6-17-2021

# A Quantitative Validation of Multi-Modal Image Fusion and Segmentation for Object Detection and Tracking

Nicholas LaHaye
*California Institute of Technology*

Michael J. Garay
*California Institute of Technology*

Brian D. Bue
*California Institute of Technology*

Hesham el-Askary
*Chapman University*, elaskary@chapman.edu

Erik Linstead
*Chapman University*, linstead@chapman.edu



## Recommended Citation

LaHaye N.; Garay M. J.; Bue, B. D.; El-Askary, H.; Linstead, E. A Quantitative Validation of Multi-Modal Image Fusion and Segmentation for Object Detection and Tracking. *Remote Sens.* 2021, *13*, 2364. http://doi.org/10.3390/rs13122364

# A Quantitative Validation of Multi-Modal Image Fusion and Segmentation for Object Detection and Tracking

*remote sensing*

MDPI

*Article*

# A Quantitative Validation of Multi-Modal Image Fusion and Segmentation for Object Detection and Tracking

**Nicholas LaHaye [1,2,3], Michael J. Garay [1], Brian D. Bue [1], Hesham El-Askary [2,4,5,*] and Erik Linstead [3,6]**

1 Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91101, USA; nlahaye@jpl.nasa.gov (N.L.); michael.j.garay@jpl.nasa.gov (M.J.G.); bbue@jpl.nasa.gov (B.D.B.)
2 Schmid College of Science and Technology, Chapman University, Orange, CA 92866, USA
3 Machine Learning and Assistive Technology Lab (MLAT), Chapman University, Orange, CA 92866, USA; linstead@chapman.edu
4 Center of Excellence in Earth Systems Modeling and Observations, Chapman University, Orange, CA 92866, USA
5 Department of Environmental Sciences, Faculty of Science, Alexandria University, Moharem Bek, Alexandria 21522, Egypt
6 Fowler School of Engineering, Chapman University, Orange, CA 92866, USA
* Correspondence: elaskary@chapman.edu

**Abstract:** In previous works, we have shown the efficacy of using Deep Belief Networks, paired with clustering, to identify distinct classes of objects within remotely sensed data via cluster analysis and qualitative analysis of the output data in comparison with reference data. In this paper, we quantitatively validate the methodology against datasets currently being generated and used within the remote sensing community, as well as show the capabilities and benefits of the data fusion methodologies used. The experiments run take the output of our unsupervised fusion and segmentation methodology and map them to various labeled datasets at different levels of global coverage and granularity in order to test our models' capabilities to represent structure at finer and broader scales, using many different kinds of instrumentation, that can be fused when applicable. In all cases tested, our models show a strong ability to segment the objects within input scenes, use multiple datasets fused together where appropriate to improve results, and, at times, outperform the pre-existing datasets. The success here will allow this methodology to be used within use concrete cases and become the basis for future dynamic object tracking across datasets from various remote sensing instruments.

**Keywords:** big data applications; clustering; computer vision; restricted Boltzmann machines (RBMs); unsupervised machine learning; image segmentation; multi-modal data fusion



**Citation:** LaHaye N.; Garay M. J.; Bue, B. D.; El-Askary, H.; Linstead, E. A Quantitative Validation of Multi-Modal Image Fusion and Segmentation for Object Detection and Tracking. *Remote Sens.* **2021**, *13*, 2364. http://doi.org/10.3390/rs13122364

Academic Editors: Fahimeh Farahnakian, Jukka Heikkonen and Dimitrios Makris

Received: 3 May 2021
Accepted: 9 June 2021
Published: 17 June 2021



## 1. Introduction

Remote sensing and instrumentation are constantly improving and increasing in capability, including an increase in the amount of different instrument types, with various combinations of spatial and spectral resolutions, pointing angles, and various other instrument-specific qualities. While the increase in instruments (and, therefore, datasets) is a boon for those aiming to study the complexities of various Earth systems, it has also led to a large number of new challenges. With this information in mind, our group has set our aims on combining data sets with different spatial and spectral resolutions in an effective and as-general-as-possible way, with as little pre-existing per-instrument or per-dataset bias as possible, in order to create a system that can use pre-existing instrumentation/data sets as a sensor web of sorts. To begin, we have leveraged both unsupervised machine learning, specifically restricted Boltzmann machines (RBMs), and clustering techniques, in order to effectively separate or segment different kinds of objects within data obtained from various spectral imagers [1]. Other works considering the general problem of unsupervised image segmentation appear to have had success in separating the foreground

from the background [2,3], or have only used single bands of input from one type of instrumentation, which is effective for their applications, but does not cover the breadth required here [4]. Other works have aimed to perform tasks, such as outlining buildings and roadways [5], which is not the goal here. A similar study, which used autoencoders and a form of clustering—an overall architecture that is close to ours—attained an accuracy of 83% on Landsat imagery alone [6], whereas this work attains higher accuracy (and balanced accuracy, in some cases) across many different instrument sets, including fused data. In this paper, we quantitatively compare the performance of our output, both when using single instruments and the fusion of multiple collocated data sets, against pre-existing classification products; in doing so, we comprehensively show the value of the RBM-cluster methodology for detailed structural understanding of the data sets tested. Within these experiments, data sets from both satellite-based and airborne instrumentation were used. Table 1 details the satellite-based instruments and data sets used, while Table 2 details the airborne instruments used.

**Table 1.** Satellite instruments and their products.

| Platform | Instruments | Science Products | Spatial Resolution |
|---|---|---|---|
| Terra | Multi-angle Imaging SpectroRadiometer (MISR) | Spectral intensities in 446 nm, 558 nm, 672 nm, and 867 nm | 1.1 km and 275 m, all resampled to 1.1 km |
| Terra | MODerate resolution Imaging SpectroRadiometer (MODIS) | Spectral intensities in 38 bands in 445 nm–967 nm and 1.616 µm–14.062 µm spectral range | 1 km resampled to 1.1 km |
| Sentinel-2 Constellation | Multi Spectral Instrument (MSI) | Spectral intensities in 10 443 nm–2190 nm spectral range | All resampled to 100 m |
| Landsat-8 | Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) | Spectral intensities in 9 bands in 0.43 µm–0.88 µm, 1.57 µm–2.29 µm, and 10.6 µm–12.51 µm spectral ranges | All resampled to 100 m |
| EO-1 | Hyperion | Spectral intensities in 220 bands in 400–2500 nm spectral range | 30 m |

**Table 2.** Airborne instruments and their products.

| Platform | Instruments | Science Products | Spatial Resolution |
|---|---|---|---|
| NASA ER-2 | Enhanced MODIS Airborne Simulator (eMAS) | Spectral intensities in 38 bands in 445 nm–967 nm and 1.616 µm–14.062 µm spectral ranges | 50 m |
| NASA DC-8 | MODIS/ASTER Airborne Simulator (MASTER) | Spectral intensities in 50 bands in 0.44–12.6 µm spectral range | 10–30 m |

## 2. Materials and Methods

### *2.1. Methods*

Restricted Boltzmann machines (RBMs) are simple two-layer learning architectures that can be trained in an unsupervised or supervised fashion. In this work, we use unsupervised RBMs. The RBMs can be stacked, thus forming a deep learning model, called a Deep Belief Network (DBN) [7]. The work described here only uses RBMs with a single layer.

An RBM is a variation of a hidden Markov field, whose energy function is linear in its free parameters. RBMs are "restricted", due to the fact that edges can only make connections between adjacent layers. Each unit in the visible layer is connected to each unit in the hidden layer, but no other intra-layer connections are allowed. The energy function used for RBMs is:

$$E(v,h) = -b^T v - c^T h - hWv \quad (1)$$

where $v$ is the set of visible units, $h$ is the set of hidden units, $b$ and $c$ are the sets of offsets for the visible and hidden units, respectively, and $W$ is the set of weights for each of the edges that connect the layers. The initial energy function can be translated into the free energy formula:

$$F(v) = -b^T v - \sum_i log(\sum_{h_i}(e^{h_i(c_i + W_i v)})) \quad (2)$$

This allows us, given the definition of energy-based models with hidden units, to define the probability distribution as:

$$P = \frac{e^{-F(x)}}{\sum_x (e^{-F(x)}} \quad (3)$$

RBMs are trained using a process called contrastive divergence, instead of performing gradient descent on the second derivative of the negative log-likelihood, as is done with traditional feed-forward neural networks. Contrastive divergence is used, in this case, as a way to speed up training, as an RBM is a special case of a Markov field, and the RBM would have to be run to convergence on its equilibrium distribution for each parameter update in order to use expected values from that distribution and calculate new updates. This is a computationally complex process, and the variance within the values sampled from the equilibrium distribution is typically high enough to cause issues when training. Instead of comparing the input or initial distribution with the equilibrium distribution, contrastive divergence runs an initial number $N$ of Gibbs sampling steps. In order to keep updates from causing the new distribution to deviate significantly from the initial distribution, the Kullback–Leibler (KL) divergence is measured for each parameter update. In addition, bias constraints have been recommended for the contrastive divergence process, in order to account for the sparsity and selectivity for sets of hidden units [8]. This allows for activation diversity, meaning that each hidden unit only activates when necessary, not simply when an instance reaches the hidden layer of the RBM. Along with this, sets of hidden units, while sparsely activating, should not all activate at the same time, allowing for selectivity.

While patterns can be recognized by RBMs, the end-user cannot interpret the output of an RBM's hidden layer as it is generated. In order to translate the output back into a human-readable format, we use a form of an agglomerative clustering technique called BIRCH clustering. A general clustering problem can be seen as a multi-objective optimization problem. The input is a set of $N$ data points with $M$ features. The goal is to group the data into a desired number of clusters, $K$, while minimizing the given error (or distortion) function. In agglomerative or hierarchical clustering, each data point belongs to a cluster $s_j$. At each step, all clusters are compared, and a merge operation is performed: $sa = sa \cup sb$, where $a$ and $b$ are cluster indices at step $i$. This merge operation is performed on the two clusters whose merge minimally affects the error function. BIRCH clustering achieves the goal of pattern recognition while also being memory efficient, by performing the clustering through a tree-based approach [9]. For the clustering process, the same pixels that are used to train the RBM are also used to train the clustering model.

### *2.2. Materials and Tools*

The software was developed with Python 3.6.8. All of the RBM training and testing was implemented using Lrn2 Deep Learning Framework [10], utilizing Theano with a GPUArray back-end (https://github.com/Theano/Theano/wiki, accessed on 11 June 2021). These packages are no longer being supported, so future work will be moved to using Learnergy [11], which utilizes PyTorch [12], but the libraries aforementioned worked well for this study. The hardware utilized was an NVIDIA GeForce Titan X GPU with 12 GB memory, as well as the NCCS Prism GPU Cluster (https://www.nccs.nasa.gov/systems/ADAPT/Prism, accessed on 11 June 2021). As for the clustering, it was performed using Scikit-Learn [13] on a machine running Ubuntu 14.04.5.

Our RBMs used two types of input sets. The first consists of geolocated orthorectified L1 data from a single instrument. The other consists of collocated orthorectified L1 data sets for spatially and temporally overlapping targets from multiple instruments with similar spatial resolution. The fusion techniques are described in the results section, along with examples. Each sample consists of itself (i.e., a pixel) and all of its neighboring pixels. This allows for a small amount of spatial context to be included as input, along with the spectral information. All pixels that are set to fill values or are out of specified valid ranges were not used. Regarding the spectral bands used, all spectral bands were used, with the exception of bands that were extremely noisy or known to be non-functional for the time period tested. For each RBM, at least 1,000,000 samples were used for training, and at least another 1,000,000 samples were used for testing. All input was also standardized (by channel) before being used as input to the RBM, and again before being used as input to the clustering model. Below, the reader will see that only $>$80,000 pixels were evaluated in the Landsat-8/Sentinel-2 tests. We believe this is still an adequate amount of data to evaluate the performance, but a smaller number of pixels was used because only a small percentage of the images were labeled (when labels were given). The same pixels were used to train both the RBMs and the clustering models. Once the clusters are generated, they then must be assigned a context in order to be used. For this experiment, we needed to assign a context relative to pre-existing products, such as pixel classifiers, fire masks, or aerosol optical depth (AOD) data sets, in order to properly compare within the same context. If there are already products, they can be used as a reference for automated mapping. Within the automated mapping process, a full image mask is built from the pre-existing product. This is either provided within the product, or there are instructions on how to compute one, given various certainty levels for each possible label. Given the full label set for the pre-existing product, spanning all test-set scenes, each cluster from our clustering product was mapped to the label it best agrees with. In some cases, as with the finer-scale evaluation of fire detection in the second subset of experiments, an automated mapping was paired with a secondary manual pixel labeling process, in order to account for clusters that correctly identified parts of an object (e.g., a fire), but were not identified in pre-existing products. This manual assignment process is much like that of the manual pixel-labeling process for training supervised learning models. However, our methodology utilizes the strong pattern-matching and data shape understanding capabilities that RBMs and BIRCH clustering offer, before human intervention occurs. These capabilities allow for a much simpler and less error-prone manual intervention technique. The separation of context assignment from the image segmentation/clustering itself is also valuable, as it allows the image segmentation product to be used for many other studies.

### *2.3. Large Scale Coarse Full Scene Evaluations*

The first few experiments performed were large-scale multi-scene investigations, which aimed to quantitatively answer the following questions: Can we capture coarse-scale information across a large set of scenes inside and outside of the area, in order to train the models? Can we provide fused data based on a generic set of methodologies, such that they provide added value when fused, while the data based on singular instruments are still viable enough for use when collocation or overlap does not occur? Can we achieve this in a way that is not extremely resource-hungry? To answer these questions, we evaluated three separate sets of data sets. Our first set was from the Multi-angle Imaging SpectroRadiometer (MISR) and the Moderate resolution Imaging Spectroradiometer (MODIS)—two instruments aboard the same satellite, Terra. These were compared against classification data sets that have been previously produced using the science data-processing pipelines of the respective instruments. The second set was a data set from the IEEE GRSS Data Fusion Challenge from 2017, which consists of imagery from the Landsat-8 and Sentinel-2 satellites. These were compared against local climate zones, provided as labels. Finally, for this first set of experiments, we used data from the Hyperion instrument, a multispectral imager aboard the EO-1 satellite.

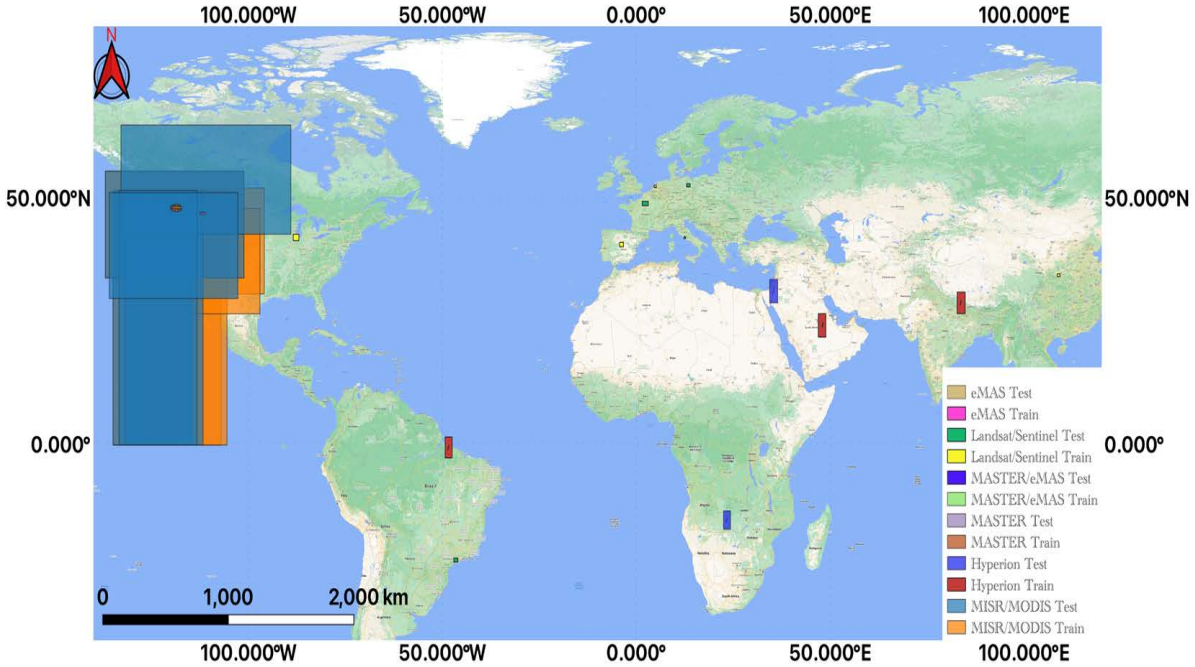
For all models generated, the architecture and parameterization remained the exact same. We did not want model variation to play a part in performance variation. We evaluated these comparisons in a few ways. The first was agreement, which considers the total percentage of labels that the pre-existing label sets and our mapped data sets agree upon, measured as:

$$\frac{total\_mapped\_pixles\_agreed\_upon}{total\_pixels}. \quad (4)$$

Note that, due to the fact that there is inherent uncertainty within most of the pre-existing classification products and, as we will show, they are not always completely correct, we named this metric agreement, and not accuracy. The second metric is balanced agreement, which also measures the total percentage of agreement, but takes into account the imbalance of pixel counts across the different labels [14]. In order to evaluate the structural understanding of the data through the output received from the models, we used a clustering metric called the Davies–Bouldin score [15]. This metric is much like the more commonly used silhouette score, but is much less computationally complex and, therefore, more feasible to use, given the amount of data. The aim is to measure the compactness of each cluster and the separation between each cluster, as good clustering performance is assumed to provide compact clusters that are far away from one another in the actual feature space, and not the (line, sample) image space. For the Davies–Bouldin score, a lower score indicates a better clustering performance and, therefore, a better structural understanding of the data (especially when coupled with higher agreement percentages). Finally, we wanted to look at the computational cost of training the RBMs, and whether there was any extreme increase in processing incurred when the fusion was done. To this end, we measured the amount of time necessary to train the model and the number of iterations each RBM required to reach convergence. It should be noted that an early stopping condition was added to the training of these RBMs, which is a common practice [16]. With this in mind, if the model's reconstruction error, or the difference between the output distribution and the input distribution (as discussed in the Methods Section) remains the same or increases for three iterations, the training is stopped and convergence is assumed. We also generally know that RBMs perform well with only a few training iterations [8], so a low number of iterations minimizes the chance of overfitting. Within the table, we provide evaluations for singular instruments and fusion sets, as well as results for clustering without passing the data through an RBM first (which we label here as "Raw", as only the raw orthorectified radiances were used). We provide information of the latter in order to show that the RBM enhances the structural understanding of the fused data. For all cases, the single instrument data sets passed through the RBM and clustered were viable, and can definitely be used for image segmentation when fusion/collocation is not available; however, the RBM-based fusion product always performed the best.

MISR is an instrument onboard the Terra satellite, which consists of nine different cameras, one of which points at the nadir. The other eight are split into two equal groups of forward and aft cameras. Each group has cameras that point at matching angles, relative to the local normal at the Earth's surface: 25.8°, 45.6°, 60.0°, and 72.5° [17]. The Moderate resolution Imaging Spectroradiometer (MODIS) is an imager with 36 spectral bands, whose resolutions span from 250 to 1000 m. For this test, we used only the 1000 m bands, which are measured continuously during both day and night [18]. In order to evaluate performance, we chose a region over the west coast of North America as the region to use for scenes to train the models with, and a partially overlapping region as the testing region, as depicted in Figure 1. In this way, we allowed for the evaluation of performance inside and outside the training extents. The reference imagery used an as example within this paper, is from outside the training extent. As there was no large difference in agreement inside or outside the training extent, all confusion matrices shown are a combination of all test scenes. Within this first experiment, we trained models for one MISR camera on its own, the fusion of all nine MISR cameras, MODIS, and nine MISR cameras + MODIS fusion. We also looked

at the raw MISR-9 camera + MODIS fusion product generated from clustering, without passing the data through the RBM. We compared against a couple of pre-existing products by mapping the classes from our RBM-based clustered product to classes with pre-existing MISR and MODIS pixel classification products that best agreed, using the pre-existing product as a label set. The first product used in this was from the MISR Support Vector Machine (SVM) classifier. This classifier is able to efficiently distinguish between clouds, aerosols, water, land, smoke/dust, and snow/ice, with an impressive global accuracy of 81% over all defined classes [19]. The only drawback with the MISR SVM product that arose in the tested scenes is that the snow label is often applied to cloudy areas. This is a difficult problem to solve with classification alone, as some clouds and snow/ice contain the same materials, only at different altitudes. The other product used here was the MODIS cloud mask [20]. This data set has classes for clouds, aerosols, land, desert, snow/ice, and water. It appears to have issues when identifying large areas of aerosols, due to a thresholding issue identified in some studies using this data set [21]. The multiple land classes are extremely useful here, as it breaks land up into land and desert classes. There is also more detail in the inland water. One drawback is that the land/water identification is based on static data sets and, thus, it may not fully reflect what is seen in a given scene; nonetheless, the granularity is still useful.




(**a**) Training and Testing Extents

**Figure 1.** Summary of training and testing extents for all experiments described in this paper. All test regions contain at least a subset of area disparate from that of the area trained on. Extents for eMAS, MASTER, and eMAS + MASTER fusion overlap.

The second coarser-scale experiment involved Landsat-8 and Sentinel-2 data provided by IEEE GRSS. Landsat-8 is a satellite platform that contains 2 instruments, one of which is the Operational Land Imager (OLI), and the other is the Thermal Infrared Sensor [22,23]. Alongside the Landsat-8 data, data from the Sentinel-2 constellation of satellites, specifically a subset of channels from the Multi-Spectral Instruments (MSIs), was provided [24]. There was also OpenStreetMap data available, providing information about land use, buildings, and water, but this was not used. For labels, hand labeled local climate zone (lcz) data was provided by WUDAPT [25] and GeoWiki (http://www.geo-wiki.org/,

accessed on 11 June 2021). The scenes chosen have completely clear skies, and the goal is to be able to classify objects on the ground. Additionally, the imagery is not necessarily from the exact same time period, but due to the fact that the sky is clear, this does not effect the fusion effort. The LCZ label data was only provided for a subset of the scenes, as the data was released as part of a competition, so some labels were withheld. These LCZs were only given for small parts of each scene, and were broken up into 17 different labels: Urban/compact high-rise, urban/compact mid-rise, urban/compact low-rise, urban/open high-rise, urban/open mid-rise, urban/open low-rise, urban/lightweight low-rise, urban/large low-rise, urban/sparsely built, urban/heavy industry, land cover/dense trees, land cover/sparse trees, land cover/brush and scrub, land cover/low plants, land cover/bare rock or paved, land cover/bare soil or sand, and land cover/water. We did not think it feasible to attain this level of specificity, so we reduced them into five classes: urban, tree cover, low plants/brush/scrub, bare soil/dirt/pavement/sand, and water.

As our task was unsupervised, we used the scenes with LCZ label data as our test set, such that we could evaluate performance, and we trained on the scenes where the information was withheld. The training and testing extents, depicted in Figure 1, are completely separate from one another; thus, we only evaluated scenes not used in the training process. One of these scenes, over Berlin, is the one we show as reference. On top of providing the mapped clusters for areas where the labels exist, we also provided a complete mapping of the scene for visual validation purposes. All labels (except for the bare soil/pavement/dirt/sand label) performed very well. As there was a relatively low number of labeled pixels for the soil/dirt/pavement/sand label, it was hard to conclude why the misclassification happened; however, using the confusion matrix and the mapped images, it appears as if most of these pixels appeared on or near roadways, and were labeled as part of the urban sprawl which, in this context, makes sense. This part of the experiment not only further demonstrates the structural understanding and fusion capabilities of the methodologies, but also indicates that collocated scenes over clear areas can easily be fused, even if they are not within the same temporal range.

The final experiment in this category used Hyperion data. Hyperion is a hyperspectral imager that flew aboard the Earth Observing-1 (EO-1) satellite. Using a single instrument is not traditionally thought of as data fusion; however, using the full set of Hyperion's $>$200 channels is akin to data fusion, given the shear number of channels used for a single scene; it is a tangential use-case that is somewhat fascinating. The training and testing extents are global and completely separate, as seen in Figure 1. We do not have label data for this experiment, and only carried out imagery and cluster analysis.

### *2.4. Fine Scale Evaluation in Select Scenes*

The second set of experiments were intended to show what the improvement in large-scale agreement and cluster performance meant, in terms of finer-scale structural understanding of the data sets with large class imbalances. For this goal, we chose to look at fire and smoke detection in both MISR and MODIS, two instruments used in the previous experimental set, and two airborne instruments, MASTER and eMAS. In both cases, there are pre-existing fire detection products to be compared against. For fire and smoke, when a pre-existing product is available, a first pass is conducted with the automated mapping procedure described above. As these are finer-scale evaluations, a manual mapping process was also performed, in order to ensure that no detections were missed that the pre-existing product may not contain, but that our product did. Smoke detection (for the most part), as well as all burn scar detection was qualitative in this study, but will be further looked into in future work.

The airborne instruments whose data was used are the MODIS/ASTER airborne simulator (MASTER) and the Enhanced MODIS Airborne Simulator (eMAS). MASTER is an airborne imager, which was aboard a DC-8 aircraft for the scenes tested; it has a spatial resolution of 10–30 m/pixel with 50 spectral bands [26]. eMAS is another airborne imager aboard the high-altitude ER-2 aircraft. The eMAS instrument has 38 spectral bands and a

spatial resolution of 50 m [27]. Data from these instruments were used separately as well as together, generating a MASTER/eMAS fusion product. Both eMAS and MASTER have pre-existing fire-detection products, which were generated using the same algorithms as MODIS. The training and testing extents for MASTER, eMAS, and MASTER eMAS fusion can be seen in Figure 1. The MASTER, eMAS, and MASTER + eMAS fusion RBMs and clustering were generated and parameterized in the same way as the other models in this study, but no full-scene classification data sets are available for these instruments; hence, they are only included in this section. We show the evaluation of all three RBM-based products as well as the raw product, in this case, as these data sets were not a part of the initial experimental set above.

Over the two scenes that were almost spatiotemporally collocated for fusion, the two fire detection products were compared. The MASTER fire detection product was resampled to eMAS resolution and then quantitatively compared. The training and testing extents can be seen in Figure 1.

## 3. Results

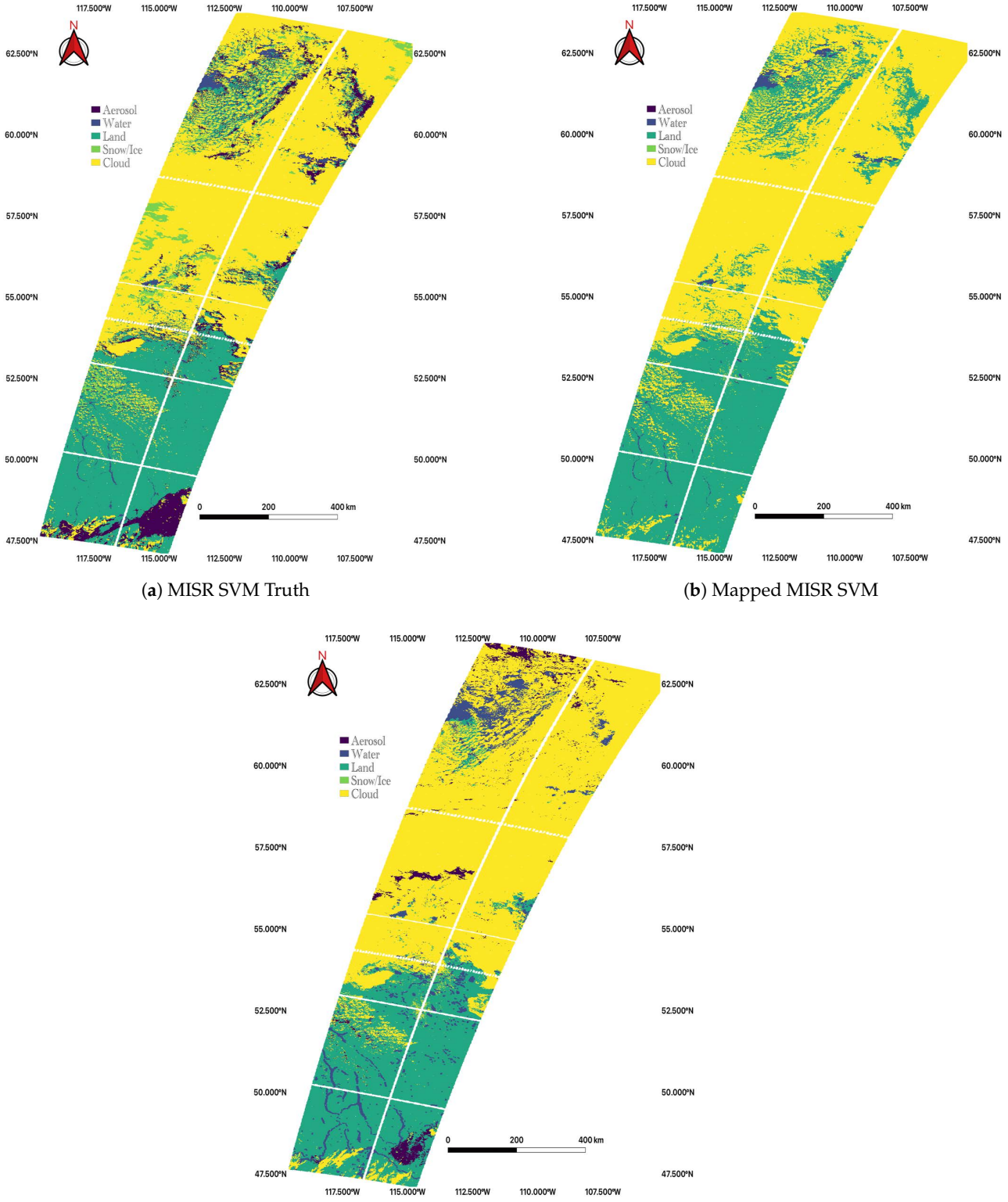
### *3.1. Large-Scale Coarse Full Scene Evaluations*

A summary of the evaluations made for each experiment can be seen in Table 3.

**Table 3.** Summary of overall performance comparison. Train N is the number of pixels used in training. Test N is the number of pixels used in testing. Training iterations is the number of iterations each RBM took to converge (N/A is put in places where no RBM was used, for comparison purposes). Training time is the number of hours taken to reach convergence for each RBM. Agreement % is the percent to which the mapped clusters agreed with pre-existing products. In the case of comparison of multiple products, the agreement % was chosen for the product that best represented the scenes tested, but all are shown in further analysis. Balanced agreement is an agreement percentage that inherently accounts for class imbalances in the data sets to pre-existing products. Cluster score is the Davies–Bouldin index value for the clusters generated. Lower Davies–Bouldin index values indicate better clustering performance.

| Data Set | Train N | Test N | # Training Iterations | Training Time (Hours) | Agreement % | Balanced Agreement % | Cluster Score |
|---|---|---|---|---|---|---|---|
| MISR 1-Camera | 1850148 | 1845900 | 6 | 29.5 | 88.4 | 68.7 | 2.9 |
| MISR 9-Camers Fusion | 1850148 | 1845900 | 4 | 24.3 | 93.3 | 73.1 | 2.4 |
| MODIS | 1850148 | 1845900 | 10 | 33.5 | 86.7 | 70.0 | 3.0 |
| MISR 9-Camera + MODIS Fusion | 1850148 | 1845900 | 6 | 34.2 | 97.2 | 77.0 | 2.1 |
| MISR 9-Camera + MODIS Fusion Raw | 1850148 | 1845900 | NA | NA | 95.0 | 72.4 | 3.2 |
| Landsat-8 | 2186449 | 81827 | 7 | 33.7 | 90.7 | 74.1 | 2.5 |
| Sentinel-2 | 2186449 | 81827 | 6 | 30.2 | 88.1 | 74.2 | 2.7 |
| Landsat-8 + Sentinel-2 Fusion | 2186449 | 81827 | 5 | 31.8 | 93.0 | 78.9 | 2.5 |
| Landsat-8 + Sentinel-2 Fusion Raw | 2186449 | 81827 | N/A | N/A | 91.8 | 77.5 | 48.1 |
| Hyperion | 2567686 | 2426716 | 4 | 26.1 | N/A | N/A | 2.0 |
| Hyperion Raw | 2567686 | 2426716 | N/A | N/A | N/A | N/A | 4.1 |

Within the experiment using MISR and MODIS, before we compared our data sets to the MODIS cloud mask and MISR SVM classification product, we compared them against each other. An example of this over the reference scene can be seen in Figure 2. As shown in the confusion matrices in Table 4, agreement was observed almost in full within land and cloud pixels, as well as a majority of water pixels. Due to this, and the issues noted above, for most classes evaluated, the MISR SVM was the product most associated with the ground truth, and we also focused on the desert/land distinction coming from the MODIS cloud mask. However, all class agreement was provided within the confusion matrices for completeness.




(**a**) MISR SVM Truth

(**b**) Mapped MISR SVM

(**c**) Actual MODIS Cloud Mask

**Figure 2.** MISR SVM Mapped to MODIS Cloud Mask (**a**–**c**).

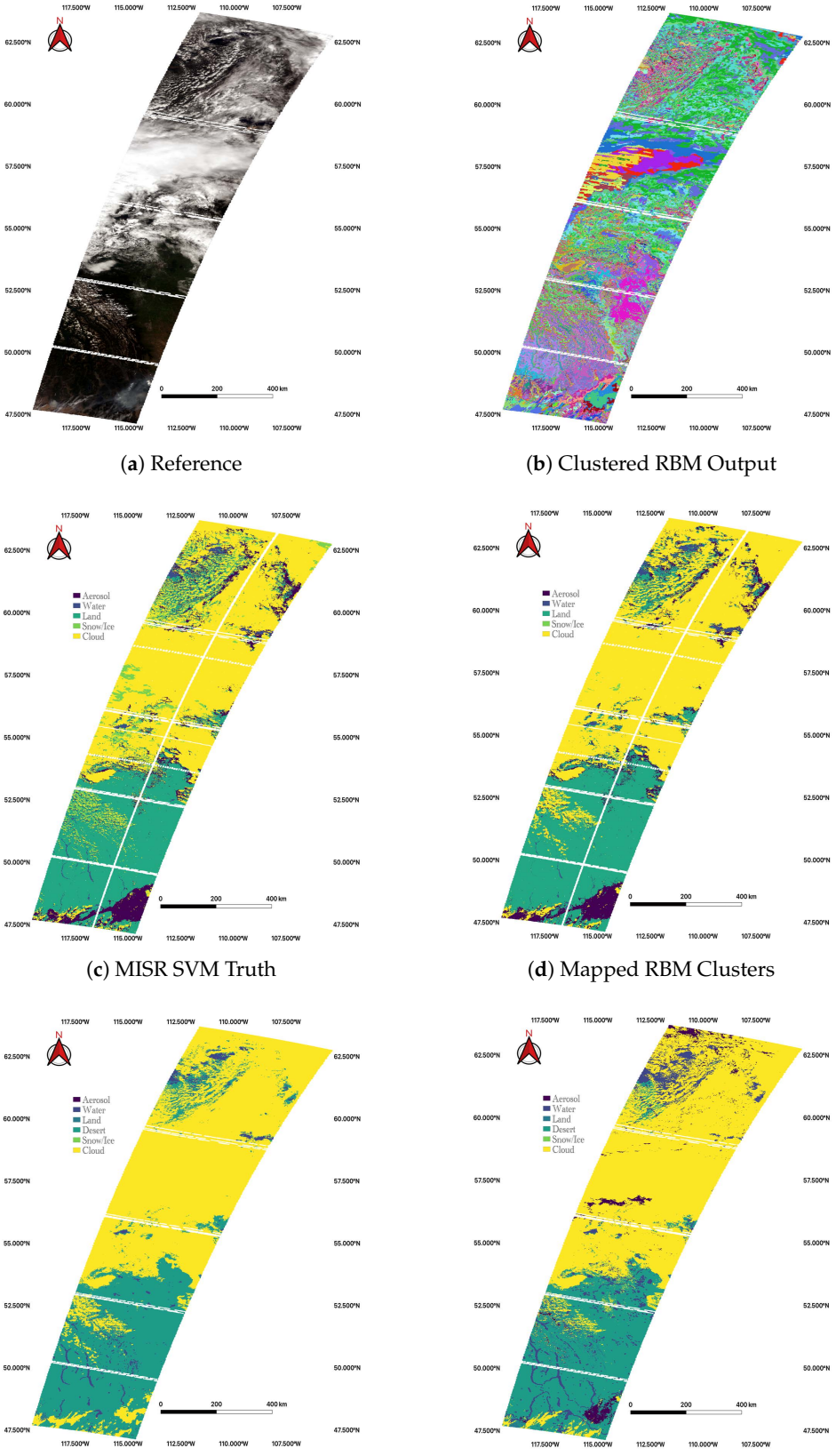
**Table 4.** Confusion matrix MISR SVM vs. MODIS cloud mask. The top section of the table shows comparisons of pixel counts, and the bottom section shows comparison of %. Column names are in bold and placed underneath the hat character '^' to denote that they are predicted labels, and to differentiate them from the pre-existing product labels along the rows, with no special formatting.

| N = 1948222 | Aerôsol | Waîer | Lând | Sñow | Clôud |
|---|---|---|---|---|---|
| Aerosol | 0 | 487 | 8358 | 0 | 29103 |
| Water | 0 | 157930 | 98087 | 0 | 8591 |
| Land | 0 | 2673 | 992812 | 0 | 16203 |
| Snow | 0 | 0 | 0 | 0 | 0 |
| Cloud | 0 | 8470 | 76085 | 0 | 549423 |
| Aerosol | 0.0 | 1.3 | 22.0 | 0.0 | 76.7 |
| Water | 0.0 | 59.7 | 37.1 | 0.0 | 3.2 |
| Land | 0.0 | 0.3 | 98.1 | 0.0 | 1.6 |
| Snow | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Cloud | 0.0 | 1.3 | 12.0 | 0.0 | 86.7 |

As shown in Table 5, our product performed very well when compared to the MISR SVM classifier product, especially with land, water, and clouds. The agreement for aerosols was also high, and we believe the minor degradation here is due to something similar to the effects shown in the next subsection, where both data sets identified parts of entities, such as aerosol plumes, with finer details, while overlapping on the majority, but detecting separate parts of the whole. When compared against the MODIS cloud mask (in Table 6), large-scale agreement was found in land, desert, and clouds. Water was agreed upon in most cases, but the disagreement was likely due to a mixture of fine-scale detail in the MODIS cloud mask, as well as seasonal water being included. The agreement on water between our data and the MODIS cloud mask was increased from that of the comparison between the MISR SVM and MODIS cloud mask, which, when paired with visual verification, as in Figure 3, indicated an increase in fine-scale water bodies being picked up within our product.

**Table 5.** Confusion matrix MISR 9-Camera MODIS Fusion RBM vs. MISR SVM. The top section of the table shows comparisons of pixel counts, and the bottom section shows comparison of %. Column names are in bold and placed underneath the hat character '^' to denote they are the predicted labels, and to differentiate them from the pre-existing product labels along the rows, with no special formatting.

| N = 1845900 | Aerôsol | Waîer | Lând | Sñow | Clôud |
|---|---|---|---|---|---|
| Aerosol | 61248 | 582 | 14732 | 11 | 6094 |
| Water | 475 | 151643 | 9477 | 1 | 3502 |
| Land | 8888 | 3908 | 1001212 | 30 | 24437 |
| Snow | 480 | 18 | 2443 | 1529 | 26375 |
| Cloud | 6268 | 4729 | 28211 | 622 | 488916 |
| Aerosol | 74.1 | 0.7 | 17.8 | 0.0 | 7.4 |
| Water | 0.3 | 91.9 | 5.7 | 0.0 | 2.1 |
| Land | 0.9 | 0.4 | 96.4 | 0.0 | 2.4 |
| Snow | 1.6 | 0.0 | 7.9 | 5.0 | 85.5 |
| Cloud | 1.2 | 0.9 | 5.3 | 0.1 | 92.5 |




(**a**) Reference

(**b**) Clustered RBM Output

(**c**) MISR SVM Truth

(**d**) Mapped RBM Clusters

(**e**) MODIS Cloud Mask Truth

(**f**) Mapped RBM Clusters

**Figure 3.** MISR 9-Camera MODIS Fuse Clustered RBM Output Mapped to Pre-Existing Classification Products (**a**–**f**).

**Table 6.** Confusion matrix MISR 9-Camera MODIS fusion RBM vs. MODIS cloud mask. The top section of the table shows comparisons of pixel counts, and the bottom section shows comparison of %. Column names are in bold and placed underneath the hat character '^' to denote they are the predicted labels, and to differentiate them from the pre-existing product labels along the rows, with no special formatting.

| N = 1845900 | Aerôsol | Waîer | Lând | Deŝert | Sn̂ow | Clôud |
|---|---|---|---|---|---|---|
| Aerosol | 10315 | 577 | 195 | 1422 | 0 | 23363 |
| Water | 856 | 180758 | 5717 | 63630 | 0 | 9641 |
| Land | 191 | 1078 | 128486 | 46640 | 0 | 3982 |
| Desert | 1083 | 11000 | 34652 | 746302 | 0 | 15344 |
| Snow | 0 | 0 | 0 | 0 | 0 | 0 |
| Cloud | 2608 | 16747 | 291 | 27752 | 0 | 557515 |
| Aerosol | 28.8 | 1.6 | 0.5 | 4.0 | 0.0 | 6.5 |
| Water | 0.3 | 69.4 | 2.2 | 24.4 | 0.0 | 3.7 |
| Land | 0.1 | 0.6 | 71.2 | 25.9 | 0.0 | 2.2 |
| Desert | 0.1 | 1.4 | 4.3 | 92.3 | 0.0 | 1.9 |
| Snow | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Cloud | 0.4 | 2.8 | 0.1 | 4.6 | 0.0 | 92.2 |

In terms of agreement, as well as clustering performance, the raw version of the fusion product performed slightly worse than the RBM-based version of the product (in Tables 7 and 8). There was a slight uptick of agreement in the cloud class, when compared to both pre-existing products; however, in terms of overall agreement and balanced agreement, the raw product performed worse. These changes in performance were slight but, as we will see through the finer-scale investigations in the next subsection, they are indicators of a less precise understanding of the data's structure.

There appeared to have been no real increase in training complexity or time required when fusing MISR and MODIS together. In fact, Figure 4 shows that the MISR 9-Camera and MODIS fusion RBM converged better, and required less iterations than when training with MODIS data alone.

**Table 7.** Confusion matrix MISR 9-Camera MODIS fusion raw vs. MISR SVM. The top section of the table shows comparisons of pixel counts, and the bottom section shows comparison of %. Column names are in bold and placed underneath the hat character '^' to denote they are the predicted labels, and to differentiate them from the pre-existing product labels along the rows, with no special formatting.

| N = 1845900 | Aerôsol | Waîer | Lând | Sn̂ow | Clôud |
|---|---|---|---|---|---|
| Aerosol | 57868 | 788 | 16919 | 0 | 7092 |
| Water | 253 | 149412 | 9551 | 2 | 5880 |
| Land | 5737 | 3708 | 994747 | 16 | 34267 |
| Snow | 142 | 4 | 2034 | 985 | 27680 |
| Cloud | 2128 | 2716 | 27265 | 409 | 496228 |
| Aerosol | 70.0 | 1.0 | 20.4 | 0.0 | 8.6 |
| Water | 0.2 | 90.5 | 5.8 | 0.0 | 4.6 |
| Land | 0.6 | 0.4 | 95.8 | 0.0 | 3.3 |
| Snow | 0.5 | 0.0 | 6.6 | 3.2 | 89.7 |
| Cloud | 0.4 | 0.5 | 5.2 | 0.0 | 93.8 |

**Table 8.** Confusion matrix MISR 9-Camera MODIS fusion raw vs. MODIS cloud mask. The top section of the table shows comparisons of pixel counts, and the bottom section shows comparison of %. Column names are in bold and placed underneath the hat character '^' to denote they are the predicted labels, and to differentiate them from the pre-existing product labels along the rows, with no special formatting.

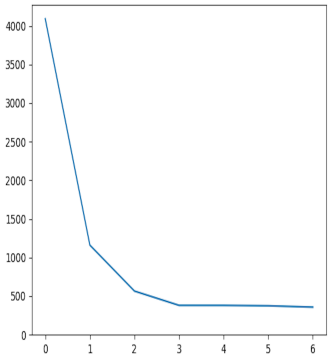
| N = 1845900 | Aerôsol | Waîer | Lând | Deŝert | Sn̂ow | Clôud |
|---|---|---|---|---|---|---|
| Aerosol | 3048 | 366 | 300 | 3050 | 0 | 29108 |
| Water | 253 | 173668 | 9027 | 64295 | 0 | 13359 |
| Land | 152 | 440 | 135881 | 35062 | 0 | 8842 |
| Desert | 123 | 8890 | 43911 | 734557 | 0 | 21100 |
| Snow | 0 | 0 | 0 | 0 | 0 | 0 |
| Cloud | 801 | 8616 | 1603 | 30458 | 0 | 563435 |
| Aerosol | 8.5 | 1.0 | 0.8 | 8.5 | 0.0 | 81.1 |
| Water | 0.1 | 66.6 | 3.5 | 24.6 | 0.0 | 5.1 |
| Land | 0.1 | 0.2 | 75.3 | 19.4 | 0.0 | 4.9 |
| Desert | 0.0 | 1.1 | 5.4 | 90.8 | 0.0 | 2.6 |
| Snow | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Cloud | 0.1 | 14.2 | 0.3 | 5.0 | 0.0 | 93.1 |

For the second coarse-scale experiment, we evaluated RBM-based products for Landsat-8, Sentinel-2, and the fusion of the two, as well as a raw clustering of the combined data (in Figure 5).

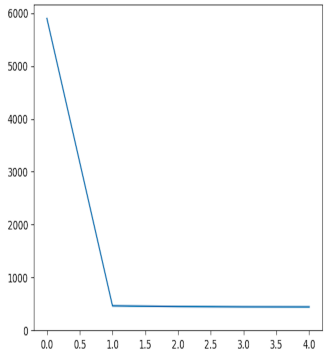
As with the previous experiment, the raw clustering product performed worse than that of the RBM-based product. In this case, the RBM-based product outperformed over all classes, shown in Tables 9 and 10 and Figure 6, as well as agreement and balanced agreement.

Once again, there did not appear to be much trade-off, in terms of the number of iterations or training time, when comparing the effort needed for the singular instrument RBMs against the fused one, as seen in Figure 6. As with MODIS, Sentinel-2's RBM-based product appeared to produce a product that was a bit more noisy, and the RBM convergence showed a higher reconstruction error, although the Sentinel-2 and MODIS products are still accurate enough to be used in cases where fusion is not possible.

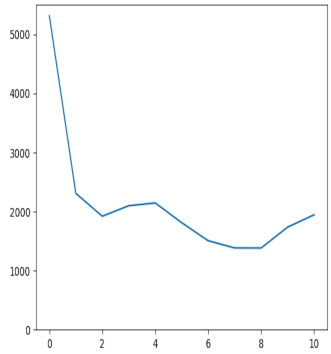
With regard to Hyperion, a slightly better performance was qualitatively observed when using the RBM, compared to only using clustering, as shown in Figure 7. Finally, the training appeared to quickly converge, although the reconstruction error was quite high; however, given the increased number of channels and the overall performance, we believe that it was acceptable. Given the increased channel number, this could also be improved through use of a multi-layer RBM or a deep belief network, but we wanted to keep all the architectures and parameterizations the same for all the models in this study.



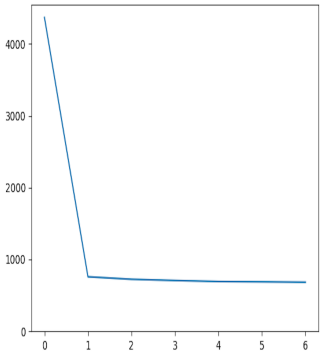

(**a**) MISR 1-Camera Cost Curve



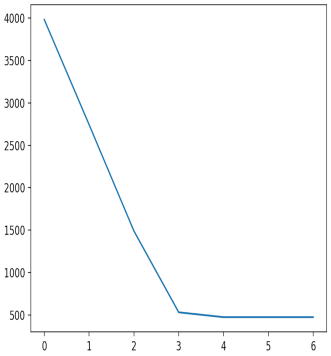

(**b**) MISR 9-Camera Fusion Cost Curve



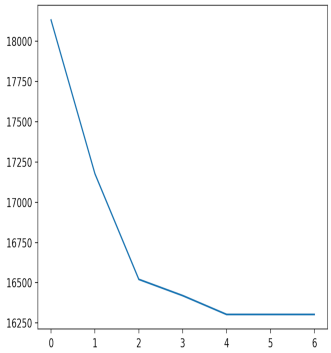

(**c**) MODIS Cost Curve



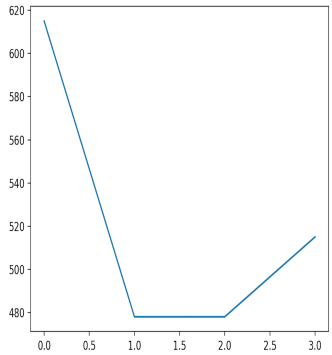

(**d**) MISR 9-Camera MODIS Fusion



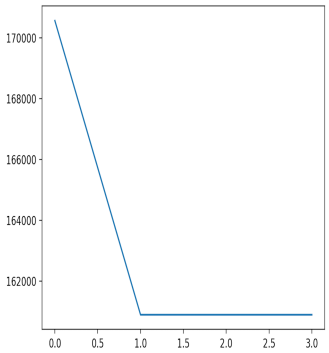

(**e**) Landsat-8 Cost Curve




(**f**) Sentinel-2 Fusion Cost Curve

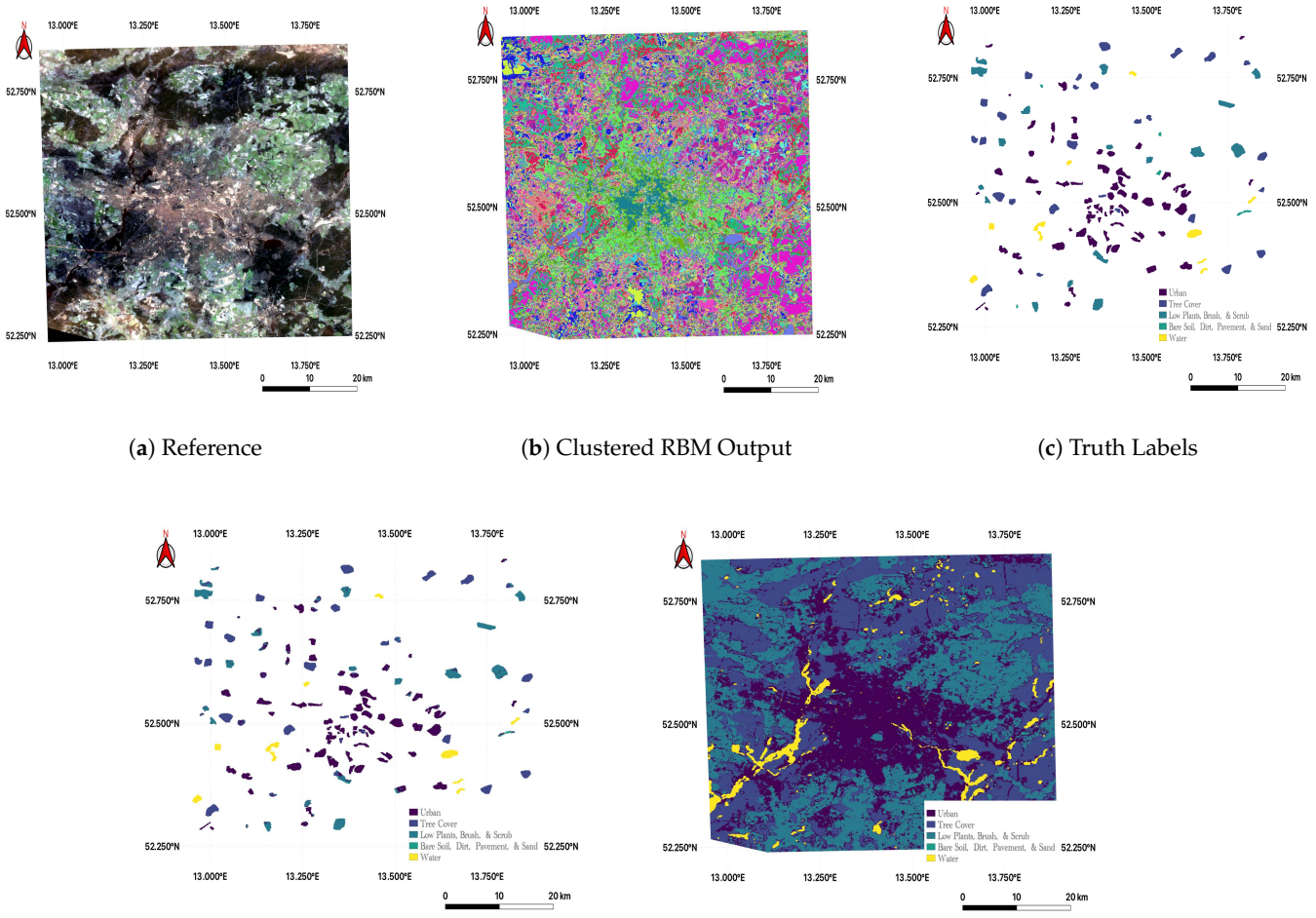



(**g**) Landsat-8 Sentinel-2 Fusion Cost Curve




(**h**) Hyperion Cost Curve

**Figure 4.** Cost Curves for All Coarse-Scale Experiments. Where applicable, the cost curve for single instruments' RBMs as well as the fusion RBM is provided, for comparison (**a**–**h**). Neither fusion RBM incurs a significant cost or processing time increase. This means it is not more costly, generally, to use the fusion output where appropriate.




(**a**) Reference (**b**) Clustered RBM Output (**c**) Truth Labels

(**d**) Mapped RBM Clusters (**e**) Expanded Mapped Clusters

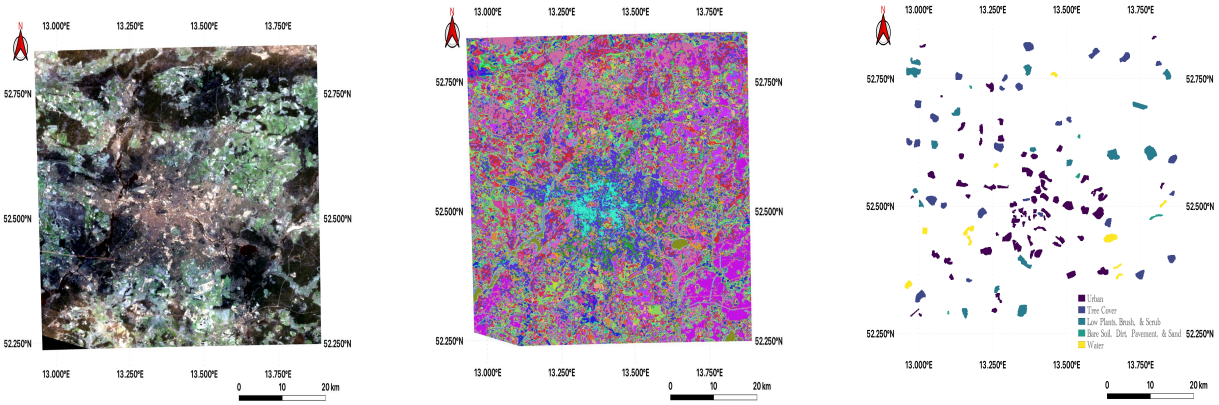
**Figure 5.** Landsat-8 Sentinel-2 Fusion Clustered RBM Output Mapped to Labels from IEEE GRSS Data Fusion Challenge 2017 (**a**–**e**).

**Table 9.** Confusion matrix Landsat-8 + Sentinel-2 fusion RBM vs. IEEE GRSS Data Fusion Challenge 2017 labels. The top section of the table shows comparisons of pixel counts, and the bottom section shows comparison of %. Column names are in bold and placed underneath the hat character '^' to denote they are the predicted labels, and to differentiate them from the pre-existing product labels along the rows, with no special formatting.

| **N = 81827** | **Urb̂an** | **Tree Ĉover** | **Low Plaǹts, Brush, and Scrub** | **Bare Sôil, Dirt, Pavement, and Sand** | **Waẗer** |
|---|---|---|---|---|---|
| Urban | 33768 | 850 | 949 | 93 | 59 |
| Tree Cover | 650 | 19073 | 789 | 0 | 23 |
| Low plants, brush, and scrub | 707 | 884 | 14575 | 13 | 19 |
| Bare soil, dirt, pavement, and sand | 533 | 6 | 136 | 148 | 3 |
| Water | 26 | 10 | 13 | 0 | 8497 |
| Urban | 94.5 | 2.4 | 2.7 | 0.3 | 0.2 |
| Tree Cover | 3.2 | 92.9 | 3.8 | 0.0 | 0.1 |
| Low plants, brush, and scrub | 4.4 | 5.5 | 90.0 | 0.0 | 0.1 |
| Bare soil, dirt, pavement, and sand | 64.5 | 0.7 | 16.5 | 17.9 | 0.4 |
| Water | 0.3 | 0.1 | 0.2 | 0.0 | 99.4 |

**Table 10.** Confusion matrix Landsat-8 + Sentinel-2 fusion raw vs. IEEE GRSS Data Fusion Challenge 2017 labels. The top section of the table shows comparisons of pixel counts, and the bottom section shows comparison of %. Column names are in bold and placed underneath the hat character '^' to denote they are the predicted labels, and to differentiate them from the pre-existing product labels along the rows, with no special formatting.

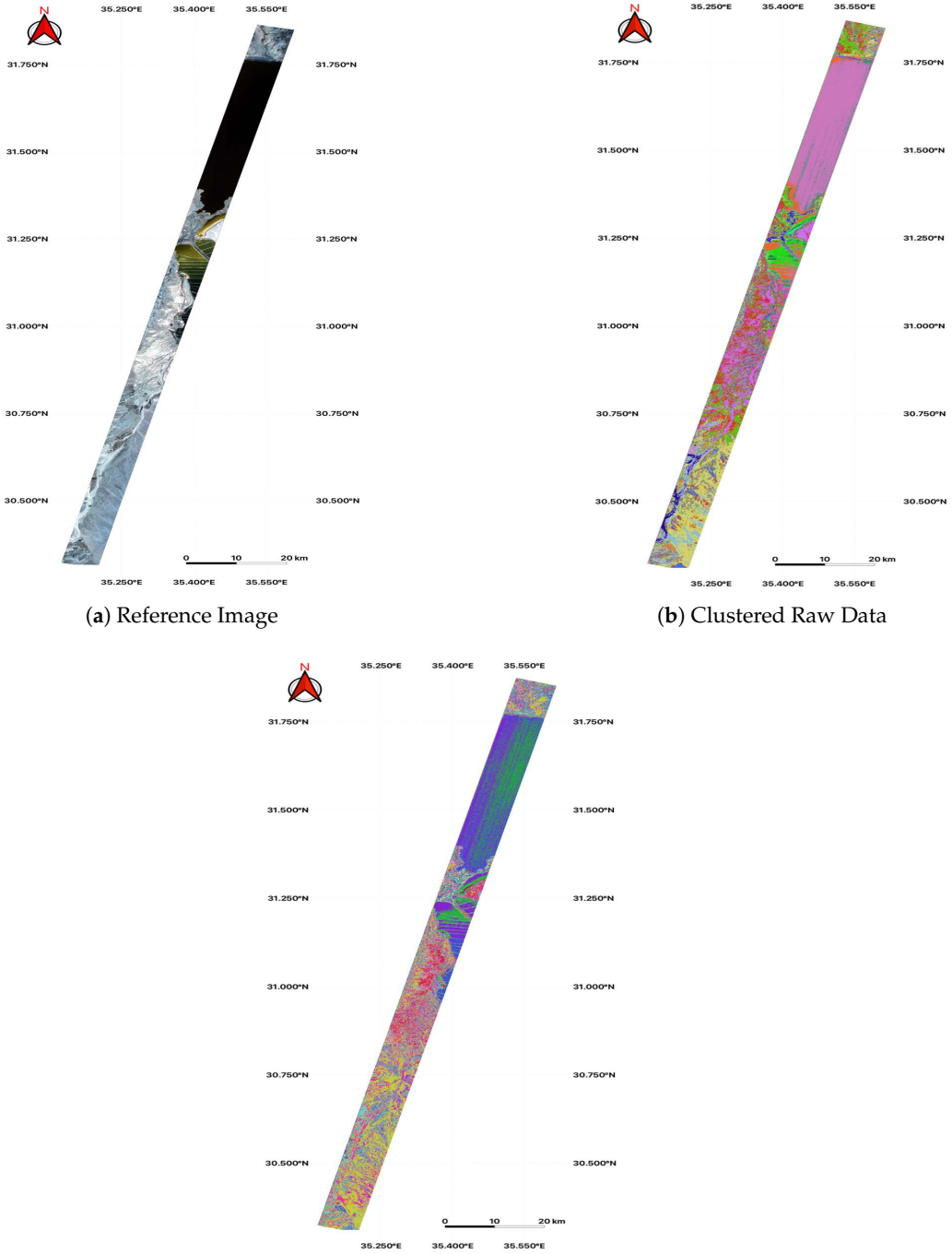
| N = 81827 | Urb̂an | Tree Ĉover | Low Plaǹts, Brush, and Scrub | Bare Sôil, Dirt, Pavement, and Sand | Waẗer |
|---|---|---|---|---|---|
| Urban | 33610 | 965 | 1033 | 46 | 65 |
| Tree Cover | 663 | 18573 | 1283 | 0 | 16 |
| Low plants, brush, and scrub | 920 | 872 | 14363 | 34 | 9 |
| Bare soil, dirt, pavement, and sand | 920 | 6 | 115 | 123 | 0 |
| Water | 45 | 4 | 0 | 0 | 8497 |
| Urban | 94.1 | 2.7 | 2.9 | 0.3 | 0.2 |
| Tree Cover | 3.2 | 90.4 | 6.2 | 0.0 | 0.0 |
| Low plants, brush, and scrub | 5.7 | 5.4 | 88.7 | 0.2 | 0.0 |
| Bare soil, dirt, pavement, and sand | 70.5 | 0.7 | 13.9 | 14.9 | 0.0 |
| Water | 0.5 | 0.0 | 0.0 | 0.0 | 99.4 |




(**a**) Reference (**b**) Clustered Output (**c**) Truth Labels

(**d**) Mapped Clusters (**e**) Expanded Mapped Clusters

**Figure 6.** Landsat-8 Sentinel-2 Fusion Clustered Raw Data Mapped to Labels from IEEE GRSS Data Fusion Challenge 2017 (**a**–**e**).




(**a**) Reference Image

(**b**) Clustered Raw Data

(**c**) Clustered RBM Output

**Figure 7.** Hyperion Reference Image, Clustered Raw Data, and Clustered RBM Output (**a**–**c**).

### *3.2. Fine Scale Evaluation in Select Scenes*

The summary of the fire detection comparison results can be seen in Table 11.

**Table 11.** Summary of fire detection comparison to pre-existing products. Total pixel count is the total number of pixels tested. Pre-existing fire pixel count is the number of pixels labeled as fire in the pre-existing product. RBM-based fire pixel count is the number of pixels labeled as fire in our product. The % Agreement is the percentage of fire pixels in the pre-existing product that are also identified as fire in our product. The % False positive is the % of fire pixels in our product that are clearly mislabeled. $\delta$ % True positive is the % change in fire pixel count within the tested pixels from the pre-existing products to our products.

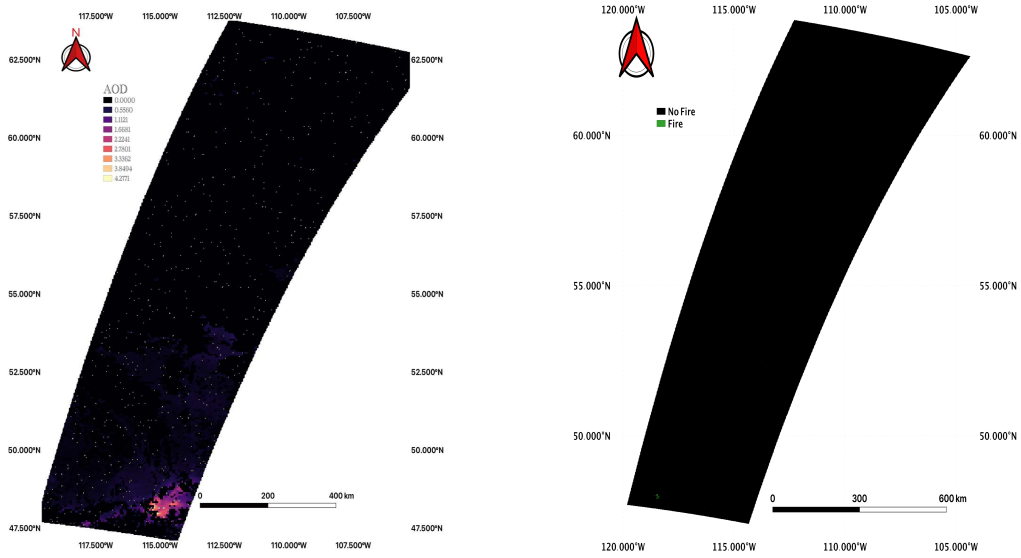
| Data Set | Total Pixel Count | Pre-Existing Product Fire Pixel Count | RBM-Based Fire Pixel Count | % Agreement | % False Positive | $\delta$ % True Positive |
|---|---|---|---|---|---|---|
| MISR + MODIS Fusion | 539345 | 77 | 55 | 33.8 | 0.0 | −28.8 |
| MISR + MODIS Raw Fusion | 539345 | 77 | 12 | 15.3 | 0.0 | −84.4 |
| MASTER | 18255496 | 84861 | 128214 | 90.1 | 0.0 | 151.1 |
| eMAS | 10895688 | 2822 | 13211 | 87.5 | 1.9 | 468.1 |
| eMAS + MASTER Fusion (eMAS) | 2492030 | 174 | 15500 | 84.4 | 0.0 | 8908.0 |
| eMAS + MASTER Fusion (MASTER) | 2492030 | 742 | 15500 | 83.5 | 0.0 | 2098.0 |
| eMAS + MASTER Raw Fusion (eMAS) | 2492030 | 174 | 2363 | 75.0 | 0.0 | 1357.5 |
| eMAS + MASTER Raw Fusion (MASTER) | 2492030 | 742 | 2363 | 74.93 | 0.0 | 318.3 |

The RBM-based and raw MISR-9-camera/MODIS fusion products were chosen for evaluation here, as the RBM-based fusion product performed the best, and we wanted to continue to answer the question regarding the need for the RBM for structural understanding. The scene used as reference above is the same one used for fire detection evaluation, as there was a large fire in the southwest region of the scene. We used a few validation methodologies. Within the aerosol classification in the MISR SVM, there is also a certainty measurement for dust, smoke, or other. Within the 74.1% agreement of overall aerosols, we were also able to attain 83.4% agreement on the smoke subclass, when compared to our smoke mask. With the raw product, we were only able to achieve 78.6% agreement with the smoke subclass. When our fire mask was compared to the operational MODIS fire mask [28], there was only a 33.8% agreement and a 28.8% reduction in pixel count. The pixels contained in our fire mask, but not in the MODIS fire mask appeared to be true fire pixels, and the remaining pixels that the MODIS fire mask identified were identified as smoke pixels in our product. The reference data can be seen in Figure 8 and the smoke and fire masks from the RBM-based fusion product can be seen in Figure 9. With the raw product, there was only a 15.3% agreement, with an 84.4% decline in identified pixels. The 12 fire pixels in the raw product were all accounted for within the MODIS fire mask. These can be seen in Figure 10.

The next experiment performed utilized the fusion of MASTER and eMAS. The first step performed was to compare the pre-existing fire products for each instrument against one another. Within the scenes evaluated, there was a 7.1% agreement between the two fire detection products, with a 50% increase in fire detection pixels in the MASTER product, compared to the eMAS product.

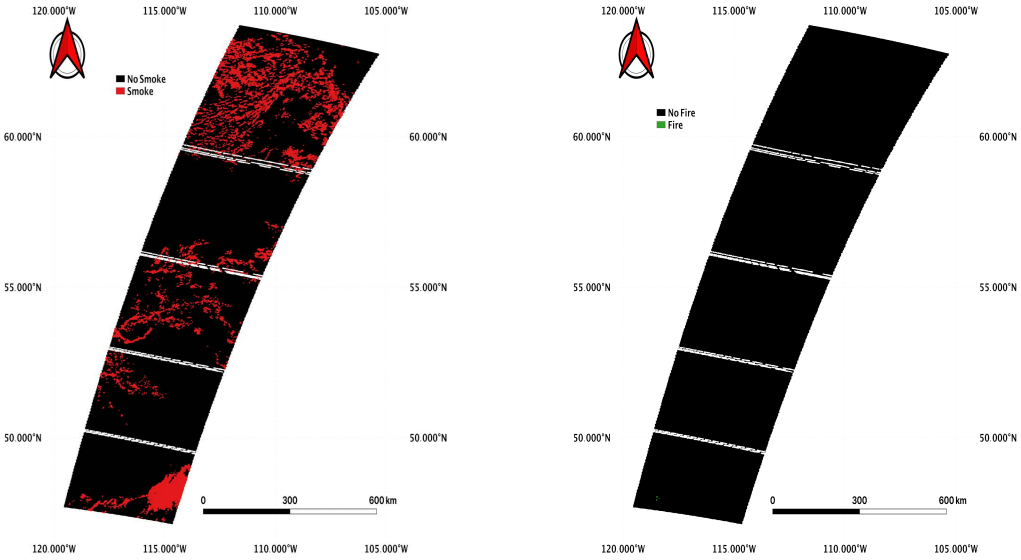
Reference imagery and existing fire products can be seen in Figure 11. Examples of the MASTER RBM-based products for the same scene can be seen in Figure 12. With respect to the MASTER RBM-based fire mask, when compared to the existing MASTER fire detection product at MASTER's native spatial resolution, there was 90.1% agreement, with the RBM-based fire mask having a 151.1% increase in fire-labeled pixel count. While the RBM-based fire mask did not identify all of the same pixels that the existing product did, it appeared to still correctly identify areas where the fire was burning, especially to the southwest of the main hot spot. The smoke mask appeared to correctly identify areas highly inundated with smoke. However, sun glinting off of water seemed to be misclassified as

smoke (see the dark green pixels in Figure 11a). With regards to burn scar detection, the MASTER RBM product was not able to clearly distinguish the burn scar from the rest of the scene.



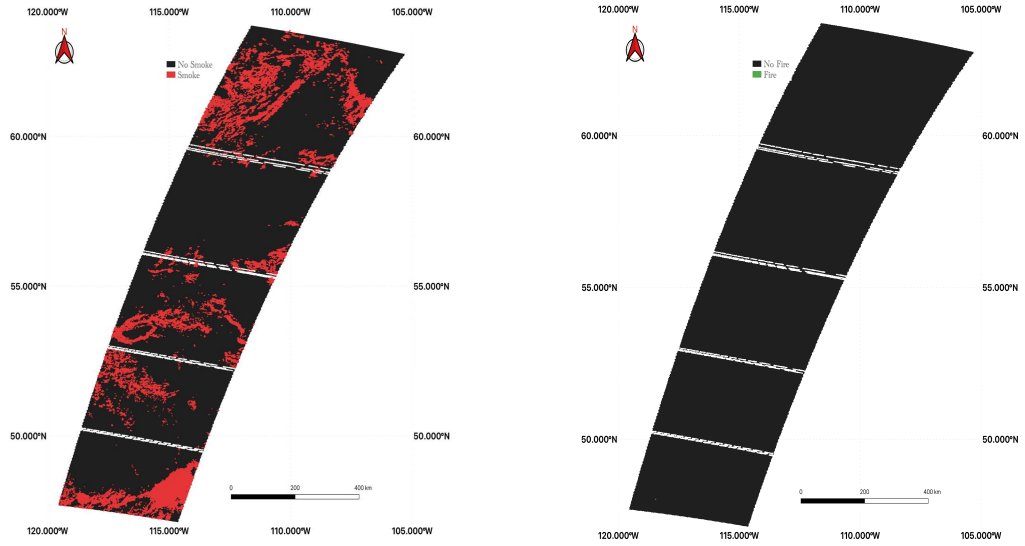

(**a**) Reference AOD (**b**) Reference Fire Mask

**Figure 8.** MISR MODIS fuse reference images (**a**–**b**).



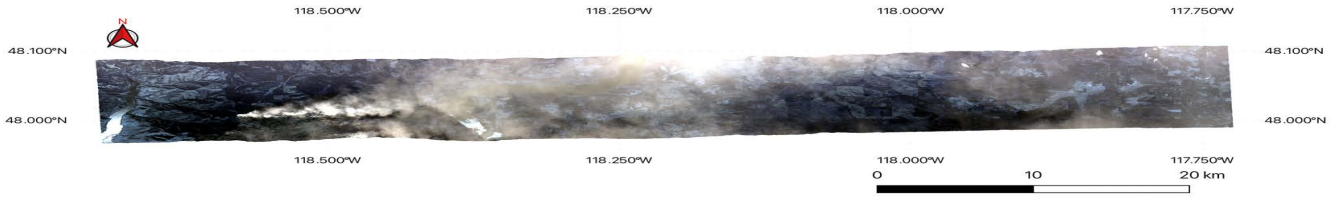

(**a**) RBM-based Smoke Mask (**b**) RBM-based Fire Mask

**Figure 9.** MISR MODIS fuse clustered RBM output, and generated smoke and fire masks (**a**–**b**).




(**a**) Raw Smoke Mask (**b**) Raw Fire Mask

**Figure 10.** MISR MODIS fuse clustered raw data, and generated smoke and fire masks (**a**–**b**).

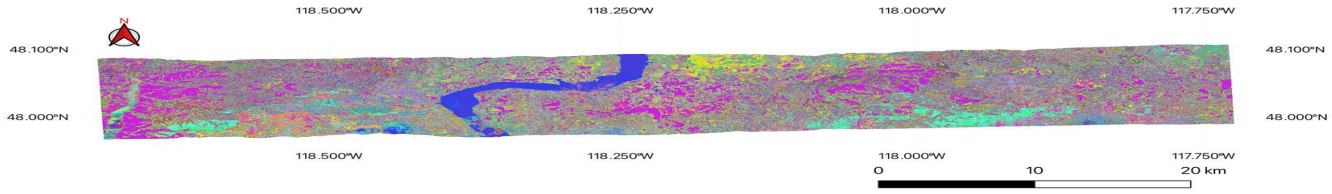



(**a**) Reference

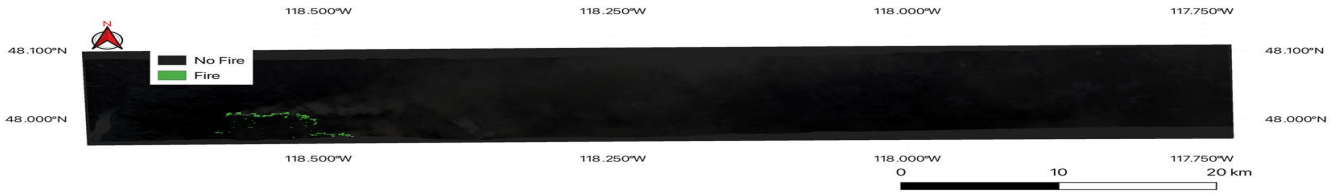

(**b**) Reference Fire Mask

**Figure 11.** MASTER reference images (**a**–**b**).

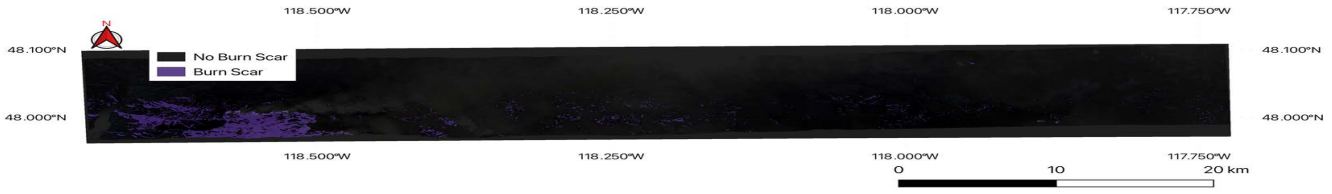
The reference imagery and pre-existing fire mask for eMAS can be found in Figure 13, and the associated RBM-based products can be found in Figure 14. The fire mask from the eMAS-based clustered RBM output had an 87.5% agreement with the existing fire detection product, at the eMAS native spatial resolution, but a 468.1% increase in number of pixels labeled as fire. Of the pixels labeled as fire in the RBM-based fire mask, 1.9% appeared to be false positives. The eMAS RBM output, like the MASTER RBM output, was not able to clearly distinguish the burn scar. The smoke mask reasonably identified areas inundated with smoke, although it seems that sun glinting off of water in the eMAS scenes was also incorrectly identified as smoke. The eMAS RBM-based products did not segment the burn scar well, similar to the MASTER RBM-based output.




(**a**) Clustered RBM Output



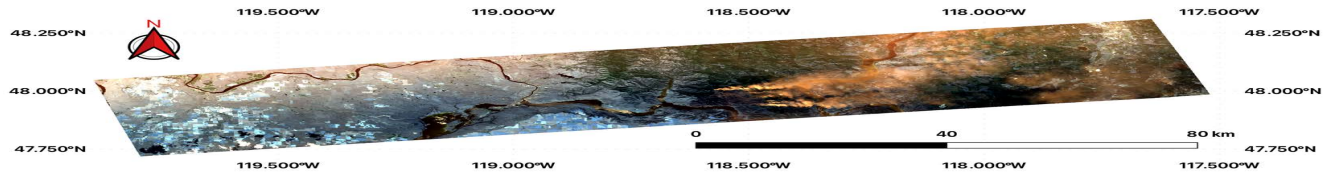

(**b**) RBM-based Fire Mask



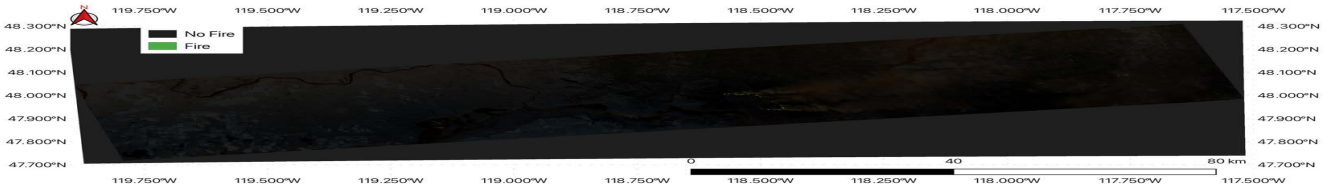

(**c**) RBM-based Fire Mask with Burn Scar


(**d**) RBM-based Smoke Mask

**Figure 12.** MASTER RBM clustered output and associated fire and smoke masks (**a**–**d**).

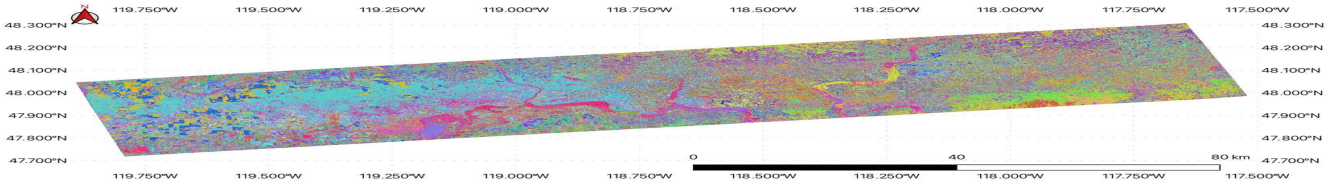



(**a**) Reference

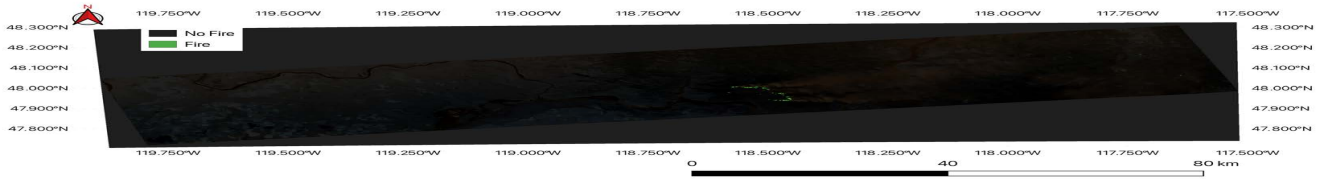



(**b**) Reference Fire Mask

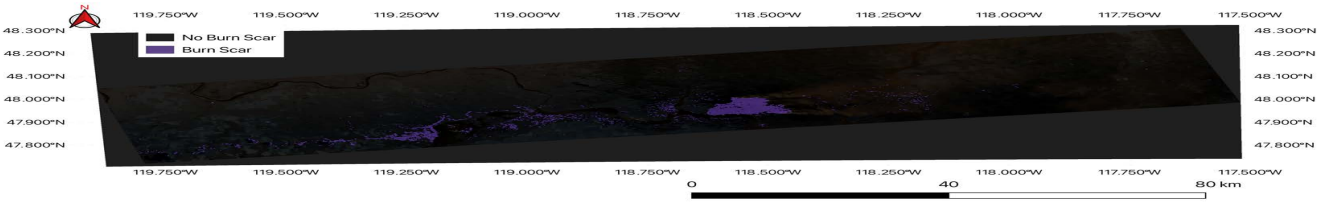
**Figure 13.** eMAS reference images (**a**–**b**).



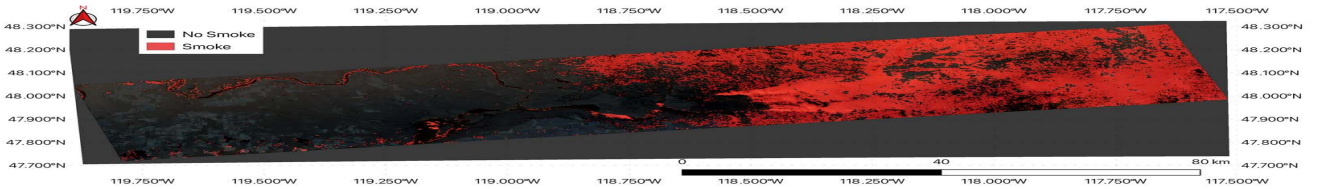

(**a**) Clustered RBM Output




(**b**) RBM-based Fire Mask



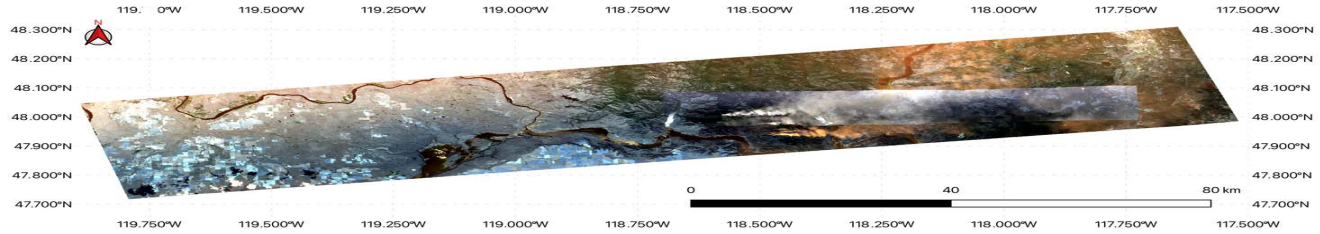

(**c**) RBM-based Fire Mask with Burn Scar



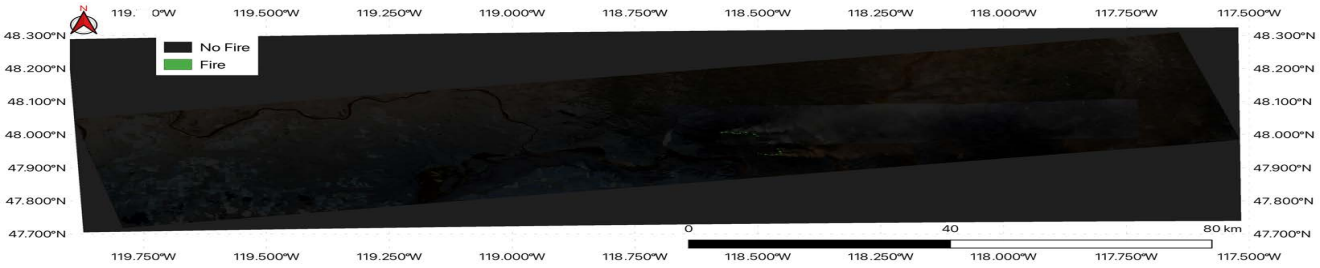

(**d**) RBM-based Smoke Mask

**Figure 14.** eMAS RBM clustered output and associated fire and smoke masks (**a**–**d**).

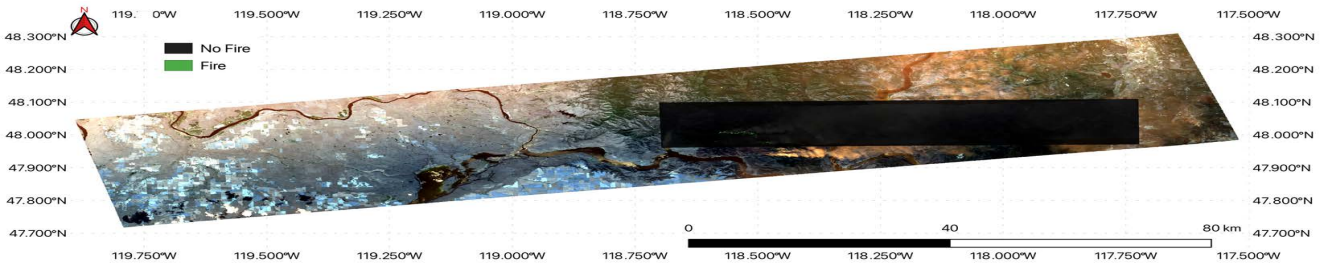
The last part of the experiment using MASTER and eMAS involved generating a fusion product, with the MASTER data being resampled to eMAS resolution over the two collocated scenes, as shown in Figure 15a. The number of scenes spatiotemporally appropriate for fusion was far smaller than when using the instrument data sets separately, but there were still more than enough pixels to evaluate the efficacy. The training and testing extents can be seen in Figure 15. With data from both instruments, the resulting fire mask had an 84.4% agreement with MASTER's pre-existing fire detection product, and an 83.5% agreement with that of eMAS. The RBM-based mask also had a 2098% increase in pixel count from the MASTER fire detection product and an 8908% increase from the eMAS product, seen in Figure 16. Although the pixel count increase was steep, the fire mask correctly identified areas that were burning, and uniquely identified smoldering fire, not just the active fire. Additionally, with data of both instruments, the fusion product was able to properly identify the burn scar more clearly and reasonably, with a small number of false positives. Finally, the smoke mask generated again appeared to correctly identify areas inundated with smoke. However, the sun glinting on water in the western end of the domain was still identified as smoke. The smoke masks generated here are of significant value, and we will examine the remaining issue of separating sun glint from smoke in future work.



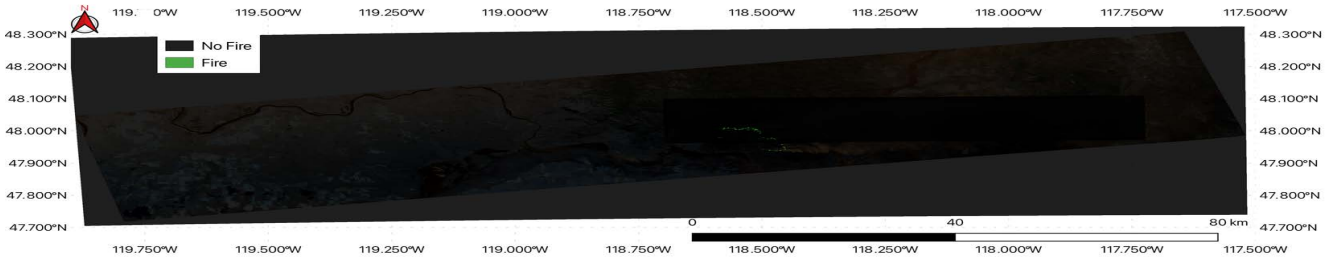

(**a**) Reference




(**b**) Reference with eMAS Reference Fire Mask

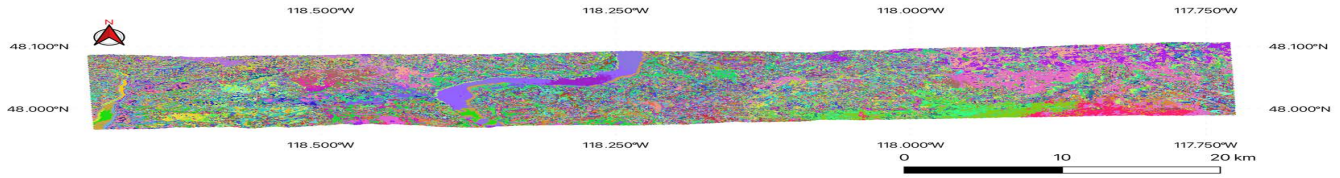



(**c**) Reference with MASTER Reference Fire Mask




(**d**) Reference with Both Reference Fire Masks

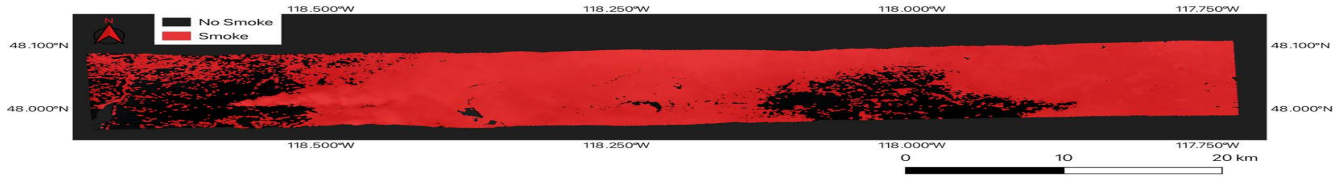
**Figure 15.** MASTER eMAS fuse reference images (**a**–**d**).




(**a**) Clustered RBM Output


(**b**) RBM-based Fire Mask

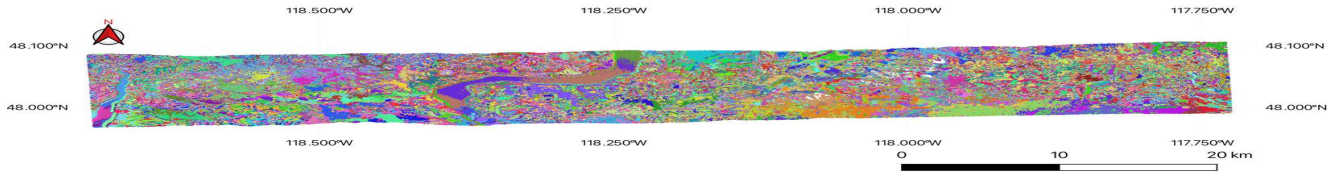

(**c**) RBM-based Fire Mask with Burn Scar



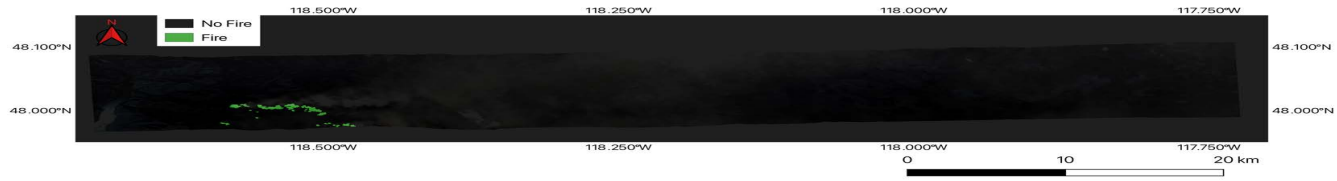

(**d**) RBM-based Smoke Mask

**Figure 16.** MASTER eMAS fuse clustered RBM output and generated smoke and fire masks (**a**–**d**).

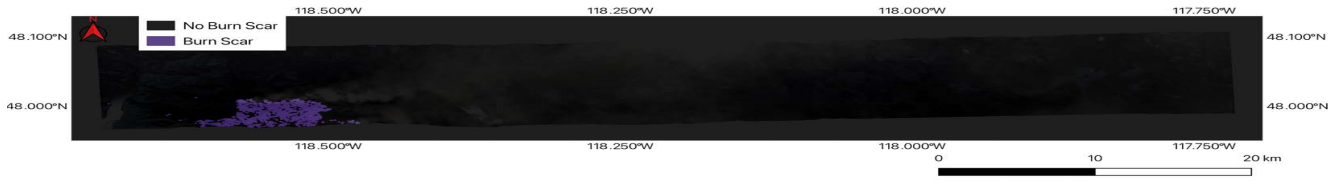
The raw clustering products were also degraded in this last case. There were still no false positives in the fire detection, but agreement fell between this product and the eMAS and MASTER fire masks, to 75% and 74.9%, respectively; there were far less fire pixels detected, the burn scar detection contained more false positives and less of the actual burn scar, and the smoke mask appeared to identify more glint and clear areas. These can be seen in Figure 17.



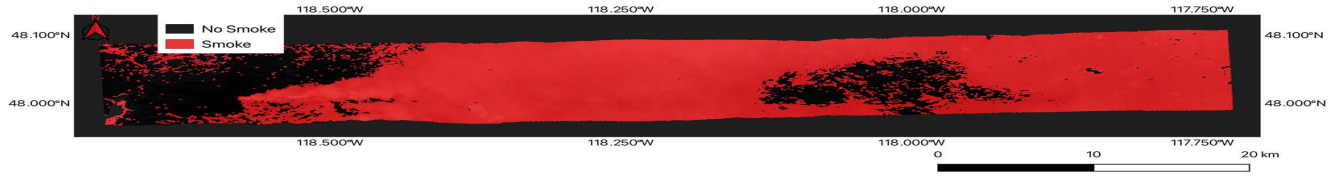

(**a**) Clustered Output




(**b**) Raw Fire Mask




(**c**) Raw Fire Mask with Burn Scar




(**d**) Raw Smoke Mask

**Figure 17.** MASTER eMAS fuse clustered raw output and generated smoke and fire masks (**a**–**d**).

## 4. Discussion

In all cases, both within large full-scene evaluations and finer, small-scale segmentation evaluations, the RBM-based fusion clustering product performed the best. Each individual instrument's RBM-based clustering product maintained an agreement level in most cases, remaining viable for use when fusion is not possible. This is a valuable methodology for the fusion and identification of various geophysical objects in an instrument and modality-agnostic way, providing a valuable first step towards the larger goal of automated object detection and tracking across data sets derived from multiple instruments. Further studies into the concrete uses of fire detection and the development of other products, such as harmful algal bloom detection, are underway, as well as using the data as input to other pieces within the larger object detection and tracking architecture that we are researching.

## 5. Conclusions

This study evaluated the ability of our methodology to generate models that can accurately represent the detailed structures within remotely sensed data sets in a way that allows for multi-sensor use-cases and fusion, where appropriate. To do this, we clustered

the output of RBM models separately trained on geolocated and orthorectified radiance data from seven different satellite and airborne instruments. These instruments collect data with typical spectral resolutions. Three fusion data sets from different pairs of the seven aforementioned data sets, and one data set from a hyperspectral imager, were also included. Coarse-scale, broad tests were performed with labeled data sets spanning large areas, but providing more coarse detail, and fine-scale tests were performed against data sets with smaller, more detailed classes and objects—such as smoke plumes—that are traditionally hard to fully segment. In all coarse-scale experiments, the agreement between our product and pre-existing products was always >80% and the balanced agreement was always >65%, with cases noted where the structure represented within data sets produced by our models provided a more accurate representation of the objects in the scenes tested. For the finer-scale comparisons, many of the comparisons yielded an agreement of >70% and, where agreement was lower, we were able to identify parts of the label-sets that were not identified in pre-existing products. Finally, when compared to simply using clustering, the RBM plus clustering method outperformed the other methods in all cases, showing that the RBM provides added value by allowing for a more detailed representation of the structure and latent patterns within the data. These results provide a valuable foundation that allows us to venture further into concrete applications, such as the identification of fire and smoke, or harmful algal blooms. The results also allow us to continue to further the methodology, in order to utilize the presented structure as a part of a larger object detection and tracking system.

**Author Contributions:** Conceptualization, N.L., M.J.G., B.D.B., and E.L.; data curation, N.L.; formal analysis, N.L.; funding acquisition, N.L., M.J.G., and H.E.-A.; investigation, N.L.; methodology, N.L., B.D.B., and E.L.; project administration, N.L., H.E.-A., and E.L.; resources, N.L. and H.E.-A.; software, N.L.; supervision, N.L., H.E.-A., and E.L.; validation, N.L., M.J.G., B.D.B., and E.L.; visualization, N.L., M.J.G., B.D.B., and H.E.-A.; writing—original draft, N.L.; writing—review and editing, N.L., M.J.G., B.D.B., H.E.-A., and E.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Data Science Working Group at the Jet Propulsion Laboratory. Computing resources were leveraged at both the NASA Center for Climate Simulation (NCCS), and the Machine Learning and Affiliated Technologies (MLAT) Lab at Chapman University.

**Data Availability Statement:** The data generated are not currently publicly available, but we have acquired funding to make this software, and example data sets, open source. This will be available soon, and accessibility announcements will be made in upcoming publications.

**Acknowledgments:** The authors would like to thank JPL, NCCS, MLAT Lab., and the Schmid College of Science and Technology, Chapman University, for supporting this research. The authors would also like to thank Phil Dennison, from the Geography Department at the University of Utah, as well as NASA for providing the data, without which this research would not have been possible. Finally, the authors would like to thank the anonymous reviewers for taking the time to read this paper and provide valuable feedback.

**Conflicts of Interest:** The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. LaHaye, N.; Ott, J.; Garay, M.J.; El-Askary, H.M.; Linstead, E. Multi-Modal Object Tracking and Image Fusion With Unsupervised Deep Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3056–3066. [CrossRef]
2. Kanezaki, A. Unsupervised Image Segmentation by Backpropagation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
3. Chen, M.; Artières, T.; Denoyer, L. Unsupervised Object Segmentation by Redrawing. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Nice, France, 2019; Volume 32.
4. Aganj, I.; Harisinghani, M.G.; Weissleder, R.; Fischl, B. Unsupervised Medical Image Segmentation Based on the Local Center of Mass. *Sci. Rep.* **2018**, *8*, 2045–2322. [CrossRef] [PubMed]

5. Soares, A.R.; Körting, T.S.; Fonseca, L.M.G.; Neves, A.K. An Unsupervised Segmentation Method For Remote Sensing Imagery Based On Conditional Random Fields. In Proceedings of the 2020 IEEE Latin American GRSS ISPRS Remote Sensing Conference (LAGIRS), Santiago, Chile, 22–26 March 2020; pp. 1–5. [CrossRef]
6. Zhang, R.; Yu, L.; Tian, S.; Lv, Y. Unsupervised remote sensing image segmentation based on a dual autoencoder. *J. Appl. Remote Sens.* **2019**, *13*, 1–19. [CrossRef]
7. Hinton, G.E.; Salakhutdinov, R. Deep Boltzmann Machines. In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater, FL, USA, 16–19 April 2009; Volume 5.
8. Hinton, G.E. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Comput.* **2002**, *14*, 1771–1800. [CrossRef] [PubMed]
9. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In Proceedings of the SIGMOD '96: 1996 ACM SIGMOD International Conference on Management of Data, Montreal, QC, Canada, 4–6 June 1996; Association for Computing Machinery: New York, NY, USA, 1996; pp. 103–114. [CrossRef]
10. Lrn2 Cre8 Learning to Create. Available online: https://web.archive.org/web/20160722213851/http://lrn2cre8.eu:80/ (accessed on 12 June 2021).
11. Roder, M.; de Rosa, G.H.; Papa, J.P. Learnergy: Energy-based Machine Learners. *arXiv* **2020**, arXiv:2003.07443.
12. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Nice, France, 2019; pp. 8024–8035.
13. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
14. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The Balanced Accuracy and Its Posterior Distribution. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 3121–3124. [CrossRef]
15. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [CrossRef]
16. Prechelt, L., Early Stopping - But When? In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 55–69. [CrossRef]
17. Diner, D. MISR Experiment Overview. 1999. Available online: https://trs.jpl.nasa.gov/handle/2014/18644 (accessed on 12 June 2021).
18. King, M.D.; Kaufman, Y.J.; Menzel, W.P.; Tanre, D. Remote Sensing of Cloud, Aerosol, and Water Vapor Properties from the Moderate Resolution Imaging Spectrometer (MODIS). *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 2–27. [CrossRef]
19. Mazzoni, D.; Garay, M.J.; Davies, R.; Nelson, D. An operational MISR pixel classifier using support vector machines. *Remote Sens. Environ.* **2007**, *107*, 149–158. [CrossRef]
20. Ackerman, S.A.; Strabala, K.I.; Menzel, W.P.; Frey, R.A.; Moeller, C.C.; Gumley, L.E. Discriminating clear sky from clouds with MODIS. *J. Geophys. Res. Atmos.* **1998**, *103*, 32141–32157. [CrossRef]
21. Van Donkelaar, A.; Martin, R.V.; Levy, R.C.; da Silva, A.M.; Krzyzanowski, M.; Chubarova, N.E.; Semutnikova, E.; Cohen, A.J. Satellite-based estimates of ground-level fine particulate matter during extreme events: A case study of the Moscow fires in 2010. *Atmos. Environ.* **2011**, *45*, 6225–6232. [CrossRef]
22. Tuia, D.; Gabriele, M.; Le Saux, B.; Bechtel, B. Data Fusion Contest 2017 (DFC2017). Available online: https://doi.org/10.21227/e56j-eh82 (accessed on 11 June 2021).
23. Roy, D.P.; Wulder, M.A.; Loveland, T.R.; Woodcock, C.E.; Allen, R.G.; Anderson, M.C.; Helder, D.; Irons, J.R.; Johnson, D.M.; Kennedy, R.; et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sens. Environ.* **2014**, *145*, 154–172. [CrossRef]
24. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
25. Ching, J.; Mills, G.; Bechtel, B.; See, L.; Feddema, J.; Wang, X.; Ren, C.; Brousse, O.; Martilli, A.; Neophytou, M.; et al. WUDAPT: An Urban Weather, Climate, and Environmental Modeling Infrastructure for the Anthropocene. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 1907–1924. [CrossRef]
26. Hook, S.J.; Myers, J.J.; Thome, K.J.; Fitzgerald, M.; Kahle, A.B. The MODIS/ASTER airborne simulator (MASTER)— A new instrument for earth science studies. *Remote Sens. Environ.* **2001**, *76*, 93–102. [CrossRef]
27. Guerin, D.C.; Fisher, J.; Graham, E.R. The enhanced MODIS airborne simulator hyperspectral imager. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVII*; Shen, S.S., Lewis, P.E., Eds.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2011; Volume 8048, pp. 214–224. [CrossRef]
28. Giglio, L.; Justice, C. MOD14A2 MODIS/Terra Thermal Anomalies/Fire 8-Day L3 Global 1km SIN Grid V006. 2015. Available online: https://doi.org/10.5067/MODIS/MOD14A2.006 (accessed on 11 June 2021).