

Article

Not peer-reviewed version

# Unsupervised Characterization of Water Composition with UAV-based Hyperspectral Imaging and Generative Topographic Mapping

[John Waczak](#) , Adam Aker , [Lakitha O. H. Wijeratne](#) , [Shawhin Talebi](#) , Ashen Fernando , Prabuddha M. H. Dewage , Mazhar Iqbal , Matthew Lary , David Schaefer , [Gokul Balagopal](#) , [And David J. Lary](#) \*

Posted Date: 3 June 2024

doi: 10.20944/preprints202405.1792.v1

Keywords: Hyperspectral Imaging; Remote Sensing; Unsupervised Classification; Endmember Extraction; Generative Topographic Mapping










Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Article

# Unsupervised Characterization of Water Composition with UAV-based Hyperspectral Imaging and Generative Topographic Mapping

John Waczak , Adam Aker , Lakitha O. H. Wijeratne , Shawhin Talebi , Ashen Fernando , Prabuddha M. H. Dewage , Mazhar Iqbal, Matthew Lary, David Schaefer, Gokul Balagopal, and David J. Lary \* 

Hanson Center for Space Sciences, University of Texas at Dallas, Richardson, TX 75080, USA; john.waczak@utdallas.edu (J.W.); adam.aker@utdallas.edu (A.A.); lhw150030@utdallas.edu (L.O.H.W.); shawhintalebi@gmail.com (S.T.); ashen.fernando@utdallas.edu (B.F.); pxh180012@utdallas.edu (P.M.H.D.); mazhar.iqbal@utdallas.edu (M.I.); MDL210001@utdallas.edu (M.L.); captdaveschaefer@gmail.com (D.S.); gokul.balagopal@utdallas.edu (G.B.);

\* Correspondence: david.lary@utdallas.edu

**Abstract:** Unmanned Aerial Vehicles (UAVs) equipped with hyperspectral imagers have emerged as an essential technology for the characterization of inland water bodies. The high spectral and spatial resolutions of these systems enable the retrieval of a plethora of optically-active water quality parameters via band ratio algorithms and machine learning methods. However, fitting and validating these models requires access to sufficient quantities of in situ reference data which are time-consuming and expensive to obtain. In this study, we demonstrate how the Generative Topographic Mapping (GTM), a Bayesian realization of the Self-organizing Map, can be used to visualize high-dimensional hyperspectral imagery and extract spectral signatures corresponding to unique endmembers present in the water. Using data collected across a North Texas pond, we first apply the GTM to visualize the distribution of captured reflectance spectra revealing small-scale spatial variability of water composition. Next, we demonstrate how the nodes of the fitted GTM can be interpreted as unique spectral endmembers. Using extracted endmembers together with the normalized spectral similarity score, we are able to efficiently map the abundance of near shore algae as well as the evolution of a rhodamine tracer dye used to simulate water contamination by a localized source.

**Keywords:** Hyperspectral Imaging; Remote Sensing; Unsupervised Classification; Endmember Extraction; Generative Topographic Mapping

## 1. Introduction

Inland water bodies present a unique challenge to characterization by remote sensing imagery due to their complex spectral characteristics and small-scale spatial variability. The broad bands of multispectral imagers coupled with the irregular shape of lakes and rivers result in pixels with highly mixed signals that are easily dominated by reflectance from shore and nearshore vegetation sources [1,2]. Recently, the combination of hyperspectral imaging with unmanned aerial vehicles (UAVs), such as drones, has emerged as a powerful approach to simultaneously address the spectral, spatial and temporal limitations of traditional high-altitude and satellite-based collection [3,4]. UAVs are significantly less expensive to deploy than their satellite or aircraft based remote sensing counterparts, and low-altitude flights enable centimeter-scale sampling while limiting the need for complicated atmospheric corrections [5]. However, the significant increase in the data volume generated by these systems presents a new challenge, namely, how to efficiently extract water quality parameters of interest from intricate pixel spectra.

Significant research efforts have focused on the development of techniques and algorithms to retrieve water quality parameters, e.g. from UAV-captured hyperspectral images (HSI). On-board compute installed alongside hyperspectral imagers can enable the rapid evaluation of spectral indices from HSI band ratios [6]. These band ratios and polynomial combinations of bands have been used to successfully invert optically active water quality parameters such as turbidity directly from

UAV acquired imagery [7,8]. Supervised machine learning techniques such as tree-based models, support vector machines, and neural networks have also been used to estimate a wide range of parameters such as colored dissolved organic matter, chlorophyll A, blue-green algae, and suspended sediment concentrations [9,10]. The calibration and evaluation of these data-driven models require a significant volume of coincident in situ data. This can be addressed by coordinating UAV flights with reference data collection using autonomous robotic boats [11,12]. However, this approach relies on prior knowledge of expected sources in order to select appropriate reference instruments for model validation. The presence of unanticipated contaminants cannot be directly identified in this sensing paradigm.

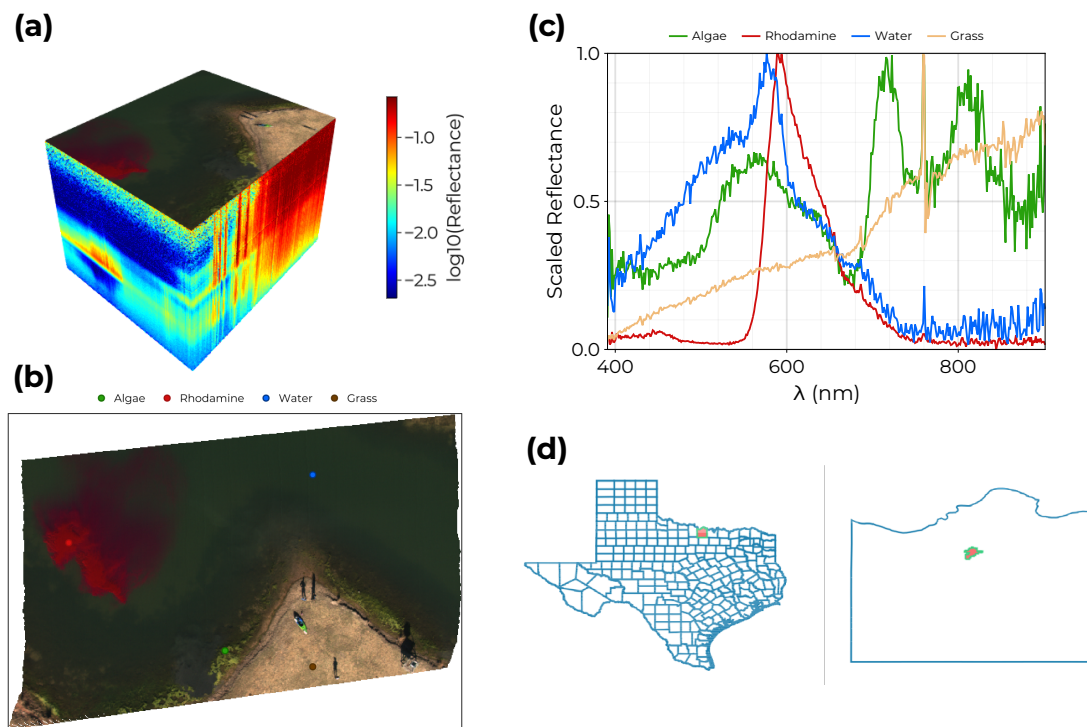
Extending the capabilities of UAV-based hyperspectral imaging to enable water quality monitoring in real-time scenarios where contaminant sources may not be known in advance requires two additional capabilities: dimensionality reduction techniques to permit the visual comparison of HSI, and endmember extraction techniques which can identify spectral signatures corresponding to unique sources within the imaging scene. In remote sensing where reference data are typically sparse, many approaches have been explored. For example, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (tSNE) are dimensionality reduction methods commonly used to reduce HSI to two or three dimensions for visualization [13,14]. Similarly for endmember extraction there are a variety of established approaches including geometric methods like vertex component analysis, statistical methods like k-nearest neighbors and non-negative matrix factorization (NMF), and deep learning methods based on autoencoder architectures [15–21]. Methods based on linear mappings like PCA and NMF are often too restrictive for HSI where the assumption of linear mixing is easily broken. However, the increased complexity of nonlinear methods like tSNE and autoencoders often lead to significant increases in computation time. An ideal approach should enable both visualization and nonlinear identification of relevant spectral endmembers.

The self-organizing map (SOM) developed by Teuvo Kohonen is an unsupervised machine learning method which non-linearly maps high dimensional data to the nodes of a two-dimensional grid [22]. By preserving the topological relationship between nodes during training, the SOM ensures that similar spectral signatures are mapped together such that related HSI pixels naturally cluster together. This presents an attractive compromise by enabling the simultaneous visualization of HSI data and endmember extraction via the weight vector associated with each SOM node [23–25]. When reference data are available, the SOM can be utilized to provide semi-supervised labeling of HSI spectra [26]. Furthermore, Danielsen et al. demonstrated that the dimensionality reduction offered by the SOM can even be used for on-board data compression of HSIs acquired by a CubeSat [27]. Despite these clear capabilities, the SOM relies on a heuristic training algorithm with hyperparameters that can be challenging to tune and offers no direct probabilistic interpretation. To address these limitations, Bishop et al. developed the generative topographic mapping (GTM), a Bayesian latent-variable model inspired by the SOM [28]. The GTM has been utilized in a variety of domains including drug design and chemical data visualization but has yet to see adoption for the analysis of hyperspectral imagery [29–31].

In this paper, we explore the application of the GTM to UAV-acquired HSI for the characterization water quality. Using data collected at a pond in Montague, North Texas, we first train a GTM using water-only pixels to produce a low-dimensional representation of the collected HSI. We use this mapping to explore the highly detailed small-scale variability within the pond and discuss how this can be used to guide reference data collection. Next, we demonstrate how the GTM can be utilized to identify relevant spectral endmembers from a combined dataset including land pixels, algae, water, and a simulated contaminant plume using a rhodamine tracer dye. Once identified, these endmembers can be used to rapidly map the abundance of spectral features within the pond. We demonstrate this capability by using a GTM to map the abundance of algae near the shore as well as the dispersion of a rhodamine dye plume.

## 2. Materials and Methods

In this study, we explore the use of the GTM as a tool for dimensionality reduction and non-linear endmember extraction of hyperspectral imagery. To this end, a dataset of HSI were collected at a pond in Montague, North Texas (shown in panel d of Figure 1) on 23 November 2020 using a UAV mounted hyperspectral imager configured as described in [11,12]. In this section we first provide a detailed overview of the GTM algorithm. Next we describe the UAV platform used for HSI collection and the various steps used to process captured HSI. Finally, we describe the two case studies presented in this paper for the utilization of the GTM as a HSI visualization tool and as an endmember extraction technique.



**Figure 1.** (a) Sample hyperspectral data cube. Spectra are plotted using their geographic position with the  $\log_{10}$ -reflectance colored along the z axis and a pseudo-color image on top. (b) Points taken from a sample hyperspectral data cube corresponding to algae, rhodamine dye, water, and dry grass. (c) Reflectance spectra for the exemplar points scaled so the peak value of each spectrum is 1.0. (d) The location of the pond in Montague, Texas where data were collected for this study.

### 2.1. Generative Topographic Mapping

The GTM is a probabilistic latent variable model inspired by the SOM for visualizing and clustering high-dimensional data. Like the SOM, the GTM assumes vectors  $\mathbf{x}$  in the  $d$ -dimensional data space (here representing reflectance spectra) are constrained to a low-dimensional embedded manifold. The SOM describes this manifold using a regular grid of nodes, each having an associated weight vector defining the mapping from the manifold to the data space. The position of each data record  $\mathbf{x}$  in the manifold is assigned to the position of the node whose weight vector has the minimum Euclidean distance to  $\mathbf{x}$ . An iterative training procedure updates the weight vectors of each node to fit the manifold to the data such that nodes near each other in the SOM grid correspond to similar records in the data space.



The GTM mimics the grid of the SOM by assuming data are generated from latent variables  $\xi$  which are constrained to the  $K$ -many nodes of a regular grid. This assumption corresponds to establishing a prior distribution on the latent space of the form

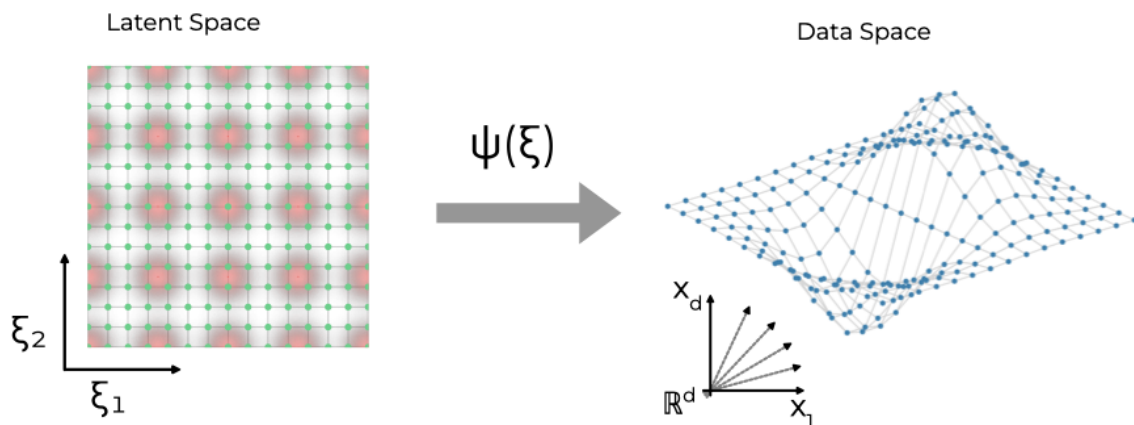
$$p(\xi) = \frac{1}{K} \sum_k^K \delta(\xi - \xi_k) \quad (1)$$

where  $\delta(\cdot)$  is the Dirac delta function.

Points  $\xi$  in this latent space are mapped to the embedded manifold by a non-linear function  $\psi$  parameterized by weights  $W$  as illustrated in Figure 2. However, since real data is rarely noise-free, this embedded manifold will not be perfectly thin. To account for this, the points  $\xi$  are described in the data space by a radially symmetric Gaussian distribution,  $\mathcal{N}(\psi(\xi), \beta^{-1})$ , with mean  $\psi(\xi)$  and variance  $\beta^{-1}$ . For a dataset  $\{\mathbf{x}_n\}_{n=1}^N$  consisting of  $N$ -many records, this choice yields a log-likelihood function of the form

$$\mathcal{L}(W, \beta) = \sum_n^N \ln \left( \frac{1}{K} \sum_k^K p(\mathbf{x}_n | \xi_k, W, \beta) \right) \quad (2)$$

which can be maximized to yield optimal  $W$  and  $\beta^{-1}$ .



**Figure 2.** Illustration of the GTM algorithm. On the left is a regular grid of  $K$  many points (green) in the latent space represented by their coordinates  $\xi_1$  and  $\xi_2$ . In red are the  $M$ -many RBF which define the mapping  $\psi$  from the latent space to the data space. Points in the latent space are mapped non-linearly to the data space yielding an embedded manifold in  $\mathcal{R}^d$ , here illustrated in three dimensions.

The function  $\psi$  is typically chosen to be given by the sum of  $M$ -many radial basis functions (RBF) evenly distributed in the latent space such that  $\psi(\xi) = W\phi(\xi)$  with centers  $\mu_m$  and width  $\sigma$  so that

$$\phi_m(\xi) = \exp \left( -\frac{\|\xi - \mu_m\|^2}{2\sigma^2} \right). \quad (3)$$

The width  $\sigma$  is taken to be the distance between neighboring RBF centers multiplied by a scale factor,  $s$ . Together the number of RBFs,  $M$ , and  $s$  are hyperparameters for the model which govern the smoothness of the resulting manifold in the data space. Additionally, sparsity of  $W$  can be enforced by introducing an additional hyperparameter  $\alpha$  corresponding to a prior distribution over the weights given by

$$p(W | \alpha) = \left( \frac{\alpha}{2\pi} \right)^{MD/2} \exp \left( -\frac{\alpha}{2} \|W\|_F^2 \right). \quad (4)$$

The use of RBFs for  $\psi$  enables maximizing  $\mathcal{L}(W, \beta)$  via an Expectation-Maximization routine. During each step of the fitting process, the responsibility of the  $k$ th node for the  $n$ th data record is computed as

$$R_{kn} = p(\xi_k | \mathbf{x}_n, W, \beta) = \frac{p(\mathbf{x}_n | \xi_k, W, \beta)}{\sum_{k'}^K p(\mathbf{x}_n | \xi_{k'}, W, \beta)}. \quad (5)$$

Together these responsibilities form a matrix with entries  $R_{kn}$  which are kept fixed during the maximization step. The maximization step is then performed by updating the weights  $W$  and variance  $\beta^{-1}$  according to

$$W_{\text{new}} = \left( \Phi^T G \Phi + \frac{\alpha}{\beta} I \right)^{-1} \Phi^T R X \quad (6)$$

$$\frac{1}{\beta_{\text{new}}} = \frac{1}{ND} \sum_n^N \sum_k^K R_{kn} \|\psi_k - \mathbf{x}_n\|^2 \quad (7)$$

where  $\Phi_{km} = \phi_m(\xi_k)$ ,  $X$  is the data matrix formed by concatenating the records  $\mathbf{x}_n$ , and  $G$  is a diagonal matrix with  $G_{kk} = \sum_n^N R_{kn}$ . This process is repeated until the log-likelihood converges to a predetermined tolerance level.

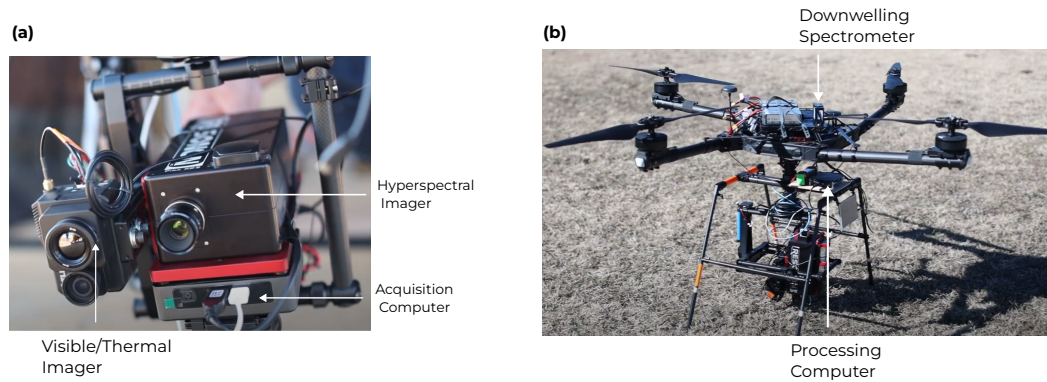
Slices  $R_{[:,n]}$  of the matrix  $R$  define the responsibility of each latent node  $\xi_k$  for the  $n$ th data record  $\mathbf{x}_n$ . Therefore the final responsibility matrix after GTM training can be used to represent each record in the latent space by the mean:

$$\hat{\xi}_n = \sum_k^K R_{kn} \xi_k. \quad (8)$$

A freely available implementation of the GTM algorithm was developed for this study and is accessible at [32]. The code is written in the Julia programming language and complies with the Machine Learning in Julia (MLJ) common interface [33,34].

## 2.2. UAV-Based Hyperspectral Imaging

A Freefly Alta-X autonomous quadcopter was used as the UAV platform in this study. This UAV was equipped with a Resonon Pika XC2 visible+near-infrared (VNIR) hyperspectral imager to capture HSI with 462 wavelengths per pixel ranging from 391 to 1011 nm. This imager is in a pushbroom configuration so that HSI are captured one scan-line at a time resulting in data cubes consisting of 1000 scan-lines with 1600 pixels each. Additionally, the camera includes an embedded GPS/INS unit to enable georectification of collected HSI. An upward facing Ocean Optics UV-Vis NIR spectrometer with a cosine corrector was also included to provide measurements of the incident solar irradiance spectrum. The configuration of the UAV with the attached hyperspectral imager is shown in Figure 3. Data collection and processing was controlled by an attached Intel NUC small-form-factor computer.



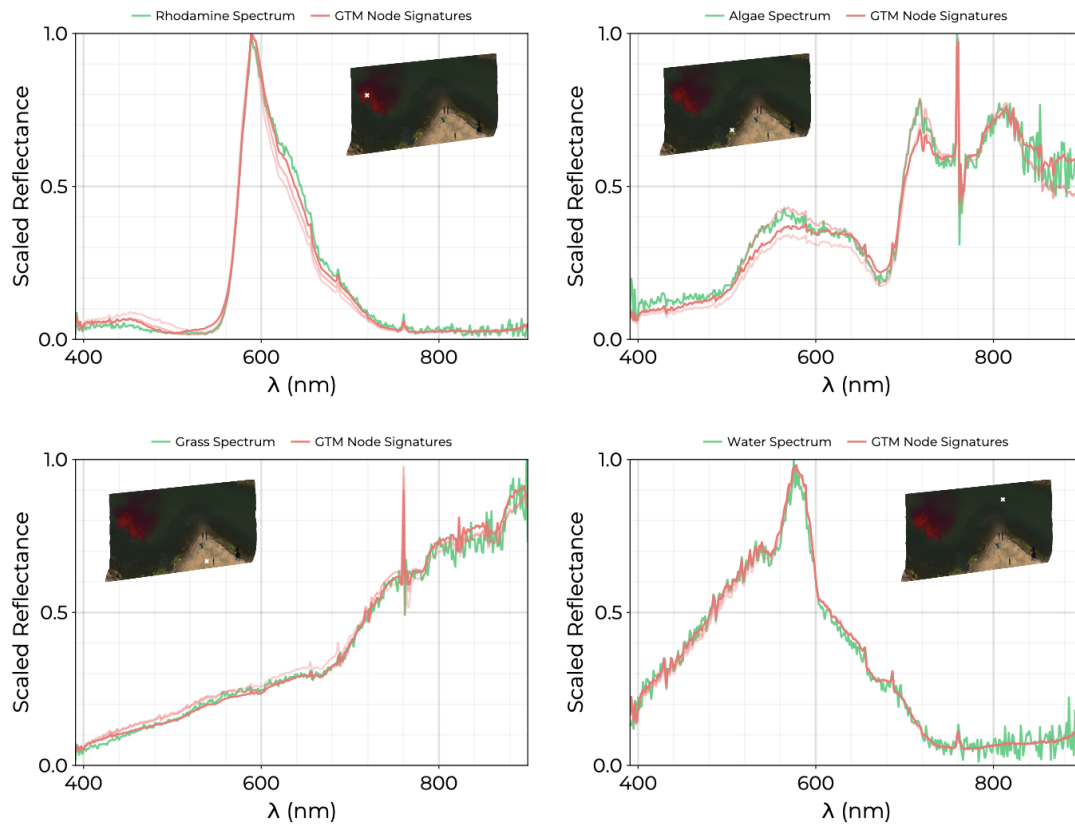
**Figure 3.** The UAV platform: (a) The Resonon Pika XC2 hyperspectral imager. (b) The Freefly Alta-X with the attached hyperspectral imager, processing computer, and downwelling irradiance spectrometer.

To account for the variability of incident light, raw HSI are converted to units of reflectance using the downwelling irradiance spectrum simultaneously captured with each HSI. With the hyperspectral imager oriented to nadir, the reflectance is given by

$$R(\lambda) = \pi L(\lambda) / E_d(\lambda) \quad (9)$$

where  $L$  is the spectral radiance,  $E_d$  is the downwelling irradiance, and a factor of  $\pi$  steradians results from assuming the water surface is Lambertian (diffuse) [35]. HSI collection was performed near solar noon to maximize the amount of incident sunlight illuminating the water. For the site in North Texas, this corresponded to an average solar zenith angle of  $54.9^\circ$  for 23 November 2020 resulting in HSI with negligible sunglint effects.

After conversion to reflectance, each HSI must also be georectified to assign geographic coordinates to each pixel. The UAV flights were performed at an approximate altitude of 50 m above the water so that the imaging surface can be considered to be flat. Consequently, HSI were rapidly georectified to a 10 cm resolution using position and orientation data from the embedded GPS/INS as outlined in [36–38]. An example hyperspectral data cube is illustrated in panel (a) of Figure 4 where the log10-reflectance values are plotted along the z-axis for each pixel in the scene and a psuedo-color image at the top of the data cube illustrates how the water would appear to the human eye from the perspective of the UAV.



**Figure 4.** Spectral signatures  $\psi(\xi_k)$  corresponding to GTM nodes with non-zero responsibility for exemplar spectra corresponding to the rhodamine dye plume (top left), algae (top right), dry grass (bottom left), and open water (bottom right). A pseudo-color image is inset into each plot with the location of the exemplar spectrum marked with a white circle.

As a final processing step before training each GTM, we limit the wavelengths of each HSI to  $\lambda \leq 900$  nm as wavelengths above 900 nm showed significant noise. Additionally, each spectrum was re-scaled to a peak value of 1.0 to account for incident light variability between HSI.

### 2.3. GTM Case Studies

To explore the ability of the GTM to segment HSI pixels and aid in source identification we first consider a dataset of water-only spectra consisting of  $> 36,000$  records identified from collected HSI by a normalized difference water index (NDWI) greater than 0.25 where the NDWI is defined as

$$\text{NDWI} = \frac{R(550) - R(860)}{R(550) + R(860)}. \quad (10)$$

We then apply the trained GTM to all water-only pixels to visualize the distribution learned by the GTM and examine spectra associated with a subset of nodes in the latent space in order to assess the small-scale spatial variability within the pond.

Next, we consider a dataset of combined HSI pixels including both water and land. To simulate the dispersion of a potential contaminant source, a Rhodamine tracer dye was released into the western portion of the pond and two additional UAV flights were used to collect HSI capturing the evolution of the resulting plume. From these HSI a collection of  $> 145,000$  pixels were sampled for model training. Exemplar spectra for water, grass, algae, and rhodamine dye were identified from a sample HSI as shown in panels (b) and (c) of Figure 1. Using these spectra, we explore the trained GTM and use it to extract spectral signatures corresponding to these endmember categories. Values for the



hyperparameters  $m$ ,  $s$ , and  $\alpha$ , are determined by training multiple GTM models and selecting values which minimize the Bayesian Information Criterion (BIC) given by

$$\text{BIC} = 2P \ln(N) - 2\mathcal{L} \quad (11)$$

where  $P$  is the total number of model parameters,  $N$  is the number of records in the dataset, and  $\mathcal{L}$  is the log-likelihood defined in Eq 2.

The normalized spectral similarity score (NS3) introduced by Nidamanuri and Zbell combines the root-mean-square (RMS) difference together with the spectral angle to provide a spectral distance function[39]. For two spectra  $R_1(\lambda)$  and  $R_2(\lambda)$  it is defined by

$$\text{NS3}(R_1, R_2) = \sqrt{\text{RMS}(R_1, R_2)^2 + (1 - \cos \theta)^2} \quad (12)$$

where

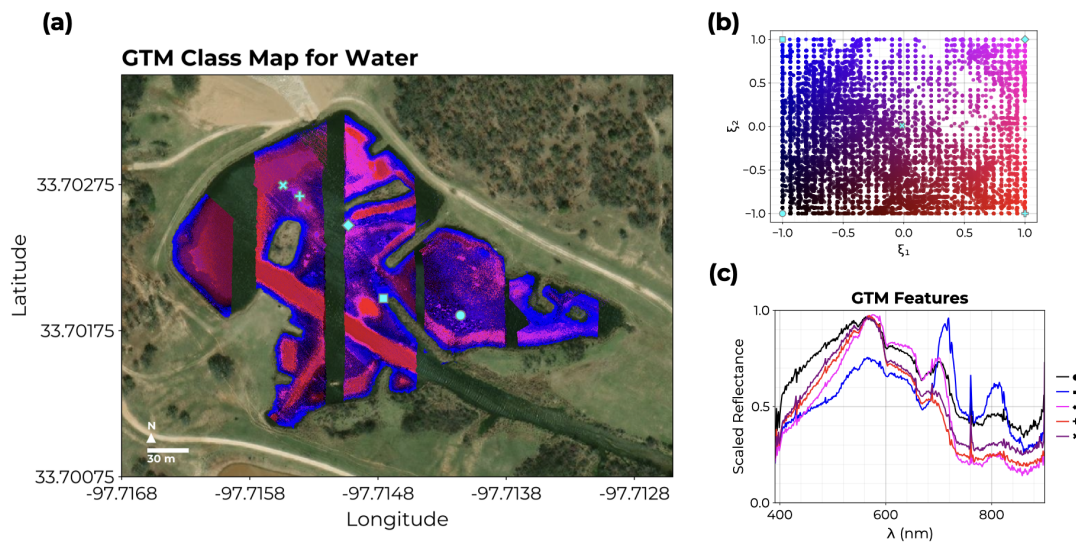
$$\text{RMS}(R_1, R_2) = \sqrt{\frac{1}{N-1} \sum_{\lambda} (R_1(\lambda) - R_2(\lambda))^2} \quad (13)$$

$$\cos \theta = \frac{\langle R_1, R_2 \rangle}{\|R_1\| \|R_2\|}. \quad (14)$$

We use the NS3 together with identified spectral endmembers to map the abundance of algae near the shore as well as the evolution of the Rhodamine dye plume.

### 3. Results

#### 3.1. Water-only Pixel Segmentation



**Figure 5.** Classification map for GTM trained solely on water pixels (no land and no rhodamine plume). **(a)** GTM applied to all water pixels colored by their associated position in the latent space. **(b)** Representation of the points from the training set in the latent space. Color is assigned to each point by mapping  $\xi_1$  to red and  $\xi_2$  to blue. **(c)** GTM spectral signatures,  $\psi(\xi_k)$ , corresponding to nodes from the four corners and center of the GTM latent space.

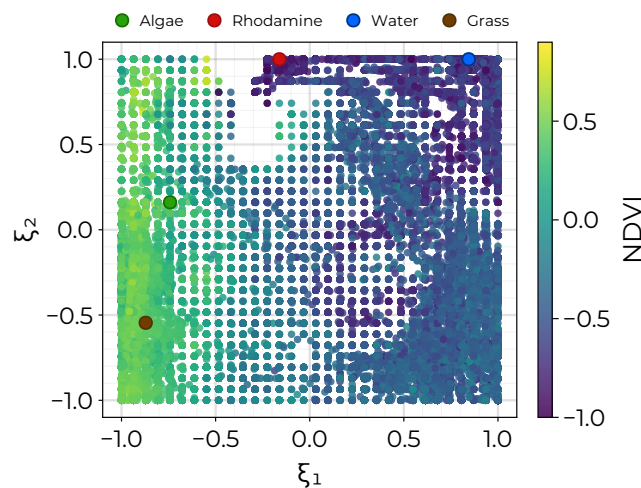
A GTM with  $K = 32 \times 32$  nodes was trained on the dataset of water-only pixels described in Section 2.3 in order to explore the distribution of reflectance spectra captured by the HSI. The resulting GTM is visualized in Figure 5. First, we note that the GTM has utilized most of the latent space as

illustrated panel (b) where the position corresponding to the mean responsibility of each data point  $\hat{\xi}_n$  has been plotted. For this water-only GTM, the spectra appear to be largely clustered toward the left edge, bottom, and right edge of the latent space. Spectral signatures corresponding to GTM nodes from the four corners and center of the latent space are shown in panel (c) illustrating the spectral variability represented across the latent space.

To visualize the distribution of HSI learned by the GTM, we can associate a color with each dimension of the latent space. In Figure 5 we have used the red channel to represent  $\xi_1$  and the blue channel to represent  $\xi_2$ . Applying the trained GTM to compute mean node responsibilities for all water-pixels in collected HSI allows us to illustrate the spatial distribution of the spectra on a map as shown in panel (a). Here we observe a clear distinction between water near the shore and water in the middle of the pond. Additionally, the eastern alcove of the pond is significantly more blue than the rest of the water reflecting the limited flow through this region. The close proximity of highly dissimilar GTM classes illustrated by sharp color gradients in the map captures the small-scale spatial variability typical of inland water bodies like this pond.

### 3.2. Endmember Extraction

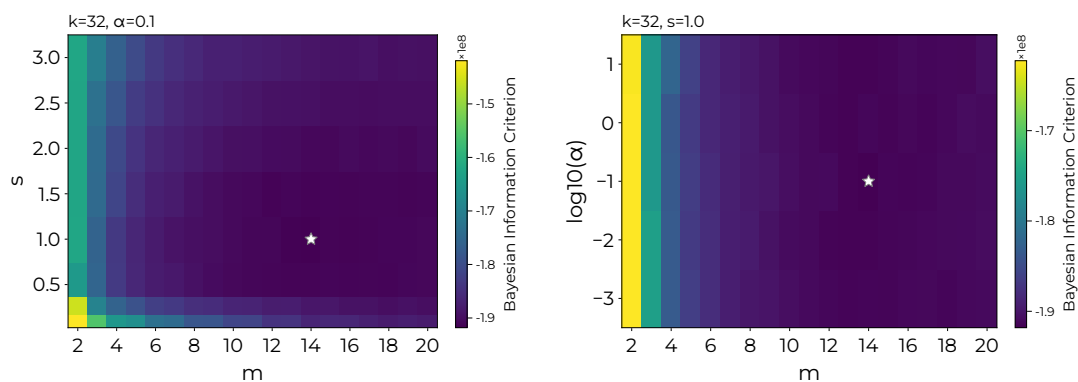
A second GTM was trained on the combined dataset of HSI covering shore, water, and the rhodamine tracer dye release. The resulting class map is shown in Figure 6 where each sample has been plotted at the position of the mean node responsibility and colored by the normalized difference vegetation index (NDVI) computed from the original reflectance spectrum. The NDVI is a spectral index sensitive to variations in vegetation health with negative values corresponding to water, small positive values corresponding to sparse vegetation, and large values near 1 corresponding to dense vegetation [40]. From this, we see that the GTM can clearly separate spectra by vegetation content with high values concentrated to the left of the latent space and negative, water-based pixels concentrated to the right. Additionally, the position of exemplar spectra for algae, rhodamine, water, and grass are included as color-filled circles further illustrating the classification obtained by the GTM.



**Figure 6.** GTM latent space visualization: Each sample spectrum from the dataset of combined HSI covering shore, water, and the rhodamine dye release are plotted in the GTM latent space at the location of the mean node responsibility,  $\hat{\xi}_n$ . Points are colored according to the NDVI computed from the original reflectance spectra. The locations of exemplar spectra for algae, rhodamine, water, and grass in the latent space are included as color-filled circles. Spectra from the shore and water are clearly separated to different regions of the latent space.

As outlined in Section 2.1, there are three hyperparameters that need to be chosen to fit a GTM model: the number of RBF centers  $M = m^2$  with  $m$  the number along each axis, the scale factor  $s$

which controls the RBF overlap, and the regularization parameter  $\alpha$ . To choose appropriate values we performed a grid search for  $m$  values between 2 and 20,  $s$  values between 0.1 and 3.0, and  $\alpha$  values between 0.001 and 10.0. The best values were determined as those which minimized the BIC, with  $m = 14$ ,  $s = 1.0$ , and  $\alpha = 0.1$ , respectively. Heatmaps comparing the BIC for different hyperparameter values are provided in Figure 7 and a table of the top 25 models is given in Appendix A. Since the matrix of RBF activations  $\Phi$  need only be computed once at the GTM initialization step, the number of latent nodes given by  $K = k^2$  can be chosen to be large enough to ensure a smooth mapping. We found that a value of  $k = 32$  provided a sufficient number of GTM nodes without significantly impacting training time.

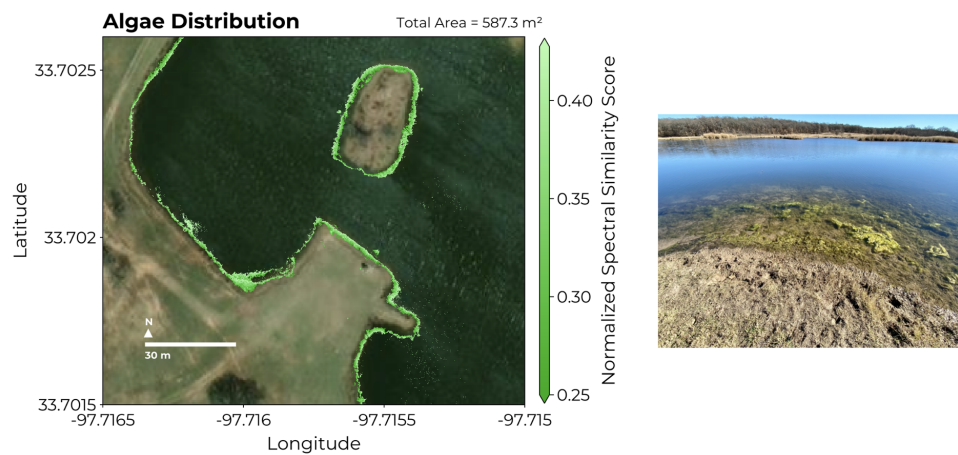


**Figure 7.** Results of the hyperparameter search. **(left)** Variation of BIC with  $m$  and  $s$  for fixed  $\alpha = 0.1$ . **(right)** Variation of the BIC with  $m$  and  $\alpha$  for fixed  $s = 1.0$ . The white star in each plot indicates the parameters with the lowest BIC across the entire parameter search.

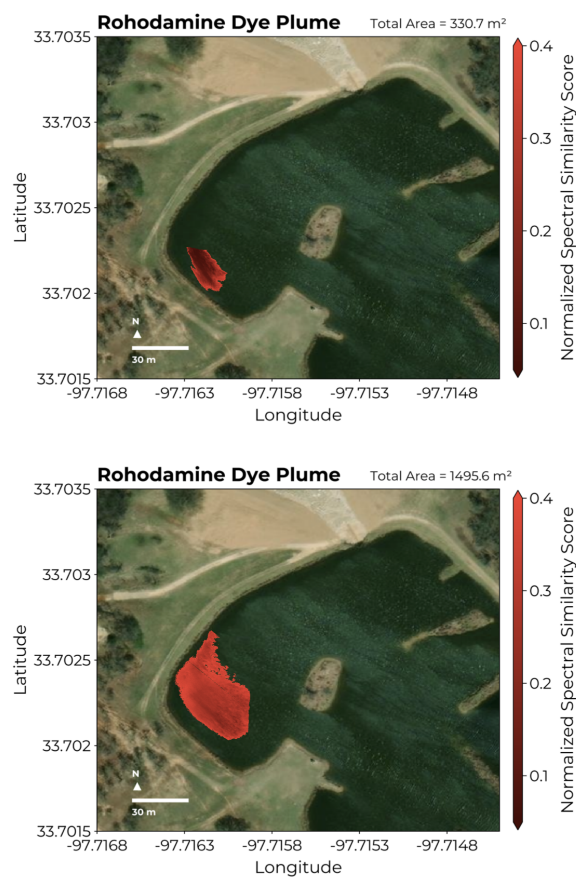
Using the trained GTM we can compute spectral signatures for each node in the latent space via the non-linear mapping  $\psi$ . The responsibilities  $R_{kn}$  therefore correspond to the contributions of the  $k$ th node  $\zeta_k$  to the  $n$ th sample spectrum in the dataset. Endmembers are identified by nodes with non-zero responsibility values for each of the selected exemplar spectra and are plotted in Figure 4. We note that these extracted endmembers are able to accurately capture the shape of exemplar spectrum including thin reflectance peaks. The endmembers also manage to smoothly interpolate through noisy wavelengths as shown by the water and algae endmembers for  $\lambda > 800$  nm.

### 3.3. Abundance Mapping with NS3

Spectral endmembers extracted from the GTM can be used to map the abundance of water constituents in HSI by selecting HSI pixels with a NS3 value below a given threshold. In Figure 8 we demonstrate this by using the extracted algae endmember spectrum to map algal abundance near the western shore of the pond. Identifying algae pixels as those with an NS3 value below 0.4725, we estimate the area subsumed by algae in Figure 8 to be  $587.3 \text{ m}^2$ . Similarly, we are able to track the evolution of the rhodamine dye release into the western portion of the pond by computing the NS3 for HSI collected across multiple flights. In the left panel of Figure 9, we see that the dye plume initially encompasses an area of roughly  $330.7 \text{ m}^2$ . A second flight performed 15 minutes later reveals the dye plume to have increased to an area of  $1495.6 \text{ m}^2$ . The increase in NS3 values between flights reflects the dilution of dye as it diffused into the water.



**Figure 8.** Using the spectral endmember assigned to algae from the trained GTM together with the NS3, we are able to estimate the algal abundance near the shore. **(Top)** NS3 values of pixels below a threshold of 0.4275 corresponding to a total area of 587.3 m<sup>2</sup>. **(Bottom)** A picture of the pond near the shore showing the algae.



**Figure 9.** Using the spectral endmember assigned to Rhodamine from the trained GTM together with the NS3, we are able to map the evolution of the dye plume across two UAV flights. **(left)** The initial dye plume corresponding to a total area of 330.7 m<sup>2</sup>. **(right)** The same dye plume imaged approximately 15 minutes later extends to a total area of 1495.6 m<sup>2</sup>.



#### 4. Discussion

The application of UAV-based hyperspectral imaging (HSI) for water quality assessment is gaining significant traction due to its ability to efficiently capture detailed spectral data at high spatial resolutions. Most studies using UAVs have employed supervised techniques that map the captured spectra to specific water quality parameters of interest. For example Lu et al. evaluated a variety of machine learning methods for the extraction of chlorophyll-a and suspended sediment concentrations from HSI using data from UAV flights over 33 sampling locations [10]. However, this approach relies on the collection of paired in situ data which can be challenging to obtain in sufficient quantities to facilitate model training. Additionally, supervised methods require prior knowledge of expected sources in order to select and calibrate suitable reference instruments. Using models purpose-built for specific water quality parameters like chlorophyll a or turbidity discards potential information contained in HSI which can reveal unanticipated sources. Therefore, unsupervised methods which aid in the visualization of HSI and enable the identification of spectral endmembers within the imaging scene are needed to complement these supervised approaches.

Recently some researchers have begun to explore endmember extraction using UAV-based imaging where the increased spatial resolution provided by UAV platforms is hypothesised to yield more pure pixels than their satellite counterparts. For instance, Alvarez et al. used multi-spectral UAV imagery to extract endmembers which were then applied to unmix remote sensing data for plant abundance estimation across a broad region of France [41]. Similarly, Gu et al. explored UAV-to-satellite hyperspectral unmixing by applying endmembers extracted from UAV-based HSI to satellite observations [42]. These studies underscore the potential of combining endmember extraction techniques with UAV-based HSI which has yet to see widespread adoption for water quality analysis.

In this study, we explored the GTM as a unsupervised approach to simultaneously enable visualization of HSI and perform endmember extraction of spectral signatures. First, we showed how the representation of data in the latent space of the GTM given by the mean responsibility can be used to visualize the small-scale structures within inland water bodies as evidenced in Figure 5. In particular, we note that the sharp gradients in colors found when visualizing the spatial distribution of GTM nodes across the pond reflects significant variability in water composition at the sub-meter scale. This has important consequences for water quality assessment where the location of in situ data collection will have a strong impact on the ability of any model to predict water quality parameters from HSI. Visualizing the distribution of spectra from collected HSI is therefore highly relevant to guide in situ data collection as shifting the sampling location by as little as a meter can lead to significant differences in measured values.

Based on these observations, one clear application of the GTM for real-time water quality assessment is for intelligently guided reference data collection. In our previous work, we showed that coordinating UAV-based hyperspectral imaging with in situ data collection by an autonomous boat can dramatically improve the inversion of water quality parameters from HSI pixels [12]. However, the spatial distributions of parameters such as chlorophyll-a, blue-green algae, and temperature are often highly dissimilar posing a challenge for optimal route planning. Since the GTM estimates the full distribution of reflectance spectra and not a single water quality parameter, one could construct a prize collecting travelling salesman problem (PC-TSP) which seeks to find the optimal route maximizing the area explored in the GTM latent space [43]. Similar approaches have been used to guide data collection with autonomous vehicles to optimize data quality subject to resource constraints [44].

The second application of the GTM presented in this study is for the extraction of endmembers corresponding to unique sources observed in the HSI. Here the GTM is an attractive choice as it does not rely on the assumption of linear mixing and the presence of pure pixels which are easily broken by the effects of multiple scattering in realistic scenarios [45]. Additionally, because the GTM is a probabilistic model, the values of model hyperparameters can be selected objectively using information criteria like the BIC. This is a clear advantage over similar methods like the SOM on which the GTM is based. Endmembers identified for exemplar spectra corresponding to rhodamine, algae, grass, and

water demonstrate that the method can successfully identify spectral signatures corresponding to diverse sources. If a set of labeled spectra for known sources are available, their representation in the GTM latent space could be used to perform as semi-supervised classification similar to the SOM approach developed by Riese et al. [26].

Once endmembers are identified, estimating their abundance using the NS3 provides a quick method to map the distribution of sources in water. We note that the spatial distribution of algae mapped in Figure 8 realistically reflects clustering of algae near the shore. Additionally, the ability for the UAV to quickly survey the same area in rapid succession is highly advantageous for tracking the diffusion of point sources as demonstrated in Figure 9 for the rhodamine dye release.

The primary limitation of the GTM is that the number of nodes increases exponentially with the dimension of the latent space. However when constrained to two dimensions in order to enable visualization of the resulting map, the number of latent nodes has a negligible impact compared to the size of the dataset on which it is trained. Additionally, the GTM considers individual HSI pixels and does not exploit spatial structure like other methods such as convolutional autoencoders and non-negative matrix factorization using super-pixels [17,21]. Extensions to the GTM have been proposed to enable batch training for large datasets as well as manifold-aligned noise models which replace the fixed variance parameter  $\beta^{-1}$  with a full covariance matrix [46].

Finally, we note that the representation obtained by the GTM can be used for non-linear feature extraction to improve supervised models. Traditionally, PCA is used as a common preprocessing technique to reduce high-dimensional HSI by keeping the first  $r$ -many principal components. For example, Uddin et al. report improved classification of HSI by using PCA to extract features for a support vector machine [47]. The GTM can similarly be used to provide a sparse representation of the input data via the latent node responsibilities  $R_{kn}$  obtained for each record.

## 5. Conclusions

In this study we present the GTM as an useful unsupervised method for the visualization of UAV-based hyperspectral imagery and associated extraction of spectral endmembers. Using data collected at a North Texas pond, we demonstrate how the latent space of the GTM can be used to visualize the distribution of observed reflectance spectra revealing the small-scale spatial variability of water composition. Spectral signatures extracted from GTM nodes are used to successfully map the abundance of algae near the shore and to track the evolution of a rhodamine tracer dye plume. These examples illustrate the power of combining unsupervised learning with UAV-based HSI collection for the characterization of water composition.

**Author Contributions:** Methodology, J.W. and D.J.L.; conceptualization, D.J.L.; software, J.W.; field deployment and preparation, J.W., A.A., L.O.H.W., S.T., B.F., P.M.H.D., M.I., M.L., D.S., G.B., and D.J.L.; validation, J.W.; formal analysis, J.W.; investigation J.W.; resources, D.J.L.; data curation, J.W., A.A., L.O.H.W., and D.J.L.; writing—original draft preparation, J.W.; writing—review and editing, J.W. and D.J.L.; visualization, J.W.; supervision, D.J.L.; project administration, D.J.L.; funding acquisition, D.J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the following grants: the Texas National Security Network Excellence Fund award for Environmental Sensing Security Sentinels; the SOFWERX award for Machine Learning for Robotic Teams and NSF Award OAC-2115094; support from the University of Texas at Dallas Office of Sponsored Programs, Dean of Natural Sciences and Mathematics, and Chair of the Physics Department is gratefully acknowledged; TRECIS CC\* Cyberteam (NSF #2019135); NSF OAC-2115094 Award; and EPA P3 grant number 84057001-0.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** The original data presented in the study are made openly available via the Open Storage Network at <https://ncsa.osn.xsede.org/ees230012-bucket01/RobotTeam/unsupervised/gtm-data/>. The open-source implementation of the GTM used in this study is available at <https://github.com/john-waczak/GenerativeTopographicMapping.jl>.

**Acknowledgments:** Don MacLaughlin, Scotty MacLaughlin, and the city of Plano, TX, are gratefully acknowledged for allowing us to deploy the autonomous robot team on their property. Christopher Simmons is gratefully

acknowledged for his computational support. We thank Antonio Mannino for his advice with regard to selecting the robotic boat’s sensing suite. Annette Rogers is gratefully acknowledged for supporting the arrangement of insurance coverage. Steven Lyles is gratefully acknowledged for supporting the arrangement of a secure place for the robot team. The authors acknowledge the OIT-Cyberinfrastructure Research Computing group at the University of Texas at Dallas and the TRECIS CC\* Cyberteam (NSF #2019135) for providing HPC resources that contributed to this research; the authors also acknowledge their receipt of the NSF OAC-2115094 Award and EPA P3 grant number 84057001-0.

**Conflicts of Interest:** The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

- UAV      Unmanned Aerial Vehicle
- GTM      Generative Topographic Mapping
- SOM      Self Organizing Map
- HSI      Hyperspectral Image
- PCA      Principal Component Analysis
- tSNE     t-Distributed Stochastic Neighbor Embedding
- MLJ      Machine Learning in Julia
- VNIR     Visible + Near-Infrared
- NDWI    Normalized Difference Water Index
- NS3      Normalized Spectral Similarity Score

Appendix A. Hyperparameter Search Results

**Table A1.** The top 25 models from the hyperparameter search. A variety of GTM were trained to explore the the impact of varying  $m$ ,  $\alpha$ , and  $s$ . The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are given in the final two columns which can be used for hyperparameter selection.

m	$\alpha$	s	k	BIC	AIC
14	0.1	1.0	32	-1.918e8	-1.926e8
13	0.01	1.0	32	-1.917e8	-1.923e8
16	0.01	1.5	32	-1.917e8	-1.926e8
14	10.0	1.0	32	-1.917e8	-1.924e8
16	0.001	1.5	32	-1.917e8	-1.926e8
13	1.0	1.0	32	-1.917e8	-1.923e8
13	10.0	1.0	32	-1.917e8	-1.923e8
14	0.001	1.5	32	-1.916e8	-1.924e8
13	0.1	1.0	32	-1.916e8	-1.923e8
14	0.01	1.0	32	-1.916e8	-1.924e8
15	0.01	1.5	32	-1.916e8	-1.925e8
14	0.01	1.5	32	-1.916e8	-1.923e8
15	1.0	1.0	32	-1.916e8	-1.924e8
18	0.01	1.5	32	-1.916e8	-1.928e8
12	0.01	1.0	32	-1.916e8	-1.921e8
15	0.01	0.5	32	-1.915e8	-1.924e8
17	1.0	1.0	32	-1.915e8	-1.926e8
16	0.1	1.0	32	-1.915e8	-1.925e8
18	0.001	1.5	32	-1.915e8	-1.928e8
13	0.001	1.0	32	-1.915e8	-1.922e8
12	1.0	1.0	32	-1.915e8	-1.921e8
17	0.001	1.5	32	-1.915e8	-1.926e8
15	0.001	1.5	32	-1.915e8	-1.923e8
15	10.0	1.0	32	-1.915e8	-1.923e8
12	0.1	1.5	32	-1.915e8	-1.92e8

## References

1. Koponen, S.; Pulliainen, J.; Kallio, K.; Hallikainen, M. Lake water quality classification with airborne hyperspectral spectrometer and simulated MERIS data. *Remote Sensing of Environment* **2002**, *79*, 51–59.
2. Ritchie, J.C.; Zimba, P.V.; Everitt, J.H. Remote sensing techniques to assess water quality. *Photogrammetric engineering & remote sensing* **2003**, *69*, 695–704.
3. Adão, T.; Hruška, J.; Pádua, L.; Bessa, J.; Peres, E.; Morais, R.; Sousa, J.J. Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote sensing* **2017**, *9*, 1110.
4. Arroyo-Mora, J.P.; Kalacska, M.; Inamdar, D.; Soffer, R.; Lucanus, O.; Gorman, J.; Naprstek, T.; Schaaf, E.S.; Ifimov, G.; Elmer, K.; others. Implementation of a UAV-hyperspectral pushbroom imager for ecological monitoring. *Drones* **2019**, *3*, 12.
5. Banerjee, B.P.; Raval, S.; Cullen, P. UAV-hyperspectral imaging of spectrally complex environments. *International Journal of Remote Sensing* **2020**, *41*, 4136–4159.
6. Horstrand, P.; Guerra, R.; Rodríguez, A.; Díaz, M.; López, S.; López, J.F. A UAV platform based on a hyperspectral sensor for image capturing and on-board processing. *IEEE Access* **2019**, *7*, 66919–66938.
7. Vogt, M.C.; Vogt, M.E. Near-remote sensing of water turbidity using small unmanned aircraft systems. *Environmental Practice* **2016**, *18*, 18–31.
8. Zhang, D.; Zeng, S.; He, W. Selection and quantification of best water quality indicators using UAV-mounted hyperspectral data: a case focusing on a local river network in Suzhou City, China. *Sustainability* **2022**, *14*, 16226.
9. Keller, S.; Maier, P.M.; Riese, F.M.; Norra, S.; Holbach, A.; Börsig, N.; Wilhelms, A.; Moldaenke, C.; Zaake, A.; Hinz, S. Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. *International journal of environmental research and public health* **2018**, *15*, 1881.
10. Lu, Q.; Si, W.; Wei, L.; Li, Z.; Xia, Z.; Ye, S.; Xia, Y. Retrieval of water quality from UAV-borne hyperspectral imagery: A comparative study of machine learning algorithms. *Remote Sensing* **2021**, *13*, 3928.
11. Lary, D.J.; Schaefer, D.; Waczak, J.; Aker, A.; Barbosa, A.; Wijeratne, L.O.; Talebi, S.; Fernando, B.; Sadler, J.; Lary, T.; others. Autonomous learning of new environments with a robotic team employing hyper-spectral remote sensing, comprehensive in-situ sensing and machine learning. *Sensors* **2021**, *21*, 2240.
12. Waczak, J.; Aker, A.; Wijeratne, L.O.; Talebi, S.; Fernando, B.; Hathurusinghe, P.; Iqbal, M.; Schaefer, D.; Lary, D.J. Characterizing Water Composition with an Autonomous Robotic Team Employing Comprehensive In-Situ Sensing, Hyperspectral Imaging, Machine Learning, and Conformal Prediction **2024**.
13. Tyo, J.S.; Konsolakis, A.; Diersen, D.I.; Olsen, R.C. Principal-components-based display strategy for spectral imagery. *IEEE transactions on geoscience and remote sensing* **2003**, *41*, 708–718.
14. Zhang, B.; Yu, X. Hyperspectral image visualization using t-distributed stochastic neighbor embedding. MIPPR 2015: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications. SPIE, 2015, Vol. 9815, pp. 14–21.
15. Heylen, R.; Parente, M.; Gader, P. A review of nonlinear hyperspectral unmixing methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2014**, *7*, 1844–1868.
16. Nascimento, J.M.; Dias, J.M. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing* **2005**, *43*, 898–910.
17. Feng, X.R.; Li, H.; Wang, R.; Du, Q.; Jia, X.; Plaza, A.J. Hyperspectral Unmixing Based on Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2022**, *15*, 4414–4436.
18. Cariou, C.; Chehdi, K. Unsupervised nearest neighbors clustering with application to hyperspectral images. *IEEE Journal of Selected Topics in Signal Processing* **2015**, *9*, 1105–1116.
19. Su, Y.; Li, J.; Plaza, A.; Marinoni, A.; Gamba, P.; Chakravortty, S. DAEN: Deep autoencoder networks for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *57*, 4309–4321.
20. Borsoi, R.A.; Imbiriba, T.; Bermudez, J.C.M. Deep generative endmember modeling: An application to unsupervised spectral unmixing. *IEEE Transactions on Computational Imaging* **2019**, *6*, 374–384.
21. Palsson, B.; Ulfarsson, M.O.; Sveinsson, J.R. Convolutional autoencoder for spectral-spatial hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing* **2020**, *59*, 535–549.
22. Kohonen, T. The self-organizing map. *Proceedings of the IEEE* **1990**, *78*, 1464–1480.



23. Cantero, M.; Perez, R.; Martinez, P.J.; Aguilar, P.; Plaza, J.; Plaza, A. Analysis of the behavior of a neural network model in the identification and quantification of hyperspectral signatures applied to the determination of water quality. *Chemical and Biological Standoff Detection II*. SPIE, 2004, Vol. 5584, pp. 174–185.
24. Duran, O.; Petrou, M. A time-efficient method for anomaly detection in hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing* **2007**, *45*, 3894–3904.
25. Ceylan, O.; Kaya, G.T. Feature Selection Using Self Organizing Map Oriented Evolutionary Approach. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS 2021*, pp. 4003–4006.
26. Riese, F.M.; Keller, S.; Hinz, S. Supervised and semi-supervised self-organizing maps for regression and classification focusing on hyperspectral data. *Remote Sensing* **2019**, *12*, 7.
27. Danielsen, A.S.; Johansen, T.A.; Garrett, J.L. Self-organizing maps for clustering hyperspectral images on-board a cubesat. *Remote Sensing* **2021**, *13*, 4174.
28. Bishop, C.M.; Svensén, M.; Williams, C.K. GTM: The generative topographic mapping. *Neural computation* **1998**, *10*, 215–234.
29. Kireeva, N.; Baskin, I.; Gaspar, H.; Horvath, D.; Marcou, G.; Varnek, A. Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison. *Molecular informatics* **2012**, *31*, 301–312.
30. Gaspar, H.A.; Baskin, I.I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *Journal of chemical information and modeling* **2015**, *55*, 84–94.
31. Horvath, D.; Marcou, G.; Varnek, A. Generative topographic mapping in drug design. *Drug Discovery Today: Technologies* **2019**, *32*, 99–107.
32. Waczak, J. GenerativeTopographicMapping.jl, 2024. doi:10.5281/zenodo.11061258.
33. Bezanson, J.; Karpinski, S.; Shah, V.B.; Edelman, A. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145* **2012**.
34. Blaom, A.D.; Kiraly, F.; Lienart, T.; Simillides, Y.; Arenas, D.; Vollmer, S.J. MLJ: A Julia package for composable machine learning. *arXiv preprint arXiv:2007.12285* **2020**.
35. Ruddick, K.G.; Voss, K.; Banks, A.C.; Boss, E.; Castagna, A.; Frouin, R.; Hieronymi, M.; Jamet, C.; Johnson, B.C.; Kuusk, J.; others. A review of protocols for fiducial reference measurements of downwelling irradiance for the validation of satellite remote sensing data over water. *Remote Sensing* **2019**, *11*, 1742.
36. Muller, R.; Lehner, M.; Muller, R.; Reinartz, P.; Schroeder, M.; Vollmer, B. A program for direct georeferencing of airborne and spaceborne line scanner images. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* **2002**, *34*, 148–153.
37. Bäumker, M.; Heimes, F. New calibration and computing method for direct georeferencing of image and scanner data using the position and angular data of an hybrid inertial navigation system. *OEEPE Workshop, Integrated Sensor Orientation*, 2001, pp. 1–17.
38. Mostafa, M.M.; Schwarz, K.P. A multi-sensor system for airborne image capture and georeferencing. *Photogrammetric engineering and remote sensing* **2000**, *66*, 1417–1424.
39. Nidamanuri, R.R.; Zbell, B. Normalized Spectral Similarity Score (NS<sup>3</sup>) as an Efficient Spectral Library Searching Method for Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2011**, *4*, 226–240.
40. Thenkabail, P.S.; Lyon, J.G.; Huete, A. *Hyperspectral indices and image classifications for agriculture and vegetation*; CRC press, 2018.
41. Alvarez-Vanhard, E.; Houet, T.; Mony, C.; Lecoq, L.; Corpetti, T. Can UAVs fill the gap between in situ surveys and satellites for habitat mapping? *Remote Sensing of Environment* **2020**, *243*, 111780.
42. Gu, Y.; Huang, Y.; Liu, T. Intrinsic Decomposition Embedded Spectral Unmixing for Satellite Hyperspectral Images With Endmembers From UAV Platform. *IEEE Transactions on Geoscience and Remote Sensing* **2023**.
43. Balas, E. The prize collecting traveling salesman problem and its applications. In *The traveling salesman problem and its variations*; Springer, 2007; pp. 663–695.
44. Suryan, V.; Tokekar, P. Learning a spatial field in minimum time with a team of robots. *IEEE Transactions on Robotics* **2020**, *36*, 1562–1576.

45. Han, T.; Goodenough, D.G. Investigation of nonlinearity in hyperspectral remotely sensed imagery — a nonlinear time series analysis approach. 2007 IEEE International Geoscience and Remote Sensing Symposium, 2007, pp. 1556–1560. doi:10.1109/IGARSS.2007.4423107.
46. Bishop, C.M.; Svensén, M.; Williams, C.K. Developments of the generative topographic mapping. *Neuro-computing* **1998**, *21*, 203–224.
47. Uddin, M.P.; Mamun, M.A.; Hossain, M.A. PCA-based feature reduction for hyperspectral remote sensing image classification. *IETE Technical Review* **2021**, *38*, 377–396.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.