EasyChair Preprint

№ 8527

Mapping of Waterlogged Areas and Silt-Affected Areas After the Flood Using the Random Forest Classifier on the Sentinel-2 Dataset

Shivam Rawat, Rashmi Saini and Annapurna Singh

July 25, 2022

Mapping of Waterlogged Areas and Silt-Affected Areas After the flood Using the Random Forest Classifier on the Sentinel-2 Dataset

Shivam Rawat[1], Dr. Rashmi Saini[2],

Dr. Annapurna Singh[3]

[1] G.B.P.I.E.T, Pauri Garhwal, 246194, India
[2] G.B.P.I.E.T, Pauri Garhwal, 246194, India
[3]G.B.P.I.E.T, Pauri Garhwal, 246194, India

Abstract. Flood Mapping is an important activity that helps in understanding the spatial extent of the flood over the impacted region thereby helping emergency responders in chalking out plans for future emergencies. The main of this study is mapping waterlogged areas and silt-affected areas after the submergence of floods. In this study, Random Forest (RF) classifier is used for mapping water-logged areas and silt affected areas using a pixel-based supervised classification approach. For the classification process, six land use/cover classes covering a total area of 1491.84 km^2 of the Khagaria district of Bihar, India have been used. A four-band Sentinel-2 dataset at 10 m spatial resolution has been used for both pre-flood and post-flood datasets. The overall accuracy (OA) and Kappa score (K) for pre-flood classified data acquired using RF are (OA=84.95%, k=0.817). Whereas overall accuracy and kappa score for post-flood classified data using RF are (OA=83.325%, k=0.798) respectively. The results of post-flood classified data have shown that waterlogged areas and silt-affected areas have increased significantly from 22.40 km^2, 7.22 km^2 to 245.60 km^2, 81.53 km^2 respectively. Also, the classifier has shown fair Producer's and User's accuracy for the affected class that consists of Water-logged areas and Silt-affected areas. Furthermore, quantitative analysis of post-flood classified data shows there is a significant in-crease in waterlogged areas and silt-affected areas.

Keywords: Flood Mapping, Waterlogged, Remote Sensing, Sentinel-2, Sen2cor, Land Use Land Cover, Random Forest (RF).

1 Introduction

Every year floods cause huge economic losses wreaking havoc resulting in loss of life, infrastructure damage, and loss of livelihood of many people around the world. They may occur due to a variety of reasons such as erratic rainfall, cloudbursts, and climate change [1-2]. Bihar is one of the flood-prone states of India that experiences flood every year during the monsoon season. The monsoon season in Bihar lasts from the onset of June to the fall of November. It consists of Southwest Monsoon (June-September) and Post-monsoon (October-November). Two of the main problems that occur after the submergence of floods are the problem of waterlogging and silting. The problem of water-logging continues till the fall of December. More than 73.06% of Bihar's area is prone to flood [3].

In the last few years, the Remote Sensing domain has proven to be extremely useful for mapping the extent of floods by leveraging the potential of very high-quality satellite data and various advanced state of art modern machine learning algorithms.

The main of this study is the mapping of waterlogged areas and silt-affected areas after the submergence of floods using the pixel-based supervised machine learning approach on the Sentinel-2 dataset. This paper is structured as follows: Section 1 presents the basic introduction; Section 2 gives the literature review; Section 3 describes the Study area; Section 4 describes the dataset preparation and pre-processing; Section 5 describes the methodology and the classifiers used; Section 6 describes the results; As the conclusion, Section 7 summarises the study's main findings.

2 Literature Review

Flood mapping involves the use of various GIS techniques integrated with modern state of art machine learning algorithms to delineate the water body and flooded areas. The conventional technique for mapping flood extent involves in situ visit which is a time-consuming and costly process [4]. Some of the techniques employed by researchers for mapping flood are masking and thresholding [5], rule-based classification [6], and optimum thresholding based on spectral bands [7].

Satellite-based land use/cover maps have become very popular owing to the availability of high-quality [8]. Two of the very popular satellite data are LANDSAT and Sentinel-2 that provide a spatial resolution of 30m and 10m respectively.

A few of the most popular indices used for mapping flood and water areas are NDWI (Normalized Difference Water Index), MNDWI (modified NDWI), and AWEI (Automated Water Extraction Index) [9-11]. The first two indices have been widely used for mapping flood and water areas. However, these two methods cannot capture floods in urban areas [9-10]. Mobley et. al used Random Forest (RF) to generate a flood map using scikit library [12]. Landsat data owing to its large archive has been extensively used by researchers for mapping the extent of the flood. Wang et al. [13] used Landsat 7 TM images to determine the flooded areas by separating the water features from non-water features.

Gianinetto et al. used MNDWI by using LANDSAT data for mapping flood and determining the extent of flood damages [14]. Li et al. used Sentinel-2 data for flood mapping by using independent component analysis after separating water and land using modified NDWI (MNDWI) [15].
LULC maps can be produced by using different methods and classification algorithms [16-17]. The results and the performance of these classification algorithms are influenced by various factors such as training/testing samples used, the sensor's data employed, and the number of classes [18-19].

3 Study Area

The study area chosen is the Khagaria district located in Bihar, India. It is among one of the most affected districts in the flash floods in Bihar in 2020. The rectangular extent of the khagaria district is given by 25° 15' 14.1048"N to 25° 43' 54.4512"N latitude and 86°16' 44.076"E to 86° 51' 24.696"E longitude with a total spatial area of 1491.84 sq. km. The district is bordered on the west by the River Kosi. Fig. 1 describes the study area. Fig. 2 (a) display the pre-flood and (b) post-flood images respectively in the false-color composite format.

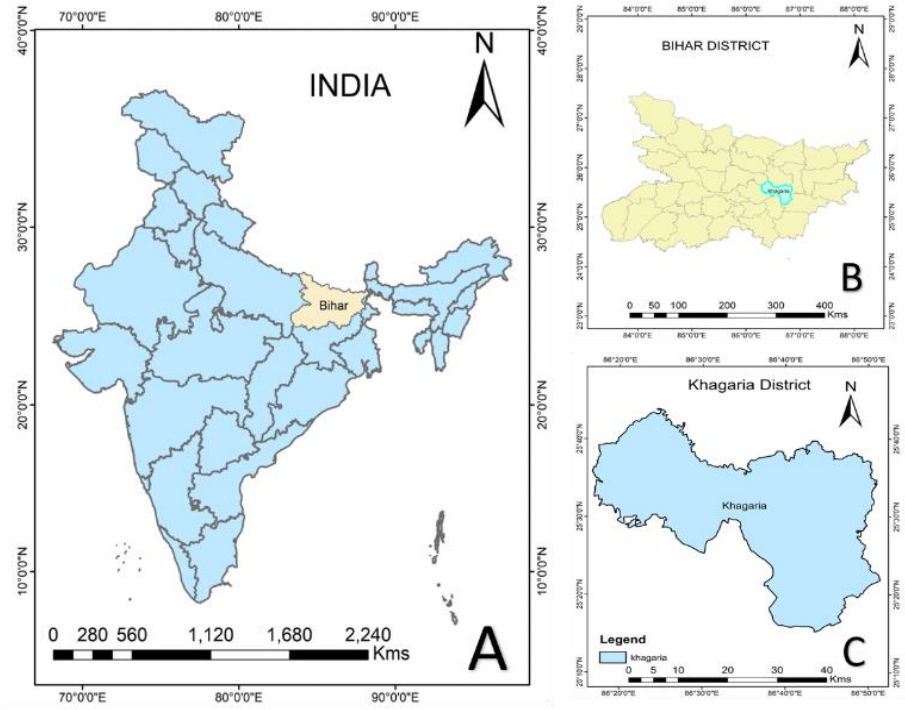

Fig. 1. (a) The Study Area located in India **(b)** Khagaria district located in the Bihar District **(c)** The study Area boundary.

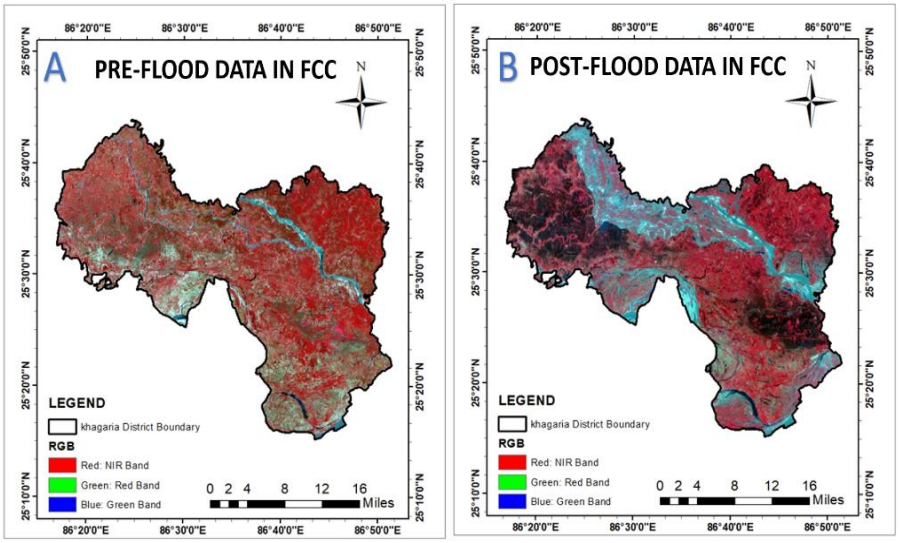

Fig. 2. Saharsa **(a)** Pre-flood and **(b)** Post-flood Data in False Color Composite.

4 Dataset Preparation and Pre-processing

4.1 Sentinel 2 Data

The European Space Agency (ESA) launched two satellites (Sentinel-2A/2B) in 2015 and 2017 respectively. Sentinel-2 data finds applications in many fields such as vegetation mapping, land cover mapping, geological remote sensing, water mapping, and many other applications [20-24]. The images taken are in WGS 1984 UTM projection, zone 45N.

TABLE I Band characteristics of Sentinel-2.

Band No.	Band Name	Resolution(m)	Wavelength (mm)
$(B)_1$	Coastal Aerosol Band	60	0.433–0.453
$(B)_2$	Blue Band	10	0.458–0.523
$(B)_3$	Green Band	10	0.543–0.578
$(B)_4$	Red Band	10	0.650–0.680
$(B)_5$	Vegetation $(Red\text{-}Edge)_{Band\text{-}1}$	20	0.698–0.713
$(B)_6$	Vegetation $(Red\text{-}Edge)_{Band\text{-}2}$	20	0.733–0.748
$(B)_7$	Vegetation $(Red\text{-}Edge)_{Band\text{-}3}$	20	0.773–0.793
$(B)_8$	Near Infra-Red (NIR) Band	10	0.785–0.900
$(B)_{8A}$	Narrow-Near Infrared Band	20	0.855–0.875
$(B)_9$	Water vapour Band	60	0.935–0.955
$(B)_{10}$	Short Wave Infrared Cirrus Band	60	1.360–1.390
$(B)_{11}$	$(Short\ Wave\ IR)_{Band\text{-}1}$	20	1.565–1.655
$(B)_{12}$	$(Short\ Wave\ IR)_{Band\text{-}2}$	20	2.100–2.280

4.2 Image Pre-Processing

Sentinel-2 data needs to be atmospherically corrected to remove the effect of scattering and absorption from the given dataset. The Sen2Cor processor is used to rectify data from the effects of atmospheric conditions. It is a Level 2A processor which converts images from Top-Of-Atmosphere (L1C) form to Bottom-Of-Atmosphere (L2A) form [25].

5 The Methodology and Classifiers Used

5.1 Methodology

In this study, the pixel-based supervised classification method is used for mapping of waterlogged areas and silt affected areas using pixel-based supervised classification technique. Fig. 3 describes the classification process methodology. The training data samples are generated using a polygon shapefile with the help of high-resolution Google Earth Imagery. Two sets of training data before the flood and one after the flood have been created for the raster dataset. After performing the classification, the area of each class before and after the flood is calculated for the quantitative assessment of post-flood impacts.

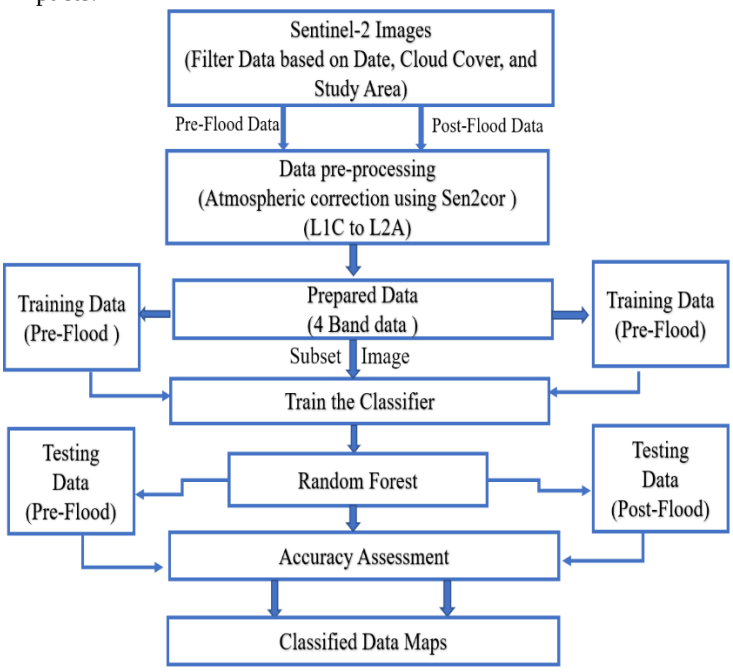

Fig. 3. Classification Process Methodology.

5.2 Classifiers Used

Random Forest

Random forest is a very popular and highly effective machine learning algorithm that uses bootstrapping techniques to build a group of decision trees [26-27]. In this technique, the different tree produces a subset of feature and training data with replacement. It combines the techniques of bootstrapping and random feature selection. Some of the samples in bootstrapping will be chosen frequently while some of the samples will not be picked at all. The unselected samples as such are used to evaluate the model performance. Each tree participates in the voting process and based on the votes from the trees, the most popular tree is selected as the output. It is ideally suited for classification and regression, however, it is most commonly used for classification-related tasks. There are two important parameters in the RF algorithm that need to be tuned which are mtry (number of independent variables sampled at each split) and ntree (number of trees to grow) [26].

6 Results and Discussion

The main aim of this study was to map the waterlogged areas and Silt-affected areas after the submergence of floods using the Sentinel-2 dataset. In this study, a four-band dataset consisting of the Red, Blue, Green, and Near-Infrared bands is used to classify the pre-crisis and post-crisis images. The training and testing samples are generated manually using the stratified random sampling technique with the help of high-resolution Google Earth Imagery. The number of testing samples taken for pre-flood and post-flood data is shown in Table II. The classification process is carried out in the R framework using the caret package using the Random Forest classifier. Here, the default values of all the tuning parameters of RF have been utilized for the pre-flood and post-flood classification. For the classification process, six land use/cover classes were identified after the careful analysis of the study area namely (1) Water-Body, (2) Vegetation, (3) Fallow Land, (4) Built-up, (5) Waterlogged Areas, and (6) Silt-Affected Areas covering a total area of 1491.84 km^2 of the Khagaria district, Bihar were determined.

For the Accuracy assessment of the classifier, a few accuracy measures are used (Overall Accuracy (OA), Kappa Score (k), Precision (P), and Recall (R)). For further details, Table IV and Table V give the Confusion matrix for the individual class for the pre-flood and post-flood classified data respectively. It can be observed from the confusion matrix (Table V) that there is misclassification between the waterlogged areas and the waterbody. Also, a few of the silt-affected areas have been misclassified as built-up areas. Table VI gives a quantitative analysis of pre-flood and post-flood classified data using the random forest classifier. The impacted class consisted of waterlogged areas and silt-affected areas. The quantitative analysis of post-flood classified data shows that there is a significant increase in waterlogged areas and silt-affected areas after the submergence of floods which can also be realized by the obtained maps as shown in Figure. The results of post-flood classified data have shown that waterlogged areas have increased significantly from 22.40 km^2 to 245.60 km^2. Whereas, the silt-affected areas have increased from 7.22 km^2 to 81.53 km^2. The classifier shows good producer accuracy and user accuracy for waterlogged areas and silt-affected areas.

The resulting recall and precision values achieved using RF for waterlogged areas are 81%, 100% for pre-flood classified data, and 67.5%, 92.2% for post-flood classified data respectively. Whereas the resulting recall and precision values achieved for silt-affected areas are 60.4%, 86.2% for pre-flood classified data, and 82.1%, 87.2% for post-flood classified data respectively. It can be visualized from the post-classified data maps that most of the impact is along the river channel along the northwestern part and southeastern part of the district.

TABLE II Number of Testing Samples for Pre-flood/Post-flood Data

Number of Samples	Water Body	Vege-tation	Fallow Land	Built-up Areas	Water-logged Areas	Silt-Affected Areas
Pre-flood	800	900	800	500	500	500
Post-flood	800	900	600	500	600	600

TABLE III Accuracy Metrics.

Classifiers	Pre-flood Data		Post-Flood Data	
	Overall Accuracy	Kappa Score	Overall Accuracy	Kappa Score
RF	84.95	0.817	83.325	0.798

Table IV Confusion Matrix for Pre-flood RF Classified Data.

Class	Water Body	Vege-tation	Fallow Land	Built-up Areas	Water Logged Areas	Silt Affected Areas	Classifica-tion Overall	User's Accuracy (Precision)
Water Body	784	0	0	0	95	0	879	0.891
Vegetation	0	815	0	0	0	0	815	1.0
Fallow Land	0	85	640	0	0	0	725	0.882
Built-up Areas	16	0	160	452	0	198	826	0.547
Water-Logged Areas	0	0	0	0	405	0	405	1.0
Silt Affected Ar-eas	0	0	0	48	0	302	350	0.862
Truth Overall	800	900	800	500	500	500	4000	
Producer's Accuracy(Recall)	0.98	0.905	0.8	0.904	0.81	0.604		

Table VI Confusion Matrix for Post-flood RF classified Data.

Class	Water Body	Vege-tation	Fallow Land	Built-up Areas	Water Logged Areas	Silt Affected Areas	Classifica-tion Overall	User's Accuracy (Precision)
Water Body	735	0	0	0	175	3	913	0.805
Vegetation	2	811	6	0	4	0	823	0.985
Fallow Land	0	79	453	36	16	0	584	0.775
Built-up Areas	0	2	134	436	0	104	676	0.644
Water-Logged Areas	26	8	0	0	405	0	439	0.922
Silt Affected Ar-eas	37	0	7	28	0	493	565	0.872
Truth Overall	800	900	600	500	600	600	4000	
Producer's Accuracy(Recall)	0.918	0.901	0.755	0.872	0.675	0.821		

Table VI Quantitative Analysis of pre-flood and post-flood classified Data

Class Description	Pre-Flood Data (Area in km^2)	Post-Flood Data (Area in km^2)
Water Body	40.18	108.62
Vegetation	708.11	575.31
Fallow Land	573.70	354.22
Built-up Areas	140.23	126.56
Water Logged Areas	22.40	245.60
Silt-Affected Areas	7.22	81.53

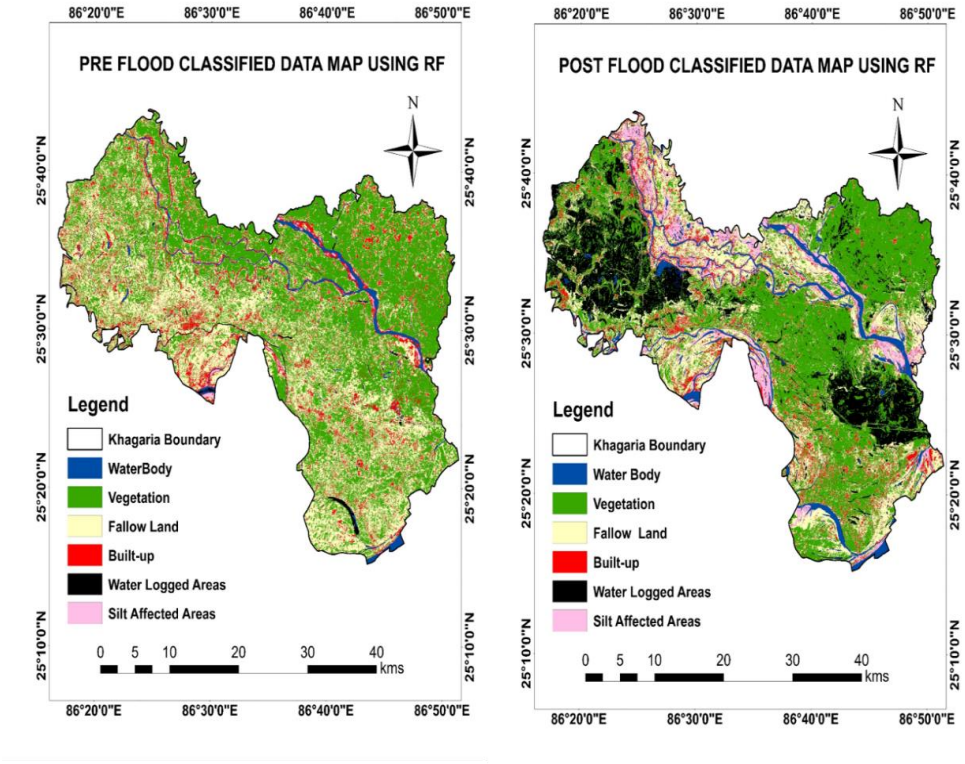

Fig. 4. RF pre-flood and post-flood Classified Data Maps.

7 Conclusion

The main aim of this study was to map the waterlogged areas and Silt-affected areas after the submergence of floods using the Sentinel-2 dataset. The overall accuracy (OA) and kappa score (k) for pre-flood classified data obtained using RF are (OA=84.95%, k=0.817). Whereas overall accuracy and kappa score for post-flood classified data using RF are (OA=83.325%, k=0.798). The results of post-flood classified data show that waterlogged areas and silt-affected areas have increased significantly from 22.40 km^2, 7.22 km^2 to 245.60 km^2, 81.53 km^2 respectively. Also, the classifier shows fair Producer's and User's accuracy for the affected class that consists of Water-logged areas and Silt-affected areas. Waterlogging and Siltation make areas inaccessible and unusable for agricultural activity, causing severe damage to the soil and the ecosystem. It can be visualized from post-classified data maps that most of the impact is along the river channel in the northwestern part and southeastern part of the district. As a future scope, one can use advanced machine learning techniques such as convolutional neural networks, deep learning to get better results.

References

1. N. Bezak, M. Mikos, "Investigation of Trends, Temporal Changes in Intensity-Duration-Frequency (IDF) Curves and Extreme Rainfall Events Clustering at Regional Scale Using 5 min Rainfall Data" in Water 11, pp. 2167, 2019.
2. G.J.P Schumann, D.K Moller, "Microwave remote sensing of flood inundation" in Physics and Chemistry of the Earth, vol. 83–84, pp. 84–95, 2015.
3. M. L. Kansal, K. A Kishore, and P. Kumar, "Impacts of floods and its management–A case study of Bihar" in International Journal of Advanced Research, vol. 5, pp. 1695-1706, 2017.
4. R. Brackenridge, E. Anderson, "MODIS-based flood detection, mapping, and measurement: The potential for operational hydrological applications" in NATO Science Series IV: Earth and Environmental Sciences, 2006.
5. K. Voormansik, J. Praks, O. Antropov, J. Jagomägi, and K. Zalite, "Flood mapping with TerraSAR-X in forested regions in Estonia" in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 562–577, 2013.
6. M. Carroll, J. R. Townshend, C. M. DiMiceli, P. Noojipady, R. Sohlberg, "A new global raster water mask at 250 m resolution" in Int. J. Digit.Earth, vol.2, no.4, pp. 291–308, 2009.
7. Z. Wang, J. Liu, J. Li, and D. Zhang, "Multi-Spectral Water Index (MuWI): A Native 10-m Multi-Spectral Water Index for Accurate Water Mapping on Sentinel-2" in Remote Sensing, vol. 10, no. 10, pp. 1643, 2018.
8. M. Tanguy, K. Chokmani, M. Bernier, J. Poulin, S. Raymond, "River flood mapping in urban areas combining Radarsat-2 data and flood return period data" in Remote Sens. Environ, vol. 198, pp. 442–459, 2017.
9. S. K. McFeeters, "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features" in International journal of remote sensing 17, 7 (1996), pp. 1425–1432.
10. H. Xu., "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery" in International journal of remote sensing 27, 14 (2006), pp. 3025–3033.

11. G.L. Feyisa, H. Meilby, R. Fensholt, S.R. Proud, “Automated water extraction index: A new technique for surface water mapping using Landsat imagery” in Remote Sens. Environ, vol. 140, pp. 23–35, 2014.
12. W. Mobley, A. Sebastian, R. Blessing, W. E. Highfield, L.Stearns, and S. D. Brody “Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: a pilot study in southeast Texas” in Nat. Hazards Earth Syst. Sci., 21, 807–822, 2021.
13. Y. Wang, J.D. Colby, and K.A. Mulcahy “An efficient method for mapping flood extent in a coastal floodplain using Landsat TM and DEM data” in Int. J. Remote Sens, 23 (18), pp. 3681–3696, 2002.
14. Villa, Paolo & Gianinetto, Marco, “Monsoon Flooding Response: a Multi-scale Approach to Water-extent Change Detection”, 2006.
15. R. Hostache, P. Matgen, G. Schumann, C. Puech, L. Hoffmann, and L. Pfister, “Water Level Estimation and Reduction of Hydraulic Model Calibration Uncertainties Using Satellite SAR Images of Floods,” in IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 2, pp. 431-441, 2009.
16. B. Waske and M. Braun, “Classifier ensembles for land cover mapping using multitemporal SAR imagery,” in ISPRS Journal of Photogrammetry and Remote Sensing, vol. 64, pp. 450–457, 2009.
17. C. Li, J. Wang, L. Wang, L. Hu, and P. Gong, “Comparison of classification algorithms and training sample sizes in urban land classification with Landsat Thematic Mapper imagery,” in Remote Sensing, vol. 6, pp. 964–983, 2014.
18. S.S Heydari and G. Mountrakis, “Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per- pixel classification accuracy using 26 Landsat sites,” in Remote Sensing of Environment, vol. 204, pp. 648–658, 2018.
19. R. Hamad, “An assessment of artificial neural networks, support vector machines and decision trees for land cover classification using sentinel- 2A data,” in Applied Ecology and Environmental Sciences, vol. 8, pp. 459–464, 2020.
20. R. Saini and S. K. Ghosh, “Exploring Capabilities of Sentinel-2 for Vegetation Mapping Using Random Forest,” in International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 42, pp. 1499-1502, 2018.
21. R. Saini and S. K. Ghosh, “Crop classification in a heterogeneous agricultural environment using ensemble classifiers and single-date Sentinel-2A imagery,” in Geocarto International, 2019.
22. X. Luo, X. Tong, and H. Pan, “Integrating Multiresolution and Multitemporal Sentinel-2 Imagery for Land-Cover Mapping in the Xiongan New Area, China,” in IEEE Transactions on Geoscience and Remote Sensing, vol. 59, pp. 1029-1040, 2020.
23. H. Werff and F. Meer, “Sentinel-2A MSI and Landsat 8 OLI provide data continuity for geological remote sensing,” in Remote Sensing, vol. 8, 2016.
24. Y. Du, Y. Zhang, F. Ling, Q. Wang, W. Li, and X. Li, “Water Bodies’ Mapping from Sentinel-2 Imagery with Modified Normalized Difference Water Index at 10-m Spatial Resolution Produced by Sharpening the SWIR Band,” in Remote Sensing, vol. 8, no. 4, p. 354, 2016.
25. M. Main-Knorn, B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon, “Sen2Cor for Sentinel-2,” in Conference-proceedings-of-spie, Image and Signal Processing for Remote Sensing, vol. 10427, 2017.
26. M. Belgiu, and L. Drăguţ, “Random Forest in Remote Sensing: A Review of Applications and Future Directions. ISPRS Journal of Photogrammetry and Remote Sensing”, 114, pp. 24-31, 2016.

27. L. Breiman, “Random forests. Machine Learning”, 45(1), pp. 5-32, 2001.