

GIScience & Remote Sensing

ISSN: 1548-1603 (Print) 1943-7226 (Online) Journal homepage: www.tandfonline.com/journals/tgrs20

Multi-task learning for simultaneous surface water and burned area mapping from earth observation data

Aiym Orynbaikyzy , Marc Wieland, Moritz Rösch, Michael Nolde & Sandro Martinis

To cite this article: Aiym Orynbaikyzy , Marc Wieland, Moritz Rösch, Michael Nolde & Sandro Martinis (2026) Multi-task learning for simultaneous surface water and burned area mapping from earth observation data, GIScience & Remote Sensing, 63:1, 2737497, DOI: 10.1080/15481603.2026.2737497

To link to this article: https://doi.org/10.1080/15481603.2026.2737497

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 01 Oct 2026.

Submit your article to this journal

Article views: 198

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tgrs20

GISCIENCE & REMOTE SENSING
2026, VOL. 63, NO. 1, 2737497
https://doi.org/10.1080/15481603.2026.2737497

Taylor & Francis
by informa

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Multi-task learning for simultaneous surface water and burned area mapping from earth observation data

Aiym Orynbaikyzy, Marc Wieland, Moritz Rösch, Michael Nolde and Sandro Martinis

German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

Multi-task learning has emerged as a promising approach for combining related tasks into a unified framework with reduced computational costs through shared representations. However, its application to hazard mapping using Earth Observation data remains underexplored. In this study, we provide the first systematic evaluation of multi-task deep learning architectures for simultaneous detection of crisis information (surface water and burned area) using Sentinel-2 and Landsat-8 optical imagerytwo societally critical hazards with strong operational demand for rapid and accurate detection. We evaluate five semantic segmentation architectures: three U-Net variants with different encoders (MobileNetV2, EfficientNet-B4, ResNet50) and two transformer-based SegFormer models (SegFormer-B0, SegFormer-B4). Each architecture was trained in both single-task learning and multi-task learning modes. For reference data, we employed the benchmark S1S2-Water dataset for water body detection and constructed a novel burned area dataset by integrating diverse global datasets with already available in-house annotations. Both reference datasets cover heterogeneous geographic regions and temporal periods. The results show that multi-task learning effectiveness is highly architecture-dependent. Burned area detection achieved F1-scores ranging from 0.624 to 0.701, with three of five architectures showing improvements under multi-task learning. SegFormer-B4 reached the best performance (F1-score: 0.67) with gains of +8.24% over its single-task baseline, while EfficientNet-B4 achieved the highest absolute accuracy (F1-score: 0.701). Conversely, ResNet50 exhibited negative transfer with an F1-score decrease of 8.24%, demonstrating that not all architectures benefit from multi-task learning. Water detection performance remained consistently high across all models (F1-score > 0.97), with SegFormer-B4 achieving an F1-score of 0.98. Importantly, the multi-task framework reduces inference time by 89–91% and parameter count by 17–48% compared to maintaining two separate single-task models, demonstrating its practical value for operational emergency mapping services where rapid hazard assessment is critical.

ARTICLE HISTORY
Received 29 April 2026
Accepted 17 September 2026

KEYWORDS
Burned area mapping; surface water mapping; natural hazards; deep learning; image segmentation

1. Introduction

Climate change is increasing the frequency and intensity of natural hazards such as floods, landslides, and wildfires worldwide (Figure 1), making rapid hazard mapping critical for disaster response (IPCC 2023; Jain et al. 2020). In this context, Earth Observation (EO) data have become an invaluable tool for monitoring these hazards, providing wide-area coverage and near-real-time information over large areas (Mishra et al. 2026; Tralli et al. 2005). Among the various hazard monitoring applications, surface water and burned area detection represent two particularly important tasks with diverse operational purposes: surface water mapping is crucial for monitoring flood extent and water resource availability, while burned area detection is essential for assessing wildfire impacts and tracking burn progression. The accuracy and efficiency of both tasks have been substantially improved by recent advances in deep learning, particularly through the development of semantic segmentation architectures that enable automated generation of high-quality and detailed hazard maps from EO data (Akhyar et al. 2024; Yuan, Shi, and Gu 2021).

CONTACT Aiym Orynbaikyzy aiym.orynbaikyzy@dlr.de; aiym.orynbai@gmail.com; Marc Wieland marc.wieland@dlr.de

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

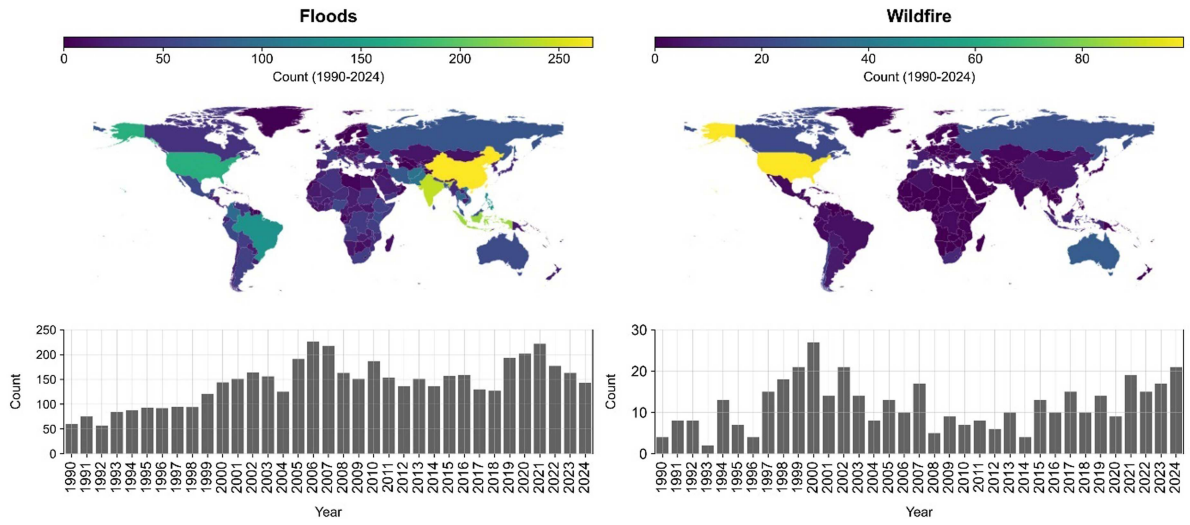


Figure 1. Global distribution (top) and temporal trends (bottom) of flood and wildfire disaster events (1990–2024). Data source: EM-DAT International Disaster Database (Delforge et al. 2025). Accessed: 27 January, 2026.

One of the recent developments in deep learning for EO applications is multi-task learning. Multi-task learning is a machine learning paradigm in which a single model is trained simultaneously on multiple related tasks using shared representations, enabling the model to learn common features across tasks while maintaining task-specific components (Crawshaw 2020; Zhang and Yang 2018). This approach offers several significant advantages beyond potential improvements in accuracy. First, multi-task learning provides substantial computational efficiency by reducing the number of model parameters and inference time, as shared feature representations need to be computed only once for multiple tasks rather than training separate models for each task (Vandenhende et al. 2020). Second, it enables more efficient resource utilisation by consolidating multiple models into a single unified architecture, thereby reducing memory requirements and deployment complexity. Additionally, the shared learning process can improve model generalisation through implicit regularisation, in which knowledge gained from one task helps prevent overfitting on related tasks (Caruana 1997). These benefits make multi-task learning particularly attractive for remote sensing applications, where tasks like water detection and burned area mapping exploit similar spectral and spatial patterns.

Multi-task learning has been increasingly applied to various remote sensing applications with varying degrees of success. Several studies have demonstrated the benefits of multi-task learning when tasks share complementary information. For instance, Rodriguez et al. (2025) showed that incorporating land cover segmentation as an auxiliary task improved burned area delineation from SPOT-6/7 imagery, as both tasks benefit from understanding vegetation boundaries and terrain characteristics. Similarly, Hong et al. (2023) successfully combined building extraction and change detection tasks using a Swin Transformer backbone, achieving an IoU improvement from 70% to 81%, demonstrating that shared feature extraction between spatially related tasks can enhance performance for both objectives. Feng et al. (2023) demonstrated that joint semantic segmentation and height estimation from remote sensing imagery can be improved by leveraging cross-task interactions, thereby improving performance on both tasks. However, multi-task learning effectiveness is highly task-dependent and can fail when tasks conflict or require fundamentally different feature representations. Xu et al. (2025) demonstrated that semantic change detection in remote sensing suffers from negative transfer when tasks have different output dimensions and loss functions, leading to shared parameters becoming biased toward the dominant task. Similarly, Wan et al. (2025) found that tasks with different optimisation priorities—such as detection emphasising object contours, segmentation focusing on boundary fineness, and depth estimation relying on geometric priors—exhibit conflicting optimisation directions that lead to negative transfer in traffic scene understanding. The key insight from the mentioned studies is that multi-task learning succeeds when tasks complement each other—such as edge detection supporting segmentation (Volpi and Tuia 2018) or land cover classification guiding

burned area delineation (Rodriguez et al. 2025)—but fails when tasks require incompatible feature representations. These findings highlight the importance of task compatibility in multi-task learning.

Multi-hazard monitoring is increasingly critical as natural disasters create cascading risks that threaten communities and infrastructure. Wildfires fundamentally affect watershed hydrology by altering vegetation and soil properties, making burned landscapes highly susceptible to flooding during subsequent rainfall events (Shakesby and Doerr 2006). Additionally, the removal of protective vegetation exposes bare soil and destabilises hillslopes, making them prone to landslides and debris flows when triggered by post-fire flooding and intense rainfall (Cannon et al. 2008). Given these interconnections, international emergency management services—including Copernicus Emergency Management Service (EMS) (European Commission and J 2026) and the International Charter "Space and Major Disasters" (International Charter "Space and Major Disasters", 2026)—routinely provide rapid mapping for multiple hazard types within integrated operational frameworks. Multi-task learning frameworks offer valuable advantages for such systems by consolidating multiple detection capabilities into unified models, enabling shared feature representations while reducing deployment complexity compared to maintaining separate models for each hazard. While multi-task learning has been successfully applied to various remote sensing tasks—including building extraction combined with change detection (Hong et al. 2023) and burned area mapping with land cover classification (Rodriguez et al. 2025)—no existing work has addressed simultaneous detection of multiple hazard types using multi-task learning. We fill this gap by providing the first systematic evaluation of a multi-task framework for simultaneous surface water detection and burned area mapping from EO data. While surface water segmentation does not constitute flood mapping in the strict hydrological sense, it is a critical operational component of flood rapid-mapping workflows routinely employed by services such as Copernicus EMS (Ajmar et al. 2017). The selection of surface water and burned area as the two target tasks is motivated by both operational and machine learning considerations: both are routinely mapped within integrated rapid mapping frameworks, and the abundance of high-quality surface water training data can provide regularisation benefits for the more data-scarce burned area task through shared feature representations.

The primary objective of this work is to develop and evaluate a multi-task deep learning framework that enables efficient, simultaneous large-scale monitoring of multiple hazard types (floods and wildfire) from freely available EO data. We address two research questions: (RQ1) whether multi-task learning improves detection accuracy compared to single-task models, and (RQ2) which architectures are most effective for multi-task hazard detection. We implement and compare five semantic segmentation architectures: three U-Net variants with different encoder backbones (MobileNetV2, EfficientNet-B4, ResNet50) and two transformer-based SegFormer models (SegFormer-B0 and SegFormer-B4). Each architecture is trained in both single-task learning (separate models per hazard) and multi-task learning (shared encoder with task-specific decoders) configurations. Input features comprise six Sentinel-2 and Landsat-8 multispectral bands supplemented with terrain slope information. Our multi-source validation strategy incorporates emergency mapping products from Copernicus EMS, global burned area datasets from C3S BARD and FireCCI, HLS USA and DLR in-house reference data, enabling the assessment of model generalisation across diverse geographic contexts and temporal scales.

2. Reference datasets

Two reference datasets were used in this study: a water body dataset and a burned area dataset (Figure 2). Both datasets were constructed from Sentinel-2 Level-2A (bottom-of-atmosphere surface reflectance) and Landsat-8 Collection 2 Level-2 (atmospherically corrected surface reflectance) optical imagery. Both processing levels provide analysis-ready, atmospherically corrected surface reflectance data, ensuring radiometric consistency across sensors without needing additional harmonisation. Training, validation, and test samples were distributed globally to ensure representation of diverse geographic and environmental contexts.

2.1. Water dataset

The water body reference dataset (hereafter "water dataset") is a globally distributed, pixel-wise-annotated benchmark (Figure 3) for deep learning-based segmentation of surface water (Wieland et al. 2024).

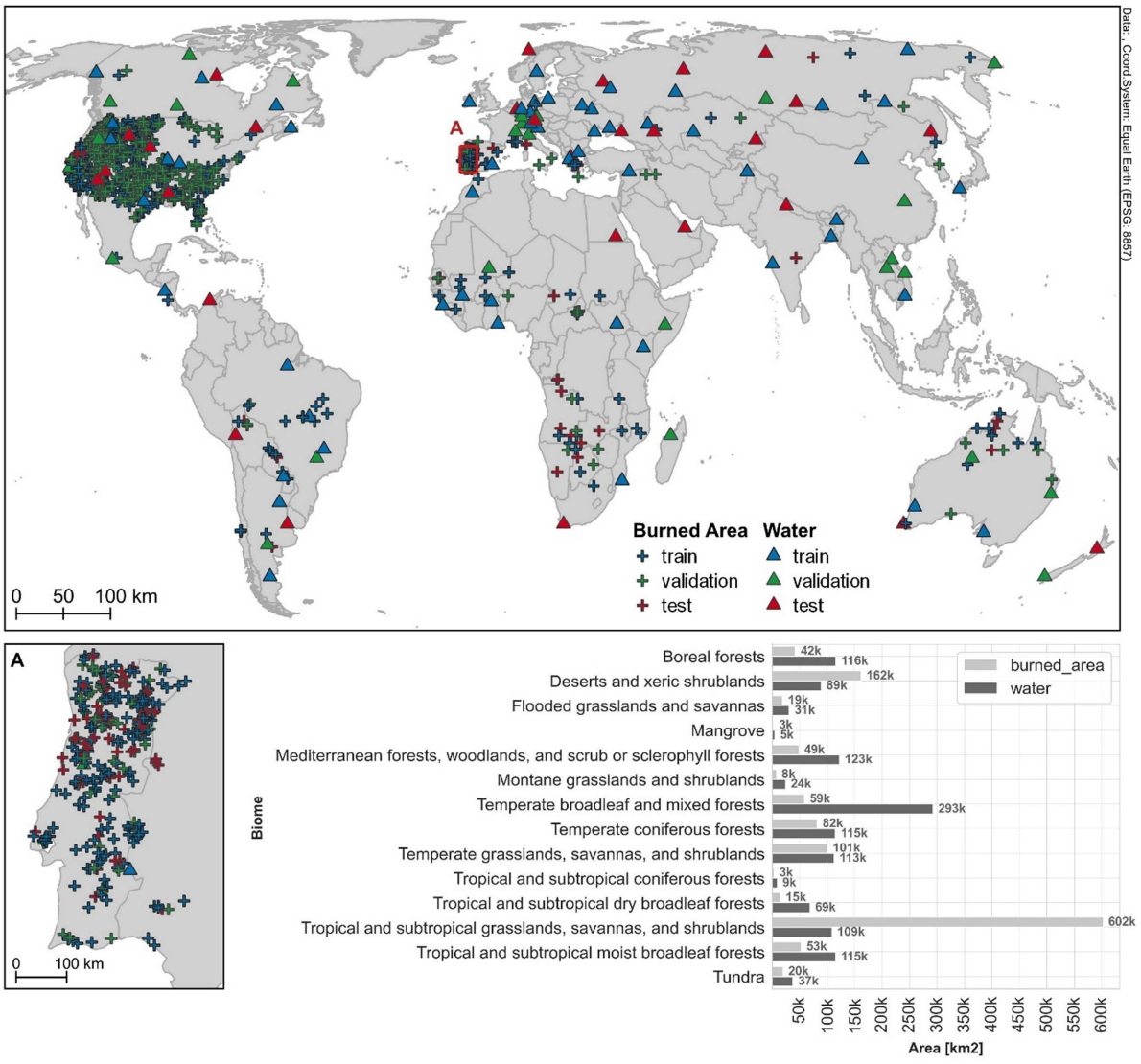


Figure 2. Geographic distribution of surface water (triangles) and burned area (crosses) reference samples. Each water body sample represents a full Sentinel-2 scene. Burned area samples are derived from five heterogeneous sources with varying spatial extents, ranging from full Sentinel-2 and Landsat-8 scenes to 512 × 512 pixel patches (HLS dataset).

The dataset consists of Sentinel-2 A/B optical scenes with 10 980m × 10 980m tiles at 10m spatial resolution. Scenes were selected across multiple continents, land cover types, and hydrological regimes to ensure geographic, environmental, and seasonal diversity. Each scene includes a binary water mask (water vs non-water) and a slope layer, both calculated from Copernicus Digital Elevation Model (DEM) at 30 m resolution (European Space Agency and Airbus 2022). For detailed information on dataset construction and validation, refer to Wieland et al. (2024).

2.2. Burned area dataset

A key challenge in deep learning-based burned area mapping is the scarcity of large, diverse reference datasets. To address this limitation, we build a comprehensive reference dataset (hereafter "burned area dataset") by integrating five complementary data sources spanning multiple continents and biomes (Table 1). This multi-source approach enhances both the geographic representativeness and environmental diversity necessary for robust model training and generalisation (Figure 4).

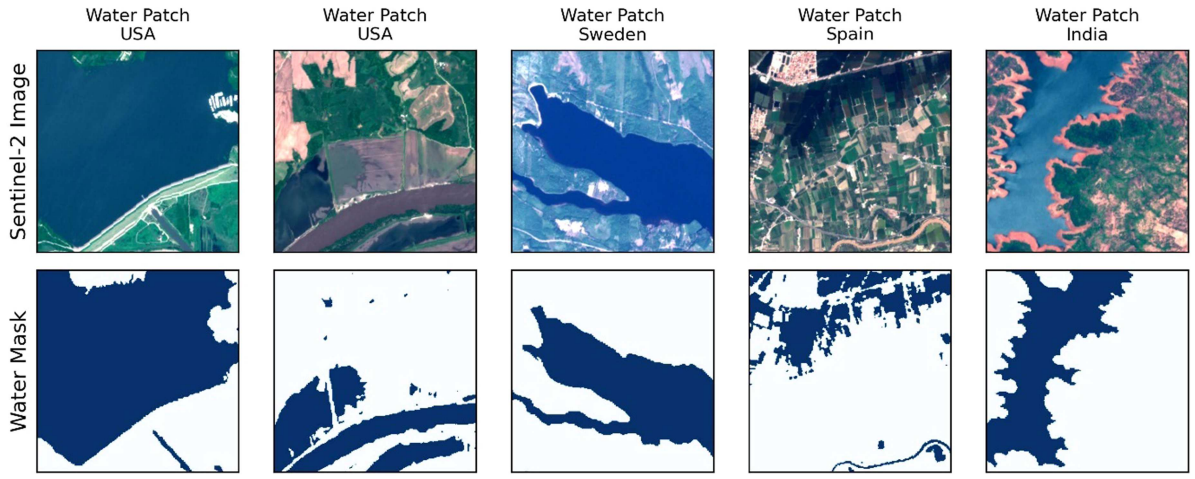


Figure 3. Water reference samples. Each sample (column) shows a Sentinel-2 image (top) and the corresponding water mask (bottom), where dark blue indicates water and white indicates non-water areas.

Table 1. Burned area dataset composition.

Dataset	*Sensor*	*Train Patches**	*Validation Patches**	*Test Images*	Reference
Burned area segmentation data	Sentinel-2	1 463	470	72	Knopp et al. (2020)
C3S dataset	Landsat-8	8 183	2 581	17	Franquesa et al. (2020)
FireCCI	Sentinel-2	12 915	4 821	3	Franquesa et al. (2020)
EMS	Sentinel-2	447	147	5	Luca et al. (2022)
HLS USA	Sentinel-2	2 160	1 056	–	Phillips et al. (2023)
Sum		*25 168*	*9 075*	*97*	

* *patches of size 256 × 256 pixels.*

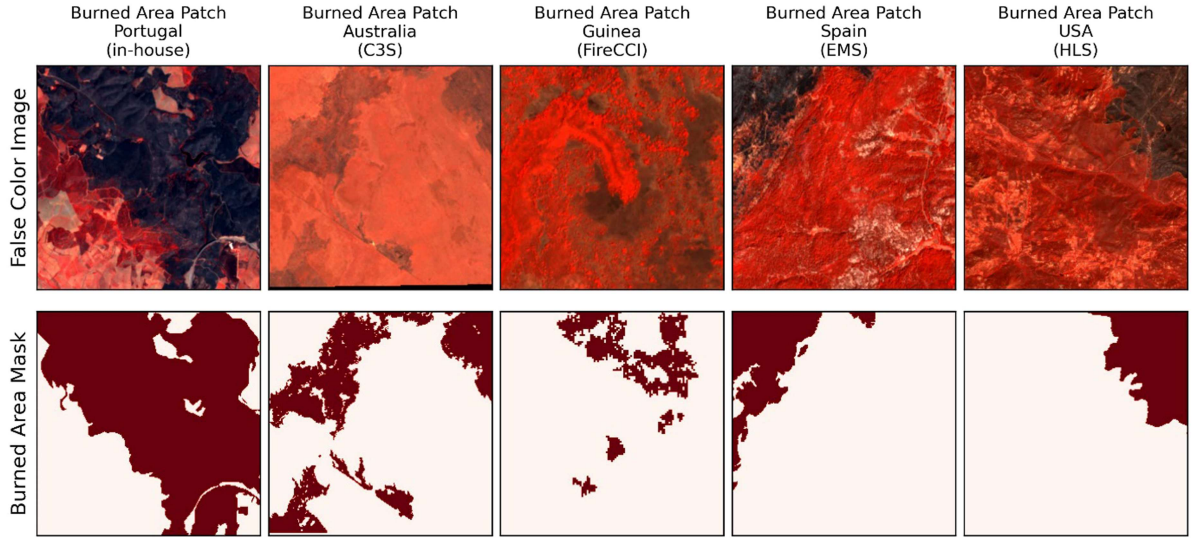


Figure 4. Burned area reference samples showing satellite imagery (top) from Sentinel-2 or Landsat-8 (top) and corresponding burned area masks (bottom).

The first component is an in-house burned area segmentation dataset developed by Knopp et al. (2020) using Sentinel-2 imagery at 10 m resolution. This dataset combines burned area masks from three sources: the California Department of Forestry & Fire Protection (CAL FIRE), the Portuguese Institute for Nature Conservation and Forests (ICNF), and fire scars processed with a DLR Sentinel-2 burned area processing chain (Rättich and Martinis, 2017). The California and Portuguese regional databases together comprise 83% of the first dataset, predominantly covering Mediterranean and temperate ecosystems from 2017–2018.

All spectral bands were resampled to 10 m resolution, and label masks were spatially aligned and visually verified to ensure consistency across sources.

The second and third components were derived from the Burned Area Reference Database (BARD), a harmonised collection of regional and global burned area reference datasets designed to validate satellite-derived fire products (Franquesa et al. 2020). BARD compiles reference data produced by a wide range of international research initiatives. From BARD, we selected two reference datasets: the C3S (Copernicus Climate Change Service) reference dataset and the FireCCI Sentinel-2 Africa dataset. The C3S dataset, based on Landsat-8 imagery, provides reference polygons generated through visual interpretation of high-resolution imagery, ensuring accurate delineation of burned and unburned surfaces across diverse biomes. The FireCCI Sentinel-2 Africa dataset was based on Sentinel-2 imagery using semi-automatic classification methods with manual refinement. For our study, we visually assessed the prominence of burn scars on the corresponding Sentinel-2 and Landsat-8 imagery. We selected only those scars that were clearly visible and no older than two months from the image acquisition date, ensuring optimal spectral contrast between burned and unburned areas.

The fourth component consists of rapid mapping products from the Copernicus EMS, which provide near-real-time burned area delineations in response to major wildfire events in Europe (Luca et al. 2022). These operational products offer high-quality reference data for recent fires across Europe and other regions where EMS activations occurred.

The fifth component is the Harmonised Landsat Sentinel-2 (HLS) burn scar dataset (Phillips et al. 2023), which provide 804 manually verified burned area scenes (512×512 pixels) from 2018–2021 across the contiguous United States. Burn scar delineations were sourced from the Monitoring Trends in Burn Severity (MTBS) programme, and all scenes were filtered for cloud cover and data quality. Original 512×512 pixel scenes were subdivided into non-overlapping 256×256 pixel patches for consistency with other datasets. The dataset provides only training and validation partitions; therefore, we used it only for model development.

To ensure spatial and thematic consistency across these heterogeneous sources, all datasets underwent standardised preprocessing. We resampled all imagery to 10 m resolution where necessary, and harmonised class definitions to a binary burned/non-burned schema. This standardised preprocessing pipeline was applied uniformly across all five datasets, lling protocol of each source. For the BARD dataset, the polygon boundaries were visually verified against Sentinel-2 composites, overlapping features were merged, and small artefacts were manually corrected.

Despite these harmonisation efforts, combining annotations derived from different sensors and labelling protocols—ranging from manual visual interpretation (EMS) to semi-automatic classification (FireCCI, C3S) and processing chain outputs (DLR in-house)—introduces inherent label uncertainty that cannot be fully eliminated. Differences in burn severity thresholds, mapping scales, and temporal proximity to fire events across sources may result in annotation inconsistencies, particularly at burn scar boundaries. We acknowledge this as a limitation of multi-source dataset construction and note that its impact on model training and evaluation was not systematically quantified. Further details on individual dataset and validation protocols can be found in the respective source publications (Table 1).

3. Methods

3.1. Input reference data

Mono-temporal post-event imagery from Sentinel-2 and Landsat-8, processed to Level-2A surface reflectance, was used as the primary input source in this study. From both sensors, six spectral bands were selected: blue (B2: ~490 nm), green (B3: ~560 nm), red (B4: ~665 nm), near-infra-red (NIR, B8: ~842 nm), shortwave infra-red 1 (SWIR1, B11: 1610 nm), and shortwave infra-red 2 (SWIR2, B12: ~2190 nm). These bands were chosen based on their established sensitivity to water absorption features and vegetation moisture content, which are relevant for both surface water mapping and burned area detection. Additionally, a slope layer derived from the Copernicus DEM at 30 m spatial resolution was included as a seventh input channel, as it has been shown to improve performance in both water body and burned area mapping (Zhou et al. 2026).

Data splitting was performed independently for each dataset at the scene level to prevent data leakage between the training, validation, and test sets. Each dataset, including the sub-datasets such as C3S, was divided into a 60/20/20 for training, validation, and testing, respectively. To ensure geographic separation between partitions, scene-level splitting was stratified by biome class, guaranteeing that no geographic region or biome type was represented exclusively in a single partition. This stratification prevents the model from being evaluated on ecosystems that were entirely absent from training, ensuring a more realistic assessment of its generalisation capability. Temporal leakage was prevented by applying the split independently for each sub-dataset so that no acquisition date or fire event contributed scenes to more than one partition. The test set includes real operational wildfire events mapped during active Copernicus EMS emergency activations, as well as diverse fire events across multiple continents and biome types, providing an operationally representative evaluation of model generalisation to unseen real-world scenarios. Training and validation scenes were subsequently tiled into non-overlapping 256×256 pixel patches, while test scenes were retained in their original spatial extent to enable scene-level evaluation. For the water body dataset, the split yielded 57 scenes for training, 28 for validation, and 28 for testing, resulting in 80 833 and 39 647 patches for training and validation, respectively. For the burned area dataset, 794 images of varying spatial extent—including HLS tiles of 512×512 pixels and full Sentinel-2 scenes from the FireCCI dataset—were assigned to training, 360 to validation, and 97 scenes to testing. Tiling of the burned area training and validation images yielded 25 168 and 9 075 patches, respectively.

A pronounced imbalance existed between the task-specific datasets, with the surface water dataset contributing substantially more patches than the burned area dataset (80 833 versus 25 168 training patches, respectively). To prevent the multi-task models from being disproportionately optimised toward the surface water mapping task, the training set was balanced at the patch level: all available burned area patches were retained, and the equal number of water body patches was drawn via simple random sampling without replacement, using fixed random seed to ensure reproducibility. After balancing both tasks were represented by 25 168 training patches, yielding a combined training set of 50 336 patches. This balancing ensures that the observed performance differences between single-task and multi-task configurations reflect the effect of the learning paradigm itself rather than differences in data volume.

Class imbalance within the burned area dataset was pronounced, with non-burned pixels substantially outnumbering burned pixels across all sources. To address this, a weighted loss function was employed during training, with class weights automatically calculated based on the inverse frequency of each class in the training set, ensuring that the model does not become biased toward the dominant background class.

3.2. Semantic segmentation models

To evaluate the potential benefits of multi-task learning for natural hazard monitoring, two distinct training configurations were implemented and compared (Figure 5). In the single-task learning configuration, separate models were trained independently for surface water and burned area segmentation, each optimised exclusively on its respective task-specific dataset. In the multi-task learning configuration, a single model with a shared encoder was trained simultaneously on both tasks, using task-specific decoder heads to produce predictions for surface water and burned areas. The shared encoder in the multi-task learning setup enables feature representations learned from one task to potentially benefit the other. At the same time, the task-specific decoders enable specialised processing of the shared features for each segmentation objective. This comparison allows for assessment of whether joint training improves performance relative to independently trained baselines.

3.2.1. Model architectures

Four semantic segmentation architectures were evaluated across both training configurations: three convolutional U-Net variants with different encoder backbones and one transformer-based architecture. These architectures were selected for their demonstrated effectiveness in surface water mapping and burned area detection tasks (Seydi, Hasanlou, and Chanussot 2022; Wieland et al. 2023).

The U-Net architecture follows the canonical encoder-decoder structure with skip connections that concatenate feature maps from the encoding layers to the corresponding decoding layers, thereby preserving spatial information that would otherwise be lost during downsampling (Ronneberger and Fischer, 2015). The skip

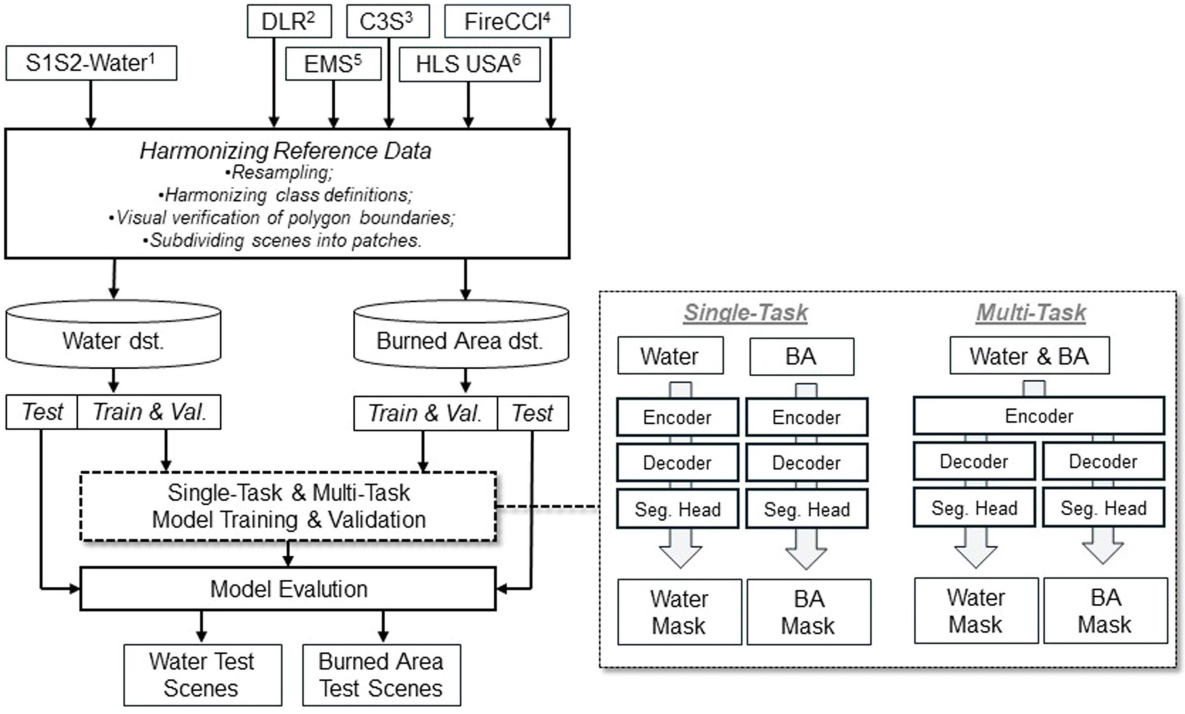


Figure 5. Overview of the study workflow (left) and single-task and multi-task architecture comparison (right).

connection mechanism makes the U-Net particularly suitable for EO applications, as it enables precise localisation of spatial features while maintaining semantic understanding from deeper layers (Hoeser and Kuenzer 2020). We selected three widely used encoder variants. (1) *MobileNet-V2* (Sandler et al. 2018) is a computationally efficient architecture designed for rapid inference with reduced memory requirements. This encoder has been successfully applied in remote sensing image classification tasks and has demonstrated effectiveness in satellite image segmentation under resource-constrained conditions (Wang et al. 2021). (2) *EfficientNet-B4* (Tan and Le 2019) is a balanced architecture that scales model complexity through coordinated adjustments to network depth, width, and input resolution. This encoder has demonstrated strong performance in EO tasks, with its efficient multi-level feature extraction proving well-suited for large-scale satellite imagery analysis and land use classification (Wieland et al. 2023). (3) *ResNet50* (He et al. 2015) is a deep convolutional neural network (CNN) architecture with 50 layers that enables effective training through residual connections. This architecture has been widely adopted in remote sensing applications for its ability to learn complex spatial patterns without performance degradation, making it suitable for natural hazard detection tasks (Fan et al. 2025).

In addition to the convolutional U-Net variants, SegFormer-B0 and SegFormer-B4 were employed as transformer-based architectures (Dosovitskiy et al. 2020; Xie et al. 2021). Unlike U-Net, which relies on a convolutional encoder-decoder structure, SegFormer replaces the traditional convolutional encoder with a hierarchical vision transformer backbone (Xie et al. 2021). This architecture processes the input image through hierarchical stages that progressively extract features at multiple scales, capturing both fine-grained details and broader contextual information. The decoder then combines these multi-scale features to generate the final segmentation map. The B0 and B4 variants differ in model depth and width, with B0 optimised for computational efficiency and B4 providing enhanced capacity for complex scene understanding.

3.2.2. Single-task models

In the single-task learning setup, each of the five architectures was trained independently for water body segmentation and burned area segmentation, resulting in ten separate models. Each model was optimised exclusively on its respective task-specific dataset using the corresponding loss function. The encoder, decoder, and segmentation head were trained end-to-end without any parameter sharing across tasks, establishing a baseline for comparison with multi-task approaches.

3.2.3. Multi-task models

Multi-task models were designed to perform simultaneous segmentation of surface water and burned area through shared feature extraction and task-specific prediction heads. The multi-task architecture consists of three components: (1) a shared encoder that processes input imagery and learns common representations across both tasks, (2) two task-specific decoders that process the shared features independently, and (3) two segmentation heads that generate the final binary predictions for water and burned area classes.

In U-Net-based multi-task models, the encoder is shared across both tasks to extract common features, while separate decoders process these features independently for water and burned area segmentation. Each decoder progressively reconstructs the spatial resolution and generates the final segmentation output for its respective task. This design enables the model to learn general features applicable to both tasks in the encoder, while allowing task-specific processing in the decoder stages.

For the SegFormer-B0 and B4 models, the transformer encoder is shared across tasks, with two independent decoder modules for water and burned area segmentation. Each decoder combines features from different encoder stages to produce task-specific segmentation maps. SegFormer-B4 decoders are initialised with ImageNet-1K pre-trained weights (nvidia/mit-b4), while SegFormer-B0 decoders use weights fine-tuned on ADE20K semantic segmentation dataset (nvidia/segformer-b0-finetuned-ade-512-512). These weights are subsequently adapted during training to specialise for their respective tasks while benefiting from the shared feature extraction.

During training, mini-batches alternate between the two tasks, with each forward pass computing predictions for one task using its corresponding decoder head. The shared encoder is updated using gradients from both tasks, enabling knowledge transfer through the common feature representation. This training strategy ensures balanced learning across both tasks while maintaining computational efficiency compared to training two completely separate models.

3.3. Training configuration

All models were trained for a fixed 50 epochs using the Adam optimiser with an initial learning rate of 0.001 and weight decay of 0.001. The number of training epochs was set to 50, balancing computational feasibility with sufficient exposure to the training data, and is consistent with settings reported in comparable semantic segmentation studies (Rangel et al. 2025; Sui et al. 2024). The Adam optimiser (Kingma and Ba 2014) was selected for its well-established effectiveness and computational efficiency in semantic segmentation tasks (Wieland et al. 2023). The initial learning rate and weight decay were set based on exploratory experiments and values widely adopted in semantic segmentation studies using similar architectures. The selected values were subsequently applied uniformly across all architectures to ensure comparability of results.

We employed the ReduceLROnPlateau scheduler to dynamically adjust the learning rate based on validation loss, reducing the learning rate by a factor of 0.5 after 5 epochs without improvement. A batch size of 16 was used across all experiments, with input patches of 256×256 pixels containing six spectral channels (blue, green, red, NIR, SWIR1, SWIR2) and slope derived from DEM.

All models were trained using a balanced dataset described in sub-chapter 3.1., where the training data consisted of 25 168 patches per task and the validation data contained 9 075 patches per task. To address the inherent class imbalance within each dataset, particularly pronounced in the burned area segmentation task, a weighted loss function was employed with class weights automatically calculated based on the inverse class frequencies in the training set. The combined Cross-Entropy and Lovász loss function (Berman, Triki, and Blaschko 2018) was used for all experiments, providing both pixel-wise classification accuracy via Cross-Entropy and smooth IoU optimisation via the Lovász-hinge loss. For multi-task learning configurations, the total loss was computed as the unweighted sum of individual task losses: $L = L_0 + L_1$, where L_0 and L_1 represent the losses for water body and burned area segmentation, respectively. Model checkpoints were saved whenever validation loss improved by a minimum delta of 0.0001, and the best-performing model was selected for final evaluation based on the lowest validation loss.

All experiments were conducted in deterministic mode with a fixed random seed to ensure reproducibility. Training was performed on a single NVIDIA GPU using the PyTorch Lightning framework with 12 parallel data loading workers. Data augmentation was not applied during training.

3.4. Model evaluation

To evaluate the segmentation quality, several commonly used metrics in the field of semantic segmentation were selected, namely Intersection over Union (IoU), F1-score, precision and recall. The raw prediction outputs were converted to binary segmentation masks using a probability threshold of 0.5: pixel with a probability equal to or greater than this value were labelled as the target class, either surface water or burned area, and all remaining pixels were assigned to the background class.

In the single-task configuration, each model was assessed on the test scenes of its corresponding task. For the multi-task configuration, the joint model was evaluated on the task-specific test scenes independently, enabling a direct comparison of the segmentation results between the single-task and multi-task setups for both water body and burned area mapping. The performance metrics were computed by accumulating confusion matrix elements (true positive - TP, true negative - TN, false positive - FP, false negative - FN) across all test scenes before calculating precision, recall, F1-score and IoU. This accumulation-based approach prevents individual scenes with sparse positive samples from disproportionately affecting the overall performance assessment.

4. Results

4.1. Single-task vs. multi-task model performance

The overall results showed that multi-task learning delivered meaningful improvements for burned area mapping task while maintaining excellent surface water detection performance across all configurations. (Table 2). However, the impact of multi-task learning varied substantially between the two segmentation tasks. For surface water segmentation, multi-task models showed minimal performance changes across most architectures, with an average F1-score change of –0.67%. The most notable decline occurred with SegFormer-B0, where F1-score decreased by 4.08%. On the contrary, U-Net with MobileNetV2 and SegFormer-B4 exhibited small gains (+0.20% and +0.93% F1-score respectively).

Burned area segmentation showed more favourable responses to multi-task learning, with an average F1-score improvement of +0.55% (Table 2). Three of five architectures demonstrated performance gains under multi-task learning, with SegFormer-B4 achieving the most substantial improvement: F1-score increased from 0.619 to 0.670 (+8.24%). Similarly, SegFormer-B0, despite its poor water segmentation performance in multi-task mode, showed improved burned area segmentation with F1-score gains of +4.12%. EfficientNet-B4 demonstrated the most balanced performance across both tasks, achieving consistently high accuracies for surface water detection and burned area detection, ranking among the top performers for both objectives (Figure 6).

Despite requiring longer training times compared to individual single-task models (Appendix B)—a well-documented characteristic of joint training frameworks—the multi-task approach offers substantial efficiency gains at inference time. Despite the longer training times (Appendix B), The multi-task framework reduces the total number of trainable parameters by 17–48% across all evaluated architectures compared to maintaining two separate single-task models, with transformer-based architectures showing the largest gains (Appendix B). Multi-task learning reduces average inference time per Sentinel-2 scene by approximately 89–91%, from 8–19 minutes for two sequential single-task model predictions to under 2 minutes for a single multi-task model (Appendix C).

Table 2. Accuracy metrics of the single-task and multi-task models for water and burned area segmentation. The precision and recall values are provided in appendix A.

		Water				Burned Area			
		STM		MTM		STM		MTM	
Decoder	Encoder	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Transformer	SegFormer-B0	0.981	0.963	0.941	0.888	0.607	0.436	0.632	0.462
	SegFormer-B4	0.971	0.943	0.980	0.960	0.619	0.448	0.670	0.503
U-Net	EfficientNet-B4	0.987	0.973	0.986	0.973	0.694	0.531	0.701	0.539
	MobileNet-v2	0.986	0.973	0.988	0.977	0.717	0.558	0.700	0.538
	ResNet50	0.989	0.977	0.986	0.972	0.680	0.516	0.624	0.453

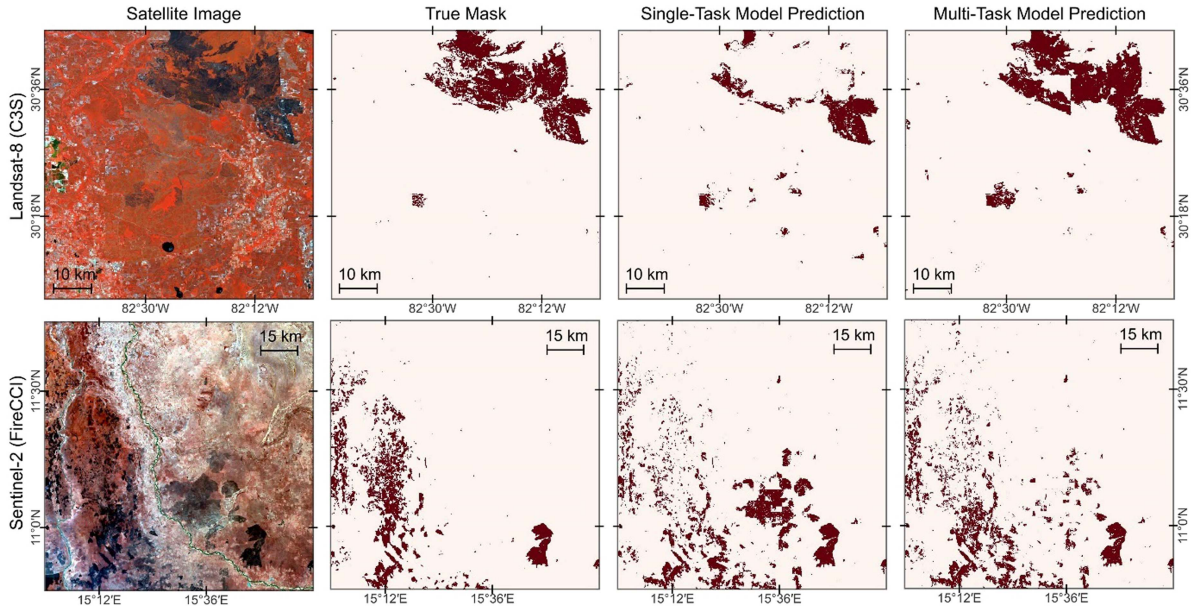


Figure 6. Burned area segmentation results from two test scenes (rows) based on the U-Net with EfficientNet-B4 encoder.

4.2. Water segmentation

The water segmentation achieved consistently excellent performance across all models in both multi-task and single-task configurations. As shown in Table 1, IoU scores ranged from 0.888 to 0.977 across all models.

The U-Net architectures showed minimal performance differences between multi-task and single-task approaches, with changes typically remaining under 1%. U-Net with a MobileNetV2 encoder even showed slight improvements in multi-task mode, achieving +0.41% higher IoU and +0.61% higher precision compared to the corresponding single-task model (Figure 7).

In contrast, the transformer-based SegFormer-B0 exhibited substantial performance degradation when trained in a multi-task setting. The IoU dropped by 7.79 percent (from 0.963 to 0.888), and recall decreased by 6.40 percent (see Appendix A). This significant drop suggests that smaller transformer architecture may lack sufficient capacity to learn both tasks simultaneously.

Among all tested architectures, U-Net with ResNet50 showed a moderate performance decline in multi-task mode, with IoU decreasing by 0.51% and recall by 1.31%. However, precision improved slightly by 0.81%, indicating that while the model became more conservative in predictions, it maintained high confidence in detected water pixels.

4.3. Burned area segmentation

Burned area detection proved to be a more challenging task than water segmentation. As shown in Table 2, F1-scores for single-task models ranged from 0.607 to 0.717, which are significantly lower than the water segmentation performance. Unlike water segmentation, the performance differences between multi-task and single-task approaches showed more variability across architectures.

Architectural responses to multi-task learning varied substantially for burned area segmentation. U-Net with ResNet50 exhibited severe negative transfer, with the F1-score declining by 8.24%, suggesting that encoder features optimised for water detection became detrimental to burned area segmentation. Whereas, U-Net with EfficientNet-B4 showed modest improvements (IoU: +1.51%, F1-score: +1.01%).

Among transformer architectures, SegFormer-B4 achieved substantial improvements in burned area (IoU: +12.28%, F1-score: +8.24%) and is the only architecture to improve both tasks simultaneously (Figure 7). In contrast, SegFormer-B0 showed inverse patterns across tasks: while water detection degraded

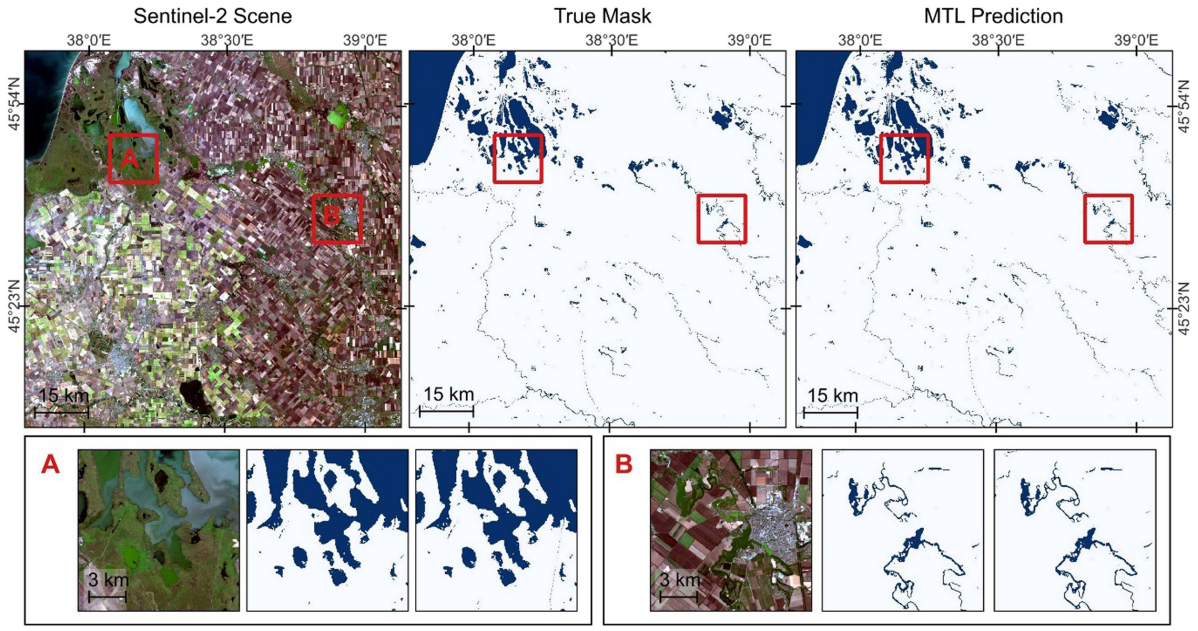


Figure 7. Surface water segmentation results of a Sentinel-2 test scene in northern Colombia using multi-task U-Net with MobileNet-v2.

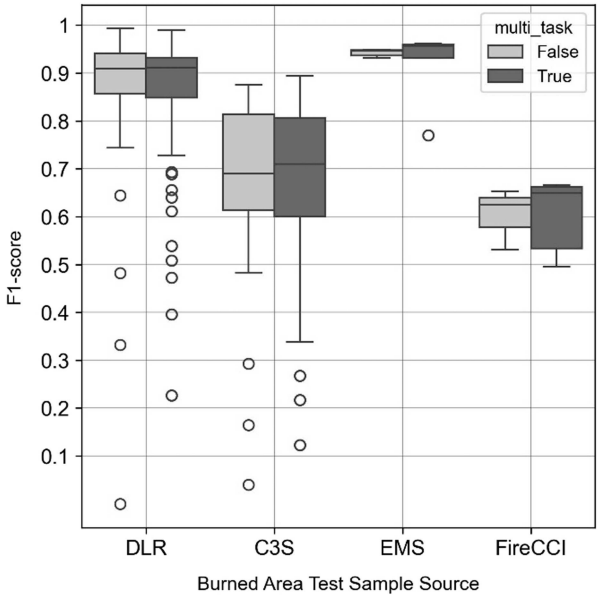


Figure 8. F1-score values (y-axis) stratified by burned area data sources (x-axis) of the U-Net models with EfficientNet-B4.

severely (−7.79% IoU), burned area performance improved moderately (IoU +5.96%, recall +8.99%), suggesting capacity reallocation toward the more challenging task. U-Net with MobileNetV2 demonstrated a moderate decline (IoU −3.58%) with precision-recall trade-offs (precision −6.57%, recall +3.12%).

Performance varied substantially across burned area test datasets, reflecting differences in fire characteristics, geographic contexts, and data quality (Figure 8). Models achieved highest F1-scores on the in-house DLR dataset (median ~0.90), which comprises predominantly Mediterranean (Portugal, Spain) and Californian fires characterised by clearly visible perimeters, high burn severity, and relatively homogeneous fire scars in forest and shrublands. The C3S global dataset showed considerably lower and more variable performance, with numerous outliers below an F1-score of 0.30 (predominantly test scenes from southern

Africa and northern Australia), likely reflecting its broader geographic and ecological diversity, including challenging detection scenarios such as low-severity burns in tropical savannas, small agricultural fires, and complex reference labels. As the C3S dataset was a large contributor to the overall dataset size (Table 1), this explains the lower overall performance of all tested model setups (Table 2). For instance, a test scene from central Angola (Figure 9), dominated by Angolan Miombo woodlands, exemplifies these challenges: the region experiences extremely high fire frequency with short fire return intervals (1.9-2.2 years), accumulated smoke from agricultural burning during peak dry season, and spectral confusion between burned areas and bare agricultural soils (Catarino et al. 2020). Despite these adverse conditions, multi-task learning improved detection performance for this scene increasing F1 from 0.748 to 0.778 using the SegFormer-B4 model. The EMS rapid mapping products demonstrated the most consistent performance with minimal variance, attributed to recent, high-quality operational fire perimeters derived from expert visual interpretation. FireCCI products yielded intermediate results with moderate spread, reflecting a mix of fire types across Africa with varying severity levels and complex spectral backgrounds. Notably, multi-task models demonstrated slightly higher median F1-scores across the in-house, C3S, and FireCCI datasets than single-task models, suggesting that joint training provides modest but consistent generalisation benefits across diverse fire regimes.

Prediction confidences showed strong positive correlation with temporal proximity, with recent scars yielding high probabilities that progressively decreased as vegetation recovered (Figure 10). Despite this, models still frequently detected older burned areas, which appeared as false positives against the temporally restricted reference mask, resulted in low precision and numerous F1 < 0.30 outliers in the C3S dataset.

5. Discussion

5.1. Multi-task vs. single-task learning

Our results demonstrate that surface water and burned area segmentation exhibit meaningful task compatibility when appropriate architectures are employed, with notable benefits emerging particularly for the more challenging burned area detection task. Transformer-based architectures, particularly SegFormer-B4, achieved improved performance for both tasks under multi-task learning while also demonstrating high absolute accuracies (F1-score: 0.98 for water, 0.67 for burned area). Similarly, U-Net with

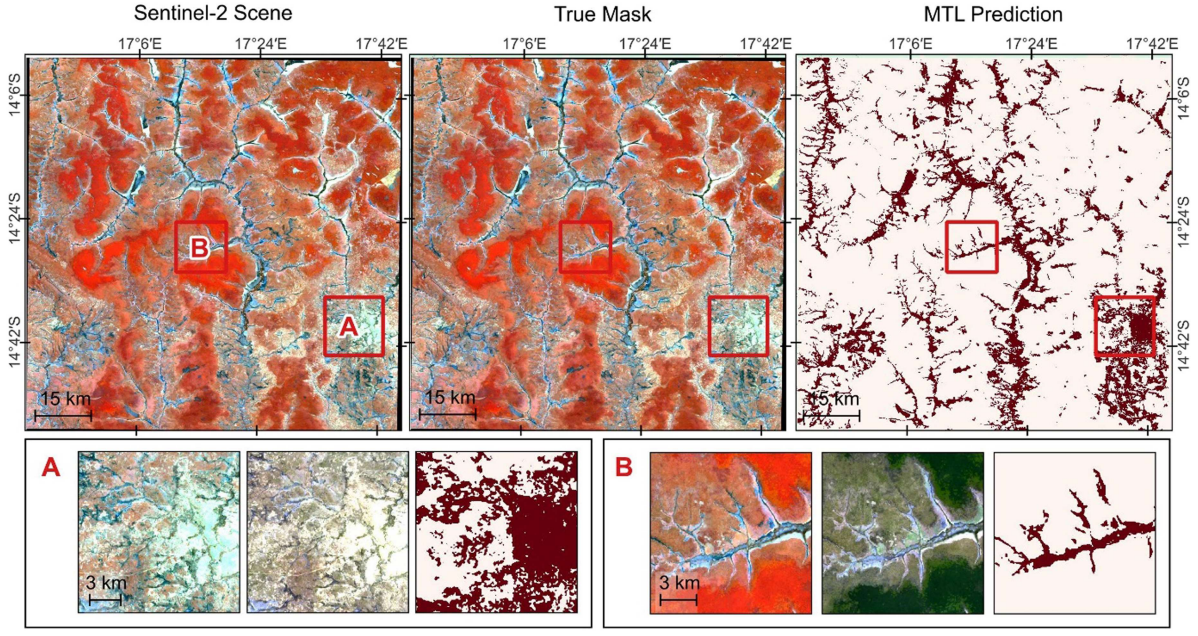


Figure 9. Burned area segmentation results of a Landsat-8 test scene in central Angola using multi-task SegFormer-B4.

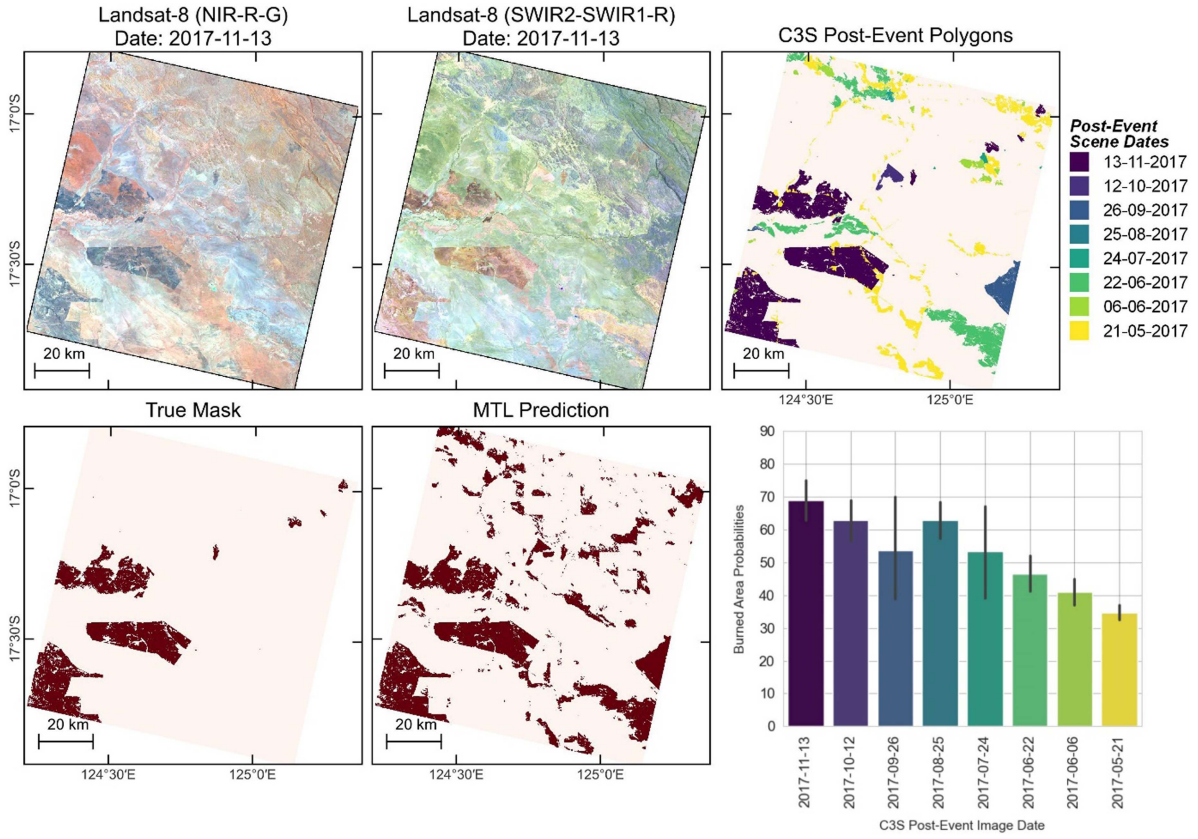


Figure 10. Burned area segmentation results for Landsat-8 test scenes in northern Australia using multi-task U-Net with EfficientNet-B4 (F1-score for this scene—0.622).

EfficientNet-B4 encoder maintained strong performance in multi-task mode with F1-scores exceeding 0.98 for surface water detection and reaching 0.70 for burned area mapping. The consistency of these results across architecturally distinct models—from lightweight CNNs to hierarchical vision transformers—and across geographically diverse test scenes suggests that the observed benefits reflect a general property of joint training rather than an architecture-specific artefact, providing initial evidence for the robustness and generalisability of the multi-task approach for remote sensing-based multi-hazard mapping. Nevertheless, repeated trials and formal significance testing remain important directions for future work to further confirm these findings.The positive results for burned area mapping under multi-task learning are particularly important given the inherent challenges of this task. The majority of tested architectures (three out of five) showed performance improvements for burned area detection in multi-task mode. This finding is significant because burned area mapping is challenging due to its high spectral variability depending on burn severity, vegetation type, and time since fire (Chuvieco et al. 2019), along with considerably limited reference data. Furthermore, in the multi-task setup, the joint prediction framework offers additional post-processing advantages: predictions from the more confident surface water detection task can be used to refine burned area outputs by filtering pixels where both decoders produce competing classifications, effectively leveraging the certainty of one task to improve the reliability of the other. This cross-task validation mechanism represents an additional operational advantage that enhances the practical value of multi-task learning for deployment in emergency mapping activities.

The observed asymmetry in multi-task benefits between tasks can be explained by their different baseline performance levels. Surface water detection achieved high accuracy in single-task mode (F1 scores above 0.97), leaving limited room for further improvement. In contrast, burned area detection showed moderate baseline performance (F1-scores ranging from 0.607 to 0.717), providing greater potential for improvement through auxiliary learning. The fact that multi-task learning maintained high surface water detection performance while improving burned area mapping demonstrates the practical value of

this approach. These results highlight that the success of multi-task learning in remote sensing depends not only on how spectrally similar the tasks are, but also on the choice of architecture. This is an important practical consideration when designing multi-hazard monitoring systems. From a computational efficiency perspective, the multi-task framework presents a favourable trade-off for operational deployment. Although multi-task models require longer training times compared to individual single-task models (Appendix D), this one-time overhead is offset by substantial gains at inference time (see Appendix C) which has been already reported in earlier studies (Standley and Zamir, 2019). In operational emergency mapping contexts, where hazard maps must be delivered rapidly to support disaster response, the ability to generate simultaneous surface water body and burned area predictions, for example, in under two minutes per Sentinel-2 scene, represents a practically meaningful and deployable efficiency gain.

5.2. Burned area mapping results and performance analysis

The multi-task learning models achieved F1-scores ranging from 0,624 to 0,701 for burned area detection. These results are lower than those reported in several recent studies: Seydi, Hasanlou, and Chanussot (2022) achieved overall accuracy of 0.97 using their Burnt-Net framework; Zhong and Gonsamo (2024) reached F1-scores above 0.7 on CanadaFire2023 dataset based on Sentinel-2 and Landsat-8/9 data; Seydi (2025) reported F1-scores above 0.74 based on AlphaEArth dataset, that comprise the high-resolution optical and thermal imagery. However, direct comparison is challenging due to substantial differences in dataset characteristics. Higher-performing studies typically focus on specific regions with consistent reference data and homogeneous ecosystems, which facilitates model learning. Our study employs globally diverse multi-source reference data spanning various biomes, with imagery from multiple sensors (Sentinel-2, Landsat-8) and diverse annotation methods (visual interpretation, semi-automatic classification, operational rapid mapping). Multiple validation studies of global burned area products have shown that performance varies substantially across different biomes (Boschetti et al. 2019; Franquesa et al. 2022). This heterogeneity therefore presents a substantially more challenging learning scenario. Nonetheless, the achieved F1-scores reaching 0.70 demonstrate promising performance for practical deployment in emergency mapping applications. The consistency of these results across geographically diverse test scenes spanning multiple continents and biome types provides additional evidence for the geographic robustness of the framework. However, transferability to additional sensor types beyond Sentinel-2 and Landsat-8, or to geographic regions and seasonal conditions underrepresented in the training data, remains to be evaluated in future work.

Stratified accuracy analysis based on burned area reference sub-datasets revealed systematic performance variations that reflect both geographic and temporal characteristics of the dataset sources. Specifically, datasets such as C3S and FireCCI demonstrated substantially lower performance metrics compared to EMS and in-house reference data; the latter being predominantly composed of samples from Mediterranean ecosystems. One reason for this, might be the large share of reference samples in tropical and subtropical ecoregions in both C3S and FireCCI datasets. Typically, burned areas in such environments are more difficult to detect due to high fire frequency, fast vegetation recovery and low spectral separability between burned area and background pixels (Boschetti et al. 2019; Franquesa et al. 2022; Melchiorre and Boschetti 2018). These findings also highlight a fundamental constraint of mono-temporal approaches: accurate burned area detection is critically dependent on the temporal proximity of the input imagery to the fire event. The C3S dataset structure, which incorporates time-series burned area polygons including scars up to several months post-fire, enabled detailed quantification of the temporal effect (Figure 10). Analysis revealed that models persistently detected older burn scars based on their enduring SWIR spectral signatures (Pleniou and Koutsias 2013). While this extended temporal detection capability could prove advantageous for applications focused on post-fire ecosystem recovery monitoring, vegetation regeneration studies, or long-term burn scar mapping, it presents challenges for operational rapid mapping scenarios where discriminating recent burns from historical scars is critical (Nolde et al. 2025). This temporal persistence of burned areas suggests that application-specific model optimisation or post-processing temporal filtering may be necessary to align model outputs with specific operational requirements.

An important pattern emerged in the precision-recall trade-off under multi-task learning. Most architectures showed increased recall for burned area detection in multi-task mode, indicating higher sensitivity to

burned pixels (see Appendix A). For instance, MobileNetV2's recall increased from 0,657 to 0,707 (+7,6%), while SegFormer-B0 showed a substantial recall gain from 0,607 to 0,687 (+13.2%), though both experienced precision decreases. This shift toward higher recall can be advantageous for emergency response applications where detecting all affected areas is prioritised over minimising false positives. Notably, SegFormer-B4 achieved simultaneous improvements: precision increased from 0,667 to 0,693 (+3,9%) and recall from 0,667 to 0,709 (+6,3%), indicating more refined decision boundaries under multi-task learning.

5.3. Deep learning architectures for multi-task learning

The choice of architecture plays a crucial role in determining multi-task learning success, with results revealing a clear capacity-dependent pattern of positive and negative transfer. Transformer-based architectures outperformed CNN-based models in multi-task scenarios: SegFormer-B4 achieved the most balanced results by simultaneously improving performance for both tasks; while smaller SegFormer-B0 improved burned area mapping at the cost of reduced surface water detection performance. Among CNN-based architectures, EfficientNet-B4 demonstrated the most stable multi-task behaviour, achieving minimal performance changes for surface water detection and modest improvements for burned area mapping. In contrast, lighter architectures such as ResNet50 and MobileNet-v2 exhibited negative transfer on the burned area task. These findings highlight that architecture selection is a critical consideration when designing multi-task frameworks for multi-hazard EO mapping.

5.4. Limitations & outlook

This study has several limitations that should be acknowledged. First, the training approach employed simple loss summation without adaptive weighting schemes, which may have contributed to the negative transfer observed in some architectures. Second, all experiments were conducted as a single training run, meaning that standard deviations cannot reported and small observed performance differences should be interpreted with caution. Third, the study examined only two hazard types and did not explore whether additional complementary tasks - such as land cover classification - might provide additional regularisation benefits or reveal different task interaction patterns. Fourth, manual refinement of the C3S reference data may have introduced annotation uncertainty, the impact of which was not systematically quantified. Fifth, the evaluation relies on benchmark datasets rather than independent validation against real operational disaster events, which limits conclusions regarding operational readiness. Furthermore, while scene-level splitting was applied to prevent geographic overlap between partitions, strict event-level independence—ensuring that no individual fire event contributes scenes to more than one data partition—was not explicitly enforced.

Building on the baseline established in this study, several promising directions for future work can be identified. From a methodological perspective, advanced multi-task learning optimisation strategies represent a natural next step. Uncertainty-based loss weighting (Kendall, Gal, and Cipolla 2017) could provide principled task balancing by automatically scaling each task's contribution to the combined loss, while gradient-based methods such as GradNorm (Chen et al. 2017) offer alternative approaches to mitigating negative transfer by directly modifying gradient magnitudes during training. Cross-task attention mechanisms and task affinity modelling could further improve feature sharing between the surface water and burned area decoders, potentially benefiting architectures such as ResNet50 that currently exhibit negative transfer. Incorporating more recent segmentation architectures—including Mask2Former, DINOv2-based models, and Swin-UNet—would allow assessment of whether the architecture-dependent patterns observed here generalise to state-of-the-art models. Additionally, formal cross-dataset generalisation experiments, such as leave-one-out, would provide stronger statistical evidence of transferability and are an important direction for future work.

From an application perspective, expanding the framework to additional natural hazards—such as landslides—would test whether the observed benefits extend to other hazard combinations relevant for integrated emergency management systems. Finally, validation against real operational disaster event records and actual emergency mapping case studies would be an important step toward confirming the framework's operational readiness and generalisability to real-world disaster scenarios.

6. Conclusion

This study evaluated the effectiveness of multi-task learning for simultaneous surface water and burned area detection from EO data, comparing five semantic segmentation architectures across single-task and multi-task configurations. Our results demonstrate that (RQ1) multi-task learning offers selective advantages for multi-hazard mapping, with effectiveness being highly architecture dependent. For burned area detection—the more challenging task—three out of five architectures showed performance improvements under multi-task learning, with F1-score gains ranging from +1.01% to +8.24%. Water body detection maintained excellent performance across various runs (F1-scores 0.941—0.988), with an average change of only −0.67% under multi-task learning. Regarding the tested model architectures (RQ2), SegFormer-B4 emerged as the optimal architecture for multi-task scenarios, achieving simultaneous improvements for both tasks (water F1 0.980, burned area F1 0.670), while EfficientNet-B4 demonstrated the most balanced performance among all architectures. These findings indicate that model capacity and architectural design are are important determinants of multi-task learning success, with lighter architectures like ResNet50 and SegFormer-B0 experiencing severe negative transfer for one or both tasks—highlighting that multi-task learning does not universally benefit all architecture types.

Performance analysis across burned area reference datasets revealed substantial variability reflecting differences in geographic context, fire characteristics, and annotation quality. Models achieved highest accuracies on the in-house DLR dataset (median F1 ~0.90), comprising predominantly clearly visible burn scars in the Mediterranean, while the globally diverse C3S dataset showed considerably lower and more variable performance (numerous F1 < 0.30 outliers), particularly for challenging scenarios including low-severity tropical savanna burns, small agricultural fires, and regions with short fire return intervals. The temporal structure of the C3S dataset exposed an important limitation: models persistently detected older burn scars, resulting in false positives when evaluated against temporally restricted reference masks. While this extended temporal detection capability could benefit post-fire ecosystem monitoring applications, it presents challenges for operational rapid mapping where discriminating recent burns from historical scars is critical.

Despite these limitations, this study establishes a meaningful baseline demonstrating the practical viability of standard multi-task learning frameworks for multi-hazard EO mapping, supported by substantial inference efficiency gains of 89–91% and parameter reductions of 17–48% compared to maintaining separate single-task models. Future work incorporating advanced task balancing strategies, more recent segmentation architectures, and validation against real operational disaster events will build directly on this foundation toward more robust and operationally deployable multi-hazard monitoring systems.

Author contributions

CRediT: **Aiym Orynbaikyzy:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Marc Wieland:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision, Writing – review & editing; **Moritz Rösch:** Conceptualization, Investigation, Validation, Writing – review & editing; **Michael Nolde:** Conceptualization, Investigation, Validation, Writing – review & editing; **Sandro Martinis:** Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through the project “Multimodale Daten und KI-Analysen für Interdisziplinäre Lagebilder im Katastrophenmanagement” (MOSAIK) under Grant 13N17560 and by DLR internal funding through the project “Resiliente Technologien für den Katastrophenschutz” (RESITEK).

ORCID

Marc Wieland http://orcid.org/0000-0002-1155-723X
Moritz Rösch http://orcid.org/0009-0003-2928-7009
Michael Nolde http://orcid.org/0000-0002-6981-9730

Data availability statement

The datasets generated within the study are openly available via Zenodo (DOI: https://doi.org/10.5281/zenodo.19879172) (Orynbaikyzy et al. 2026). The water body reference data used in this study is publicly available at Zenodo (DOI: https://doi.org/10.5281/zenodo.11278238) (Wieland et al. 2024). Burned area reference data were compiled from the following publicly accessible sources: (1) the Burned Area Reference Database (DOI: https://doi.org/10.21950/BBQQU7) (Franquesa et al. 2020), Satellite Burned Area Dataset (DOI: https://doi.org/10.5281/zenodo.6597139) (Luca et al. 2022), and HLS Burn Scar Scenes (DOI: https://doi.org/10.57967/HF/0956). Sentinel-2 and Landsat-8 data were accessed via Microsoft Planetary Computer (URL: https://planetarycomputer.microsoft.com/).

References

Ajmar, A., P. Boccardo, M. Broglia, J. Kucera, F. Giulio-Tonolo, and A. Wania. 2017. "Response to Flood Events: The Role of Satellite-Based Emergency Mapping and the Experience of the Copernicus Emergency Management Service, *Geophysical Monograph Series*, edited by Molinari, D., S. Menoni, F. Ballio, 211–228. Wiley. https://doi.org/10.1002/9781119217930.ch14.

Akhyar, A., M. Asyraf Zulkifley, J. Lee, T. Song, J. Han, C. Cho, S. Hyun, Y. Son, and B.-W. Hong. 2024. "Deep Artificial Intelligence Applications for Natural Disaster Management Systems: A Methodological Review." *Ecological Indicators* 163: 112067. https://doi.org/10.1016/j.ecolind.2024.112067.

Berman, M., A. R. Triki, and M. B. Blaschko. 2018. "The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4413–4421. Salt Lake City, UT: IEEE. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2018.00464.

Boschetti, L., D. P. Roy, L. Giglio, H. Huang, M. Zubkova, and M. L. Humber. 2019. "Global Validation of the Collection 6 MODIS Burned Area Product." *Remote Sensing of Environment* 235: 111490. https://doi.org/10.1016/j.rse.2019.111490.

Cannon, S. H., J. E. Gartner, R. C. Wilson, J. C. Bowers, and J. L. Laber. 2008. "Storm Rainfall Conditions for Floods and Debris Flows from Recently Burned Areas in Southwestern Colorado and Southern California." *Geomorphology* 96: 250–269. https://doi.org/10.1016/j.geomorph.2007.03.019.

Caruana, R. 1997. "Multitask Learning." *Machine Learning* 28: 41–75. https://doi.org/10.1023/A:1007379606734.

Catarino, S., M. M. Romeiras, R. Figueira, V. Aubard, J. M. N. Silva, and J. M. C. Pereira. 2020. "Spatial and Temporal Trends of Burnt Area in Angola: Implications for Natural Vegetation and Protected Area Management." *Diversity* 12: 307. https://doi.org/10.3390/d12080307.

Chen, Zhao, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2017. "GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks." arXiv Version 4. https://doi.org/10.48550/ARXIV.1711.02257.

Chuvieco, E., F. Mouillot, G. R. Van Der Werf, J. San Miguel, M. Tanase, N. Koutsias, M. García, et al. 2019. "Historical Background and Current Developments for Mapping Burned Area from Satellite Earth Observation." *Remote Sensing of Environment* 225: 45–64. https://doi.org/10.1016/j.rse.2019.02.013.

Crawshaw, Michael 2020. "Multi-Task Learning with Deep Neural Networks: A Survey." arXiv Version 1(2009.09796). https://doi.org/10.48550/ARXIV.2009.09796.

Delforge, D., V. Wathelet, R. Below, C. L. Sofia, M. Tonnelier, J. A. F. Van Loenhout, and N. Speybroeck. 2025. "EM-DAT: The Emergency Events Database." *International Journal of Disaster Risk Reduction* 124: 105509. https://doi.org/10.1016/j.ijdrr.2025.105509.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and M. Dehghani, et al. 2020. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, https://doi.org/10.48550/ARXIV.2010.11929.

European Commission, J. 2026. Copernicus Emergency Management Service.

European Space Agency, Airbus. 2022. Copernicus DEM. https://doi.org/10.5270/ESA-c5d3d65

Fan, J., Y. Yao, Y. Li, X. Zhang, J. Chen, J. B. Fisher, X. Zhang, et al. 2025. "Detecting Burned Vegetation Areas By Merging Spectral and Texture Features in a ResNet Deep Learning Architecture." *Remote Sensing* 17: 3665. https://doi.org/10.3390/rs17223665.

Feng, Y., X. Sun, W. Diao, J. Li, R. Niu, X. Gao, and K. Fu. 2023. "Height Aware Understanding of Remote Sensing Images Based on Cross-Task Interaction." *ISPRS Journal of Photogrammetry and Remote Sensing* 195: 233–249. https://doi.org/10.1016/j.isprsjprs.2022.11.014.

Franquesa, M., J. Lizundia-Loiola, S. V. Stehman, and E. Chuvieco. 2022. "Using Long Temporal Reference Units to Assess the Spatial Accuracy of Global Satellite-Derived Burned Area Products." *Remote Sensing of Environment* 269: 112823. https://doi.org/10.1016/j.rse.2021.112823.

Franquesa, M., M. K. Vanderhoof, D. Stavrakoudis, I. Gitas, E. Roteta, M. Padilla, and E. Chuvieco. 2020. *BARD: a global and regional validation burned area database*. https://doi.org/10.21950/BBQQU7.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. *Deep Residual Learning for Image Recognition*. https://doi.org/10.48550/ARXIV.1512.03385.

Hoeser, T., and C. Kuenzer. 2020. "Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends." *Remote Sensing* 12: 1667. https://doi.org/10.3390/rs12101667.

Hong, D., C. Qiu, A. Yu, Y. Quan, B. Liu, and X. Chen. 2023. "Multi-Task Learning for Building Extraction and Change Detection from Remote Sensing Images." *Applied Sciences* 13: 1037. https://doi.org/10.3390/app13021037.

International Charter Space and Major Disasters. 2026. International Charter Space and Major Disasters.

IPCC. 2023. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed., Cambridge University Press. https://doi.org/10.1017/9781009157896

Jain, P., S. C. P. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, and M. D. Flannigan. 2020. "A Review of Machine Learning Applications in Wildfire Science and Management." *Environmental Reviews* 28: 478–505. https://doi.org/10.1139/er-2020-0019.

Kendall, Alex, Yarin Gal, and Roberto Cipolla. 2017. "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics." arXiv Version 3(Preprint). https://doi.org/10.48550/ARXIV.1705.07115.

Kingma, D. P., and J. Ba. 2014. *Adam: A Method for Stochastic Optimization*. https://doi.org/10.48550/ARXIV.1412.6980.

Knopp, L., M. Wieland, M. Rättich, and S. Martinis. 2020. "A Deep Learning Approach for Burned Area Segmentation with Sentinel-2 Data." *Remote Sensing* 12: 2422. https://doi.org/10.3390/rs12152422.

Luca, C., F. Alessandro, M. Simone, G. Salvatore, G. Paolo, A. Daniele, B. Elena, and C. Tania. 2022. *Satellite Burned Area Dataset*. https://doi.org/10.5281/ZENODO.6597139.

Melchiorre, A., and L. Boschetti. 2018. "Global Analysis of Burned Area Persistence Time with MODIS Data." *Remote Sensing* 10: 750. https://doi.org/10.3390/rs10050750.

Mishra, M., R. Guria, S. Paul, R. M. Da Silva, and C. A. G. Santos. 2026. "Landslide Susceptibility Mapping Using a Machine Learning Approach and Different Environmental Factors in the Western Ghats Region (India)." *Earth Surf Processes Landf* 51: e70293. https://doi.org/10.1002/esp.70293.

Nolde, M., M. Rösch, T. Riedlinger, and H. Taubenböck. 2025. "Multi-Sensor Near-Realtime Burnt Area Monitoring Using a Superpixel-Based Graph Convolutional Network Approach." *GIScience & Remote Sensing* 62: 2498188. https://doi.org/10.1080/15481603.2025.2498188.

Orynbaikyzy, A., M. Wieland, M. Rösch, M. Nolde, and S. Martinis. 2026. *Multi-Hazard Mapping using Earth Observation data and Multi-Task Learning*. https://doi.org/10.5281/ZENODO.19879172.

Phillips, C., S. Roy, K. Ankur, and R. Ramachandran. 2023. *HLS Foundation Burnscars Dataset*. https://doi.org/10.57967/hf/0956.

Pleniou, M., and N. Koutsias. 2013. "Sensitivity of Spectral Reflectance Values to Different Burn and Vegetation Ratios: A Multi-Scale Approach Applied in a Fire Affected Area." *ISPRS Journal of Photogrammetry and Remote Sensing* 79: 199–210. https://doi.org/10.1016/j.isprsjprs.2013.02.016.

Rangel, A., J. Terven, D.-M. Córdova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, E. A. Chávez-Urbiola, F. Willars-Rodríguez, and G. Alfonso-Francia. 2025. "Tracking U.S. Land Cover Changes: A Dataset of Sentinel-2 Imagery and Dynamic World Labels (2016&Ndash;2024)." *Data* 10: 67. https://doi.org/10.3390/data10050067.

Rättich, Michaela, Martinis, Sandro, and Plank, Simon 2017) An Automatic Process Chain for Detecting Burn Scars Using Sentinel-2 Data.Presented at the 11th EARSeL Forest Fires Special Interest Group Workshop, September 25, 2017

Rodriguez, M., M.-T. Pham, M. Sudmanns, Q. Poterek, and O. Narvaez. 2025. *Enhancing deep learning performance on burned area delineation from SPOT-6/7 imagery for emergency management*. https://doi.org/10.48550/ARXIV.2512.12056.

Ronneberger, O., P. Fischer, and T. Brox. 2015. " U-Net: Convolutional Networks for Biomedical Image Segmentation." arXiv Version 1(Preprint). https://doi.org/10.48550/ARXIV.1505.04597.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. 2018. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. https://doi.org/10.48550/ARXIV.1801.04381.

Seydi, S. T. 2025. "Deep Learning-Based Burned Area Mapping Using Bi-Temporal Siamese Networks And." *AlphaEarth Foundation Datasets*. https://doi.org/10.48550/arXiv.2509.07852.

Seydi, S. T., M. Hasanlou, and J. Chanussot. 2022. "Burnt-Net: Wildfire Burned Area Mapping with Single Post-Fire Sentinel-2 Data and Deep Learning Morphological Neural Network." *Ecological Indicators* 140: 108999. https://doi.org/10.1016/j.ecolind.2022.108999.

Shakesby, R. A., and S. H. Doerr. 2006. "Wildfire as a Hydrological and Geomorphological Agent." *Earth-Science Reviews* 74: 269–307. https://doi.org/10.1016/j.earscirev.2005.10.006.

Standley, T., A. R. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese. 2019. "Which Tasks Should Be Learned Together in Multi-task Learning?" arXiv Version 4. https://doi.org/10.48550/ARXIV.1905.07553.

Sui, T., Q. Huang, M. Wu, M. Wu, and Z. Zhang. 2024. "BiAU-Net: Wildfire Burnt Area Mapping Using Bi-Temporal Sentinel-2 Imagery and U-Net with Attention Mechanism." *International Journal of Applied Earth Observation and Geoinformation* 132: 104034. https://doi.org/10.1016/j.jag.2024.104034.

Tan, M., and Q. V. Le. 2019. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. https://doi.org/10.48550/ARXIV.1905.11946.

Tralli, D. M., R. G. Blom, V. Zlotnicki, A. Donnellan, and D. L. Evans. 2005. "Satellite Remote Sensing of Earthquake, Volcano, Flood, Landslide and Coastal Inundation Hazards." *ISPRS Journal of Photogrammetry and Remote Sensing* 59: 185–198. https://doi.org/10.1016/j.isprsjprs.2005.02.002.

Vandenhende, S., S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool. 2020. "Multi-Task Learning for Dense Prediction Tasks: A Survey." arXiv Version 3. https://doi.org/10.48550/ARXIV.2004.13379.

Volpi, M., and D. Tuia. 2018. "Deep Multi-Task Learning for a Geographically-Regularized Semantic Segmentation of Aerial Images." *ISPRS Journal of Photogrammetry and Remote Sensing* 144: 48–60. https://doi.org/10.1016/j.isprsjprs.2018.06.007.

Wan, Z., S. Ye, F. Wang, S. Liu, and L. Peng. 2025. "Uncertainty Weighted Multi Task Learning for Robust Traffic Scene Semantic Understanding." *Scientific Reports* 15: 40993. https://doi.org/10.1038/s41598-025-24838-3.

Wang, Y., S. Li, Y. Lin, and M. Wang. 2021. "Lightweight Deep Neural Network Method for Water Body Extraction from High-Resolution Remote Sensing Images with Multisensors." *Sensors (Basel, Switzerland)* 21: 7397. https://doi.org/10.3390/s21217397.

Wieland, M., F. Fichtner, S. Martinis, S. Groth, C. Krullikowski, S. Plank, and M. Motagh. 2024. "S1S2-Water: A Global Dataset for Semantic Segmentation of Water Bodies from Sentinel- 1 and Sentinel-2 Satellite Images." *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 17: 1084–1099. https://doi.org/10.1109/JSTARS.2023.3333969.

Wieland, M., S. Martinis, R. Kiefl, and V. Gstaiger. 2023. "Semantic Segmentation of Water Bodies in Very High-Resolution Satellite and Aerial Images." *Remote Sensing of Environment* 287: 113452. https://doi.org/10.1016/j.rse.2023.113452.

Xie, E., W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. 2021. *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*

Xu, Z., H. Wu, W. Jiang, and J. Geng. 2025. *Graph Aggregation Prototype Learning for Semantic Change Detection in Remote Sensing*. https://doi.org/10.48550/ARXIV.2507.10938.

Yuan, X., J. Shi, and L. Gu. 2021. "A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery." *Expert Systems with Applications* 169: 114417. https://doi.org/10.1016/j.eswa.2020.114417.

Zhang, Y., and Q. Yang. 2018. "An Overview of Multi-Task Learning." *National Science Review* 5: 30–43. https://doi.org/10.1093/nsr/nwx105.

Zhong, Z., and Gonsamo, A. 2024. CanadaFire2023: Burned Area Mapping Datasets and Benchmarks for Canadian Wildfires in 2023.

Zhou, L., T. Chen, F. Yang, Y. Pan, L. Huang, X. Huang, and P. Lai. 2026. "MSFlood-Net: A Physically Informed Deep Learning Model Integrating Multi-Source Data for Flood Inundation Mapping." *Environmental Modelling & Software* 196: 106779. https://doi.org/10.1016/j.envsoft.2025.106779.

Appendices

Appendix A: F1-score, precision and recall values of performed experiments for single-task (STL) and multi-task (MTLM) learning models. Bold values indicate the best performance for each metric.

		Water						Burned Area					
		STL			MTL			STL			MTL		
Decoder	Encoder	*F1*	*Prec.*	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*	*Prec.*	*Rec.*
Transf.	SegFormer-B0	0.981	0.979	0.984	0.941	0.962	0.921	0.607	0.530	0.712	0.632	0.533	0.776
	SegFormer-B4	0.971	0.985	0.957	0.980	0.985	0.975	0.619	0.569	0.678	0.670	0.599	0.759
U-Net	EfficientNet-B4	0.987	**0.994**	0.979	0.986	**0.996**	0.977	0.694	0.629	**0.773**	**0.701**	**0.653**	0.756
	MobileNet-v2	0.986	0.986	0.986	**0.988**	0.992	**0.984**	**0.717**	**0.670**	0.770	0.700	0.626	**0.794**
	ResNet50	**0.989**	0.988	**0.989**	0.986	**0.996**	0.976	0.680	0.634	0.734	0.624	0.559	0.706

Appendix B: Total number of model parameters for single-task learning (STL) and multi-task learning (MTL) configurations across all evaluated architectures. STL Combined represents the total parameter count when maintaining separate water body and burned area models. Reduction (%) indicates the parameter savings achieved by the MTL framework relative to STL Combined

Decoder	Encoder	STL: Water	STL: Burned Area	STL (Combined)	MTL	Reduction [%]
Transformer	SegFormer-B0	3,721,185	3,721,185	7,442,370	4,116,706	45%
	SegFormer-B4	64,005,569	64,005,569	128,011,138	67,160,514	48%
U-Net	EfficientNet-B4	19,544,617	19,544,617	39,089,234	22,223,674	43%
	MobileNet-v2	6,666,193	6,666,193	13,332,386	11,073,250	17%
	ResNet50	32,584,913	32,584,913	65,169,826	41,599,970	36%

Appendix C: Average inference time (in minutes) per Sentinel-2 scene (10,980 × 10,980 pixels) for single-task learning (STL) and multi-task learning (MTL) configurations. STL Combined represents the total time required when running surface water and burned area models separately

Decoder	Encoder	STL: Water	STL: Burned Area	STL (Combined)	MTL
Transformer	SegFormer-B0	6.22	2.19	8.41	0.74
	SegFormer-B4	14.10	4.50	18.60	1.71
U-Net	EfficientNet-B4	6.71	2.35	9.06	1.12
	MobileNet-v2	5.74	2.32	8.06	0.81
	ResNet50	7.69	2.26	9.95	0.83

Appendix D: Training time (hours) for single-task learning (STL) and multi-task learning (MTL) configurations across all evaluated architectures. STL Combined represents the total training time required when training separate water body and burned area models independently

Decoder	Encoder	STL: Water [h]	STL: Burned Area [h]	STL: Combined [h]	MTL [h]
Transformer	SegFormer-b0	19.5	19.2	38.7	56.3
	SegFormer-b4	32.4	29.3	61.7	62.3
U-Net	EfficientNet-B4	29.3	19.2	38.5	46.0
	MobileNet-v2	19.5	19.1	38.6	50.0
	ResNet50	19.4	19.2	38.6	54.9