EARTH OBSERVATION FOUNDATION MODELS FOR REGION-SPECIFIC FLOOD SEGMENTATION

Helen Tamura-Wicks[*], **Geoffrey Dawson**[*], **Anne Jones & Paolo Fraccaro**
IBM Research Europe
The Hartree Centre STFC Laboratory, Sci-Tech Daresbury, Warrington, WA4 4AD, UK
`{helen.tamura-wicks, geoffrey.dawson}@ibm.com`

Andrew Taylor & Chris Dearden
Science and Technology Facilities Council
The Hartree Centre STFC Laboratory, Sci-Tech Daresbury, Warrington, WA4 4AD, UK

ABSTRACT

AI foundation models for earth observation are an important tool to inform and adapt to extreme weather events brought on by climate change. Here, we investigate the performance of these models for a region-specific task. We build upon the Prithvi-EO model, which uses optical imagery, and incorporate Synthetic Aperture Radar (SAR) imagery for UK and Ireland by both additional pretraining and directly fine tuning for regional flood segmentation. Incorporating SAR band imagery via either approach improved flood segmentation performance from 0.58 to 0.79 (by approximately 35%), suggesting that EOFMs can relatively easily be tuned to new locations and application-specific satellite bands.

1 INTRODUCTION

Earth Observation data is an invaluable resource for climate change action. As climate change exacerbates the intensity and frequency of weather-driven hazards, satellite-based impact monitoring is a critical tool used to inform the response: both to acute events such as floods (Fraccaro et al., 2022), and to longer term changes such as ecosystem response (Myers-Smith et al., 2020). Over recent years, Geospatial Foundation Models have emerged as an increasingly popular approach to developing AI applications for satellite data (Zhang et al., 2024). By pre-training at scale using self-supervised learning, and then fine-tuning for specific tasks, this approach aims to benefit from the sophisticated architecture of deep learning models whilst improving computational efficiency at the fine-tuning stage. A particular additional benefit for environmental applications, where labelled datasets are often difficult and expensive to produce, is the potential to fine-tune application models on relatively small amounts of labelled data.

Whilst evaluation and comparison of Earth Observation Foundation Models (EOFMs) using standard benchmark datasets is valuable and increasingly common (Lacoste et al., 2023; Dionelis et al., 2024), such metrics do not necessarily provide an indication of how such models may perform in real world applications for a given region of interest. Here, we focus on how additional pre-training of a foundation model with region-specific data might benefit downstream tasks, using the Prithvi-EO family of models (Jakubik et al., 2023; Szwarcman et al., 2024). We focus on the UK and Ireland (UKI) and pre-training to add both additional samples of the same bands for this region, and samples of new bands, in this case Synthetic Aperture Radar (SAR), as it was not included in the original model. We then fine tune for a climate change adaption task, flood segmentation for disaster relief, which should particularly benefit from the addition of SAR data which can see through clouds. We assess performance in terms of both prediction accuracy and efficiency of training. Our findings should be informative for potential users of EOFMs interested in real world climate change applications in a particular region [1].

[*] Equal contribution

[1] Material related to this study have been open sourced and can be found under Reproducibility

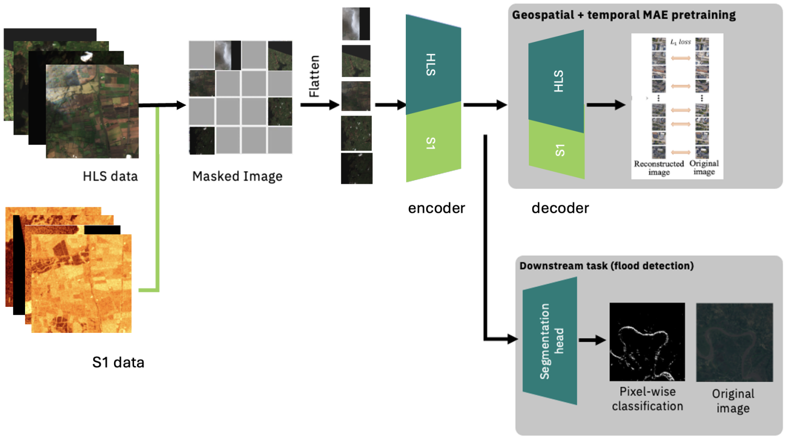

Figure 1: Model schematic of the Prithvi family of models. The light green parts indicate additional data and sections of the model which we built upon for the purposes of this study.

2 Methods

As a base model we used Prithvi-EO-1.0-100M (Jakubik et al., 2023): a transformer-based EOFM with 100 million parameters, pre-trained using a masked autoencoder (MAE) approach (He et al., 2021) to reconstruct masked images using an encoder-decoder architecture with a Vision Transformer (ViT) backbone. Full details of the architecture and pre-training process can be found in Jakubik et al. (2023). The model was pre-trained on Harmonized Landsat Sentinel-2 (HLS) data, which combines Landsat 8 and 9 with Sentinel-2 (S2) at a resolution of 30m (Claverie et al., 2018). The input image size was 224x224 and a total of 250,000 cloud-free samples covering the contiguous United States were used, each consisting of six bands: Red, Blue, Green, Narrow Infrared (NIR), Short-wave infrared 1 (SWIR 1) and Short-wave infrared 2 (SWIR 2) (Szwarcman et al., 2024). Here, we carried out additional pre-training of Prithvi-EO-1.0-100M, before fine-tuning for flood segmentation over UKI.

2.1 Additional pre-training

We performed additional pre-training using two extra datasets. Firstly, additional HLS data covering UKI from 2022 and 2023, using the same bands as Prithvi-EO-1.0-100M. Unlike in the original pre-training, we allowed images with up to 50% cloud coverage, to enable flood segmentation in cloudy conditions. Secondly, we included VV and VH bands of Synthetic Aperture Radar (SAR) backscatter σ_0 from the Sentinel-1 (S1) satellites. We normalised the backscatter using $10log(\sigma_0)$ and set pixels that fell below -35dB to -35dB and pixels which were larger than 10dB to 10dB, to remove specular reflections. The SAR images were re-sampled to the 30m resolution of HLS and added as new bands to the input, resulting in eight bands (Red, Blue, Green, NIR, SWIR 1, SWIR 2, VV and VH). To pre-train this model we started with the Prithvi-EO-1.0-100M weights, and initialised the two new SAR bands with the mean of the weight of the other channels. We then pre-trained the model using the same methodology as (Jakubik et al., 2023) with 15,448 training and 3862 validation samples. We will refer to the model that has gone through the above additional pre-training process as *granite-geospatial-uki*.

2.2 Fine-tuning

We curated an annotated flood imagery dataset for UKI using flood and water body extent maps from Copernicus Emergency Management Service (CEMS) (EC JRC, 2025) as labels. For the inputs we obtained orthorectified S1 and S2 imagery (Sinergise Solutions, 2025;Torres et al., 2012; Drusch et al., 2012) that fell within a two day window before and after the CEMS map dates, using the same bands as the pre-trained model. We also processed S2's scene classification maps to obtain a cloud mask map to be used as an additional band, resulting in a total of nine bands for fine tuning. Additional information on how we curated this dataset can be found in Appendix A.

We fine tuned three models on this curated dataset (Table1): a *baseline-S2* model which was fine tuned from Prithvi-EO-1.0-100M on S2 bands only, a second *baseline-S1-S2* model where we fine-tuned on S1 and S2 bands, and the *granite-geospatial-uki-flooddetection* model, which was fine tuned on the *granite-geospatial-uki* model with S2, S1 and cloud mask bands. All models consisted of a fully convolutional decoder and a classification head of two classes in addition to the Prithvi ViT backbone. Due to the relatively small training dataset size, we ran an ensemble of 10 runs for each model with varying seeds. For completeness, we also ran *baseline-S2-global*, which was *baseline-S2* fine tuning repeated using the newly released Prithvi-EO-2.0-300M, with results reported in Appendix D.

3 Results

Training efficiency between the models is compared in Figure 5 in Appendix C. The mIoU of the validation set against epochs shows that the *granite-geospatial-uki-flooddetection* model is able to achieve the *baseline-S2* model's best performance in fewer epochs. The *baseline-S1-S2* model also reaches similar levels of performance to the *granite-geospatial-uki-flooddetection* model. To compare flood segmentation accuracy, we evaluated mIoU and F1 scores on the whole test set (Table 1). Since this includes images with a range of cloud cover (see Figure 4 in Appendix B), we also evaluated performance for only virtually cloud-free images. For mIoU, including additional S1 bands in the *granite-geospatial-uki-flooddetection* model resulted in 21.1 and 4.1 percentage points improvement in model performance for the whole test set and cloud-free set, respectively, compared to *baseline-S2*. This shows that additional S1 bands benefited the model performance in general, but especially for cloud-covered areas. The mIoU for *baseline-S1-S2* and the *granite-geospatial-uki-flooddetection* models are comparable to within one standard deviation when evaluating on the five images reserved for testing, but it's possible that we may see more of a difference if we tested on additional images.

	baseline-S2				*baseline-S1-S2*				*granite-geospatial* *-uki-flooddetection*			
	US	UKI	S1	S2	US	UKI	S1	S2	US	UKI	S1	S2
base (Prithvi-EO-1.0-100M)	✓	-	-	✓	✓	-	-	✓	✓	-	-	✓
additional-pretraining	-	-	-	-	-	-	-	-	-	✓	✓	✓
fine-tuning	-	✓	-	✓	-	✓	✓	✓	-	✓	✓	✓
inference	-	✓	-	✓	-	✓	✓	✓	-	✓	✓	✓
mIoU on whole test set	0.575 ± 0.019				0.790 ± 0.015				0.786 ± 0.012			
mIoU on cloud-free test set	0.803 ± 0.017				0.845 ± 0.011				0.845 ± 0.009			
F1 on whole test set	0.828 ± 0.020				0.942 ± 0.006				0.938 ± 0.006			
F1 on cloud-free test set	0.952 ± 0.007				0.966 ± 0.003				0.967 ± 0.004			

Table 1: Comparisons of data used at each stage of model set-up are shown in the first four rows. The blue ticks show additional data used in each model set-up compared to what was used for the *baseline-S2* model. Comparisons of mean mIoU and mean F1 scores for the three models, along with their standard deviations are shown in the bottom four rows.

Examples of inference results in Figure 2 illustrate model performance for a range of cloud coverage. In cloud-free areas in the top and middle rows, the *baseline-S2* and *granite-geospatial-uki-flooddetection* models perform at similar levels in segmenting water bodies. In cloudy conditions, in the middle and bottom row, the *granite-geospatial-uki-flooddetection* model produces much more accurate flood segmentation, as might be expected, with the inclusion of S1. In the bottom row we note that model artefacts appear when segmenting water in areas of mismatched swaths between S1 and S2 images, indicating the model should be used with caution for this scenario.

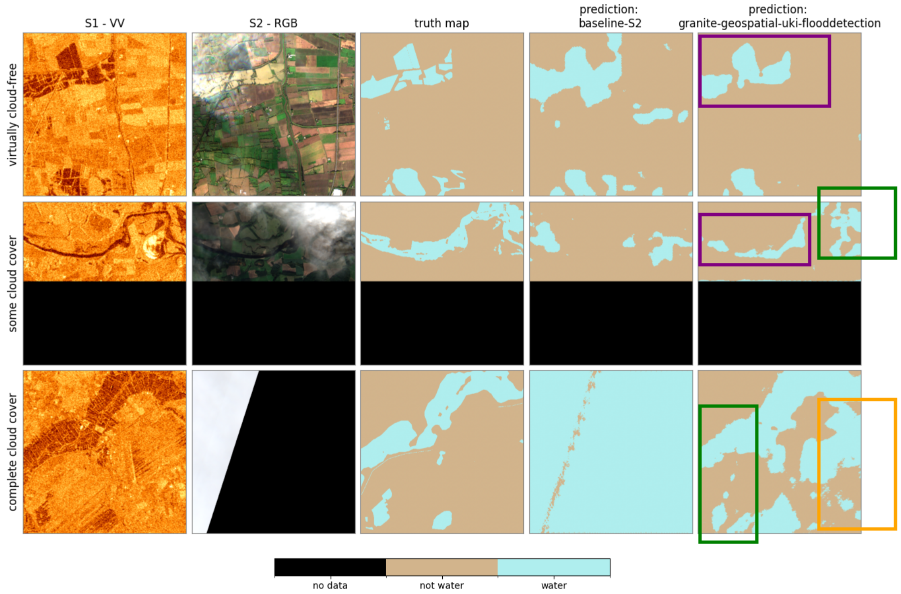

Figure 2: Examples of input imagery, label and inference results from the *baseline-S2* and *granite-geospatial-uki-flooddetection* models. Purple boxes highlight cloud-free regions. Green boxes highlight cloud covered areas. The orange box indicates areas where S1 and S2 swaths do not match.

4 Discussion

Our results show that incorporating S1 bands, with or without additional pretraining, improved the model performance by approximately 35%, however, the improvement will depend on the fraction of images that have significant cloud-cover in any given dataset. Changes in flood extent between capture dates of S1 and S2, and the labelled data will also affect performance. Due to the artefacts seen in Figure 2, we recommend to only apply the model where S1 and S2 swaths overlap. Further models could be fine-tuned to accommodate instances of missing data in one of the modalities. The close model performance between all three models for cloud-free cases may indicate that most of the necessary information was already available in the S2 bands for flood segmentation. The model performance significantly improved with the inclusion of S1 in cloud-covered areas. While it may be difficult to further improve on this task through additional pre-training as the performance is already high, the close model performance between *baseline-S1-S2* and *granite-geospatial-uki-flooddetection* models imply that this improvement may be driven by the inclusion of S1 rather than the fact that in the *granite-geospatial-uki-flooddetection* model we incorporated additional UKI data at the additional pre-training stage.

5 Conclusion

Satellite based monitoring and modelling could be an invaluable tool to understand and adapt to the rapidly changing global climate and accompanying extreme weather events. Here we developed a region-specific pre-trained EOFM for the UKI incorporating additional SAR and visible band imagery. We curated annotated flood maps for the UKI and fine-tuned on these for flood segmentation. Incorporating additional cloud-penetrating SAR imagery significantly improved flood segmentation performance. While it may be difficult to further improve on this already well performing task, there was little benefit in additional pre-training on region-specific data versus directly fine tuning the original EOFM trained in another region. This suggests that EOFMs can be easily adapted to different regions and bands during fine tuning, even with relatively little labelled data. Potential users of such models for climate change adaptation tasks should therefore be encouraged to incorporate relevant additional data for their region or application at the fine tuning stage, particularly in the case of incorporating SAR data in cloudy regions.

REPRODUCIBILITY

Material related to this study are openly available at the following locations:

Prithvi family of foundation models

https://huggingface.co/ibm-nasa-geospatial

granite-geospatial-uki **model**

https://huggingface.co/ibm-granite/granite-geospatial-uki

granite-geospatial-uki-flooddetection **model**

https://huggingface.co/ibm-granite/granite-geospatial-uki-flooddetection

Curated UKI flood dataset

https://zenodo.org/records/14216851

ACKNOWLEDGMENTS

This study contains modified Copernicus Sentinel data [2019-2024] and modified Copernicus Service information [2019-2023]. This work was supported by the Hartree National Centre for Digital Innovation, a collaboration between STFC and IBM.

REFERENCES

Martin Claverie, Junchang Ju, Jeffrey G Masek, Jennifer L Dungan, Eric F Vermote, Jean-Claude Roger, Sergii V Skakun, and Christopher Justice. The harmonized landsat and sentinel-2 surface reflectance data set. *Remote sensing of environment*, 219:145–161, 2018.

Nikolaos Dionelis, Casper Fibaek, Luke Camilleri, Andreas Luyts, Jente Bosmans, and Bertrand Le Saux. Evaluating and benchmarking foundation models for earth observation and geospatial ai, 2024. URL https://arxiv.org/abs/2406.18295.

Security Directorate Space and European Commission Joint Research Centre (EC JRC) Migration. Copernicus emergency management service on-demand mapping, 2025. URL https://mapping.emergency.copernicus.eu/. Accessed November, 2024.

Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

European Space Agency. S2 processing, 2025. URL https://sentiwiki.copernicus.eu/web/s2-process. Accessed February, 2025.

Paolo Fraccaro, Nikola Stoyanov, Zaheed Gaffoor, Laura Elena Cue La Rosa, Jitendra Singh, Tatsuya Ishikawa, Blair Edwards, Anne Jones, and Komminist Weldermariam. Deploying an artificial intelligence application to detect flood from sentinel 1 data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12489–12495, Jun. 2022. doi: 10.1609/aaai.v36i11.21517. URL https://ojs.aaai.org/index.php/AAAI/article/view/21517.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. URL https://arxiv.org/abs/2111.06377.

Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *CoRR*, 2023.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. Geo-bench: Toward foundation models for earth monitoring, 2023. URL https://arxiv.org/abs/2306.03831.

Isla H. Myers-Smith, Jeffrey T. Kerby, Gareth K. Phoenix, Jarle W. Bjerke, Howard E. Epstein, Jakob J. Assmann, Christian John, Laia Andreu-Hayles, Sandra Angers-Blondin, Pieter S. A. Beck, Logan T. Berner, Uma S. Bhatt, Anne D. Bjorkman, Daan Blok, Anders Bryn, Casper T. Christiansen, J. Hans C. Cornelissen, Andrew M. Cunliffe, Sarah C. Elmendorf, Bruce C. Forbes, Scott J. Goetz, Robert D. Hollister, Rogier de Jong, Michael M. Loranty, Marc Macias-Fauria, Kadmiel Maseyk, Signe Normand, Johan Olofsson, Thomas C. Parker, Frans-Jan W. Parmentier, Eric Post, Gabriela Schaepman-Strub, Frode Stordal, Patrick F. Sullivan, Haydn J. D. Thomas, Hans Tømmervik, Rachael Treharne, Craig E. Tweedie, Donald A. Walker, Martin Wilmking, and Sonja Wipf. Complexity revealed in the greening of the Arctic. *Nature Climate Change*, 10(2):106–117, February 2020. ISSN 1758-6798. doi: 10.1038/s41558-019-0688-1. URL https://doi.org/10.1038/s41558-019-0688-1.

Sinergise Solutions d.o.o., a Planet Labs company. Sentinel hub, 2025. URL https://www.sentinel-hub.com. Accessed November, 2024, via API.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, João Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Trevor Keenan, Paulo Arévolo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. *arXiv preprint arXiv:2412.02732*, 2024.

Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012.

Hao Zhang, Jin-Jian Xu, Hong-Wei Cui, Lin Li, Yaowen Yang, Chao-Sheng Tang, and Niklas Boers. When geoscience meets foundation models: Toward a general geoscience artificial intelligence system. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–41, 2024. doi: 10.1109/MGRS.2024.3496478.

A Flood dataset curation

A.1 Labelled flood maps

Labelled flood events over the UK and Ireland were retrieved from the Copernicus Emergency Management Service (CEMS) portal in the form of flood extents. The flood extent polygons retrieved were rasterised. Pixels that fall under areas listed as permanent waterbodies or observed flood events were assigned a value of 1 to denote presence of water. All other pixels were assigned a value of 0 to represent absence of water. Pixels with missing values were assigned a value of -1.

A.2 Input imagery

We noted the imagery dates of the CEMS maps and queried orthorectified Sentinel-1 and Sentinel-2 imagery over the same region at 10m resolution to use as inputs to the fine-tuned model. For the Sentinel-1 interferometric wide swath (IW) product we queried the VV and VH bands. For Sentinel-2 we queried L2A surface reflectance values of the B02 (blue), B03 (green), B04 (red), B8A (Narrow NIR), B11 (SWIR1), B12 (SWIR2) bands and SCL (scene classification map). We queried all images available within a two day window before and after the CEMS map dates. Pixels with missing values were assigned a value of -9999 and overwritten as 0 at the fine-tuning stage within Terratorch.

A.3 Further processing

If there were more than one Sentinel-1 or Sentinel-2 imagery available within the two day window we manually linked the best matching image to the labelled flood map. The matched Sentinel-1 and Sentinel-2 images were concatenated. The scene classification maps were processed to create a cloud mask. We did this by assigning 1 to pixels with scene classification values 8 (cloud medium probability), 9 (cloud high probability) and 10 (thin cirrus) (European Space Agency, 2025) to represent cloud presence, and we assigned 0 to pixels with all other scene classification values to represent the absence of clouds. Sentinel-1 backscatter (σ_0) was normalised using $10log(\sigma_0)$. Any pixels that fell below -35 as a result were set to -35. Any pixels that were still larger than 10 after the normalisation process were capped at 10. We sorted the dataset by waterbody size and cloud-free area and limited the number of images which had minimal or no water presence across the whole field of view. This resulted in 69 images of size 512 x 512, with the flood events and rough dates of the curated events shown in Figure 3. We split the 69 images into 50 images for the training set, 10 images for the validation set and 9 images for the test set.

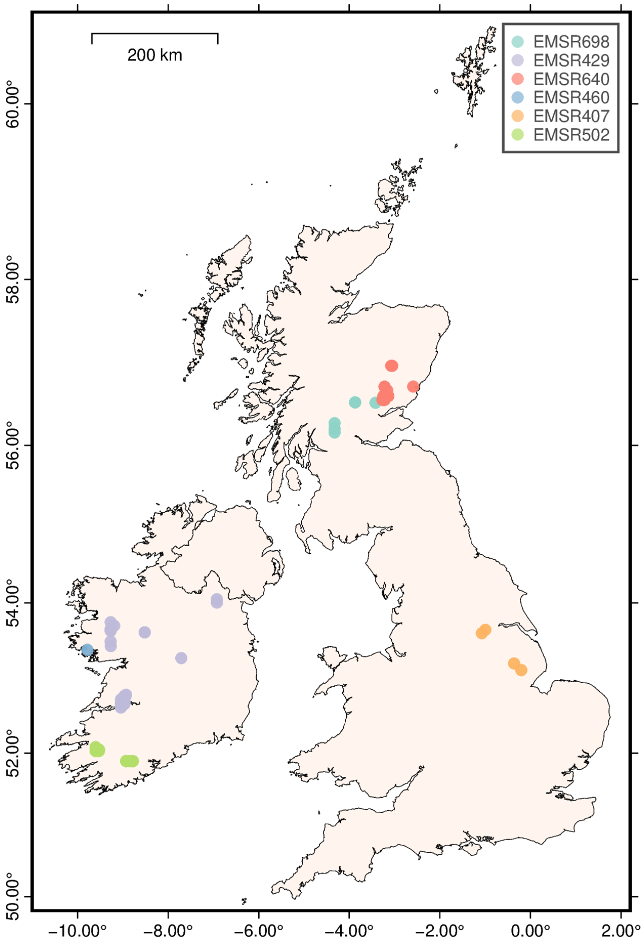

Figure 3: Locations of the 69 images used for the testing, training and validation sets. They are from six flood events, labelled with their CEMS identification codes.

CEMS flood event	approximate date
EMSR407	November 2019
EMSR429	February 2020
EMSR460	September 2020
EMSR502	February 2021
EMSR640	November 2022
EMSR698	October 2023

Table 2: CEMS flood events used in this study and their approximate date of incidence.

B Test set images

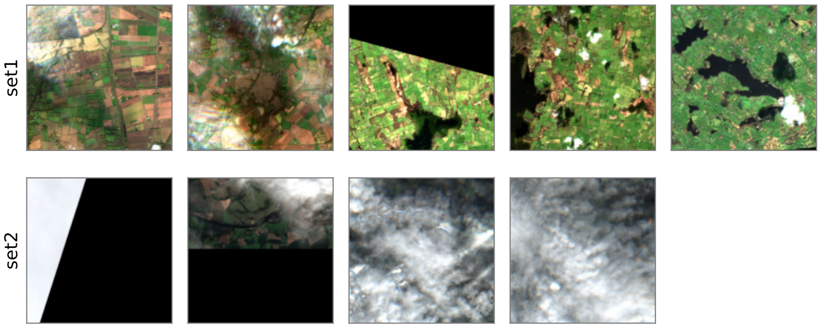

Figure 4: Breakdown of images included in the test set. Set 1 contains virtually cloud-free images, while set 2 contain cases where a model trained on Sentinel-2 or HLS bands only may find more challenging to carry out flood segmentation. We carry out model performance evaluations on set 1 only, as well as on set 1 and set 2 combined (Table 1).

C Validation set performance metric at training time

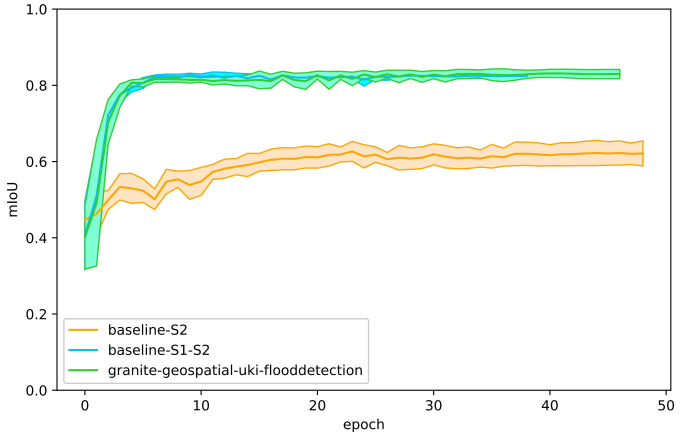

Figure 5: Comparison of mIoU against epoch for the *baseline-S2*, *baseline-S1-S2* and *granite-geospatial-uki-flooddetection* models. The solid lines show the ensemble mean for each model and the lighter shading around it shows the standard deviation of the ensemble.

D MODEL COMPARISON WITH *baseline-S2-global*

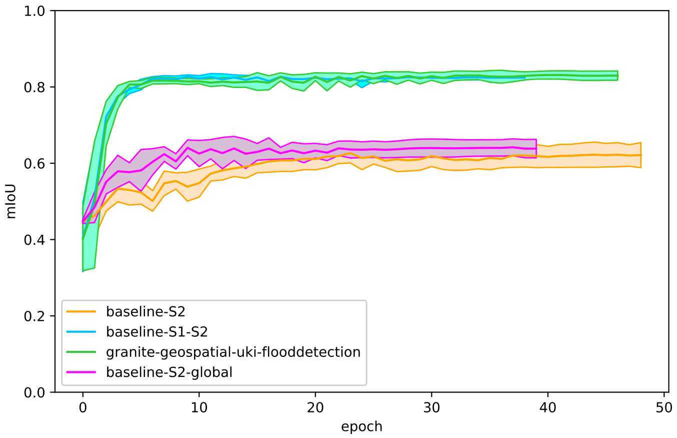

Figure 6: Comparison of mIoU of the validation set against epoch for the *baseline-S2*, *baseline-S1-S2*, *baseline-S2-global* (*baseline-S2* model using *Prithvi-EO-2.0-300M* as a base, as in Table3) and *granite-geospatial-uki-flooddetection* models. The solid lines show the ensemble mean for each model and the lighter shading around it shows the standard deviation of each ensemble.

	baseline-S2					*baseline-S2-global*					*granite-geospatial -uki-flooddetection*				
	US	UKI	global	S1	S2	US	UKI	global	S1	S2	US	UKI	global	S1	S2
base model (Prithvi-EO family)	✓	-	-	-	✓	-	-	✓	-	✓	✓	-	-	-	✓
continuous-pretraining	-	-	-	-	-	-	-	-	-	-	-	✓	-	✓	✓
fine-tuning	-	✓	-	-	✓	-	✓	-	-	✓	-	✓	-	✓	✓
inference	-	✓	-	-	✓	-	✓	-	-	✓	-	✓	-	✓	✓
mIoU on whole test set	0.575 ± 0.019					0.623 ± 0.022					0.786 ± 0.012				
mIoU on cloud-free test set	0.803 ± 0.017					0.806 ± 0.010					0.845 ± 0.009				
F1 on whole test set	0.828 ± 0.020					0.873 ± 0.024					0.938 ± 0.006				
F1 on cloud-free test set	0.952 ± 0.007					0.953 ± 0.004					0.967 ± 0.004				

Table 3: Comparisons of data used at each stage of model set-up are shown in the first four rows. Comparisons of mean mIoU and mean F1 scores for the three models, along with their standard deviations are shown in the bottom four rows.