Journal Pre-proof

Assessing Multi-source Random Forest Classification and Robustness of Predictor Variables in Flooded Areas Mapping

Cinzia Albertini, Andrea Gioia, Vito Iacobellis, George P. Petropoulos, Salvatore Manfreda

PII: S2352-9385(24)00103-4

DOI: https://doi.org/10.1016/j.rsase.2024.101239

Reference: RSASE 101239

To appear in: *Remote Sensing Applications: Society and Environment*

Received Date: 17 October 2023

Revised Date: 21 February 2024

Accepted Date: 7 May 2024

Please cite this article as: Albertini, C., Gioia, A., Iacobellis, V., Petropoulos, G.P., Manfreda, S., Assessing Multi-source Random Forest Classification and Robustness of Predictor Variables in Flooded Areas Mapping, *Remote Sensing Applications: Society and Environment*, https://doi.org/10.1016/j.rsase.2024.101239.

Assessing Multi-source Random Forest Classification and Robustness of Predictor Variables in Flooded Areas Mapping

Cinzia Albertini[1,2], Andrea Gioia[2], Vito Iacobellis[2], George P. Petropoulos[3], and Salvatore Manfreda[4]

[1] Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy (**cinzia.albertini@uniba.it**)
[2] Dipartimento di Ingegneria Civile, Ambientale, del Territorio, Edile e di Chimica, Politecnico di Bari, Bari, Italy (**andrea.gioia@poliba.it**; **vito.iacobellis@poliba.it**)
[3] Department of Geography, Harokopio University of Athens, Kallithea, Greece (**gpetropoulos@hua.gr**)
[4] Dipartimento di Ingegneria Civile, Edile e Ambientale, Università di Napoli Federico II, Naples, Italy (**salvatore.manfreda@unina.it**)

Abstract

Flood extent delineation techniques have benefited from the increasing availability of remote sensing imagery, classification techniques and the introduction of geomorphic descriptors derived from Digital Elevation Models (DEM). On the other hand, high-performing Machine Learning (ML) methods have allowed for the development of accurate flood maps by integrating several predictor variables into supervised or unsupervised algorithms. Among others, Random Forest (RF) is a powerful and widely applied ML classifier, providing accurate predictions also with complex datasets and for varying parameters set. In the present study, the effectiveness of this algorithm for mapping flooded areas was evaluated. Various geospatial data sources were integrated, including morphological indicators, such as the Geomorphic Flood Index (GFI), Sentinel-2 bands, multispectral indices, and Sentinel-1 polarizations. The reliability of the predictor variables under different training sample sizes was evaluated and the accuracy of the RF classifier was assessed. Moreover, by exploring the algorithm ability to identify the most important variables, the predictors contributing the most to the classification were identified and their stability for varying training parameters was investigated. To gauge the adaptability and consistency of these features, we applied our analyses to different study areas around the World. The results indicate that certain predictors displayed remarkable stability across different sample sizes and remained robust under various training parameters. However, some variability in the algorithm structure and the features related to the specific complexities of each considered study case was also observed.

Keywords: *flood mapping; satellite imagery; morphologic features; random forest (RF); predictors robustness; GFI; Sentinel-2; Sentinel-1*

1. Introduction

Over the last few years, changing rainfall and runoff regimes, exacerbated by land use modifications and higher population density, have raised concerns about the increased frequency and magnitudes of flood events, also accelerated by the effects of climate change and global warming (Merz *et al.* 2010, Alfieri *et al.* 2017, Blöschl *et al.* 2019, Blöschl 2022). According to the recent report by the Centre for Research on the Epidemiology of Disasters (CRED, Delforge *et al.* 2022), the recorded number of flood occurrences in 2021 was above the 2001–2020 annual average (223 and 163, respectively). As flood hazard increases, there is a growing need for methodologies able to produce reliable estimates of the flood extent to support damage assessment and disaster responses and for planning risk reduction strategies. Hazard mapping represents, indeed, a measure that can address risk reduction under current and projected future climate scenarios and help with adaptation, stakeholders' engagements and communities informing (Field *et al.* 2012, European Commission 2021).

In this context, Earth Observation (EO) data offers a unique opportunity to retrieve information and measurements regarding the flooding domain (i.e., extent, depth, volume) and water presence across large spatial scales and with various temporal resolutions (Schumann *et al.* 2018, 2022, Tsatsaris *et al.* 2021). Both microwave and multispectral satellite remote sensing techniques have been extensively used in surface water detection and flood mapping. Applications of Synthetic Aperture Radar (SAR) imagery for water detection from Sentinel-1, COSMO-SkyMed or TerraSAR-X, can be found in the field of irrigation events monitoring and surface soil moisture retrieval also linked to flood estimations (Kim *et al.* 2019, Balenzano *et al.* 2021, 2022), flooded area mapping (e.g., Oberstadler *et al.* 1997, Mason *et al.* 2009, Giustarini *et al.* 2012, Cao *et al.* 2019) and spatio-temporal inundation dynamics monitoring (e.g., Wang 2004a, Bates *et al.* 2006, Pulvirenti *et al.* 2011, Refice *et al.* 2018, 2020). Observations derived from SAR sensors are the most commonly adopted and preferred over optical imagery guaranteeing all-weather and illumination conditions (day and night) working capabilities (Volpi *et al.* 2013, Shen *et al.* 2019). Nevertheless, multispectral satellites offer a more straightforward interpretation of data through the visual inspection or simple processing of specific spectral bands or color composites (Albertini *et al.* 2022a). Several studies have been conducted to assess the ability of different multispectral sensors, including Landsat, Moderate Resolution Imaging Spectroradiometer (MODIS) and the most recent Sentinel-2 mission, for flood extent delineation and evolution monitoring or estimation of flood impacts (e.g., Wang 2004b, Chignell *et al.* 2015, Ireland *et al.* 2015, Memon *et al.* 2015, Nandi *et al.* 2017, Munasinghe *et al.* 2018). A comprehensive review of applications and methods for the detection of surface water and floods using multispectral satellites has been recently provided by Albertini et al. (2022a).

If satellite observations depict the flood situation at the event scale, geomorphic approaches based on descriptors derived from Digital Elevation Models (DEMs) provide a valuable characterization of the portion of a river basin frequently exposed to the flood hazard. Several studies have investigated the use of a variety of different morphologic features and composite indices, such as the contributing area, A, local slope, S, elevation difference to the nearest channel, H (or HAND) as first defined by Nobre *et al.* (2016), the modified topographic index, TI_m, and Geomorphic Flood Index, GFI, (Manfreda *et al.* 2011, 2014, 2015, Degiorgis *et al.* 2012, Samela *et al.* 2016, 2017, Tavares da Costa *et al.* 2019, 2020, Albertini *et al.* 2022b, Magnini *et al.* 2022), highlighting the potential to derive over large-scales and ungauged basins the flood exposure of a territory by simply exploiting the information on its morphology.

Developments of accurate flood models and inundation maps have also been possible thanks to advancements in high-performing processing algorithms, including Machine Learning (ML) methods, able to extrapolate better knowledge and new insights from existing data (Zagorecki *et al.* 2013). The advent of ML has facilitated the management of large information volumes and multi-sources data fusion, gaining great popularity among the hydrologists and emergency managers communities in the context of Earth Observation (EO) data applications to monitor natural disasters

(Zagorecki *et al.* 2013, Mosavi *et al.* 2018, Wagenaar *et al.* 2020). A wide range of supervised and unsupervised ML algorithms in remote sensing and flood hazard assessment have been investigated and compared in order to identify the best available classifier in terms of predictive results and generalizations over different study areas. Maxwell et al. (2018) provided an overview of ML classifications of remote sensing imagery comparing six different algorithms, namely the Support Vector Machines (SVM), Random Forest (RF), single decision trees (DTs), Artificial Neural Networks (ANNs), boosted DTs, and *k*-nearest neighbour (*k*-NN), applied to hyperspectral and high spatial resolution urban land cover data. The authors showed the superiority of RF, SVM and boosted DTs, in classification accuracies and computational costs, and their robustness to the characteristics of the features space. Particularly, the RF classifier has been proven to be able to handle well multicollinearity, high-dimension and complex datasets and to display stable accuracy to varying training parameters (Belgiu and Drăguţ 2016, Maxwell *et al.* 2018, Billah *et al.* 2023). Applications of the RF algorithm have been carried out in the context of ecological predictions to map tree species distribution under future climate conditions (Prasad *et al.* 2006), land cover classification using hyperspectral data (Ham *et al.* 2005) and multispectral imagery and topographic features (Gislason *et al.* 2006), to map irrigated crops with Sentinel-2 derived vegetation indices (Radulović *et al.* 2023), as well as in flood hazard and risk assessment. For example, using several predictor variables, such as topographic data, soil properties, hydrological variables and the Normalized Difference Vegetation Index (NDVI, Rouse *et al.* 1974), Wang et al. (2015) analyzed the flood hazard distribution of the Dongjiang River Basin (China) by developing a flood risk map, while Billah et al. (2023) used the RF classifier to assess flood damages to different land cover classes combining Sentinel-1 and Sentinel-2 data.

Furthermore, compared to other ML algorithms, the RF classifier can assess the contribution of each selected predictor variable in the classification. This is possible thanks to embedded functions or ad-hoc algorithms that automatically evaluate whether one of the features is essential for the classification, rank them by order of importance and eventually eliminate the less relevant from the feature space. Such an asset allows for further reducing the data dimensionality and computational cost (Lawrence *et al.* 2006, Wang *et al.* 2015, Belgiu and Drăguţ 2016) and, if considering several study areas, investigating the stability of the predictors, which may help in making generalizations.

In the literature, several studies have examined the sensitivity of the classification accuracies to the RF parameterization and training samples (e.g., Breiman 2001, Gislason *et al.* 2006, Guan *et al.* 2013, Belgiu and Drăguţ 2016), though few focused on the robustness of predictor variables to varying algorithm parameters and size of training classes. Some further investigations, for example regarding the stability of features selection under different algorithm structure, are needed as highlighted in the review on RF applications in remote sensing introduced by Belgiu and Drăguţ (2016). In this context, the present study aims to further explore the potentials of RF for flooded areas classification.

In order to assess the capabilities of the RF algorithm and a series of predictor variables in flood mapping, we carried out an in-depth investigation using a multi-source dataset that includes satellite-based data and DEM-derived features. In particular, we considered: (i) Sentinel-2 spectral bands at 20 m of spatial resolution and six derived multispectral indices, i.e., the Normalized Difference Moisture Index (NDMI, Gao, 1996), the Normalized Difference Water Index (NDWI, McFeeters, 1996), the Red and Short-Wave Infra-Red (RSWIR, Memon et al., 2015; Rogers and Kearney, 2004) index, the Modified Normalized Difference Water Index (MNDWI, Xu, 2006), the NDVI and the Normalized Difference Turbidity Index (NDTI, Lacaux *et al.* 2007); (ii) Sentinel-1 VV and VH polarizations; and (iii) some features derived from DEMs, namely the local slope, S, flow distance to the nearest stream, D, elevation difference to the nearest channel, H, and the GFI. Through a series of sensitivity analyses, the aim was to evaluate the robustness of the algorithm itself and the selected variables for flood extent mapping, to ultimately evaluate the synergy of including both satellite observations and DEM-based features.

Firstly, the present study seeks to assess the stability of features selection (and resulting RF performances) for varying training samples size required in performing the classification. Second, it aims to investigate the contribution of each predictor variable to the classification and quantify the robustness of the most important for varying training parameters. Finally, it delves into the exploration of the accuracy of the RF classification and the stability of the selected features in different study areas. To this end, a first case study was considered to address the first objective, while three additional research areas were introduced to implement the subsequent analyses and carry out an intercomparison between research areas. Using a backward features selection method, only the most relevant predictors showing the greatest contribution to the classification and, hence, the highest discriminating capabilities, were used and assessed in this case.

2. Study cases and datasets

2.1 Criteria for case studies selection and validation data

To carry out the proposed investigations, various case studies were selected from the list of floods documented by the Copernicus Emergency Management Service Rapid (EMSR) that occurred between 2020 and 2023. For each event, mapping products obtained from satellite imagery through photointerpretation, automatic methods, or with a mixed procedure involving automatic and manual classification were delivered. The use of satellite images ensures a large spatial coverage and allows overcoming some issues encountered, for example, with hydrodynamic simulations carried out to derive reference flood hazard maps, including computational costs. Although these maps are produced to provide a relatively immediate (hours to days) response to the emergency activation, they can be considered a reliable source for the retrieval of flooded areas. Therefore, for the scope of this work, Copernicus EMSR maps represent a valuable validation reference and a homogenous source among the considered case studies, thus herein chosen for validation purposes.

For the case studies selection, additional priority was given to flood events meeting the following criteria: (i) availability of satellite images of the same scene from SAR and optical sensors; (ii) shortest time lag between the satellite acquisitions and respect to the Copernicus delineations; (iii) cloud-free images for the applicability of the multispectral data; (iv) if possible, heterogeneity of the landscape complexity, particularly looking at the landform types and land cover. To this end, the Global Shuttle Radar Topographic Mission (SRTM) Landforms (Theobald *et al.* 2015) and the European Space Agency (ESA) WorldCover 10 m 2020 (Zanaga *et al.* 2021) were employed.

A total of four case studies (hereinafter referred to as CS1 to 4) were thus identified regarding floods that occurred in Italy, Australia and Malawi.

For all study areas the dataset includes Sentinel-1 Interferometric Wide (IW) Level-1 Ground Range Detected (GRD) High Resolution (HR) products, Sentinel-2 Level-2A (bottom of atmosphere reflectance) imagery and the NASA's SRTM DEM (Farr *et al.* 2007). Figure 1 depicts the locations of the four selected case studies and the false color composites of the flood events from Sentinel-2 observations. The Copernicus EMSR maps for validation are depicted in Figure 2, while in Table 1 flood events and details about the dataset used for the analyses, as well as main characteristics of the research areas are reported. In the following subsections a brief overview of each case study is provided.

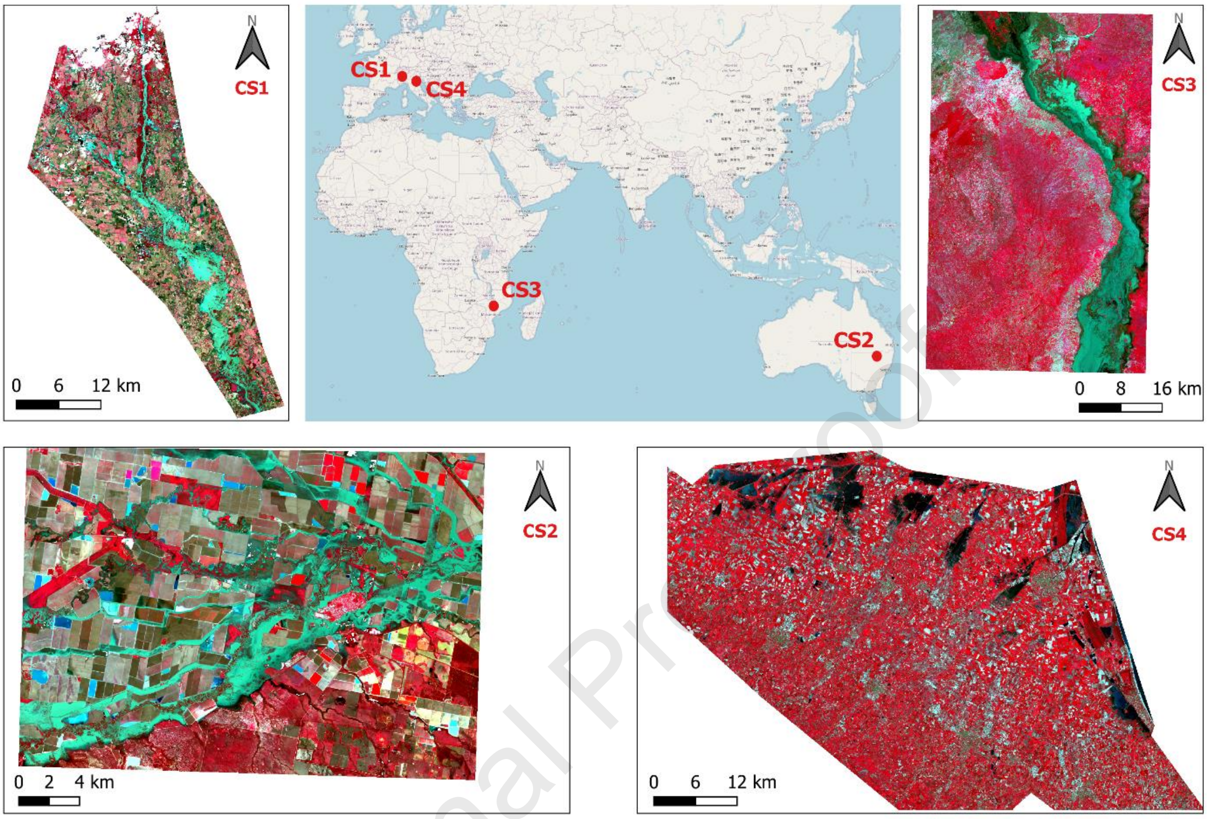

Figure 1. Location of the selected study cases and false color composites of the flood events from Sentinel-2 observations. Case Study 1 (CS1): Sesia River flood in the Piedmont region, northern Italy, 2-3 October 2020; Case Study 2 (CS2): Namoi River flood at the town of Wee Waa, New South Wales, Australia, 22 November - 4 December 2021; Case Study 3 (CS3): Shire River flood between the towns of Bangula and Nsanje in southern Malawi, 20-26 January 2022; and Case Study 4 (CS4): flood in the Emilia Romagna region, northern Italy, at multiple locations, 16-18 May 2023. (map source: Open Street Map).

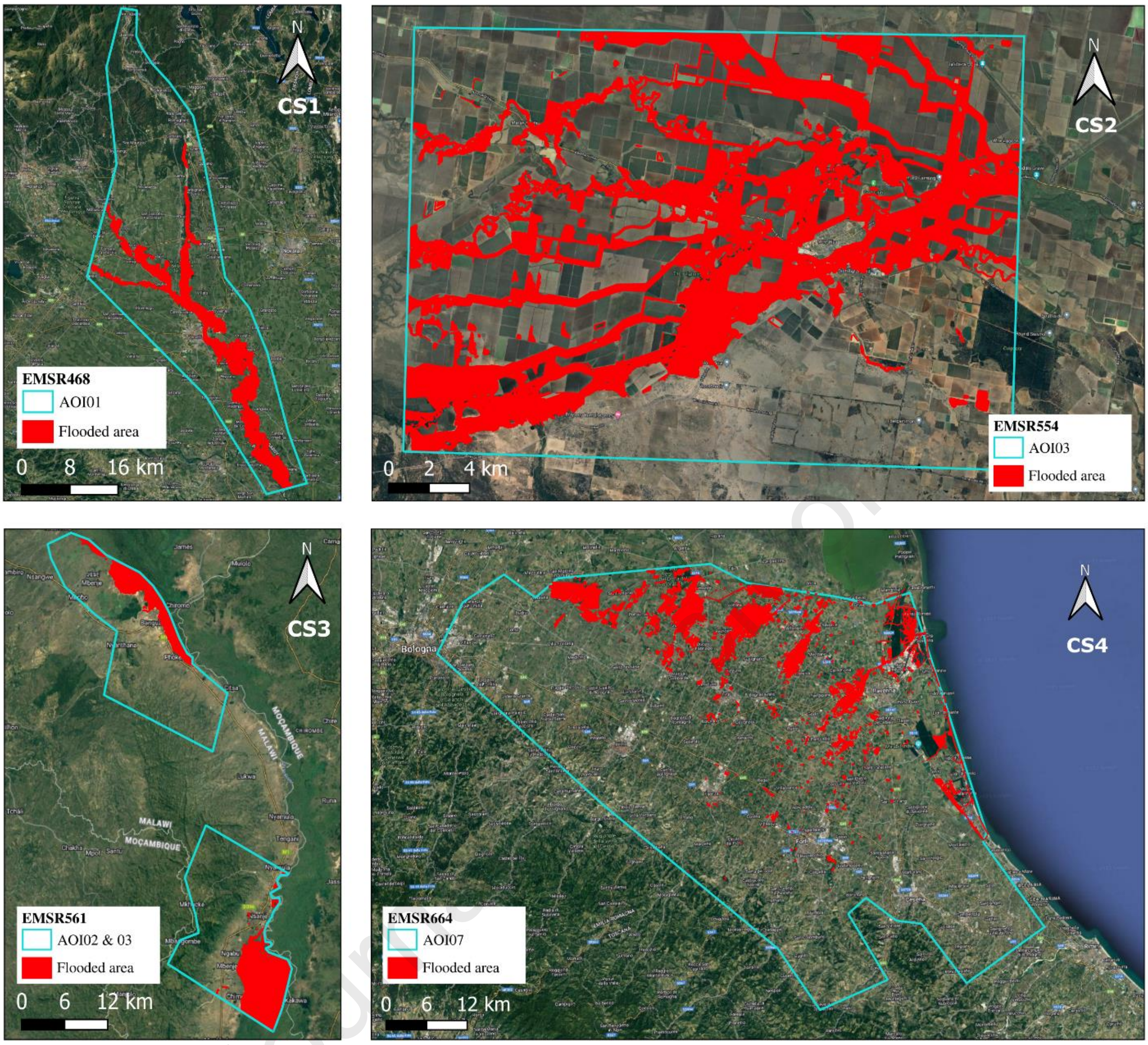

Figure 2. Validation maps as derived from the Copernicus Emergency Management Service (EMS). Rapid Mapping products EMSR468, EMSR554, EMSR561, and EMSR664 for each case study are reported: flood delineations in red and the areas of interest (AOI) in light blue (map source: Google Satellite Hybrid).

2.2 CS1: Sesia River case study (Italy)

The first case study is located in northern Italy at the border between the Piedmont and Lombardy regions. Here a flood event occurred along the Sesia River between 2 and 3 October 2020 extending for about 131 km^2. Heavy precipitations in those days reached 325 mm in the Sesia River basin, leading to exceptional flooding with pick discharge values above 3000 m^3/s and 5000 m^3/s at the "Borgosesia" and "Palestro" stations (Agenzia Regionale per la Prevenzione e la Protezione dell'Ambiente Piemonte 2020). Being located in the Padana plain, the territory is mainly flat and the predominant land cover is crops, particularly rice fields.

For the flood event, Sentinel-1 and Sentinel-2 scenes captured on 3 October are available and the delineation carried out by the Copernicus EMS (Figure 2) on the same day through visual interpretation of the Sentinel-2 image was selected for the validation process (Table 1).

2.3 CS2: Wee Waa case study (New South Wales, Australia)

At the end of November 2021, a La Niña event brought a prolonged rainfall event in southern Queensland and northeastern New South Wales (NSW), Australia, causing the flooding of multiple rivers. Different rapid mapping activations from the Copernicus EMS were issued, including the monitoring of the town of Wee Waa (NSW) where the Namoi River remained under major flooding conditions for around two weeks (22 November-4 December). The delineation was carried out through visual interpretation of a post-event Sentinel-2 image acquired on 3 December, depicting a flood extent of about 154 km^2 in the area of Wee Waa (Figure 2). In addition, the Sentinel-1 satellite captured the flood situation on 4 December. The study region is flat with cropland and grassland as the main cover types (Table 1).

2.4 CS3: Southern Malawi case study

On 25 January 2022, a tropical storm named Ana passed over southern Malawi, bringing heavy rain and widespread flooding in many districts. The monitoring of the flood situation by the Copernicus EMS was activated on 27 January, particularly in the hardest-hit districts of Chikwawa, Bangula and Nsanje where the Shire River overflowing was observed.

In the present investigations, the latter two were considered as the areas of interest (AOI) for which Copernicus flood extent maps to be used in the validation process were obtained through visual interpretation of a post-event Sentinel-2 acquired on 30 January. For the analyses, we considered Sentinel-1 and Sentinel-2 images respectively acquired on 2 and 4 February. Flooding in the study area extended for approximately 90 km^2 and affected a flat territory mainly covered by shrubs and grass (Table 1).

2.5 CS4: Emilia-Romagna case study (Italy)

In May 2023 the south-east territory of the Emilia-Romagna region (northern Italy) was affected by severe weather conditions that triggered intense rainfall, generating rivers overflowing and considerable floods on 2 May in the Bologna, Ravenna and Forlì-Cesena provinces. A subsequent perturbation between 16 and 18 May exacerbated the dramatic condition of these territories, where the breaking of several river embankments caused the inundation of different towns, roads and people displacement. The Copernicus EMS activated the monitoring of the emergency and produced a detailed delineation of the flood extent.

The present analyses were carried out considering the Forlì, Lugo and Ravenna territories as areas of interest, in which the flooding covered a total area of about 160 km^2. Copernicus maps depicting the flood situation as of 21 May are available, while Sentinel-1 and Sentinel-2 images were acquired on 22 and 23 May, respectively. The study region is flat and mainly characterized by cropland and grassland (Table 1).

Table 1. Selected case studies (CS) with information about flood occurrence locations and dates, data used for the validation and analyses with details on production/acquisition dates, original spatial resolutions and data portals for download, and main characteristics of the study areas.

			Case Study			
			CS1	**CS2**	**CS3**	**CS4**
			Sesia River (Northern Italy)	*Wee Waa (New South Wales, Australia)*	*Southern Malawi*	*Emilia-Romagna Region (Northern Italy)*
		Flood event date	2-3 October 2020	22 November – 4 December 2021	20-26 January 2022	16-18 May 2023
Dataset	**Copernicus flood map**	**Date**	3 October 2020	3 December 2021	30 January 2022	21 May 2023
		Activation Number	EMSR468	EMSR554	EMSR561	EMSR664
	Sentinel-2 acquisition (Level-2A product)	**Date**	3 October 2020	3 December 2021	4 February 2022	23 May 2023
		Data download	Sentinel Scientific Data Hub https://scihub.copernicus.eu/ (last accessed on 22 June 2023)			
		Spatial resolution	20 m			
	Sentinel-1 acquisition	**Date**	3 October 2020	4 December 2021	2 February 2022	22 May 2023
		Data download	Sentinel Scientific Data Hub https://scihub.copernicus.eu/ (last accessed on 22 June 2023)			
		Spatial resolution	10 m			
	SRTM DEM	**Spatial resolution**	1 arc-second (~ 30 m)			
		Data download	NASA Earthdata Search https://search.earthdata.nasa.gov/search (last accessed on 22 June 2023)			
Main characteristics	**Global SRTM Landforms**		Lower slope (flat)	Lower slope (flat)	Lower slope (flat)	Lower slope (flat)
	ESA WorldCover		Cropland	Cropland Grassland	Shrubland Grassland	Cropland Grassland

3. Methodology

The RF classifier is a supervised ML algorithm for prediction problems proposed by Breiman (2001). It consists of a collection of trees built through a random selection of both predictor variables, *p*, and subsets of the training dataset. Each tree provides a prediction on the class membership and the final choice is made based on the most popular vote among all the trees. The algorithm requires the user to specify some parameters that ensure the best classification accuracies, in particular the number of trees to be grown, *Ntree*, and the number of variables randomly selected to split each node of the tree, *mtry*. These can be fine-tuned through optimization procedures, such as *k*-fold cross-validation (Hastie *et al.* 2009), or set to default values since RF performances have been shown to be quite robust with varying parameters set. In the literature, several authors recommended 500 to be a reasonable value for *Ntree* (Gislason *et al.* 2006, Belgiu and Drăguţ 2016) and the square root of the number of predictor variables for *mtry* (Gislason *et al.* 2006).

As with all supervised ML algorithms, the RF classifier needs labelled data in order to be trained and build the final prediction model. Regions of Interest (ROIs) are groups of labelled training samples, i.e., training pixels, collected through field observations or photointerpretation, and must be representative of the classes to be predicted. In this work, ROIs were selected based on visual interpretation of Sentinel-2 RGB band combinations. In particular, true (B4, B3, B2), false (B8a, B4, B3) and SWIR (B12, B8a, B4) color composites were considered and visually evaluated to derive flooded and not flooded classes by manually digitizing sample polygons which gather some pixels together. Regarding the number of samples necessary for training the algorithm, a minimum of 10-30*p* training pixels per class is suggested in the remote sensing literature (Piper 1992, Van Niel *et al.* 2005, Mather and Koch 2011, Petropoulos *et al.* 2011).

While the stability of the RF classification accuracies using different parameterization schema and its sensitivity to the size of the training samples have long been assessed (e.g., Gislason *et al.* 2006, Colditz 2015, Millard and Richardson 2015, Ramezan *et al.* 2021), specific investigations concerning the robustness of features selection to varying training samples sizes and the sensitivity of the *Ntree* parameter to the number of variables are required (Belgiu and Drăguţ 2016). To this end, a series of sensitivity analyses was carried out herein either to one or all study cases using satellite-based and geomorphic features as predictors for flood extent delineation. In particular, it was investigated the stability of predictors to varying sample sizes (subsection 3.3), their robustness for varying *Ntree* (subsection 3.4) and the stability of both predictors and RF accuracies in different study areas (subsection 3.5). A complete overview of the methodological workflow is presented in Figure 3, while in the following subsections, pre-processing of input data is illustrated, and a detailed description of each analysis is provided.

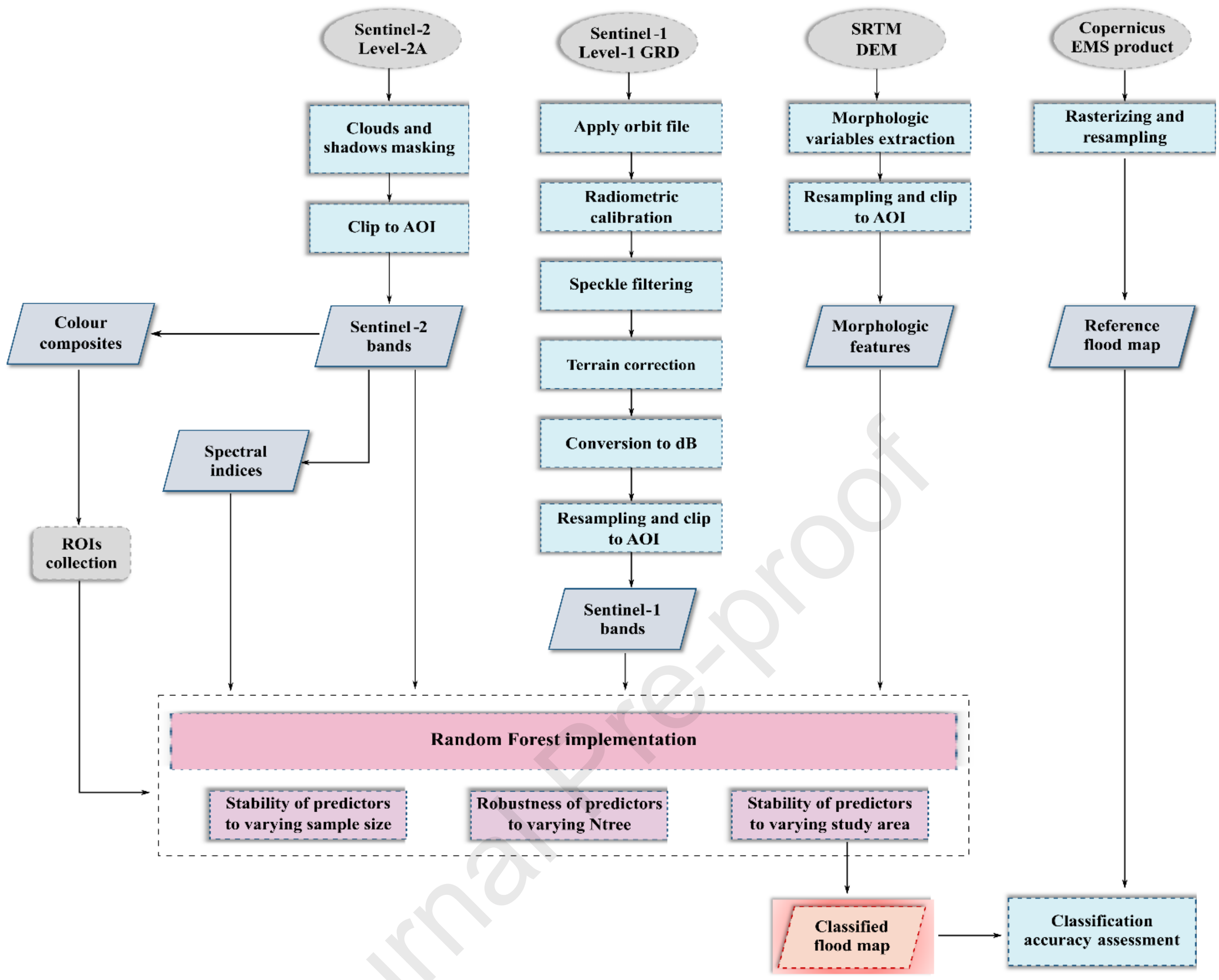

Figure 3. Methodological workflow adopted for the implementation of the Random Forest classification to delineate flooded areas and assess the stability and robustness of predictor variables.

3.1 Pre-processing

A total of *p*=21 morphologic and satellite-based features were selected as predictors in the RF classification of flooded areas in the four study cases, as listed in Table 2. A series of pre-processing steps, the same for all the investigated areas, were applied to both satellite imagery and DEM-based data before the implementation of the predictors into the RF model.

Regarding Sentinel-2 imagery, nine bands at 20 m spatial resolution ranging from the visible (i.e., blue, green, red) and near-infrared spectral domain (i.e., Red Edge 1, Red Edge 2, Red Edge 3 and NIR) to the short-wave infrared (i.e., SWIR 1 and SWIR 2) were considered and pre-processed to mask clouds and their shadows (if present in the AOI). To this end, the cloud mask layer and the Scene Classification Layer map at 20 m spatial resolution contained in the distributed Sentinel-2 Level-2A product were employed. It is worth mentioning that in the selected case studies if clouds were present, they only partially covered the scene and did not affect the flood extent. Once the nine bands were pre-processed, six multispectral indices were computed, namely the NDMI, NDWI, RSWIR, MNDWI, NDVI and NDTI (see Table 2 for the complete formulas).

Sentinel-1 data, in particular the VV and VH bands, were pre-processed in the Sentinel Application Platform (SNAP, version 9.0.0) according to the standard generic workflow suggested for GRD products (Filipponi 2019). The implemented processing steps include the application of the orbit file to update the orbit state vectors and correct the satellite position and velocity; radiometric calibration to convert digital pixel intensity into backscatter values (sigma noughts values, σ_0 [-]); speckle filtering to remove the scattering noise and improve the image quality; Range Doppler terrain correction to correct image distortions related to the side looking geometry of the satellite; and the conversion of σ_0 values to decibels (dB) using a logarithmic function (Filipponi 2019, Gašparović and Klobučar 2021). A Lee filter with a 3x3 window size was used in the speckle filtering step (Lee *et al.* 1994, Dutsenwai *et al.* 2016, Cenci *et al.* 2017, Ezzine *et al.* 2018), while the SRTM DEM and the bilinear resampling method were applied for terrain correction.

Four morphologic features were derived from the SRTM DEM at 30 m spatial resolution, namely the local slope, S (-), expressed as the tangent of the gradient, that is to say, the maximum slope among the eight possible directions connecting the pixel under exam to the neighbouring cells; the flow distance, D (m), and elevation difference, H (m), to the nearest stream, respectively defined as the length of the path hydraulically connecting the location under exam and the nearest pixel of the river network and the difference in elevation between these two cells; and the GFI, which is a composite index expressed as the natural logarithm of the ratio between the water level in the nearest element of the river network and *H*. It is computed at the river basin scale once the flow direction and flow accumulation rasters are derived from a depressionless DEM. For a complete description of this index and the processing step necessary to compute it, please refer to Samela et al., (2017).

Since input data have different spatial resolutions, Sentinel-1 bands and DEM-derived features were resampled through bilinear interpolation to 20 m after pre-processing, assuming the resolution of the Sentinel-2 bands as reference. This choice lies in the fact that ROIs collection for the subsequent algorithm training was based on the visual interpretation of Sentinel-2 scenes.

Regarding the validation products, vector data produced for the selected flood events by the Copernicus EMS were rasterized and resampled to match the 20 m resolution of the input data. Since Copernicus maps could include both flooded areas and flood traces in the delineation, depending on the selected product, the latter were excluded from the final validation map, especially if the considered data were days apart from the EMSR delineation and flood event occurrence, as in the CS3 and 4 (Malawi and Emilia-Romagna case studies). The reason for this lies in the fact that a fair domain for the comparison between maps was desirable. In fact, it should be considered that the detection of flooded areas encounters some limitations with post-flood data and, in general, the quality of the delineation decreases with the time after the flood peak (Notti *et al.* 2018).

Finally, all data were clipped to the AOI as identified in the map extent layer of each case study provided in the corresponding Copernicus vector data package (Figure 2).

Table 2. Satellite-based and morphologic features selected as predictors in the Random Forest classification of flooded areas in the four case studies.

Sentinel-2 bands	**Band name**	**Band number**
	Blue	B2
	Green	B3
	Red	B4
	Red-Edge 1	B5
	Red-Edge 2	B6
	Red-Edge 3	B7
	Nir	B8a
	SWIR 1	B11
	SWIR 2	B12
Sentinel-2 spectral indices	**Index name**	**Index formula**
	NDVI	$\frac{Nir - Red}{Nir + Red}$
	NDWI	$\frac{Green - Nir}{Green + Nir}$
	NDMI	$\frac{Nir - SWIR\ 1}{Nir + SWIR\ 2}$
	MNDWI	$\frac{Green - SWIR\ 1}{Green + SWIR\ 1}$
	NDTI	$\frac{Red - Green}{Red + Green}$
	RSWIR	$\frac{Red - SWIR\ 1}{Red + SWIR\ 1}$
Sentinel-1 bands	**Band name / polarization**	
	VV	
	VH	
DEM-derived morphologic features/indices	**Feature/index name**	**Feature/index formula**
	H	/
	D	/
	S	/
	GFI	$\ln\left(\frac{h_r}{H}\right)$

3.2 Stability of predictors to varying sample size

Considering different sizes of training samples, the sensitivity of feature selection (and consequently the RF prediction accuracy) was assessed on the CS1.

In the literature, different recommendations are given regarding the number of training samples. In general, a minimum of 10-30p training pixels per class, where p is the number of predictors, is suggested to be used to train the classifier (Piper 1992, Van Niel *et al.* 2005, Mather and Koch 2011, Petropoulos *et al.* 2011). Therefore, in the current investigation involving the two classes of flooded and not flooded pixels, a number of pixels per class, n, equal to 210 (10p), 420 (20p) and 630 (30p) were considered, where p=21 is the number of selected predictors (see subsection 3.1). In addition, sample sizes outside the suggested range, i.e., a minimum of 50 samples per class (Colditz 2015) and 945 (45p) training pixels were also explored. Such samples were collected through the aforementioned manual digitization of training polygons and by means of photointerpretation of the

Sentinel-2 scene. Figure S.1 illustrates some examples of polygons digitized based on the interpretation of the true color composite to collect the ROIs in the case n=210 and n=945 pixels per class. In particular, panel S.1(a) describes the agreement/disagreement between the Copernicus EMSR map and the ROIs pixels labelled as “flooded” and “not flooded”. The size of the training samples was increased by either digitizing new polygons (ROIs b.1-2 in panel S.1(b)) or enlarging the existing ones (ROIs a.1-3, b.3-4 in panels S.1(a,b)).

The commonly adopted Jeffries-Matusita (JM) distance was computed to assess the quality of the collected samples, that is to say, the ability of each individual wavelength in the selected ROIs to discriminate between flooded and not flooded classes. In Table S.1 JM values for each set of ROIs are reported.

The analyses were implemented in the R package “caret” (Kuhn *et al.* 2020) to build the RF models for the Sesia River flood. In this case, training parameters were set to default values, i.e., *Ntree*=500 and *mtry*=$\sqrt{p}$, and no fine-tuning was carried out. Each training sample was split into 75% for training and 25% for testing and prediction accuracies were registered. Finally, for each sample size, the stability of predictors was assessed by considering the Mean Decrease in Accuracy measure (MDA, Breiman, 2001) to rank them in order of importance. The MDA is an RF internal estimate of the contribution of each feature to the final classification and quantifies the reduction in accuracy occurring when one of the predictors is excluded from the model. Hence, higher MDA scores correspond to very important features. MDA values were scaled between 0% and 100% to provide a measure of the mean relative importance.

3.3 Robustness of predictors for varying Ntree

The robustness of feature selection for varying *Ntree* was assessed to provide an overview of the sensitivity of predictors to different RF classification schemas. This analysis was conducted in each case study to also evaluate the stability and potential transferability of predictor variables in different study areas (see subsection 3.5).

A feature selection technique was used to implement the RF classification exploiting only the most significant variables. Besides variable importance ranking methods embedded in the RF model, i.e., the MDA and the Gini Impurity metrics (Breiman 2001), wrapper approaches can be used that identify and select the most useful variables to train the classification model. Through a specified search strategy, such methods consist of evaluating different combinations of feature subsets with which the algorithm is trained (on a training dataset) and tested (on a test set or via cross-validation). For each subset, classification performances are derived and only the subset yielding the best accuracy is selected (Kohavi and John 1997, Guyon and Elisseeff 2003).

The RF classifier was applied with the recursive feature elimination (RFE) method, also known as backward feature selection, for different *Ntree* parameter values. This analysis allowed the selection of an optimum number of predictors, $p*$, among all the input features for testing their robustness in different parameter sets. The RFE trains the classifier with a recursive backward strategy that fits the model using a decreasing size of predictor subsets. The model is first trained on a training dataset using all p predictors, then model performance and variable importance are computed and only the most relevant are kept. The new subset of predictors is used to train the model once again, predictors are reranked and the least important are removed. The model with the best performance is identified and used to fit the final classifier using the corresponding optimal subset $p*$.

The RF-RFE was implemented using the "rfe" function in the R "caret" package. The algorithm was built using in each case study a number of ROIs *n* at least equal to 10*p* per class (collected following the same procedure as for the analysis of the stability of predictors to varying sample sizes), which were split into 75% for training and 25% for the final testing. JM distance was computed to evaluate the spectral separability of the selected wavelength (Table S.2). A 5-fold cross-validation was applied for model evaluation in the RFE and the Overall Accuracy (OA) metric was selected to identify the optimal model. Different values of *Ntree* were tested, ranging from 10 to 50 and 100 to 1000, while the *mtry* parameter is automatically set by the algorithm to the default value, i.e., to the square root of the number of optimum predictors *p** identified by the "rfe" function.

3.4 Stability of RF classifier and predictors to varying study area

The accuracy of the RF classifier and the robustness of the optimal feature subsets in the four study areas were assessed to identify the most stable predictors and their transferability in different contexts. To carry out this analysis, after testing the robustness of predictor variables for varying numbers of trees, the model showing the best performances was identified. To this end, for each value of *Ntree* an objective function, *obj*, defined as the sum between the false positive rate, R_{fp}, and the false negative rate, R_{fn}, was considered (Equations 1-3) that assigns equal weights to the two error rates. The model with the lowest *obj* value was chosen as the final one. The Copernicus flood maps were used as validation products in a pixel-per-pixel comparison with the RF classification maps obtained from the final selected model, and from the confusion matrices true positive (TP), true negative (TN), false negative (FN), and false positive (FP) pixels were identified.

$$obj = R_{fp} + R_{fn} \tag{1}$$

$$R_{fp} = \frac{FP}{TN + FP} \tag{2}$$

$$R_{fn} = \frac{FN}{TP + FN} \tag{3}$$

Additional error and accuracy metrics (Equations 4 to 8) were computed, including the True Positive Rate, R_{tp}, True Negative Rate, R_{tn}, OA, Precision and the F-score:

$$R_{tp} = \frac{TP}{TP + FN} \tag{4}$$

$$R_{tn} = \frac{TN}{TN + FP} \tag{5}$$

$$OA = \frac{TP + TN}{T} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F - score = \frac{2 \cdot Precision \cdot R_{tp}}{Precision + R_{tp}} \tag{8}$$

where *T* is the total number of pixels in the image.

4. Results

4.1 Stability of predictors to varying samples size

The sensitivity of the predictors to varying samples size was assessed in the Sesia River flood case study (CS1). For each configuration, high training and testing accuracies were achieved (Table 3) through the RF classification model implemented with the default parameters values, as described in Section 3.2. In particular, training accuracy values above 99% were obtained starting from n = 210 samples per class. Testing accuracies of 100% were registered in each case, mainly because in a two-class classification problem the chance of producing good results is very high (i.e., the probability of mistake is minimal) since categorizing a pixel in one of the two classes has the same probability.

Table 3. Accuracy values obtained with the training and testing datasets for varying sample sizes computed as multiples of the number of predictors p: n = 50, 210 (10p), 420 (20p), 630 (30p) and 945 (45p) number of pixels per class.

	Model Accuracy (%)	
n	*Training set*	*Test set*
50	98.69	100
210 (10*p*)	99.43	100
420 (20*p*)	99.98	100
630 (30*p*)	99.85	100
945 (45*p*)	99.86	100

Figure 4 shows the mean relative importance (top panel) and rankings (bottom panel) of the variables used to detect flooded and not flooded classes for the five considered sample sizes. In the figure, different colors refer to the number of training pixels per class, n, that is 50 (green), 210 (orange), 420 (red), 630 (light blue) and 945 (pink).

Based on the RF model, three multispectral indices, namely the NDMI, RSWIR and MNDWI, and the Sentinel-2 bands SWIR 1 and SWIR 2 (except in one case) were the most important variables with a relative importance of over 20% in each sample size configuration (Figure 4, top panel). The NDMI was the predictor with the highest importance, having a stable mean relative value of around or over 90%. In addition, these were always among the five most ranked variables (Figure 4, bottom panel) and, between them, the MNDWI showed very strong stability being classified as the third most contributing predictor for every sample size. The other four variables were also characterized by a certain stability being in the same rank except in one size out of five.

In addition to those, the NDWI, NDVI and NDTI mean relative importance was between 5% and 10% in each sample size configuration, being always within the first 12 important variables. The Green band was also ranked among the first 12 predictors in each configuration, but the mean relative importance was more unstable. All the other predictors showed a relative importance of less than 5%, or if above, they were characterized by a higher variability in the relative importance value. It is interesting to note that the VH and VV Sentinel-1 bands and the S morphologic features had the highest relative importance when the sample size was small, i.e., for n=50 pixels. Finally, the NDTI and Red Edge 1 predictors showed an opposite behavior: while the contribution of the former almost linearly decreased for increasing sample size, it increased for the latter (Figure 4, bottom panel).

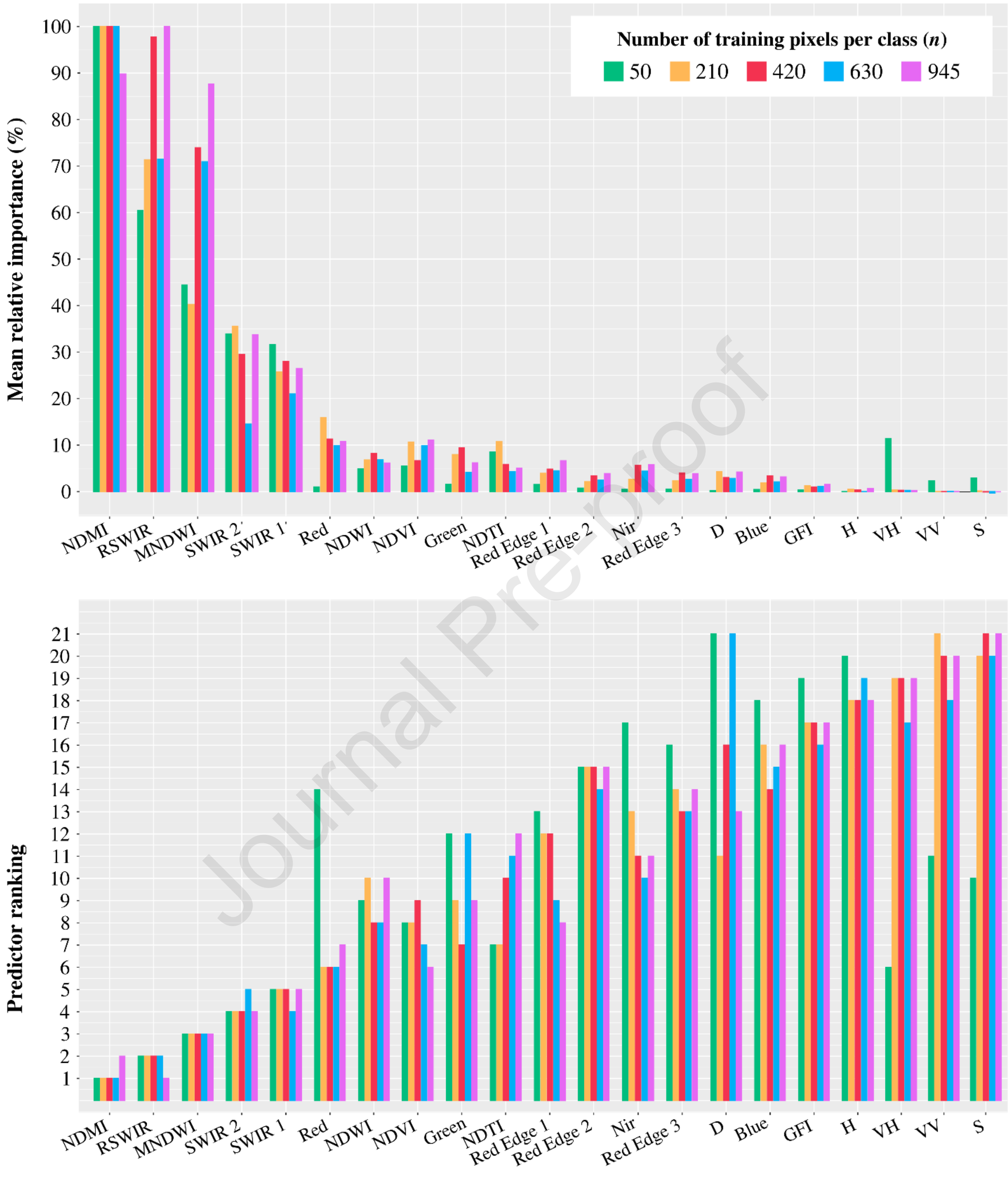

Figure 4. Stability of predictor variables used to classify flooded and not flooded areas through the Random Forest model trained with varying samples size: 50 (green), 210 (orange), 420 (red), 630 (light blue) and 945 (pink) pixels per class. In the top panel mean relative importance values (%) are reported, while the ranking of predictors according to the mean decrease accuracy metrics is shown in the bottom panel.

4.2 Robustness of predictors for varying Ntree

Applying the RF classifier with the RFE method the robustness of feature selection for varying *Ntree* was assessed in the four case studies. Figure 5 shows the number of optimum predictors (*p**, top panel) for each value of the parameter, together with the model accuracies (bottom panel). Different colors refer to the Sesia River (CS1, orange), Wee Waa (CS2, green), Southern Malawi (CS3, blue) and Emilia-Romagna (CS4, purple) study areas.

Overall, good performances were achieved with the optimum subsets size and a certain stability of the RF accuracies for varying *Ntree* in the different case studies can be observed. Regarding the subsets of predictors, in general, for CS3, a decreasing number of variables were identified as necessary for increasing number of trees, while *p** was more stable for CS2. In fact, a lower variability with *Ntree* was observed and no less than five and more than 16 variables were selected. Regarding CS1 and CS4, on average a higher number of predictors were identified as necessary to the classification and in some cases all the 21 input features were selected.

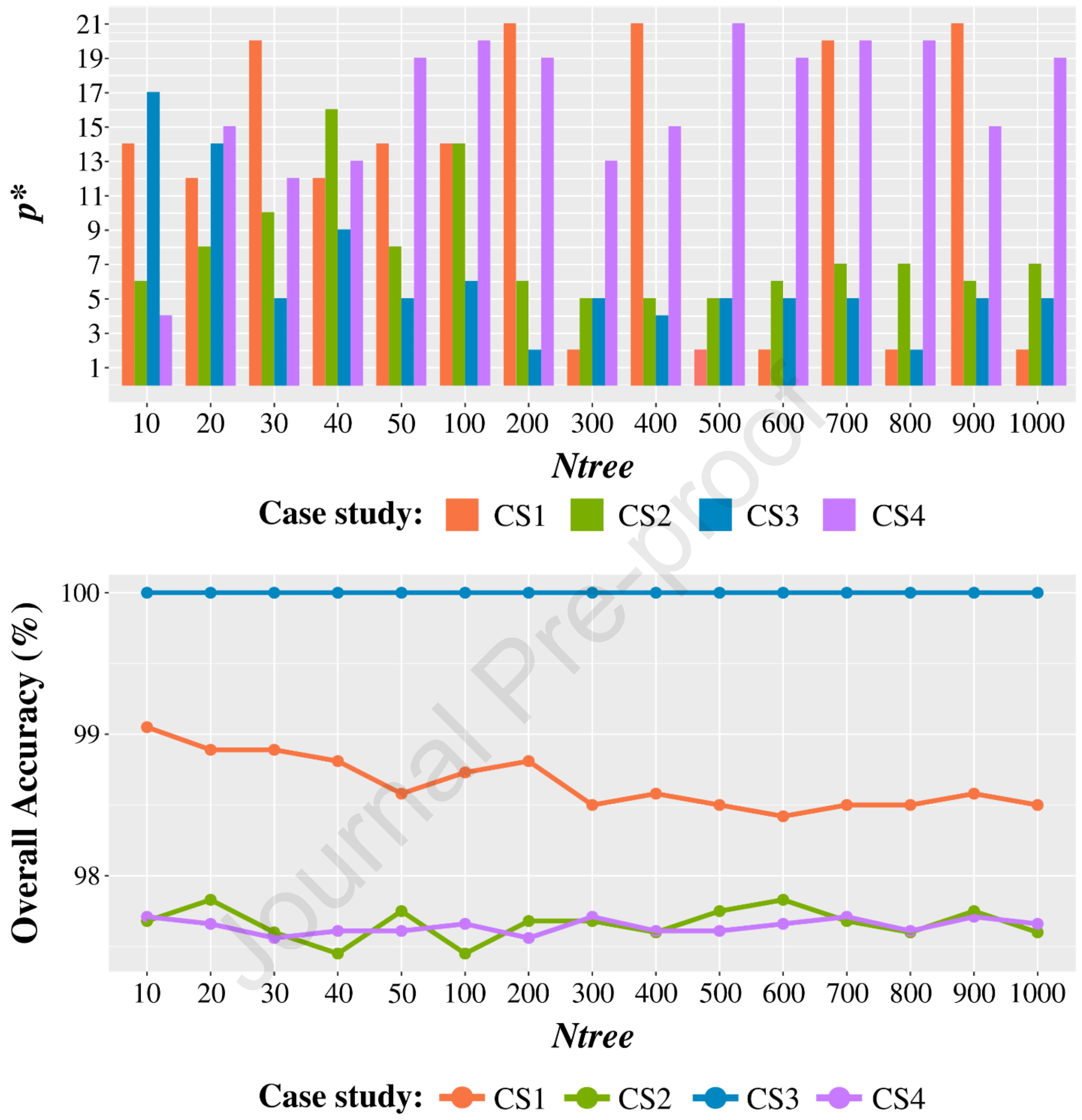

Figure 5. Optimum number of predictors, *p**, selected through the Recursive Feature Elimination Method (RFE) in the Random Forest algorithm (top panel) and accuracy values (bottom panel) for different numbers of trees, *Ntree*. Colors refer to the four case studies, which are the Sesia River (CS1, orange), Wee Waa (CS2, green), Southern Malawi (CS3, blue) and Emilia-Romagna (CS4, purple) flood events.

CS1: Sesia River case study

Analysing the selected predictors in every *Ntree* parameter configuration it is possible to assess the sensitivity of each variable and identify the most stable. The heatmap shown in Figure 6 depicts such variability for the Sesia River study area. Boxes are marked with colors if variables were selected in a specific *Ntree* model configuration. In addition, colors are graded from blue to yellow, respectively indicating a highest or lower contribution to the flooded areas classification (ranked as first to 21st most important variables), while grey color is shown if a predictor was not selected.

Regarding the Sentinel-2 related variables, the NDMI and RSWIR index were the strongest features, being selected for each *Ntree* and ranked among the first three most contributing predictors (dark blue). Together with the first two indices, the MNDWI, NDTI, NDVI and NDWI and the SWIR 2, SWIR 1 and Red bands were among the first ten important variables (blue to aqua green colors), even if not present for each *Ntree* model. Regarding the Sentinel-1 bands, VH showed a lower sensitivity to changing *Ntree* compared to VV and higher importance, especially for lower values of *Ntree*. The weakest predictors were the morphologic features S, D and H, being selected only for four or five *Ntree* values and always among the last three most important variables (green to yellow shades).

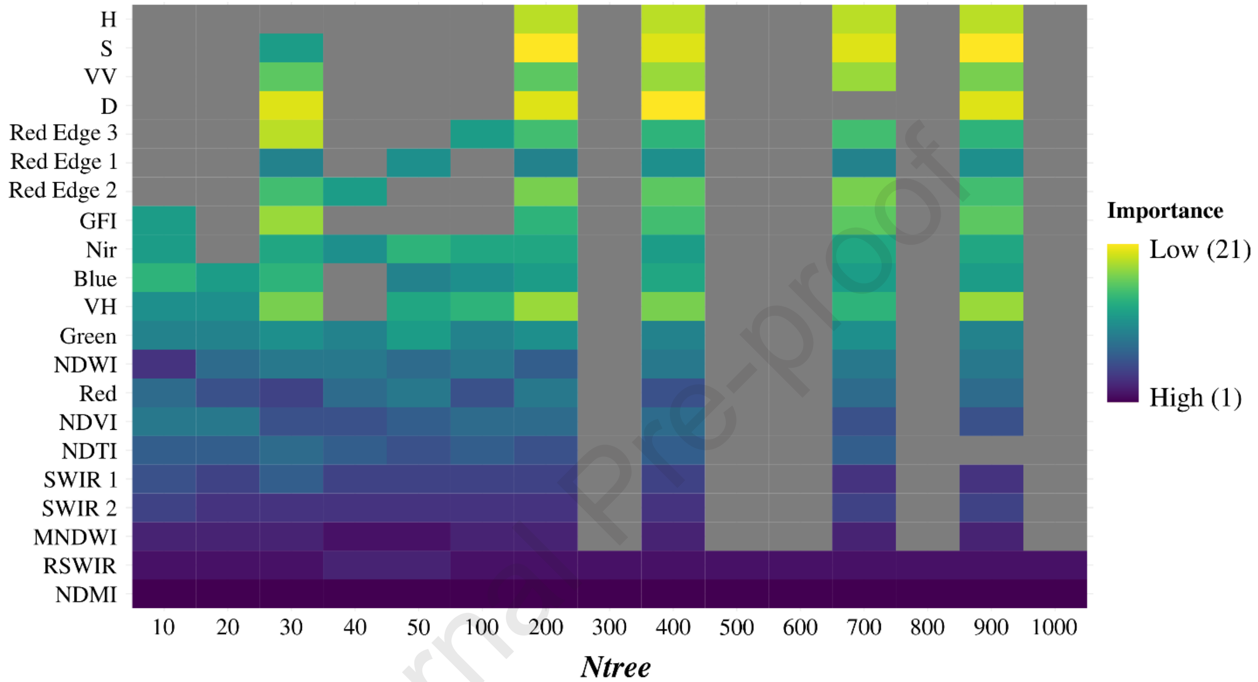

Figure 6. Heatmap for Case Study 1 (CS1) depicting the robustness of predictors for varying *Ntree* and their importance (blue to green colors).

CS2: Wee Waa case study

As reported in Figure 5, a maximum of 16 variables out of the initial 21 were selected for the Wee Waa case study. The heatmap in Figure 7 shows the robustness of the predictors to varying *Ntree* and the importance of each, which in this case goes from 1 to 16 (blue to yellow color shades).

The NDWI, RSWIR, MNDWI, and NDVI variables were the most stable and also ranked as the most important for each *Ntree* value (blue shades). Some predictors were selected only starting from certain values of the parameter, while others only for a low number of trees. For example, the SWIR 1 band was chosen from *Ntree* = 20, while the NDMI and NDTI for *Ntree* values not above 40 and 100, respectively. In addition, the NDMI was the most contributing to the classification for very low number of trees, i.e., *Ntree* = 10. The morphologic features D and H were never selected among the optimum predictors, while the GFI only once (*Ntree* = 40). The same occurred with the Sentinel-1 polarization VV (never chosen) and VH (chosen only for *Ntree* = 40).

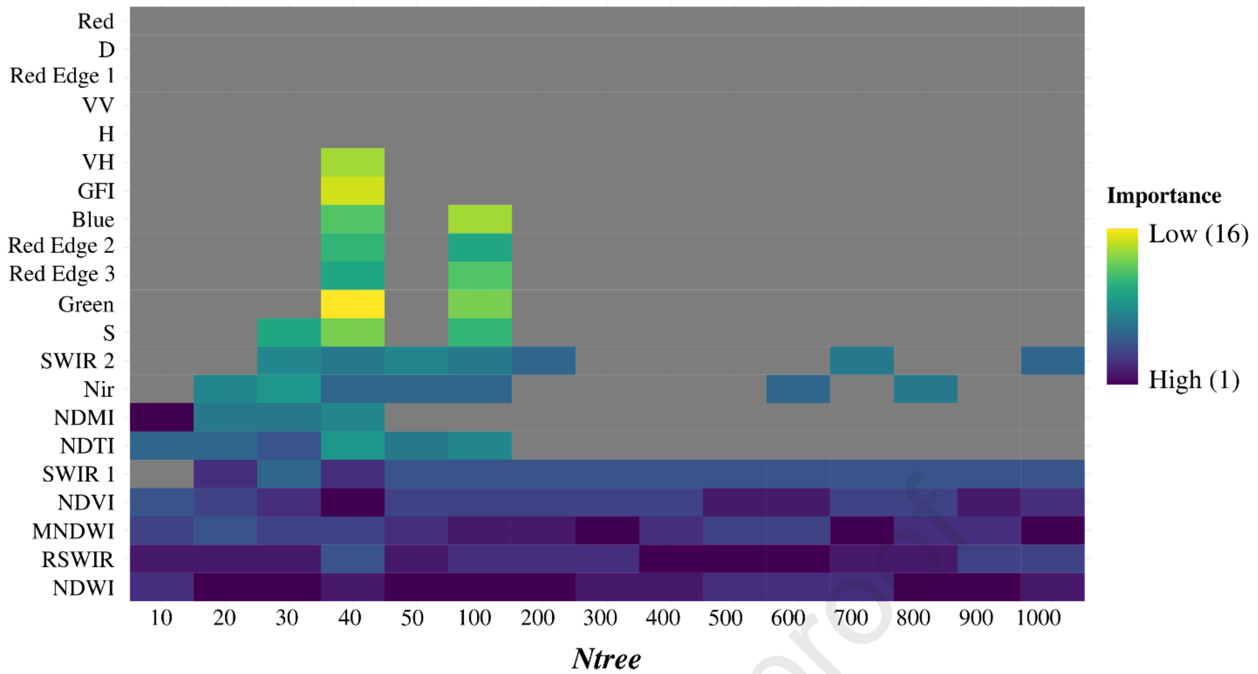

Figure 7. Heatmap for Case Study 2 (CS2) depicting the robustness of predictors for varying *Ntree* and their importance (blue to green colors).

489

490 *CS3: Southern Malawi case study*

491 Five predictors were the most stable being selected almost in each *Ntree* configuration, thus showing
492 a high stability to varying numbers of trees (Figure 8). These are the NDWI, Red Edge 2, Red Edge
493 3, Nir and NDVI, which were always classified among the first eight variables most contributing to
494 the classification. The weakest features were the NDTI, Blue, Green, SWIR 2, S, D, SWIR 1, VV and
495 VH, chosen only one time, while the Red band was never selected. The morphologic features GFI
496 and H contributed to the classification for low values of *Ntree* (not above 40) and were among the
497 first eight most important variables.

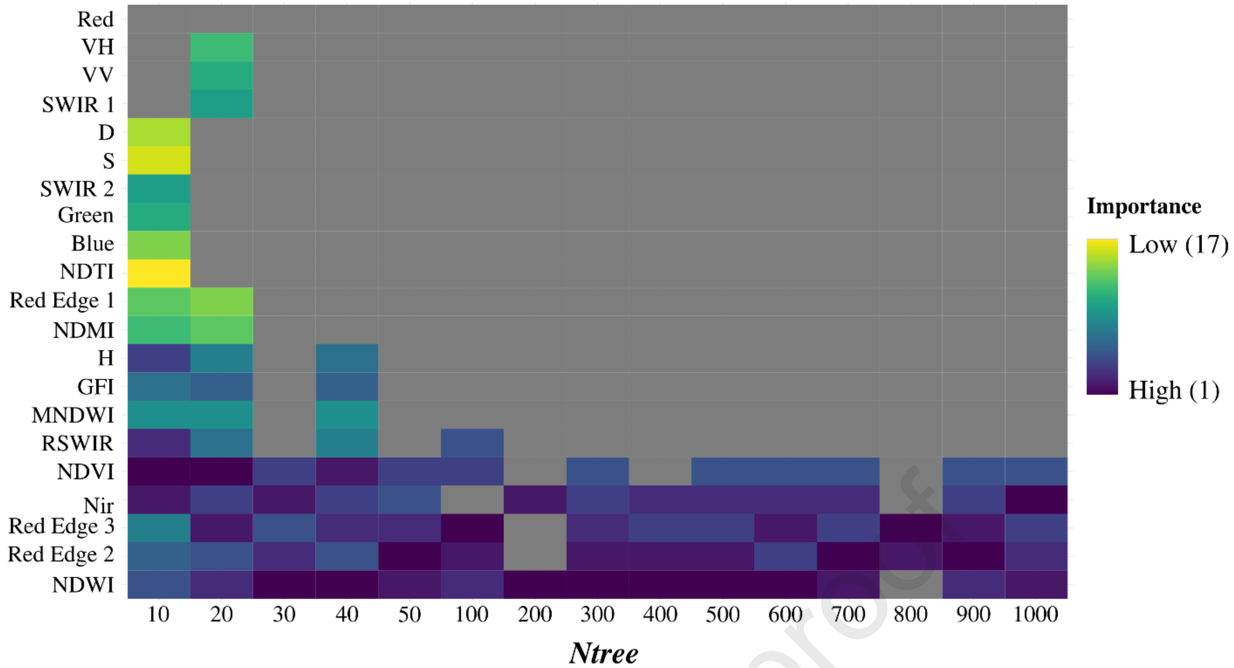

Figure 8. Heatmap for Case Study 3 (CS3) depicting the robustness of predictors for varying Ntree and their importance (blue to green colors).

CS4: Emilia-Romagna case study

The heatmap reported in Figure 9 depicts the robustness of the predictor variables to varying *Ntree* for the Emilia-Romagna study case. Five variables, namely the MNDWI and RSWIR indices, the Sentinel-1 VV polarization, the morphologic features H and the Red Edge 3 Sentinel-2 band showed the highest stability to the parameter being selected for each value of *Ntree*. The former two were also ranked among the first two features mostly contributing to the RF classification (dark blue color), while the latter three had the highest importance only for a very low number of trees, i.e., *Ntree* = 10. In addition, the SWIR 2, SWIR 1, VH, GFI, Nir, Red Edge 2 and NDWI variables showed modest robustness to varying *Ntree* and were also classified among the first 14 most important predictors. The weakest variable was the morphologic feature S chosen only once and ranked as the least important together with the NDTI, D, and the Sentinel-2 bands in the visible range of the electromagnetic spectrum (green to yellow shades in the figure).

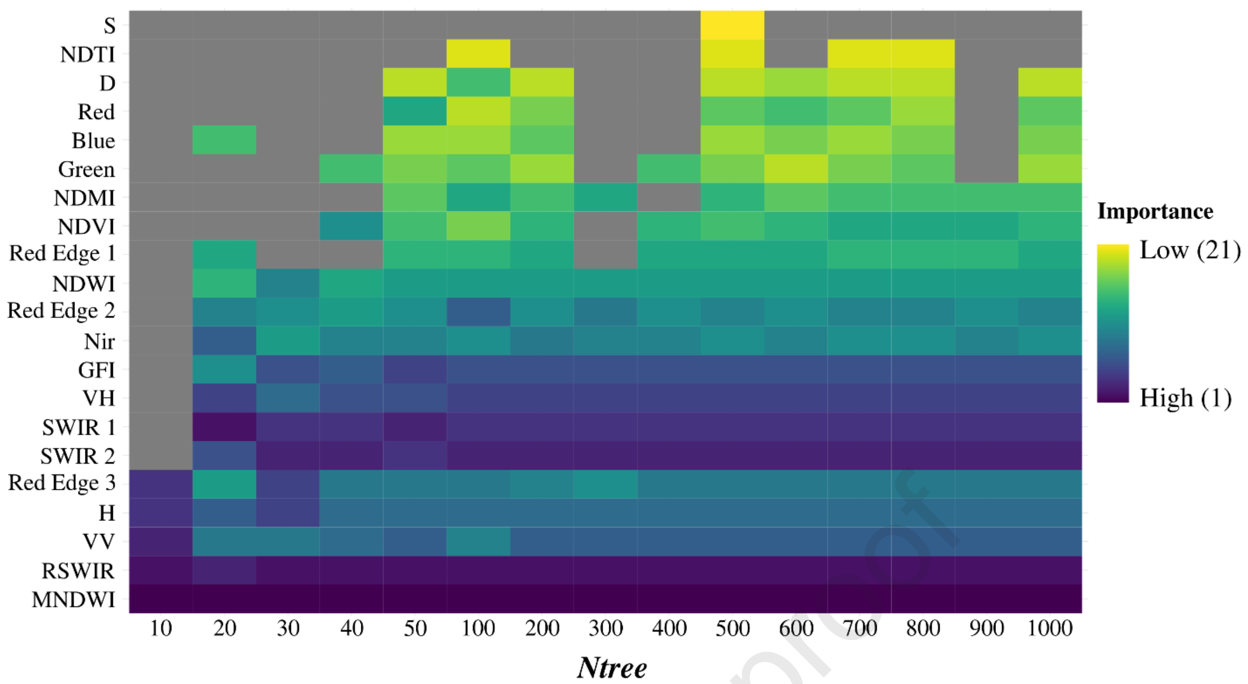

Figure 9. Heatmap for Case Study 4 (CS4) depicting the robustness of predictors for varying Ntree and their importance (blue to green colors).

4.3 Stability of RF classifier and predictors to varying study area

In the four study areas, for each value of *Ntree* the corresponding RF model was applied to the whole scene to derive the classification maps depicting flooded and not flooded classes. Each map was compared with the Copernicus EMSR flood delineation and the RF model minimizing the objective function (*obj*, Equation 1) was identified and selected as the best final classification scheme (Table S.3). In Table 4 details about the optimum models for CS1 to 4 are reported. In particular, for each study area, *obj* values of the optimum RF classification scheme and the corresponding number of trees and optimum predictors are shown. Performance metrics obtained from the pixel-per-pixel comparison between the final flooded areas map and the Copernicus map are also reported in the table.

Regarding CS1, the model minimizing the error function was characterized by 20 trees and 12 predictors (*obj* = 0.3849) and the validation of the detected flood extent showed an OA value of 92.67%, while for CS2 *Ntree* was equal to 100 and *p** to 14 (*obj* = 0.2030) delineating flooded areas with an OA of 94.61%. Very high accuracy was achieved with the best RF model for CS3 (OA = 96.02%), characterized by 20 trees (*obj* = 0.0905) and 14 predictors. Finally, the model with the best *obj* value for CS4 was characterized by *Ntree* = 50 and *p** = 19 (*obj* = 0.2904) and the flooded areas were detected with an OA of 95.82%.

Table 4 also reports the best set of predictors employed in the implementation of the final RF models across the four case studies. It is important to emphasize that these are not listed in order of importance; rather, they are organized to help in identifying the most consistently reliable predictors within distinct settings. Figure S.2 in the supplementary material, instead, provides a visual understanding of the variables selected in each case study, as well as of similarities and differences in their contribution to the classification in the respective RF models.

Among the 21 variables, five multispectral indices, namely the MNDWI, RSWIR, NDMI, NDWI and NDVI, and the SWIR 1 band were the most stable, being selected in all the study areas, while seven predictors in three out of the four (Table 4 and Figure S.2). These include the Sentinel-2 bands Blue, Green, Red Edge 2, Red Edge 3, Nir and SWIR 2, as well as the VH Sentinel-1 polarization. The Red, Red Edge 1, NDTI, VV, GFI and H variables were found in two case studies, while the remaining two morphologic features, i.e., D and S, were specific to individual cases. Finally, five predictors were ranked in the same class of importance in two case studies out of four (Figure S.2). These are the RSWIR index and SWIR 2 band (second and fourth most important features both in CS1 and CS4), SWIR 1 (fifth most contributing variable in both CS1 and CS2), H and Red Edge 1 (respectively classified as the eighth and fourteenth most important predictors both in CS3 and CS4). Finally, a visual comparison between the flooded area maps derived through the selected models in the four case studies and the Copernicus flood delineations is reported in Figure 10. Common areas detected by both the maps are shown in blue, in green areas included in the generated flood maps but not in the reference one (overestimations) are depicted, while in red the areas included in the reference maps but not in the generated flooded areas (underestimations).

Table 4. Results of the pixel-per-pixel comparison between the final Random Forest (RF) models and the Copernicus flood maps for the four case studies (CS1-4). Details about each optimum model are provided: objective function (*obj*) values, number of trees (*Ntree*), optimum predictors (p^*) number, False Positive Rate (R_{fp}), False Negative Rate (R_{fp}), True Positive Rate (R_{fp}), True Negative Rate (R_{fp}), Overall Accuracy (OA), Precision, F-score, and selected predictors subsets.

	CS1	**CS2**	**CS3**	**CS4**
	Ntree=20, p^*=12 *obj* = 0.3849	*Ntree*=100, p^*=14 *obj* = 0.2030	*Ntree*=20, p^*=14 *obj* = 0.0905	*Ntree*=50, p^*=19 *obj* = 0.2904
R_{fp} (%)	1.78	0.88	3.71	3.06
R_{fn}(%)	36.71	19.42	5.34	25.98
R_{tp}(%)	63.29	80.57	94.66	74.03
R_{tn}(%)	98.21	99.12	96.29	96.94
OA (%)	92.67	94.61	96.02	95.82
Precision (%)	86.99	96.72	83.48	55.38
F-score (%)	73.27	87.91	88.72	63.36
Selected Predictors	Blue	Blue		Blue
	Green	Green		Green
	Red			Red
			Red Edge 1	Red Edge 1
		Red Edge 2	Red Edge 2	Red Edge 2
		Red Edge 3	Red Edge 3	Red Edge 3
		Nir	Nir	Nir
	SWIR 1	SWIR 1	SWIR 1	SWIR 1
	SWIR 2	SWIR 2		SWIR 2
	MNDWI	MNDWI	MNDWI	MNDWI
	RSWIR	RSWIR	RSWIR	RSWIR
	NDMI	NDMI	NDMI	NDMI
	NDWI	NDWI	NDWI	NDWI
	NDTI	NDTI		
	NDVI	NDVI	NDVI	NDVI
			VV	VV
	VH		VH	VH
			GFI	GFI
			H	H
				D
		S		

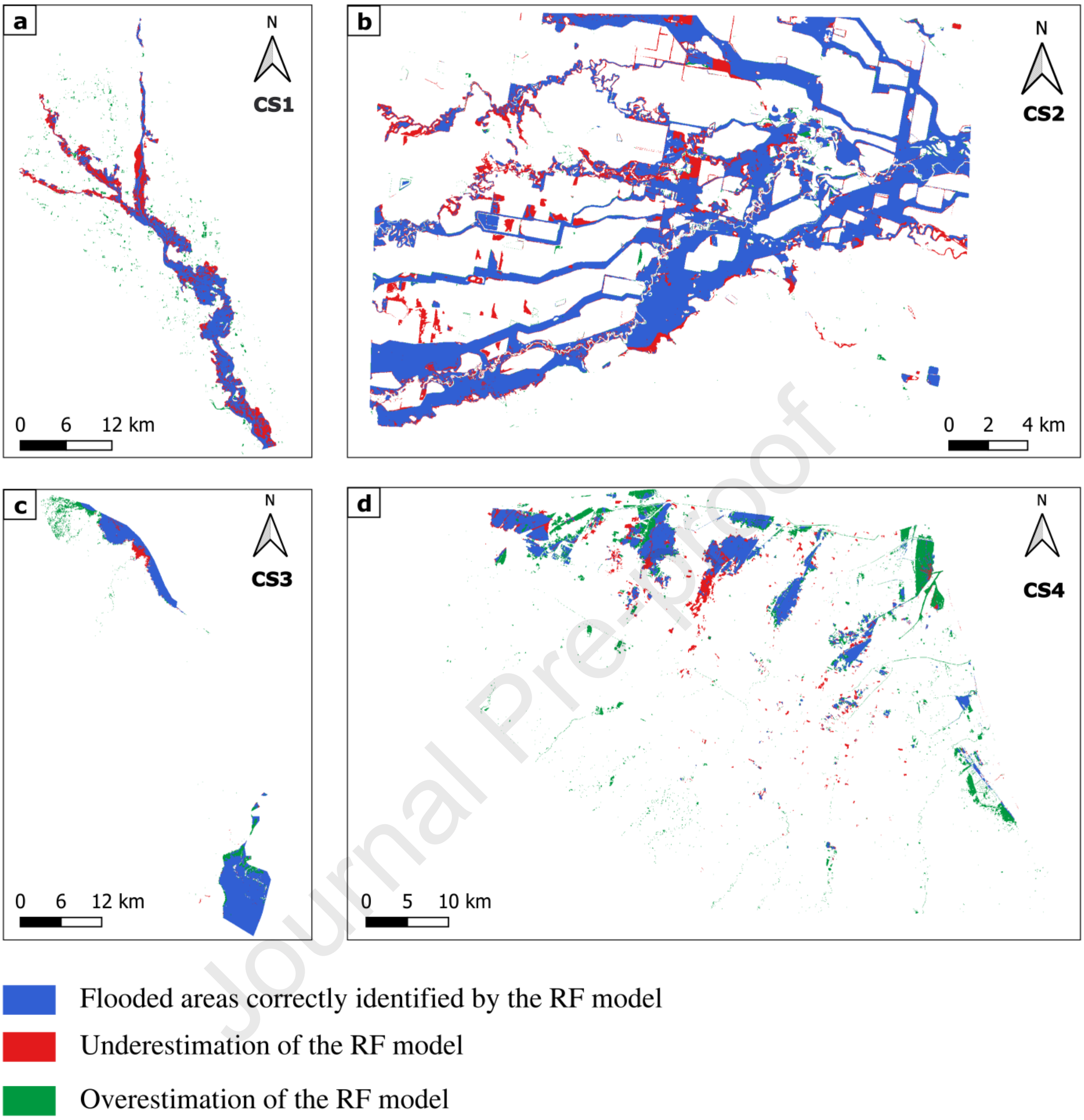

Figure 10. Classification results derived from the implementation of the optimal Random Forest (RF) models in the four case studies : a) Sesia River (CS1); b) Wee Waa (CS2); c) Southern Malawi (CS3) and d) Emilia-Romagna (CS4).

5. Discussion

In this work, RF classification capabilities for flood mapping using a multi-source dataset were evaluated. Predictors included morphologic descriptors, Sentinel-2 bands, derived multispectral indices, and Sentinel-1 polarizations. Rather than focusing mainly on the algorithm accuracies, which have been shown herein and in previous works (e.g., Billah *et al.* 2023) to be as higher as 90%, the primary objective of this study was to carry out an in-depth investigation on the predictive power of several input variables, their robustness and stability to varying training sample sizes and RF parameters, as well as to different contextual settings.

RF classification accuracies have been long assessed in a variety of research experiments and under different parameterizations (e.g., Gislason *et al.* 2006, Wang *et al.* 2015, Ghansah *et al.* 2021, Ramezan *et al.* 2021, Billah *et al.* 2023), though no studies have specifically examined the sensitivity of predictive variables to the algorithm architecture.

5.1 Stability of predictors to varying samples size

A first analysis of the stability of predictors to varying training samples sizes was carried out using the RF classifier on the Sesia River case study (CS1). Five different sizes of pixels per class, *n*, were considered based on recommendations reported in the literature (i.e., at least 50 samples per class, 10*p*, 20*p*, 30*p*, 45*p*, with *p* indicating the number of predictors). Results concerning the model accuracies confirmed that at least 10*p* training samples per class should be used to achieve good performances (in the proposed analysis accuracy above 99% was registered starting from *n*=210 pixels per class, i.e., *n*=10*p*).
Regarding the stability of predictors for each sample size, the internal RF MDA measure showed that three multispectral indices are quite insensitive to the training sizes, namely the MNDWI, RSWIR and NDMI. In particular, the former is the most stable variable, being classified as the third most important feature for each value of *n*, while the latter two can either be ranked as first or second variables. In addition, the NDMI is shown to be the predictor with the highest mean relative importance (equal or over 90% in each sample size configuration). Similarly, the Sentinel-2 bands SWIR 1 and SWIR 2 show a certain stability, being ranked as either the fourth or fifth most important variables for varying *n*. Concerning the morphologic descriptors, the GFI and H exhibit quite constant behavior for changing training set size, even if are not among the most important variables. If results are rather stable for higher *n* values, they slightly differ when the dimension of the training dataset is small (i.e., *n* = 50). A higher variability of the predictors ranking (and MDA measure) is, in fact, observed. On the one hand, a lower *n* led to a lower number of features classified as most important (e.g., nine predictors are ranked among the first ten most important variables for *n*=50 and 10 for all the other cases, while a MDA at least equal to 10% was observed for six predictors in the case *n*=50 and for seven or eight variables in the other cases), but the accuracy value was lower. On the other hand, some predictor variables acquired more relevance than with higher sample sizes. Indeed, the contribution of the local slope, S, is more significant when *n*=50. Likewise, the Sentinel-1 VV and VH polarization show a higher importance for small sample sizes. Such behavior may be explained considering that ROIs were collected manually by digitizing sample polygons based on visual interpretation of Sentinel-2 scenes. Therefore, they mainly reflect Sentinel-2 related variables patterns, which do not necessarily correspond to those of morphologic features and Sentinel-1 bands. If considering for example the predictor S, in which variations at the pixel scale reflect changes in local slope, smaller polygon dimensions or numbers imply a higher chance to capture the local patterns.

5.2 Robustness of predictors for varying Ntree

Considering different case studies, the robustness of predictors was assessed under different configurations of number of trees. In this case, the algorithm was trained using at least 10*p* samples per class and the RFE method for features selection was applied to build RF models exploiting only the *p** most important variables. Such a procedure aids in reducing data dimensionality excluding those that do not significantly contribute to the classification. In general, results showed that RF performances obtained with different subset sizes are stable with respect to changing number of trees. Nevertheless, no clear patterns across the case studies can be detected in terms of relationships between p^* and model accuracies or *Ntree*. Every study area and flooding event needs an exploratory assessment of the variables and training parameters best suited for predicting the flood extent. Regarding the stability of predictors to varying *Ntree*, the RSWIR index was found to be the most robust and insensitive to the parameter as it was selected as one of the best predictors (among the first five most important variables) for each value of *Ntree* and in three case studies out of four (CS1, CS2, CS4). The MNDWI also showed moderate stability, being chosen in every *Ntree* configuration in two case studies out of four (CS32, CS4) and among the first five most contributing variables in three case studies (CS1, CS2, CS4). This confirms that the MNDWI is a reliable index for flood mapping (see e.g., Albertini *et al.* 2022a), being not only insensitive to the training sample size and to some extent to the algorithm architecture (i.e., the *Ntree* parameter) but also one of the best predictor in different study areas.

5.3 Stability of RF classifier and predictors to varying study area

The investigation on the stability of the RF classifier and predictors to varying study areas was carried out by identifying the best model in terms of minimization of the error function as defined in Equation 1 and through the comparison with the reference maps from the Copernicus EMSR. It is interesting to note that for every study area no less than 12 predictors and a number of trees between 20 and 100 were found necessary for an accurate delineation (above 92% overall accuracy) of flooded areas (p^*=14 in two out of four case studies). In most cases (three out of four), the best combination of predictors included Sentinel-2 bands in the visible range of the electromagnetic spectrum and multispectral indices and Sentinel-1 polarizations. This is in agreement with findings from the literature, according to which spectral indices are more stable than other variables when applied to new study areas (Belgiu and Drăguţ 2016). Regarding morphologic descriptors, the GFI and H also appear to be robust predictors significantly contributing to the classification. Differences between the case studies, especially in the selection of geomorphic predictors were mainly linked to limitations related to the available DEM. In fact, some errors regarding deviations of the DEM-derived hydrologic network from the actual river flow were observed in the Sesia River and Wee Waa case studies, most likely due to active alluvial and erodible rived beds that during floods lead to changes in the watercourse and the creation or reactivation of channels (Fugazza *et al.* 2008, Wray 2009). Such deviations affect the estimation of geomorphic descriptors, which inevitably cannot capture the morphology of the territory with fidelity. Figure S.3 in the supplementary material aims to explain this mechanism, by depicting the flood extents as derived from the RF models, the river network extracted from the DEM and the GFI computed based on it. Whenever differences between the river channels and the drainage system at the time of the floods exist (Figure S.3(a.1), (a.2) and (b.2)), deviations between the GFI configuration and the actual flood patterns exist as well and geomorphic features become less relevant for the classification (CS1 and 2). If the GFI description of floodable areas better matches the flood imprint (Figure S.3(b.1), (c.1), (c.2), (d.1) and (d.2)), as follows from a more accurate representation of the river network, hence these contribute to the classification (CS3 and 4). This obviously highlights the need for updated morphological descriptors which may become rapidly outdated, especially in alluvial systems, where every flood may potentially lead to significant modifications of the water course trajectories and position. Issues related to DEMs accuracy and hydraulic consistency of the extracted river channels have also been recently considered by Magnini *et al.* (2023) who highlighted the need for reliable river network extraction to effectively use DEM-based flood hazard indicators.

In conclusion, this work proved that minor changes to the RF algorithm allow its transferability to different study areas. In addition, the findings of the analyses underlined how the joint use of both optical and SAR features, as well as geomorphic descriptors, allows for achieving a fair delineation of flooded areas with minimum errors. In particular, the use of geomorphic data can help reduce false alarms and missed interventions and solve some issues related to satellite imagery, such as those linked to the presence of vegetation, turbid water, clouds, shadow areas, or to the time span between the satellite overpass and the flood peak, which can reduce the ability of interpreting and reconstructing the phenomena. On the other hand, morphologic features strictly depend on the input elevation data. In the current work, a global and open-source product was used (i.e., the SRTM DEM) to provide a unified and homogeneous modelling framework among case studies. Further future investigations could concern the use of national and local data for the classification, or the selection of DEMs tailored to specific requirements and geographical features, as highlighted by Moges et al. (2023).

It is worth underlining that the current study was carried out considering the Sentinel-2 imagery as the reference for collecting training samples, which may have somewhat influenced the outcomes of the implemented analyses and favored Sentinel-2-related features over the other variables. Furthermore, some over or underestimations observed in the final classified flood maps (Figure 10) might also be linked to the (dis)agreement between the collected ROIs pixels and the Copernicus delineations, as illustrated, for example, in Figure S.1 for CS1. In fact, the visual interpretation of Sentinel-2 scenes can lead to some misinterpretations. However, considering that Copernicus EMSR maps are also obtained through a mixture of photointerpretation and classification, the implemented methodology and comparison can be considered robust.

6. Conclusions

In this study, the RF algorithm was employed for flood mapping using a multi-source dataset. This included satellite-based data (both optical and SAR) and morphologic features to ultimately assess the robustness of the algorithm and predictors to varying training schemes and landscape contexts.

Overall, generalizations between different study areas are difficult to be made and the identification of predictor variables suitable for different settings requires ad-hoc investigations. Every flood event is dominated by the combination of several factors (turbidity, initial soil moisture conditions, land cover and vegetation status at the time of the flood, and geomorphologic dynamics), which makes flood mapping case specific. Nonetheless, some key conclusions can be drawn from the current work which can be summarized as follows:

- The MNDWI is one of the most powerful variables for flooded areas detection, as it was proven to be highly stable to changing training dataset size, number of trees in the RF algorithm and study areas. Likewise, the RSWIR index was found to be a robust index to varying *Ntree* and context.

- Morphologic descriptors can be important if updated morphological data are available, otherwise they do not significantly contribute to the classification also because of errors in the DEM.
- In all the considered study areas, the RF accuracy across different subset sizes of the predictor variables was quite stable for varying *Ntree*. Furthermore, a RF model built with no less than 12 predictors was found to provide the best flood delineation in terms of reduction of false positives (overestimation errors) and false negatives (underestimation errors).
- RF classifier exhibits very high predictive capabilities in flooded areas mapping with accuracy values above 92% especially when the synergy between Sentinel-2, Sentinel-1 and geomorphic data (mainly the GFI and H features) is exploited.

The study provided an exploration of the predictive power of a variety of predictors used in flooded area mapping which can straightforwardly be incorporated in RF models. Further investigations may be needed in order for the results to be confirmed and the possibility of using high-resolution satellite images may be explored in future studies when those imagery are timely available. Nonetheless, this work proved that the identification of the most robust and stable variables, as well as the synergetic use of multi-source data, allows for enhancing classification accuracy and for they transferability to new study areas.

Author contributions

Cinzia Albertini: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing - Original draft preparation. **Andrea Gioia**: Investigation, Supervision, Writing – review & editing. **Vito Iacobellis**: Investigation, Supervision, Writing – review & editing. **George Petropoulos**: Investigation, Supervision, Writing – review & editing. **Salvatore Manfreda**: Software, Investigation, Supervision, Writing – review & editing.

Competing interests

The authors declare that they have no conflict of interest.

References

Agenzia Regionale per la Prevenzione e la Protezione dell'Ambiente Piemonte, 2020. *EVENTO DEL 2-3 OTTOBRE 2020*. Available at: https://www.arpa.piemonte.it/news/rapporto-evento-alluvionale-del-2-3-ottobre-2020-in-piemonte (accessed on 24 February 2022)

Albertini, C., Gioia, A., Iacobellis, V., and Manfreda, S., 2022. Detection of Surface Water and Floods with Multispectral Satellites. *Remote Sensing*, 14 (23), 6005. https://doi.org/10.3390/rs14236005

Albertini, C., Miglino, D., Iacobellis, V., De Paola, F., and Manfreda, S., 2022. Delineation of flood-prone areas in cliffed coastal regions through a procedure based on the geomorphic flood index. *Journal of Flood Risk Management*, 15 (1), e12766. https://doi.org/10.1111/jfr3.12766

Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., de Roo, A., Salamon, P., Wyser, K., and Feyen, L., 2017. Global projections of river flood risk in a warmer world. *Earth's Future*, 5 (2), 171–182. https://doi.org/10.1002/2016EF000485

Balenzano, A., Mattia, F., Satalino, G., Lovergine, F.P., Palmisano, D., Peng, J., Marzahn, P., Wegmüller, U., Cartus, O., and Dąbrowska-Zielińska, K., 2021. Sentinel-1 soil moisture at 1 km resolution: a validation study. *Remote Sensing of Environment*, 263, 112554. https://doi.org/10.1016/j.rse.2021.112554

Balenzano, A., Satalino, G., Lovergine, F.P., D'Addabbo, A., Palmisano, D., Grassi, R., Ozalp, O., Mattia, F., Nafría García, D., and Paredes Gómez, V., 2022. Sentinel-1 and Sentinel-2 Data to Detect Irrigation Events: Riaza Irrigation District (Spain) Case Study. *Water*, 14 (19), 3046. https://doi.org/10.3390/w14193046

Bates, P.D., Wilson, M.D., Horritt, M.S., Mason, D.C., Holden, N., and Currie, A., 2006. Reach scale floodplain inundation dynamics observed using airborne synthetic aperture radar imagery: Data analysis and modelling. *Journal of Hydrology*, 328 (1–2), 306–318. https://doi.org/10.1016/j.jhydrol.2005.12.028

Belgiu, M. and Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Billah, M., Islam, A.K.M.S., Mamoon, W. Bin, and Rahman, M.R., 2023. Random forest classifications for landuse mapping to assess rapid flood damage using Sentinel-1 and Sentinel-2 data. *Remote Sensing Applications: Society and Environment*, 30, 100947. https://doi.org/10.1016/j.rsase.2023.100947

Blöschl, G., 2022. Three hypotheses on changing river flood hazards. *Hydrology and Earth System Sciences*, 26 (19), 5015–5033. https://doi.org/10.5194/hess-26-5015-2022Blöschl, G., Hall, J., Viglione, A., Perdigão, R.A.P., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G.T., and Bilibashi, A., 2019. Changing climate both increases and decreases European river floods. *Nature*, 573 (7772), 108–111. https://doi.org/10.1038/s41586-019-1495-6

Breiman, L., 2001. Random forests. *Machine learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cao, H., Zhang, H., Wang, C., and Zhang, B., 2019. Operational flood detection using Sentinel-1 SAR data over large areas. *Water*, 11 (4), 786. https://doi.org/10.3390/w11040786

Cenci, L., Boni, G., Pulvirenti, L., Squicciarino, G., Gabellani, S., Gardella, F., Pierdicca, N., and Chini, M., 2017. Monitoring reservoirs' water level from space for flood control applications. A case study in the Italian Alpine region. *In*: *2017 IEEE International Geoscience and Remote*

Sensing Symposium (IGARSS). IEEE, 5617–5620. https://doi.org/10.1109/IGARSS.2017.8128279

Chignell, S.M., Anderson, R.S., Evangelista, P.H., Laituri, M.J., and Merritt, D.M., 2015. Multi-Temporal Independent Component Analysis and Landsat 8 for Delineating Maximum Extent of the 2013 Colorado Front Range Flood. *Remote Sensing* . *Remote Sensing* . https://doi.org/10.3390/rs70809822

Colditz, R.R., 2015. An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sensing*, 7 (8), 9655–9681. https://doi.org/10.3390/rs70809655

Copernicus Emergency Management Service. (n.d.). (© European Union, 2012-2023). Available online: https://emergency.copernicus.eu/ (accessed on 18 July 2023)

Copernicus Emergency Management Service (© 2020 European Union), EMSR468. Available online: https://emergency.copernicus.eu/mapping/list-of-components/EMSR468 (accessed on 12 September 2022)

Copernicus Emergency Management Service (© 2021 European Union), EMSR554. Available online: https://emergency.copernicus.eu/mapping/list-of-components/EMSR554 (accessed on 20 March 2023)

Copernicus Emergency Management Service (© 2022 European Union), EMSR561. Available online: https://emergency.copernicus.eu/mapping/list-of-components/EMSR561 (accessed on 17 April 2023)

Copernicus Emergency Management Service (© 2023 European Union), EMSR664. Available online: https://rapidmapping.emergency.copernicus.eu/EMSR664/download (accessed on 18 July 2023)

Degiorgis, M., Gnecco, G., Gorni, S., Roth, G., Sanguineti, M., and Taramasso, A.C., 2012. Classifiers for the detection of flood-prone areas using remote sensed elevation data. *Journal of Hydrology*, 470–471, 302–315. https://doi.org/10.1016/j.jhydrol.2012.09.006

Delforge, D., Below, R., Wathelet, V., Jones, R., Tubeuf, S., and Speybroek, N., 2022. *2021 Disasters in Numbers*. Brussels. Available at: https://cred.be/sites/default/files/2021_EMDAT_report.pdf (accessed on 26 June 2023)

Dutsenwai, H.S., Ahmad, B. Bin, Mijinyawa, A., and Tanko, A.I., 2016. Fusion of SAR images for flood extent mapping in northern peninsula Malaysia. *International Journal of Advanced and Applied Sciences*, 3 (12), 37–48. http://dx.doi.org/10.21833/ijaas.2016.12.006

Earthdata Search—NASA. Available at: https://search.earthdata.nasa.gov/search (accessed on 22 June 2023).

European Commission, 2021. *Current practice in flood risk management in the European Union*. Luxembourg: Publications Office of the European Union. Available online: https://op.europa.eu/en/publication-detail/-/publication/21d8c5c2-2199-11ec-bd8e-01aa75ed71a1/language-en/format-PDF/source-232431484 (Accessed on 7 September 2023)

Ezzine, A., Darragi, F., Rajhi, H., and Ghatassi, A., 2018. Evaluation of Sentinel-1 data for flood mapping in the upstream of Sidi Salem dam (Northern Tunisia). *Arabian Journal of Geosciences*, 11, 1–9. https://doi.org/10.1007/s12517-018-3505-7

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., and Roth, L., 2007. The shuttle radar topography mission. *Reviews of geophysics*, 45 (2). https://doi.org/10.1029/2005RG000183

Field, C.B., Barros, V., Stocker, T.F., Qin, D., Dokken, D.J., Ebi, K.L., Mastrandrea, M.D., Mach, K.J., Plattner, G.-K., Allen, S.K., Tignor, M., and Midgley, P.M., 2012. *IPCC, 2012: Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press. Available at: https://www.ipcc.ch/report/managing-the-risks-of-extreme-events-and-disasters-to-advance-climate-change-adaptation/ (Accessed on 22 August 2023)

Filipponi, F., 2019. Sentinel-1 GRD preprocessing workflow. *In*: *International Electronic Conference on Remote Sensing*. MDPI, 11. https://doi.org/10.3390/ECRS-3-06201

Fugazza, M., Magri, P., Natale, L., and Savi, F., 2008. The October 2000 flood at the Po river and Sesia river junction. *WIT Transactions on Ecology and the Environment*, 118, 37–45. http://dx.doi.org/10.2495/FRIAR080041

Gao, B., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58 (3), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Gašparović, M. and Klobučar, D., 2021. Mapping floods in lowland forest using sentinel-1 and sentinel-2 data and an object-based approach. *Forests*, 12 (5), 553. https://doi.org/10.3390/f12050553

Ghansah, B., Nyamekye, C., Owusu, S., and Agyapong, E., 2021. Mapping flood prone and Hazards Areas in rural landscape using landsat images and random forest classification: Case study of Nasia watershed in Ghana. *Cogent Engineering*, 8 (1), 1923384. https://doi.org/10.1080/23311916.2021.1923384

Gislason, P.O., Benediktsson, J.A., and Sveinsson, J.R., 2006. Random forests for land cover classification. *Pattern recognition letters*, 27 (4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Giustarini, L., Hostache, R., Matgen, P., Schumann, G.J.-P., Bates, P.D., and Mason, D.C., 2012. A change detection approach to flood mapping in urban areas using TerraSAR-X. *IEEE transactions on Geoscience and Remote Sensing*, 51 (4), 2417–2430. https://doi.org/10.1109/TGRS.2012.2210901

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z., and Yang, X., 2013. Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. *International Journal of Remote Sensing*, 34 (14), 5166–5186. http://dx.doi.org/10.1080/01431161.2013.788261

Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, 3 (Mar), 1157–1182.

Ham, J., Chen, Y., Crawford, M.M., and Ghosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43 (3), 492–501. http://dx.doi.org/10.1109/TGRS.2004.842481

Hastie, T., Tibshirani, R., Friedman, J.H., and Friedman, J.H., 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Ireland, G., Volpi, M., and Petropoulos, G.P., 2015. Examining the capability of supervised machine learning classifiers in extracting flooded areas from Landsat TM imagery: A case study from a Mediterranean flood. *Remote sensing*, 7 (3), 3372–3399. https://doi.org/10.3390/rs70303372

Kim, S., Zhang, R., Pham, H., and Sharma, A., 2019. A review of satellite-derived soil moisture and its usage for flood estimation. *Remote Sensing in Earth Systems Sciences*, 2, 225–246. https://doi.org/10.1007/s41976-019-00025-7

Kohavi, R. and John, G.H., 1997. Wrappers for feature subset selection. *Artificial intelligence*, 97 (1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., and Team, R.C., 2020. Package 'caret'. *The R Journal*, 223 (7).

Lacaux, J.P., Tourre, Y.M., Vignolles, C., Ndione, J.A., and Lafaye, M., 2007. Classification of ponds from high-spatial resolution remote sensing: Application to Rift Valley Fever epidemics in Senegal. *Remote sensing of environment*, 106 (1), 66–74. https://doi.org/10.1016/j.rse.2006.07.012

Lawrence, R.L., Wood, S.D., and Sheley, R.L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). *Remote Sensing of Environment*, 100 (3), 356–362. https://doi.org/10.1016/j.rse.2005.10.014

Lee, J.-S., Jurkevich, L., Dewaele, P., Wambacq, P., and Oosterlinck, A., 1994. Speckle filtering of synthetic aperture radar images: A review. *Remote sensing reviews*, 8 (4), 313–340. https://doi.org/10.1080/02757259409532206

Magnini, A., Lombardi, M., Bujari, A., Mattivi, P., Shustikova, I., Persiano, S., Patella, M., Bitelli, G., Bellavista, P., and Lo Conti, F., 2023. Geomorphic flood hazard mapping: from floodplain delineation to flood hazard characterization. *Hydrological Sciences Journal*, 1–16. https://doi.org/10.1080/02626667.2023.2269909

Magnini, A., Lombardi, M., Persiano, S., Tirri, A., Lo Conti, F., and Castellarin, A., 2022. Machine-learning blends of geomorphic descriptors: value and limitations for flood hazard assessment across large floodplains. *Natural Hazards and Earth System Sciences*, 22 (4), 1469–1486. https://doi.org/10.5194/nhess-22-1469-2022

Manfreda, S., Di Leo, M., and Sole, A., 2011. Detection of flood-prone areas using digital elevation models. *Journal of Hydrologic Engineering*, 16 (10), 781–790. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000367

Manfreda, S., Nardi, F., Samela, C., Grimaldi, S., Taramasso, A.C., Roth, G., and Sole, A., 2014. Investigation on the use of geomorphic approaches for the delineation of flood prone areas. *Journal of hydrology*, 517, 863–876. https://doi.org/10.1016/j.jhydrol.2014.06.009

Manfreda, S., Samela, C., Gioia, A., Consoli, G.G., Iacobellis, V., Giuzio, L., Cantisani, A., and Sole, A., 2015. Flood-prone areas assessment using linear binary classifiers based on flood maps obtained from 1D and 2D hydraulic models. *Natural Hazards*, 79 (2), 735–754. https://doi.org/10.1007/s11069-015-1869-5

Mason, D.C., Speck, R., Devereux, B., Schumann, G.J.-P., Neal, J.C., and Bates, P.D., 2009. Flood detection in urban areas using TerraSAR-X. *IEEE transactions on geoscience and remote sensing*, 48 (2), 882–894. https://doi.org/10.1109/TGRS.2009.2029236

Mather, P.M. and Koch, M., 2011. *Computer processing of remotely-sensed images: an introduction*. John Wiley & Sons.

Maxwell, A.E., Warner, T.A., and Fang, F., 2018. Implementation of machine-learning classification in remote sensing: An applied review. *International journal of remote sensing*, 39 (9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

McFeeters, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17 (7), 1425–1432. https://doi.org/10.1080/01431169608948714

Memon, A.A., Muhammad, S., Rahman, S., and Haq, M., 2015. Flood monitoring and damage assessment using water indices: A case study of Pakistan flood-2012. *The Egyptian Journal of Remote Sensing and Space Science*, 18 (1), 99–106. http://dx.doi.org/10.1016/j.ejrs.2015.03.003

Merz, B., Hall, J., Disse, M., and Schumann, A., 2010. Fluvial flood risk management in a changing world. *Natural Hazards and Earth System Sciences*, 10 (3), 509–527. https://doi.org/10.5194/nhess-10-509-2010

Millard, K. and Richardson, M., 2015. On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote sensing*, 7 (7), 8489–8515. https://doi.org/10.3390/rs70708489

Moges, D.M., Virro, H., Kmoch, A., Cibin, R., Rohith, A.N., Martínez-Salvador, A., Conesa-García, C., Uuemaa, E., 2023. How does the choice of DEMs affect catchment hydrological modeling? Sci. Total Environ. 164627. https://doi.org/10.1016/j.scitotenv.2023.164627

Mosavi, A., Ozturk, P., and Chau, K., 2018. Flood prediction using machine learning models: Literature review. *Water*, 10 (11), 1536. https://doi.org/10.3390/w10111536

Munasinghe, D., Cohen, S., Huang, Y., Tsang, Y., Zhang, J., and Fang, Z., 2018. Intercomparison of satellite remote sensing-based flood inundation mapping techniques. *JAWRA Journal of the American Water Resources Association*, 54 (4), 834–846. https://doi.org/10.1111/1752-1688.12626

Nandi, I., Srivastava, P.K., and Shah, K., 2017. Floodplain mapping through support vector machine and optical/infrared images from Landsat 8 OLI/TIRS sensors: Case study from Varanasi. *Water Resources Management*, 31 (4), 1157–1171. https://doi.org/10.1007/s11269-017-1568-y

Van Niel, T.G., McVicar, T.R., and Datt, B., 2005. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. *Remote sensing of environment*, 98 (4), 468–480. https://doi.org/10.1016/j.rse.2005.08.011

Nobre, A.D., Cuartas, L.A., Momo, M.R., Severo, D.L., Pinheiro, A., and Nobre, C.A., 2016. HAND contour: a new proxy predictor of inundation extent. *Hydrological Processes*, 30 (2), 320–333. https://doi.org/10.1002/hyp.10581

Notti, D., Giordan, D., Caló, F., Pepe, A., Zucca, F., and Galve, J.P., 2018. Potential and Limitations of Open Satellite Data for Flood Mapping. *Remote Sensing*. http://dx.doi.org/10.3390/rs10111673

Oberstadler, R., Hönsch, H., and Huth, D., 1997. Assessment of the mapping capabilities of ERS-1 SAR data for flood mapping: a case study in Germany. *Hydrological processes*, 11 (10), 1415–1425.

Petropoulos, G.P., Kontoes, C., and Keramitsoglou, I., 2011. Burnt area delineation from a uni-temporal perspective based on Landsat TM imagery classification using Support Vector Machines. *International Journal of Applied Earth Observation and Geoinformation*, 13 (1), 70–80. https://doi.org/10.1016/j.jag.2010.06.008

Piper, J., 1992. Variability and bias in experimentally measured classifier error rates. *Pattern Recognition Letters*, 13 (10), 685–692. https://doi.org/10.1016/0167-8655(92)90097-J

Prasad, A.M., Iverson, L.R., and Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9, 181–199.

https://doi.org/10.1007/s10021-005-0054-1

Pulvirenti, L., Chini, M., Pierdicca, N., Guerriero, L., and Ferrazzoli, P., 2011. Flood monitoring using multi-temporal COSMO-SkyMed data: Image segmentation and signature interpretation. *Remote Sensing of Environment*, 115 (4), 990–1002. https://doi.org/10.1016/j.rse.2010.12.002

Radulović, M., Brdar, S., Pejak, B., Lugonja, P., Athanasiadis, I., Pajević, N., Pavić, D., and Crnojević, V., 2023. Machine learning-based detection of irrigation in Vojvodina (Serbia) using Sentinel-2 data. *GIScience & Remote Sensing*, 60 (1), 2262010. https://doi.org/10.1080/15481603.2023.2262010

Ramezan, C.A., Warner, T.A., Maxwell, A.E., and Price, B.S., 2021. Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. *Remote Sensing*, 13 (3), 368. https://doi.org/10.3390/rs13030368

Refice, A., D'Addabbo, A., Lovergine, F.P., Tijani, K., Morea, A., Nutricato, R., Bovenga, F., and Nitti, D.O., 2018. Monitoring flood extent and area through multisensor, multi-temporal remote sensing: The Strymonas (Greece) River Flood. *Flood Monitoring through Remote Sensing*, 101–113. https://doi.org/10.1007/978-3-319-63959-8_5

Refice, A., Zingaro, M., D'Addabbo, A., and Chini, M., 2020. Integrating C-and L-band SAR imagery for detailed flood monitoring of remote vegetated areas. *Water*, 12 (10), 2745. https://doi.org/10.3390/w12102745

Rogers, A.S. and Kearney, M.S., 2004. Reducing signature variability in unmixing coastal marsh Thematic Mapper scenes using spectral indices. *International Journal of Remote Sensing*, 25 (12), 2317–2335. https://doi.org/10.1080/01431160310001618103

Rouse, J.W., Haas, R.H., Schell, J.A., and Deering, D.W., 1974. *Monitoring Vegetation Systems in the Great Plains with ERTS*. Greenbelt, ON, Canada.

Samela, C., Manfreda, S., Paola, F. D., Giugni, M., Sole, A., & Fiorentino, M. 2016. DEM-based approaches for the delineation of flood-prone areas in an ungauged basin in Africa. J*ournal of Hydrologic Engineering*, 21(2), 06015010. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001272

Samela, C., Troy, T.J., and Manfreda, S., 2017. Geomorphic classifiers for flood-prone areas delineation for data-scarce environments. *Advances in Water Resources*, 102, 13–28. https://doi.org/10.1016/j.advwatres.2017.01.007

Schumann, G., Giustarini, L., Tarpanelli, A., Jarihani, B., and Martinis, S., 2022. Flood Modeling and Prediction Using Earth Observation Data. *Surveys in Geophysics*, 1–26. https://doi.org/10.1007/s10712-022-09751-y

Schumann, G.J.-P., Brakenridge, G.R., Kettner, A.J., Kashif, R., and Niebuhr, E., 2018. Assisting flood disaster response with earth observation data and products: A critical assessment. *Remote Sensing*, 10 (8). https://doi.org/10.3390/rs10081230

Sentinel Scientific Data Hub. Available at: https://scihub.copernicus.eu/ (Accessed on 22 June 2023)

Shen, X., Wang, D., Mao, K., Anagnostou, E., and Hong, Y., 2019. Inundation Extent Mapping by Synthetic Aperture Radar: A Review. *Remote Sensing*. https://doi.org/10.3390/rs11070879

Tavares da Costa, R., Manfreda, S., Luzzi, V., Samela, C., Mazzoli, P., Castellarin, A., and Bagli, S., 2019. A web application for hydrogeomorphic flood hazard mapping. *Environmental Modelling & Software*, 118, 172–186. https://doi.org/10.1016/j.envsoft.2019.04.010

Tavares da Costa, R., Zanardo, S., Bagli, S., Hilberts, A.G.J., Manfreda, S., Samela, C., and Castellarin, A., 2020. Predictive Modeling of Envelope Flood Extents Using Geomorphic and Climatic-Hydrologic Catchment Characteristics. *Water Resources Research*, 56 (9), e2019WR026453. https://doi.org/10.1029/2019WR026453

Theobald, D.M., Harrison-Atlas, D., Monahan, W.B., and Albano, C.M., 2015. Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning. *PloS one*, 10 (12), e0143619. https://doi.org/10.1371/journal.pone.0143619

Tsatsaris, A., Kalogeropoulos, K., Stathopoulos, N., Louka, P., Tsanakas, K., Tsesmelis, D.E., Krassanakis, V., Petropoulos, G.P., Pappas, V., and Chalkias, C., 2021. Geoinformation technologies in support of environmental hazards monitoring under climate change: An extensive review. *ISPRS International Journal of Geo-Information*, 10 (2), 94. https://doi.org/10.3390/ijgi10020094

Volpi, M., Petropoulos, G.P., and Kanevski, M., 2013. Flooding extent cartography with Landsat TM imagery and regularized kernel Fisher's discriminant analysis. *Computers & Geosciences*, 57, 24–31. https://doi.org/10.1016/j.cageo.2013.03.009

Wagenaar, D., Curran, A., Balbi, M., Bhardwaj, A., Soden, R., Hartato, E., Mestav Sarica, G., Ruangpan, L., Molinario, G., and Lallemant, D., 2020. Invited perspectives: How machine learning will change flood risk and impact assessment. *Natural Hazards and Earth System Sciences*, 20 (4), 1149–1161. https://doi.org/10.5194/nhess-20-1149-2020

Wang, Y., 2004a. Seasonal change in the extent of inundation on floodplains detected by JERS-1 Synthetic Aperture Radar data. *International Journal of Remote Sensing*, 25 (13), 2497–2508. https://doi.org/10.1080/01431160310001619562

Wang, Y., 2004b. Using Landsat 7 TM data acquired days after a flood event to delineate the maximum flood extent on a coastal floodplain. *International Journal of Remote Sensing*, 25 (5), 959–974. https://doi.org/10.1080/0143116031000150022

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., and Bai, X., 2015. Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008

Wray, R.A.L., 2009. Palaeochannels of the Namoi River Floodplain, New South Wales, Australia: the use of multispectral Landsat imagery to highlight a Late Quaternary change in fluvial regime. *Australian Geographer*, 40 (1), 29–49. https://doi.org/10.1080/00049180802656952

Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International journal of remote sensing*, 27 (14), 3025–3033. https://doi.org/10.1080/01431160600589179

Zagorecki, A.T., Johnson, D.E.A., and Ristvej, J., 2013. Data mining and machine learning in the context of disaster and crisis management. *International Journal of Emergency Management*, 9 (4), 351–365. http://dx.doi.org/10.1504/IJEM.2013.059879

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., and Vergnaud, S., 2021. ESA WorldCover 10 m 2020 v100 (Version v100)[Data set]. Zenodo.

Highlights

- Multi-source Random Forest classification of floods exhibits accuracies above 90%
- Predictors stability to the algorithm architecture and study areas was assessed
- Models built with the most important predictors provides the best flood delineation
- The MNDWI is robust to training sample sizes, number of trees and study areas
- Morphologic descriptors are relevant predictors under updated morphological data

Ethical statement

The authors declare that all ethical practices have been followed in relation to the development, writing, and publication of this paper.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: