remote sensing

MDPI

Article

GL-FCR: A Global–Local Feature Context and Reconstruction Decoder for Transferable Foundation Model-Based Flood Mapping

Tarek Gherib [1], Mohamed Said Guettouche [1], Boularbah Souissi [2], Abdelaziz Rabehi [3], Mustapha Habib [4,*], Abdelkader Riche [5] and Mawloud Guermoui [6]

[1] Faculty of Earth Sciences, Geography and Territorial Planning, University of Sciences and Technology Houari Boumediene, BP 32 Bab Ezzouar, Algiers 16111, Algeria; tgherib@usthb.edu.dz (T.G.); medsaid.guettouche@usthb.edu.dz (M.S.G.)
[2] Department of Telecommunication, University of Sciences and Technology Houari Boumediene, BP 32 Bab Ezzouar, Algiers 16111, Algeria; boulerbah.souissi@usthb.edu.dz
[3] Telecommunication and Smart Systems Laboratory, Faculty of Sciences and Technology, Ziane Achour University, Djelfa 17000, Algeria; abdelaziz.rabehi@univ-djelfa.dz
[4] Division of Building Design and Technology, Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden
[5] Department of Earth Sciences, University of Florence, Via La Pira, 4–50121 Florence, Italy; abdelkader.riche@unifi.it
[6] Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 9-38123 Trento, Italy; mawloud.guermoui@unitn.it
* Correspondence: mushab@kth.se

Highlights

What are the main findings?

- A novel GL-FCR downstream decoder was proposed for foundation model-based flood segmentation.
- Prithvi-EO-2.0 + GL-FCR showed strong geographic and cross-dataset generalization, while SAR provided a complementary alternative for flood mapping.

What are the implications of the main findings?

- Global–local feature reconstruction effectively adapts EO foundation-model representations for accurate and transferable flood mapping.
- Optical imagery performed best under clear conditions, while SAR remains valuable when optical observations are limited or degraded.

Abstract

Accurate and transferable flood segmentation from Earth observation (EO) imagery remains challenging because models trained on specific geographic regions and environmental conditions often exhibit limited generalization. This study proposes Global–Local Feature Context and Reconstruction (GL-FCR), a task-specific downstream decoder designed to adapt compact EO foundation-model representations for spatially detailed flood mapping. Integrated with Prithvi-EO-2.0, GL-FCR enhances the reconstruction of dense flood maps while maintaining contextual information from the pretrained representation. On the SEN1Floods11 standard test set, Prithvi + GL-FCR achieved a flood F1 score of 91.08%, compared with 90.02% on the geographically held-out Bolivia subset in the original experiment. Across three random seeds, GL-FCR achieved mean flood IoU scores of 83.07 ± 0.50% and 81.53 ± 0.27% on the standard test and Bolivia subsets, respectively,

Academic Editor: Teodosio Lacava

Received: 8 September 2026
Revised: 16 October 2026
Accepted: 19 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

demonstrating reproducible performance across training runs. Additional SAR-only and optical–SAR experiments showed that TerraMind-B + GL-FCR achieved the strongest SAR-only performance, with F1 scores of 81.53% and 82.28% on the standard test and Bolivia subsets, respectively, whereas adding SAR to strong clear-scene optical representations provided limited additional benefit. Furthermore, zero-shot cross-dataset evaluation across all 509 WorldFloods v2 scenes yielded a flood F1 score of 84.56% and a flood IoU of 73.24% for Prithvi + GL-FCR, compared with 83.35% and 72.15%, respectively, for the Prithvi native decoder, without additional training or adaptation. Overall, the results demonstrate that task-specific global–local reconstruction can effectively leverage EO foundation-model representations for accurate and transferable flood segmentation across geographic regions and datasets.

Keywords: flood segmentation; earth observation foundation models; Prithvi-EO-2.0; global–local feature reconstruction; multimodal remote sensing

1. Introduction

Floods are among the most frequent and damaging natural hazards worldwide, affecting human populations, infrastructure, economic activities, and ecosystems, while growing exposure and climate-related changes are expected to further increase flood risk in many regions [1,2]. Satellite Earth observation (EO) has consequently become an essential source of information for flood monitoring, providing spatially consistent observations over large and often inaccessible areas and supporting rapid inundation mapping, emergency response, damage assessment, and flood-risk management [3,4]. Optical sensors, particularly Sentinel-2, provide rich multispectral information for distinguishing water from surrounding land-cover classes, whereas Sentinel-1 Synthetic Aperture Radar (SAR) offers day-and-night observations and the ability to acquire useful measurements through cloud cover [4]. At the operational level, the Copernicus Emergency Management Service (CEMS) Global Flood Monitoring (GFM) service demonstrates the capability of Sentinel-1 for fully automated, near-real-time flood monitoring at the global scale, combining complementary SAR-based flood-mapping algorithms with reference-water and contextual information [5]. Nevertheless, reliable and transferable flood delineation remains challenging because flood signatures vary with land cover and environmental conditions, while optical observations are affected by clouds and surface obstruction and SAR measurements can be complicated by speckle, vegetation, urban structures, and terrain effects [4].

The increasing availability of EO data and annotated flood inventories has promoted a transition from conventional machine-learning approaches toward deep-learning-based semantic segmentation, enabling hierarchical spatial features to be learned directly from satellite observations. Convolutional neural networks (CNNs), particularly encoder–decoder architectures such as U-Net, DeepLabv3+, residual and attention-based variants, have been widely applied to Sentinel-1, Sentinel-2, and combined optical–SAR observations for flood and water segmentation, demonstrating their ability to extract contextual features and recover spatially coherent inundation patterns across heterogeneous landscapes [6–9]. More recently, Transformer-based and hybrid CNN–Transformer architectures have been increasingly investigated for flood segmentation because their attention mechanisms enable broader spatial relationships and contextual information to be modelled beyond the local receptive fields characteristic of conventional convolutional operations. Applications based on Vision Transformers, SegFormer, Swin Transformer, and CNN–Transformer hybrid architectures have demonstrated their potential for extracting

inundated areas from optical, SAR, and multimodal EO imagery [10–14]. Despite these advances, conventional task-specific deep networks generally depend on sufficiently representative annotated data, and high performance within the training distribution does not necessarily translate into equivalent performance across unseen flood events, geographic regions, sensors, or environmental conditions. This sensitivity to domain and geographic shifts remains particularly important for operational flood mapping, where models must generalize beyond the locations and conditions represented during training.

To improve the transferability of task-specific models and reduce their dependence on large amounts of task-specific labelled data, recent research has increasingly shifted toward Earth observation foundation models (EO-FMs). By pretraining large neural networks on extensive and geographically diverse satellite archives, EO-FMs learn general-purpose representations that can subsequently be adapted to a range of downstream remote-sensing tasks with comparatively limited task-specific supervision. Representative models include SatMAE [15], Scale-MAE [16], Prithvi [17], CROMA [18], DOFA [19], SatlasPretrain [20], Galileo [21], and TerraMind [22], which differ in their pretraining objectives, sensor inputs, spatial scales, and mechanisms for accommodating heterogeneous EO data. Among these developments, Prithvi has demonstrated the potential of large-scale pretraining on multispectral Harmonized Landsat and Sentinel-2 observations for downstream geospatial applications, including flood segmentation [17,22], while TerraMind extends the foundation-model paradigm toward multimodal EO representation learning across optical, SAR, and other geospatial modalities [22]. This multimodal capability is particularly relevant to flood mapping because optical imagery provides rich spectral information for water delineation, whereas SAR can maintain observations under cloud cover; however, effectively exploiting these complementary representations within downstream flood-segmentation architectures remains an important challenge.

Despite the strong representation capabilities of EO foundation models, their effectiveness for flood mapping also depends on how their compact latent features are adapted to dense pixel-level prediction. Existing studies have explored different downstream strategies, including conventional segmentation decoders with Prithvi-EO-2.0 [17], the U-Prithvi architecture [23], and the Prithvi-CAFE framework, which enhances Prithvi features through complementary adaptive fusion [23]. Although these approaches demonstrate the effectiveness of pretrained EO representations for flood segmentation, translating coarse foundation-model features into spatially detailed and transferable flood maps remains challenging. In particular, existing downstream architectures do not explicitly combine adaptive directional context with progressive spatial reconstruction and repeated global–local feature integration. This motivates the development of a task-specific decoding strategy capable of preserving broad contextual information while progressively recovering fine-scale flood structures.

To address this gap, this study proposes Global–Local Feature Context and Reconstruction (GL-FCR), a task-specific downstream decoder designed to effectively adapt EO foundation-model representations for dense flood segmentation. The proposed framework is evaluated under different sensor configurations and generalization settings to investigate its effectiveness and transferability for flood mapping. The main contributions of this study are:

- A novel GL-FCR downstream decoder that integrates adaptive directional context, progressive spatial feature reconstruction, and stage-wise global–local feature fusion to improve the translation of compact foundation-model representations into spatially detailed flood maps.
- A comprehensive evaluation of optical, SAR, and optical–SAR configurations, incorporating different EO foundation-model representations to investigate their effectiveness and sensor complementarity for flood segmentation.

- An assessment of geographic and cross-dataset generalization using the geographically held-out Bolivia subset and a comprehensive zero-shot evaluation across all 509 WorldFloods scenes, comparing the proposed GL-FCR decoder with the native Prithvi decoder and SegFormer-B2 without additional training, fine-tuning, or adaptation.
- Extensive comparisons with conventional deep-learning and recent EO foundation-model approaches, complemented by ablation experiments to quantify the contribution of the principal GL-FCR components.

2. Materials and Methods

2.1. Dataset

Two publicly available datasets, SEN1Floods11 and WorldFloods, were used to evaluate the proposed framework under complementary geographic and cross-dataset conditions. SEN1Floods11 was used for model training, validation, and primary testing, including a geographically held-out evaluation over Bolivia, whereas WorldFloods was used exclusively for external zero-shot evaluation. The geographic distribution of the samples considered in this study and their respective roles in the experimental protocol are illustrated in Figure 1.

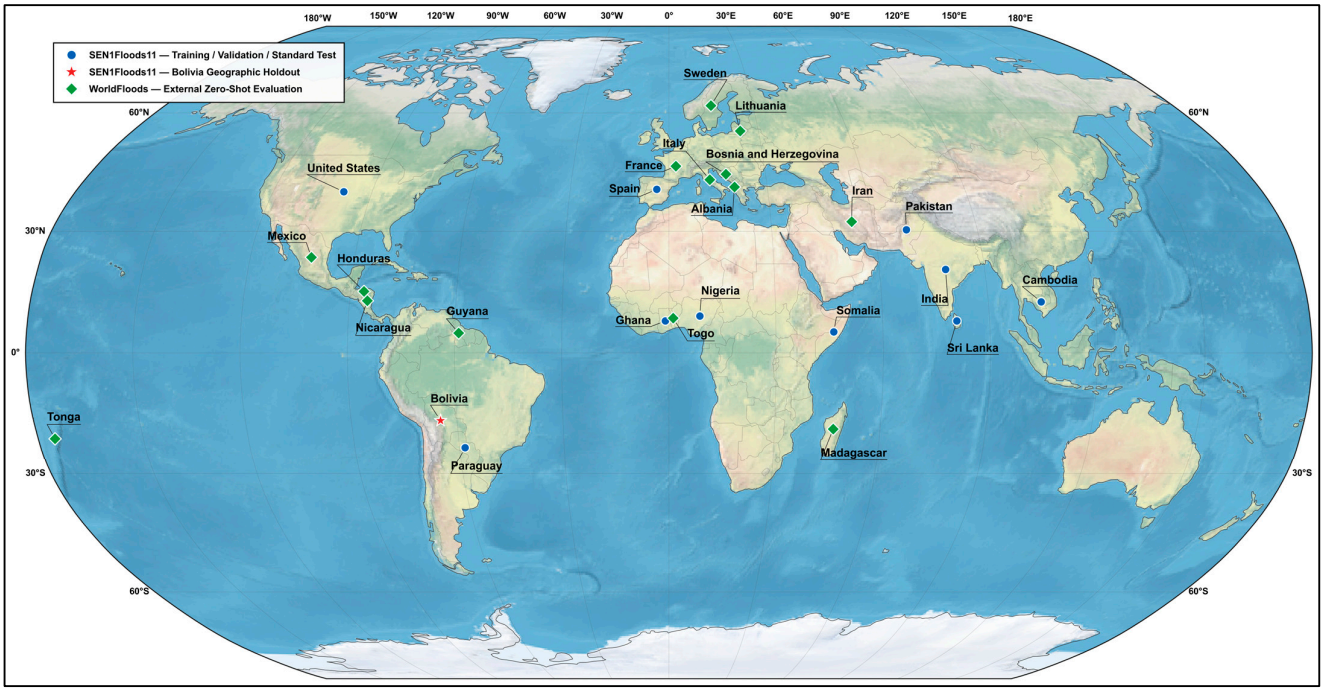


Figure 1. Global distribution of the datasets used in this study, showing the SEN1Floods11 training, validation, and the geographically held-out Bolivia subset, and the WorldFloods scenes selected for external zero shot evaluation.

SEN1Floods11 was originally introduced by Bonafilia et al. [24] as a publicly available georeferenced benchmark for global flood-water mapping using Sentinel-1 SAR imagery. The dataset was constructed from 11 flood events and comprises 4831 non-overlapping image chips of 512 × 512 pixels, covering approximately 120,406 km^2. In addition to Sentinel-1 observations, Sentinel-2 imagery was incorporated during dataset generation and annotation. The benchmark was subsequently investigated by Bai et al. [7] in a multisensor setting combining Sentinel-1 SAR and Sentinel-2 multispectral observations, with all bands sampled to a common 10 m spatial resolution.

In this study, only the 446 manually annotated samples were considered, excluding the weakly labelled portion of SEN1Floods11. The manually annotated benchmark comprises 252 samples for training, 89 for validation, and 90 for standard testing, together with an additional 15 samples from Bolivia that are geographically held out to evaluate model performance on an unseen flood event [7,24]. The flood events represented in SEN1Floods11 are located in Bolivia, Ghana, India, Cambodia, Nigeria, Pakistan, Paraguay, Somalia, Spain, Sri Lanka, and the United States. As shown in Figure 1, Bolivia is distinguished from the remaining SEN1Floods11 locations because it constitutes the geographic holdout in the experimental protocol.

WorldFloods is a publicly available global dataset designed for flood-extent mapping from Sentinel-2 multispectral imagery [25]. In its extended release, the dataset contains 509 flood-extent maps associated with 144 flood events, providing geographically diverse observations for assessing flood-mapping models beyond a single benchmark. In the present study, WorldFloods was used exclusively as an independent external dataset: no WorldFloods samples were used during model training, validation, fine-tuning, or adaptation. All 509 WorldFloods scenes were evaluated under a zero-shot cross-dataset protocol using Prithvi + GL-FCR, the Prithvi native decoder, and SegFormer-B2. A fixed randomly sampled subset of 20 scenes was retained for detailed scene-level comparison. These scenes represent 14 countries: France, Lithuania, Tonga, Madagascar, Sweden, Iran, Bosnia and Herzegovina, Nicaragua, Togo, Mexico, Honduras, Italy, Albania, and Guyana as illustrated in Figure 1. The complete 509-scene evaluation was additionally used to quantify global performance and scene-level variability. This external evaluation complements the SEN1Floods11 Bolivia holdout by examining model transfer to an independent dataset acquired across different geographic and environmental conditions.

Prior to model training and evaluation, the manually annotated SEN1Floods11 samples were partitioned into non-overlapping 256 × 256-pixel patches, while preserving the official training, validation, standard test, and geographically held-out Bolivia partitions. For the optical experiments, six Sentinel-2 spectral bands (B2, B3, B4, B8A, B11, and B12) were used, whereas the SAR experiments employed the Sentinel-1 VV and VH channels. The Sentinel-1 patches were extracted from the original SEN1Floods11 SAR imagery without additional radiometric conversion or percentile clipping during dataset preparation, preserving the dataset-provided VV and VH backscatter values. The reference masks were converted to a common binary representation, with 0 representing non-water and 1 representing water/flood, while invalid or ignored pixels were assigned a value of 255 and excluded from both model optimization and metric computation. The positive class (1) includes both permanent water bodies and temporary floodwater, as represented in the reference annotations. Accordingly, "water" refers broadly to water-covered pixels, whereas "flood" refers specifically to temporary inundation. The term "water/flood" describes the combined positive class adopted in this study. The reported "Flood IoU" and "Flood F1" metrics therefore quantify segmentation performance for the combined water/flood class rather than temporary floodwater alone.

Input imagery was normalized according to the statistics associated with the corresponding pretrained foundation model. Specifically, the Prithvi optical configuration used its six-band normalization statistics, whereas SegFormer-B2 employed band-wise means and standard deviations calculated from valid pixels in the training partition. For the SAR-CNN configuration, VV and VH were standardized using channel-wise means of −10.394032 and −17.241774 and standard deviations of 4.038712 and 4.754204, respectively. These statistics were applied during model input preparation rather than during the initial dataset extraction. For the external WorldFloods evaluation, the reference masks were converted to the same binary convention, with invalid or cloud-contaminated pixels treated as ignored regions. Because the WorldFloods scenes are larger than the model

input size, zero-shot inference was performed using 256 × 256-pixel tiles with a stride of 128 pixels, and the predicted probabilities in overlapping regions were averaged to reconstruct the final scene-level flood maps. No WorldFloods samples were used for training, fine-tuning, or model adaptation.

2.2. Proposed GL-FCR Framework

The proposed Global–Local Feature Context and Reconstruction (GL-FCR) decoder is designed to progressively recover spatial detail from the compact representation produced by the Prithvi-EO-2.0-300M encoder provided by Szwarcman et al. [17] while preserving its large-scale contextual information. As illustrated in Figure 2, the encoder representation is processed by a Global Directional Context (GDC) module and subsequently reconstructed through four decoding stages, each combining PixelShuffle-based upsampling with a Stage Global–Local Gate (SGLG). The internal structures of GDC and SGLG are detailed in Figure 3.

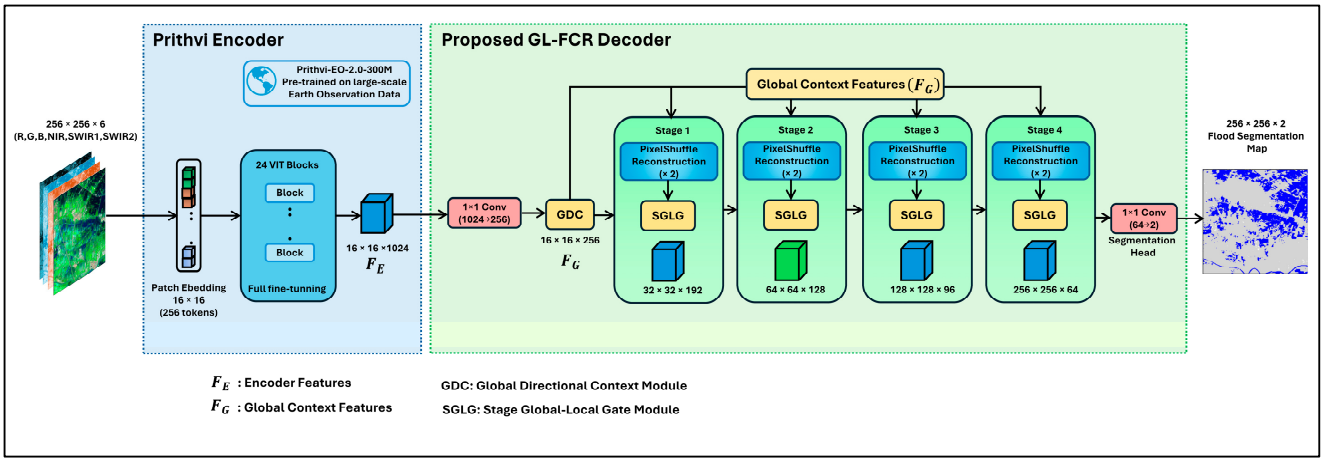


Figure 2. Overall architecture of the proposed GL-FCR framework for flood segmentation, comprising the Prithvi-EO-2.0-300M encoder and the proposed GL-FCR decoder, which progressively reconstructs spatial features while integrating global context across multiple decoding stages.

The combination of these components is motivated by the spatial characteristics of flooded areas, which may range from extensive inundated regions to narrow, fragmented water structures with irregular boundaries. Multi-scale feature reconstruction is intended to recover spatial information at different resolutions, while contextual feature processing helps distinguish flooded surfaces from spectrally similar non-flooded areas. Within GDC, directional convolutions capture horizontal and vertical spatial relationships that may be relevant to elongated flood structures and complex inundation boundaries. During progressive decoding, PixelShuffle upsampling restores spatial resolution, while SGLG adaptively balances the reconstructed local features with the broader contextual representation. These mechanisms are therefore designed to address complementary requirements of flood segmentation: preserving large-scale spatial consistency, resolving local flood structures, and reducing ambiguity during pixel-level classification. Their combined effectiveness is subsequently examined through the ablation experiments presented in Section 3.5.

For an input patch X, the fully fine-tuned Prithvi encoder produces

$$F_E = \varepsilon(X), \quad F_E \in \mathbb{R}^{16\times16\times1024}, \tag{1}$$

where $\varepsilon(.)$ denotes the Prithvi encoder and F_E represents its output features. These features are projected to 256 channels before entering the contextual decoder:

$$F_0 = \phi\left[\mathcal{N}\left(Conv_{1\times1}(F_E)\right)\right], \quad F_0 \in \mathbb{R}^{16\times16\times256}, \tag{2}$$

where $\mathcal{N}$ and ϕ denote normalization and GELU activation, respectively.

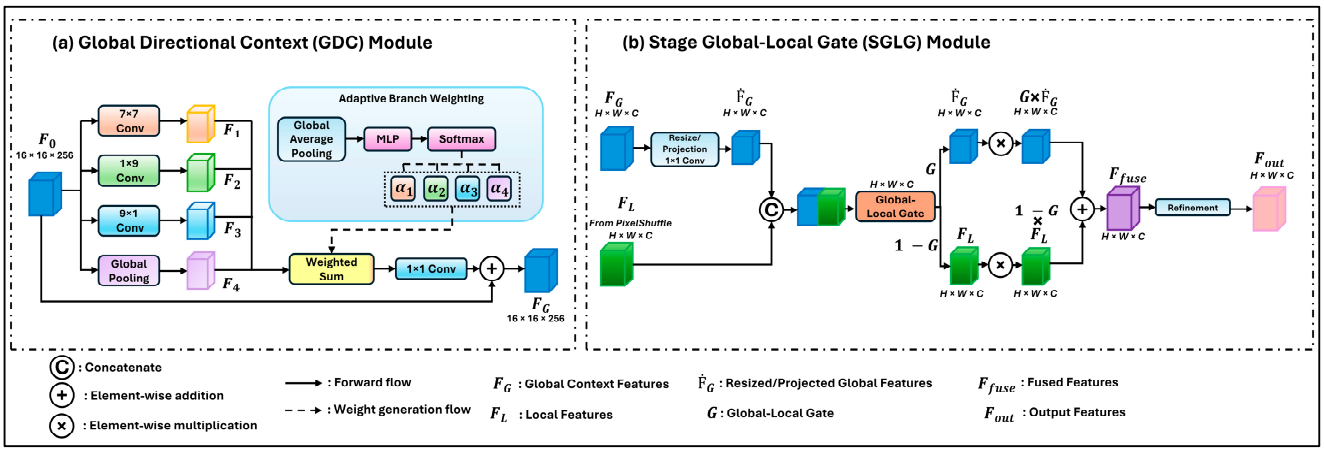


Figure 3. Detailed architecture of the proposed GL-FCR modules: (**a**) Global Directional Context (GDC) module for adaptive multi directional context extraction and (**b**) Stage Global–Local Gate (SGLG) module for gated integration of global context with progressively reconstructed local features.

2.2.1. Global Directional Context (GDC)

The GDC module extracts complementary contextual representations using three spatial branches and one global branch, as shown in Figure 3a:

$$F_1 = \mathcal{B}_{7\times7}(F_0), F_2 = \mathcal{B}_{1\times9}(F_0), F_3 = \mathcal{B}_{9\times1}(F_0), F_4 = \mathcal{B}_G(F_0), \tag{3}$$

where F_1, F_2, F_3, and F_4 are the contextual feature maps generated by the four parallel GDC branches. $\mathcal{B}_{7\times7}(.)$ denotes the two-dimensional contextual branch used to capture broader spatial patterns, while $\mathcal{B}_{1\times9}(.)$ and $\mathcal{B}_{9\times1}(.)$ denote the horizontal and vertical contextual branches, respectively. $\mathcal{B}_G(.)$ represents the global-context branch based on global average pooling and subsequent feature transformation. F_0 is the projected encoder representation defined previously in Equation (2).

The contribution of these branches is determined adaptively from the input representation:

$$[\alpha_1, \alpha_2, \alpha_3, \alpha_4] = Softmax\left[MLP\left(GAP(F_0)\right)\right], \quad \sum_{i=1}^{4}\alpha_i = 1, \tag{4}$$

where $GAP(.)$ denotes global average pooling, $MLP(.)$ is the branch-weight prediction network, and α_i is the normalized adaptive weight assigned to the i_{th} contextual branch.

The weighted contextual features are then aggregated and combined with the original projected representation through a residual connection:

$$F_G = F_0 + \mathcal{P}_G\left(\sum_{i=1}^{4}\alpha_i F_i\right), \quad F_G \in \mathbb{R}^{16\times16\times256}, \tag{5}$$

where $\mathcal{P}_G(.)$ denotes the GDC output projection and F_G is the resulting global context representation. This adaptive formulation allows the relative importance of the different contextual patterns to vary according to the input scene.

2.2.2. Progressive Reconstruction and Global–Local Fusion

The decoder progressively restores spatial resolution through four PixelShuffle-based reconstruction stages. At stage s_i, the local representation is obtained as:

$$F_L^{(s)} = \mu_s\left(F_{out}^{(s-1)}\right), \quad s = 1, \ldots, 4, \tag{6}$$

where s denotes the decoding stage, $\mu_s(.)$ is the corresponding PixelShuffle-based reconstruction operator, and $F_L^{(s)}$ represents the reconstructed local features at that stage. For the first stage, $F_{out}^{(0)} = F_G$. The four stages progressively produce feature maps of $32 \times 32 \times 192, 64 \times 64 \times 128, 128 \times 128 \times 96$, and $256 \times 256 \times 64$.

At each stage, the original global context representation is resized and projected to match the spatial and channel dimensions of the corresponding local features:

$$\tilde{F}_G^{(s)} = \mathcal{P}_S[Resize(F_G; H_s, W_s)], \tag{7}$$

where H_s and W_s are the spatial dimensions of stages, $Resize(.)$ denotes spatial resizing, $\mathcal{P}_S(.)$ denotes the stage-specific channel projection, and $\tilde{F}_G^{(s)}$ is the resized and projected global representation.

As illustrated in Figure 3b, Stage Global-Local Gate (SGLG) jointly considers the global and local features to estimate an adaptive gate:

$$G^{(s)} = \sigma\left\{\mathcal{G}_s\left[F_L^{(s)} \parallel \tilde{F}_G^{(s)}\right]\right\}, \tag{8}$$

where $\parallel$ denotes channel-wise concatenation, $\mathcal{G}_s(.)$ is the learnable gating transformation at stage s, $\sigma(.)$ denotes the sigmoid activation, and $G^{(s)}$ is the resulting global–local gate, with values bounded between 0 and 1.

The gate controls the complementary contributions of the global and local representations:

$$F_{fuse}^{(s)} = G^{(s)} \odot \tilde{F}_G^{(s)} + \left(1 - G^{(s)}\right) \odot F_L^{(s)}, \tag{9}$$

where $\odot$ denotes element-wise multiplication and $F_{fuse}^{(s)}$ represents the fused features. Thus, larger values of $G^{(s)}$ favor the global context, whereas smaller values favor the reconstructed local information.

The fused representation is subsequently refined through a lightweight residual operation:

$$F_{out}^{(s)} = F_{fuse}^{(s)} + \mathcal{R}_s\left(F_{fuse}^{(s)}\right), \tag{10}$$

where $\mathcal{R}_s(.)$ denotes the stage-specific refinement operation and $F_{out}^{(s)}$ is the final output of stage s, which is forwarded to the next reconstruction stage.

After the fourth stage, the full-resolution representation is converted into pixel-wise class logits using the segmentation head:

$$Z = Conv_{1\times1}\left(F_{out}^{(4)}\right), \quad Z \in \mathbb{R}^{256\times256\times2}. \tag{11}$$

where $F_{out}^{(4)} \in \mathbb{R}^{256\times256\times64}$ is the final decoder representation and Z denotes the two-class segmentation logits.

Overall, the two proposed modules serve complementary functions within GL-FCR: GDC constructs an adaptively weighted contextual representation from the encoder features, whereas SGLG controls its contribution relative to the progressively reconstructed local features at each decoder stage. This enables the decoder to retain broad contextual information while recovering the spatial detail required for accurate flood delineation.

2.3. Model Configurations and Experimental Setup

In addition to the proposed optical configuration described in Section 2.2, a series of SAR-only and optical–SAR configurations were investigated to assess the contribution of different sensor modalities and feature representations to flood segmentation. The principal optical configuration employed Prithvi-EO-2.0-300M, a geospatial foundation model pretrained on large-scale Harmonized Landsat and Sentinel-2 observations [17], as the encoder of the proposed GL-FCR framework. All six Sentinel-2 bands considered in this

study (B2, B3, B4, B8A, B11, and B12) were provided to this configuration. For the architectural ablation study, a Prithvi baseline was implemented using the same pretrained encoder coupled with a lightweight decoder, referred to here as the native MLP decoder. This decoder projects the final 1024-channel encoder representation to 256 channels through a 1 × 1 convolution followed by batch normalization and GELU activation. Two subsequent 3 × 3 convolutional refinement layers, incorporating batch normalization, GELU activation, and dropout, are followed by a 1 × 1 classification layer producing two-class segmentation logits. Bilinear interpolation restores the predictions to the input spatial resolution. Unlike GL-FCR, this baseline uses only the final encoder representation and does not incorporate multi-scale reconstruction, directional context, or global–local feature fusion. The pretrained Prithvi encoder was fine-tuned jointly with the decoder using AdamW, with learning rates of 1×10^{-6} and 1×10^{-4} for the encoder and decoder, respectively. The segmentation objective combined cross-entropy and Dice loss with equal weights, and checkpoint selection was based on validation flood-class F1 score. SegFormer with the MiT-B2 backbone was additionally employed as a fully supervised transformer-based baseline [26]. The SegFormer backbone was initialized using the pretrained NVIDIA MiT-B2 checkpoint (nvidia/mit-b2). To accommodate the six Sentinel-2 spectral bands, the first patch-embedding projection was expanded from three to six input channels. The pretrained blue, green, and red kernels were assigned to B2, B3, and B4, respectively, while the additional B8A, B11, and B12 channels were initialized using the mean of the three pretrained RGB kernels. All other shape-compatible pretrained weights were retained, and the resulting six-channel SegFormer model was fine-tuned for binary flood segmentation. For the independent WorldFloods evaluation, three optical configurations were compared: Prithvi-EO-2.0 with the proposed GL-FCR decoder, Prithvi-EO-2.0 with its native MLP decoder, and SegFormer-B2. All three models were evaluated on the same 509 scenes using their SEN1Floods11-trained checkpoints, without additional training or adaptation. Six-band variant (B2, B3, B4, B8A, B11, and B12) were evaluated, with the latter also serving as the optical branch in the corresponding optical–SAR experiments.

For the SAR experiments, Sentinel-1 VV and VH observations were evaluated using two complementary feature extractors. The first was a lightweight convolutional neural network (CNN) designed to progressively extract spatial features from the two-channel SAR input [27]. The second employed TerraMind-v1-B, a multimodal Earth observation foundation model pretrained across multiple geospatial modalities, including Sentinel-1 SAR observations [22]. For the latter configuration, TerraMind-B was coupled with the proposed GL-FCR decoder, allowing the same contextual reconstruction strategy used with Prithvi to be evaluated with SAR foundation-model representations.

To investigate whether SAR information provides complementary information to optical observations, several dual-branch configurations were additionally evaluated. In the Prithvi–CNN configuration, the optical branch consisted of Prithvi-EO-2.0-300M with GL-FCR, whereas Sentinel-1 VV/VH observations were processed by the SAR CNN branch. The two branches were independently decoded before their segmentation logits were combined using a learned residual decision-level fusion mechanism. Specifically, the optical and SAR logits were concatenated and passed through a convolutional correction module, whose output was scaled by a learnable sigmoid-constrained coefficient and added to the optical logits before the final prediction. In particular, late-fusion and feature-level variants were examined to assess whether the radar representation could complement the strong optical representation. For the SegFormer-B2 + SAR-CNN feature-level configuration, the final optical and SAR feature maps were projected into a common 256-dimensional space and combined using bidirectional eight-head cross-attention. A learned sigmoid gate integrated the cross-modal representations, followed by spatial refinement and residual injection into the optical features before segmentation decoding. A

second dual-foundation configuration replaced the CNN SAR branch with TerraMind-B coupled with GL-FCR, resulting in a Prithvi optical branch and a TerraMind SAR branch, each retaining its own decoder before fusion. This configuration employed the same residual decision-level fusion strategy, combining the two branches segmentation logits through a learned convolutional correction. These configurations were used as complementary experiments rather than modifications to the principal GL-FCR architecture, allowing the contribution of SAR information to be evaluated independently of the proposed optical model.

For each model configuration, the best-performing checkpoint was selected exclusively according to the highest flood-class F1 score on the validation set. The standard SEN1Floods11 test set and geographically held-out Bolivia subset were not used for model training, optimization, or checkpoint selection. The ablation experiments subsequently examined the effects of input resolution (256 × 256 and 512 × 512 pixels) and the contributions of the proposed decoder components, including feature reconstruction, directional context, and global–local fusion. Performance comparisons across the standard test and Bolivia subsets were used to assess the effectiveness and geographic transferability of these configurations and informed the selection of the final GL-FCR framework. Consequently, although these subsets remained independent of model training and checkpoint selection, they were not strictly independent of the final configuration-selection process. All models were trained using the official SEN1Floods11 training and validation partitions described in Section 2.1, while the standard test and geographically held-out Bolivia subsets were reserved exclusively for evaluation. During training, data augmentation consisted of independent random horizontal and vertical flips, each applied with a probability of 0.5, followed by a random rotation selected from 0°, 90°, 180°, and 270°. Identical geometric transformations were applied to the input images and their corresponding segmentation masks to preserve pixel-level alignment. For optical–SAR configurations, the same transformations were also applied consistently across both sensor modalities. These augmentations were restricted to the training partition and were not applied during validation or testing. Training was performed for a maximum of 80 epochs with a batch size of 2 and gradient accumulation over four steps. Optimization was performed using the AdamW optimizer [28], with a weight decay of 1×10^{-4}. For configurations incorporating a pretrained foundation-model encoder, learning rates of 1×10^{-6} and 1×10^{-4} were assigned to the pretrained encoder and task-specific decoder, respectively. A linear warm-up over the first 5% of the optimization steps was followed by cosine learning-rate decay. Mixed-precision training and gradient clipping with a maximum norm of 1.0 were also employed. The segmentation objective combined pixel-wise cross-entropy (CE) and Dice loss [29] with equal contributions. The cross-entropy loss was defined as:

$$l_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\, log\left(p_{i,c}\right), \tag{12}$$

where N is the number of valid pixels, $C = 2$ is the number of segmentation classes, $y_{i,c}$ denotes the ground-truth indicator for pixel i and class C, and $p_{i,c}$ is the corresponding predicted probability. Pixels assigned the ignore value of 255 were excluded from N. The Dice loss was formulated as:

$$l_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_{i,1} y_{i,1} + \varepsilon}{\sum_{i=1}^{N} p_{i,1} + \sum_{i=1}^{N} y_{i,1} + \varepsilon}, \tag{13}$$

where $p_{i,1}$ and $y_{i,1}$ represent the predicted probability and reference value for the flood class, respectively, and ε is a small constant introduced for numerical stability. The resulting segmentation loss was

$$l_{seg} = 0.5 l_{CE} + 0.5 l_{Dice}. \tag{14}$$

All experiments were implemented in PyTorch 2.14.1 and conducted on an NVIDIA GeForce RTX 4090 GPU. The original model comparisons and architectural ablation experiments were performed using a fixed random seed of 42. To assess the reproducibility of the two principal optical configurations, Prithvi + GL-FCR and SegFormer-B2 were additionally trained using random seeds of 123 and 2026, resulting in three independent training runs per model (42, 123, and 2026). The same dataset partitions and model-specific training protocols were maintained across the corresponding runs. For each run, the best-performing checkpoint was selected based on the highest flood-class F1 score on the validation set. The multi-seed results are reported as mean ± standard deviation in Section 3.5.

3. Results

This section presents the quantitative and qualitative evaluation of the investigated flood-mapping models. Results are first reported on the SEN1Floods11 test set and on the geographically independent Bolivia subset, considering optical, SAR, and combined optical–SAR inputs. The Bolivia subset is reported separately to assess performance on geographically held-out imagery. Finally, zero-shot cross-dataset evaluation is performed on all 509 scenes of the WorldFloods v2 dataset, using Prithvi + GL-FCR, the Prithvi native decoder, and SegFormer-B2, without additional training or fine-tuning. Flood-class F1 score, flood Intersection over Union (IoU), and mean Intersection over Union (mIoU) are reported where available. In Tables 1–3, results reproduced from previous studies are presented alongside those obtained from models trained and evaluated in this work. These two groups are distinguished by their references and the "Ours" designation, respectively. Since preprocessing procedures, training configurations, and evaluation protocols may differ across studies, the published results are considered contextual benchmarks rather than strictly controlled comparisons, and small numerical differences should be interpreted cautiously.

Table 1. Quantitative comparison with state-of-the-art optical flood segmentation methods on the SEN1Floods11 standard test set and geographically held-out Bolivia subset.

Input	Ref.	Model	Test			Bolivia		
			Flood F1 (%)	Flood IoU (%)	mIoU (%)	Flood F1 (%)	Flood IoU (%)	mIoU (%)
Optical	[30]	U-Net	—	**84.03**	**90.80**	—	70.57	82.54
		Prithvi	—	81.98	89.59	—	76.62	86.02
		SegFormer-B0	—	81.57	89.36	—	71.46	83.05
		SegFormer-B5	—	81.89	89.54	—	75.82	85.59
	[31]	DeepLabV3+	—	79.02	—	—	—	—
	[17]	Prithvi-EO-2.0-300M	—	82.20	89.70	—	—	—
		Prithvi-EO-2.0-600M	—	83.10	90.30	—	—	—
	[22]	TerraMindv1-B	—	82.26	89.70	—	—	—
	[19]	DOFA	—	76.83	86.58	—	—	—
	[23]	S12-DINO	—	80.25	88.61	—	—	—
		U-Prithvi	—	82.22	89.73	—	79.68	87.70
		Prithvi-CAFE	—	83.41	90.50	—	<u>81.37</u>	<u>88.87</u>
	Ours	Prithvi + GL-FCR	**91.08**	<u>83.62</u>	<u>90.54</u>	**90.02**	**81.83**	**89.08**
		TerraMind B + GL-FCR	89.71	81.33	89.24	86.59	76.35	85.78
		SegFormer-B2	<u>90.93</u>	83.37	90.43	<u>89.53</u>	81.04	88.64

The **Bold** indicates the best, the <u>underlined</u> indicates the second best, and the (_) indicates not reported.

Table 2. Quantitative comparison with state-of-the-art SAR-based flood segmentation methods on the SEN1Floods11 standard test set and geographically held-out Bolivia subset.

Input	Ref	Model	Test					Bolivia
			Flood F1 (%)	Flood IoU (%)	mIoU (%)	Flood F1 (%)	Flood IoU (%)	mIoU (%)
SAR	[32]	FM-Mamba	—	65.63	80.07	—	—	—
	[31]	DeepLabV3+ supervised	—	67.63	—	—	—	—
		Cross-modal distillation	—	**72.74**	—	—	—	—
	[27]	Attentive U-Net	80.20	67.20	—	—	—	—
		Fusion Network (S1 + DEM + PW)	**82.00**	<u>69.50</u>	—	—	—	—
	[22]	TerraMindv1-B	—	68.00	<u>81.06</u>	—	—	—
	Ours	TerraMind B + GL-FCR	<u>81.53</u>	68.81	**81.84**	**82.28**	**69.89**	**81.27**
		CNN	77.64	63.46	78.76	<u>78.02</u>	<u>63.96</u>	<u>78.23</u>

The **Bold** indicates the best, the <u>underlined</u> indicates the second best, and the (_) indicates not reported.

Table 3. Quantitative comparison with state-of-the-art optical–SAR flood segmentation methods on the SEN1Floods11 standard test set and geographically held-out Bolivia subset.

Input	Ref.	Model	Test					Bolivia
			Flood F1 (%)	Flood IoU (%)	mIoU (%)	Flood F1 (%)	Flood IoU (%)	mIoU (%)
Optical+ SAR	[31]	DeepLabV3+	—	79.25	—	—	—	—
	[22]	TerraMindv1-B			**90.62**			
	Ours	SegFormer-B2+ SAR-CNN single-level cross-modal fusion	**90.98**	**83.45**	<u>90.46</u>	<u>89.67</u>	<u>81.28</u>	<u>88.77</u>
		TerraMindv1_B _Native Multimodal_GL_FCR	89.25	80.58	88.81	86.79	76.67	85.97
		Prithvi_GLFCR_TerraMindB_S1 GLFCR_DualDecoder_LateFusion	90.58	82.78	90.08	89.59	81.14	88.69
		Prithvi_GLFCR_SAR_CNN_ DualDecoder_LateFusion	<u>90.86</u>	<u>83.26</u>	90.37	**90.05**	**81.91**	**89.15**

The **Bold** indicates the best, the <u>underlined</u> indicates the second best, and the (_) indicates not reported.

The qualitative examples presented in Figures 4–10 were manually selected from the corresponding evaluation subsets to illustrate different flood extents, spatial patterns, and segmentation outcomes, including both successful predictions and challenging cases. These examples are intended for qualitative interpretation rather than as a statistically representative sample; quantitative performance is assessed using the complete corresponding evaluation subsets.

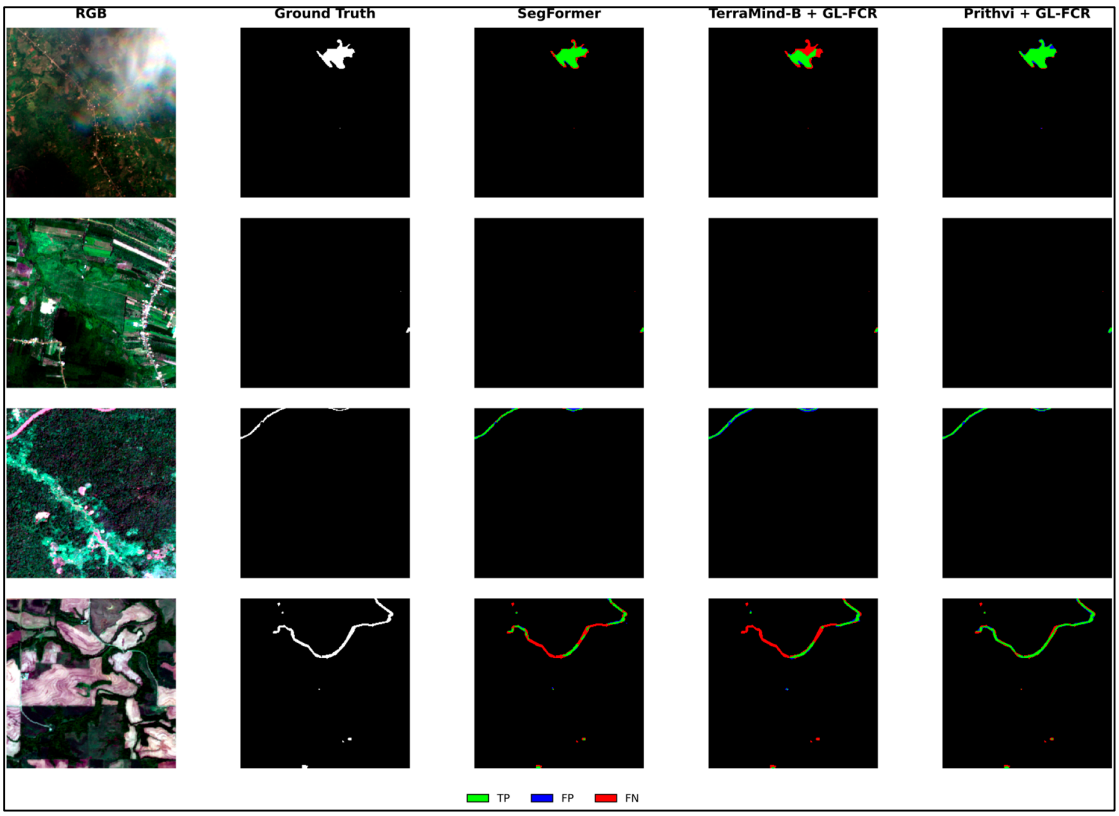


Figure 4. Qualitative comparison of optical flood segmentation on the SEN1Floods11 standard test set.

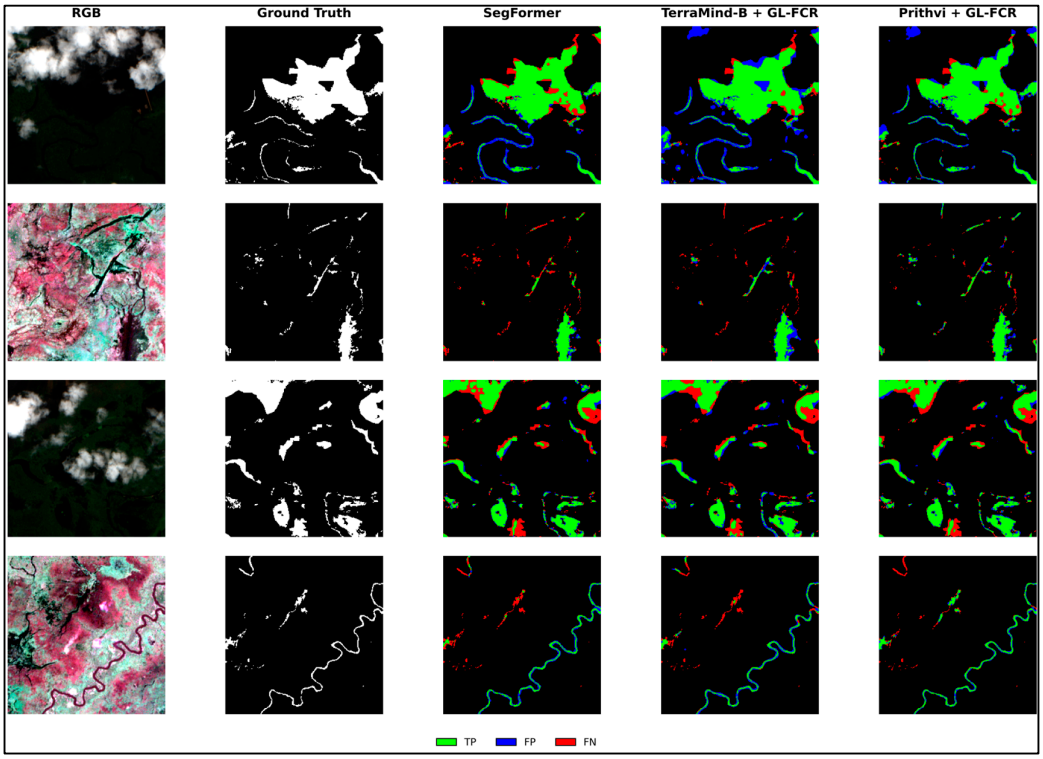


Figure 5. Qualitative comparison of optical flood segmentation on the geographically held-out Bolivia subset.

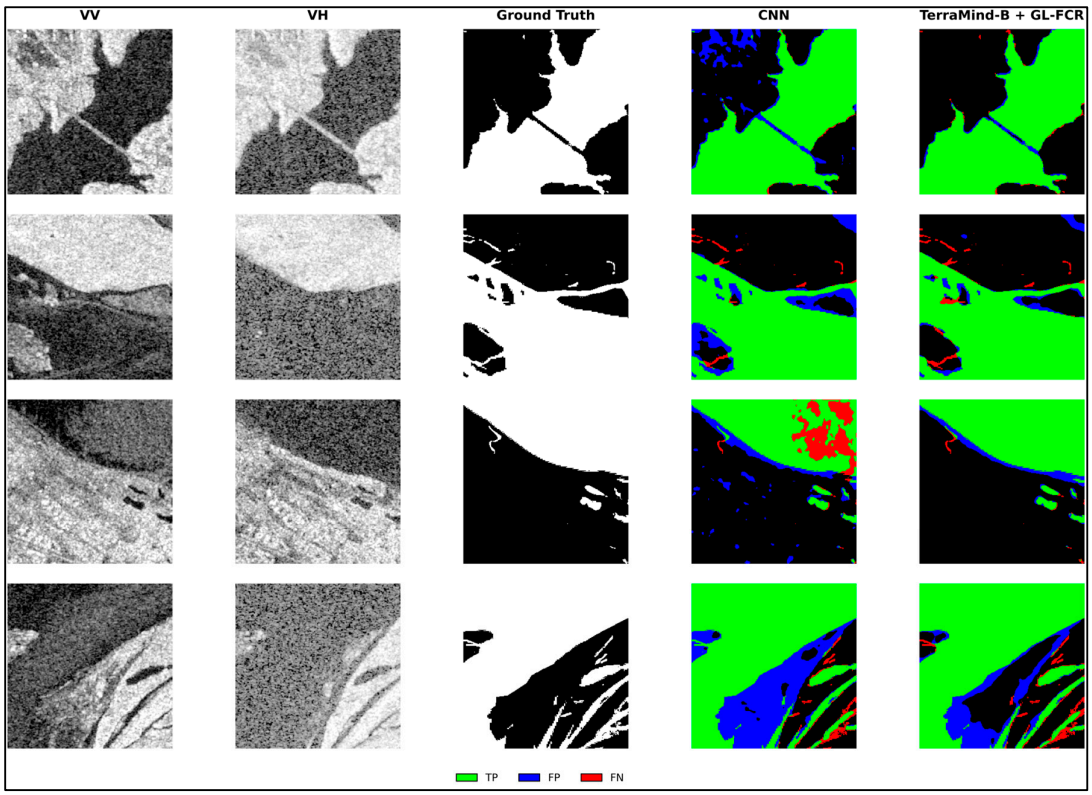


Figure 6. Qualitative comparison of SAR-based flood segmentation on the SEN1Floods11 standard test set.

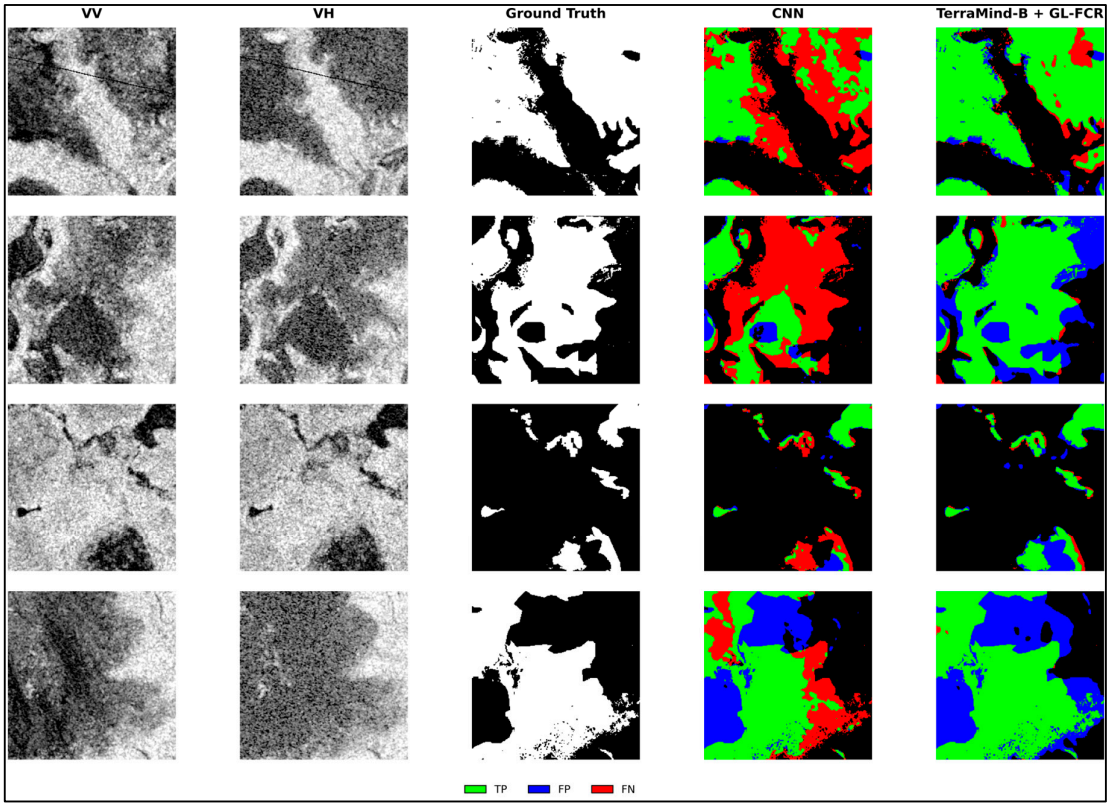


Figure 7. Qualitative comparison of SAR-based flood segmentation on the geographically held-out Bolivia subset.

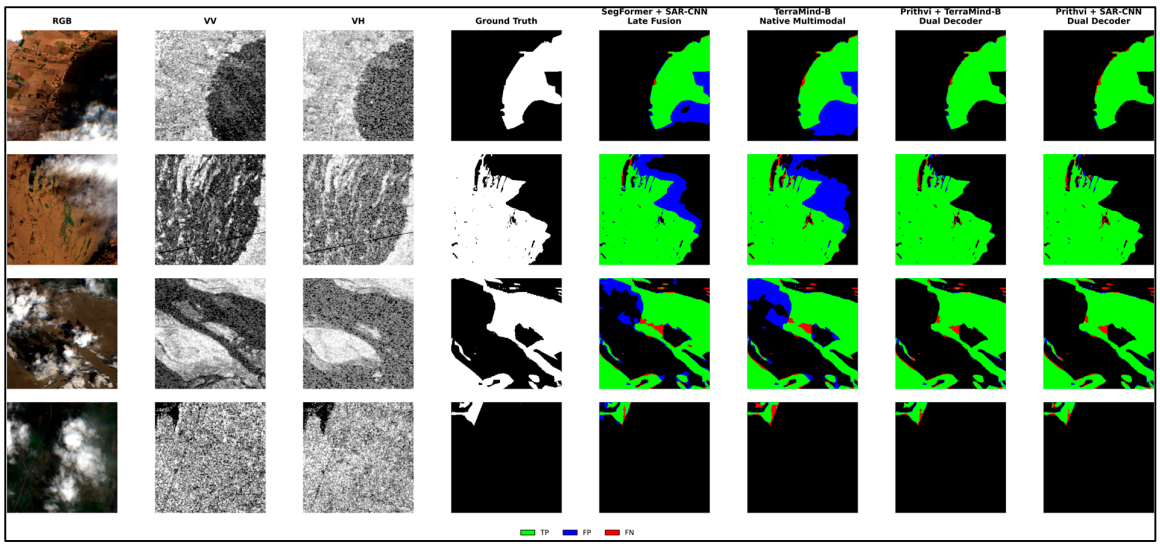


Figure 8. Qualitative comparison of optical–SAR flood segmentation on the SEN1Floods11 standard test set.

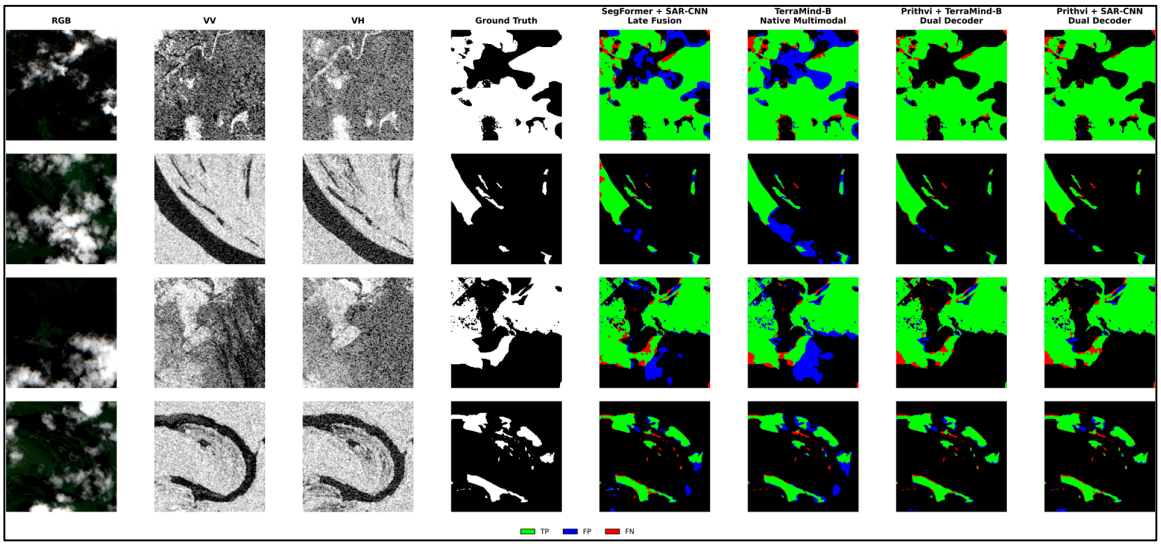


Figure 9. Qualitative comparison of optical–SAR flood segmentation on the geographically held-out Bolivia subset.

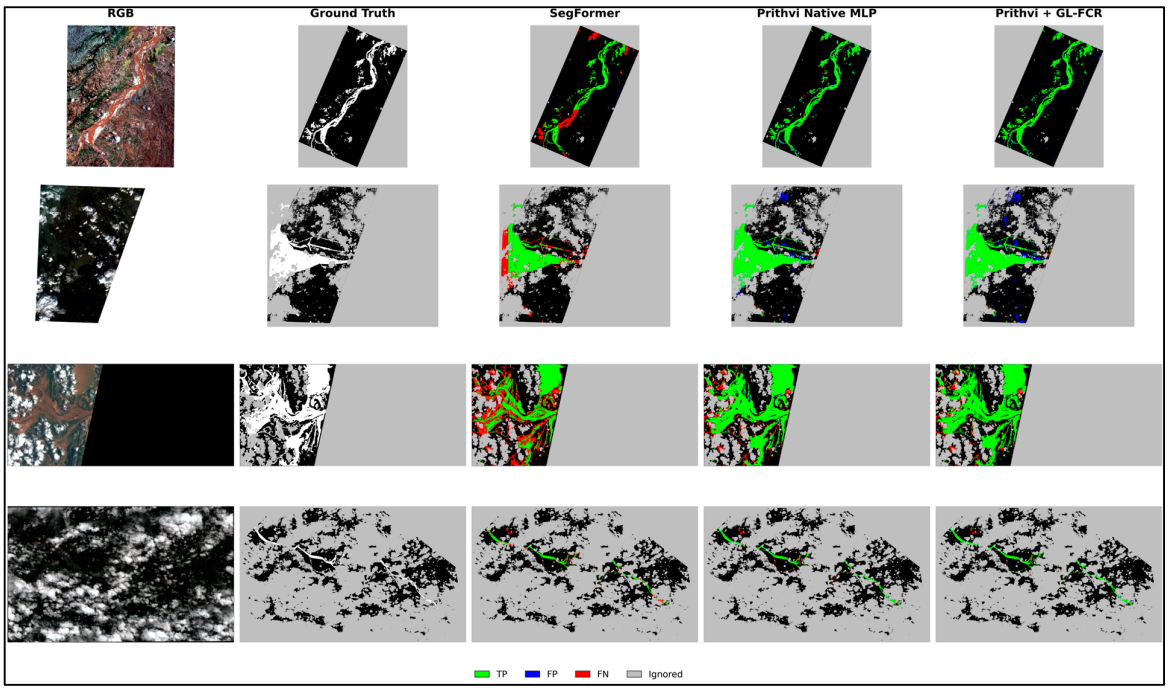


Figure 10. Qualitative zero-shot cross-dataset comparison on WorldFloods.

3.1. Optical Inputs

The optical experiments evaluate the proposed framework under two complementary settings: the standard SEN1Floods11 test set and the geographically independent Bolivia subset. Table 1 summarizes the quantitative results together with previously reported methods.

3.1.1. SEN1Floods11 Test Set

On the standard SEN1Floods11 test set, Table 1 shows that several existing approaches achieve flood IoU values above 80%. Among the reported methods, A Unet records the highest test-set result. The foundation-model-based methods also show strong performance: Prithvi-EO-2.0, TerraMind, U-Prithvi, and Prithvi-CAFE obtain flood IoUs in the low-80% range, with Prithvi-CAFE reaching 83.41%. These results provide the reference against which the models implemented in this study are evaluated.

In the original experiments conducted using random seed 42, Prithvi + GL-FCR achieved a flood F1 of 91.08%, flood IoU of 83.62%, and mIoU of 90.54%. SegFormer-B2 produced closely comparable values of 90.93%, 83.37%, and 90.43%, respectively, while TerraMind-B + GL-FCR obtained 89.71% flood F1, 81.33% flood IoU, and 89.24% mIoU. Thus, Prithvi + GL-FCR achieved the highest test-set scores among the three optical configurations in the original single-seed evaluation. However, the multi-seed analysis presented in Section 3.5 revealed comparable average performance between Prithvi + GL-FCR and SegFormer-B2, with SegFormer-B2 achieving a slightly higher mean flood IoU (83.36% versus 83.07%) across three random seeds.

The corresponding qualitative results are presented in Figure 4. The selected samples cover different flood geometries, ranging from compact inundated regions to narrow and fragmented water features. The error maps represent correctly detected flood pixels as true positives (green), overestimated flood regions as false positives (blue), and missed flood pixels as false negatives (red). The spatial patterns in Figure 4 complement the quantitative results in Table 1 by showing the distribution of detection and delineation errors produced by the three evaluated optical models.

3.1.2. SEN1Floods11 Bolivia External Set

The geographically independent Bolivia subset provides an additional evaluation of the optical models outside the geographic regions represented by the standard SEN1Floods11 training data. As shown in Table 1, fewer published approaches provide results for this subset. Among the available literature results, the reported flood IoUs include 70.57% for U-Net, 76.62% for Prithvi, 71.46% for SegFormer-B0, 75.82% for SegFormer-B5, 79.68% for U-Prithvi, and 81.37% for Prithvi-CAFE.

For the models evaluated in this study using random seed 42, Prithvi + GL-FCR achieved a flood F1 of 90.02%, flood IoU of 81.83%, and mIoU of 89.08%. SegFormer-B2 reached 89.53%, 81.04%, and 88.64%, respectively, whereas TerraMind-B + GL-FCR obtained 86.59% flood F1, 76.35% flood IoU, and 85.78% mIoU. Prithvi + GL-FCR therefore produced the highest flood IoU among the methods reported in Table 1 for the Bolivia subset in the original single-seed comparison, while also recording the highest F1 and mIoU among our evaluated optical configurations. The multi-seed evaluation presented in Section 3.5 further showed that Prithvi + GL-FCR maintained a higher mean flood IoU than SegFormer-B2 across three random seeds (81.53 ± 0.27% versus 80.68 ± 0.46%). This result indicates a modest advantage for GL-FCR on the Bolivia subset, although the limited evaluation sample size and number of training runs warrant cautious interpretation.

The qualitative results for this geographically held-out evaluation are shown in Figure 5. Compared with the test examples in Figure 4, the selected Bolivia scenes contain more extensive and spatially heterogeneous flood patterns. The TP, FP, and FN maps show the correspondence between the predicted and reference flood regions and reveal differences among the models along complex boundaries, fragmented inundated areas, and narrow water structures. These spatial results complement the Bolivia metrics reported in Table 1.

3.2. SAR Inputs for Early Flood Mapping

SAR imagery was evaluated as an alternative input source for flood mapping, with particular relevance to early mapping conditions in which optical observations may be limited by cloud cover or illumination conditions. The SAR results on the standard SEN1Floods11 test set and the geographically independent Bolivia subset are reported in Table 2, together with previously published SAR-based approaches. The corresponding qualitative results are presented in Figures 6 and 7.

3.2.1. SEN1Floods11 Test Set

The SAR-based methods reported in Table 2 exhibit a wider range of flood IoU values than the optical approaches. Among the previous studies included in the comparison, FM-Mamba reports a flood IoU of 65.63%, while the supervised DeepLabV3+ model reaches 67.63%. The cross-modal distillation approach achieves 72.74%, whereas the Attentive U-Net and the S1–DEM–permanent-water Fusion Network report 67.20% and 69.50%, respectively. The foundation-model-based TerraMind-v1-B reports a flood IoU of 68.00%.

For the SAR configurations evaluated in this study, TerraMind-B + GL-FCR achieved a flood F1 of 81.53%, flood IoU of 68.81%, and mIoU of 81.84% on the standard test set. The SAR-CNN baseline obtained 77.64% flood F1, 63.46% flood IoU, and 78.76% mIoU. TerraMind-B + GL-FCR therefore yielded the highest values across all three metrics between the two SAR-only configurations evaluated in this work.

The qualitative examples in Figure 6 further show the differences between the two SAR-based configurations. Across the selected test scenes, TerraMind-B + GL-FCR generally provides a more complete delineation of the flooded areas than the SAR-CNN, with a larger proportion of the reference flood extent represented by true-positive predictions. This difference is particularly visible in extended and spatially connected inundated

regions, where TerraMind-B + GL-FCR recovers a greater portion of the flood extent and exhibits fewer false-negative areas. The SAR-CNN predictions show comparatively more missed flood pixels and, in some examples, a less continuous representation of the inundated regions. False-positive detections remain visible for both approaches in areas presenting SAR responses similar to those of flooded surfaces. Overall, the qualitative patterns observed in Figure 6 are consistent with the higher flood F1 and IoU obtained by TerraMind-B + GL-FCR on the test set.

3.2.2. SEN1Floods11 Bolivia External Set

The geographically held-out Bolivia subset was additionally used to evaluate the SAR models under geographic transfer. Since corresponding Bolivia results are not available for the previously published SAR approaches included in Table 2, the quantitative comparison on this subset is restricted to the SAR models evaluated in the present study.

On the Bolivia subset, TerraMind-B + GL-FCR achieved a flood F1 of 82.28%, flood IoU of 69.89%, and mIoU of 81.27%. The SAR-CNN obtained 78.02% flood F1, 63.96% flood IoU, and 78.23% mIoU. TerraMind-B + GL-FCR consequently maintained higher values across all three metrics on the geographically held-out data.

This difference is also evident in the qualitative examples presented in Figure 7. TerraMind-B + GL-FCR detects a larger proportion of the reference flooded areas and more consistently preserves their spatial extent, whereas the SAR-CNN exhibits more extensive false-negative regions within several inundated areas. The difference is especially apparent for flood regions characterized by irregular boundaries and heterogeneous SAR responses, where TerraMind-B + GL-FCR produces more spatially complete predictions. In several examples, the TerraMind-based model also maintains greater continuity across the flooded areas, while the SAR-CNN predictions are more fragmented or omit portions of the reference extent. Nevertheless, both approaches retain some false-positive and false-negative regions, reflecting the remaining errors visible in the SAR-only predictions. These qualitative observations correspond to the higher flood F1, flood IoU, and mIoU obtained by TerraMind-B + GL-FCR on the Bolivia subset.

3.3. Multimodal Optical and SAR Inputs

The multimodal experiments combine optical Sentinel-2 observations with Sentinel-1 SAR information to evaluate whether complementary information from the two sensing modalities improves flood delineation. The quantitative results for the standard SEN1Floods11 test set and the geographically held-out Bolivia subset are summarized in Table 3, together with previously reported multimodal approaches. The evaluated configurations include a SegFormer-B2 model combined with a SAR-CNN through cross-modal fusion, the native multimodal TerraMind configuration, and two dual-decoder architectures coupling Prithvi-based optical features with either a TerraMind-B or CNN-based SAR branch. Representative predictions for the test and Bolivia subsets are presented in Figure 8 and Figure 9, respectively.

3.3.1. SEN1Floods11 Test Set

Among the previously published multimodal approaches reported in Table 3, the optical–SAR DeepLabV3+ configuration reports a flood IoU of 79.25%, while the native S1 + S2 configuration of TerraMind-v1-B reports an mIoU of 90.62%. The latter result is reported under its corresponding multimodal evaluation protocol and therefore provides an additional reference for the performance of joint Sentinel-1 and Sentinel-2 representations.

For the multimodal configurations evaluated in this study, SegFormer-B2 + SAR-CNN with single-level cross-modal fusion achieved the highest test-set flood F1 and flood IoU, reaching 90.98% and 83.45%, respectively, together with an mIoU of 90.46%. The

Prithvi + GL-FCR + SAR-CNN dual-decoder late-fusion configuration obtained a flood F1 of 90.86%, flood IoU of 83.26%, and mIoU of 90.37%. The dual-foundation configuration combining Prithvi + GL-FCR for the optical branch and TerraMind-B + GL-FCR for the SAR branch reached 90.58% flood F1, 82.78% flood IoU, and 90.08% mIoU. The native multimodal TerraMind-B S1 + S2 + GL-FCR configuration obtained 89.25% flood F1, 80.58% flood IoU, and 88.81% mIoU.

The qualitative examples in Figure 8 show that the multimodal configurations successfully delineate the major flooded regions across the selected test scenes, although differences remain in the completeness and spatial continuity of the predictions. The SegFormer-B2 + SAR-CNN and Prithvi + GL-FCR + SAR-CNN configurations generally recover a larger proportion of the reference inundated areas, with relatively limited false-negative regions across the principal flood extents. The dual-foundation Prithvi–TerraMind configuration also captures the dominant flooded regions, although additional omission and commission errors are visible in some spatially heterogeneous areas. The native TerraMind multimodal predictions exhibit comparatively larger missed or fragmented flood regions in several examples. These qualitative patterns are consistent with the quantitative ranking observed for the evaluated multimodal models on the test set.

3.3.2. SEN1Floods11 Bolivia External Set

The multimodal models were subsequently evaluated on the geographically held-out Bolivia subset to assess their performance under geographic transfer. For the configurations evaluated in this study, the results in Table 3 remain closely grouped, particularly among the models incorporating the Prithvi optical representation.

The Prithvi + GL-FCR + SAR-CNN dual-decoder late-fusion configuration achieved the highest multimodal performance on Bolivia, with a flood F1 of 90.05%, flood IoU of 81.91%, and mIoU of 89.15%. The SegFormer-B2 + SAR-CNN cross-modal fusion model reached 89.67%, 81.28%, and 88.77%, respectively. The dual-foundation Prithvi + GL-FCR + TerraMind-B + GL-FCR configuration obtained a flood F1 of 89.59%, flood IoU of 81.14%, and mIoU of 88.69%, whereas the native multimodal TerraMind-B S1 + S2 + GL-FCR configuration reached 86.79%, 76.67%, and 85.97%, respectively.

The qualitative results in Figure 9 further illustrate the behavior of the multimodal configurations on the Bolivia scenes. The Prithvi + GL-FCR + SAR-CNN dual-decoder model generally provides a complete delineation of the main flooded areas, with strong spatial correspondence between the predicted and reference extents. The SegFormer-B2 + SAR-CNN configuration similarly detects most of the principal inundated regions, although differences are visible along irregular boundaries and within locally fragmented flood areas. The Prithvi–TerraMind dual-foundation configuration also preserves the overall flood geometry in the selected scenes, while the native TerraMind multimodal model exhibits more visible false-negative regions in some examples. Across the four configurations, the remaining errors are concentrated mainly around complex flood boundaries and spatially heterogeneous regions, consistent with the relatively close quantitative performance of the leading multimodal models reported in Table 3.

3.4. External Zero-Shot Cross-Dataset Evaluation on WorldFloods

To further evaluate the cross-dataset generalization capability of the optical models, an external zero-shot experiment was conducted using the complete WorldFloods dataset comprising 509 scenes. In contrast to the geographically held-out Bolivia subset of SEN1Floods11, WorldFloods represents an independent dataset, providing a more challenging setting for assessing model transfer across datasets. For this evaluation, all 509 WorldFloods scenes were selected and processed using the spectral bands required by the optical models. SegFormer-B2, Prithvi with its native MLP decoder, and Prithvi + GL-FCR

were transferred directly from SEN1Floods11 to WorldFloods without any retraining, fine-tuning, or adaptation. Invalid and cloud-contaminated pixels in the WorldFloods reference masks were excluded from the evaluation. The scene-level and global quantitative results are reported in Table 4, with additional scene-level performance statistics across all 509 scenes presented in Table 5, while representative predictions are presented in Figure 10.

The zero-shot results on WorldFloods show a substantial difference between the two evaluated optical models. Across all 509 WorldFloods scenes, Prithvi + GL-FCR achieved a global flood F1 of 84.56%, flood IoU of 73.24%, and mIoU of 84.08%. The Prithvi native MLP decoder achieved a flood F1 of 83.35%, flood IoU of 72.15%, and mIoU of 83.23%. In comparison, SegFormer-B2 obtained a flood F1 of 57.38%, flood IoU of 40.23%, and mIoU of 65.22%. The corresponding global precision and recall were 87.27% and 82.00% for Prithvi + GL-FCR, 86.17% and 81.36% for the Prithvi native decoder, compared with 92.71% and 41.55% for SegFormer-B2.

Table 4. Quantitative zero-shot cross-dataset comparison on WorldFloods. Scene-level results are reported for the 20 selected scenes, while global metrics are computed over all 509 WorldFloods scenes.

	Flood F1 (%)			Flood IoU (%)			mIoU (%)		
Scene ID	**SegFormer-B2**	**Prithvi + Native MLP**	**Prithvi+ GL-FCR**	**SegFormer-B2**	**Prithvi + Native MLP**	**Prithvi+ GL-FCR**	**SegFormer-B2**	**Prithvi + Native MLP**	**Prithvi+ GL-FCR**
0042	82.04	90.12	**91.56**	69.55	83.13	**84.43**	80.10	89.20	**89.25**
0043	42.77	94.11	**95.03**	27.20	88.87	**90.54**	25.55	76.80	**79.33**
0046	4.26	92.54	**93.04**	2.18	86.12	**86.99**	40.68	91.09	**91.65**
0064	9.53	92.64	**93.43**	5.01	86.64	**87.67**	16.16	79.96	**80.68**
0100	18.17	90.30	**90.85**	9.99	82.99	**83.23**	43.86	88.91	**89.47**
0213	51.47	88.41	**89.43**	34.65	79.23	**80.89**	62.07	87.60	**88.61**
0216	1.64	70.50	**70.53**	0.83	54.44	**54.48**	42.04	**72.70**	72.61
0229	60.89	74.78	**75.54**	43.78	60.01	**60.69**	70.57	78.32	**79.18**
0259	74.29	74.81	**75.51**	59.10	59.68	**60.66**	76.52	**77.09**	76.06
0263	79.11	88.57	**92.73**	65.44	79.48	**86.45**	80.29	88.04	**92.13**
0322	2.34	17.37	**21.52**	1.18	9.51	**12.06**	46.08	50.49	**51.87**
0346	0.00	89.06	**90.12**	0.00	81.28	**82.02**	27.46	84.22	**84.56**
0361	50.12	78.59	**79.62**	33.44	64.73	**66.14**	59.62	78.28	**79.10**
0369	61.66	85.79	**87.15**	44.57	75.12	**77.22**	65.36	83.80	**85.13**
0380	35.34	93.50	**93.83**	21.47	87.79	**88.39**	52.38	92.29	**92.64**
0382	49.80	**74.57**	74.12	33.16	**59.70**	58.88	62.91	**76.58**	76.04
0396	71.32	87.63	**88.46**	55.42	78.58	**79.31**	71.85	85.56	**86.36**
0425	42.01	86.79	**87.28**	26.59	77.24	**77.44**	46.96	80.96	**81.35**
0426	77.64	84.88	**89.94**	63.46	73.73	**81.72**	80.81	86.20	**90.39**
0488	70.39	89.27	**91.19**	54.31	80.61	**83.81**	61.59	81.93	**84.78**
Global (509 scenes)	57.38	83.35	**84.56**	40.23	72.15	**73.24**	65.22	83.23	**84.08**

The **Bold** indicates the best.

The scene-level results in Table 4 further demonstrate the consistency of this difference across the 20 selected WorldFloods scenes. Prithvi + GL-FCR achieved a higher flood F1 and flood IoU than SegFormer-B2 in all 20 selected WorldFloods scenes. Its flood IoU exceeded 80% in 11 of the 20 scenes, whereas the performance of SegFormer-B2 varied considerably across individual scenes and reached a flood IoU of 0% for scene 0346. Considering mIoU, Prithvi + GL-FCR achieved the higher value in 19 of the 20 scenes. The

only exception was scene 0259, for which SegFormer-B2 obtained an mIoU of 76.52%, compared with 76.06% for Prithvi + GL-FCR. Compared with the Prithvi native MLP decoder, GL-FCR achieved higher flood F1 and flood IoU in 19 of the 20 selected scenes, with scene 0382 being the exception. Across all 509 WorldFloods scenes, the scene-level statistics in Table 5 show mean flood F1 scores of 74.83%, 72.87%, and 48.87% for Prithvi + GL-FCR, the Prithvi native decoder, and SegFormer-B2, respectively. The corresponding standard deviations were 20.37, 21.88, and 25.38 percentage points, highlighting substantial variability in performance across scenes.

Table 5. Scene-level performance statistics across 509 WorldFloods scenes (All values are %).

Metric	Model	Mean	Median	Std. Dev.
Flood F1	SegFormer-B2	48.87	50.37	25.38
	Prithvi Native MLP	72.87	79.73	21.88
	Prithvi + GL-FCR	**74.83**	**80.10**	**20.37**
Flood IoU	SegFormer-B2	36.17	33.66	23.22
	Prithvi Native MLP	62.61	65.68	23.77
	Prithvi + GL-FCR	**63.40**	**66.80**	**22.62**
mIoU	SegFormer-B2	60.79	61.41	15.96
	Prithvi Native MLP	76.99	78.23	12.86
	Prithvi + GL-FCR	**77.33**	**79.19**	**12.65**

The **Bold** indicates the best.

The qualitative examples from the WorldFloods dataset presented in Figure 10 visually reflect the differences observed in the quantitative evaluation. Across the selected scenes, Prithvi + GL-FCR detects a substantially larger proportion of the reference water/flood regions than SegFormer-B2, while exhibiting performance broadly comparable to the Prithvi native MLP decoder, and provides a more complete delineation of their spatial extent. The predicted regions generally remain continuous across large inundated areas, while narrow, irregular, and spatially fragmented target regions are also more consistently recovered. Compared with the native MLP decoder, GL-FCR exhibits relatively small visual differences, with localized improvements in water/flood delineation and reductions in false-positive predictions in some regions. These observations are consistent with the modest quantitative improvements reported across the complete 509-scene WorldFloods dataset.

In contrast, the SegFormer-B2 predictions contain considerably larger false-negative regions, with substantial portions of the reference water/flood extent remaining undetected in several WorldFloods examples. This visual pattern corresponds to the large difference in recall between SegFormer-B2 (41.55%) and Prithvi + GL-FCR (82.00%). Although SegFormer-B2 achieves a high precision of 92.71%, its predictions cover a substantially smaller fraction of the reference target regions. Prithvi + GL-FCR provides a more complete spatial representation of these regions while maintaining a precision of 87.27%. The Prithvi native MLP decoder exhibits a similar behavior, achieving 81.36% recall and 86.17% precision. The relatively small differences between the two Prithvi-based models indicate that GL-FCR provides an incremental improvement over the native decoder, rather than a substantial change in cross-dataset performance.

To further investigate the substantial scene-to-scene performance variations observed in Table 6, four representative WorldFloods scenes were analyzed in terms of flood coverage, spatial fragmentation, NDWI-based spectral separability, and segmentation errors (Table 6). Scene 0322 exhibited relatively low flood coverage (9.12%) and the weakest NDWI spectral separation (0.106), coinciding with severe flood omission across all three models and suggesting that limited spectral contrast may contribute to poor segmentation

performance. In contrast, scene 0346 showed higher flood coverage (45.09%) and strong spectral separation (0.653), yet SegFormer-B2 failed to detect any flood pixels, indicating that spectral contrast alone cannot explain the observed performance differences. Scene 0043, characterized by extensive flood coverage (81.26%), demonstrated substantially better flood recovery with the Prithvi-based models than with SegFormer-B2. Finally, scene 0382 illustrated an omission–commission trade-off, with GL-FCR achieving relatively high flood recall (84.74%) but lower precision (65.87%). Overall, these findings suggest that scene-dependent performance variability reflects the combined influence of flood extent, spectral characteristics, spatial distribution, and model-specific segmentation errors, rather than a single dominant factor.

Table 6. Scene-level flood characteristics and spectral separability across four representative WorldFloods scenes.

WorldFloods Scene	Flood Coverage (%)	Flood Fragmentation (Components/km²)	NDWI Spectral Separation
0043	81.26	6.39	0.350
0322	9.12	3.63	0.106
0346	45.09	0.22	0.653
0382	10.48	2.00	0.598

3.5. Ablation Study and Multi-Seed Results

An ablation study was conducted on the optical configuration to evaluate the effects of input patch size and the progressive decoder components leading to the proposed GL-FCR architecture. The ablation experiments were performed using random seed 42 and evaluated on both the standard SEN1Floods11 test set and the geographically held-out Bolivia subset, with flood IoU used as the primary evaluation metric. As summarized in Table 7, the experiments start from the Prithvi baseline and progressively incorporate multi-scale decoding, feature context, directional context, and global–local feature fusion. Additionally, to assess training variability and the reproducibility of the model comparison, Prithvi + GL-FCR and SegFormer-B2 were evaluated across three random seeds (42, 123, and 2026). The corresponding multi-seed results are presented in Table 6 as mean ± standard deviation.

Table 7. Ablation study of the proposed GL-FCR components on the SEN1Floods11 standard test set and geographically held-out Bolivia subset.

Model	Dataset	Patch Size	Multi-Scale Decoding	Feature Context	Directional Context	Global–Local Fusion	PixelShuffle	Pre Flood (%)	Rec (%)	F1 Flood (%)	Flood IoU (%)	mIoU (%)
Prithvi baseline	Test	256	—	—	—	—	—	88.11	84.90	86.48	76.20	86.23
	Bolivia	256	—	—	—	—	—	88.63	85.63	87.10	77.15	86.25
Prithvi baseline	Test	512	—	—	—	—	—	88.16	86.92	87.54	77.84	87.18
	Bolivia	512	—	—	—	—	—	87.51	83.89	85.66	74.92	84.89
Prithvi + MD-FCR	Test	256	✓	✓	—	—	—	**91.91**	86.73	89.24	80.58	88.82
	Bolivia	256	✓	✓	—	—	—	88.61	88.95	88.78	79.82	87.84
Prithvi + FCR	Test	256	✓	✓	✓	—	—	91.93	88.62	90.25	82.21	89.77
	Bolivia	256	✓	✓	✓	—	—	89.61	88.46	89.03	80.22	88.10
Prithvi + GL-FCR (without PixelShuffle)	Test	256	✓	✓	✓	✓	—	91.29	87.95	89.97	81.63	89.01
	Bolivia	256	✓	✓	✓	✓	—	87.89	88.27	88.14	80.41	87.59

Prithvi + GL-FCR	Test	256	✓	✓	✓	✓	✓	91.84	**90.34**	**91.08**	**83.62**	**90.13**
	Bolivia	256	✓	✓	✓	✓	✓	**90.74**	**89.28**	**90.00**	**81.83**	**89.07**

The **Bold** indicates the best, ✓ indicates that the corresponding component is included.

The Prithvi baseline in Table 7 corresponds to the native MLP decoder implementation described in Section 2.3, trained and evaluated within the experimental framework of this study. At 256 × 256 pixels, this configuration achieved a Flood IoU of 76.20% on the standard test set, compared with approximately 82% reported for Prithvi-based approaches in Table 1. However, these published results correspond to independently reported implementations rather than exact reproductions under our experimental protocol. Differences in decoder architecture, training configurations, preprocessing, and evaluation procedures may contribute to this discrepancy. Therefore, the internal Prithvi baseline serves as a reference for assessing the progressive contributions of the GL-FCR decoder components rather than as a direct reproduction of previously published Prithvi results.

The effect of the input patch size was first examined using the Prithvi baseline without the additional decoder components. Increasing the patch size from 256 × 256 to 512 × 512 increased the flood IoU on the test set from 76.20% to 77.84%. However, the same change reduced the flood IoU on the Bolivia subset from 77.15% to 74.92%. This difference was also reflected in flood recall, which increased from 84.90% to 86.92% on the test set but decreased from 85.63% to 83.89% on Bolivia. The subsequent decoder experiments were therefore conducted using the 256 × 256 patch configuration.

Starting from this configuration, the introduction of multi-scale decoding and feature context in Prithvi + MD-FCR increased the flood IoU from 76.20% to 80.58% on the test set and from 77.15% to 79.82% on Bolivia. The subsequent FCR configuration, which additionally incorporates directional context, further increased the flood IoU to 82.21% on the test set and 80.22% on the Bolivia subset. These improvements were accompanied by increases in flood F1-score from 86.48% for the baseline to 89.24% for MD-FCR and 90.25% for FCR on the test set, while mIoU increased from 86.23% to 88.82% and 89.77%, respectively. As multi-scale decoding and feature context were introduced together, their individual contributions cannot be fully separated through this comparison.

The complete GL-FCR configuration, incorporating multi-scale decoding, feature context, directional context, global–local fusion, and PixelShuffle upsampling, achieved the highest flood IoU in the ablation study, reaching 83.62% on the SEN1Floods11 test set and 81.83% on Bolivia. Relative to the 256 × 256 Prithvi baseline, this corresponds to absolute increases of 7.42 percentage points and 4.68 percentage points, respectively. The full configuration also improved flood recall from 84.90% to 90.34% on the test set and from 85.63% to 89.28% on Bolivia, while increasing flood precision from 88.11% to 91.84% and from 88.63% to 90.74%, respectively. These results indicate reduced flood omission without increased commission errors relative to the baseline. Compared with FCR, the complete GL-FCR configuration increased flood IoU by a further 1.41 percentage points on the test set and 1.60 percentage points on Bolivia. A controlled ablation replacing PixelShuffle upsampling reduced flood IoU to 81.63% on the test set and 80.41% on Bolivia, with corresponding reductions in flood F1-score from 91.08% to 89.97% and from 90.00% to 88.14%, respectively. Overall, the progressive results in Table 7 demonstrate improvements across multiple segmentation metrics, culminating in the full GL-FCR configuration on both evaluation sets.

To assess the stability of the two principal optical configurations across different random initializations, Prithvi + GL-FCR and SegFormer-B2 were evaluated over three independent training runs using seeds 42, 123, and 2026. As reported in Table 8, both models demonstrated comparable performance on the standard SEN1Floods11 test set. Prithvi + GL-FCR achieved a mean flood IoU of 83.07 ± 0.50% and flood F1 of 90.75 ± 0.30%, while

SegFormer-B2 obtained 83.36 ± 0.34% and 90.93 ± 0.20%, respectively. On the geographically held-out Bolivia subset, comprising only 15 scenes, Prithvi + GL-FCR achieved a higher mean flood IoU (81.53 ± 0.27% versus 80.68 ± 0.46%) and flood F1 (89.83 ± 0.16% versus 89.31 ± 0.28%). The reported standard deviations quantify variability across the three independent training runs rather than variability among individual Bolivia scenes. These results indicate relatively limited performance variability across the evaluated seeds and suggest that the proposed GL-FCR architecture maintains competitive performance on the standard test set while exhibiting a modest advantage on the Bolivia subset. Nevertheless, given the small number of training runs and the limited size of the Bolivia evaluation subset, these differences should be interpreted cautiously rather than as evidence of statistically significant superiority.

Table 8. Multi-seed performance comparison of Prithvi + GL-FCR and SegFormer-B2 on the SEN1Floods11 standard test set and geographically held-out Bolivia subset. Values are mean ± sample standard deviation (%) across three training seeds (42, 123, and 2026).

Dataset	Model	Pre (%)	Rec (%)	Flood F1 (%)	Flood IoU (%)	mIoU (%)
Test	Prithvi + GL-FCR	92.22 ± 0.56	**89.34** ± 1.04	90.75 ± 0.30	83.07 ± 0.50	90.25 ± 0.28
Test	SegFormer-B2	**92.81** ± 0.23	89.12 ± 0.29	**90.93** ± 0.20	**83.36** ± 0.34	**90.43** ± 0.20
Bolivia	Prithvi + GL-FCR	90.24 ± 0.83	**89.42** ± 0.81	**89.83** ± 0.16	**81.53** ± 0.27	**88.89** ± 0.17
Bolivia	SegFormer-B2	**92.16** ± 0.74	86.63 ± 0.60	89.31 ± 0.28	80.68 ± 0.46	88.43 ± 0.28

The **Bold** indicates the best.

3.6. Computational Complexity

To assess the computational requirements of the investigated architectures, we compared the total parameter count, inference latency, and peak GPU memory allocation of SegFormer-B2, Prithvi-EO-2.0-300M with its native MLP decoder, and Prithvi-EO-2.0-300M with the proposed GL-FCR decoder. All measurements were performed on an NVIDIA GeForce RTX 4090 using FP32 precision, a batch size of one, and six-band input patches of 256 × 256 pixels. Inference latency was measured over 100 iterations following 20 warm-up iterations.

As shown in Table 9, SegFormer-B2 has the smallest parameter count (27.36 M) and GPU memory allocation (0.22 GB), whereas Prithvi with its native MLP decoder achieves the lowest measured inference latency (14.79 ms per patch). Replacing the native decoder with GL-FCR increases the parameter count from 305.33 M to 308.34 M, corresponding to an increase of approximately 0.99%. However, inference latency increases from 14.79 to 23.27 ms per patch, while peak GPU memory allocation rises from 1.18 to 1.30 GB. These results indicate that GL-FCR introduces a relatively small parameter overhead but a more noticeable runtime cost, highlighting the trade-off between decoder complexity and segmentation performance.

Table 9. Computational complexity comparison.

Metric	SegFormer-B2	Prithvi + Native MLP	Prithvi + GL-FCR
Parameters (M)	27.358	305.330	308.341
Trainable parameters (M)	27.358	305.330	308.341

Inference time (ms/patch)	20.078 ± 2.751	14.794 ± 1.719	23.266 ± 4.469
Peak GPU allocated memory (GB)	0.220	1.184	1.303

4. Discussion

The results of this study provide several insights into the use of Earth observation foundation models for flood mapping across different sensor modalities and geographic conditions. Overall, the experiments demonstrate the effectiveness of the proposed GL-FCR decoder when coupled with foundation-model representations, while also highlighting important differences between optical, SAR, and multimodal configurations. Beyond performance on the standard SEN1Floods11 test set, the geographically held-out Bolivia evaluation and the zero-shot transfer to the independent WorldFloods dataset provide complementary evidence regarding model generalization. At the same time, the SAR-only and optical–SAR experiments offer additional insights into the role of radar observations for early flood mapping and their contribution when combined with high-quality optical imagery. These findings are discussed below in relation to existing state-of-the-art approaches, the architectural design of GL-FCR, and the practical requirements of operational flood mapping.

4.1. Main Findings and Strengths

The proposed Prithvi + GL-FCR framework demonstrated competitive performance across the standard SEN1Floods11 test set and the geographically held-out Bolivia subset. In the original evaluation conducted using random seed 42, Prithvi + GL-FCR achieved a Flood IoU of 83.62% on the standard test set, positioning it among the strongest optical approaches included in the comparison. Earlier results reported Flood IoUs of 84.03% for U-Net, 81.98% for Prithvi, 81.57% for SegFormer-B0, and 81.89% for SegFormer-B5 [30], while DeepLabV3+ achieved 79.02% [31]. Among more recent Earth observation foundation models, Prithvi-EO-2.0-300M and Prithvi-EO-2.0-600M-TL achieved 82.20% and 83.10%, respectively [17], while TerraMind-v1-B reached 82.26% [22] and DOFA obtained 76.83% [18]. More recently, Flood IoUs of 80.25% for S12-DINO, 82.22% for U-Prithvi, and 83.41% for Prithvi-CAFE were reported [23]. Although U-Net retains the highest Flood IoU on the standard test set among the methods considered (84.03%), the proposed Prithvi + GL-FCR remains highly competitive and slightly exceeds the recent Prithvi-CAFE (83.41%) and Prithvi-EO-2.0-600M-TL (83.10%) configurations. More importantly, a different pattern emerges under geographic transfer. On the Bolivia subset, Prithvi + GL-FCR achieved 81.83% Flood IoU, compared with 70.57% for U-Net, 76.62% for Prithvi, 71.46% for SegFormer-B0, and 75.82% for SegFormer-B5 [30], as well as 79.68% for U-Prithvi and 81.37% for Prithvi-CAFE [23]. The proposed model therefore records the highest Bolivia Flood IoU among the methods in the comparison for which results on this geographic holdout are available. Particularly noteworthy is the limited reduction from 83.62% on the standard test set to 81.83% on Bolivia, indicating that the performance of Prithvi + GL-FCR is largely preserved when transferred to geographically unseen conditions.

The multi-seed evaluation provides additional insight into the robustness of these performance comparisons. Although Prithvi + GL-FCR slightly outperformed SegFormer-B2 in the original single-seed evaluation on the standard SEN1Floods11 test set, the three-seed results revealed comparable average performance, with SegFormer-B2 achieving a marginally higher mean flood IoU (83.36% versus 83.07%). In contrast, Prithvi + GL-FCR maintained a higher mean flood IoU on the geographically held-out Bolivia subset (81.53% versus 80.68%), with lower variability across the evaluated seeds (0.27 versus 0.46 percentage points). These findings suggest that the proposed architecture offers competitive segmentation accuracy and potentially more stable performance on the Bolivia subset. However, the observed differences are modest, and the limited number of training runs

and evaluation samples prevents definitive conclusions regarding statistical superiority or broader geographic generalization.

This geographic robustness is further supported by the external zero-shot evaluation on WorldFloods, which provides a more stringent cross-dataset assessment than the Bolivia holdout. WorldFloods contains Sentinel-2 observations of flood events across diverse geographic and environmental settings [25]. Without any retraining, fine-tuning, or adaptation to WorldFloods, Prithvi + GL-FCR achieved 73.24% Flood IoU and 84.56% Flood F1 across all 509 WorldFloods v2 scenes, compared with 40.23% Flood IoU and 57.38% Flood F1 for SegFormer-B2 under the same zero-shot protocol. The Prithvi native MLP decoder achieved 72.15% Flood IoU and 83.35% Flood F1, indicating that GL-FCR provides a modest improvement over the native decoder under cross-dataset transfer. Moreover, Prithvi + GL-FCR achieved a higher Flood IoU in all 20 selected scenes compared with SegFormer-B2, demonstrating that the difference was consistently observed across these illustrative external samples, rather than being driven by a limited number of favorable cases. The scene-level statistics across all 509 scenes (Table 5) further show mean Flood F1 values of 74.83%, 72.87%, and 48.87% for Prithvi + GL-FCR, the native MLP decoder, and SegFormer-B2, respectively, with substantial scene-to-scene variability. The contrasting precision–recall profiles provide further insight into this behavior: SegFormer-B2 maintained a high precision of 92.71% but reached only 41.55% recall, reflecting substantial omission of the reference water/flood extent, whereas Prithvi + GL-FCR maintained a considerably more balanced 86.30% precision and 84.07% recall. Together, the Bolivia geographic holdout and the independent WorldFloods zero-shot evaluation strengthen the evidence that the proposed framework maintains its flood-mapping capability under both geographic and cross-dataset distribution shifts.

The SAR-based experiments provide a complementary perspective on the applicability of foundation models to flood mapping when optical observations are unavailable or compromised by cloud cover. On the SEN1Floods11 test set, TerraMind-B + GL-FCR achieved a Flood IoU of 68.81% and an mIoU of 81.84%, improving over the 68.00% Flood IoU and 81.06% mIoU reported for the original TerraMind-v1-B [22]. Its Flood IoU was also higher than the 65.63% obtained by FM-Mamba [32], the 67.20% of Attentive U-Net and 69.50% of the Fusion Network combining Sentinel-1, DEM, and permanent-water information [27], while the supervised DeepLabV3+ reached 67.63% [31]. The highest Flood IoU in the SAR comparison, 72.74%, was reported by the cross-modal distillation approach [31]; however, this method exploits paired optical–SAR information during training, whereas the final inference is performed from SAR imagery. Within our experiments, TerraMind-B + GL-FCR clearly outperformed the CNN baseline, increasing Flood IoU from 63.46% to 68.81% and Flood F1 from 77.64% to 81.53%. More importantly, this performance was preserved on the geographically held-out Bolivia subset, where TerraMind-B + GL-FCR reached 69.89% Flood IoU, 82.28% Flood F1, and 81.27% mIoU, compared with 63.96%, 78.02%, and 78.23%, respectively, for the CNN. The slight increase in Flood IoU from 68.81% on the test set to 69.89% on Bolivia is particularly encouraging under geographic transfer and is consistent with the qualitative results, where the TerraMind-based model recovered more complete and spatially continuous inundated areas with fewer omissions than the CNN. Although SAR-only performance remains below that obtained from clear optical imagery, these results demonstrate its practical relevance for early flood mapping, where the all-weather and day–night acquisition capability of Sentinel-1 can provide observations when timely cloud-free optical imagery is not available.

The integration of optical and SAR observations provides further insight into the contribution of multimodal information when high-quality optical imagery is already available. On the SEN1Floods11 test set, the proposed multimodal configurations achieved Flood IoUs between 82.78% and 83.45%, exceeding the 79.25% reported for the optical–

SAR DeepLabV3+ configuration [31]. The SegFormer + SAR-CNN single-level cross-modal fusion achieved the highest Flood IoU among our multimodal models on the standard test set, reaching 83.45%, together with a Flood F1 of 90.98% and an mIoU of 90.46%. The Prithvi + GL-FCR + SAR-CNN dual-decoder late-fusion configuration followed closely with 83.26% Flood IoU, 90.86% Flood F1, and 90.37% mIoU, while the dual-foundation configuration combining Prithvi + GL-FCR and TerraMind-B + GL-FCR obtained 82.78% Flood IoU, 90.58% Flood F1, and 90.08% mIoU. TerraMind-v1-B [22] reported an mIoU of 90.62% for its multimodal S1 + S2 configuration, which remains slightly above the mIoU obtained by our evaluated multimodal models. A particularly relevant finding emerges when these results are considered alongside the optical-only experiments. The optical Prithvi + GL-FCR configuration already achieved 83.62% Flood IoU on the test set, and none of the evaluated optical–SAR combinations surpassed this value. A similar pattern was observed on the geographically held-out Bolivia subset, although the Prithvi + GL-FCR + SAR-CNN dual-decoder late-fusion model reached 81.91% Flood IoU, marginally exceeding the 81.83% obtained by optical-only Prithvi + GL-FCR. This gain of only 0.08 percentage points, accompanied by a slight decrease on the standard test set from 83.62% to 83.26%, indicates that the inclusion of SAR does not systematically translate into improved segmentation when the optical observations already provide a strong representation of the flooded surface. Moreover, replacing the SAR-CNN branch with the stronger SAR-only TerraMind-B + GL-FCR representation did not improve the multimodal result, with the dual-foundation configuration reaching 82.78% and 81.14% Flood IoU on the test and Bolivia sets, respectively. This is an important finding because it suggests that the limited multimodal gain cannot be attributed solely to an insufficient SAR representation; rather, stronger standalone representations from two modalities do not necessarily yield additive benefits when fused. Consequently, the value of SAR in a multimodal flood-mapping framework may lie less in systematically increasing accuracy under clear optical conditions and more in providing complementary and operationally robust information when optical observations are degraded, cloud-contaminated, or unavailable.

4.2. Limitations and Future Perspectives

Despite the strong performance obtained across the different evaluation settings, several limitations should be acknowledged. The external zero-shot evaluation on WorldFloods was conducted on the complete WorldFloods v2 dataset comprising 509 scenes, rather than a limited subset. This experiment was intended to provide an independent cross-dataset assessment of models trained exclusively on SEN1Floods11, without retraining or fine-tuning on WorldFloods. Although Prithvi + GL-FCR achieved higher global Flood IoU than both SegFormer-B2 and the Prithvi native MLP decoder, the improvement over the native decoder was relatively modest, and substantial scene-level variability remained across the 509 scenes. These findings demonstrate the importance of distinguishing overall cross-dataset performance from consistency across individual flood events and environmental conditions. Future work should therefore extend the zero-shot evaluation to additional independent flood datasets, with particular attention to challenging scenes and variations in acquisition conditions, to provide a more comprehensive assessment of cross-dataset generalization.

An additional consideration concerns the comparability of the results reported in Tables 1–3. Although the proposed configurations achieve competitive performance relative to previously published approaches, differences in preprocessing, decoder architectures, training strategies, and evaluation protocols may influence the reported metrics. Consequently, small numerical differences between independently reported studies should not be interpreted as conclusive evidence of model superiority. The comparisons among

models evaluated within the present experimental framework provide a more consistent basis for assessing the relative effectiveness of the investigated architectures.

A second limitation concerns the dependence of the best-performing configuration on optical Sentinel-2 observations. The results obtained with Prithvi + GL-FCR demonstrate the value of multispectral optical information when suitable imagery is available; however, operational flood mapping frequently coincides with persistent cloud cover and adverse atmospheric conditions. Under such circumstances, the availability of usable optical observations cannot be guaranteed. Furthermore, the SEN1Floods11 training annotations used in this study provide water/flood segmentation targets but do not include explicit cloud-cover labels for training the model to identify or handle cloud-obstructed regions. In addition, cloud-contaminated and invalid pixels were excluded from the WorldFloods evaluation. Consequently, the reported zero-shot results support cross-dataset transferability over valid evaluated pixels but do not establish robustness to cloud obstruction or degraded optical imagery. The SAR experiments partly address this limitation, demonstrating that TerraMind-B + GL-FCR can provide geographically stable flood predictions using Sentinel-1 alone, although with lower segmentation accuracy than the optical configuration. Future investigations should therefore evaluate the models explicitly under different levels of optical degradation, including partial cloud obstruction, missing spectral information, and temporally mismatched optical acquisitions, to better characterize the conditions under which SAR becomes necessary rather than merely complementary.

The multimodal experiments further reveal a limitation of static optical–SAR fusion. Although SAR provides physically complementary information, its systematic inclusion did not improve the already strong optical predictions on the standard test set, while only a marginal improvement was observed on Bolivia. Moreover, replacing the CNN-based SAR representation with the stronger TerraMind-B + GL-FCR branch did not lead to improved multimodal performance. These findings suggest that treating both modalities as equally informative for every observation may not be optimal. A promising direction is therefore to develop quality- or reliability-aware multimodal strategies in which the contribution of each modality is dynamically adjusted according to observation conditions. Such a framework could rely predominantly on optical information when cloud-free Sentinel-2 imagery is available, progressively increase the contribution of SAR as optical quality deteriorates, and operate in a SAR-only mode when optical observations become unusable. This would more closely reflect the requirements of operational early flood mapping than a fixed fusion strategy. Future work will also investigate the integration of explicit cloud-cover annotations or cloud-quality masks during model development, together with controlled evaluations across different levels of cloud contamination and optical degradation. This would enable the development and assessment of cloud-aware flood-mapping strategies that identify unreliable optical regions and adaptively exploit Sentinel-1 SAR information where optical observations are obstructed.

Finally, the present experiments primarily assess predictive performance through a single training realization for each configuration. Although the consistent trends observed across the standard test set, Bolivia geographic holdout, and WorldFloods external evaluation provide complementary evidence, repeated training with different random seeds would allow the variability of the reported improvements to be quantified more rigorously, particularly where differences between competing methods are small. Future work should therefore incorporate multi-seed experiments and statistical uncertainty estimates, together with broader cross-event validation. From an operational perspective, further investigation should also consider computational efficiency, inference latency, and integration with near-real-time Sentinel-1 and Sentinel-2 data streams. These developments would extend the current framework from an accurate and transferable flood-segmentation approach toward a

more adaptive system capable of supporting rapid flood mapping under the heterogeneous observation conditions encountered during real-world disaster response.

5. Conclusions

This study presented GL-FCR, a decoder designed to enhance Earth observation foundation-model representations for flood mapping. The proposed approach combines multi-scale feature decoding with contextual and global–local information to improve the delineation of inundated areas. Its evaluation on SEN1Floods11 demonstrated robust performance on both the standard test data and the geographically held-out Bolivia subset, while the external zero-shot evaluation on the complete WorldFloods v2 dataset comprising 509 scenes further showed its ability to transfer to an independent dataset without model adaptation.

The broader experiments also highlighted the complementary roles of optical and SAR observations. Optical foundation-model features provided the strongest flood delineation when suitable imagery was available, whereas SAR offered an important alternative for early mapping under observation conditions that may restrict optical acquisitions. The multimodal analysis further showed that incorporating SAR does not necessarily improve an already strong optical representation, emphasizing the importance of considering data quality and modality relevance rather than systematically combining all available observations.

Overall, the findings support the use of foundation models coupled with task-specific contextual decoding as a promising direction for accurate and transferable flood mapping. Future work will focus on broader cross-dataset validation and adaptive multimodal strategies capable of dynamically exploiting optical and SAR information according to their availability and reliability, with the objective of supporting more robust operational flood mapping.

Author Contributions: T.G.: Writing—original draft, Software, Methodology, Conceptualization, Visualization. M.S.G.: Writing—review & editing, Methodology, Conceptualization, Supervision. B.S.: Writing—review & editing, Methodology, Conceptualization, Supervision. A.R. (Abdelaziz Rabehi): Writing—review & editing, Supervision. M.H.: Writing—review & editing, Supervision. A.R. (Abdelkader Riche): Writing—review & editing, Software, Methodology, Conceptualization, Visualization. M.G.: Writing—review & editing, Software, Methodology, Conceptualization, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The SEN1Floods11 dataset used for model training and evaluation is publicly available at https://github.com/cloudtostreet/Sen1Floods11 (accessed on 10 October 2026). The WorldFloods dataset used for external zero-shot evaluation is publicly available at https://doi.org/10.5281/zenodo.8153514 (accessed on 10 October 2026). The source code developed in this study will be made publicly available in an online repository upon acceptance of the manuscript.

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

References

1. Tellman, B.; Sullivan, J.A.; Kuhn, C.; Kettner, A.J.; Doyle, C.S.; Brakenridge, G.R.; Erickson, T.A.; Slayback, D.A. Satellite Imaging Reveals Increased Proportion of Population Exposed to Floods. *Nature* **2021**, *596*, 80–86. https://doi.org/10.1038/s41586-021-03695-w.
2. Boulange, J.; Hanasaki, N.; Yamazaki, D.; Pokhrel, Y. Role of Dams in Reducing Global Flood Exposure under Climate Change. *Nat. Commun.* **2021**, *12*, 417. https://doi.org/10.1038/s41467-020-20704-0.
3. Tarpanelli, A.; Mondini, A.C.; Camici, S. Effectiveness of Sentinel-1 and Sentinel-2 for Flood Detection Assessment in Europe. *Nat. Hazards Earth Syst. Sci.* **2022**, *22*, 2473–2489. https://doi.org/10.5194/nhess-22-2473-2022.

4. Martinis, S.; Groth, S.; Wieland, M.; Knopp, L.; Rättich, M. Towards a Global Seasonal and Permanent Reference Water Product from Sentinel-1/2 Data for Improved Flood Mapping. *Remote Sens. Environ.* **2022**, *278*, 113077. https://doi.org/10.1016/j.rse.2022.113077.
5. Wagner, W.; Bauer-Marschallinger, B.; Roth, F.; Raiger-Stachl, T.; Reimer, C.; McCormick, N.; Matgen, P.; Chini, M.; Li, Y.; Martinis, S.; et al. The Fully-Automatic Sentinel-1 Global Flood Monitoring Service: Scientific Challenges and Future Directions. *Remote Sens. Environ.* **2026**, *333*, 115108. https://doi.org/10.1016/j.rse.2025.115108.
6. Konapala, G.; Kumar, S.V.; Khalique Ahmad, S. Exploring Sentinel-1 and Sentinel-2 Diversity for Flood Inundation Mapping Using Deep Learning. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 163–173. https://doi.org/10.1016/j.isprsjprs.2021.08.016.
7. Bai, Y.; Wu, W.; Yang, Z.; Yu, J.; Zhao, B.; Liu, X.; Yang, H.; Mas, E.; Koshimura, S. Enhancement of Detecting Permanent Water and Temporary Water in Flood Disasters by Fusing Sentinel-1 and Sentinel-2 Imagery Using Deep Learning Algorithms: Demonstration of Sen1Floods11 Benchmark Datasets. *Remote Sens.* **2021**, *13*, 2220. https://doi.org/10.3390/rs13112220.
8. Tavus, B.; Can, R.; Kocaman, S. A CNN-based flood mapping approach using SENTINEL-1 data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *V-3–2022*, 549–556. https://doi.org/10.5194/isprs-annals-V-3-2022-549-2022.
9. Li, Z.; Demir, I. U-Net-Based Semantic Classification for Flood Extent Extraction Using SAR Imagery and GEE Platform: A Case Study for 2019 Central US Flooding. *Sci. Total Environ.* **2023**, *869*, 161757. https://doi.org/10.1016/j.scitotenv.2023.161757.
10. Sharma, N.K.; Saharia, M. DeepSARFlood: Rapid and Automated SAR-Based Flood Inundation Mapping Using Vision Transformer-Based Deep Ensembles with Uncertainty Estimates. *Sci. Remote Sens.* **2025**, *11*, 100203. https://doi.org/10.1016/j.srs.2025.100203.
11. Preetha, S.; M S, S.P.; Manikandan, P. Optimized Flood Scene Segmentation with Swin Transformer-Based Architecture. *Sci. Rep.* **2026**, *16*, 20532. https://doi.org/10.1038/s41598-026-39188-x.
12. Shaheen, M.T.; Iqbal, H.; Khurshid, N.; Sadia, H.; Saeed, N. SwinSegFormer: Advancing Aerial Image Semantic Segmentation for Flood Detection. *IEEE Open J. Comput. Soc.* **2025**, *6*, 645–657. https://doi.org/10.1109/OJCS.2025.3565185.
13. Lu, X.; Gou, T.; Huang, Z.; Yang, Y.; Jiao, L.; Li, L.; Liu, X.; Liu, F. Multimodal Segformer for Flood Rapid Mapping with Sentinel-2 Data. In Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7 July 2024; pp. 790–793.
14. Arian, Md.A.R.; Shishir, Md.M.H.; Samin, S.I.C.; Siddique, S. Flood Induced Economic Damage Assessment from Satellite Imagery Using Vision Transformers. In Proceedings of the TENCON 2025—2025 IEEE Region 10 Conference (TENCON), Kota Kinabalu, Malaysia, 27 October 2025; pp. 943–947.
15. Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.B.; Ermon, S. SatMAE: Pre-Training Transformers for Temporal and Multi-Spectral Satellite Imagery. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 197–211. https://doi.org/10.52202/068431-0015.
16. Reed, C.J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; Darrell, T. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1 October 2023; pp. 4065–4076.
17. Szwarcman, D.; Roy, S.; Fraccaro, P.; Gíslason, Þ.E.; Blumenstiel, B.; Ghosal, R.; De Oliveira, P.H.; De Sousa Almeida, J.L.; Sedona, R.; Kang, Y.; et al. Prithvi-EO-2.0: A Versatile Multitemporal Foundation Model for Earth Observation Applications. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 4400120. https://doi.org/10.1109/TGRS.2025.3642610.
18. Fuller, A.; Millard, K.; Green, J.R. CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36. https://doi.org/10.52202/075280-0241.
19. Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A.J.; Hanna, J.; Borth, D.; Papoutsis, I.; Le Saux, B.; Camps-Valls, G.; Zhu, X.X. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. *arXiv* **2024**, arXiv:2403.15356. https://doi.org/10.48550/arXiv.2403.15356.
20. Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; Kembhavi, A. SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1 October 2023; pp. 16726–16736.
21. Tseng, G.; Fuller, A.; Reil, M.; Herzog, H.; Beukema, P.; Bastani, F.; Green, J.R.; Shelhamer, E.; Kerner, H.; Rolnick, D. Galileo: Learning Global & Local Features of Many Remote Sensing Modalities. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*; Proceedings of Machine Learning Research: Ann Arbor, MI, USA, 2025; Volume 267, pp. 60280–60300.

22. Jakubik, J.; Yang, F.; Blumenstiel, B.; Scheurer, E.; Sedona, R.; Maurogiovanni, S.; Bosmans, J.; Dionelis, N.; Marsocci, V.; Kopp, N.; et al. TerraMind: Large-Scale Generative Multimodality for Earth Observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2025; pp. 7383–7394.
23. Kaushik, S.; Maurya, L.; Tellman, B. Prithvi-Complimentary Adaptive Fusion Encoder (CAFE): Unlocking Full-Potential for Flood Inundation Mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*; IEEE: Piscataway, NJ, USA, 2026; pp. 1435–1444.
24. Bonafilia, D.; Tellman, B.; Anderson, T.; Issenberg, E. Sen1Floods11: A Georeferenced Dataset to Train and Test Deep Learning Flood Algorithms for Sentinel-1. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2020; pp. 835–845.
25. Portalés-Julià, E.; Mateo-García, G.; Purcell, C.; Gómez-Chova, L. Global Flood Extent Segmentation in Optical Satellite Images. *Sci. Rep.* **2023**, *13*, 20316. https://doi.org/10.1038/s41598-023-47595-7.
26. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34.
27. Yadav, R.; Nascetti, A.; Ban, Y. Deep Attentive Fusion Network for Flood Detection on Uni-Temporal Sentinel-1 Data. *Front. Remote Sens.* **2022**, *3*, 1060144. https://doi.org/10.3389/frsen.2022.1060144.
28. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
29. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*; IEEE: Piscataway, NJ, USA, 2016; pp. 565–571.
30. Li, W.; Lee, H.; Wang, S.; Hsu, C.-Y.; Arundel, S.T. Assessment of a New GeoAI Foundation Model for Flood Inundation Mapping. In Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, Hamburg, Germany, 13 November 2023; pp. 102–109.
31. Garg, S.; Feinstein, B.; Timnat, S.; Batchu, V.; Dror, G.; Rosenthal, A.G.; Gulshan, V. Cross-Modal Distillation for Flood Extent Mapping. *Environ. Data Sci.* **2023**, *2*, e37. https://doi.org/10.1017/eds.2023.34.
32. Wang, B.; Chen, Z.; Zou, H.; Yuan, A.; Pan, Y.; Guo, H.; Li, J. FM-Mamba: End-to-End Non-Causal Mamba-Based Network for Efficient Flood Mapping. *Sci. Rep.* **2026**, *16*, 24857. https://doi.org/10.1038/s41598-026-56046-y.

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