CLOUD BASED GEO-PROCESSING PLATFORM FOR ANALYZING LARGE VOLUME TEMPORAL SATELLITE DATA: A STUDY IN PART OF GHAGHARA RIVER BASIN (INDIA) FOR SURFACE WATER SPREAD ANALYSIS

Y. Walia[1a], P. K. Gupta[2b], S. K. Srivastav[2c], A. Gulzat[3d], S. K. Saha[1e]
[1]University of Petroleum and Energy Studies, Dehradun, India
Email: [a]yash39wal@gmail.com, [e]sksaha@ddn.upes.ac.in
[2]Indian Institute of Remote sensing, Indian Space Research Organisation, Department of Space, Government of India, Dehradun, India
Email: {[b]prasun, [c]sksrivastav}@iirs.gov.in
[3]Centre for Space Science and Technology Education in Asia and the Pacific, IIRS Campus, Dehradun, India
Email: [d]gulzat.kgg@mail.ru

KEYWORDS: Cloud platform, Earth Engine, Geo-processing, Surface water, Landsat, MODIS

ABSTRACT: With the availability of large spatio-temporal Earth Observation satellite data and geospatial layers with concomitant increase in geo-computation requirement, the popularity of cloud based platform for accessing, sharing and processing large volume data is growing day by day among the geospatial community. Google's Earth Engine is one such platform, which provides access to satellite and other geospatial datasets and analysis functionalities to researchers free of cost, for various applications from local to global-scale. This study aims at exploring Google's Earth Engine to study the spatio-temporal dynamics of surface water during the Indian Summer Monsoon season over the last three decades. Top of Atmosphere (TOA) reflectance images available from Landsat-5, 7 and 8 (1984 to 2016.) and Aqua MODIS (2002 to 2016) are used in this process. A part of Ghaghara river basin covering about 8400 km^2 area is selected as the study site. The clouds and their shadows are masked using "Fmask" and "StateQA" quality bands available along with TOA reflectance images for Landsat and MODIS, respectively. Normalized Difference Water Index (NDWI) and Normalized Difference Vegetation Index (NDVI) are calculated and by applying suitable thresholds and combining them, binary surface water images for each date of acquisition are generated. By combining the binary surface water images over the entire study period, a composite long-term surface water map is generated wherein each pixel represents the frequency of the presence of surface water for cloud-free observations. The output is compared with the geomorphology map. Such composite long-term surface water map is extremely useful not only for understanding the spatio-temporal dynamics of surface water but also for other applications e.g., flood hazard zonation and developmental planning, etc. The study demonstrates the potential of cloud based online geo-processing platform in analyzing the time-series data from multiple satellites without any cost implication towards data procurement and processing; the scalability for large-area application may be explored further.

1. INTRODUCTION

Surface water bodies are fundamental water stockpiling units and play an essential role in effective tapping of large amounts of water from precipitation and overflow occasions. Surface water spread being dynamic natural process, varies from months, seasons to years. To understand it fully, there is a need to look for large datasets with long temporal coverage. With the advent of satellite imaging, long-term storage of data in digital form and availability in public domain, many strategies have emerged to delineate surface water spread from satellite pictures utilizing both optical and microwave sensors.

Several papers have reported utilization of multi-transient satellite pictures for portraying the water spread territory. Lenher *et al.*, 2004 have made National Global lakes and wetlands database at various levels viz., substantial lakes, supplies, littler water bodies, and wetlands. Li *et al.*, 2005 have published the maps of Canada's wetlands utilizing optical, radar and DEM information. Subramaniam et al., 2011 have created an automated algorithm to extract water bodies, for the depiction of surface water bodies utilizing Indian Remote Sensing (IRS) satellite Resourcesat's AWiFS and LISS III sensors, and executed on national datasets.

Google Earth Engine (GEE) is a cloud-based geo-processing platform, where large datasets of satellite images can be processed at once and countless analysis can be made. A global effort on analyzing Landsat data has been made by Pekel *et al.*, 2016 and the results are available on their portal (EU-JRC, 2017). These outputs can be utilized for disaster management, water resource management, biodiversity characterization and various other thematic applications.

In this study, temporal data from Landsat and MODIS satellites have been used on GEE to study the spatio-temporal dynamics of surface water spread during the Indian Summer Monsoon season over last three decades, from 1984 to 2016. The results are further analyzed with reference to geomorphology of the study area.

2. MATERIALS AND METHODS

2.1 Study area

A part of Ghaghara river basin in the Uttar Pradesh State of India, covering about 8400 km^2 area, is taken up for this study (Fig. 1). The area lies between 81.24^{o}E to 82.35^{o}E and 26.75^{o}N to 27.48^{o}N. With an average population density of about 800 persons per km^2, Uttar Pradesh is the State with largest population. As the area is fertile, there is a pressure on land availability. Several important government and private installations such as, offices, hospitals, schools, colleges and housing units have cropped up in the flood plains. During the Indian summer monsoon, the multiple rivers in the area tend to overflow its banks and change their course. Due to meandering of rivers and multiple river confluences, there is an urgent requirement to do a flood hazard zonation of this area.

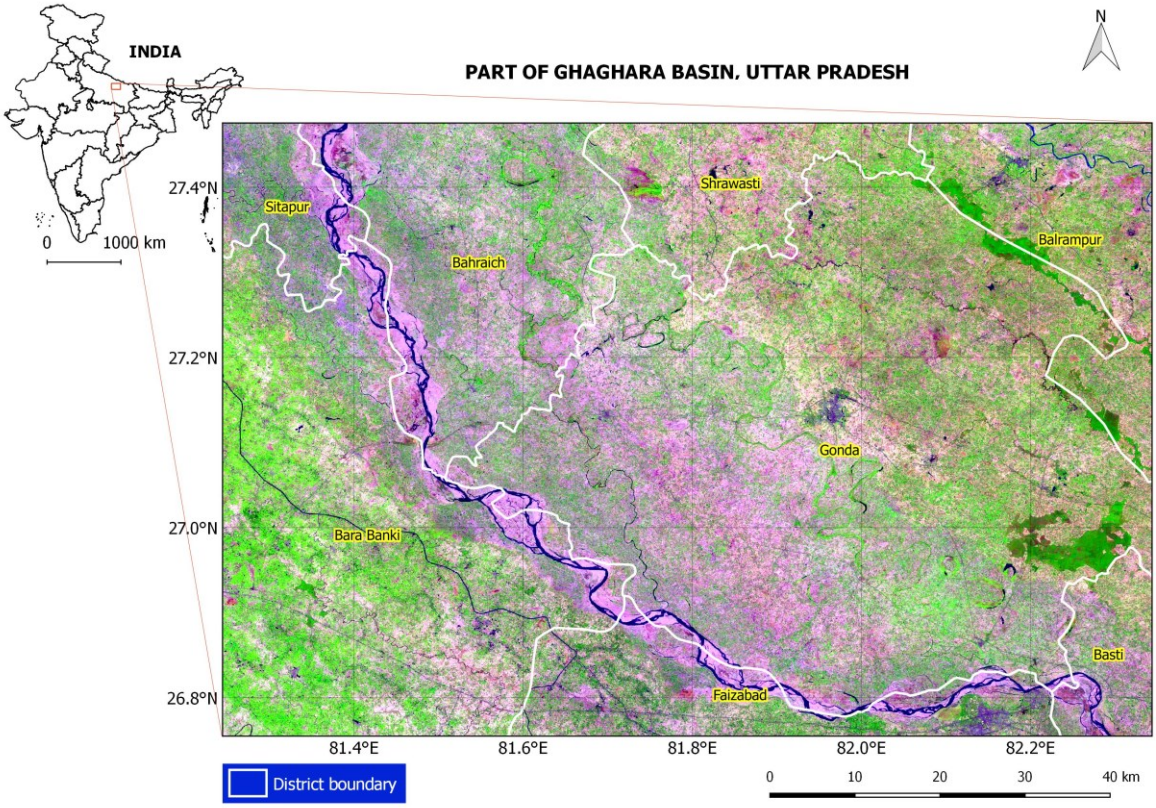

Figure 1: Study area is part of the Ghaghara river basin in the State of Uttar Pradesh, India. The selected area lies in six districts. Shown here is the Landsat 8 enhanced natural color composite, acquired on June 4, 2017.

2.2 Landsat data

Landsat has longest archive of continuous moderate resolution remote sensing data of land. Due to availability of such a vast data, researchers from many fields like geology, agriculture, water resources, soils, forestry, disaster management, etc. are using these data for various applications. With time newer applications using these datasets are being implemented for the betterment of society.

For this study, satellite imagery of Landsat 5, 7 and 8 are utilized for the time duration of 33 years from 1984 to 2016 (Table 1). Only four months (June to September) data are taken up corresponding to the Indian summer monsoon season in which 90% of the annual cumulative rainfall takes place. “Top of Atmosphere (TOA) Reflectance” images of only three bands i.e., green (0.52~0.60), red (0.63~0.69) and near infrared (0.76~0.90), are utilized as a part of the present study.

Optical remote sensing data have limitations during monsoon months due to presence of clouds. To counter this problem, the “Fmask” band developed by Zhu *et al.*, 2012 and Zhu *et al.*, 2015 along with Landsat TOA reflectance

dataset is used. This "Fmask" band can identify pixels which correspond to clear sky (0), water (1), shadow (2), snow (3) and cloud (4).

Table 1: Details of Landsat data used

Dataset	Spatial Resolution	Temporal Granularity	Temporal Coverage	Spatial Coverage
Landsat 8 OLI/TIRS	30m	16 day	2013-now	Global
Landsat 7 ETM +	30m	16 day	2000-Now	Global
Landsat 5 TM	30m	16 day	1984-2012	Global

2.3 MODIS data

MODIS (Moderate Resolution Imaging Spectroradiometer) is one of the instruments on-boad Terra (EOS AM-1) and Aqua (EOS PM-1) satellites. Terra and Aqua MODIS together cover the whole Earth within 1 to 2 days and acquire data in 36 different spectral bands. Data received from MODIS are used for understanding numerous processes of land, ocean and atmosphere, environment monitoring and studying global change.

As MODIS data are present for a relatively shorter duration of time, the temporal range used in the study is from 2002 to 2016. The surface reflectance data from Aqua satellite is used which has a temporal resolution of 8 days, spatial resolution of 500 m and is the level-3 processed product (namely, MYD09A1). The MODIS bands (sur_refl_b01, sur_refl_b02 and sur_refl_b04) used in this study are similar to that of Landsat bands. The Fmask band available in Landsat is replaced by the "StateQA" band in MODIS data, which contains the information about the cloud presence.

2.4 Geomorphology map

The geomorphology map prepared on 1:50,000 scale under the National Geomorphological and Lineament Mapping (NGLM) project, jointly carried out by Indian Space Research Organisation (ISRO) and Geological Survey of India, is also used in this study (Fig. 2) apart from satellite data. The geomorphology map is prepared using IRS LISS-III images of 2005-06 and other ancillary/ field information. The area is covered under three major geomorphic units i.e., *active flood plain*, *older flood plain* and *older alluvial plain*.

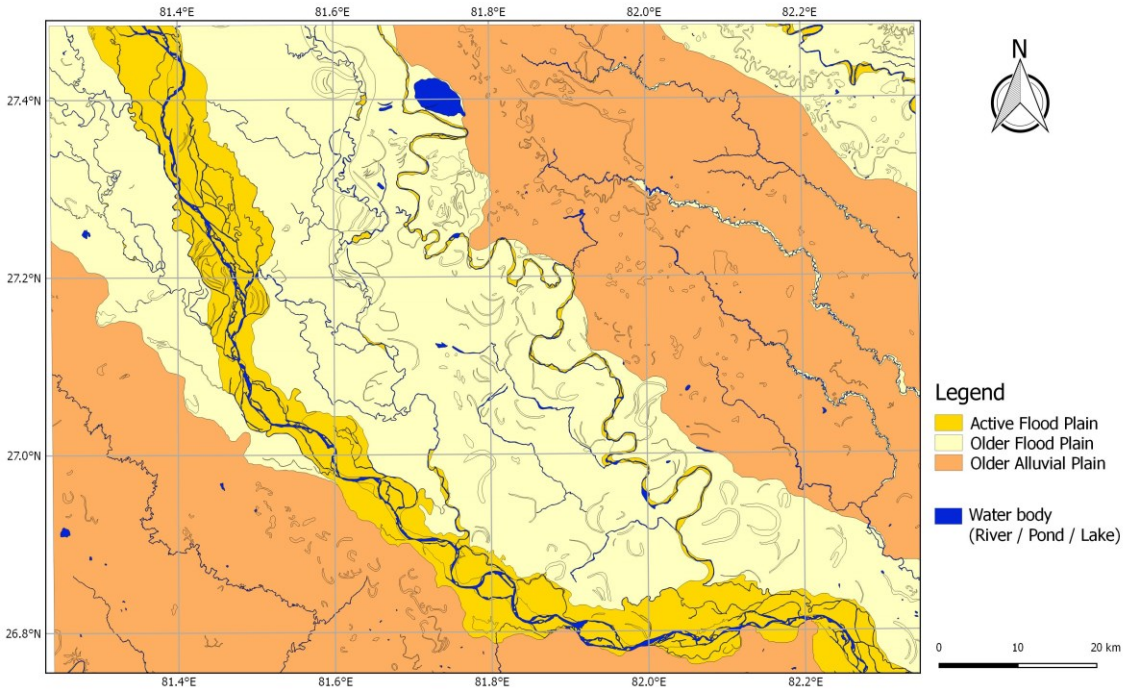

Figure 2: Geomorphogy map of the study area (Source: National Geomorphological and Lineament Mapping, NGLM, project of ISRO and GSI)

2.5 Google Earth Engine (GEE)

Google Earth Engine (GEE) is a cloud-based platform for Earth Observation data analysis that combines vast collection of publically available aerial and satellite data with a large-scale computational facility optimized for parallel processing of geospatial data. GEE has an archive of historical Earth imagery (going back more than forty years), temporal data series of continuing satellite series (such as Landsat, MODIS, etc.) and newer satellite data collections (such as Sentinel, VIIRS, etc.). GEE supports geospatial analyses such as, image processing, classification, change detection and vector-based extraction of image statistics. To enable the analysis of large datasets, GEE also provides APIs in JavaScript and Python programming languages, as well as other tools. It provides a library of built-in functions which may be applied to this geospatial data collection for display and analysis. Through the API, users can create their own algorithms by recombining existing algorithms.

In this study, GEE is used to process time-series satellite data (Landsat and MODIS) for the last three decades (1984 to 2016).

2.6 Methods and process

This section discusses various methods employed in the collection of large amount of data and processing them in GEE and desktop GIS software (Fig. 3).

2.6.1 Data pre-processing

The GEE platform has over 170 different collections ranging from aerial, satellite to modelled data products for different time periods and the associated metadata. It allows spatial, temporal and metadata based filtering of these collections for further use in analysis. The Landsat and MODIS data required for the study are filtered for the area around Ghaghara basin, for the monsoon period of 1984-2016 (for Landsat) and 2002-2016 (for MODIS). This is done using the built-in geometry based filtering in the spatial domain "filterBounds," and the metadata based filtering function in the time domain "filterDate." Landsat data for 33 years and MODIS data for 15 years are assembled in the form of an "ImageCollection" to perform time-series analysis.

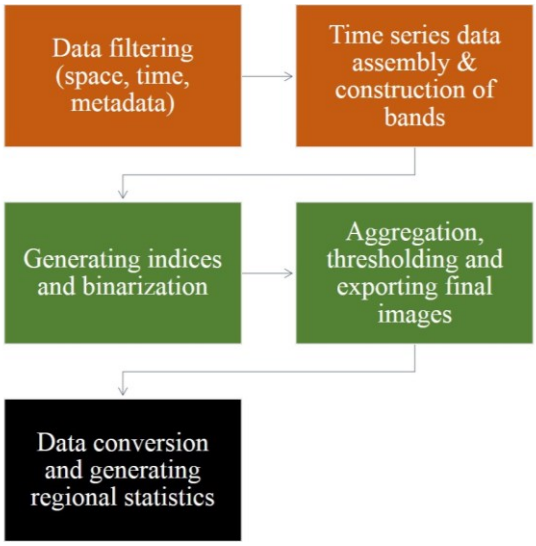

Figure 3: Generalized structure of methodology followed. The orange boxes represent data pre-processing steps performed in GEE, the green boxes represent the powerful "map" and "reduceRegion" functions in GEE which allows pixel-wise parallel time-series analysis to be performed, and the black box represents operations performed on desktop GIS software

2.6.2 Time-series analysis

Landsat bands B2, B3, B4, Fmask and MODIS bands sur_refl_b01, sur_refl_b02, sur_refl_b04, StateQA are used to create Normalized Difference Water Index (NDWI). McFeeters (1996, 2013) equations are used to derive NDWI. A pre-defined function is available in GEE to find the normalized difference, "normalizedDifference (['Green band', 'NIR band'])" to make this calculation. The power of the cloud platform is utilized here by making

the calculations in parallel for each pixel, thereby reducing the time required to perform the task. Landsat and MODIS data are also used to create Normalized Difference Vegetation Index (NDVI) images using "normalizedDifference (['NIR band', 'Red band'])" function.

To generate non-vegetated surface water binary images, thresholding of the two indices (NDWI and NDVI) is performed. Cloud information is also taken into consideration in this step. The Fmask (for Landsat) and StateQA (for MODIS) bands are used to ignore cloudy pixels. All the values above zero are changed to 1 and below zero to 0 in case of the NDWI; while in case of NDVI all the values less than zero are changed to 1 and values above zero are changed to 0. This binarization helps in capturing non-vegetated surface water. This operation is performed in parallel on Google's massive array of servers, using the "map" function to loop overall images in an image collection.

After applying the threshold to all the images in both the collections, each collection is individually summed. These are termed as "surface water aggregated image." The Fmask and StateQA datasets are separately binarized and summed to generate "cloud free aggregated image," The ratio of the surface water aggregated image and the cloud free aggregated image is calculated and termed as "normalized frequency of water occurrence image." These products are exported from GEE to local computer.

2.6.3 Analysis of surface water maps

The surface water aggregated image maps and normalized frequency of water occurrence image maps obtained from Landsat and MODIS are compared with the geomorphology map. Zonal statistics of surface water occurrence with reference to different frequency classes geomorphic units are extracted and analyzed.

3. RESULTS AND DISCUSSION

3.1 Surface water aggregated images

The surface water aggregated images show the number of days in which water is detected on each pixel i.e., a pixel value of 10 means that on 10 images water is detected on cloud-free days. The number of images from which this number is derived varies from pixel to pixel, as some areas may not have been imaged or may not be available in the GEE Landsat/MODIS archive. Figure 4a represents areas with presence of water in shades of blue as derived from Landsat data; white background means that water is not present on any of the days of satellite data used in the study. Nearly 90% of the pixels have 0 values (Table 2), hence the majority portion of the map remains free from occurrence of water and appears white (Fig. 4a). Nearly 10% of the remaining pixels show areas where the water is present in <60 days. The MODIS data show that about 91% of the area remains free from the occurrence of water, while remaining area has water on different number of days (Table 2, Fig. 4b). The difference in the maximum aggregated values on Landsat and MODIS data may be explained due to difference in the length of studied time period and spatial resolution. As MODIS has a coarser resolution, sub-pixel occurrence of water may lead to a given pixel getting classified as non-water or water pixel depending on the proportion of water in that pixel. Importantly, despite differences in the spatial resolutions in Landsat and MODIS, the overall pattern of surface water occurrence in Landsat and MODIS data is similar. However, owing to better spatial resolution, the Landsat data derived surface water map is smooth and also depict finer details.

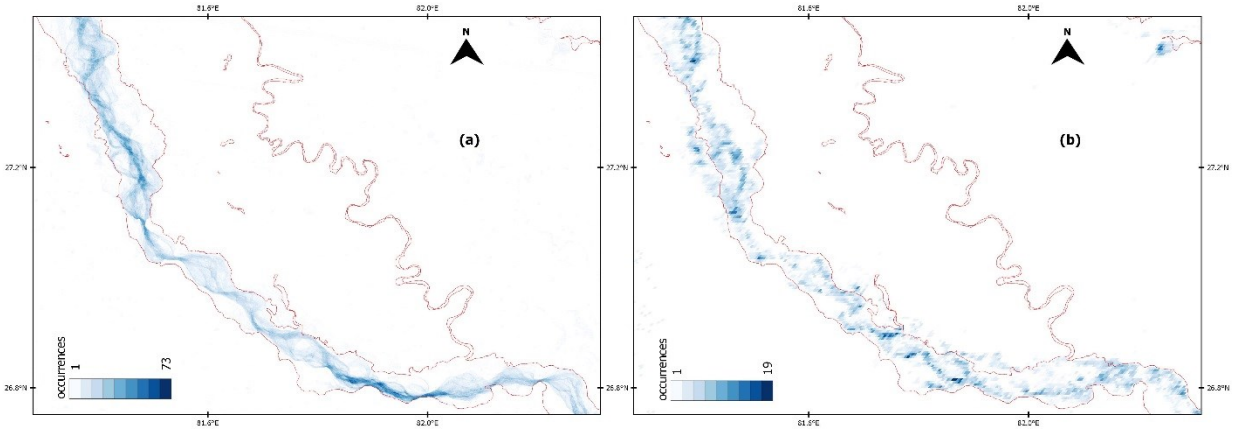

Figure 4: Surface water detection (in number of days) from (a) Landsat data during monsoon period of 1984 - 2016, and (b) MODIS data during monsoon period of 2002 – 2016. Active flood plain boundary taken from geomorphology map is shown as red lines

Table 2: Area under water on different days as observed from (a) Landsat and (b) MODIS data

(a) Landsat		(b) MODIS	
Number of days	Area (in % of total area)	Number of days	Area (in % of total area)
0	90.11	0	91.23
1-20	8.48	1-5	7.23
21-40	1.24	6-10	1.34
40-60	0.18	11-15	0.18
61-73	0.00	16-19	0.03

3.2 Cloud-free aggregated image

To ensure no distortion in the calculation of indices from the satellite data due to presence of clouds, the extra bands having the information about the clouds (Fmask and StateQA) are used. All available datasets in the study area are aggregated to get total number of cloud-free observations (NCO) for each pixel. In the said period, Landsat data have pixels with 28 to 187 days of cloud-free observations (out of total 297 days of observations) and MODIS data pixels have 11 to 80 days (out of total 214 days). Figure 5 shows the percentage of area versus availability of cloud-free observations for Landsat and MODIS data. The Landsat histogram shows a bi-modal distribution, whereas the MODIS histogram shows a uni-modal distribution.

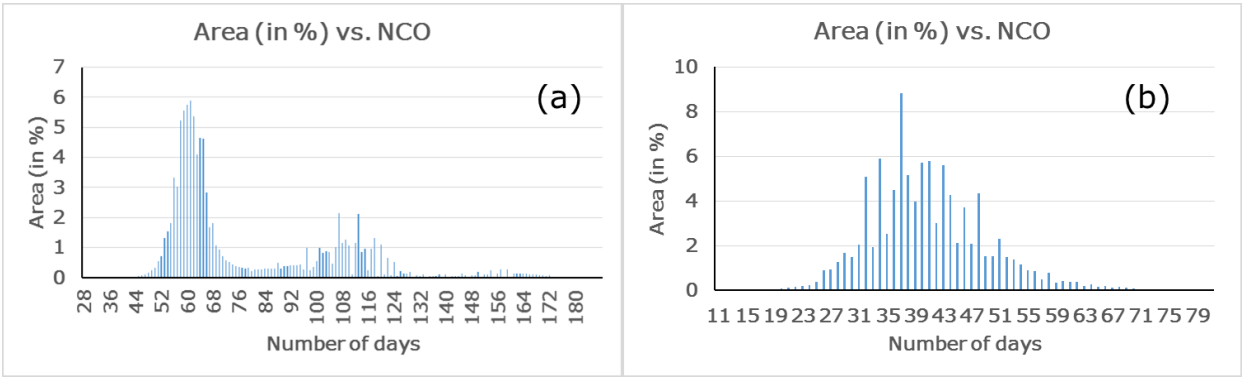

Figure 5: Histogram of NCO's for (a) Landsat and (b) MODIS

3.3 Normalized frequency of water occurrence

After finding image of number of water days and number of cloud-free days for each pixel, both the images are divided to find the water pixels for the clear sky. The values in the output image i.e., in the normalized frequency of water occurrence image, lies between 0 and 1 with most of the values near zero (Fig. 6a for Landsat data and Fig. 6b for MODIS data). The overall pattern of surface water occurrence in Landsat and MODIS data is similar. Further, the overall pattern of water occurrence as observed in the surface water aggregated image (Fig. 4) and the normalized frequency of water occurrence image (Fig. 6) is similar. As mentioned in section 3.1, about 90% of the study area can be considered as high-ground as there are no incidences of water occurrence during monsoon of 1984-2016 (Table 3). The areas which require most attention are the ones which show high frequency of water occurrence. For example, about 0.01% of the total area (approx. 84 km^2) is covered with water for 70-80% of the time as observed from Landsat data. In case of MODIS data, about 0.02% of the total area (approx. 168 km^2) has water for 28-32% of the time. The difference in the values of normalized frequency of occurrence of surface water in Landsat and MODIS appears to be mainly due to differences in the number of cloud-free observations.

The normalized frequency of water occurrence obtained by combining the results of Landsat and MODIS data is shown in Figure 6c (zoomed-in north-western portion shown in Figure 6d) and water occurrence statistics is shown in Table 3. The combined results show that 87.56% of the area remains free from occurrence of surface water and the remaining area contains water for different periods of time. As expected, the values of normalized frequency of water occurrence in the combined dataset have reduced as compared to that obtained from Landsat data due to incorporation of MODIS data.

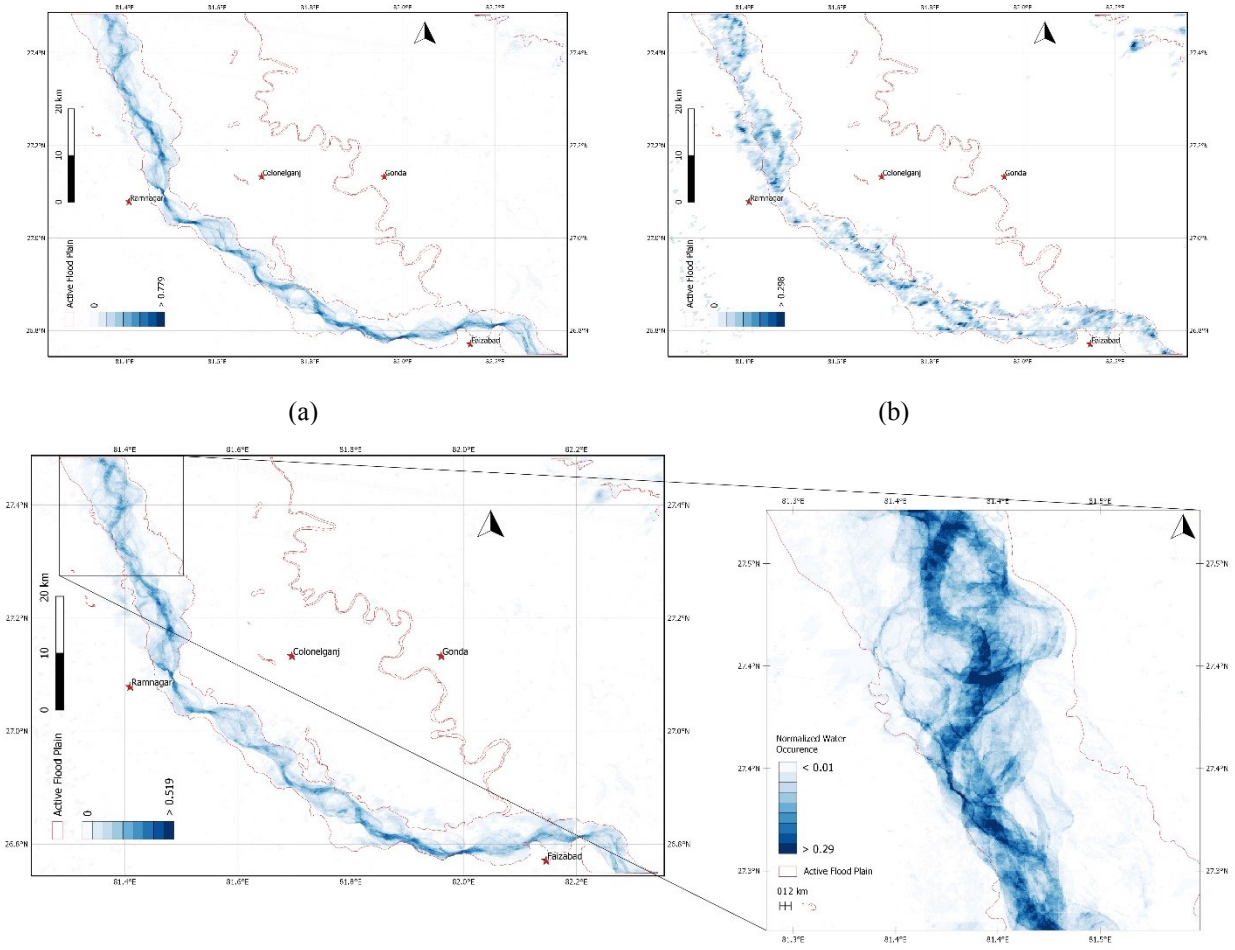

Figure 6: Normalized frequency of water occurrence from (a) Landsat data (monsoon period of 1984-2016), (b) MODIS data (monsoon period of 2002-2016), (c) Landsat and MODIS merged data (at 60m resolution with nearest neighbor resampling), and (d) zoomed-in north-west portion of the study area from Landsat and MODIS merged data. Active flood plain boundary taken from geomorphological map is shown in red

Table 3: Percentage of total area under water as a function of normalized frequency of occurrence of surface water from Landsat data, MODIS data, and combined Landsat and MODIS data (at 60m resolution with nearest neighbor resampling)

Landsat		MODIS		Combined Landsat & MODIS	
Normalized Freq.	Area (in % of total area)	Normalized Freq.	Area (in % of total area)	Normalized Freq.	Area (in % of total area)
0	90.11	0	91.23	0	87.56
>0 - 0.1	5.49	>0 - 0.04	3.56	>0 - 0.01	2.49
0.1 - 0.2	2.01	0.04 - 0.08	2.77	0.01 - 0.07	5.02
0.2 - 0.3	1.12	0.08 - 0.12	1.39	0.07 - 0.13	2.14
0.3 - 0.4	0.65	0.12 - 0.16	0.65	0.13 - 0.20	1.49
0.4 - 0.5	0.36	0.16 - 0.20	0.27	0.20 - 0.26	0.65
0.5 - 0.6	0.18	0.20 - 0.24	0.07	0.26 - 0.39	0.58
0.6 - 0.7	0.06	0.24 - 0.28	0.04	0.39 - 0.46	0.05
0.7 - 0.8	0.01	0.28 - 0.32	0.02	0.46 - 0.51	0.00
0.8 - 0.816	0.00	0.32 - 0.332	0.00	0.51 - 0.52	0.00

3.4 Relation between surface water spread and geomorphology

Zonal statistics for different classes of normalized frequency of surface water occurrence are generated with reference to three major geomorphic units i.e., active flood plain, older flood plain and older alluvial plain. Similar

results are obtained for Landsat, MODIS and combined datasets, hence only Landsat results are presented for brevity (Table 4). It is seen from this table that 98.6% area of the older alluvial plain and 95.5% area of the older flood plain remain free from occurrence of surface water, while 78.8% area of active flood plain remains covered with water for different periods of time. It is also clearly visible from Figure 4 and Figure 6 that water spread is mainly confining to the zone of active flood plain.

Table 4: Geomorphological unit-wise area under water as a function of normalized frequency of occurrence of surface water as derived from Landsat data

Normalized Frequency of Occurrence of Surface Water	Area (in % of total area in a geomorphic unit)		
	Older Alluvial Plain (%)	Older Flood Plain (%)	Active Flood Plain & River (%)
0.0	98.58	95.51	21.24
>0.0-0.1	1.40	3.45	24.22
0.1-0.2	0.02	0.76	13.35
0.2-0.3	0.01	0.20	7.14
0.3-0.4	0.00	0.06	3.76
0.4-0.5	0.00	0.01	19.50
0.5-0.6	0.00	0.01	8.05
0.6-0.7	0.00	0.00	2.32
0.7-0.8	0.00	0.00	0.40
0.8-0.9	0.00	0.00	0.01
0.9-1.0	0.00	0.00	0.00

Thus, a strong correspondence is observed between the surface water occurrence mapped from time-series satellite data and geomorphology map. At some of the places, water occurrence is also seen outside (but adjacent) the active flood plain boundary. This may also be due to error in the active plain boundary. The study highlights the potential of GEE in studying time-variant surface features, such as surface water in this case, without any constraint on historical satellite data availability and processing infrastructure. A small area occupying about 8400 km^2 is taken up in this case, therefore, the scalability aspect of the GEE for larger area application can be tested in further studies.

4. CONCLUSIONS

In this paper, an approach to monitor surface water spread using GEE is presented. The study uses about three decades (1984 to 2016) of satellite images in a part of Ghaghara river basin in the State of Uttar Pradesh of India. Over 500 satellite images pertaining to the monsoon season from Landsat and MODIS series are used. High correspondence is observed between the active flood plain and the frequency of surface water occurrence delineated from time-series satellite images. Such composite long-term surface water spread map can be used for flood hazard zonation and developmental planning. It is concluded that the studies which are data and computation intensive can be done quickly and with ease on cloud based geo-processing platforms, such as GEE, without any cost implication towards data procurement and processing. The scalability of GEE in terms of large area application can be further explored.

ACKNOWLEDGEMENTS

The first author (YW) is thankful to Indian Institute of Remote Sensing (IIRS) for permitting to do internship. Provision for cloud computing resources and interactive interface of Google Earth Engine is duly acknowledged. The second and third authors (PKG and SKS) are grateful to the Director, IIRS, Dehradun for his constant encouragement and support. The fourth author (AG) thanks the Centre for Space Science and Technology in Asia and the Pacific (CCSTEAP), Dehradun for providing opportunity to undergo the Postgraduate Diploma course in Remote Sensing and GIS.

REFERENCES

EU-JRC, 2017. Global Surface Water Explorer. Retrieved July 14, 2017, from https://global-surface-water.appspot.com/.

Lehner, B., and Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology, 296(1), pp. 1-22.

Li, J., and Chen, W., 2005. A rule-based method for mapping Canada's wetlands using optical, radar and DEM data. International Journal of Remote Sensing, 26(22), pp. 5051-5069.

Li, W., Du, Z., Ling, F., Zhou, D., Wang, H., Gui, Y., and Zhang, X., 2013. A comparison of land surface water mapping using the normalized difference water index from TM, ETM+ and ALI. Remote Sensing, 5(11), pp. 5530-5549.

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), pp. 1425-1432.

McFeeters, S. K., 2013. Using the normalized difference water index (NDWI) within a geographic information system to detect swimming pools for mosquito abatement: A practical approach. Remote Sensing, 5(7), pp. 3544-3561.

Pekel, J. F., Cottam, A., Gorelick, N., and Belward, A. S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature, 540(7633), pp. 418-422.

Subramaniam, S. and Suresh Babu, A. V., 2014. Satellite derived Information on Water Bodies Area (WBA) and Water Bodies fraction (WBF). Technical Report. Available on http://bhuvan.nrsc.gov.in/data/download/tools/document/waterbodies_fraction.pdf. Accessed on 2017-09-07.

Tang, Z., Ou, W., Dai, Y., and Xin, Y., 2012. Extraction of water body based on LandSat TM5 imagery-a case study in the Yangtze River. In International Conference on Computer and Computing Technologies in Agriculture (pp. 416-420). Springer, Berlin, Heidelberg.

Zhu, Z., and Woodcock, C. E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sensing of Environment, 118, pp. 83-94.

Zhu, Z., Wang, S., and Woodcock, C. E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and Sentinel 2 images. Remote Sensing of Environment, 159, pp. 269-277.

--- x ---