A Transformer-Based Framework for Spatiotemporal Unmixing of Land–Water Mixtures in Multispectral Satellite Data

An Bao Nguyen[1], Andreas Schenk[2], Stefan Hinz[2]

[1] Mechatronics, Biostatistics and Sensors, KU Leuven, Belgium
anbao.nguyen@kuleuven.be
[2] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Germany
(andreas.schenk, stefan.hinz)@kit.edu

Keywords: Spectral Unmixing, Spectral Variablity, Transformer, Variational Autoencoder, Spectral Analysis, Water Monitoring.

Abstract

Spectral unmixing is essential for analyzing mixed pixels in remote sensing, though it has traditionally focused on hyperspectral data. Multispectral Sentinel-2 imagery, despite its wide availability and relevance for environmental monitoring, has seen limited application in this domain and is affected by spectral variability caused by environmental conditions, atmospheric residuals, and temporal changes, which are often neglected in existing methods. We propose the time-dependent Deep Transformer MultiSpectral Unmixing Model (tDTMSUM), a multimodal deep generative framework designed to extract pure water spectra from mixed Sentinel-2 observations, particularly in narrow rivers where water pixels are frequently mixed with adjacent land. The model integrates Sentinel-2 reflectance with auxiliary variables contributing to spectral variability, including the geographical position of water bodies, to capture the spatial dynamic transition of water properties. For example, in the study area in this work, the model successfully detected the change of the water body from standing water in the southern reservoir to sediment-laden flowing water in the northern river. tDTMSUM combines a Variational Autoencoder with a channel-wise Transformer and is trained on augmented dataset derived by synthetic mixtures from real Sentinel-2 data to perform supervised endmember extraction and abundance estimation, focusing on the water endmember. Evaluation using Sentinel-2 imagery and close range spectrometer measurements demonstrates that tDTMSUM outperforms state-of-the-art methods in efficiency, robustness, and accuracy, providing a practical tool for real-world water monitoring even in the absence of extensive ground truth.

1. Introduction

Sentinel-2 is a pair of Earth observation satellites developed by the European Space Agency (ESA) under the Copernicus Program. It provides multispectral imagery across 13 spectral channels, ranging from visible to shortwave infrared, with different spatial resolutions of 10, 20, and 60 meters across the channels. The spectral information reveals the properties of the Earth’s surface, and environmental processes, making Sentinel-2 imagery a widely used and reliable data source in many domains, such as agriculture (Van Tricht et al., 2018), forestry (Immitzer et al., 2016), environmental monitoring (Bastani et al., 2023) (Phiri et al., 2020), and water monitoring (Llodra-Llabres et al., 2023) (Tian et al., 2023) (Sent et al., 2021). The medium resolution between 10 and 60 meters is appropriate for many environmental applications, but it may lack sufficient detail for assessing characteristics of small linear features such as rivers and infrastructure lines. Particularly in those cases, each spectral pixel covers several materials or land cover types on the ground due to its coarse spatial resolution. The measured spectral reflectance is then a mixture of pure spectra (*endmembers*) of the materials within the pixel, weighted by their fractional proportions (*abundances*), as modeled by the widely used *linear mixing model* (LMM) (Keshava and Mustard, 2002). This spectral mixing often degrades classification accuracy and hinders environmental monitoring. For instance, assessing the water quality of narrow rivers is often challenging, as water pixels are contaminated by spectral signatures from the adjacent riverbanks.

Spectral unmixing addresses this issue by decomposing each mixed pixel into its constituent endmember spectra and their corresponding abundances. The abundances are physically constrained by the Abundance Nonnegativity Constraint (ANC), which ensures that no abundance is negative, and the Abundance Sum-to-One Constraint (ASC), which enforces that all abundances sum up to one. Furthermore, since endmember spectra represent surface reflectance, they must satisfy the Endmember Nonnegativity Constraint, which requires reflectance values to be nonnegative and bounded to one.

Recently, deep learning-based networks have been introduced for spectral unmixing (Ghosh et al., 2022) (Bhakthan and Loganathan, 2024), particularly for hyperspectral data. However, most of the methods assume fixed endmember spectra and fail to capture spectral variability caused by factors such as residual atmospheric effects (even after atmospheric correction), topographical influences, shadows cast by clouds and terrain (Borsoi et al., 2021), temporal changes, and the natural dynamics of observed endmembers. Only a few studies have attempted to address this critical challenge (Shi et al., 2021), but overlooked the impact of shadows and temporal variations. Fig. 1 illustrates how such variability affects even homogeneous endmembers across time. Additionally, the water body in the observed area transitions from sediment-laden, flowing water in the northern river to clear, standing water in the southern reservoir, as evident in both the RGB image and the corresponding spectral signatures.

Traditional solutions, e.g., illumination scaling or parametric models like the Hapke model, are often simplistic and condition-dependent (Drumetz et al., 2016). In contrast, deep probabilistic generative models have recently gained significant attention for their ability to represent complex spectral distributions. These models can effectively capture the statist-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-573-2026 | © Author(s) 2026. CC BY 4.0 License.

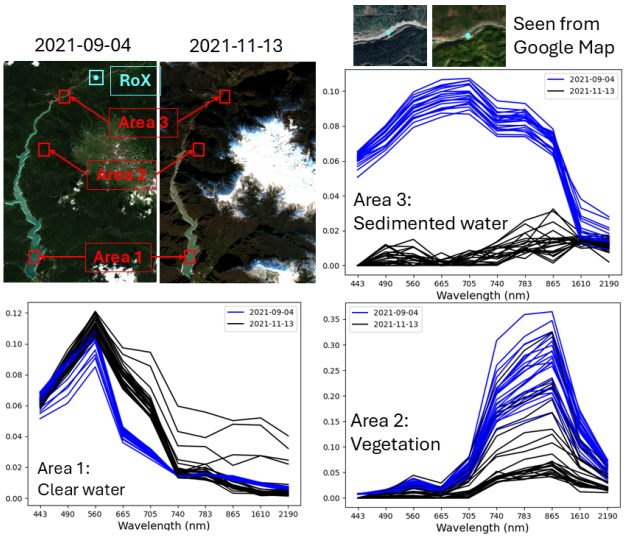

Figure 1. Spectral variability is ubiquitous in the observed area. Pixels in red rectangle regions composed of a single land cover type. The location of RoX is marked with the red circle.

ical characteristics of spectral variability and generate different spectral variants from the learned distribution. However, most existing methods rely solely on spectral inputs, neglecting auxiliary information that could further improve the modeling of variability. To address these limitations, we propose the time-dependent Deep Transformer MultiSpectral Unmixing Model (tDTMSUM), a multimodal deep generative framework designed to explicitly account for spectral variability in Sentinel-2 Level-2A data, with a particular focus on modeling the spatial and temporal spectral variations of the water body. This approach facilitates the generation of more accurate pixel-specific endmember spectra and higher-quality abundance maps. The main contribution of the proposed methodology is summarized as follows:

- A novel two-stream deep learning model based on a Variational Autoencoder (VAE) with a channel-wise transformer applied in both encoder and decoder. The encoder simultaneously estimates abundances and learns the latent distribution of the spectral input, while the decoder maps the latent variables to the corresponding endmember spectra, which are then linearly mixed with estimated abundances to reconstruct the spectral input according to the LMM. The model is trained in a supervised manner using synthetic mixtures generated from a small number of reference days and can be generalized across scenes within the same study period.
- The integration of auxiliary variables that account for sources of spectral variability as additional inputs alongside the spectral data. The auxiliary variables reflect the phenological cycle (day of year), atmospheric changes (aerosol optical thickness and water vapor pressure), varying illumination/viewing geometry (azimuth and zenith angle of sensor and sun), and the spatial context, represented by the geographic latitude of the primarily north-south oriented water body. Their relationship is captured by the transformer in the encoder through a new parallel channel-wise attention mechanism.
- An automated endmember extraction and synthetic mixing pipeline directly from Sentinel-2 imagery to generate supervised augmented training data.

The remainder of this article is organized as follows: Section 2 details the components of the proposed workflow. Section 3 presents extensive experiments on real Sentinel-2 data, including performance comparisons between the proposed approach and state-of-the-art unmixing models, as well as detailed evaluations of the extracted water spectra against in-situ RoX measurements, and assesses the proposed model’s capability to capture the temporal-spatial spectral and abundance variation of the water body. Finally, comprehensive conclusions and future outlook are drawn in Section 4.

2. Proposed Methodology

The proposed spectral unmixing methodology is illustrated in Fig. 2. For proving the validity of the framework and assessing its potentials, we applied it to the Enguri region (Fig. 1) during the period from 2021 to 2024. The study area, located in the southern Caucasus Mountain range of Georgia, spans approximately 21 × 18 km and encompasses the Enguri Reservoir, the Enguri Dam at the southern end, and the upstream Enguri River extending northeastward, where sediment-laden water is commonly observed. The Sentinel-2 dataset comprises 101 scenes acquired between September 2021 and October 2024, each containing less than 20% cloud coverage. Four endmembers are considered in the analysis: vegetation (covering approximately 82% of the study area, consisting of all vegetation types present in the scene), non-vegetation (3%), sediment-laden water (3%), and clear water (12%). Since the objective of this study is to extract water spectra from mixed pixels, and the proportion of non-vegetation is considerably small, all non-vegetation features, such as soil and built-up areas, are grouped into a single endmember class non-vegetation.

2.1 Linear Mixing Model

Let the Sentinel-2 multispectral imaging (MSI) to be denoted as $\mathbf{Y} \in \mathbb{R}^{N \times B}$, where N is the number of pixels and B is the number of spectral bands. The endmember matrix will be denoted as $\mathbf{E} \in \mathbb{R}^{R \times B}$, where R represents the number of endmembers present in the MSI. The corresponding abundance matrix will be denoted as $\mathbf{A} \in \mathbb{R}^{N \times R}$. In the Linear Mixing Model (LMM), the observed spectral reflectance is formulated as:

$$\mathbf{Y} = \mathbf{AE} + \mathbf{N} \tag{1}$$

where $\mathbf{N} \in \mathbb{R}^{N \times B}$ is the additive Gaussian noise present in $\mathbf{Y}$. Three physical constraints must be satisfied: $\mathbf{E} \geq 0$, $\mathbf{A} \geq 0$ (ANC), and $\mathbf{1}_R^T \mathbf{A}^T = \mathbf{1}_N^T$ (ASC) where $\mathbf{1}_N^T$ indicates a vector of ones with N entries.

2.2 Data Preprocessing

The input data includes 20 m spectral data (bands B02, B03, B04, B05, B06, B07, B08, B11, B12), auxiliary data (B01, AOT - Aerosol Optical Thickness, WVP - Water Vapor Pressure, solar and sensor azimuth and zenith angles, latitude coordinates of the water body), and acquisitions (DOY). All spectral and auxiliary data were cropped to the study area, and invalid pixels were removed. Auxiliary variables were resampled to 20 m spatial resolution using bilinear interpolation. To account for seasonal variability, the DOY was represented using sine and cosine transformations.

$$\mathrm{DOY}_{\sin} = \sin\left(2\pi \frac{d}{D}\right), \quad \mathrm{DOY}_{\cos} = \cos\left(2\pi \frac{d}{D}\right) \tag{2}$$

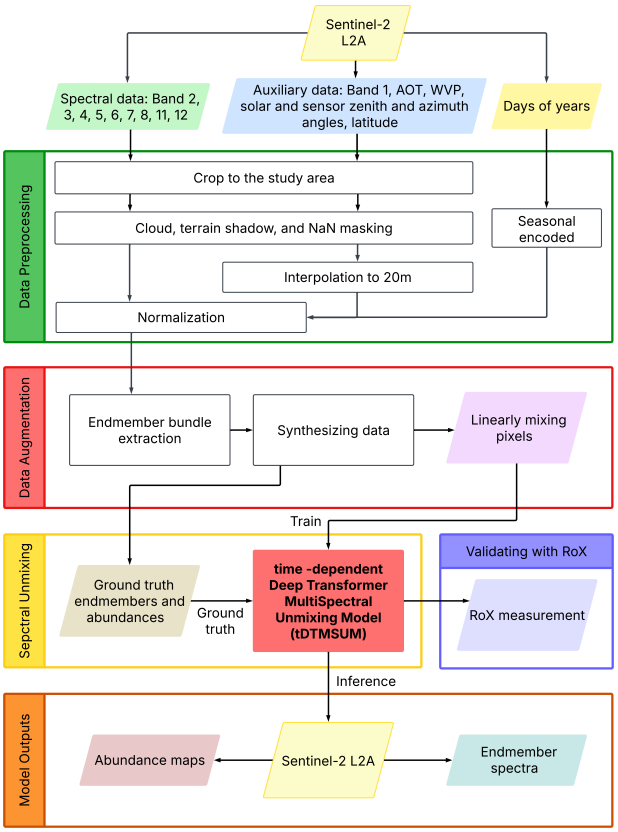

Figure 2. The flowchart of the methodology in this study.

where $d \in [1, D]$ is the day of the year, and $D = 365$ or 366 depending on the year.

2.3 Data Augmentation

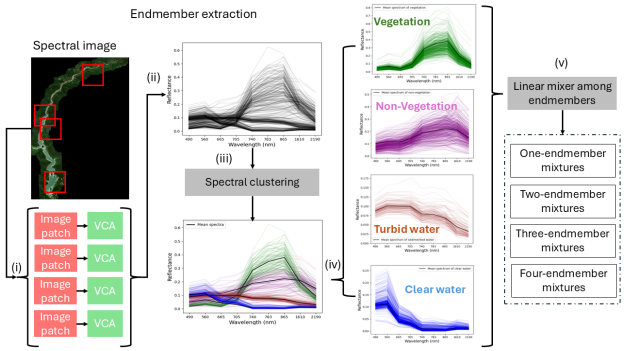

Figure 3. Processing pipeline for automatic construction of the four endmember libraries from a spectral image.

In Sentinel-2 and most satellite imagery, true endmember abundances are not available, and pixels are potentially mixed, making the actual material contributions impossible to measure directly. Many recent spectral unmixing models like (Ghosh et al., 2022), (Shi et al., 2021), and (Meng et al., 2024) rely on unsupervised reconstruction-based objectives, which work well for hyperspectral data due to their high spectral resolution and large number of bands. However, this approach is less suitable for multispectral sensors like Sentinel-2, where the limited spectral information makes different materials appear more similar and increases ambiguity in mixed pixels. To guide the model toward physically meaningful endmember spectra and abundances, supervised training signals are required. Therefore, synthetic mixtures with known abundances were generated from endmember spectra extracted from the study scene. Since these endmember spectra vary with illumination, viewing geometry, water suspended sediment concentration, and topography, data augmentation was applied to simulate natural spectral variability. This enlarged synthetic dataset enables the model to learn a stable mixing relationship and to produce reliable abundance estimates for real Sentinel-2 pixels despite the absence of true ground-truth labels.

A spectral library of four endmember bundles was constructed using a sliding kernel to extract patches from MSI, and within each patch, Vertex Component Analysis (VCA) was employed to extract local endmembers (Somers et al., 2012). To incorporate distorted spectra caused by auxiliary factors such as residual atmospheric effects and shadows, the Spectral Angle Distance (SAD) between distorted and undistorted spectra was restricted to a maximum threshold of 0.3 radians in this study. Increasing the threshold would enhance spectral diversity, but would also raise the risk of including dissimilar spectra. The extracted spectra were then clustered into four endmember bundles using k-means clustering with a combined similarity measure based on SAD and the first spectral derivative. Based on these spectral bundles, synthetic linear mixtures were generated under four scenarios: single endmembers and mixtures of two, three, and four endmembers. To satisfy ANC and ASC, abundances were drawn from a Dirichlet distribution. This process, illustrated in Fig. 3, was applied to available Sentinel-2 scenes from each season across the study years, considering only scenes with less than 20% cloud coverage. The process can be summarized as: (i)–(ii) local endmembers were extracted, (iii)–(iv) extracted endmember spectra were clustered to obtain four endmember bundles, and (v) spectra from four bundles were mixed to create augmented data.

2.4 Spectral Unmixing

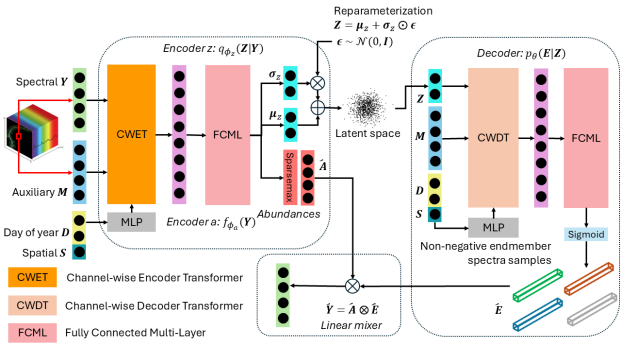

Figure 4. The framework for the proposed tDTMSUM.

The architecture of the proposed deep MSI unmixing model is shown in Fig. 4. It consists of three main modules: an encoder, a decoder, and a linear mixer.

Encoder: The encoder adopts a dual-stream design for multi-task learning: to generate abundances and endmember spectra simultaneously. One stream encodes the latent variables $\boldsymbol{\mu}_z$, $\boldsymbol{\sigma}_z \in \mathbb{R}^{N \times L_Z}$, where L_Z is the dimension of the latent space, for endmember generation via a VAE, while the other performs dimensionality reduction to estimate the abundance vector $\hat{\mathbf{A}} \in \mathbb{R}^{N \times R}$ using Sparsemax activation (Martins and Astudillo, 2016) to enforce ANC and ASC. Both streams share a backbone composed of a Channel-wise Encoder Transformer (CWET), which helps the model learn the influence of the aux-

iliary factors on the spectra via spectral channels, and a Fully Connected Multi-Layer (FCML). The module takes four inputs: spectral data $\mathbf{Y} \in \mathbb{R}^{N \times L_Y}$, auxiliary data $\mathbf{M} \in \mathbb{R}^{N \times L_M}$, day-of-year (DOY) data $\mathbf{D} \in \mathbb{R}^{N \times L_D}$, and spatial (latitude) data $\mathbf{S} \in \mathbb{R}^{N \times L_S}$. Here, L_Y, L_M, L_D, and L_S represent the number of spectral bands, auxiliary factors, temporal features, and spatial features, respectively. In CWET, $\mathbf{Y}$ and $\mathbf{M}$ are first embedded to the embedding space N_{emb} and assigned a fixed positional embedding. Self-attention is applied to $\mathbf{Y}$, followed by cross-attention between $\mathbf{Y}$ and $\mathbf{M}$. This cross-attention mechanism reveals which auxiliary factors contribute to variations in the spectra, and how each spectral channel is influenced by a given auxiliary factor. The resulting representations are aggregated using global average pooling across the embedding dimension. The DOY ($\mathbf{D}$) and spatial ($\mathbf{S}$) features are then embedded via a lightweight MLP to match this dimensionality and are added to the aggregated output. Finally, the combined representation is refined through the FCML, composed of sequential fully connected layers, each followed by batch normalization, ReLU activation, and dropout.

Decoder: The decoder generates corresponding endmember spectra from the latent representation $\mathbf{Z} \in \mathbb{R}^{N \times L_Z}$, derived from $\boldsymbol{\mu}_z$ and $\boldsymbol{\sigma}_z$ via the reparameterization trick (Kingma et al., 2019). The decoder comprises a Channel-wise Decoder Transformer (CWDT) and a FCML similar to the one in the encoder. In CWDT, $\mathbf{Z}$ and $\mathbf{M}$ are first embedded to the embedding space N_{emb}. A fixed positional embedding is applied only to $\mathbf{M}$, as $\mathbf{Z}$ represents a compressed, orderless latent vector. Then, self-attention is applied to $\mathbf{Z}$, followed by cross-attention from $\mathbf{M}$ to $\mathbf{Z}$. The outputs of the CWDT are processed similarly to those of the CWET, with global aggregation and addition with embedded $\mathbf{D}$ and $\mathbf{S}$ and feature refinement performed by the FCML. Finally, a Sigmoid activation is applied to the FCML output to produce the nonnegative endmember spectra tensor $\hat{\mathbf{E}} \in \mathbb{R}^{N \times R \times L_Y}$, representing the estimated endmember signatures for each pixel.

Linear Mixer: Given the estimated abundance matrix $\hat{\mathbf{A}} \in \mathbb{R}^{N \times R}$ and the extracted endmember spectra tensor $\hat{\mathbf{E}} \in \mathbb{R}^{N \times R \times L_Y}$, the reconstructed input $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times L_Y}$ is obtained by performing a batched linear combination of the endmembers weighted by their corresponding abundances. For each sample $\hat{\mathbf{y}}_i \in \mathbb{R}^{1 \times L_Y}, i = 1, 2, \cdots, N$:

$$\hat{\mathbf{y}}_i = \hat{\mathbf{A}}_i \cdot \hat{\mathbf{E}}_i \tag{3}$$

where $\hat{\mathbf{A}}_i \in \mathbb{R}^{1 \times R}, i = 1, 2, \cdots, N$ and $\hat{\mathbf{E}}_i \in \mathbb{R}^{R \times L_Y}, i = 1, 2, , \cdots, N$ are the abundance and endmember matrix, respectively, of the *i*th pixel.

2.5 Loss Function

The proposed model was trained end-to-end by minimizing a composite loss function $\mathcal{L}_{total}$, which is a weighted sum of several loss terms entailing the loss reconstruction loss, the abundance loss, the endmember's spectral loss, and the Kullback–Leibler (KL) divergence. Since the model was trained on the synthetic data, the ground truth for both abundances and endmember spectra was available. In general, the total loss of the model is:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_1 \cdot \text{MSE}(\mathbf{Y}, \hat{\mathbf{Y}}) + \lambda_2 \cdot \text{SAD}(\mathbf{Y}, \hat{\mathbf{Y}}) \\ + \lambda_3 \cdot \text{MSE}(\mathbf{A}, \hat{\mathbf{A}}) + \lambda_4 \cdot \text{mMSE}(\mathbf{E}, \hat{\mathbf{E}}) \\ + \lambda_5 \cdot \text{mSAD}(\mathbf{E}, \hat{\mathbf{E}}) + \lambda_6 \cdot \mathcal{L}_{\text{KL}}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z)\end{aligned} \tag{4}$$

where $\lambda_i, i = 1, 2, \cdots, 6$ are the hyperparameters controlling the contribution of each term in the total loss. MSE and SAD are Mean Squared Error and the Spectral Angle Distance. mMSE and mSAD stand for mean MSE and mean SAD, respectively, which compute the mean value across the second dimension (R) as $\hat{\mathbf{E}}$ and $\mathbf{E}$ are 3D matrices of $\mathbb{R}^{N \times R \times L_Y}$. On the left side of Eq. 4, the first two terms refer to the reconstruction losses, while the fourth and the fifth terms are the endmember's spectral losses. The third term is the abundance loss, and $\mathcal{L}_{\text{KL}}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z)$ is known as the KL divergence, a regularization in VAE. Given the matrix of the mean and standard deviation $\boldsymbol{\mu}_z$ and $\boldsymbol{\sigma}_z \in \mathbb{R}^{N \times L_Z}$, the KL-divergence is calculated as:

$$\mathcal{L}_{KL} = -\frac{1}{2N} \sum_{n=1}^{N} \sum_{d=1}^{L_Z} \left(1 + \log \sigma_{nd}^2 - \mu_{nd}^2 - \sigma_{nd}^2\right) \tag{5}$$

2.6 Validation with Hyperspectral RoX Measurements

RoX (reflectance box from JB Hyperspectral Devices GmbH) is a field spectrometer that records reflected radiation across wavelengths from 341 to 1016 nm, comprising a total of 1021 spectral channels. It was installed on the bridge upstream of the Enguri River (indicated by the blue dot in Fig. 1) to measure the reflectance of the Enguri River's sediment-laden water. The water spectra extracted from the spectral unmixing model were compared with the corresponding close range measurement acquired on the same day with the RoX instrument. The unmixing was performed at the device's location, which is a mixed pixel in Sentinel-2 imagery. For this comparison, the RoX spectra were spectrally resampled to the Sentinel-2 band configuration using Sentinel-2's Spectral Response Functions (SRF).

3. Experimental Results

3.1 Experimental Setup

The model was trained on augmented spectral mixtures obtained from Subsection 2.3, preprocessed as indicated in Subsection 2.2. Hyperparameters were optimized via grid search. Training was done over 200 epochs using the OneCycle scheduler (Smith and Topin, 2019), with Early Stopping (patience = 20) and Adam optimizer (weight decay = $4e-5$). The composite loss uses weights $\lambda_1 = 40.0$, $\lambda_2 = 5.0$, $\lambda_3 = 30.0$, $\lambda_4 = 70.0$, $\lambda_5 = 9.0$, and $\lambda_6 = 1e-3$. The embedding N_{emb} and latent L_Z dimensions are 16 and 4, respectively. In Subsection 3.2, the comparison between different versions of the proposed model is conducted, whereas Subsection 3.3 compares the performance of the proposed model with other state-of-the-art approaches and evaluates the generated water spectra with the corresponding close-range RoX measurements. Finally, Subsection 3.4 validates the proposed model's capability to monitor the spatio-temporal variation of endmembers.

3.2 Quantitative Study and Analysis of Model Variants

	tDTMSUM	Baseline	Model1	Model2
Transformer	✓	✗	✓	✗
Auxiliary	✓	✗	✓	✓
Temporal	✓	✗	✓	✓
Spatial	✓	✗	✗	✗

Table 1. Overview of the models used in this study.

The performance of the proposed tDTMSUM is assessed against several architectural variants that were trained on the

same dataset as the proposed model. The description of these models is provided in Table 1, where a tick indicates that a specific part of the model is included in the architecture. All considered models were built with VAE and take the spectral data as the main input.

The RMSE and SAD evaluations across pixel samples from another synthetic dataset, separate from the one used for training, are presented in Table 2. The best results are highlighted in green cells. *rec* and *abd* refer to the reconstruction and abundance, while *vege*, *nveg*, *sedw*, and *clrw* denotes the endmember Vegetation, Non-Vegetation, Sedimented Water, and Clear Water, respectively. As shown in the table, the proposed model consistently achieves the best performance on the unseen data, particularly for the main objectives: abundance estimation (RMSE_{abd}) and water spectra generation (RMSE_{sedw}, SAD_{clrw}, RMSE_{sedw}, and SAD_{clrw}). Model1 ranks second, highlighting the benefit of incorporating a Transformer to capture the complex relationships between spectral and auxiliary information. In contrast, removing the Transformer, as in Model2, leads to a clear decline in performance, even worse than the baseline model that relies solely on spectral data.

	tDTMSUM	Baseline	Model1	Model2
RMSE_{rec}	0.0115	**0.0093**	0.0120	0.0140
SAD_{rec}	**0.0358**	0.0391	0.0379	0.0500
RMSE_{abd}	**0.0485**	0.0826	0.0508	0.0621
RMSE_{vege}	**0.0341**	0.0372	0.0356	0.0393
RMSE_{nveg}	**0.0474**	0.0488	0.0499	0.0511
RMSE_{sedw}	**0.0160**	0.0231	0.0167	0.0164
RMSE_{clrw}	**0.0160**	0.0194	0.0162	0.0162
SAD_{vege}	**0.0690**	0.0821	0.0761	0.0862
SAD_{nveg}	**0.1435**	0.1531	0.1482	0.1515
SAD_{sedw}	0.0903	0.1352	**0.0896**	0.0925
SAE_{clrw}	**0.1125**	0.1692	0.1242	0.1135

Table 2. Quantitative evaluation of RMSE and SAD metrics for each model on the synthetic dataset.

Day	tDTMSUM	Baseline	Model1	Model2
20210904	0.0077	**0.0066**	0.0075	0.0101
20220703	0.0116	**0.0091**	0.0122	0.0239
20230417	0.0090	**0.0075**	0.0111	0.0089
20230718	0.0083	**0.0092**	0.0096	0.0121
20240705	0.0084	**0.0061**	0.0094	0.0111

Table 3. The reconstruction RMSE on the evaluated days.

Day	tDTMSUM	Baseline	Model1	Model2
20210904	**0.0253**	0.0255	0.0270	0.0358
20220703	**0.0268**	0.0283	0.0341	0.0550
20230417	0.0292	0.0293	**0.0290**	0.0349
20230718	**0.0227**	0.0314	0.0276	0.0400
20240705	**0.0212**	0.0218	0.0251	0.0281

Table 4. The reconstruction SAD on the evaluated days.

To evaluate the models on real Sentinel-2 data, several scenes were randomly selected: 2021-09-04, 2022-07-03, 2023-04-17, 2023-07-18, and 2024-07-05. The first two scenes (2021-09-04 and 2022-07-03) are included in the training set, while the remaining three are entirely new and unseen. The ground truth in this case is unavailable because pixels in real satellite imagery are potentially mixed and the abundances of endmembers vary spatially and temporally (e.g., a pixel is mixed with water and bare soil on one day but with bare soil and trees on another day due to a plummet in the water level), making it arduous to acquire the ground truth for the abundances and endmember spectra. For this reason, evaluation focuses solely on reconstruction quality, measured using RMSE and SAD. The average RMSE_{rec} and SAD_{rec} across all valid pixels are reported in Table 3 and 4, respectively. The baseline achieves the lowest RMSE, while the proposed model and Model1 show comparable performance. However, for reconstruction SAD, which reflects spectral-shape similarity, the proposed tDTMSUM model performs best, whereas Model2 performs worst. This indicates that atmospheric, seasonal, and spatial variations in real Sentinel-2 data challenge models that do not incorporate the relationship among auxiliary variables, DOY, and spatial context. For applications prioritizing spectral fidelity, the proposed model is superior.

Fig. 5 shows the abundance maps of the four endmembers, estimated with the proposed model for 2021-09-01. Fig. 6 compares extracted water spectra from mixed pixels along the reservoir and the Enguri River with their nearest pure water pixels. The results indicate that the proposed model can reliably identify n pure water pixels even under thin clouds or cloud shadows, and can extract realistic water spectra that capture the transition from clear reservoir water in the south to sediment-laden river water towards the north.

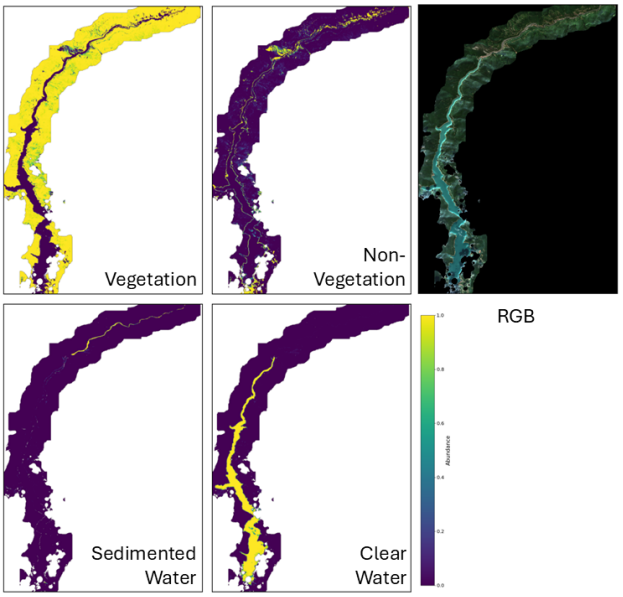

Figure 5. The abundance maps generated by the proposed model on 2021-09-01.

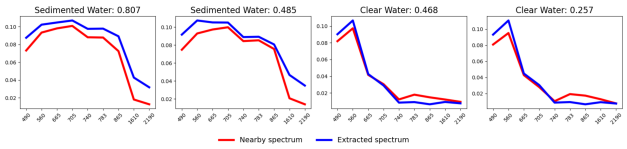

Figure 6. Comparison of reflectance water spectra obtained by unmixing with the proposed model and nearby pure water spectra on 2021-09-01. The title indicates the respective abundance in the mixed pixel.

3.3 Comparisons with State-of-the-art Approaches

The proposed model was compared against two unmixing approaches: Fully Constrained Least Squares Unmixing

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-573-2026 | © Author(s) 2026. CC BY 4.0 License.

(FCLSU) (Heinz et al., 2001) and Deep Transformer Network (DeepTrans) (Ghosh et al., 2022). Since these models are trained on a single scene at a time, the comparison was performed on one day, which was 2021-09-01, a day with moderate atmospheric residuals and dense cloud areas. The comparison metrics are the reconstruction RMSE and SAD, due to the lack of ground truth for abundances and endmember spectra, exhibited in Table 5 and supplemented with metrics on the models' complexity. The inference time was measured for processing the entire scene. The best and second-best error values are highlighted in green and blue, respectively. tDTMSUM achieves the lowest RMSE and SAD, outperforming DeepTrans by 63% (RMSE) and 66% (SAD). These results not only underscore the effectiveness of Transformer architectures in modeling spectral variability but also highlight the added value of incorporating auxiliary and temporal information, which further enhances the accuracy of spectral reconstruction. Although the proposed model is smaller than DeepTrans, its slower inference time results from using only dense layers instead of CNNs.

	tDTMSUM	FCLSU	DeepTrans
RMSE_{rec}	**0.0075**	0.0265	**0.0202**
SAD_{rec}	**0.0222**	0.0787	**0.0661**
MFLOPS	0.182	N/A	0.233
Train. params (M)	0.013	N/A	345.4
Model size (MB)	0.049	N/A	1317
Inference time (s)	0.443	251.83	0.04

Table 5. Reconstruction performance and model complexity across all evaluated models.

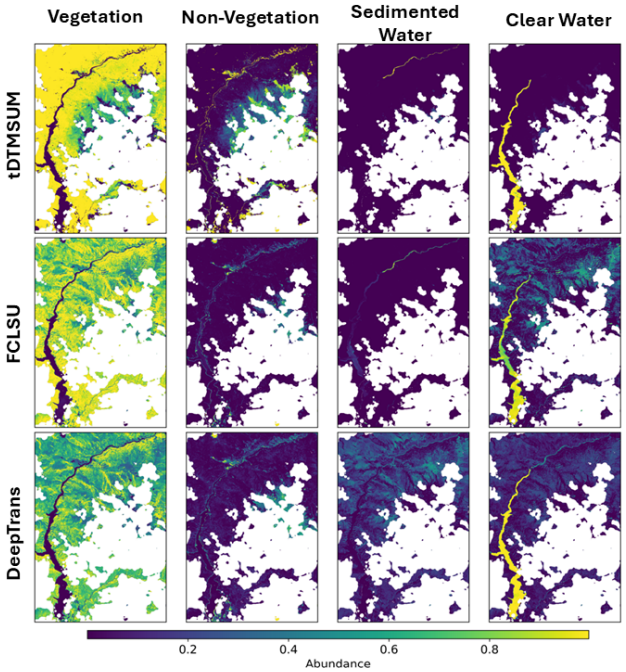

Figure 7. Visual comparison of the abundance maps obtained by different unmixing models.

The abundance maps of all four endmembers are shown in Fig. 7, with cloudy pixels masked as NaN. Among the methods, only tDTMSUM and FCLSU can clearly distinguish sedimented water from clear water. While DeepTrans provides a homogeneous water region, it misclassifies many vegetation pixels as sedimented water. Although FCLSU performs reasonably well, it is less accurate than tDTMSUM. Overall, tDTMSUM demonstrates the most reliable abundance estimation.

Unlike DeepTrans, which estimates shared endmember spectra, tDTMSUM generates pixel-specific endmembers. Fig. 8, following the same setup as in Fig. 6, compares the extracted endmember spectra from both models (blue lines) to manually selected pure spectra (orange lines). For tDTMSUM, mean spectra across all relevant pixels are shown. The proposed model outperforms DeepTrans, particularly for sedimented water and clear water, while performance on vegetation and non-vegetation is comparable between the two models.

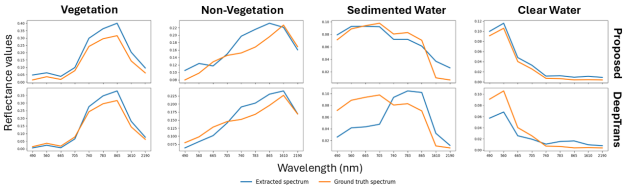

Figure 8. Visual comparison of the abundance maps obtained by different unmixing models.

Fig. 9 shows a comparison of four randomly selected acquisitions between unmixed water spectra from tDTMSUM and in-situ RoX measurements with the indication of the RoX device (red dot) and the nearby water pixel (blue dot). The results demonstrate that the extracted water spectra closely match the in-situ RoX measurements, highlighting the effectiveness of our approach. This agreement is particularly valuable for monitoring narrow rivers where mixed pixels dominate, enabling reliable water quality assessment without the need to install numerous monitoring stations or sensors.

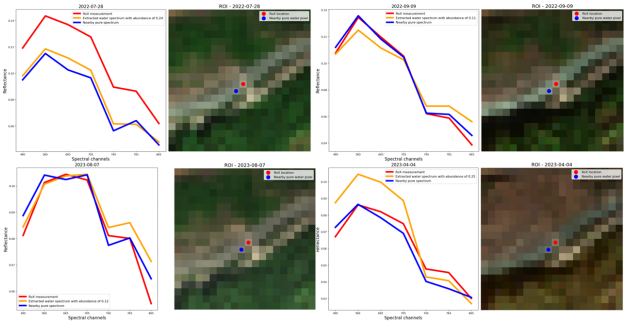

Figure 9. The comparison between the RoX measurement (red), the nearby water spectrum (blue), and the extracted spectrum (orange) from the mixed pixel at the RoX's location.

3.4 Capability to Monitor Tempo-Spatial variation of endmembers

The proposed spectral unmixing framework enables continuous monitoring of how endmembers evolve across both time and space, which is an essential capability in highly dynamic river–reservoir systems. By estimating endmember abundances (i.e., endmember coverage) for each acquisition and location, the model captures variations driven by changing hydrological conditions, fluctuating sediment loads, and temporal water discharge from the reservoir, which directly affects water levels and the distribution of mixed pixels along the riverbank line. Seasonal changes here refer not only to the annual cycle of water availability and flow regimes but also to broader seasonal hydrological dynamics that influence sediment transport, vegetation presence, and overall water composition. This integrated representation provides a detailed view of how endmembers

respond to both short-term fluctuations and longer-term seasonal patterns. As a result, the framework can detect gradual trends, such as progressive seasonal increases in water turbidity and sudden sediment inflows. Consequently, this tempo-spatial tracking enhances the robustness of spectral unmixing in highly variable environments and supports more accurate, long-term monitoring and interpretation of underlying environmental processes.

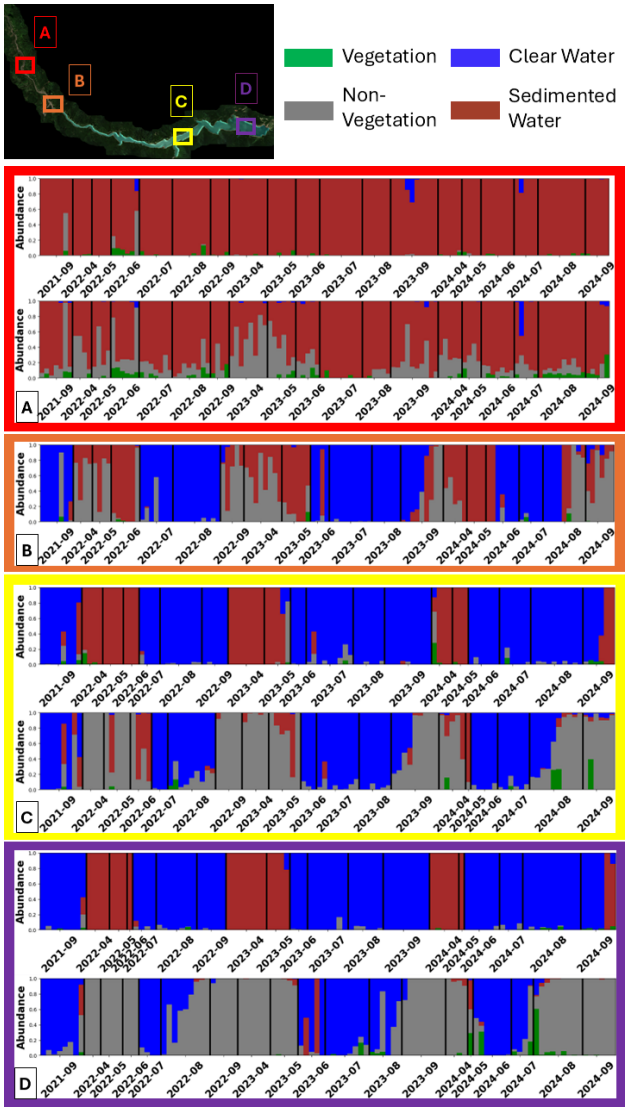

Figure 10. The spatiotemporal dynamic of selected pixels in the study area.

Fig. 10 illustrates the model’s capability to monitor tempo-spatial variations of water composition across the study region. Each plot shows the temporal evolution of endmember abundances for a selected pixel, covering the period from 2021 to 2024 on a monthly basis. The vertical axis represents abundance values, which sum to one, while the horizontal axis indicates the month of each acquisition. Abundance values for Vegetation, Non-Vegetation, Sedimented Water, and Clear Water are shown in green, gray, brown, and blue, respectively. Four key areas are examined:

- Area A in the northern part, characterized by flowing sediment-laden water;
- Area B, the river floor where the Enguri River enters the reservoir;
- Area C and Area D, representing the upper and lower sections of the reservoir.

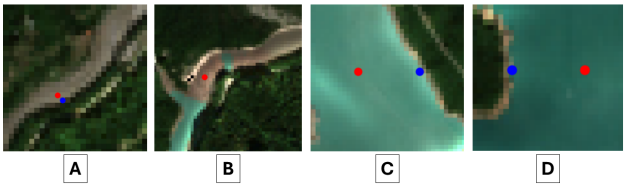

Figure 11. The location of examined pixels in four areas.

For each region, the central pure pixel, representative of that region, is selected. Additionally, in Areas A, C, and D, a nearby riverbank pixel is also included. For Areas A, C, and D in Fig. 10, the upper plot corresponds to the area’s pure pixel, and the lower plot corresponds to the selected riverbank pixel. Their locations are marked with the red (center pixel) and blue (riverbank pixel) dot in Fig. 11. The following observations can be made:

- **Area A:** In the upper plot, the pixel consistently exhibits high Sedimented Water abundance, reflecting the persistent presence of sediment-laden flow. Occasional increases in Clear Water abundance indicate lighter sediment conditions on certain days. Small contributions from Vegetation and Non-Vegetation arise mainly from cloud interference. In the lower plot, the temporal fluctuations between Sedimented Water and Non-Vegetation abundances reflect changes in water level at the riverbank. Vegetation signatures appear as expected due to partial land exposure along the riverbank.
- **Area B:** This region, where the Enguri River enters the reservoir, is highly dynamic. A clear seasonal pattern of reservoir operation is evident: from June to August, reservoir storage increases, raising the water level and resulting in dominant Clear Water abundance. From September to May, the reservoir discharges water, causing a sharp decline in Clear Water abundance. During this discharge phase, the riverbed becomes exposed on some days, while on others the inflowing Enguri River contributes sedimented water to the area.
- **Areas C and D:** In the upper plots, the pure water pixels show seasonal variations that align well with the reservoir discharge cycle observed in Area B. In the lower riverbank pixels, clear shifts between exposed soil (Non-Vegetation) during the discharge period and Clear Water during the storage period are evident, demonstrating how changing water levels affect shoreline composition.

Overall, Fig. 10 clearly demonstrates the seasonal hydrological dynamics of the Enguri Dam system. The observed patterns reflect both reservoir storage and discharge cycles, which drive fluctuations in water level and consequently alter pixel composition along the reservoir margins. Furthermore, the model successfully tracks transitions in water type: from the static, clear, sediment-accumulating waters in the southern reservoir to the highly dynamic, sediment-laden flows of the Enguri River in the north. This showcases the system’s capability to capture both spatial and temporal variations in water composition.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-573-2026

4. Conclusions

In this article, we introduced tDTMSUM, a novel multimodal deep generative model for supervised spectral unmixing. The model incorporates auxiliary factors contributing to spectral variability, such as residual atmospheric effects and spatio-temporal changes, and captures their interaction with spectral data through a channel-wise transformer architecture. tDTMSUM simultaneously estimates abundance maps and latent representations of endmembers in the encoder, effectively modeling their pixel-wise variability. The decoder reconstructs endmember spectra from the latent variables, which are linearly combined using the estimated abundances to approximate the observed spectra. By leveraging the nonlinear representational capacity of deep neural networks, the model can flexibly approximate complex and arbitrary endmember distributions with varying endmember spectra, enabling more accurate and robust unmixing. The model is trained on synthetic linear mixtures derived from endmember bundles extracted from selected days across different seasons, making it generalizable to any real Sentinel-2 scene of the region of interest within the study period without any further fine-tuning. Experimental results demonstrate that tDTMSUM consistently outperforms both classical and state-of-the-art LMM-based models that account for spectral variability, offering improved accuracy and computational efficiency without requiring model retraining or fine-tuning for each acquisition date. Future work will focus on generating more realistic synthetic mixtures, developing more expressive probabilistic generative models. Finally, consideration of nonlinear mixing effects may enhance this approach and its transferability to other complex aquatic and coastal environments.

5. Acknowledgments

This work was partially funded by the BMFTR (Federal Ministry of Research, Technology and Space) through the projects DAMAST and DAMAST-Transfer (Dams and induced Seismicity Technologies for Risk Reduction), within the framework of “CLIENT II - International Partnership for Sustainable Innovations”.

References

Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., Kembhavi, A., 2023. Satlaspretrain: A large-scale dataset for remote sensing image understanding. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16772–16782.

Bhakthan, S. M., Loganathan, A., 2024. A hyperspectral unmixing model using convolutional vision transformer. *Earth Science Informatics*, 17(3), 2255–2273.

Borsoi, R. A., Imbiriba, T., Bermudez, J. C. M., Richard, C., Chanussot, J., Drumetz, L., Tourneret, J.-Y., Zare, A., Jutten, C., 2021. Spectral variability in hyperspectral data unmixing: A comprehensive review. *IEEE geoscience and remote sensing magazine*, 9(4), 223–270.

Drumetz, L., Veganzones, M.-A., Henrot, S., Phlypo, R., Chanussot, J., Jutten, C., 2016. Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability. *IEEE Transactions on Image Processing*, 25(8), 3890–3905.

Ghosh, P., Roy, S. K., Koirala, B., Rasti, B., Scheunders, P., 2022. Hyperspectral unmixing using transformer network. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–16.

Heinz, D. C. et al., 2001. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE transactions on geoscience and remote sensing*, 39(3), 529–545.

Immitzer, M., Vuolo, F., Atzberger, C., 2016. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. *Remote sensing*, 8(3), 166.

Keshava, N., Mustard, J. F., 2002. Spectral unmixing. *IEEE signal processing magazine*, 19(1), 44–57.

Kingma, D. P., Welling, M. et al., 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4), 307–392.

Llodra-Llabres, J., Martinez-Lopez, J., Postma, T., Perez-Martinez, C., Alcaraz-Segura, D., 2023. Retrieving water chlorophyll-a concentration in inland waters from Sentinel-2 imagery: Review of operability, performance and ways forward. *International Journal of Applied Earth Observation and Geoinformation*, 125, 103605.

Martins, A., Astudillo, R., 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. *International conference on machine learning*, PMLR, 1614–1623.

Meng, F., Sun, H., Li, J., Xu, T., 2024. CTNet: an efficient coupled transformer network for robust hyperspectral unmixing. *International Journal of Remote Sensing*, 45(17), 5679–5712.

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V. R., Murayama, Y., Ranagalage, M., 2020. Sentinel-2 data for land cover/use mapping: A review. *Remote sensing*, 12(14), 2291.

Sent, G., Biguino, B., Favareto, L., Cruz, J., Sá, C., Dogliotti, A. I., Palma, C., Brotas, V., Brito, A. C., 2021. Deriving water quality parameters using sentinel-2 imagery: A case study in the Sado Estuary, Portugal. *Remote sensing*, 13(5), 1043.

Shi, S., Zhao, M., Zhang, L., Altmann, Y., Chen, J., 2021. Probabilistic generative model for hyperspectral unmixing accounting for endmember variability. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15.

Smith, L. N., Topin, N., 2019. Super-convergence: Very fast training of neural networks using large learning rates. *Artificial intelligence and machine learning for multi-domain operations applications*, 11006, SPIE, 369–386.

Somers, B., Zortea, M., Plaza, A., Asner, G. P., 2012. Automated extraction of image-based endmember bundles for improved spectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2), 396–408.

Tian, S., Guo, H., Xu, W., Zhu, X., Wang, B., Zeng, Q., Mai, Y., Huang, J. J., 2023. Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms. *Environmental Science and Pollution Research*, 30(7), 18617–18630.

Van Tricht, K., Gobin, A., Gilliams, S., Piccard, I., 2018. Synergistic use of radar Sentinel-1 and optical Sentinel-2 imagery for crop mapping: A case study for Belgium. *Remote Sensing*, 10(10), 1642.