Off to new Shores: A Dataset & Benchmark for (near-)coastal Flood Inundation Forecasting

Brandon Victor
La Trobe University
b.victor@latrobe.edu.au

Mathilde Letard
University of Rennes
mathilde.letard@univ-rennes.fr

Peter Naylor
ESA Φ-lab
peter.naylor@esa.int

Karim Douch
ESA Science Hub
karim.douch@esa.int

Nicolas Longépé
ESA Φ-lab
nicolas.longepe@esa.int

Zhen He
La Trobe University
z.he@latrobe.edu.au

Patrick Ebel
ESA Φ-lab
patrick.ebel@esa.int

Abstract

Floods are among the most common and devastating natural hazards, imposing immense costs on our society and economy due to their disastrous consequences. Recent progress in weather prediction and spaceborne flood mapping demonstrated the feasibility of anticipating extreme events and reliably detecting their catastrophic effects afterwards. However, these efforts are rarely linked to one another and there is a critical lack of datasets and benchmarks to enable the direct forecasting of flood extent. To resolve this issue, we curate a novel dataset enabling a timely prediction of flood extent. Furthermore, we provide a representative evaluation of state-of-the-art methods, structured into two benchmark tracks for forecasting flood inundation maps i) in general and ii) focused on coastal regions. Altogether, our dataset and benchmark provide a comprehensive platform for evaluating flood forecasts, enabling future solutions for this critical challenge. Data, code & models are shared at `https://github.com/Multihuntr/GFF` under a CC0 license.

1 Introduction

Floods are among the most impactful natural disasters, both in terms of the societal as well as the economic costs they impose (15). The consent amongst climate scientists and disaster relief experts is that this trend will aggravate in the coming decades(50; 36; 27; 61), close by rivers (21) and in particular near the coastlines due to rising sea levels and more severe extreme weather events (35; 10; 45; 65; 64; 46). Yet, closeness to waterways is of economical importance such that endangered regions have grown in population, thus bringing more people at the risk of floods (66; 60).

Hence, international collaborations and efforts e.g. in the context of the *United Nations (UN) Sustainable Development Goals (SDG)* (62; 58) tackle climate change mitigation and adaptation. According to the *Early warnings for all* initiative of the World Meteorological Organization and the UN, every person on Earth shall be protected by early warning systems until 2027 (67), but to date significantly more effort is required towards covering the Global South and developing early warning systems for coastal inundation (71). In line with these needs, our contribution is a global dataset and benchmark for a timely forecasting of flood extent maps. Our novel Global Flood Forecasting (GFF) dataset represents climate zone and continent distributions of events as reported in the Dartmouth

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-3644

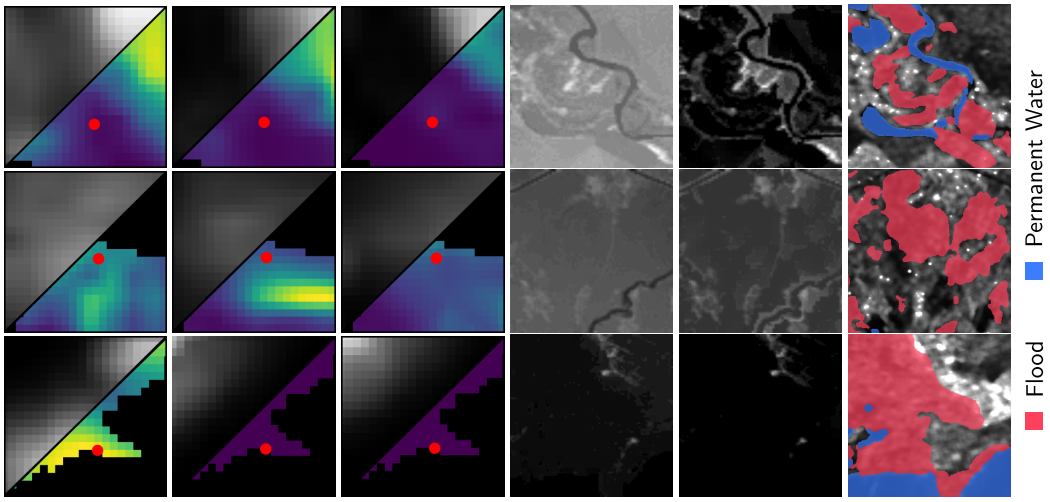


Figure 1: **Exemplary data.** Columns: Three ERA5 and ERA5-Land time series samples, DEM, Height Above Nearest Drainage (HAND). Pre-flood Sentinel-1 (S1) overlaid with target-time flood map. Columns 1-3 provide context data at a coarse scale, red dots indicates the coverage of columns 4-6 at fine scale. Rows: Three examples, showcasing floods near a river, settlement and coastline. The events are due to heavy rain, tropical storms and a storm surge, illustrating the diversity of GFF.

Flood Observatory (DFO) (66), while focusing on (near-)coastal areas and their varied drivers of flood hazards. To underline the diversity of cases covered by GFF, Fig. 1 illustrates examples of flood due to heavy rain, plus high tides in the last case. Each sample combines multi-temporal atmospheric reanalysis products, high resolution terrain models and simulated precipitation drainage, hydrological basin attributes, as well as pre-flood Sentinel-1 (S1) radar satellite observations and flood extent annotations as targets. While offering vast information, combining this multi-modal and multi-scale data poses technical challenges, especially for hand-crafted solutions common in flood forecasting.

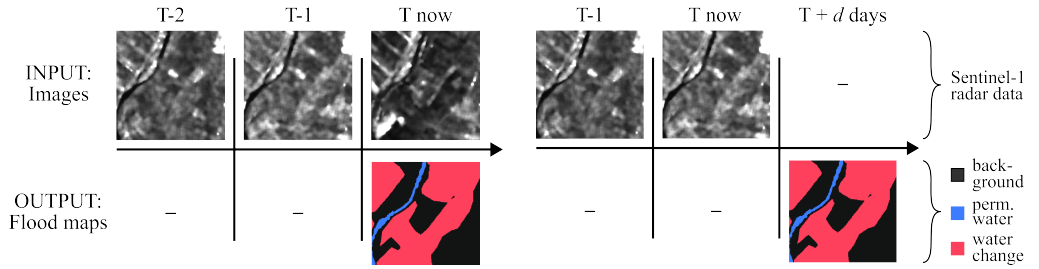
Meanwhile, data-driven machine learning has been a major cause of recent breakthroughs in modeling of ungauged rivers (38; 53) and rapid flood mapping (11; 66; 12). While the former may help anticipating river run-off and the latter can support ongoing relief efforts, there's a lack of research on ahead-of-time prediction of the inundation maps themselves and the comparability of such forecasting models on a common benchmark dataset. This is unfortunate, as the timely availability of flood extent maps would allow humanitarian agencies to undertake preparatory measures such as the evacuation of endangered population ahead of time rather than acting post-hoc. Though forecasting of inundation maps provides critical information for disaster preparedness, few prior works tackle this challenge and there is an absence of benchmarks to facilitate such developments (51). The aim of our work is to fill this gap by introducing GFF as a global dataset and benchmark for flood extent forecasting, analysis-ready for modern machine learning approaches. In sum, our main contributions are:

- The curation of GFF, a novel global multi-modal multi-temporal dataset for (near-)coastal flood forecasting, derived from six distinct sensors and products at two separate scales of resolution. Through careful stratification, we sample regions representative of climate zones and continents for which major floods have been recorded.

- The design of two tracks for i) general flood extent prediction and ii) with a focus on separating coastal versus near-coastal and inland floods. Each region experiences distinct climate, weather and flood drivers such that tailored approaches may be most fruitful.

- The benchmarking of established methods, to provide the reader with a comprehensive overview of the diverse landscape of methods and the state-of-the-art on the defined tracks.

https://doi.org/10.52202/079017-3644



(a) Rapid Flood Mapping (b) Flood Map Forecasting

Figure 2: **Conceptual similarities and differences** between the two tasks of a) in-event rapid flood mapping and b) pre-event forecasting potential flood maps. While the former focuses on detecting change of water coverage in observations at prior dates versus now, the latter is about predicting such change at a given lead time d where no observation is yet available. Beyond internalizing the physical signatures of moisture and water, this requires learning the dynamics of associated flood drivers.

2 Related Work

2.1 River streamflow & runoff forecasting

The close relation between riverine floods and a river's weather-driven run-off sustain general interests in river streamflow modeling. Classical forecasting approaches are simulation-based and build on complex, hand-crafted features calibrated to encode basin-specific properties and dynamics (3; 8). More recent data-driven models are based on deep neural networks and demonstrated the ability to forecast streamflow at ungaged sites (38; 53). The limitations of such approaches are that they lack any direct relation to actionable flood extent maps, as there is no straightforward translation from river run-off to actual inundation maps. The closest efforts pursuing this endeavor are given by the prior work of (54) who relate river run-off to historical flood extent observed via multispectral satellites by querying records in a lookup table. The shortcomings of this approach are its reliance on optical imagery which may be affected by clouds, and especially the reliance on historic flood event observations at every region of interest. Furthermore, the intermediate modeling of run-off oftentimes either relies on access to river gauges or focuses on upstream regions nearby the source, which renders it inapplicable to coastal areas. In contrast to this work and prior efforts, our dataset seeks to enable research and development for a timely and all-weather forecasting of flood inundation maps on a global scale, not constrained to sites where historic flood observations are readily available for and including (near-)coastal regions.

2.2 Rapid mapping

The spaceborne mapping of floods and their impacts in order to support disaster relief is a central mandate of several international services and charters (9; 2). While rapid mapping and its automation have a long history (22; 14; 26), recent progress in machine learning offers access to products at a pace and accuracy outmatching expert hand-labeled annotations (48; 12). Key to this is the availability of large-scale annotated datasets (47; 66; 12). Our work builds on the recent Kuro Siwo dataset of S1 observations and models (12) to obtain reference flood maps on par in terms of quality with post-event expert annotations (12), subsequently post-processed and used as forecasting targets herein. The task tackled in our work is related to rapid mapping, but demands the timely forecasting of inundation maps and thus allows for pre-disaster preparatory measures. Accordingly, the setup of available observations and target date of predictions differs across both tasks, as conceptualized in Fig. 2. While rapid mapping is about the detection of physical properties such as surface soil moisture and water mass, our forecasting task requires translating atmospheric dynamics and their hydro-meteorological impact onto land surface while taking into account factors like local topography and basin properties. Independent of their similarities and differences, forecasting and rapid mapping are complementary in their function and both are crucial for disaster mitigation and relief, respectively.

https://doi.org/10.52202/079017-3644

Dataset	Task	Sample size	Resolution	Sample count	Static input	Dynamic input	Event count	Timestamps
SEN12-FLOOD (59)	classification	512 × 512	10 m	336	-	S1, S2	3	circa 9-14
OMBRIA (18)	segmentation	256 × 256	10 m	1,688	-	S1, S2	23	1 Pre + Post
S1GFloods (63)	segmentation	256 × 256	10 m	5,360	-	S1	46	1 Pre + Post
CAU-Flood (31)	segmentation	256 × 256	10 m	18,302	-	S1, S2	18	1 Pre + Post
Kuro Siwo (12)	segmentation	224 × 224	10 m	67,490	DEM	S1	43	2 Pre + Post
GRDC GRDB (13)	regression	1D sequence	in-situ	10,000+	-	river gauges	-	10,000+
HYSETS (6)	regression	1D sequence	in-situ, basin & 10 - 30 km	14,425	basin properties	river gauges, NRCan + SCDNA + Livneh + ERA5(-Land)	-	10,000+
Caravan (39)	regression	1D sequence	in-situ & basin	10,000+	HydroATLAS	river gauges, ERA5-Land	-	10,000+
Global Flood Forecasting (ours)	**segmentation**	**224 × 224**	**10 m & 5 - 30 km**	**163,873**	**DEM, HAND, HydroATLAS**	**S1, GloFAS, ERA5(-Land)**	**298**	**20 Pre + Post**

Table 1: **Overview of datasets** for flood mapping (top) and flood forecasting (bottom) purposes. The former feature high resolution imaging while the latter focus on in-situ time series. GFF enables forecasting of flood extent by curating sequences of gridded products at high spatial resolution.

3 Data

The GFF dataset focuses on (near-)coastal regions characterized by a diversity of causes underpinning flood hazard—ranging from pluvial, fluvial or coastal drivers such as storm surges, to potential compound events. The resulting dataset includes 298 Regions of Interest (ROI) experiencing an equal count of spatially and temporally separated flood events in the years 2014-2020. For each flood event, the dataset contains flood drivers as input and flood segmentation maps at event time as targets. The dataset is accompanied by a pre-defined 5-fold cross-validation benchmark split to enable fair comparison between models and drive innovation. All data are provided in a rasterized *TIFF* file format (72) and prepared to facilitate developing and evaluating data-driven machine learning models. Beyond the flood maps, observations and products provided herein, the GFF dataset is extendable and comes with all scripts needed to expand to further regions, flood events or to include new modalities.

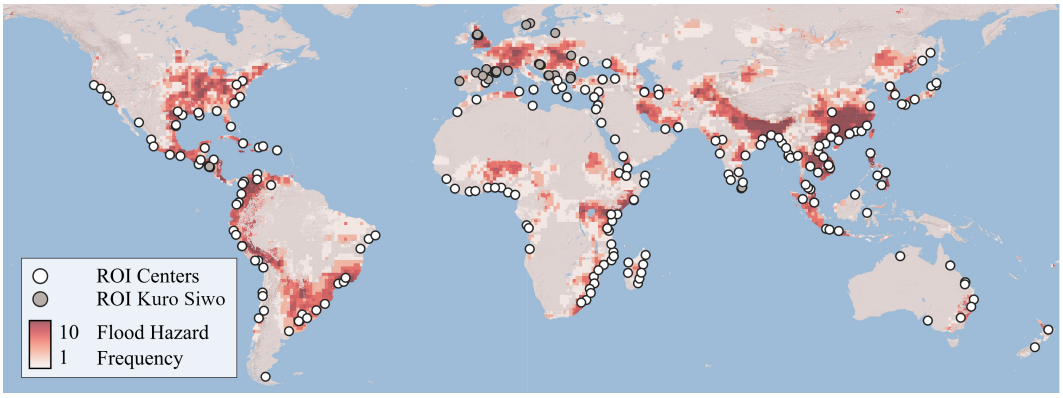


Figure 3: **Map** of dataset. Points are centers of curated ROI and their associated flood events. Red shadings indicate the distribution of global flood hazard frequency (15). Many of the endangered regions are close to the sea, especially in the (sub-)tropics. Unlike most prior work addressing well-monitored or upstream regions, these areas are particularly well represented in our dataset.

3.1 Identifying candidate Regions Of Interest

To collect a diverse set of floods on a global scale, we cross-reference the exhaustive records of DFO with HydroBASIN level 8 basins (40) underlying events featured in three prior works: the recent Kuro Siwo dataset for rapid mapping (12), a list of 280 landfalling tropical cyclones which caused extreme precipitation and subsequent flooding (68), and finally any (near-)coastal basins characterized by HydroATLAS (42; 41)—a global compendium of river basins and their attributes.

First, all of Kuro Siwo's expert-labelled data in the years of our analysis are included for training and testing. While its distribution spans six continents, its focus is on Europe in the well-monitored northern extratropical lattitudes. Complementary, the *second* group of basins feature inland level 8 basins at the epicenters of cyclones' land footprints, which mostly cover (sub-)tropical areas. Finally,

the *third* group of basins are those indicated by HydroATLAS to be at or near the coast, filling in any properties not yet sufficiently captured by the preceding two sources of events.

To prioritize on the subset of most relevant ROI and focus on hazardous events, candidate areas of all three sources are sorted by impact and affected population as given by DFO and HydroATLAS. All candidate ROI and dates are iterated through and checked for coincidence with S1 orbits. For creating global flood extent labels at the desired quality, at least one S1 image during the flood time period and two S1 pre-event images are required. Floodmaps were not generated for a ROI within $< 5\,^{\circ}$proximity of a previously selected ROI or if its climate zone is already represented sufficiently.

3.2 Capturing the diverse drivers of flood & flood extent forecasting

Each sample in the dataset consists of multi-modal inputs as shown in Fig. 1, paired with floodmap segmentation labels. The various input types are split into two scales: ERA5, ERA5-Land, HydroATLAS and downsampled CopDEM30 are provided as a *coarse context*. Pre-event Sentinel-1, CopDEM30 and HAND are given at the fine *local scale* of the floodmaps. Most forcings are static, but at the coarse scale ERA and ERA5-Land are multi-temporal. In summary, the inputs of GFF are:

- **ERA5** and **ERA5-Land** (32; 52), containing daily atmospheric state reanalysis. The provided $9 + 14$ atmospheric variables are corresponding to the forcings used in established data-driven river-runoff models (37; 54; 53). ERA5 and ERA5-Land complement each other: the first provides vital information of weather conditions over both the land and sea. The latter focuses on land at a significantly finer horizontal resolution for better forcings.
- **HydroATLAS** (42), providing static multi-level basin attributes including a subset of 87 hydro-environmental key properties like 'aridity index' and 'average wet season' as selected by stream-flow experts (37; 54; 53). HydroATLAS is licensed to be freely available for any research project, and is provided in this dataset as raster data.
- **Sentinel-1** (S1) Ground Range Detected (GRD) Synthetic Aperture Radar images (25), preprocessed via ESA's SNAP toolbox (76) and replicating the pipeline of Kuro Siwo (12). In the temporal proximity of excessive rainfall, cloud-penetrating radar measurements are more useful than optical imagery. Observations at the closest time *before* the event's recorded onset are used as input, while images *during* the event are used for the label generation.
- **Copernicus Digital Elevation Model** at global 30 meters resolution (CopDEM30) (1) and **Height Above Nearest Drainage** (HAND) maps (7) derived from the surface model via NASA's HydroSAR package (49). The former provides a representation of the ROI's terrain, while the latter is a downstream product specifying the local topology's drainage potentials.
- Daily aggregated **river discharge and runoff water** history modeled via the Global Flood Awareness System (GloFas) (29), part of the operational flood forecasting within the Copernicus Emergency Management Service. Explicit, hydrological modeling of such variables is a key step in operational flood forecasting systems (4; 17; 54). To explore the forecasting of extent maps, we include this modality in our dataset and encourage investigating models with and without it for exploring one versus two-stage modeling.

Conceptually, in an initial step the contextual information allows a model to integrate weather, soil and elevation data over the surrounding topography for a time window starting 20 days before the target date. In a second stage, the model can then use the processed contextual representation together with finer resolved information to fill in details of exactly where the water will go in the local area. The data are preprocessed as detailed in the corresponding section of the accompanying *Datasheet*.

3.3 Generating floodmaps anywhere

A central design objective of GFF is to cover a diverse set of continents, climate zones and land cover types beyond ROI which are already monitored and served well. For this sake, floodmaps are computed by ensembling two rapid mapping models pretrained on the Kuro Siwo dataset and further refined in a post-processing step. The two utilized models are a masked auto-encoder pre-trained ViT (30; 16) and a SNUNet change detector (20). Both models use pre- and in-event Sentinel-1 images to classify whether a pixel of the in-event Sentinel-1 image became flooded or not. For rapid mapping,

 https://doi.org/10.52202/079017-3644

the models classify no-water pixels with F1 scores of 99.07 and 98.97 and water pixels with F1 scores of 87.58 and 86.52 on hold-out data, respectively. A third set of floodmaps is generated by ensembling the logits of both models. All three types of floodmaps are released with the GFF dataset, but the ensemble labels are considered the default. The exception are ROI with hand-annotated expert labels collected in Kuro Siwo, which are utilized herein upon availability.

For creating labels at a given ROI, a uniform grid is created over the scene and tiles of 224×224 pixels size which intersect with HydroRIVER geometries are added to an initial search set. The tiles in this set are sorted following the upstream flow of any adjacent river and the 200 most downstream cells are selected. Orthogonal to the grid-covered river stream, a buffer of 2 additional tiles are added to both sides of this set, resulting in up to $200 \times (1 + 2 \times 2)$ initial tiles per ROI. Starting from these at most 1000 tiles, a conditional floodfill algorithm is used to search for affected areas. Specifically, floodmaps are generated for each tiles and if any tile shows significant flooding (defined as $> 5\%$ of the tile), then a buffer of 3 neighboring tiles is added to expand the open set of search tiles. This process terminates at a maximum of 2500 tiles per ROI or when no more flooding is encountered.

3.4 Label post-processing

The models utilized for generating labels achieve high performance in the rapid mapping scenario (12), but their outputs were found to exhibit artifacts when deployed in practice. These include tiling artifacts and speckled segmentation artifacts at multiple levels, which we corrected for as follows.

First, as is common for neural networks operating on individual local tiles one at a time, the generated outputs initially displayed serious tiling artifacts. These artifacts are cleaned for by generating logits of 50%-overlapping tiles and then taking a weighted average at each pixel. The resulting averaged logits smoothly transition between adjacent tiles. This method entirely removes all tiling artifacts.

Second, the initial maps displayed strong speckling patterns induced by the upconvolution operations used in the pre-trained models (55). We used a 5×5 pixels majority filter to determine a more reliable boundary between classes and applied a contour-finding algorithm to remove any remaining blobs smaller than 50 pixels (43; 56). Importantly, these post-processing steps were not performed on a tile-by-tile basis but on the full area after combining tiles.

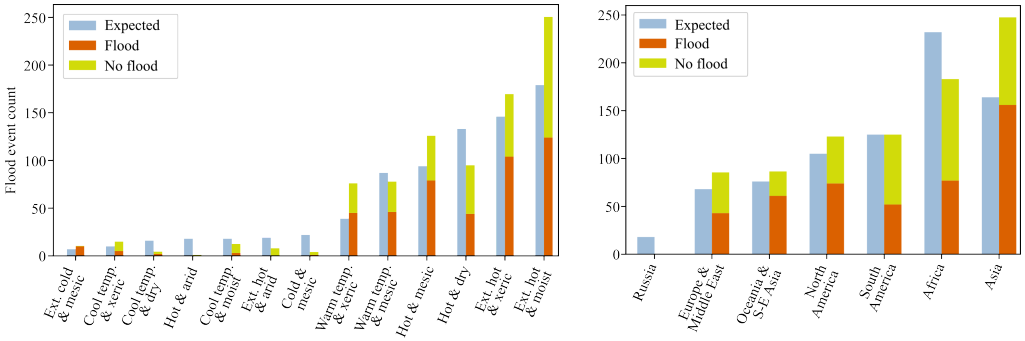
Moreover, it was found during preliminary testing that the pretrained SNUNet model (20) performed better when driven by CopDEM30 compared to the DEM originally used in (12), potentially due to CopDEM30's better quality, and thus we use CopDEM30 for generating floodmaps with SNUNet.

Finally, to improve the delineation between flood and permanent water pixels in the ensembled generated maps, we merge both classes into the former and superimpose the permanent water labels from ESA WorldCover (73) for the latter. This was found to improve separation of both water classes.

3.5 Selecting representative ROI via stratification

Following the selection process of section 3.1, candidate sites are proactively filtered to ensure a representative distribution of prior flood events across continents and climate zones. To obtain a reference, we estimate the true distribution of worldwide flood events using the Cartesian product of DFO events and level 8 basins. Each flood $\times$ basin pairing is counted for the basin's continent/climate zone. An expected distribution is created by scaling down all bins in the true distribution such that we expect to produce floodmaps for 2000 level 8 basins. While iterating over all potential ROI, a ROI is skipped if its climate zone is already well represented. The distributions and ROI iterations are calculated within each continent separately to allow for continents with different proportions of climate zones affected by floods, and ensure that the dataset has global coverage.

Stratification is complicated by the fact that any ROI's floodmap may extend to multiple level 8 basins but which ones are covered isn't known prior to rapid mapping. Therefore, a ROI is skipped if over half of its area is covered by climate zones which are already sufficiently well represented. While generating, a ROI may or may not show flooding. As ROI without flooding are less relevant than ROI with flooding for the creation of our dataset, they were included but counted at half for the statistics. Flooded ROI may vary in their extent if DFO indicates long time intervals. In this case, multiple S1 passes may be available and we pick the timestamp exhibiting the largest flood extent.



(a) Climate zone distributions

(b) Continent distributions

Figure 4: **Empirical distributions of floods** in our dataset regarding their frequency across different a) climate zones and b) continents. In both regards, the histogram of cumulative flood (orange) and no flood observations (green) qualitatively mirrors the expected occurrence of all global flood events reported by the DFO (blue), with minor discrepancies due to the (near-)coastal focus.

The outlined procedure results in 298 ROI of flood events depicted in the map of Fig. 3 (99 of which are listed by both DFO and the database of (68), indicating cyclone-driven floods). The selected regions cover 1049 level 8 sub-basins and a total of 164845 tiles with 9.5 % of the tiles featuring flood extent, for all of which input forcings and segmentation labels as described in sections 3.2 and 3.3 are provided. The histograms of climate zone and continent coverage are depicted in Fig. 4 and confirm that the distribution of global flood history recorded by the DFO is well represented.

3.6 Data splits

The GFF dataset is accompanied by a suggested split into five partitions for a 5-fold cross-validation setup. Partitions are defined to not leak any test information into optimization. This is accomplished by partitioning the whole world by HydroATLAS level 4 basins, ensuring hydrologically well-defined groupings. ROI are assigned to the partition of largest overlap and adjacency is avoided by excluding ROI closer than circa 500 km to one another during the candidate selection stage. Finally, the defined splits are backward-compatible to river gauge splits of the popular Caravan (39) dataset for river streamflow modeling, allowing the community to explore synergies between both contributions.

4 Benchmarks

To highlight the value of the GFF dataset, we benchmark a diverse and representative set of state-of-the-art models on two distinct tracks: global flood extent forecasting and a separate focus on coastal floods. In both cases, the task is to predict a flood segmentation map of binary values (B: background vs W: water) at a given lead time d days after the pre-flood Sentinel-1 observation, with d varying per flood event. For evaluating the performance of flood extent forecasts, we mask the permanent water pixels given by ESA WorldCover labels as described in section 3.4 and then evaluate the F1 score of the binary forecasts over all other pixels, checking whether a pixel became flooded or not. This is to focus on the pixels potentially undergoing class change rather than on permanent water bodies whose dynamics are comparably static. We report overall F1 score and scores for both classes individually.

For every track, the experimental setup is a 5-fold cross-validation scheme, with one partition as the test set, another as the validation set and the remaining three for training. For each baseline, all F1 scores are reported as the means of performances of model instances across partitions, plus their cross-fold standard deviation in F1 scores in order to provide an estimate of cross-run spread.

The first track measures the performance of forecasting inundation maps across all ROI included. The second track evaluates models specifically on ROI separated into *coastal* versus *near-/non-coastal* areas, with tiles categorized whether they are less or more than 10 km distant from the nearest coast. This distinction is to differentiate between areas directly impacted by coastal hazards compared to regions only affected by near-shore weather dynamics, which may require differences in modeling.

https://doi.org/10.52202/079017-3644

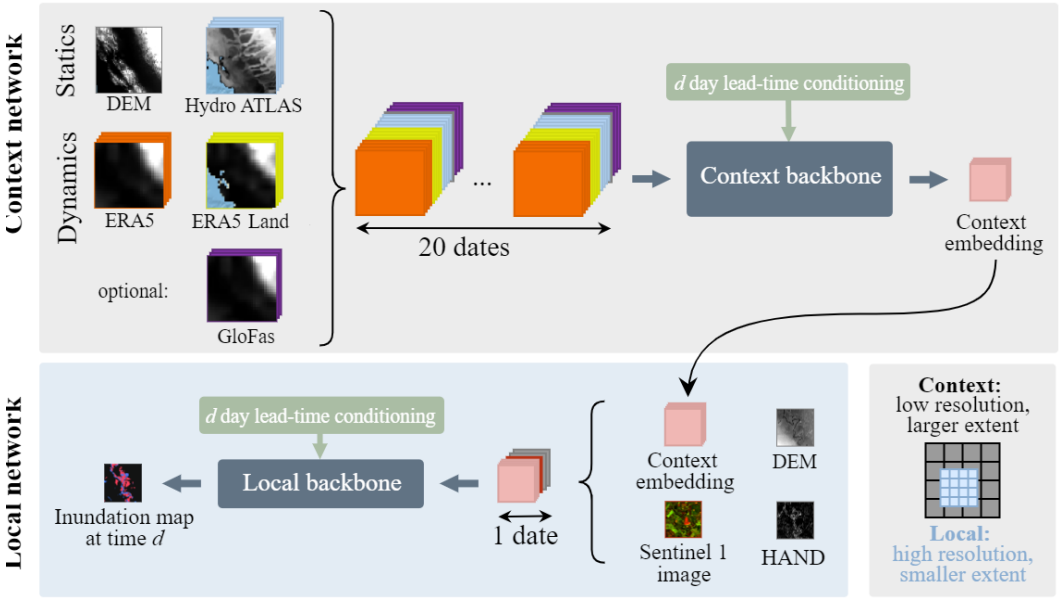


Figure 5: **Baseline model platform design,** accommodating for i) a *context network (top)* which processes a spatio-temporal sequence of coarse resolution context data and whose output feature embeddings are then processed by ii) a *local network (bottom)*, concatenated with local high resolution data. The two network backbones (in dark grey) are placeholders for the different baselines benchmarked herein. The final output is a flood segmentation forecast at a given lead time.

4.1 Baselines

To provide baselines for our benchmark, we propose a novel two-level pipeline which serves as a platform to combine existing state-of-the-art spatio-temporal architectures with minimal extras.

The platform is designed to process input forcings of coarse context and local information. The horizontal resolution of context forcings such as GloFas is at $0.05°$ (circa 5.5 km at the equator), but local information such as S1 images and the flood maps are resolved to a 10 m resolution, finer by three orders of magnitude. To our knowledge there are no existing solutions processing spatial information at such difference in scale. Although there exist works that utilise both S1 and ERA5 at the same time (69), they do not do so while modelling spatial relationships. This challenge invites for innovative technical contributions from the broader machine learning audience.

Per setup, our platform hosts any of the diverse established baselines—ranging from convolutional models, LSTM (44), to temporal attention (23) and vision transformers (70; 5). In all cases, i) the context inputs (ERA5, ERA5-Land, HydroATLAS & CopDEM) are embedded using any of the aforementioned multi-temporal architectures. Then, ii) a local segmentation model takes the aggregated context embedding along with the local data (Sentinel-1, CopDEM30 & HAND) to predict a floodmap, conditioned on a given lead time of d days via Feature-wise Linear Modulation (FiLM) (57). For each setup incorporating a specific baseline, the context and the local modules share the same backbone architecture, but do not share weights. The described pipeline is illustrated in Fig. 5. As a fifth baseline, we implement a simple logistic regression model (33; 24). Analogous to the prior work of (24), we train it on the mono-temporal local information and report the learned channel weightings to provide interpretation of feature importance in the Supplementary Material.

Each baseline is evaluated following the 5-fold cross-validation protocol, trained via the ADAM optimizer (34) at a batch size of 16 for 10 epochs with an initial learning rate of 10^{-3} and an exponential decay of 0.8. The cost function is a binary cross entropy loss with class weightings of 0.5 and 5 for background versus water classes — roughly equal to each category's inverse frequency. Using the same loss, models are evaluated on the validation partition every epoch and the checkpoint with best validation loss is used for testing. Computations have been performed on-premises at ESA, training on NVIDIA RTX A6000 GPU. All code and checkpoints are made public.

4.2 Track 1: Global flood map prediction

This track is on translating all contextual spatio-temporal and local high-resolution data detailed in section 3.2 to flood extent forecasts at a given lead time, as described in the preceding paragraphs.

The results are reported in Table 2, whose left half shows the overall and class-wise F1 scores for each of the five baselines. U-TAE performs best overall, closely followed by LSTM U-Net. Interestingly, the strength of LSTM-based models for flood extent forecasting mirrors the method's competitiveness in the related task of river streamflow forecasting (38), for which it likewise is the state-of-the-art (37). Complementary to these outcomes, the right side of the table reports F1 scores when only evaluating on the hand-labeled data of Kuro Siwo (KS). Compared to the previous results, there is a trend of performance decrease but most differences are within a standard deviation, which are generally higher on the Kuro Siwo labels. Even though it now is LSTM U-Net obtaining best results, we conclude that performances are roughly comparable for annotated and our derived labels. In summary, the best baselines can forecast flood inundation at an F1 score of circa 0.77, with strong performances on the background class. However, the prediction of flooded pixels is significantly more challenging, as this is the minority class and may benefit most from future research and modeling.

Table 2: **Track 1.** Benchmarking of five baselines on the GFF dataset. U-TAE performs best overall, while LSTM U-Net is best when only evaluating on the subset of KS hand-annotated ROI.

Model	F1	F1-B	F1-W	$F1_{KS}$	F1-B_{KS}	F1-W_{KS}
U-TAE (23)	$\mathbf{0.77 \pm 0.04}$	$\mathbf{0.97 \pm 0.00}$	$\mathbf{0.57 \pm 0.07}$	0.72 ± 0.02	0.97 ± 0.01	0.48 ± 0.05
LSTM U-Net (44)	0.76 ± 0.04	$\mathbf{0.97 \pm 0.00}$	0.55 ± 0.08	$\mathbf{0.73 \pm 0.03}$	0.97 ± 0.01	$\mathbf{0.50 \pm 0.07}$
3DConv U-Net (44)	0.76 ± 0.04	0.97 ± 0.01	0.54 ± 0.08	0.70 ± 0.08	0.97 ± 0.02	0.43 ± 0.16
MaxViT U-Net (70; 5)	0.75 ± 0.03	0.96 ± 0.01	0.53 ± 0.06	0.73 ± 0.05	$\mathbf{0.98 \pm 0.01}$	0.49 ± 0.09
logistic regression (33)	0.66 ± 0.04	0.93 ± 0.02	0.40 ± 0.07	0.65 ± 0.10	0.94 ± 0.04	0.36 ± 0.16
U-TAE GloFAS (23)	0.76 ± 0.05	$\mathbf{0.97 \pm 0.00}$	0.55 ± 0.08	0.73 ± 0.05	0.97 ± 0.01	0.49 ± 0.10

4.3 Track 2: Coastal versus near-coastal & inland flood map prediction

The focus of this track is on disentangling the distinct challenges of forecasting coastal versus near-coastal floods, separately analyzed and benchmarked herein. (Near-)coastal floods are difficult to model for established approaches not only due to their potential compound nature (74), but also due to being most distant from the headwater basin at which river-runoff models typically excel (38; 39). To provide further insight, we follow the preceding experimental protocol but evaluate the baselines distinctly on the dataset's subset of coastal versus near-coastal and inland areas. Regions 10 km or closer to the nearest shore are considered as coastal, a distance about the size of one ERA5-Land cell.

The outcomes are shown in Table 3, the left and right sides contain outcomes for coastal versus near-coastal and inland areas, respectively. Overall, baseline performances are higher at coastal regions. This may be due to the incidental presence of coastal wetlands whose re-occurring inundation are more predictable, or due to the high count of cyclone-driven floods in the dataset whose extreme precipitation is most destructive near the coasts even more so than right at the shores (28; 19; 75). Overall, the regional separation reveals performance differences, which deserve further research.

Table 3: **Track 2.** Benchmarking of five baselines on the GFF dataset, separating (c)oastal versus (n)ear-coastal & inland areas. U-TAE and 3DConv U-Net perform best at the coasts, while U-TAE performs best on inland regions. Altogether, baseline performances are higher at coastal regions.

Model	$F1_c$	F1-B_c	F1-W_c	$F1_n$	F1-B_n	F1-W_n
U-TAE (23)	$\mathbf{0.80 \pm 0.06}$	$\mathbf{0.95 \pm 0.02}$	0.65 ± 0.11	$\mathbf{0.76 \pm 0.03}$	$\mathbf{0.98 \pm 0.00}$	$\mathbf{0.55 \pm 0.07}$
LSTM U-Net (44)	0.78 ± 0.07	0.94 ± 0.03	0.63 ± 0.11	0.75 ± 0.04	0.97 ± 0.00	0.53 ± 0.07
3DConv U-Net (44)	0.80 ± 0.08	$\mathbf{0.95 \pm 0.02}$	$\mathbf{0.65 \pm 0.09}$	0.74 ± 0.04	0.97 ± 0.01	0.50 ± 0.09
MaxViT U-Net (70; 5)	0.78 ± 0.06	0.94 ± 0.03	0.63 ± 0.10	0.74 ± 0.03	0.97 ± 0.01	0.50 ± 0.06
logistic regression (33)	0.73 ± 0.04	0.92 ± 0.02	0.54 ± 0.06	0.65 ± 0.05	0.93 ± 0.02	0.36 ± 0.09
U-TAE GloFAS (23)	0.78 ± 0.08	0.95 ± 0.02	0.62 ± 0.10	0.75 ± 0.04	0.97 ± 0.0	0.52 ± 0.09

https://doi.org/10.52202/079017-3644

5 Discussion

Potential societal impact: Our work and the future research it may facilitate closely align with the UN *Sustainable Development Goals* (62; 58). We do not foresee any direct adversarial impact of our work on society, and all information utilized herein is already made publicly accessible by the responsible authorities (e.g. ASF & NASA, Copernicus, ECMWF, ESA) in line with their mandates.

Known limitations & future work: The definition of what is considered a flood is contested in the context of ecosystems such as salt marshes, swamps and practices like irrigation farming. Rather than focusing on such non-hazardous scenarios, disasters listed by the DFO and close to populated areas as indicated via WorldPop guide our dataset curation. However, non-hazardous and semi-persistent land submersion may still be included in our dataset in case they occur in a harmful event's periphery.

The post-event mapping algorithm, although we build upon the state-of-the-art and apply further post-processing, is not flawless. While being more sophisticated than established heuristic approaches, our exchange with operational flood forecasting teams clarified that a separate and regularly re-running of permanent water detection is oftentimes employed in practice. Accordingly, we recommend a combination of our approach with such a complementary step to meet the best forecasting practices.

Finally, the meteorological forcings included in the dataset are reanalysis rather than reforecasts. This is to disentangle challenges in the quickly advancing field of weather forecasting from the task at the heart of our work, so benchmarking outcomes can be isolatedly attributed to the latter. However, using reanalysis as forcings may overestimate downstream task performances compared to utilizing forecasts during operational deployment. Hence, we recommend evaluating developed models using forecasts as forcings in case they shall be deployed in practice.

6 Conclusions

To tackle the challenge of timely flood extent forecasting and empower researchers to address this matter on a global scale, we curate the first dataset and public benchmark on this task. The dataset features multi-modal and multi-temporal data of (near-)coastal flood events distributed across six continents and 13 climate zones, encompassing a variety of events driven by pluvial, fluvial, coastal or compound hazards. As diverse as these cases, the collected observations and products pose unique technical challenges inviting for innovative contributions from the scientific community—ranging from optimal solutions for sensor fusion to multi-scale Earth observation modeling. While we consider these technical challenges stimulating for a broader machine learning audience, it is the societal importance of disaster risk reduction via timely forecasting that we wish to highlight at last.

Acknowledgments and Disclosure of Funding

We thank SmartSat CRC for funding the research visit of Brandon Victor at ESA Φ-lab. Furthermore, we would like to thank our colleagues at ESA, as well as Frederik Kratzert and Adi Gerzi Rosenthal at Google Research's Flood Forecasting Department for the fruitful discussions and precious feedback.

References

[1] DS Airbus. Copernicus dem copernicus digital elevation model validation report. *Airbus Defence and Space—Intelligence: Potsdam, Germany*, 2020.

[2] Andrea Ajmar, Piero Boccardo, Marco Broglia, Jan Kucera, Fabio Giulio-Tonolo, and Annett Wania. Response to flood events: The role of satellite-based emergency mapping and the experience of the copernicus emergency management service. *Flood damage survey and assessment: New insights from research and practice*, pages 211–228, 2017.

[3] Lorenzo Alfieri, Peter Burek, Emanuel Dutra, Blazej Krzeminski, David Muraro, Jutta Thielen, and Florian Pappenberger. Glofas–global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, 17(3):1161–1175, 2013.

[4] Lorenzo Alfieri, Peter Salamon, Alessandra Bianchi, Jeffrey Neal, Paul Bates, and Luc Feyen. Advances in pan-european flood hazard mapping. *Hydrological processes*, 28(13):4067–4077, 2014.

[5] Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alex Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*, 2023.

[6] Richard Arsenault, François Brissette, Jean-Luc Martel, Magali Troin, Guillaume Lévesque, Jonathan Davidson-Chaput, Mariana Castañeda Gonzalez, Ali Ameli, and Annie Poulin. A comprehensive, multi-source database for hydrometeorological modeling of 14,425 north american watersheds. *Scientific Data*, 7(1):243, 2020.

[7] Alaska Satellite Facility (ASF). GlobalHAND/GLO30_HAND (ImageServer). `Website.` https://gis.asf.alaska.edu/arcgis/rest/services/GlobalHAND/GLO30_HAND/ImageServer, 2021. Accessed: 2024-05-03.

[8] Hylke E Beck, Ming Pan, Peirong Lin, Jan Seibert, Albert IJM van Dijk, and Eric F Wood. Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *Journal of Geophysical Research: Atmospheres*, 125(17):e2019JD031485, 2020.

[9] J-L Bessis, Jerome Bequignon, and Ahmed Mahmood. The international charter "space and major disasters" initiative. *Acta Astronautica*, 54(3):183–190, 2004.

[10] Emanuele Bevacqua, Michalis I Vousdoukas, Giuseppe Zappa, Kevin Hodges, Theodore G Shepherd, Douglas Maraun, Lorenzo Mentaschi, and Luc Feyen. More meteorological events that drive compound coastal flooding are projected under climate change. *Communications earth & environment*, 1(1):47, 2020.

[11] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 210–211, 2020.

[12] Nikolaos Ioannis Bountos, Maria Sdraka, Angelos Zavras, Ilektra Karasante, Andreas Karavias, Themistocles Herekakis, Angeliki Thanasou, Dimitrios Michail, and Ioannis Papoutsis. Kuro siwo: 12.1 billion m^2 under the water. a global multi-temporal satellite dataset for rapid flood mapping. *arXiv preprint arXiv:2311.12056*, 2023.

[13] Global Runoff Data Centre. Global Runoff Data Base (GRDB). `Website.` https://grdc.bafg.de/GRDC/EN/01_GRDC/13_dtbse/database_node.html , 1980. Accessed: 2024-08-14.

[14] Tom De Groeve. Flood monitoring and mapping using passive microwave remote sensing in namibia. *Geomatics, Natural Hazards and Risk*, 1(1):19–35, 2010.

[15] Maxx Dilley. *Natural disaster hotspots: a global risk analysis*, volume 5. World Bank Publications, 2005.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] Francesco Dottori, Milan Kalas, Peter Salamon, Alessandra Bianchi, Lorenzo Alfieri, and Luc Feyen. An operational procedure for rapid flood risk assessment in europe. *Natural Hazards and Earth System Sciences*, 17(7):1111–1126, 2017.

[18] Georgios I Drakonakis, Grigorios Tsagkatakis, Konstantina Fotiadou, and Panagiotis Tsakalides. Ombrianet—supervised flood mapping via convolutional neural networks using multitemporal sentinel-1 and sentinel-2 data fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2341–2356, 2022.

[19] Rebecca Emerton, Hannah Cloke, Andrea Ficchi, Laurence Hawker, Sara de Wit, Linda Speight, Christel Prudhomme, Philip Rundell, Rosalind West, Jeffrey Neal, et al. Emergency flood bulletins for cyclones idai and kenneth: A critical evaluation of the use of global flood forecasts for international humanitarian preparedness and response. *International Journal of Disaster Risk Reduction*, 50:101811, 2020.

[20] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[21] Luc Feyen, Rutger Dankers, Katalin Bódis, Peter Salamon, and José I Barredo. Fluvial flood risk in europe in present and future climates. *Climatic change*, 112:47–62, 2012.

[22] Bo-Cai Gao. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266, 1996.

[23] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[24] Sebastian Gerard, Yu Zhao, and Josephine Sullivan. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. *Advances in Neural Information Processing Systems*, 36:74515–74529, 2023.

[25] Dirk Geudtner, Ramón Torres, Paul Snoeij, Malcolm Davidson, and Björn Rommen. Sentinel-1 system capabilities and applications. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 1457–1460. IEEE, 2014.

[26] Laura Giustarini, Renaud Hostache, Patrick Matgen, Guy J-P Schumann, Paul D Bates, and David C Mason. A change detection approach to flood mapping in urban areas using terrasar-x. *IEEE transactions on Geoscience and Remote Sensing*, 51(4):2417–2430, 2012.

[27] Avantika Gori, Ning Lin, Dazhi Xi, and Kerry Emanuel. Tropical cyclone climatology change greatly exacerbates us extreme rainfall–surge hazard. *Nature Climate Change*, 12(2):171–178, 2022.

[28] Timothy M Hall and James P Kossin. Hurricane stalling along the north american coast and implications for rainfall. *Npj Climate and Atmospheric Science*, 2(1):17, 2019.

[29] Shaun Harrigan, Ervin Zsoter, Lorenzo Alfieri, Christel Prudhomme, Peter Salamon, Fredrik Wetterhall, Christopher Barnard, Hannah Cloke, and Florian Pappenberger. Glofas-era5 operational global river discharge reanalysis 1979–present. *Earth System Science Data*, 12(3):2043–2060, 2020.

 https://doi.org/10.52202/079017-3644

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
[31] Xiaoning He, Shuangcheng Zhang, Bowei Xue, Tong Zhao, and Tong Wu. Cross-modal change detection flood extraction based on convolutional neural network. *International Journal of Applied Earth Observation and Geoinformation*, 117:103197, 2023.
[32] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.
[33] Fantine Huot, R Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.
[34] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
[35] Ebru Kirezci, Ian R Young, Roshanka Ranasinghe, Sanne Muis, Robert J Nicholls, Daniel Lincke, and Jochen Hinkel. Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st century. *Scientific reports*, 10(1):11629, 2020.
[36] Thomas Kleinen and Gerhard Petschel-Held. Integrated assessment of changes in flooding probabilities due to climate change. *Climatic Change*, 81(3):283–312, 2007.
[37] Frederik Kratzert, Martin Gauch, Grey Nearing, and Daniel Klotz. Neuralhydrology — a python library for deep learning research in hydrology. *Journal of Open Source Software*, 7(71):4050, 2022.
[38] Frederik Kratzert, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018.
[39] Frederik Kratzert, Grey Nearing, Nans Addor, Tyler Erickson, Martin Gauch, Oren Gilon, Lukas Gudmundsson, Avinatan Hassidim, Daniel Klotz, Sella Nevo, et al. Caravan-a global community dataset for large-sample hydrology. *Scientific Data*, 10(1):61, 2023.
[40] Bernhard Lehner and Günther Grill. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27(15):2171–2186, 2013.
[41] Bernhard Lehner, Mathis L Messager, Maartje C Korver, and Simon Linke. Global hydro-environmental lake characteristics at high spatial resolution. *Scientific Data*, 9(1):351, 2022.
[42] Simon Linke, Bernhard Lehner, Camille Ouellet Dallaire, Joseph Ariwi, Günther Grill, Mira Anand, Penny Beames, Vicente Burchard-Levine, Sally Maxwell, Hana Moidu, et al. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific data*, 6(1):283, 2019.
[43] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998.
[44] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition workshops*, pages 75–82, 2019.
[45] Alexandre K Magnan, Michael Oppenheimer, Matthias Garschagen, Maya K Buchanan, Virginie KE Duvat, Donald L Forbes, James D Ford, Erwin Lambert, Jan Petzold, Fabrice G Renaud, et al. Sea level rise risks and societal adaptation benefits in low-lying coastal areas. *Scientific reports*, 12(1):10677, 2022.
[46] Cristian Martinez-Villalobos and J David Neelin. Regionally high risk increase for precipitation extreme events under global warming. *Scientific Reports*, 13(1):5579, 2023.
[47] Gonzalo Mateo-Garcia, Joshua Veitch-Michaelis, Lewis Smith, Silviu Vlad Oprea, Guy Schumann, Yarin Gal, Atılım Güneş Baydin, and Dietmar Backes. Towards global flood mapping onboard low cost satellites with machine learning. *Scientific reports*, 11(1):7249, 2021.
[48] Patrick Matgen, Sandro Martinis, Wolfgang Wagner, Vahid Freeman, Peter Zeil, Niall McCormick, et al. Feasibility assessment of an automated, global, satellite-based flood-monitoring product for the copernicus emergency management service. *Luxembourg: Publications Office of the European Union*, 2020.
[49] Franz Josef Meyer, Andrew Molthan, Jordan Robert Bell, Lori Ann Schultz, Ronan Lucey, Batuhan Osmanoglu, Minjeong Jo, David B McAlpin, Alex Lewandowski, Thomas Meyer, et al. Hydrosar: A cloud-based sar data analysis service to monitor hydrological disasters and their impact on population and agriculture. In *AGU Fall Meeting Abstracts*, volume 2020, pages NH026–06, 2020.
[50] P Christopher D Milly, Richard T Wetherald, KA Dunne, and Thomas L Delworth. Increasing risk of great floods in a changing climate. *Nature*, 415(6871):514–517, 2002.
[51] Hafiz Suliman Munawar, Ahmed WA Hammad, and S Travis Waller. Remote sensing methods for flood prediction: A review. *Sensors*, 22(3):960, 2022.
[52] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data*, 13(9):4349–4383, 2021.
[53] Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Frederik Kratzert, Asher Metzger, Sella Nevo, et al. AI increases global access to reliable flood forecasts. *arXiv preprint arXiv:2307.16104*, 2023.

[54] Sella Nevo, Efrat Morin, Adi Gerzi Rosenthal, Asher Metzger, Chen Barshai, Dana Weitzner, Dafi Voloshin, Frederik Kratzert, Gal Elidan, Gideon Dror, et al. Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, 26(15):4013–4032, 2022.
[55] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.
[56] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
[57] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
[58] Claudio Persello, Jan Dirk Wegner, Ronny Hänsch, Devis Tuia, Pedram Ghamisi, Mila Koeva, and Gustau Camps-Valls. Deep learning and Earth observation to support the Sustainable Development Goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):172–200, 2022.
[59] Clément Rambour, Nicolas Audebert, E Koeniguer, Bertrand Le Saux, M Crucianu, and Mihai Datcu. Flood detection in time series of optical and SAR images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43(B2):1343–1346, 2020.
[60] Jun Rentschler, Paolo Avner, Mattia Marconcini, Rui Su, Emanuele Strano, Michalis Vousdoukas, and Stéphane Hallegatte. Global evidence of rapid urban growth in flood zones since 1985. *Nature*, 622(7981):87–92, 2023.
[61] Matthew Rodell and Bailing Li. Changing intensity of hydroclimatic extreme events revealed by grace and grace-fo. *Nature Water*, 1(3):241–248, 2023.
[62] Jeffrey D Sachs, Christian Kroll, Guillame Lafortune, Grayson Fuller, and Finn Woelm. *Sustainable Development Report 2022*. Cambridge University Press, 2022.
[63] Tamer Saleh, Xingxing Weng, Shimaa Holail, Chen Hao, and Gui-Song Xia. DAM-net: Flood detection from SAR imagery using differential attention metric-based vision transformers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212:440–453, 2024.
[64] Sonia I Seneviratne, Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, Alejandro Di Luca, Subimal Ghosh, I Iskander, James Kossin, Sophie Lewis, et al. Weather and climate extreme events in a changing climate (chapter 11). 2021.
[65] Mohsen Taherkhani, Sean Vitousek, Patrick L Barnard, Neil Frazer, Tiffany R Anderson, and Charles H Fletcher. Sea-level rise exponentially increases coastal flood frequency. *Scientific reports*, 10(1):6466, 2020.
[66] Beth Tellman, Jonathan A Sullivan, Catherine Kuhn, Albert J Kettner, Colin S Doyle, G Robert Brakenridge, Tyler A Erickson, and Daniel A Slayback. Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, 596(7870):80–86, 2021.
[67] The UN Global Early Warning Initiative for the Implementation of Climate Adaptation. Early warnings for all: Executive action plan 2023-2027. Website: https://www.preventionweb.net/publication/early-warnings-all-executive-action-plan-2023-2027, 2023. Accessed: 2024-05-03.
[68] Helen A Titley, Hannah L Cloke, Shaun Harrigan, Florian Pappenberger, Christel Prudhomme, JC Robbins, EM Stephens, and Ervin Zsótér. Key factors influencing the severity of fluvial flood hazard from tropical cyclones. *Journal of Hydrometeorology*, 22(7):1801–1817, 2021.
[69] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.
[70] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022.
[71] United Nations Office for Disaster Risk Reduction (UNDRR). Global Status of Multi-Hazard Early Warning Systems. Website: https://www.undrr.org/reports/global-status-MHEWS-2023#download, 2023. Accessed: 2024-05-03.
[72] Richard H Wiggins, H Christian Davidson, H Ric Harnsberger, Jason R Lauman, and Patricia A Goede. Image file formats: past, present, and future. *Radiographics*, 21(3):789–798, 2001.
[73] D Zanaga, R Van De Kerchove, W De Keersmaecker, N Souverijns, C Brockmann, R Quast, J Wevers, A Grosu, A Paccini, S Vergnaud, et al. ESA WorldCover 10 m 2020 v100. 2021, 2021.
[74] Wei Zhang, Ming Luo, Si Gao, Weilin Chen, Vittal Hari, and Abdou Khouakhi. Compound hydrometeorological extremes: Drivers, mechanisms and methods. *Frontiers in Earth Science*, 9:673495, 2021.
[75] Laiyin Zhu, Kerry Emanuel, and Steven M Quiring. Elevated risk of tropical cyclone precipitation and pluvial flood in houston under global warming. *Environmental Research Letters*, 16(9):094030, 2021.
[76] Marco Zuhlke, Norman Fomferra, Carsten Brockmann, Marco Peters, Luis Veci, Julien Malik, and Peter Regner. SNAP (Sentinel Application Platform) and the ESA Sentinel 3 toolbox. In *Sentinel-3 for Science Workshop*, volume 734, page 21, 2015.

 https://doi.org/10.52202/079017-3644

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] , our two key contributions are the curation of a novel dataset for flood inundation map forecasting and the subsequent benchmarking on it.
   (b) Did you describe the limitations of your work? [Yes] , see section 5. While we discuss potential limitations, we consider these to accompany our contribution's focus on a given scope per definition. We hope the reader agrees that our work's benefits and opportunities outweigh potential limitations, which may be addressed in the future.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] , see section 5. We can't foresee any direct negative impact. We are in contact with teams running flood mapping and forecasting in an operational setting, and in case of concerns may discuss any potential risks with our colleagues.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] and we have ensured conformity.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] , no theoretical results are included.
   (b) Did you include complete proofs of all theoretical results? [N/A] , no theoretical results are included.
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] . All code, data and instructions are provided at `https://github.com/Multihuntr/GFF`. This includes all code to reproduce the main experimental results, as well as everything needed to re-download, process and extend the dataset with new regions or modalities.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] , see section 3.6 and the accompanying Datasheet for Datasets for information on the splits. Training hyperparameters are specified in section 4.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] , standard deviations are reported in all of our outcomes due to the 5-fold cross-validation experimental protocol as specified in section 3.6.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] , this is specified in section 4.1. The utilized resources are of modest scale and will not pose a bottleneck for reproducibility.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] , our work integrates data across several existing assets, co-registering them in time and space. We cite the creators of all prior work.
   (b) Did you mention the license of the assets? [Yes] , we release the code and checkpoints created as well as data derived in our work under a CC0 license. Furthermore, we specify each individual asset's license in the supplementary material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] , new assets (code, data, model checkpoints) are provided under the URL `https://github.com/Multihuntr/GFF`.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All data we utilize is already made publicly accessible by the responsible authorities (e.g. ASF & NASA, Copernicus, ECMWF, ESA) in line with their mandates.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our data has a spatial resolution ranging from circa 30 meters to 30 kilometers, based on which no personally identifyable information can be derived.
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] , no crowdsourcing or experiments involving human subjects were conducted.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] , no crowdsourcing or experiments involving human subjects were conducted.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] , no crowdsourcing or experiments involving human subjects were conducted.

https://doi.org/10.52202/079017-3644