4. Results for the Oroville Dam site

As discussed, we fused the red and NIR reflectance bands of MODIS and Landsat for the selected study region. In Figure 3, we show the fused red (Figure 3a) and NIR (Figure 3b) reflectances as well as the acquired red and NIR reflectance values from MODIS and Landsat. Acquisition dates are displayed in the top labels at each column with a character, M for MODIS and L for Landsat, indicating the image used in the fusion algorithms. We recall that only the first and last Landsat images were used in the fusion process, keeping the remaining three images as ground-truth for evaluation purposes. Analyzing the

results we can see that the images estimated by the proposed Kalman filter and smoother methods, under different assumptions, produce better visual similarity with the Landsat (ground-truth) images for both bands. For instance, the increase in the island and the expansion of other land parts are clearly visible for the proposed methods. In contrast, analyzing ESTARFM results we note that land parts remain mainly constant through time until a new Landsat image is observed. Although lighter areas on the water portions can be noticed, specially for $k > 8$, its distribution does not resemble the ground-truth. This is expected since ESTARFM is not designed to acknowledge prior information or historical data. PSRFM results show an improvement compared with ESTARFM results, since it uses both the first and the last Landsat images. However, the PSRFM results does not resemble the ground-truth very closely, and significant blurring occurs around the edge of the island when $k < 8$. The blurring results in PSRFM are caused by the fact that the reconstructions provided by this algorithm are based on a form of interpolation which does not consider any information about the transition of the pixel reflectance values, whereas in our proposed methods we use the historical data to calibrate the time-varying dynamical model by means of matrix $\mathbf{Q}_k$, which can increase the accuracy of the estimations.

Note that the images estimated by KF-F and SM-F (which used a full state covariance matrix) contained more artifacts when compared to the ones obtained by KF-B, SM-B, KF-D and SM-D (which constrained the state covariance matrix to be diagonal or block diagonal). This occurs due to the high-dimensionality of the state vector (i.e., equivalent to a vectorized Landsat image) when compared to the MODIS measurements, as this leads to the amount of measurements not being sufficient to provide an accurate estimate of the full state vector and its covariance matrix, as shown in [60]. Thus, the extra degrees of freedom of KF-F and SM-F end up impacting their performance negatively. By setting the covariance matrix of the Kalman filter and smoother to be block diagonal or fully diagonal, the amount of parameters to be estimated is greatly reduced in KF-B, SM-B, KF-D and SM-D, leading to better results.

The results discussed above are corroborated by the absolute error maps displayed in Figure 4, and SAM results shown in Table 1 for dates in which ground-truth is available. Analyzing Figure 4 we highlight that SM-B and SM-D clearly present the smallest errors (i.e., overall darker pixels) for both bands and all dates. KF-B also presents low absolute error except for contour regions. PSRFM is the third overall darker image, followed by KF-B, KF-D, SM-F, KF-F and ESTARFM with exception of the results on 07/19 (first column), where ESTARFM is close to the ground-truth. Similar conclusions can be achieved by analyzing Table 1. The difference between the images estimated by the Kalman filter and smoother under the different approximations for the state covariance matrices (which are discussed in Section 4 and illustrated for this example in Figure 7) is shown in Figure 5. It can be seen that the approximations had a more pronounced effect on the Kalman filter compared to the smoother. Moreover, the differences between the filter with a diagonal (assumption *i*) and block diagonal state covariance with one block per Landsat pixel (assumption *ii*) was relatively small. Taking in to consideration the quantitative metrics in Table 1, this indicates that using a diagonal or block diagonal assumption on the state

covariance matrix with small blocks has a positive effect on the estimation performance, which likely occurs since it drastically reduces the amount of unknowns in the model that have to be estimated by the methods.

The left panel in Figure 6 presents the water maps for the ground-truth (first row) and all studied algorithms obtained using K-means clustering, while the right panel in Figure 6 shows the misclassification maps (i.e., the absolute error between the water maps obtained by each algorithm and the ground-truth). When comparing the resulting classification maps and the misclassification error with the ground-truth, the proposed methods present classification maps that are semantically better than the competing methods. This conclusion is also reached by considering the quantitative misclassification results presented in Table 2, in which the Kalman filter- and smoother-based methods led to smaller misclassification rates for all images except the ones on 07/19 and 09/21. A closer analysis reveals that the SM-D and SM-B methods hold the first and second best performance on average, followed by SM-F, KF-D, KF-B, PSRFM, KF-F and ESTARFM. Note that the PSRFM method requires access to the ground-truth (Landsat image) on 09/21 in order to produce an estimation for the MODIS image observed in this same date (i.e., measurement $k = 16$), which is why the corresponding misclassification percentage is not reported. We also remark that KF-D and KF-B also obtained competitive misclassification performance (i.e., better than PSRFM), despite using no knowledge of the Landsat image at 09/21. Moreover, comparing the results in Table 1 and 2, it can be seen that the higher SAM results observed for all methods at date 08/20 does not translate into a worse classification performance. This indicates that the SAM results at this date were influenced by the acquisition conditions of the Landsat image which was used for ground truth, making the classification performance more straightforward to interpret.

Finally, we plotted the percentage of pixels classified as water over the time index k in Figure 8, as well as a hydrograph which serves as an indicative of the dynamical evolution of the true level of the reservoir over time. It can be seen that ESTARFM was not able to properly identify the dynamical evolution of the reservoir level, leading to an estimation that was almost constant for all $k < 17$ and very different from the hydrograph curve. PSRFM led to results that, although showing relatively high day-to-day variations, were closer to the hydrograph curve. The Kalman filter and smoother-based algorithms, particularly those with the diagonal and block diagonal state covariance assumption (KF-D, KF-B, SM-B and SM-D) led to curves that were very close to the hydrograph. Thus, the Kalman filter methods captured the general trends of the hydrograph curves, even without having access to information from the Landsat image at the end of the sequence (like the smoothers and PSRFM). We note, however, that the connection between the hydrograph and the water surface area is indirect; thus, small differences between the algorithms have to be interpreted with proper care.

5.5. Contribution of the temporal dynamics calibration strategy

This subsection aims to show the impact of the proposed calibration strategy, which learns the temporal dynamical model parameters $\mathbf{Q}_k$ using historical data, on the performance of the proposed KF and SM algorithms. To this end, we compared the proposed

KF-D and SM-D (which estimate $\mathbf{Q}_k$ and use a diagonal assumption on the state covariance matrix), to a Kalman filter and smoother with a fixed $\mathbf{Q}_k = 10^{-2}\mathbf{I}$, which we denote by KF-I and SM-I, respectively. In Figure 9, we show the fused red (Figure 9a) and NIR (Figure 9b) reflectance images, as well as the acquired red and NIR reflectance values from MODIS and Landsat. Acquisition dates are displayed in the top labels at each column with a character, M for MODIS and L for Landsat indicating the image used in the fusion algorithms. We recall that only the first and last Landsat images were used in the fusion process, keeping the remaining three images as ground-truth for evaluation purposes. Analyzing the results, we can see that the images estimated by the proposed KF-D and SM-D methods produce significantly better visual similarity with the Landsat (ground-truth) images for both bands. For instance, the increase in the island and the expansion of other land parts at date 08/20 are clearly visible for the proposed methods. On the other hand, analyzing the results of the KF-I and SM-I methods, where the temporal dynamics matrix $\mathbf{Q}_k$ was kept constant and independent of past data, we observe that the results appear very blurry, with a resolution that is comparable to that of the MODIS images. This shows that the proposed weakly supervised calibration strategy is key in order for the KF- and SM-based strategies to obtain high quality reconstructions.

5.6. Results for larger scale Elephant Butte site

In this subsection, we compare the proposed strategies to ESTARFM and PSRFM in the Elephant Butte example, which comprises a larger geographical area. For simplicity and to reduce the use of space, we compare only proposed Kalman filter and smoother methods with the block diagonal assumption on the state covariance matrices (i.e., KF-B and SM-B).

The fusion results for both bands and all algorithms are shown in Figure 10, while Figure 11 shows the corresponding water mapping results. To measure the performances of different methods in this large area, the Landsat images at dates 06/07 and 06/23 were chosen as a ground truth to evaluate the quality of the reconstructed images at dates 06/14 and 06/27 (we remark that the MODIS images at dates 06/07 and 06/23 were not available due to the presence of cloud cover). It can be seen that the proposed KF-B, SM-B and the PSRFM methods provide estimates that are close to the ground truth images, whereas the ESTARFM method shows an inferior performance. This can be seen more clearly for the image at date 06/14 ($k = 5$), in which the smoother method better captured the increase in the area of the reservoir. To evaluate the performances of different methods more clearly, Figure 12 shows the absolute error of water maps of images compared with the ground truth, and Figures 13 and 14 show a zoomed-in area of the image of the fused image and water mapping result, respectively. It can be seen from Figure 12 that the misclassification errors are concentrated at the borders of the reservoir, which is the area that undergoes the largest amounts of changes over time, and consequently the hardest to classify correctly. The SM-B algorithm shows the best results, followed by KF-B, PSRFM and ESTARFM. Nevertheless, PSRFM provides results that contain less artifacts compared to KF-B, despite the lower classification accuracy. The superior visual quality of the results of SM-B and

PSRFM is explained by their use of Landsat images both at the beginning and at the end of the image sequence, whereas KF-B and ESTARFM do not have access to the last Landsat image.

Table 3 presents the SAM results, and Table 4 shows the corresponding percentage of misclassified pixels for the different methods. It can be seen that in terms of SAM, the SM-B method obtained the best results for both dates, followed by PSRFM and ESTARFM. However, the KF-B strategy was able to obtain a better water mapping performance compared to PSRFM. This indicates that the artifacts seen in the (comparatively noisier) reconstructions of KF-B impact the the classification performance in a less substantial way compared to the SAM. This shows that the proposed Kalman-filter based strategy can provide meaningful water mapping results in a real-time setting, in which we do not have access to future Landsat images, precluding smoothing-based algorithms (such as SM-B and PSRFM) to be used.

Table 3: Spectral angle mapper between the estimated high-resolution image and the Landsat measurement for the Elephant Butte example (note that the Landsat images at dates 06/07 and 06/23 were not supplied to the algorithms and only used for evaluation purposes).

Method	KF-B	SM-B	ESTARFM	PSRFM
Image (06/07)	5.5416	2.9993	9.2678	4.2698
Image (06/23)	5.7514	1.9923	6.2158	4.8719
Average	5.6465	2.4958	7.7418	4.5709

Table 4: Percentage of misclassified pixels for the Elephant Butte example (note that the Landsat images at dates 06/07 and 06/23 were not supplied to the algorithms and only used for evaluation purposes).

Method	KF-B	SM-B	ESTARFM	PSRFM
Image (06/07)	5.3593	1.4289	9.2678	6.6606
Image (06/23)	5.9233	0.8250	10.8330	7.8675
Average	5.6413	1.1269	10.0504	7.2640

5.7. Discussion

The results presented above clearly indicate that the proposed weakly supervised smoother-based image fusion strategy outperforms the ESTARFM and PSRFM algorithms in terms of image reconstruction when an appropriate covariance structure is selected (SM-D and SM-B). This highlights that having less model parameters to estimate (i.e., a more constrained state covariance model) can lead to better results. Moreover, even the Kalman filter strategies (particularly KF-B and KF-D), which estimate high-resolution images from MODIS without having access to any future data, have shown very competitive performance, with great potential for tasks in which high-resolution estimates are required online and one cannot wait for another Landsat image to be available before computing the high-resolution reconstructions.

The advantage of the proposed filter and smoother strategies is more clear when evaluated semantically by means of the water classification performance. For instance, the

growth of the island portion over time in regions that are semantically meaningful leads to more meaningful results that cannot be entirely captured by one standard metric such as the SAM. This can be observed more clearly through the spatial distribution of the misclassification error maps in Figure 6, which for ESTARFM and PSRFM are significantly more concentrated on the borders between land and water. In general, the proposed filtering-based strategies clearly outperformed both the ESTARFM and PSRFM algorithms, a standard and a state of the art remote sensing image fusion algorithms. Moreover, the proposed distributed implementation, described in Section 4, is able to reduce the computational power and memory demand of the standard Kalman filter and smoother when applied for large images.

6. Conclusions

In this paper, an online Bayesian approach for fusing multi-resolution space-borne multispectral images was proposed. By formulating the image acquisition process as a linear and Gaussian measurement model, the proposed method leveraged the Kalman filter and smoother to perform image fusion by estimating the latent high resolution image from the different observed modalities. Moreover, a weakly supervised strategy is also proposed to define an informative time-varying dynamical image model by leveraging historical data, which leads to a better localization of changes occurring in the high-resolution image even in intervals where only coarse resolution observations are available. Experimental results indicate that the proposed strategy can lead to considerable improvements compared to both classical and state-of-the-art image fusion algorithms.

7. Acknowledgments

The authors would like to thank the support of the National Geographic Society under Grant NGS-86713T-21, the National Science Foundation under Award ECCS-1845833, and NASA – GRACE-FO Science Team (80NSSC20K0742).



Figure 3: Fused bands from MODIS and Landsat for the Oroville Dam example using different strategies over time. The first two rows of each subfigure depict MODIS and Landsat bands acquired at dates displayed on top labels. At each time index estimation with KF and SM under different model assumptions, ESTARFM and PSRFM are presented. Some Landsat images were omitted from the estimation process and used solely as ground-truth. Images used at each update step are indicated on top labels where “M” stands for MODIS and “L” for Landsat.

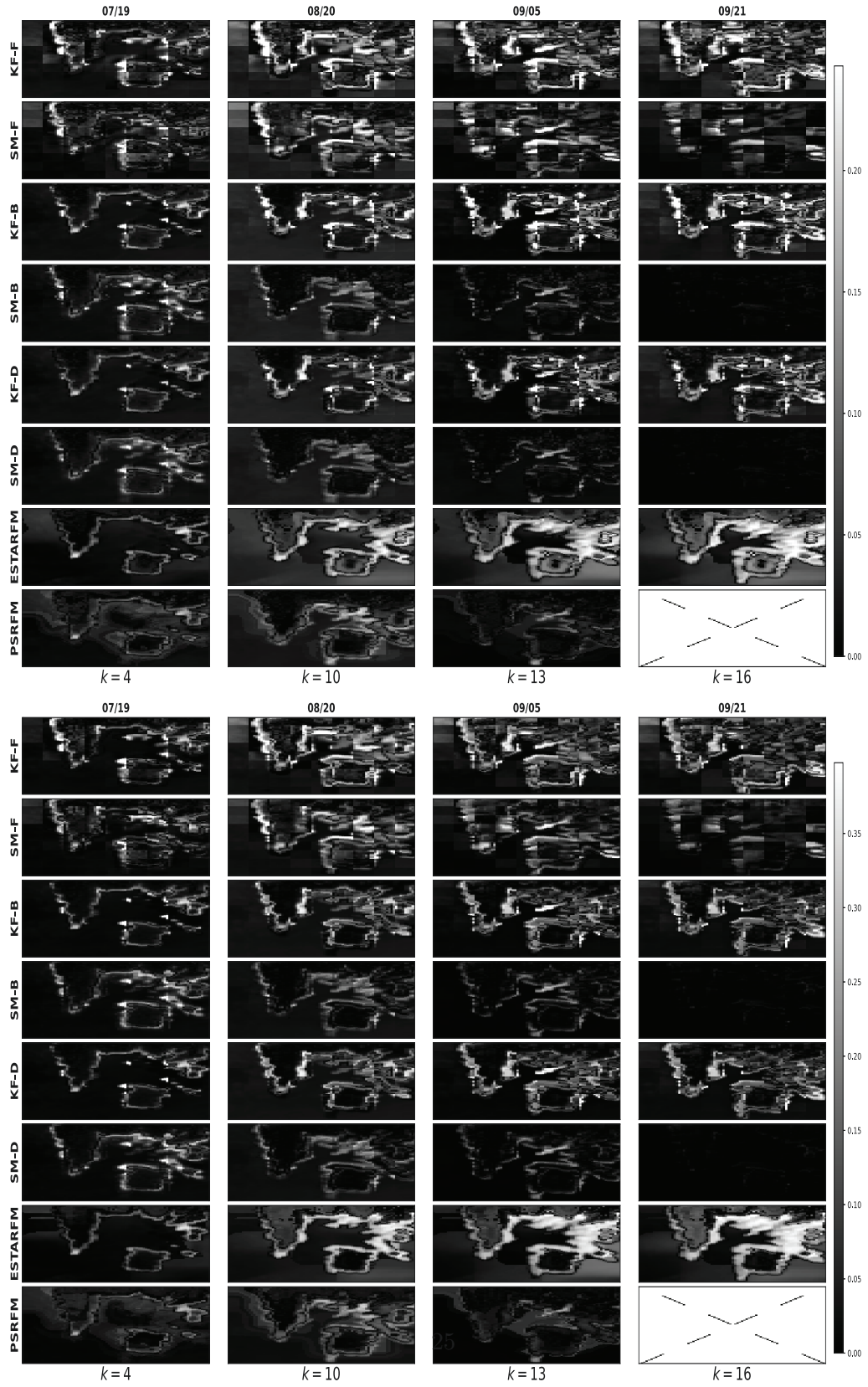


Figure 4: Absolute difference between the estimated and ground truth (Landsat) images for the Oroville Dam example in the red (**upper panel**) and NIR (**lower panel**) bands.

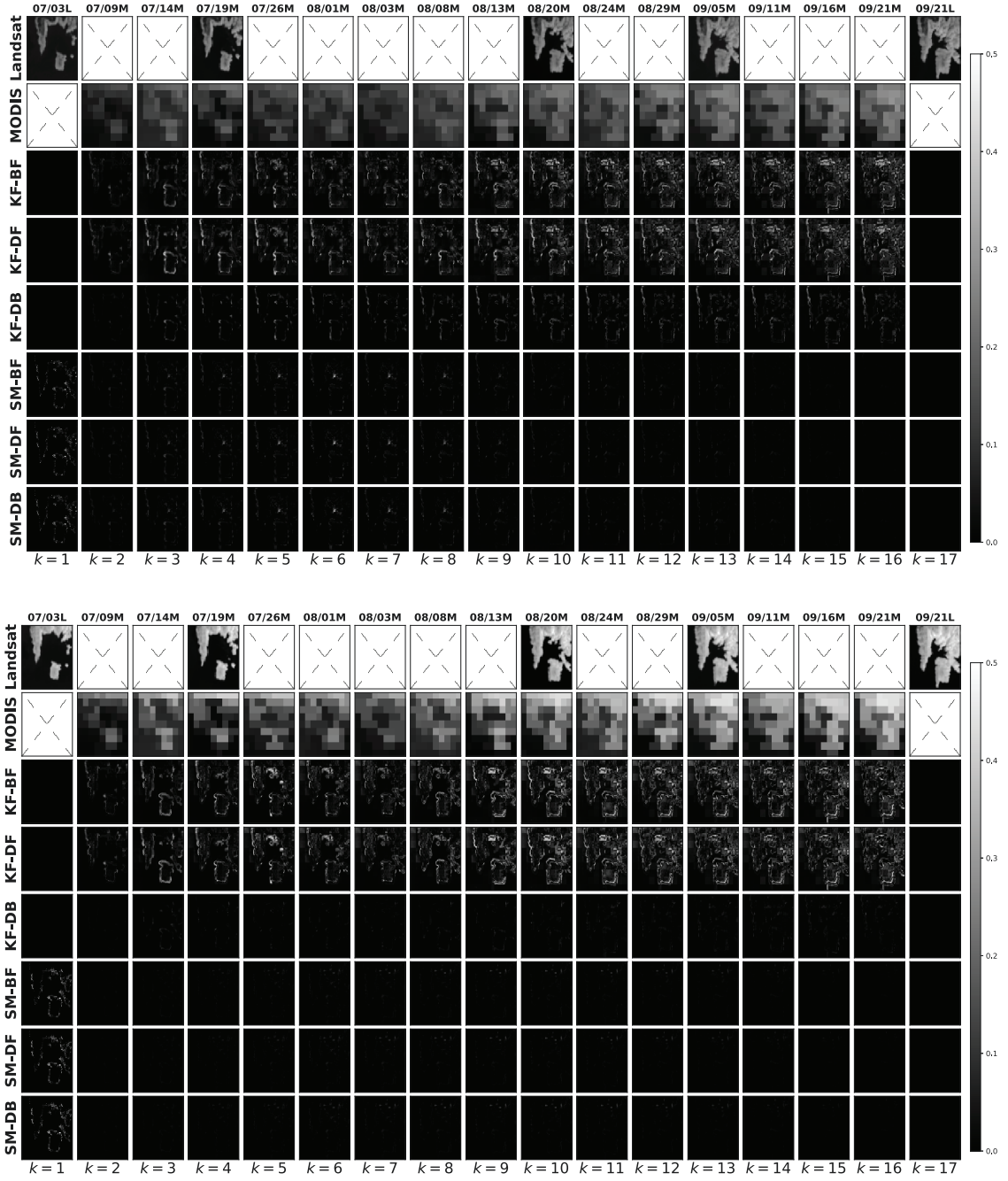


Figure 5: Absolute differences between the images estimated by the KF and Smoother under different model assumptions for red (**upper panel**) and NIR (**lower panel**) bands, for the Oroville Dam example. KF-BF: difference between the estimates of KF-B and KF-F. KF-DF: difference between the estimates of KF-D and KF-F. KF-DB: difference between the estimates of KF-D and KF-B. An analogous notation holds for the smoother (SM) estimates.

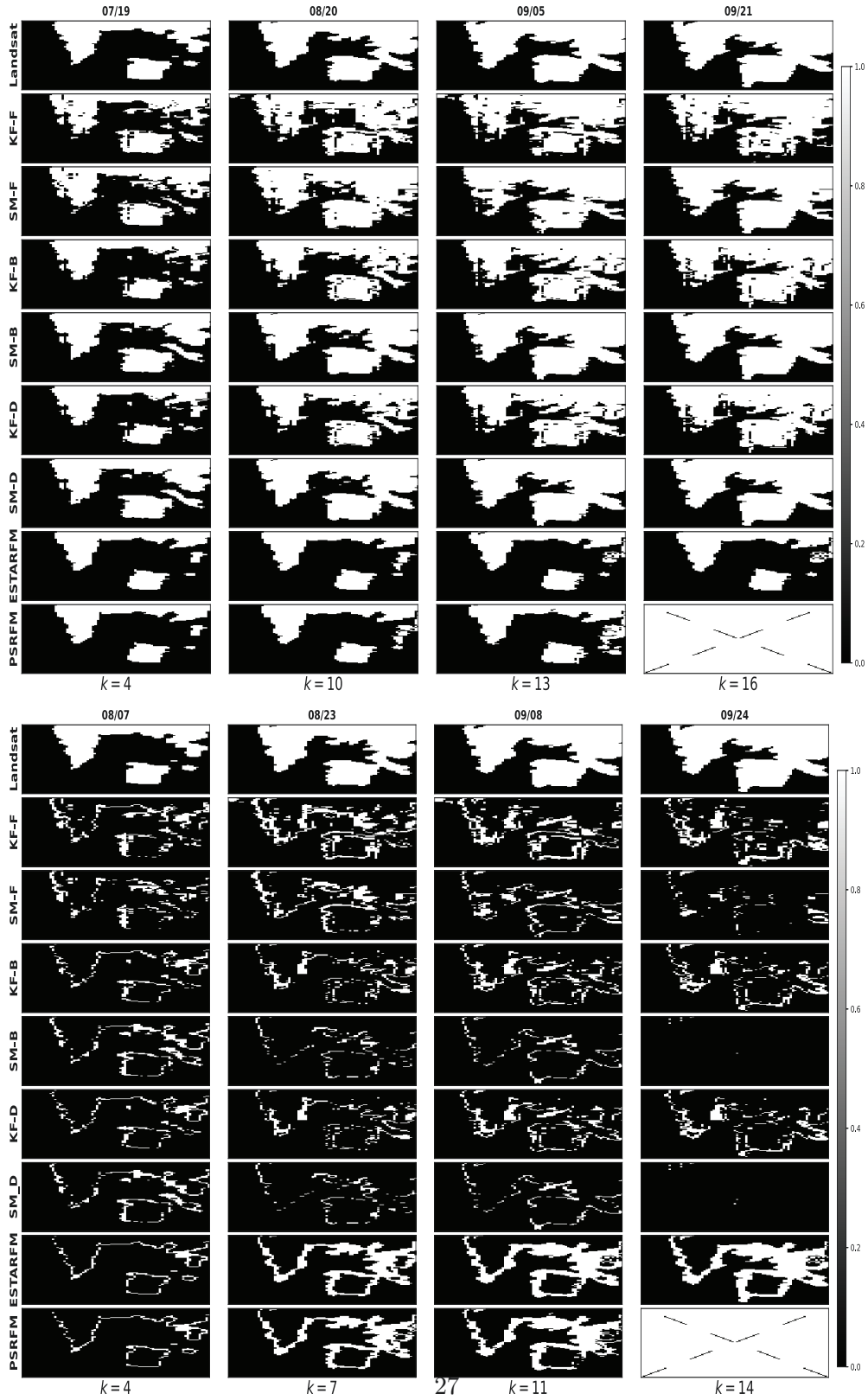


Figure 6: (**Upper Panel**) Water map of the reconstructed images of the Oroville Dam example based on K-means clustering strategy, where 1 indicates land and 0 indicates water pixels. Classification maps obtained from Landsat images not observed by the image fusion algorithms establish the ground-truth (first row). (**Lower Panel**) Absolute error of Water map of images based on K-means clustering strategy, where 0 indicates correctly classified pixels and 1 indicates misclassifications. The ground-truth is shown in the first row.

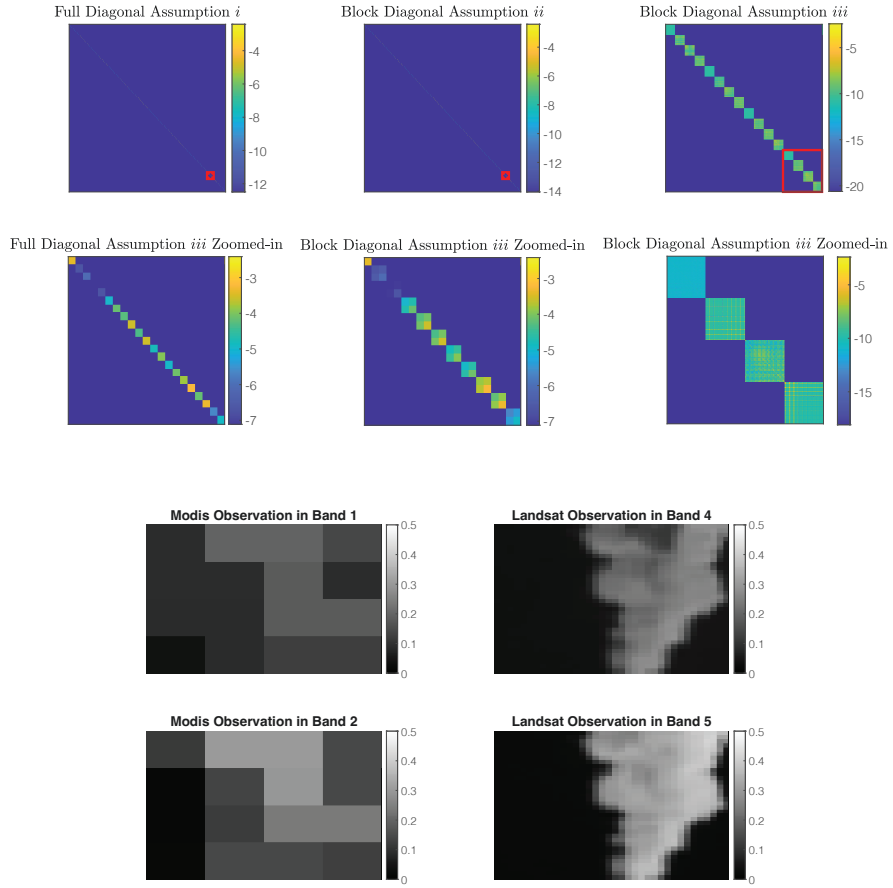


Figure 7: **(Top Colored Panel)** Estimated state covariance structure of the Kalman filter under model assumptions i , ii and iii for a small image area in the Oroville Dam example and $k = 13$. Top row depicts the whole covariance matrix with a red square indicating the zoomed part displayed on the bottom row. The plots indicate that correlations are present when assuming block diagonal covariance matrices. **(Bottom Panel)** Zoom of the MODIS image for bands 1 and 2 (left), and the corresponding Landsat observations for bands 4 and 5 (right) corresponding to the covariance matrices plotted in the right panels.

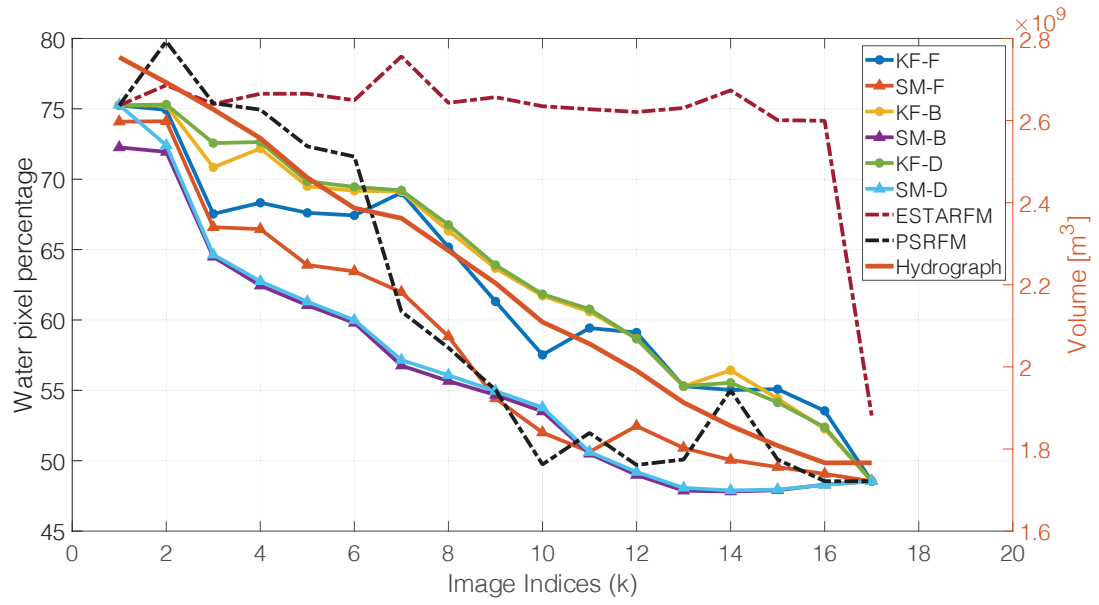


Figure 8: Percentage of water pixels in the estimated images over image index (time) and the reservoir volume in m^3 (hydrograph) for the Oroville Dam example. Classification of water was done by performing clustering on the estimated bands for each method and time index. High resolution Landsat images were observed at indices $k \in \{1, 17\}$.

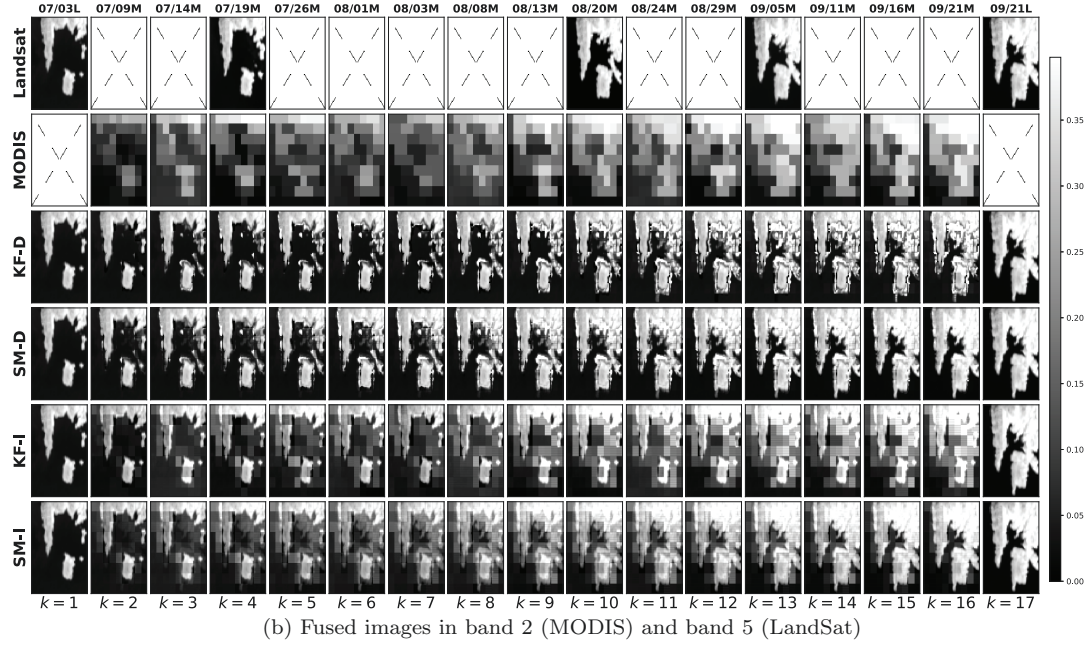
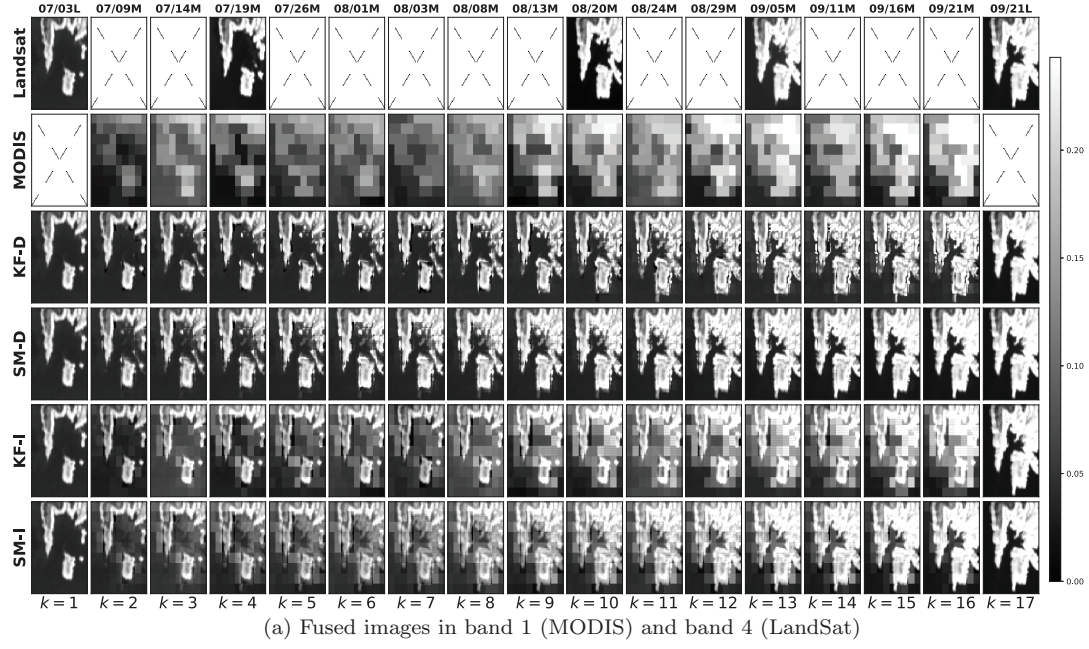


Figure 9: Fused bands from MODIS and Landsat for the Oroville Dam example using different strategies over time. The first two rows of each subfigure depict MODIS and Landsat bands acquired at dates displayed on top labels. At each time index estimation results of the diagonal Kalman filter and smoother with the proposed weakly supervised calibration strategy (KF-D and SM-D) are compared to the result of a Kalman filter and smoother with Q_k being proportional to the identity (denoted by KF-I and SM-I). Landsat images at dates 07/19, 08/20 and 08/29 were omitted from the estimation process and used solely as ground-truth. Images used at each update step are indicated on top labels where “M” stands for MODIS and “L” for Landsat.

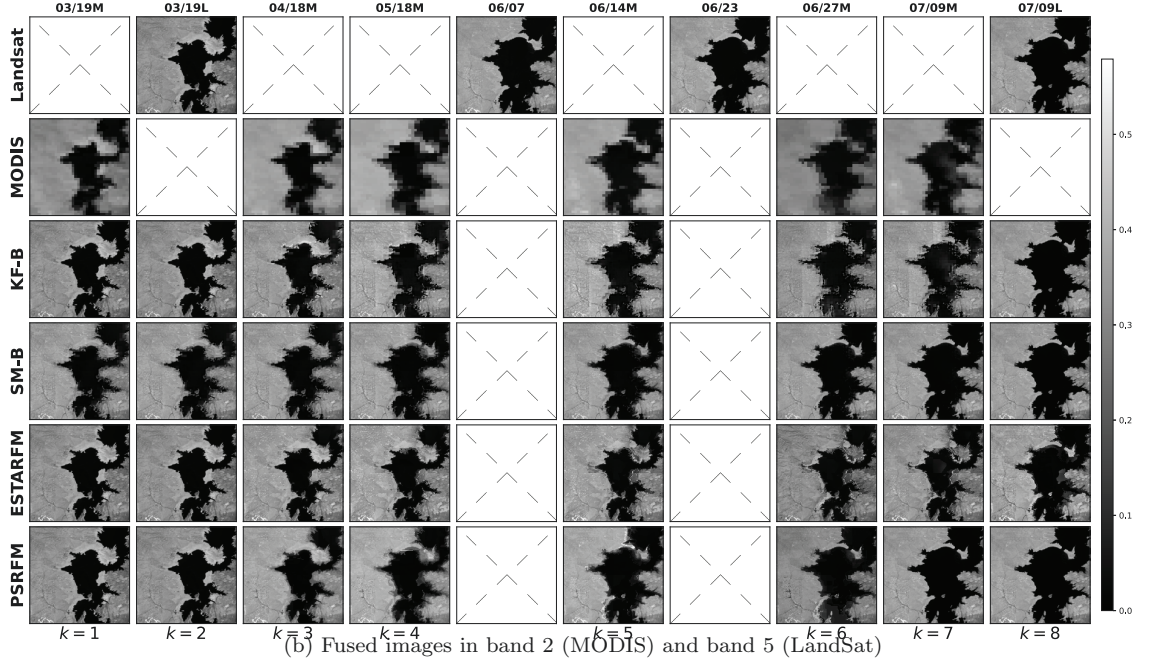
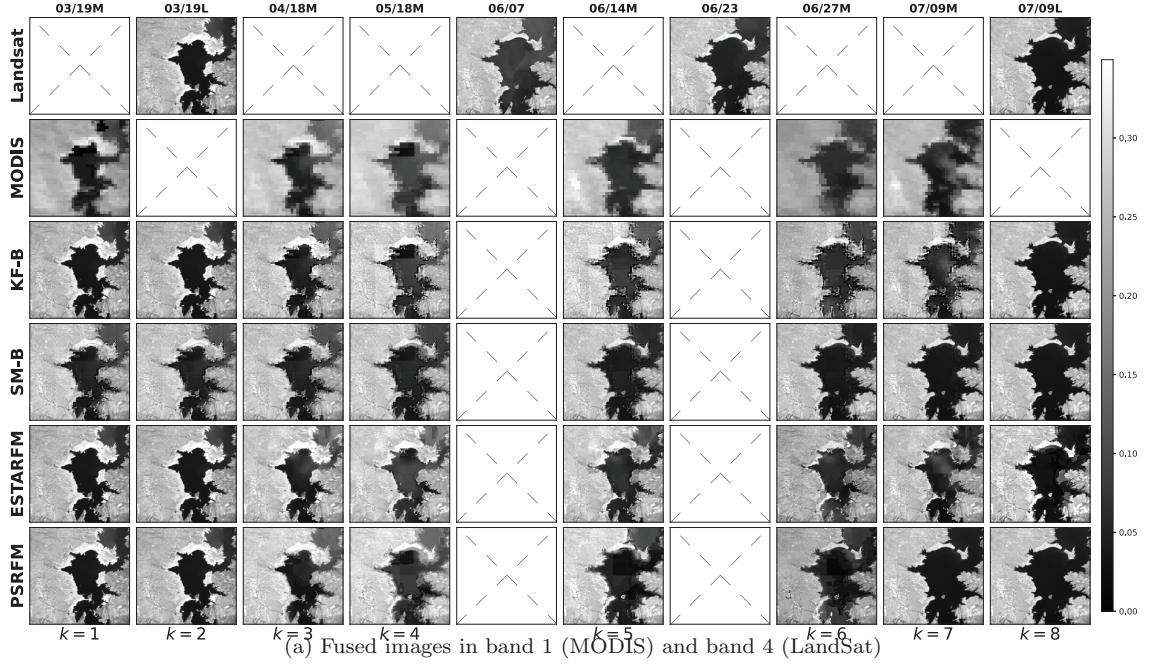


Figure 10: Fused bands from MODIS and Landsat for the Elephant Butte example using different strategies over time. The first two rows of each subfigure depict MODIS and Landsat bands acquired at dates displayed on top labels. At each time index estimation with KF and SM under block diagonal model assumptions, ESTARFM and PSRFM are presented. Some Landsat images were omitted from the estimation process and used solely as ground-truth. Images used at each update step are indicated on top labels where “M” stands for MODIS and “L” for Landsat.

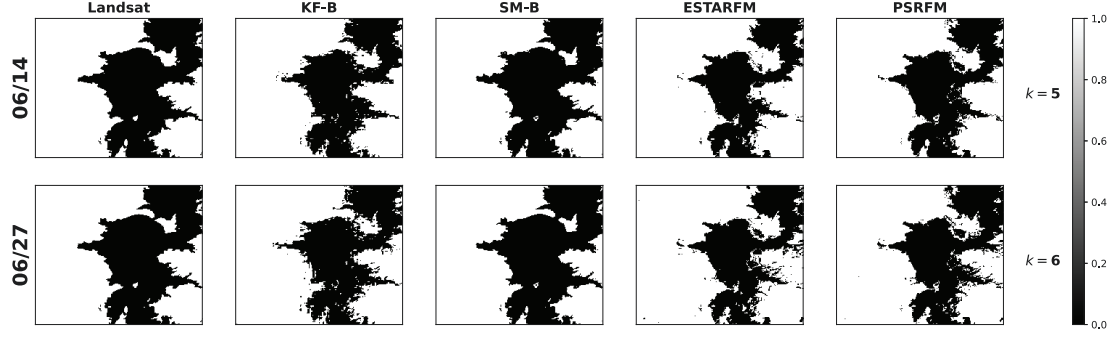


Figure 11: Water map of images for the Elephant Butte example based on K-means clustering strategy where 1 indicates land and 0 indicates water pixels. Unused Landsat classification maps establish the ground-truth (first column).

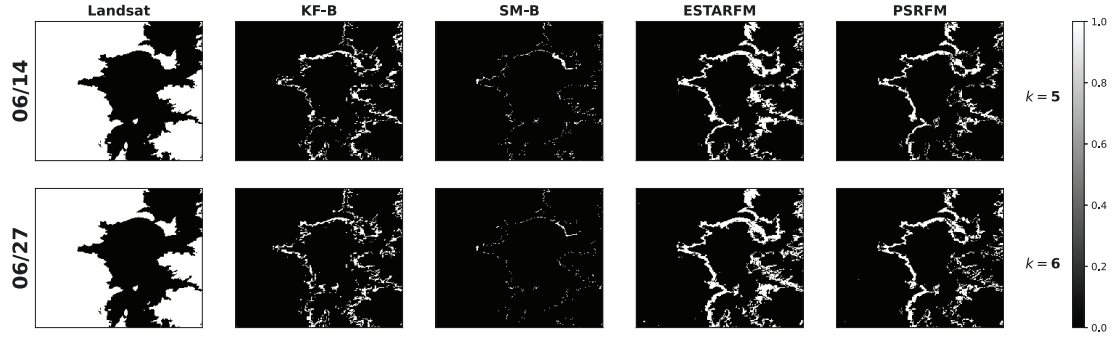


Figure 12: Absolute error of Water map of images for the Elephant Butte example based on K-means clustering strategy. Unused Landsat classification maps establish the ground-truth (first column).

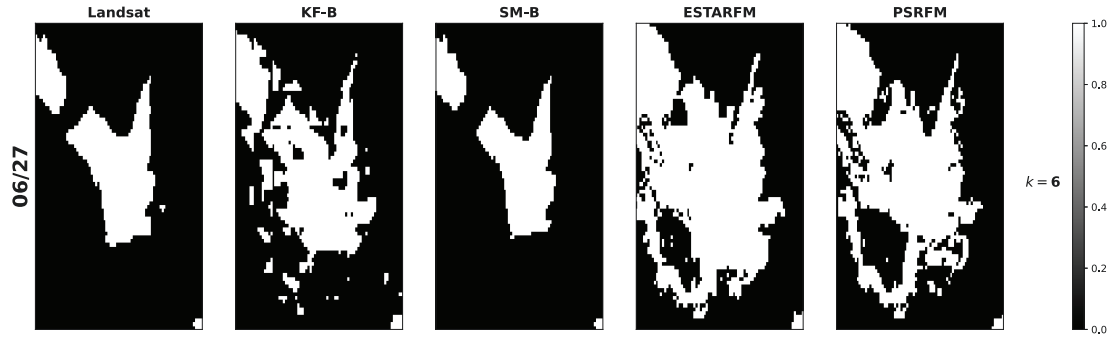


Figure 13: Zoomed-in water map of images for the Elephant Butte example based on K-means clustering strategy where 1 indicates land and 0 indicates water pixels. Unused Landsat classification map at date 06/23 establish the ground-truth (first column).

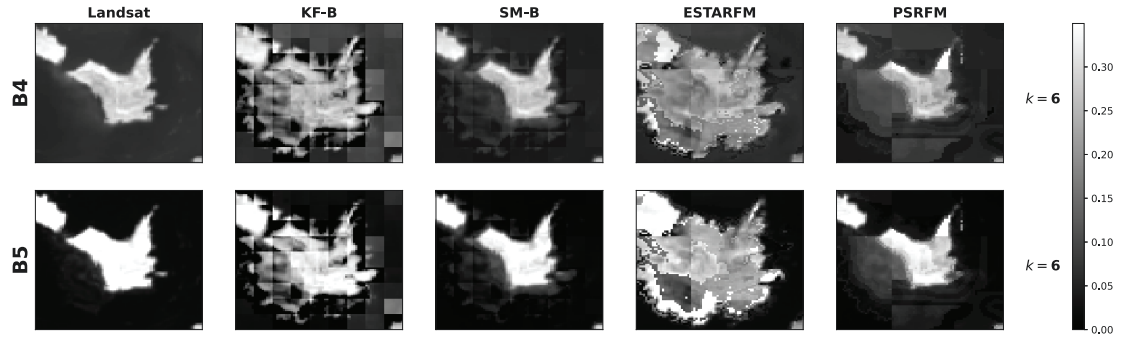


Figure 14: Zoomed-in version of the fused bands from MODIS and Landsat for the Elephant Butte example using different strategies at date 06/27 (ground-truth at 06/23 is shown in the first column).

References

- [1] M. Lu, J. Chen, H. Tang, Y. Rao, P. Yang, and W. Wu, "Land cover change detection by integrating object-based data blending model of landsat and modis," *Remote Sensing of Environment*, vol. 184, pp. 374–386, 2016.
- [2] Z. Zhu and C. E. Woodcock, "Continuous change detection and classification of land cover using all available landsat data," *Remote sensing of Environment*, vol. 144, pp. 152–171, 2014.
- [3] C. Portillo-Quintero, A. Sanchez, C. Valbuena, Y. Gonzalez, and J. Larreal, "Forest cover and deforestation patterns in the northern andes (lake maracaibo basin): a synoptic assessment using modis and landsat imagery," *Applied Geography*, vol. 35, no. 1-2, pp. 152–163, 2012.
- [4] M. Schultz, J. G. Clevers, S. Carter, J. Verbesselt, V. Avitabile, H. V. Quang, and M. Herold, "Performance of vegetation indices from landsat time series in deforestation monitoring," *International journal of applied earth observation and geoinformation*, vol. 52, pp. 318–327, 2016.
- [5] D. Kim, H. Lee, A. Laraque, R. M. Tshimanga, T. Yuan, H. C. Jung, E. Beighley, and C.-H. Chang, "Mapping spatio-temporal water level variations over the central congo river using palsar scansar and envisat altimetry data," *International Journal of Remote Sensing*, vol. 38, no. 23, pp. 7021–7040, 2017.
- [6] Y. Yoon, E. Beighley, H. Lee, T. Pavelsky, and G. Allen, "Estimating flood discharges in reservoir-regulated river basins by integrating synthetic swot satellite observations and hydrologic modeling," *Journal of Hydrologic Engineering*, vol. 21, no. 4, p. 05015030, 2016.
- [7] M. H. Gholizadeh, A. M. Melesse, and L. Reddi, "A comprehensive review on water quality parameters estimation using remote sensing techniques," *Sensors*, vol. 16, no. 8, p. 1298, 2016.
- [8] D. P. Roy, M. A. Wulder, T. R. Loveland, C. E. Woodcock, R. G. Allen, M. C. Anderson, D. Helder, J. R. Irons, D. M. Johnson, R. Kennedy *et al.*, "Landsat-8: Science and product vision for terrestrial global change research," *Remote sensing of Environment*, vol. 145, pp. 154–172, 2014.
- [9] Y. Li, Y. Zhou, Y. Zhang, L. Zhong, J. Wang, and J. Chen, "DKDFN: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 186, pp. 170–189, 2022.
- [10] Y. Yuan, L. Lin, Z.-G. Zhou, H. Jiang, and Q. Liu, "Bridging optical and SAR satellite image time series via contrastive feature extraction for crop classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 222–232, 2023.

- [11] V. S. F. Garnot, L. Landrieu, and N. Chehata, “Multi-modal temporal attention models for crop mapping from satellite time series,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 187, pp. 294–305, 2022.
- [12] J. Wu, L. Xia, T. O. Chan, J. Awange, and B. Zhong, “Downscaling land surface temperature: A framework based on geographically and temporally neural network weighted autoregressive model with spatio-temporal fused scaling factors,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 187, pp. 259–272, 2022.
- [13] A. Sharma, X. Liu, and X. Yang, “Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks,” *Neural Networks*, vol. 105, pp. 346–355, 2018.
- [14] Z. Fang, Y. Wang, L. Peng, and H. Hong, “Predicting flood susceptibility using LSTM neural networks,” *Journal of Hydrology*, vol. 594, mar 2021.
- [15] N. Yokoya, C. Grohnfeldt, and J. Chanussot, “Hyperspectral and multispectral data fusion: A comparative review of the recent literature,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 2, pp. 29–56, 2017.
- [16] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, “Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability,” *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 116–127, 2020.
- [17] L. Loncan, L. B. De Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes *et al.*, “Hyperspectral pansharpening: A review,” *IEEE Geoscience and remote sensing magazine*, vol. 3, no. 3, pp. 27–46, 2015.
- [18] M. Belgiu and A. Stein, “Spatiotemporal image fusion in remote sensing,” *Remote sensing*, vol. 11, no. 7, p. 818, 2019.
- [19] Q. Wang and P. M. Atkinson, “Spatio-temporal fusion for daily Sentinel-2 images,” *Remote Sensing of Environment*, vol. 204, pp. 31–42, 2018.
- [20] K. Rittger, M. Krock, W. Kleiber, E. H. Bair, M. J. Brodzik, T. R. Stephenson, B. Rajagopalan, K. J. Bormann, and T. H. Painter, “Multi-sensor fusion using random forests for daily fractional snow cover at 30 m,” *Remote Sensing of Environment*, vol. 264, p. 112608, 2021.
- [21] Y. Yang, M. C. Anderson, F. Gao, J. D. Wood, L. Gu, and C. Hain, “Studying drought-induced forest mortality using high spatiotemporal resolution evapotranspiration data from thermal satellite imaging,” *Remote Sensing of Environment*, vol. 265, p. 112640, 2021.

- [22] X. Zhu, F. Cai, J. Tian, and T. K.-A. Williams, "Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions," *Remote Sensing*, vol. 10, no. 4, p. 527, 2018.
- [23] F. Gao, T. Hilker, X. Zhu, M. Anderson, J. Masek, P. Wang, and Y. Yang, "Fusing Landsat and MODIS data for vegetation monitoring," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 47–60, 2015.
- [24] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance," *IEEE Transactions on Geoscience and Remote sensing*, vol. 44, no. 8, pp. 2207–2218, 2006.
- [25] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sensing of Environment*, vol. 114, no. 11, pp. 2610–2623, 2010.
- [26] Y. Zhang, G. M. Foody, F. Ling, X. Li, Y. Ge, Y. Du, and P. M. Atkinson, "Spatial-temporal fraction map fusion with multi-scale remotely sensed images," *Remote Sensing of Environment*, vol. 213, pp. 162–181, 2018.
- [27] T. Hilker, M. A. Wulder, N. C. Coops, J. Linke, G. McDermid, J. G. Masek, F. Gao, and J. C. White, "A new data fusion model for high spatial-and temporal-resolution mapping of forest disturbance based on Landsat and MODIS," *Remote Sensing of Environment*, vol. 113, no. 8, pp. 1613–1627, 2009.
- [28] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE signal processing magazine*, vol. 19, no. 1, pp. 44–57, 2002.
- [29] R. A. Borsoi, T. Imbiriba, J. C. M. Bermudez, and C. Richard, "A fast multiscale spatial regularization for sparse hyperspectral unmixing," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 4, pp. 598–602, April 2019.
- [30] R. Zurita-Milla, J. G. Clevers, and M. E. Schaepman, "Unmixing-based landsat TM and MERIS FR data fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 453–457, 2008.
- [31] J. Amorós-López, L. Gómez-Chova, L. Alonso, L. Guanter, R. Zurita-Milla, J. Moreno, and G. Camps-Valls, "Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring," *International journal of Applied earth observation and Geoinformation*, vol. 23, pp. 132–141, 2013.
- [32] M. Wu, Z. Niu, C. Wang, C. Wu, and L. Wang, "Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic landsat data using a spatial and temporal reflectance fusion model," *Journal of Applied Remote Sensing*, vol. 6, no. 1, p. 063507, 2012.

- [33] R. A. Borsoi, T. Imbiriba, J. C. M. Bermudez, C. Richard, J. Chanussot, L. Drumetz, J.-Y. Tournieret, A. Zare, and C. Jutten, "Spectral variability in hyperspectral data unmixing: A comprehensive review," *IEEE Geoscience and Remote Sensing Magazine*, 2021, doi: 10.1109/MGRS.2021.3071158.
- [34] R. A. Borsoi, T. Imbiriba, and J. C. Moreira Bermudez, "A data dependent multiscale model for hyperspectral unmixing with spectral variability," *IEEE Transactions on Image Processing*, vol. 29, pp. 3638–3651, 2020.
- [35] X. Li, G. M. Foody, D. S. Boyd, Y. Ge, Y. Zhang, Y. Du, and F. Ling, "SFSDAF: An enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion," *Remote Sensing of Environment*, vol. 237, p. 111537, 2020.
- [36] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 140–158, 2019.
- [37] R. A. Borsoi, T. Imbiriba, J. C. M. Bermudez, and C. Richard, "Fast unmixing and change detection in multitemporal hyperspectral data," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 975–988, 2021.
- [38] A. Ertürk, M.-D. Iordache, and A. Plaza, "Sparse unmixing-based change detection for multitemporal hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 708–719, 2015.
- [39] Q. Wang, Y. Tang, X. Tong, and P. M. Atkinson, "Virtual image pair-based spatio-temporal fusion," *Remote Sensing of Environment*, vol. 249, p. 112009, 2020.
- [40] W. Shi, D. Guo, and H. Zhang, "A reliable and adaptive spatiotemporal data fusion method for blending multi-spatiotemporal-resolution satellite images," *Remote Sensing of Environment*, vol. 268, p. 112770, 2022.
- [41] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3707–3716, 2012.
- [42] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 821–829, 2018.
- [43] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio-temporal-spectral fusion of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7135–7148, 2016.

- [44] B. Huang, H. Zhang, H. Song, J. Wang, and C. Song, “Unified fusion of remote-sensing imagery: Generating simultaneously high-resolution synthetic spatial–temporal–spectral earth observations,” *Remote sensing letters*, vol. 4, no. 6, pp. 561–569, 2013.
- [45] J. Xue, Y. Leung, and T. Fung, “A bayesian data fusion approach to spatio-temporal fusion of remotely sensed images,” *Remote Sensing*, vol. 9, no. 12, p. 1310, 2017.
- [46] F. Zhou and D. Zhong, “Kalman filter method for generating time-series synthetic landsat images and their uncertainty from Landsat and MODIS observations,” *Remote Sensing of Environment*, vol. 239, p. 111628, 2020.
- [47] F. Sedano, P. Kempeneers, and G. Hurtt, “A Kalman filter-based method to generate continuous time series of medium-resolution NDVI images,” *Remote Sensing*, vol. 6, no. 12, pp. 12 381–12 408, 2014.
- [48] S. Xu and J. Cheng, “A new land surface temperature fusion strategy based on cumulative distribution function matching and multiresolution Kalman filtering,” *Remote Sensing of Environment*, vol. 254, p. 112256, 2021.
- [49] R. A. Borsoi, T. Imbiriba, P. Closas, J. C. M. Bermudez, and C. Richard, “Kalman filtering and expectation maximization for multitemporal spectral unmixing,” *IEEE Geoscience and Remote Sensing Letters*, 2020.
- [50] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013, no. 3.
- [51] G. Kitagawa, “Non-Gaussian state-space modeling of nonstationary time series,” *Journal of the American statistical association*, vol. 82, no. 400, pp. 1032–1041, 1987.
- [52] D. Simon, “Kalman filtering with state constraints: a survey of linear and nonlinear algorithms,” *IET Control Theory & Applications*, vol. 4, no. 8, pp. 1303–1318, 2010.
- [53] P. Closas, C. Fernandez-Prades, and J. Vila-Valls, “Multiple quadrature Kalman filtering,” *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6125–6137, 2012.
- [54] J. Vilà-Valls, P. Closas, and Á. F. García-Fernández, “Uncertainty exchange through multiple quadrature Kalman filtering,” *IEEE signal processing letters*, vol. 23, no. 12, pp. 1825–1829, 2016.
- [55] J. Vilà-Valls, P. Closas, Á. F. García-Fernández, and C. Fernández-Prades, “Multiple sigma-point Kalman smoothers for high-dimensional state-space models,” in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2017, pp. 1–5.
- [56] D. Zhong and F. Zhou, “Improvement of clustering methods for modelling abrupt land surface changes in satellite image fusions,” *Remote Sensing*, vol. 11, no. 15, p. 1759, 2019.

- [57] —, “A prediction smooth method for blending landsat and moderate resolution imagine spectroradiometer images,” *Remote Sensing*, vol. 10, no. 9, p. 1371, 2018.
- [58] B.-C. Gao, “NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space,” *Remote sensing of environment*, vol. 58, no. 3, pp. 257–266, 1996.
- [59] E. F. Vermote, J. C. Roger, and J. P. Ray, “MODIS Surface Reflectance User’s Guide,” NASA, Tech. Rep., May 2015.
- [60] R. Furrer and T. Bengtsson, “Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants,” *Journal of Multivariate Analysis*, vol. 98, no. 2, pp. 227–255, 2007.