

Generating Aerial Flood Prediction Imagery with Generative Adversarial Networks

Natasha Randall¹^a, Gernot Heisenberg¹^b and Juan Luis Ramirez Duval²^c

¹*Institute of Information Science, Technical University of Applied Sciences Cologne, Germany*

²*Institute for Natural Resources Technology and Management, Technical University of Applied Sciences Cologne, Germany*
fi

Keywords: Flood Forecasting, Generative Adversarial Networks, Image Generation, Deep Learning.

Abstract: Floods are one of the most dangerous, impactful natural disasters, and flood forecasting is a critical component of effective pre-flooding preparedness. In this paper a data-driven approach to flood forecasting is presented, which provides photorealistic predictions that are less computationally expensive to generate than traditional physically-based models. A ‘PairedAttention’ generative adversarial network (GAN) was developed, that combines attention and content mask subnetworks, and was trained on paired sets of pre- and post-flooding aerial satellite images aligned with topographical data. The PairedAttention GAN achieved 88% accuracy and an F1 score of 0.8 at flood predictions on three USA flood events, and an ablation study determined that the digital elevation model was the most significant factor to improving the GAN’s performance. Although the model is a successful proof-of-concept for the effectiveness of a data-driven GAN to generate photorealistic, accurate aerial flood prediction imagery, it nevertheless struggled with generalisation, indicating an important avenue for future research.

1 INTRODUCTION

Floods are an extremely deadly and costly natural hazard, accounting for 47% of all weather-related disasters between 1995 and 2015, leading to immense economic damages and losses of life (CRED, 2015). Climate change is also increasing the prevalence and intensity of storm and flooding events (UNDP, 2023). Developing effective flood risk management strategies is thus more important than ever, and flood forecasting is a critical component to supporting pre-flooding preparedness (Jain et al., 2018).

The traditional approach to making detailed flood inundation forecasts begins with a numerical weather prediction (Ming et al., 2020). The precipitation forecast is then input into hydrological models to create hydrographs, which depict water level information over time (Arduino et al., 2005). Using an external coupling system, hydraulic or hydrodynamic models use the inflow hydrographs as boundary conditions to simulate the flow of waters, by solving partial differential equations of continuity and momentum (Allaby,

2014). Although these ‘process-based’ models can make highly detailed and accurate flood inundation predictions, they have a very high computational requirement, as the prediction must be recomputed for each new area (Henonin et al., 2013). The calibration process of a hydrological model is also very time consuming, reaching hours or even days in length (Mihon et al., 2013). Furthermore, process-based models require huge amounts of data that explicitly represent the underlying physical characteristics of the flood (Devia et al., 2015), and in-depth knowledge and expertise is required to work with the hydrological parameters (Mosavi et al., 2018).

Alternatively, data-driven approaches to flood forecasting use deep learning, neural network based models that learn from historical flood data, to approximate the outputs of the computationally expensive hydraulic/hydrodynamic models (Guo et al., 2021). At inference time, neural networks can make much faster predictions than process-based models in complex environments, as they can easily handle the impact of factors like buildings or trees on diverting fluid flow, which conversely add a lot of complexity to the calculations of the hydraulic models (Liu et al., 2019). Neural network based models can also be directly applied to unseen areas, whereas process-based

^a <https://orcid.org/0009-0008-0937-7417>

^b <https://orcid.org/0000-0002-1786-8485>

^c <https://orcid.org/0000-0003-2239-8921>

models must always be retrained on the new topography.

Furthermore, generative deep learning models are also capable of creating photorealistic predictions and visualisations of floodwaters, whereas the traditional process-based models can output only numerical predictions. Hama et al. (2021) argue that colour-coded flood hazard maps are not engaging or intuitive, and that alone, they are insufficient to enhance situational awareness and eliminate cognitive bias (Lütjens et al., 2024). In contrast, realistic aerial imagery of predicted flooding outcomes can help orient rescue personnel to landscape features (Sivanpillai et al., 2021) and improve situational readiness for efficient post-flooding interventions (Goswami et al., 2022). Compelling flood imagery also effectively raises awareness of the potential impact of climate change on individuals' personal environments (Schmidt et al., 2021).

The contributions of our work include the development of a data-driven flood prediction model, which uses a generative adversarial network (GAN) architecture to generate aerial flood imagery that is both photorealistic and accurate, by modelling the flood forecasting process. Quantitative state-of-the-art results are limited, as related works have predominantly focused on photorealism and not predictive accuracy (Luccioni et al., 2021) and require running additional process-based numerical simulations to provide the accuracy to the data-driven model (Lütjens et al., 2024). However, our GAN was trained on only open observational data, using only tools distributed under open licences, and without the use of any expensive simulations, thus demonstrating the viability of a more accessible approach to developing flood prediction models.

The availability of datasets on which to train generative flood prediction models is currently very limited, with no benchmark dataset, hence we also release the dataset developed for this paper. It includes over 1000 pixel-aligned sets of pre-flooding satellite images, a high resolution digital elevation model, flow accumulation, distance to rivers, cartographical map, and ground truth post-flooding satellite images.

The research questions explored in this paper are:

- How accurately can GANs generate aerial flood prediction imagery?
- Which GAN architecture generates the most accurate flood predictions?
- Which topographic factors most improve the accuracy of the flood predictions?
- How accurately can the GANs generalise to new flood events and features?

2 RELATED WORK

The first generative models that were used to create images relied on a convolutional neural network architecture (LeCun et al., 1998). Flood susceptibility mapping using convolutional and autoencoder deep learning architectures has been successfully demonstrated by Wang et al. (2020) and Ahmadlou et al. (2021) respectively. In these studies, a range of variables were input into the models, including altitude, slope, land use and lithology, and the probability of flooding occurring at each pixel was subsequently output.

The generative adversarial network (GAN) exploits the strength of neural networks as universal function approximators (Hornik et al., 1989) to learn a mapping from random noise to a synthetic image that is indistinguishable from real images (Goodfellow et al., 2014). The basic GAN consists of two networks - a generator and a discriminator - that are trained simultaneously and work adversarially. The discriminator is a classifier, which trains to better discriminate between real images and the synthetic images created by the generator, thus incentivising the generator to train to generate more convincingly realistic synthetic images, which are capable of 'fooling' the discriminator. The loss function of a GAN comprises the probability that the discriminator correctly predicts that real images are real ($\log D(x)$), and that synthetic images are synthetic ($\log(1 - D(G(z)))$). It is therefore described as a 'minmax' function, because the discriminator wants to maximise this function, whereas the generator wants to minimise it (Goodfellow et al., 2014).

In 2017, Isola et al. introduced the Pix2Pix architecture for image-to-image translation, which generates images that are conditioned on an additional input image. Pix2Pix therefore needs access to paired image sets in order to train using a supervised learning approach. Hofmann and Schüttrumpf (2021) and do Lago et al. (2023) utilised a conditional GAN architecture to generate binary flood predictions, by training their models on data generated by hydrodynamic simulations.

The CycleGAN architecture was developed to use an unsupervised learning strategy without the need for paired images, by utilising the concept of cycle consistency. The CycleGAN model (Zhu et al., 2017) consists of two generators and two discriminators. The first generator transforms a real image from domain X to domain Y , and the second generator takes the generated domain Y image, and transforms it back to domain X again. The cycle consistency loss then evaluates whether the reconstructed synthetic domain

X image correctly matches the original real domain X image. CycleGAN's loss function is formed from the cycle consistency losses from both the forwards and backwards cycles, as well as the adversarial losses contributed by the discriminators, which assess the realism of the synthetic images in both domains (Zhu et al., 2017).

Rui et al. (2021) and Luccioni et al. (2021) used GAN architectures, including the CycleGAN (Schmidt et al., 2019), to create engaging, realistic 'street-view' images of floods. Lütjens et al. (2024) and Goswami et al. (2022) focused on generating accurate and photorealistic aerial flood imagery, but their models relied on pre-segmented labelled masks to define the correct floodwater locations, and did not use the GAN model to make the predictions.

Hama et al. (2021), Schmidt et al. (2021) and Goswami et al. (2022) all concluded that training a model on labelled flood masks greatly improves the generated images of floods. However, obtaining labelled data is extremely time consuming and expensive. 'Attention' mechanisms represent an alternative approach to achieve a similar result, without the need for the labelled mask data. AttentionGAN (Tang et al., 2021) trains subnetworks to produce attention masks and content (image transformation) masks; the attention masks are then applied to the content masks, which are then added to the original image pixel-wise, hence only the area of the image indicated by the attention mask is transformed. AttentionGAN also uses the same unsupervised learning approach and cycle consistency loss function as CycleGAN.

3 METHODOLOGY

3.1 Dataset

The study area consists of three flood events in the USA, caused by Hurricane Harvey in the city of Houston in 2017, Hurricane Florence in North Carolina near the Northeast Cape Fear river in 2018, and heavy rainfall in the Midwestern United States along the Arkansas river in 2019. This set of events comprises a variety of geographical regions, from the countryside to urban settlements, as well as a range of different precipitation intensities.

The selection of features for the dataset was determined by the most impactful factors on the likelihood of fluvial (river) and pluvial (rainfall) flooding occurring in a particular area. The formal definition of a flood is "a body of water which rises to overflow land which is not normally submerged" (Ward, 1978), hence the quantity of runoff is a function of

rainfall intensity and the infiltration capacity of the ground, which in itself is modified by the characteristics of the surface (Bolt et al., 2013). Also important are the attributes of the overall catchment area, which determines the dynamics of the water flow. Key features therefore include the shape, slope, aspect, altitude, climate, geology, soil type, infiltration, vegetation cover, and channel factors of a region (Smith and Ward, 1998).

To assist the model in making flood predictions, five factors were selected as model inputs. These include a pre-flooding optical satellite image, a digital elevation model (DEM), representation of flow accumulation, distance to nearby rivers, and a cartographical map. The data sources and attributions are provided in section 6. Having access to high-resolution, optical satellite imagery of the region before the flood event had occurred, was essential for the model to be able to generate photorealistic post-flooding aerial images. A DEM is critical for accurate flood forecasting, as it encapsulates key information on topographical features such as elevation, slope, aspect, roughness and curvature (Mujumdar and Kumar, 2012). Land use, land cover and imperviousness data were not directly input into the model, because the openly available datasets were of poor resolution, and were typically created by automated models classifying satellite imagery (Kontgis, 2021), thus it was determined to be feasible for the GAN to extract this information itself from the pre-flooding images. The cartographical map input also provides some indication of land use, as well as the locations of roads and buildings. It was also important to input into the model data that described the wider context of the study area; the distance to nearby rivers is a basic but crude method to achieve this, whereas the flow accumulation provides a detailed deterministic prediction of which cells drain into other downslope cells, hence determining the catchment areas (Mark, 1983).

Optical Satellite Images: The 'xBD' dataset contains paired satellite images captured before and after flood events, in the exact same locations. The images are taken from the Maxar Open Data program, and were made by the GeoEye-1, WorldView-2 and WorldView-3 satellites. Each image has 1024×1024 pixels in a three-band RGB format, with a resolution of approximately 0.5 metres/pixel. The pre-flooding satellite image is input into the model, whereas the paired post-flooding image functions as the ground truth target, used to evaluate the predictions made by the model.

Digital Elevation Model: The USGS 3D Elevation Program DEM, which has a $1/3$ arc-second (10 metre) resolution, describes the topography of the

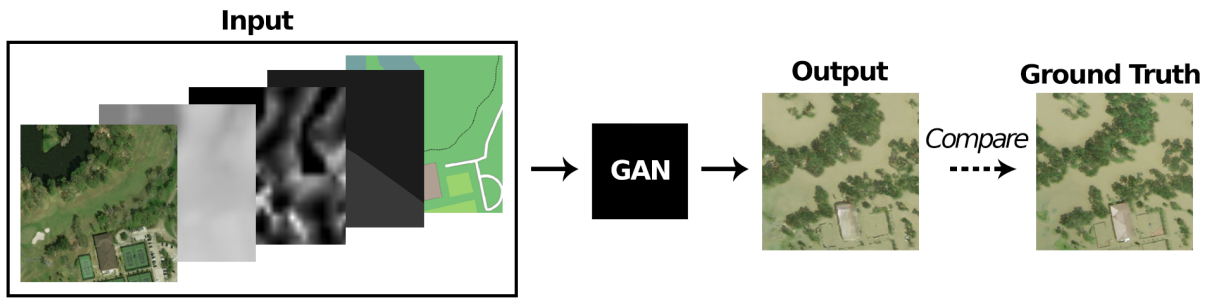


Figure 1: A single image stack comprised of a pre-flooding satellite image, DEM, flow accumulation, distance to rivers, and map, is input into the GAN model. The GAN outputs a synthetic, predicted post-flooding satellite image, which can be evaluated by comparing it to the real ground truth post-flooding image.

study areas. The elevations in this DEM represent the topographic bare-earth surface, hence excluding objects like trees and buildings, which are instead represented in the pre-flooding satellite images.

Flow Accumulation: We calculated the flow accumulation from DEM rasters that each covered approximately 2000km^2 , containing a study area and its surrounding areas. It was firstly necessary to fill the sinks, to prevent artificial depressions from trapping water and preventing continuous flow (Wang and Liu, 2006). Next, the flow direction and subsequent flow accumulation was calculated using the multiple flow direction algorithm (Wolock and McCabe Jr, 1995), in the SAGA GIS tool (Conrad et al., 2015). Rather than applying a threshold to the output, logarithm base 10 was taken of the flow accumulation values, in order to retain the most detail for the model.

Distance to Rivers: The QGIS software was used to visualise buffer distances (at 0.5km intervals) to all major rivers and waterways, as classified by the Open Street Map.

Map: OpenStreetMap data was downloaded from Planet OSM and processed using the Osmium tool. The Maperitive software was then used to apply a custom ruleset to the maps’ appearance, removing all of the text, and enhancing the clarity of the land use types.

Input Stack: All of the inputs were georeferenced, and projected to the same WGS 84 coordinate system that the satellite images originally used. The satellite images did not compose a continuous image of the study area, but rather captured separate, relevant locations. As such, the four input features of the DEM, flow accumulation, distance to rivers and map (henceforth referred to collectively as the “topographical factors”) were cropped to create individual image stacks. Figure 1 depicts how a single sample image stack is input into the model, which then outputs a synthetic post-flooding satellite image, that can hence be compared to the ground truth post-flooding image. Each image set contains a single pre-flooding

satellite image and its associated topographical factors concatenated as tensor channels, with each pixel representing the same $0.5\text{m} \times 0.5\text{m}$ geographical area in all of the channels. All of the input stacks had their alignment manually reviewed and adjusted to correct for any errors or skewness.

The data were transformed by resizing each 1024×1024 pixel image to 512×512 pixels using bicubic resampling, and then cropped into 4 separate images of 256×256 pixels each, because the model architectures worked most optimally with exactly 256×256 pixel images. Although many techniques have been developed to better handle higher resolution imagery (Karras et al., 2017), (Wang et al., 2018), the inclusion of these would have added yet another layer of complexity to an already highly complex task. The only augmentation applied to the data was a horizontal flip; the images were not rotated or flipped vertically, as such a transformation would be unrealistic for this domain.

The final dataset contained a total of 5736 image stacks, of which 2680 corresponded to Hurricane Harvey, 1880 image stacks of Hurricane Florence, and 1176 image stacks of the Midwest floods. The data were split into 80% in the training dataset, 10% in the validation set, and 10% in the test set. The splits were stratified by the flood events, so that each flood was represented proportionally in each set. All of the following results, metrics and sample generated images are taken from the test dataset, which was held-out during the model training process.

3.2 Models

Although the current state-of-the-art in generative networks tends towards the vision transformer (Dosovitskiy et al., 2020) and diffusion model (Ho et al., 2020) architectures, GANs are in comparison less computationally intensive and require less data to train. They are also much faster at inference time, requiring only one forward pass, as opposed to the

multiple de-noising steps of a diffusion model.

GAN architectures were therefore used to generate the aerial flood prediction imagery in this work, as they best fulfilled the requirement of a data-driven model that could be faster than process-based models, and could train efficiently on the comparably small size of the flood dataset. The performance of three existing model architectures was compared; Pix2Pix, representing the supervised approach, CycleGAN, representing the unsupervised approach, and AttentionGAN, representing the guided approach. The attention mechanism in the AttentionGAN architecture is very suited for flood prediction, because the attention masks can identify the location for the flooding, and the content masks can generate floodwater visualisations. However, AttentionGAN's unsupervised learning strategy does not make effective use of the paired pre- and post-flooding images available in the dataset. A new model called 'PairedAttention' was hence created, which utilises the architecture of AttentionGAN, but the supervised training approach and loss function of Pix2Pix.

The training procedures and model architectures were replicated from their original papers, with the exception that all of the generators and discriminators in the CycleGAN and AttentionGAN architectures were modified to input the additional topographical factors. A comprehensive hyperparameter tuning analysis determined the following optimal hyperparameters for the models: The model weights were initialised from a Gaussian distribution with mean 0 and standard deviation 0.02. The optimizer used the Adam algorithm (Kingma, 2014) with a learning rate of 0.0002 and beta decay rates of 0.5 and 0.999. The learning rate scheduler maintained the initial learning rate for the first half of epochs, then reduced the learning rate linearly for the second half of epochs. Each model trained for 200 epochs, as performance was found to plateau beyond this point. The batch size was set to 1.

3.3 Evaluation Metrics

Because GANs generate photorealistic imagery, evaluating their outputs is much more difficult than evaluating the performance of traditional classifier neural networks (Betzalet et al., 2022). Determining the most appropriate metrics is therefore key to any evaluation of GANs, and is highly dependent on the characteristics of the data and the goal of the research. For the task of generating aerial flood prediction imagery, there are three core objectives: achieving a high level of photorealism, predictive performance, and efficiency.

All three of these goals are critical to making a good flood prediction. The images must be photorealistic in order to create a convincing representation that is useful for enhancing situational awareness. The photorealism metrics that were used to evaluate the GANs are: PSNR, SSIM, MS-SSIM, and LPIPS (Arabboev et al., 2024). These metrics compare how close a generated image is to a reference image from their signal-to-noise ratio, by the similarity of their structure, luminance and contrast, on multiple resolution scales, and by their feature maps, respectively.

As well as generating photorealistic imagery, it is also important that the GANs accurately position the floodwaters within the generated images. The performance metrics (MSE, accuracy, F1, precision and recall scores) hence evaluate the predictive power of the models, by comparing binary flood masks extracted from both a generated image and its corresponding ground truth post-flooding image. The flood masks were produced by a trained flood segmentation model, which could identify the flooded pixels in the images with an accuracy (evaluated on a held-out test set of flood masks) of 94.9% and MSE of 0.051. Although the additional degree of error induced by the segmentation model should be taken into consideration when assessing the absolute performance of the GANs, this error is small, and relative comparisons can still be dependably made.

Finally, one of the main advantages of using a data-driven approach over traditional flood forecasting methods, is that it should be faster and use fewer computational and human resources. The two measurements used to evaluate the efficiency of the models were the training time (the total time needed to train the model) and the inference time (the time needed for the model to make a flood prediction and generate an image).

Nevertheless, there are limitations inherent to all of the evaluation metrics. For example, if a flood prediction is off by simply one pixel, it would be penalised by the performance metrics, even though to a human eye there may be no visible difference. Chlis (2019) argues that the most reliable way to evaluate the performance of a GAN is therefore to manually inspect the quality of the generated images. As such, sample images are also included throughout the presentation of the results.

4 RESULTS

4.1 Model Architectures Performance

Table 1 presents the performance of the four model architectures PairedAttention, Pix2Pix, AttentionGAN, and CycleGAN. The two models that use a supervised training approach (PairedAttention and Pix2Pix) achieved better results than the two cycle-based models (AttentionGAN and CycleGAN) on all of the photorealism and performance metrics, with PairedAttention performing the best overall. Figure 2 displays sample images generated by each of the model architectures, and demonstrates that the PairedAttention model is able to consistently generate post-flooding imagery that is not only photorealistic, but makes accurate flood predictions.

There is no strong pattern depicted in figure 3(b), suggesting that the GAN is capable of making accurate predictions at varying flood amounts, from 0% to almost 100% of the image being flooded. Figure 3(c) plots the distribution of the MSE for only instances where 1% or less of the ground truth image contains flooded pixels. The histogram is extremely right skew, indicating that the model is able to correctly identify when the area depicted in an image should have little to no flooding. The models were also able to successfully learn the key features of river floods, including how rivers break their banks and flood adjacent low-lying areas (as in figure 2(c)).

The cycle-consistency terms in the loss functions of the cycle-based models assisted these models in recreating the fine details in the images, such as the clean lines of buildings. However, AttentionGAN and CycleGAN have substantially lower recall scores than PairedAttention and Pix2Pix, corresponding to failures to generate flooding in areas of the images that should have been flooded. This error occurs because cycle-based methods assume a one-to-one mapping instead of a many-to-one mapping (Schmidt et al., 2019). When the first generator transforms many different features - such as roads, grass, or mud - all into floodwaters, then the second generator is unable to determine which feature the floodwaters should be transformed back into, when attempting to reconstruct the original image (Luccioni et al., 2021). In order to improve the cycle consistency loss, the cycle-based methods therefore tend to avoid modifying and hiding as much of the original pre-flooding image as possible, as can be seen in figure 2(b), where AttentionGAN and CycleGAN do not place floodwaters over the lake, so that the lake can be more easily reconstructed later.

Due to the two generators and two discriminators

required by the AttentionGAN and CycleGAN models, they also use the most computational resources to train, whereas the PairedAttention and Pix2Pix models were trained in approximately a third of the time (table 1). After the models have been trained just once on the historical data however, they can be applied to any new unseen areas, all taking only approximately 0.004 seconds to generate an image prediction. In comparison, as reported in research comparing data-driven models to traditional methods, Hofmann and Schüttrumpf’s (2021) process-based hydrodynamic model took 7 hours to perform a detailed simulation, whereas Kabir et al. (2020) and do Lago et al. (2023) required 1.3 hours and 1.5 hours to simulate a single event with a 2D hydrodynamic and hydraulic model respectively. The advantage of the data-driven approach is evident, as once trained, the GANs can be applied to a variety of previously unseen images with extremely fast inference times, whereas the process-based models must always rerun their simulations.

4.2 Impact of Topographical Factors

An ablation study was carried out in order to ascertain the impact of the different topographical factors on the performance of the GANs. Specifically, the pre-flooding satellite image was always input into the model, (as it is a requisite to be able to generate a photorealistic image), and in each experiment just one of the topographical factors was additionally input into the model. Table 2 presents the results from models using the PairedAttention architecture, the first of which utilised all of the input factors as a baseline. The remaining models input a combination of only the pre-flooding satellite image and the DEM, or only the satellite image and the flow accumulation representation, etc. The final model, labelled ‘None’, was input with the pre-flooding satellite image alone.

With regards to the photorealism metrics, all of the models performed similarly. This is not unexpected, as the realism of the generated post-flooding image depends predominantly on a transformation of the pre-flooding image, which was input to all of the models. In all of the performance metrics, ‘All’ ‘DEM’ and ‘Flow accumulation’ outperformed ‘Distance to rivers’ ‘Map’ and ‘None’. This result indicates that the elevation values (as the flow accumulation was also derived from the DEM) are the most important to making accurate flood predictions. This is also evident in figures 4(a) and (b), where the models without access to elevation information often either under-flood or over-flood areas in the images. ‘Distance to rivers’ and ‘Map’ did not perform better than

Table 1: The results from evaluating the performance of the four model architectures on predicting the flood events.

Model Architectures	Photorealism Metrics				Performance Metrics					Efficiency Metrics	
	PSNR $\uparrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	LPIPS $\downarrow$	MSE $\downarrow$	Accuracy $\uparrow$	F1 $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	Inference $\downarrow$ (s)	Training $\downarrow$ (h)
PairedAttention	22.681	0.573	0.683	0.228	0.123	0.877	0.804	0.824	0.786	0.004	15.3
Pix2Pix	22.442	0.557	0.644	0.250	0.134	0.866	0.781	0.824	0.742	0.004	11.3
AttentionGAN	21.394	0.524	0.581	0.261	0.157	0.843	0.738	0.796	0.689	0.004	43.6
CycleGAN	20.841	0.510	0.578	0.269	0.156	0.844	0.735	0.808	0.673	0.004	38.1

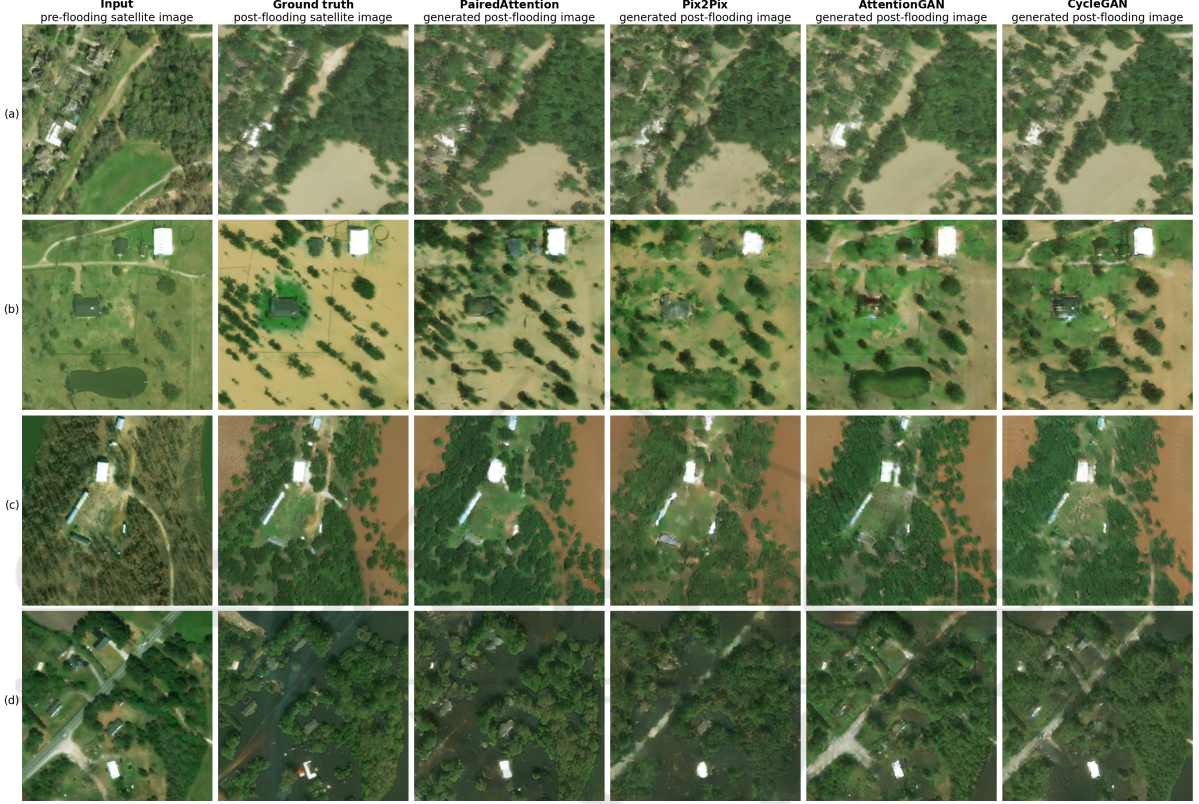


Figure 2: Sample generated images of flood events from each of the four model architectures.

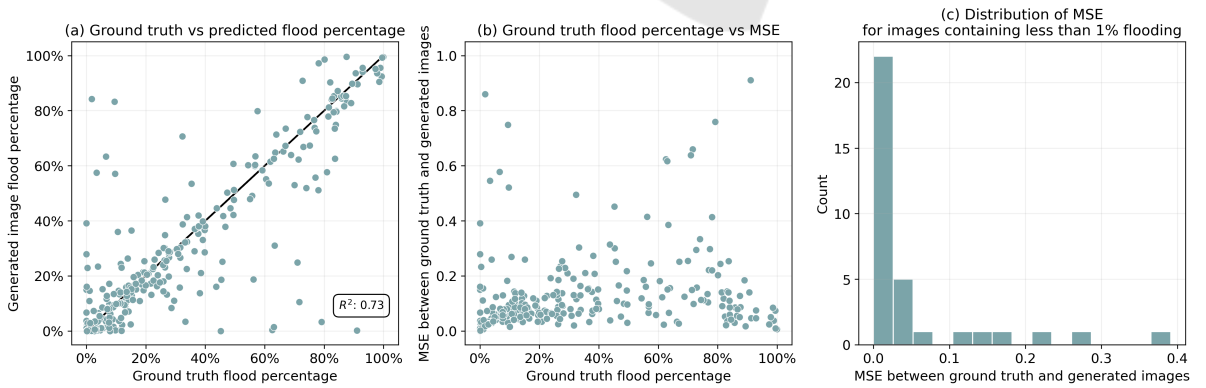


Figure 3: The performance of the PairedAttention GAN at different percentages of flooding amounts in the post-flood images.

Table 2: The results from the ablation study, comparing the models using different topographical factors as inputs.

Input Topography	Photorealism Metrics				Performance Metrics				
	PSNR $\uparrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	LPIPS $\downarrow$	MSE $\downarrow$	Accuracy $\uparrow$	F1 $\uparrow$	Precision $\uparrow$	Recall $\uparrow$
All	22.681	0.573	0.683	0.228	0.123	0.877	0.804	0.824	0.786
DEM	23.134	0.585	0.701	0.220	0.116	0.884	0.812	0.844	0.782
Flow accumulation	23.126	0.590	0.701	0.222	0.126	0.874	0.798	0.822	0.776
Distance to rivers	22.913	0.584	0.693	0.229	0.141	0.859	0.770	0.809	0.735
Map	22.834	0.597	0.691	0.269	0.140	0.860	0.772	0.811	0.736
None	22.900	0.583	0.690	0.231	0.140	0.860	0.772	0.812	0.736

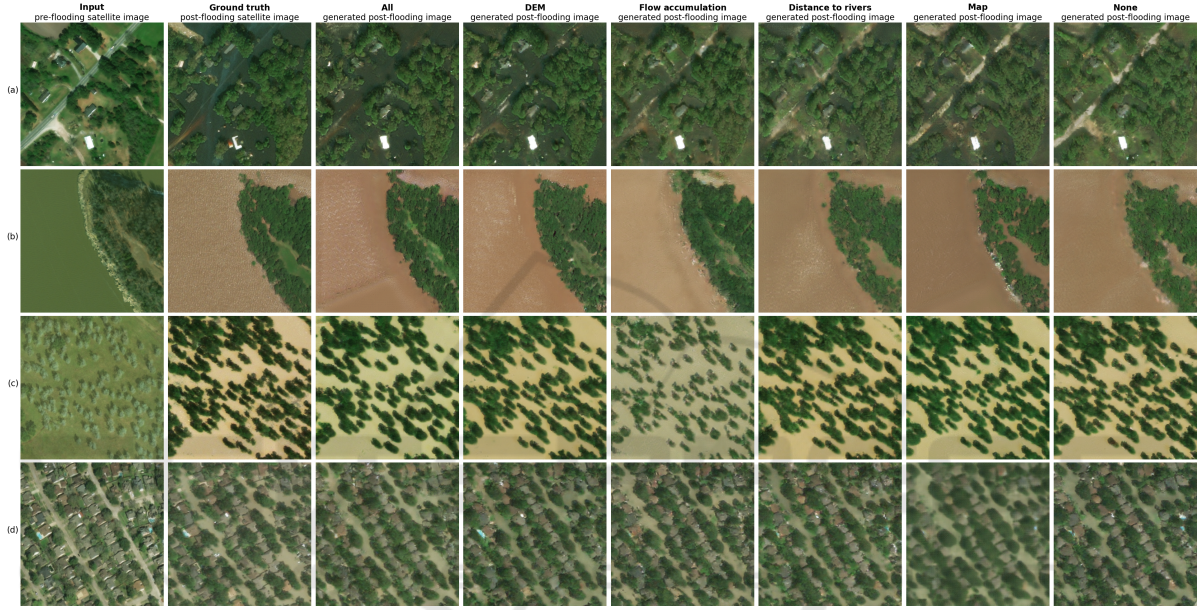


Figure 4: Sample generated images of flood events from models trained on different topographical factors. For example, ‘DEM’ means the model was input only with the pre-flooding satellite image and the ‘DEM’ factor.

‘None’ - it is likely that the descriptions of land use provided by the map can already be learned by the model from the pre-flooding satellite image, and the ‘distance to rivers’ is too simplistic of a measure in comparison to the more detailed flow accumulation.

Although access to elevation values consistently improved the performance of the models, the absolute differences between the metric scores is still fairly small, and the model can achieve an accuracy of 86% even when no topographical factors are input. There are two explanations which could potentially account for this result. Firstly, the model may be capable of inferring more information than expected from the pre-flooding satellite image alone, including approximate elevation values. Secondly, an investigation of the full test dataset revealed a large number of post-flooding satellite images where simply the entire ground area of the image is flooded, (as in figures 4(c) and (d)), and thus a detailed analysis of the topography is unnecessary for the model to make an accurate flood

prediction. Nevertheless, in more complex images for which the topography is relevant to the positioning of the floodwaters, having access to the elevation values makes a significant difference to the performance of the model, such as in figures 4(a) and (b). However, because the contents of the dataset instead tends to be biased towards images like 4(c) and (d), the metrics therefore do not properly reflect the importance of the DEM.

4.3 Evaluating Model Generalisability

Although the test dataset contains images of geographical areas never before seen by the model during training, these images are still from the same flood events that the model previously trained on. Thus to examine the true generalisability of the GANs, a PairedAttention model was trained on images from Hurricane Harvey and Hurricane Florence, and evaluated on images of the Midwest floods. The general-

Table 3: The generalisability study results: The model was trained on data from Hurricane Harvey and Hurricane Florence, and then evaluated on unseen data from the Midwest floods.

Photorealism Metrics				Performance Metrics				
PSNR $\uparrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	LPIPS $\downarrow$	MSE $\downarrow$	Accuracy $\uparrow$	F1 $\uparrow$	Precision $\uparrow$	Recall $\uparrow$
19.843	0.497	0.536	0.373	0.220	0.780	0.655	0.887	0.519

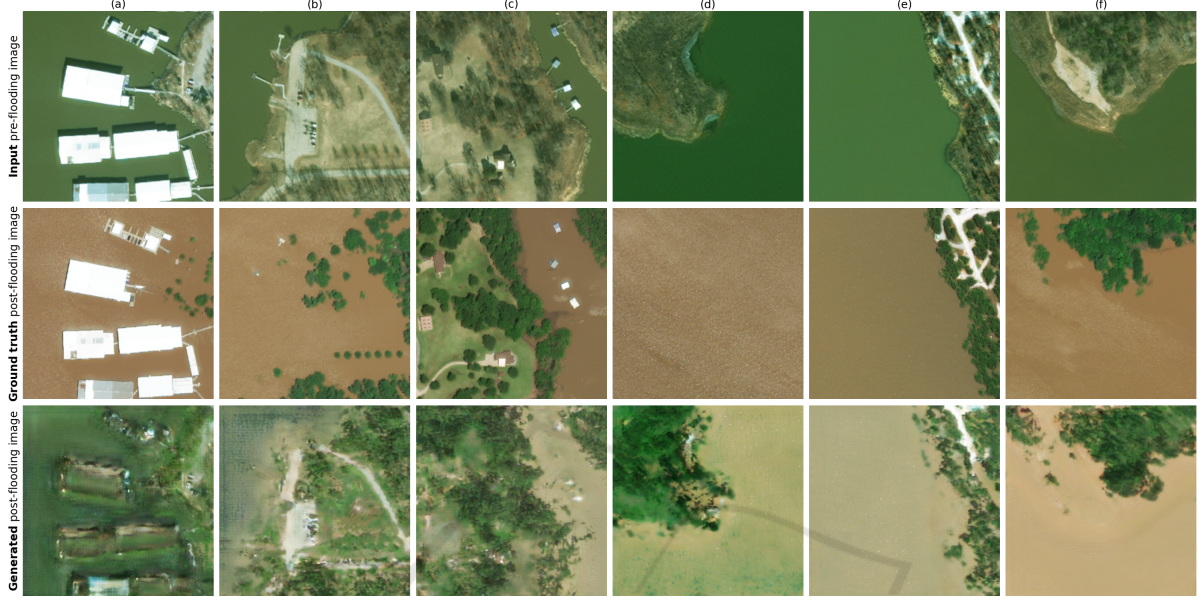


Figure 5: Sample generated images of the Midwest flood event, produced by a model trained only on the Hurricane Harvey and Hurricane Florence flood events.



Figure 6: The changes to the generated images after different modifications have been made to a pre-flooding satellite image of the Hurricane Florence flood event, such as editing in more trees (a) or increasing the elevation values of the road (g).

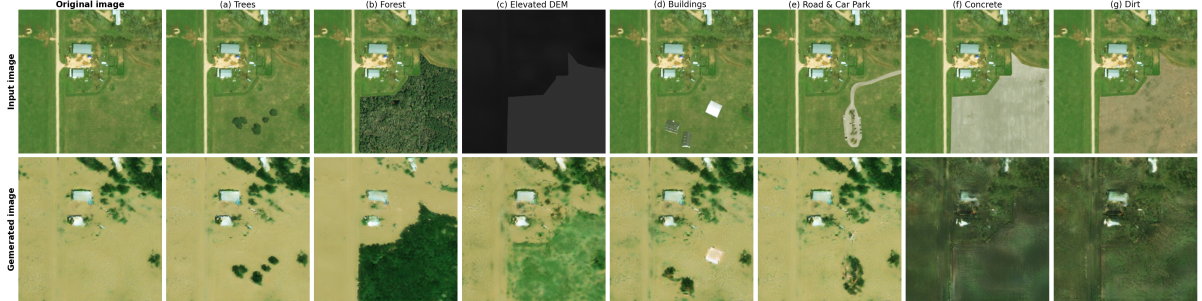


Figure 7: The changes to the generated images after different modifications have been made to a pre-flooding satellite image of the Hurricane Harvey flood event.

isability study results are presented in table 3.

The photorealism metrics are fairly poor, because the model had not learned how to appropriately transform the features and objects unique to the Midwest floods, thus creating artifacts and blurry sections in the generated images, as can be seen in figures 5(a), (b) and (c). The model was also unable to predict the correct colour of the floodwaters (dark brown) instead replicating the yellow and blue appearances of the waters from Hurricane Harvey and Hurricane Florence.

The performance metrics also reveal the weak predictive power of the model, with an F1 score of 0.66 and a low recall score of 0.52, indicating that the model significantly underpredicts flooding (as in figure 5(d)) even though the Midwest floods experienced less total rainfall (45-55cm) than the Hurricane Harvey flood event that the model had trained on (75-100cm). The representations of the core features of each flood event, such as the associated amount of precipitation and corresponding runoff, are therefore entangled with each of the flood events. The model has overfit to each flood event rather than only having learned general principles regarding the contributions of different features to the likelihood of flooding.

Nevertheless, there is evidence that the model has learned some generalisable principles. The appearance of the post-flooding satellite images in figure 5 possess a modicum of coherence, with recognisable recreations of trees, buildings and roads. The model has also learned some of the characteristics of fluvial flooding - likely from the Hurricane Florence data - as it is able to accurately flood the banks of the river in figures 5(e) and (f).

The capability of the model to generalise to new features in the images was also investigated, in conjunction with an analysis of the explainability of the model's behaviour. Because a GAN is a black-box model, it can be interpreted only via post-hoc methods; by changing the input to the model, and evaluating the subsequent change to its outputs (Ribeiro et al., 2016). Figure 6 demonstrates how manual alterations to a pre-flooding satellite image of Hurricane Florence change the post-flooding image generated by the GAN. The generated images show how the model has not only learned to classify different objects, but that it understands how floodwaters physically interact with each of them, flowing around trees and buildings (figures 6(a), (b) and (c)) but over cropland (figure 6(d)). The model also demonstrates consistency, as the flooding elsewhere in the image is unchanged when an independent element of the image is altered. In figure 6(g), the elevation of the road was increased by modifying the values of the DEM, and the GAN subsequently no longer floods the road, in-

dicating that it has learned the relationship between elevation and flooding. The model thus utilises features from both the pre-flooding satellite image and the topographical factors to make flood predictions.

Figure 7 depicts similar manual alterations to a pre-flooding satellite image of Hurricane Harvey. In figures 7(a) and (b) the model is able to correctly identify and avoid covering the trees with floodwaters. Increasing the elevation values of the DEM (figure 7(c)) results in the model appropriately reducing the amount of flooding within the modified area, suggesting that the GAN has learned a general function connecting elevation and floods, that is applicable to different contexts. However, when the model encounters an unfamiliar combination of objects and settings, such as a car park or residential housing in a rural region, it does not know how to handle them correctly, instead transforming the objects into trees (figures 7(d) and (e)). When a large surface of concrete or dirt is added, (figures 7(f) and (g)), the model floods the entire image with the blue floodwaters typically associated with the Hurricane Florence flood event. This outcome suggests that the GAN does not contain knowledge of impervious concrete surfaces that is disentangled from the flood events; rather the model may use cues in the form of particular colours and textures within the image in order to determine the flood event that the image likely depicts, and accordingly makes assumptions regarding the probable locations and amounts of flooding.

5 DISCUSSION

With regards to the research questions, the results can be summarised:

How accurately can GANs generate aerial flood prediction imagery, and which GAN architecture generates the most accurate flood predictions? The PairedAttention GAN architecture had the best efficiency, photorealism and performance metrics, with an accuracy of 88%, F1 score of 0.80, precision of 0.82, and recall of 0.79. Overall, the supervised training approach on paired images was more effective and trained more quickly than the cycle-based approach.

Which topographic factors most improve the accuracy of the flood predictions? The ablation study revealed that the elevation values were the most significant factor for improving flood predictions. Although the metrics suggested that the impact of the DEM on model performance was limited, these results were attenuated by the GAN's capability to infer information from the pre-flooding satellite image alone, as well as the dataset's bias towards images for

which the knowledge of detailed elevation values was unnecessary to make an accurate flood prediction.

How accurately can the GANs generalise to new flood events and features? Manual alterations to pre-flooding satellite images demonstrated that the GAN model has learned how floodwaters physically interact with different types of objects in an image, as well as the connection between flooding and elevation values, providing some explanation for the model's decisions when generating a post-flooding image. However, the model sometimes acted unexpectedly when handling out-of-distribution inputs. Although there were indications that the GAN had learned some general principles regarding flooding, overall, the model performed poorly when applied to a previously unseen flood event. These results suggest that the core features of each flood event are highly entangled within the model, rather than the model having learned a generalisable function.

Unlike hydraulic or hydrodynamic models, a GAN is not only learning to make flood predictions, but also to generate photorealistic imagery. The GAN's training process thus incorporates both adversarial and L1 loss terms in the loss function, which respectively regulate the overall photorealism of the generated image, and the similarity of the entire image to the ground truth image. Consequently, optimising the accuracy of the flood predictions is only a small part of the GAN's objective, and is a part that is only ever learned indirectly.

This training approach also resulted in a highly entangled representation of concepts within the model. In a disentangled data representation, individual factors (such as the amount of precipitation) are isolated and captured by separate, distinct elements of the model's representations, and hence can be varied independently (Bengio et al., 2013). However the flood prediction GANs had strongly overfit to the flood events that they were trained on, with the flood factors inextricable from the events themselves.

Future work should therefore focus on improving the generalisability of flood prediction models. A generalisable data-driven model also amplifies its advantages over numerical process-based models, as it can quickly generate predictions for a wider variety of areas and scenarios. The model thus needs to learn disentangled data representations, which can be achieved by adapting the model architecture, loss functions and the training dataset (Wang et al., 2022). Isolating the independent influence of factors such as precipitation would allow the model to dynamically condition the generated flood predictions on input weather forecasts. Due to the significantly faster inference times of data-driven models in comparison to

traditional physically-based models, a fully disentangled model could even be extended to flood risk management applications, by revealing how flood predictions change when the pre-flooding input image is modified, such as through the addition of flood relief channels or barriers (Pender and Faulkner, 2010). A GAN model could also be utilised in conjunction with traditional flood inundation maps that reduce subjectivity, with consideration of the described limitations of the flood prediction GAN, in order to avoid false alarms during operational usage.

The performance of the GANs was evaluated by using a separate segmentation model to derive a binary flood mask from the photorealistic images. The accuracy of the flood predictions could hence be improved by additionally training models on inundation or water depth labels, which can be estimated from the combination of flooded area boundaries and digital elevation models (Poterek et al., 2025).

6 CONCLUSION

Floods are an extremely deadly natural hazard, and flood forecasting is a critical component to pre-flooding preparedness. A data-driven approach to flood forecasting provides faster, less computationally expensive predictions than traditional process-based models. GANs are capable of generating photorealistic images of flood predictions, that can hence improve situational readiness for post-flooding interventions.

This work developed a successful proof-of-concept GAN based on the PairedAttention architecture, that was capable of both generating photorealistic aerial flood prediction imagery and making accurate flood predictions for the flood events that it was trained on, achieving 88% accuracy and an F1 score of 0.80. The model architecture utilised a supervised training approach on paired images (aligned sets of pre- and post-flooding satellite images with topographical factors) in combination with attention mask and content mask subnetworks. An ablation study determined that the elevation values provided by the DEM was the most important factor to improving predictive performance.

However, the model struggled to generalise when applied to previously unseen flood events and out-of-distribution features. Nevertheless, it demonstrated knowledge of general principles connecting flooding and topography, and future work to develop new modelling approaches that learn disentangled data representations, could improve the effectiveness of data-driven models for flood prediction even further.

ACKNOWLEDGEMENTS

The code was implemented in Python 3.10 and PyTorch 2.0.1, and the model training was executed on a single NVIDIA RTX A6000 GPU with 48GB RAM. The code is open source under the MIT licence, and can be found on GitHub at <https://github.com/Natasha-R/Flood-Prediction-GAN>. The pre-processed datasets and corresponding metadata are available under the Creative Commons Attribution-Noncommercial-Sharealike 4.0 International licence (CC BY-NC-SA 4.0) on Zenodo at <https://zenodo.org/doi/10.5281/zenodo.13366121>.

The xBD dataset is released under the Creative Commons Attribution-Noncommercial-Sharealike 4.0 International licence (CC BY-NC-SA 4.0). The data are sourced from the Maxar Open Data Program: <https://www.maxar.com/open-data/>.

The USGS 3D Elevation Program DEM is released by the U.S. Geological Survey, 2023, 1/3rd arc-second Digital Elevation Models (DEMs) - USGS National Map 3DEP Downloadable Data Collection. Distributed by OpenTopography. <https://doi.org/10.5069/G98K778D>. All 3DEP products are public domain.

The OpenStreetMap (OSM) data is distributed under the Open Database License (ODbL). <https://www.openstreetmap.org/copyright>.

REFERENCES

- Ahmadlou, M., Al-Fugara, A., Al-Shabeeb, A. R., Arora, A., Al-Adamat, R., Pham, Q. B., Al-Ansari, N., Linh, N. T. T., and Sajedi, H. (2021). Flood susceptibility mapping and assessment using a novel deep learning model combining multilayer perceptron and autoencoder neural networks. *Journal of Flood Risk Management*, 14(1):e12683.
- Allaby, M. (2014). *Floods*. Infobase Publishing.
- Arabboev, M., Begmatov, S., Rikhsivoev, M., Nosirov, K., and Saydiakbarov, S. (2024). A comprehensive review of image super-resolution metrics: classical and ai-based approaches. *Acta IMEKO*, 13(1):1–8.
- Arduino, G., Reggiani, P., and Todini, E. (2005). Recent advances in flood forecasting and flood risk assessment. *Hydrology and Earth System Sciences*, 9(4):280–284.
- Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.
- Betzalel, E., Penso, C., Navon, A., and Fetaya, E. (2022). A study on the evaluation of generative models. *arXiv preprint arXiv:2206.10935*.
- Bolt, B. A., Horn, W., MacDonald, G. A., and Scott, R. (2013). *Geological Hazards*. Springer Science & Business Media.
- Chlis, N. K. (2019). Image-to-image translation with a pix2pix gan and keras. https://nchlis.github.io/2019_11_22/page.html.
- Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. (2015). System for automated geoscientific analyses (saga) v. 2.1. 4. *Geoscientific model development*, 8(7):1991–2007.
- CRED (2015). The human cost of weather-related disasters 1995-2015. *United Nations Office for Disaster Risk Reduction*.
- Devia, G. K., Ganasri, B. P., and Dwarakish, G. S. (2015). A review on hydrological models. *Aquatic procedia*, 4:1001–1007.
- do Lago, C. A., Giacomoni, M. H., Bentivoglio, R., Taormina, R., Junior, M. N. G., and Mendiondo, E. M. (2023). Generalizing rapid flood predictions to unseen urban catchments with conditional generative adversarial networks. *Journal of Hydrology*, 618:129276.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Goswami, S., Verma, S., Gupta, K., and Gupta, S. (2022). Floodnet-to-floodgan: Generating flood scenes in aerial images. *HAL Open Science*.
- Guo, Z., Leita, J. P., Simões, N. E., and Moosavi, V. (2021). Data-driven flood emulation: Speeding up urban flood predictions by deep convolutional neural networks. *Journal of Flood Risk Management*, 14(1):e12684.
- Hama, K., Siritanawan, P., and Kazunori, K. (2021). Synthesis of localized flooding disaster scenes using conditional generative adversarial network. In *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*, pages 393–398. IEEE.
- Henonin, J., Russo, B., Mark, O., and Gourbesville, P. (2013). Real-time urban flood forecasting and modelling—a state of the art. *Journal of Hydroinformatics*, 15(3):717–736.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Hofmann, J. and Schüttrumpf, H. (2021). Floodgan: Using deep adversarial learning to predict pluvial flooding in real time. *Water*, 13(16):2255.
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.
- Jain, S. K., Mani, P., Jain, S. K., Prakash, P., Singh, V. P., Tullos, D., Kumar, S., Agarwal, S. P., and Dimri, A. P.

- (2018). A brief review of flood forecasting techniques and their applications. *International journal of river basin management*, 16(3):329–344.
- Kabir, S., Patidar, S., Xia, X., Liang, Q., Neal, J., and Pender, G. (2020). A deep convolutional neural network model for rapid prediction of fluvial flood inundation. *Journal of Hydrology*, 590:125481.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.
- Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kontgis, C. (2021). Mapping the world in unprecedented detail. <https://medium.com/impactobservatoryinc/mapping-the-world-in-unprecedented-detail-7c0513205b90>.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Liu, Z., Zhang, H., and Liang, Q. (2019). A coupled hydrological and hydrodynamic model for flood simulation. *Hydrology Research*, 50(2):589–606.
- Luccioni, A., Schmidt, V., Vardanyan, V., and Bengio, Y. (2021). Using artificial intelligence to visualize the impacts of climate change. *IEEE Computer Graphics and Applications*, 41(1):8–14.
- Lütjens, B., Leshchinskiy, B., Boulais, O., Chishtie, F., Diaz-Rodriguez, N., Masson-Forsythe, M., Mata-Payerro, A., Requena-Mesa, C., Sankaranarayanan, A., Pina, A., et al. (2024). Generating physically-consistent satellite imagery for climate visualizations. *IEEE Transactions on Geoscience and Remote Sensing*.
- Mark, D. M. (1983). Automated detection of drainage networks from digital elevation models. In *Proceedings of Auto-Carto*, volume 6, pages 288–298.
- Mihon, D., Bacu, V., Rodila, D., Stefanut, T., Abbaspour, K., Rouholahnejad, E., and Gorgan, D. (2013). Grid based hydrologic model calibration and execution. *Advances in Intelligent Control Systems and Computer Science*, pages 279–293.
- Ming, X., Liang, Q., Xia, X., Li, D., and Fowler, H. J. (2020). Real-time flood forecasting based on a high-performance 2-d hydrodynamic model and numerical weather predictions. *Water Resources Research*, 56(7):e2019WR025583.
- Mosavi, A., Ozturk, P., and Chau, K.-w. (2018). Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536.
- Mujumdar, P. and Kumar, D. N. (2012). *Floods in a changing climate: hydrologic modeling*. Cambridge University Press.
- Pender, G. and Faulkner, H. (2010). *Flood Risk Science and Management*. John Wiley & Sons.
- Poterek, Q., Caretto, A., Braun, R., Clandillon, S., Huber, C., and Ceccato, P. (2025). Interpolated flood surface (inflos), a rapid and operational tool to estimate flood depths from earth observation data for emergency management. *Remote Sensing*, 17(2):329.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- Rui, X., Cao, Y., Yuan, X., Kang, Y., and Song, W. (2021). Disastergan: Generative adversarial networks for remote sensing disaster image generation. *Remote Sensing*, 13(21):4284.
- Schmidt, V., Luccioni, A., Muckavilli, S. K., Balasooriya, N., Sankaran, K., Chayes, J., and Bengio, Y. (2019). Visualizing the consequences of climate change using cycle-consistent adversarial networks. *arXiv preprint arXiv:1905.03709*.
- Schmidt, V., Luccioni, A. S., Teng, M., Zhang, T., Reynaud, A., Raghupathi, S., Cosne, G., Juraver, A., Vardanyan, V., Hernandez-Garcia, A., et al. (2021). Climategan: Raising climate change awareness by generating images of floods. *arXiv preprint arXiv:2110.02871*.
- Sivanpillai, R., Jacobs, K. M., Mattilio, C. M., and Piskorski, E. V. (2021). Rapid flood inundation mapping by differencing water indices from pre-and post-flood landsat images. *Frontiers of Earth Science*, 15:1–11.
- Smith, K. and Ward, R. (1998). *Floods: physical processes and human impacts*. John Wiley & Sons.
- Tang, H., Liu, H., Xu, D., Torr, P. H., and Sebe, N. (2021). Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE transactions on neural networks and learning systems*, 34(4):1972–1987.
- UNDP (2023). Climate change’s impact on coastal flooding to increase 5-times over this century, putting over 70 million people in the path of expanding floodplains, according to new undp and cil data. *United Nations Development Programme*.
- Wang, L. and Liu, H. (2006). An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *International Journal of Geographical Information Science*, 20(2):193–213.
- Wang, M., Wang, H., Xiao, J., and Liao, L. (2022). A review of disentangled representation learning for remote sensing data. *CAAI Artificial Intelligence Research*, 1(2).
- Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.
- Wang, Y., Fang, Z., Hong, H., and Peng, L. (2020). Flood susceptibility mapping using convolutional neural network frameworks. *Journal of Hydrology*, 582:124482.
- Ward, R. C. (1978). *Floods: a geographical perspective*. The Macmillan Press Ltd.
- Wolock, D. M. and McCabe Jr, G. J. (1995). Comparison of single and multiple flow direction algorithms for computing topographic parameters in topmodel. *Water Resources Research*, 31(5):1315–1324.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.