



# Mapping active paddy rice area over monsoon asia using time-series Sentinel – 2 images in Google earth engine; a case study over lower gangetic plain

Arabinda Maiti<sup>a</sup>, Prasenjit Acharya<sup>a</sup>, Srikanta Sannigrahi<sup>b</sup>, Qi Zhang<sup>c,d</sup>, Somnath Bar<sup>e</sup> , Suman Chakraborti<sup>f</sup>, Bijoy K Gayen<sup>g</sup>, Gunadhar Barik<sup>a</sup>, Surajit Ghosh<sup>h</sup>  and Milap Punia<sup>f</sup>

<sup>a</sup>Department of Geography, Vidyasagar University, Midnapore, West Bengal, India; <sup>b</sup>School of Architecture, Planning and Environmental Policy, University College Dublin Richview, Clonskeagh, Dublin, Ireland; <sup>c</sup>Department of Earth and Environment, Boston University, Boston, MA, USA; <sup>d</sup>Frederick S. Pardee Center for the Study of the Longer-Range Future, Frederick S. Pardee School of Global Studies, Boston University, Boston, MA, USA; <sup>e</sup>Department of Geoinformatics, School of Natural Resource Management, Central University of Jharkhand, Ranchi, India; <sup>f</sup>Center for the Study of Regional Development, Jawaharlal Nehru University, New Delhi, Delhi, India; <sup>g</sup>Department of Remote Sensing, Vidyasagar University, Midnapore, West Bengal, India; <sup>h</sup>International Water Management Institute (IWMI), Colombo, Sri Lanka

## ABSTRACT

We proposed a modification of the existing approach for mapping active paddy rice fields in monsoon-dominated areas. In the existing PPPM approach, LSWI higher than EVI at the transplantation stage enables the identification of rice fields. However, it fails to recognize the fields submerged later due to monsoon floods. In the proposed approach (IPPPM), the submerged fields, at the maximum greenness time, were excluded for better estimation. Sentinel-2A/2B time-series images were used for the year 2018 to map paddy rice over the Lower Gangetic Plain (LGP) using Google earth engine (GEE). The overall accuracy (OA) obtained from IPPPM was 85%. Further comparison with the statistical data reveals the IPPPM underestimates (slope ( $\beta_1$ ) = 0.77) the total reported paddy rice area, though  $R^2$  remains close to 0.9. The findings provide a basis for near real-time mapping of active paddy rice areas for addressing the issues of production and food security.

## ARTICLE HISTORY

Received 12 October 2021  
Accepted 17 January 2022

## KEYWORDS

Paddy rice mapping; Sentinel-2; Google earth engine; lower gangetic plain; random forest

## 1. Introduction

Among the most populous countries of the world, India stands second, residing 1.6 billion people that contribute 17 per cent of the world population with the massive demand for food crops (Lehane 2014). Among the major cereal crops, paddy rice is among the vital staple food grains for over half of the world population and occupies around 12% of the global cropland area (Elert 2014; Dong et al. 2016). The global rice production, as per

**CONTACT** Prasenjit Acharya  [prasenjit.ac@gmail.com](mailto:prasenjit.ac@gmail.com); [prasenjit.ac@mail.vidyasagar.ac.in](mailto:prasenjit.ac@mail.vidyasagar.ac.in)

 Supplemental data for this article is available online at <http://doi.org/10.1080/10106049.2022.2032396>

© 2022 Informa UK Limited, trading as Taylor & Francis Group

2018s FAO estimates, is about 759 million tons (FAO 2018), and the forecast pattern shows it will continue to increase with demand keeping in pace with the increasing global population. The combined share of China, India, Bangladesh and Indonesia to the total global production of rice is 66%. It is expected to increase in the next decade due to the growing population. However, there is an interannual shortfall in the production that is happening due to weather severity, especially during the monsoon season. Periodic occurrence of droughts and floods in major rice-growing areas in the south and south-east Asia due to changing climate poses an impending threat to agriculture (Mendelsohn 2014; Miyan 2015). The IPCC report on climate change in South Asia shows a likely increase of 2–4° C in temperature in the coming decades, and that results in an approximate decrease of 30% in crop production (Hijioka et al. 2014). Being an important production area of paddy rice in south Asia, the Gangetic plain - a part of the northern Indian plains - may also face a decrease in paddy rice production. As India shares a significant portion of the global rice production and its international marketing, the uncertainty of the paddy rice production in changing climate may lead to future food security issues. In such context, it is imperative to have adequate information on the total cropped area and its dynamics over time in south and south-east Asia with special reference to paddy rice.

In the last two decades, a series of methods have been proposed to mapping cropped areas, especially the irrigated rice area mapping from remotely sensed images. Several remote sensing-based agricultural layers, including Moderate Resolution Imaging Spectroradiometer (MODIS), Medium Resolution Imaging Spectrometer (MERIS), and Finer Resolution Observation and Monitoring of Global Land Cover (FROM-GLC), are available for mapping of paddy cultivation area from local to global scales. A range of different studies was done in the past to map the cropped area under paddy cultivation through phenology-based information, derived from time-series satellite images (Dong et al. 2016; G. Zhang et al. 2017; Zhang et al. 2015). Additionally, paddy-rice mapping was done by using supervised Maximum Likelihood (Oguro et al. 2001), Support Vector Machine (Xu et al. 2018), Random Forest (Qadir and Mondal 2020) and unsupervised classifications such as Iso-cluster or K-means (Pan et al. 2010; Nguyen et al. 2012). The use of the time-series spectral vegetation indices for crop mapping, e.g., Enhanced vegetation index (EVI), Normalized difference vegetation index (NDVI), Land surface water index (LSWI), have an added advantage over the multispectral images because time-driven vegetation indices represent phenological events of a crop cycle which could be used effectively in the threshold-based algorithm for mapping paddy-rice (Shew and Ghosh 2019; Yin et al. 2019). The time series of flooding signal (FS), derived from LSWI, was used in several studies for identifying the rice pixels during the transplantation phase. In contrast, time-series of NDVI and EVI were used in the early growing to harvesting stage for mapping paddy-rice (Xiao et al. 2005; Xiao et al. 2006; Teluguntla et al. 2015; Dong et al. 2016; Zhang et al. 2018). All of these studies evidenced effective pixel-based mapping of rice areas using phenology-driven time-series spectral indices. A summary of phenological pixel-based paddy-rice mapping (PPPM) using moderate to fine resolution images, such as MODIS, Landsat 7 and 8, and Sentinel 1 and 2, is given in Table 1. Many of these reported studies, which used coarse spatial (MODIS, 500 m) and low temporal resolution (Landsat, 16-days), have faced challenges in classifying and delineating paddy rice pixels in the heterogeneous landscape (Gallego 2004; Zhen et al. 2013; Zhou et al. 2013). As the south and south-east Asian regions are mostly characterized by small and fragmented paddy rice fields, the use of MODIS often results in low classification accuracy emanating from mixed pixels containing many other types of crops (Teluguntla et al. 2015). However, the accuracy of the classification could be improved by using Sentinel-2

**Table 1.** A literature summary of existing studies based on the Phenology and Pixel-based Paddy-rice Mapping (PPPM) method using the reflectance, vegetation indices, and LST data for paddy rice mapping.

Refs.	Sensor and product	Variables	Periods of data	Methods	Study area	Accuracy
Xiao et al. (2005)	MODIS	NDVI, EVI, LSWI, NDSI	Time series images in entire year	PPPM	Southern China	–
Xiao et al. (2006b)	MODIS	NDVI, EVI, LSWI, NDSI	Time series images in entire year	PPPM	South and Southeast Asia	–
Sun et al. (2009)	MODIS	EVI, LSWI	Time series images in entire year	PPPM	China	>73%
Peng et al. (2011)	MODIS	LSWI, EVI	Time series images in entire year	PPPM	Hunan, China	>84%
Teluguntla et al. (2015)	MODIS (MOD09A1)	NDVI, EVI, LSWI, LST	June 2000 to May 2010	PPPM	Krishna River Basin, India	–
J. Wang et al. (2015)	Landsat-8 OLI, MODIS, PALSAR	LSWI, EVI, NDVI, LST	Flooding periods (June 24 & July 10, 2013)	PPPM	Jiangsu Province, China	OA = 89.8%, Kappa = 79%
G. Zhang et al. (2015)	MODIS (MOD09A1, MOD11A2, MYD11A2)	NDVI, EVI, LSWI, NDSI, LST	Rice transplanting phase in 2014	PPPM	North-eastern China	OA = 97%, Kappa = 92
Dong et al. (2016)	MODIS (MOD09A1, MOD11A2), Landsat-8 (TOA)	NDVI, EVI, LSWI, LST	Within the rice transplanting phase in 2014	PPPM	North-eastern Asia	OA = 98%
G. Zhang et al. (2017)	MODIS (MOD09A1, MOD11A2)	NDSI, NDVI, EVI, LSWI, LST	From 2000 to 2015	PPPM	China and India	R <sup>2</sup> = 0.77–0.92
X. Zhang et al. (2018)	Landsat-8, Sentinel-2 A/B, Sentinel-1	NDSI, EVI, NDVI, LSWI, GCVI, C-band	March and November of 2017	PPPM, RF	Heilongjiang, Hunan, and Guangxi provinces in China	OA = 90%
Yin et al. (2019)	Landsat (ETM, OLI), MODIS (MOD09A1, MYD11A2)	NDVI, EVI, LSWI, LST	From 2013 – 2018	PPPM, RF	North-eastern China	OA = 98.19%
G. Zhang et al. (2020)	MODIS (MOD09A1, MYD11A2, MOD12Q1)	NDSI, NDVI, EVI, LSWI, LST	From 2000–2015	PPPM	Monsoon Asia	–
Ding et al. (2020)	Landsat (ETM, OLI), MODIS (MOD09A1)	NDVI, EVI, LSWI	2015	PPPM	Poyang Lake Plain, Southern China	OA = 93.66%, Kappa = 0.85

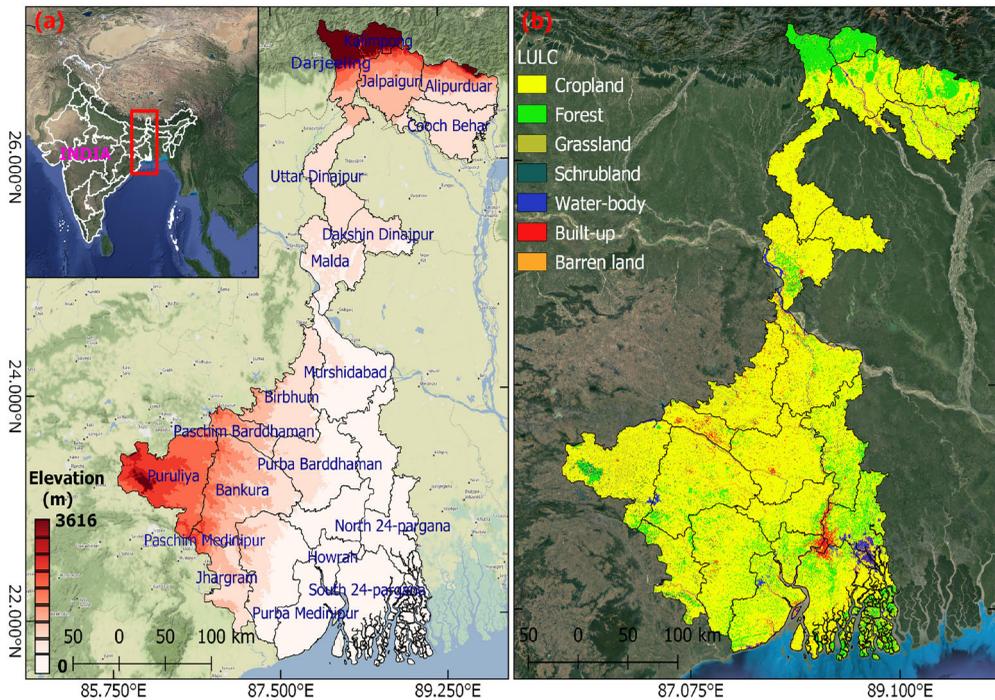
Note: MODIS = Moderate Resolution Imaging Spectroradiometer, GCVI = Green Chlorophyll Vegetation Index, EVI = Enhanced Vegetation Index, NDVI = Normalized Difference Vegetation Index, LST = Land surface temperature, NDSI = Normalized Difference Snow Index, LSWI = Land Surface Water Index, PPPM = Phenology and Pixel-based Paddy-rice Method, RF = Random Forest, OA = Overall Accuracy.

dense time-series data that provides better spectral and radiometric resolutions along with the high frequency of revisit cycle and thereby captures spatial diversity in the crop management practices (Liu et al. 2020). The PPPM was applied for paddy rice mapping in China (Xiao et al. 2005b; Wang et al. 2015; Zhang et al. 2015), South and South-east Asia (Xiao et al. 2006; Dong et al. 2016; Zhang et al. 2020), and in India (Teluguntla et al. 2015). These studies were conducted using varied datasets, starting from high temporal but coarse spatial resolution (MODIS 500 m) to finer spatial with lower temporal (revisiting interval) resolution (Landsat-8 30 m imagery) data. The magnitude of the overall accuracies that these studies were reported was high enough (89 – 98%) to comprehend the model's efficacy in differentiating the paddy-rice pixels from other crop pixels. However, it is unclear whether the PPPM model performs at the same efficacy in the Lower Gangetic Plain (LGP), which is characterized by small landholding size – smaller than a Landsat-8 pixels size, and frequent submergence of the field due to flood and heavy rain due to monsoon during sowing through physiological maturity stage. For the PPPM approach, the difference of LSWI and EVI during sowing through transplantation stages was used to select either static or dynamic thresholds for mapping paddy rice (Xiao et al. 2005b; Peng et al. 2011). Such condition is useful in mapping paddy rice fields as paddy rice needs staggged water for couple of weeks for growth after transplantation. However, the mapping also includes any surface flooded with the water. This enhances the chances of wrongly classifying the paddy rice fields. As LGP is characterized by the frequent flooding events throughout the monsoon season due to periodic heavy rain, using a threshold approach only during sowing through transplantation stage might yield an inaccurate estimation of active paddy rice fields. Therefore, in this study we propose a modification to the existing PPPM method through incorporation of the difference of LSWI and EVI later in the middle of the crop cycle. We believe that such inclusion improves the estimation of crop area under the active paddy cultivation. The objectives of the present study are (a) to map the paddy rice area during the monsoon using suggested modification of PPPM method, (b) to estimate the efficiency of the proposed method with reference to the existing PPPM and standard machine learning based random forest classification output, and (c) to compare the estimated paddy rice statistics with the national agricultural statistics.

## 2. Materials and methods

### 2.1. Study area

The LGP shares the international territory between India and Bangladesh. It is one of the world's most extensive flood plain regions (Singh et al. 1998). In this study, we took the Indian part of the LGP that primarily consists of the eastern Indian state, West Bengal (Figure 1). The state of West Bengal comprises 22 districts, as mentioned in Figure 1. Topographically, the western part of the study area is connected with the plateau fringe with undulating topography (elevation of 400 – 590 m above mean sea level), while the northern part covers Shiwalik Himalaya with elevation ranges up to 3,616 m (Figure 1a). Three major river systems, i.e., Teesta in the north, Bhagirathi-Hooghly of Ganga River in the south, and Ajay, Damodar, and Dwarkeshwar river system in the west, feed this alluvial floodplain leading to an intense cropping practice. The climate of the area is governed by the monsoon system (Attri and Tyagi 2010; Jin and Wang 2017). The study area is characterized by frequent floods throughout the monsoon season due to the combined



**Figure 1.** (a) Study area: location of the Lower Gangetic Plain. A digital elevation model (DEM) shows altitude variation from east-west and north-south. (b) Land use land cover pattern.

effect of heavy rainfall and the release of the water from reservoirs (Jha and Bairagya 2013), leading to crop yield loss because of submergence of the crop fields.

The total geographical extent of the study area is around 8.68 million ha, of which 5.25 million ha is the net sown area, comprising 68% of the geographical area. The study area is known for its intense cropping practices with an average landholding size of 0.76 ha (<http://matirkatha.gov.in/agricultural-scenario-in-west-bengal/>). Most of the croplands are under a triple cropping pattern, with rice (*Oryza sativa*) as the predominant crop. The crop calendar starts with monsoon crops (*Kharif* crops: from July to November) (Prashnani et al. 2014) followed by winter crops (*Rabi* crops: from December to April) and ends with summer or pre-monsoon crops (*Zaid* crops: from May to July). In West Bengal, the monsoon crop is dominated by paddy rice followed by redgram, grown as a secondary crop. The winter crops are also dominated by rice, followed by gram, lentil and wheat, and the summer crop is dominated by gram, blackgram and pulses, including the pre-monsoon paddy rice.

## 2.2. Sentinel-2 multispectral data and pre-processing

In this study, Sentinel-2A/B Multi-Spectral Instrument (MSI) Top-of-atmosphere (TOA) reflectance (Level-1C) data was used from July 2018 to December 2018. The Sentinel-2 MSI images with 10 m spatial resolution and a 5-day repetitive coverage over an area around the globe are provided by the European Space Agency (ESA). The TOA reflectance data was used due to the limited access of the Sentinel – 2 surface reflectance

product before 2019. Additionally, the choice of the TOA product was guided by the selected timeline for this study which is also steered by the availability of the statistical data - discussed in section 2.4 – as one of the reference data for this study.

The wavelengths of the Sentinel-2 bands range from 442.3 nm to 2,185.7 nm with 10 m (B2, B3, B4, B8), 20 m (B5, B6, B7, B8A, B11, B12), and 60 m (B1, B9, B10) spatial resolutions. The time series TOA reflectance products were called in Google earth engine (GEE) cloud computing facility - made for planetary-scale analysis (Gorelick et al. 2017), and the corrections were made for reducing the atmospheric effect and cloud cover. The TOA reflection was corrected by applying sensor invariant atmospheric correction (SIAC) module, developed by Yin et al. (2019), in the GEE. The SIAC exploits operational global datasets on bi-directional reflection distribution function (BRDF) and coarse resolution atmospheric data to estimate surface reflection from TOA product. The method uses a Bayesian framework to model surface reflection using standard BRDF product from Moderate Resolution Imaging Spectroradiometer (MODIS) and atmospheric products from Copernicus Atmospheric Monitoring Services. It gives information of aerosol optical depth and total columnar water vapor thickness with the surface reflection from TOA reflection product in due course of solving the inverse radiative transfer problem. After the atmospheric correction, the cloud pixels were removed from the scenes. Since the monsoon cloud varies in brightness over the study area, the cloud pixels were removed based on the cloud-score algorithm (Qadir and Mondal 2020; Nanshan et al. 2021). The total number of image before and after the cloud removal was counted within the selected time interval of July 2018 to December 2018 and represented in Supplement Figure 1. The removal of cloud from the images produces gap. We further processed these atmospherically corrected Sentinel-2 time series images following the gap-filling process by considering the mean of total six consecutive pixels (three pixels below and three above a gap pixel) in the time series which spans over 30 days.

The spectral indices of EVI and LSWI were computed from these corrected time series images using Equation 1 and 2. Prior to computing the spectral indices, the SWIR-1 image layers were resampled from 20 m to 10 m with reference to other spectral channels involved in the analysis. Moreover, a linear fit smoothing function was applied on time series EVI and LSWI to reduce the residual effects of atmosphere emanating from cloud, aerosols and columnar water vapour. Linear fit uses a simple moving window algorithm where a subset of data from the whole data set is used to estimate the coefficients of linear fit (Khanal et al. 2020). The window size of the linear fit smoothing was decided based on the  $R^2$  value. We used a window size of 20-time steps that has an  $R^2$  of 0.93 (Figure 2).

$$EVI = \frac{2.5(\rho_{nir} - \rho_{red})}{\rho_{nir} + 6\rho_{red} - 7.5\rho_{blue} + 1} \quad (1)$$

$$LSWI = \frac{\rho_{nir} - \rho_{swir1}}{\rho_{nir} + \rho_{swir1}} \quad (2)$$

### 2.3. Lulc data

We used the cropland land layer from a standard global Land Use Land Cover (LULC) classification (Figure 1(b)), available at the 10 m resolution based on Sentinel-2 data (Gong et al. 2019). This global LULC layer was prepared using the Random Forest (RF) classifier, where the training set contains approximately 340,000 samples across the world, with an overall accuracy of the classification of 72.76%. The cropland pixels from this

LULC layer were separated to build a thematic layer of cropland that was further used to mask the time series LSWI and EVI pixels. Prior to masking, the accuracy of the cropland mask was evaluated using the ground truth points (GTP) – discussed further in the section 2.5. The accuracy metrics of the cropland layer from the LULC product of Sentinel – 2 is given in Supplement [Table 1](#).

The global permanent surface water layer – prepared from Landsat 5, 7 and 8 at 30 m resolution - was also used in this study (Pekel et al. 2016). The purpose is to increase the classification accuracy by removing the spectral response from the aquatic weeds which replicates the spectral response of vegetation and cropland in the NIR and Red bands (Silva et al. 2008; Schmidt and Witte 2010; Singh et al. 2020; Akbari et al. 2021).

#### **2.4. Regional cropland statistics**

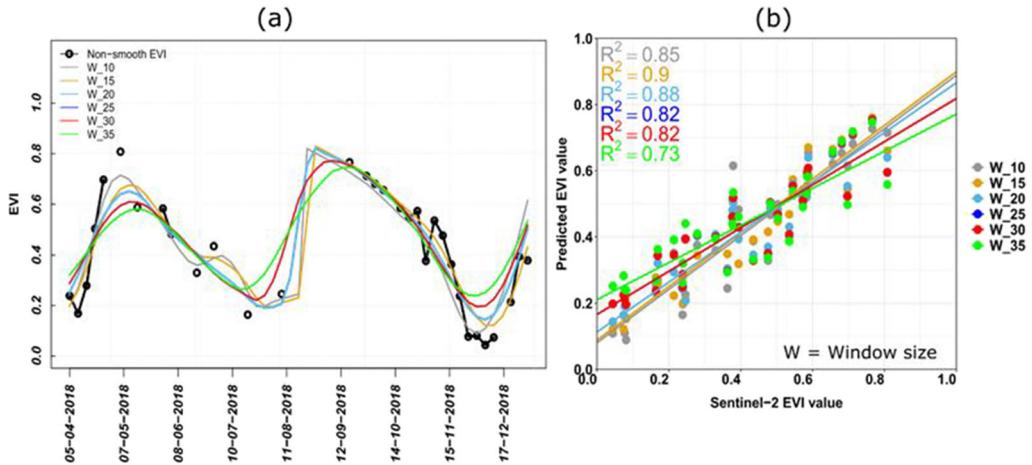
We used the district-level net sown area (NSA) and the paddy rice area statistics during monsoon season. The NSA, including the area statistics for different crop types for each cropping season, is assessed by the Directorate of Economics and Statistics (DES) under the Ministry of Agriculture (MoA), Govt. of India (<http://eands.dacnet.nic.in/>). The MoA conducts yearly field surveys on the nine-fold classification of land use and land cover. It includes the irrigated area (source wise and crop-wise) and total area under the crops at the district level for the whole country. The DES data contain seasonal crop statistics, i.e., NSA of different crops, crop yield and production per year, etc. The agricultural statistics of monsoon paddy rice was taken from DES for the year 2018.

#### **2.5. In-situ data**

A field survey was conducted in 2018 for collecting adequate training samples for conducting the rice detection from Sentinel – 2. A sum of 1263 ground truth points (GTP), which include rice and non-rice (e.g., other crops except paddy rice, vegetation, fallow land, build-up) fields, was collected by Garmin GPS (Garmin eTrex 10; accuracy  $\sim 3.65$  m). All GTPs were collected by a random sampling procedure. Out of 1263 GTPs, a total of 665 (52.9%) GTPs were collected over paddy rice fields during the monsoon season, and 598 (47.1%) GTPs form non-rice fields (Supplement [Figure 2](#)). For the purpose of reducing the uncertainty of positional vector, we selected the sample paddy fields which were within the continuous clusters of agricultural land having a size greater than 2500 m<sup>2</sup>. All of these GTPs were further visually checked using high-resolution Google Earth and Sentinel-2 RGB composition to ensure the reliability of the data.

#### **2.6. Precipitation data**

We used daily precipitation data from Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) in support of upholding the hypothesis we made at the beginning. Since heavy precipitation causes waterlogging situations at least for the low lying or poorly drained areas, precipitation anomaly can reveal the spatial distribution of potential waterlogged areas, and that helped in the detection of active paddy rice fields from Sentinel-2. The CHIRPS-based daily precipitation product is modelled from ground-based observations and satellite data (Funk et al. 2015). The precipitation data towards the end of monsoon season, 20<sup>th</sup> September to 10<sup>th</sup> October, was called in GEE-interface (Gorelick et al. 2017). We took the mean precipitation for the said time-window for the



**Figure 2.** Smoothing time series curve of spectral indices based on linear fitting. (a) An example of linear fit over enhanced vegetation index (EVI) using different window size, (b) The measurement of goodness of fit using coefficient of determination ( $R^2$ ) for different window size. For the purpose of smoothing the time series spectral index was taken from 1<sup>st</sup> April 2018 to 31<sup>st</sup> December 2018.

year 2018. The mean precipitation for 2018 was subtracted from the long-term average from 1988 to 2017 compute the precipitation anomaly.

## 2.7. PPPM Algorithm and its improvement

The PPPM is a well-recognized algorithm to detect paddy rice fields from optical remote sensing data (Xiao et al. 2005b). The paddy rice has three main phenological stages covering a period of approximately 100-120 days. The typical three stages of paddy rice over Asia during the monsoon are a) the flooding/transplanting stage (day of the year (DOY): 213 – 262 [1<sup>st</sup> August – 19<sup>th</sup> September]), b) the growth stage, which is featured by vegetative growth, flowering, and reproductive activity (DOY: 263 – 283 [20<sup>th</sup> September – 10<sup>th</sup> October]), and c) the harvest phase (DOY: 284 – 334 [11<sup>th</sup> October – 10<sup>th</sup> December]) with mixture of bare soil and rice plant residues. In the transplanting phase, land surface pixels are often mixed with water and sporadic green plants (Teluguntla et al. 2015). After 50–60 days from the transplanting date, the standing water in the field, which is approximately 2 – 10 cm deep, is mostly encroached by the rapidly growing canopy (Xiao et al. 2005; 2006). The growth curve, at its peak, is called maximum greenness time (MGT), which is characterized by closure of background soil through expansion of foliar surface. Following this physical maturity stage, the leaf chlorophyll and moisture gradually decrease and become minimum at the harvesting stage.

The spectral reflectance of the water suggests the presence of a strong water absorbance band at SWIR (1630 – 2130 nm), leading to low surface reflectance of water body in SWIR, while it is relatively higher in the NIR channel ( $\sim 856$  nm) (Jensen 2014). The spectral response of water also shows a higher reflection in Red and a lower reflection in NIR. Thus, a water surface yields a positive value in the spectral index of LSWI while it gives a negative value in the spectral index of EVI. During the transplantation stage, a high positive value of LSWI (Equation (2)) and a relatively lower EVI (Equation (1)) value suggests the presence of stagnant water in the rice field (Gao 1996). After 3 – 4 weeks of transplanting, the paddy rice fields are converted into a mixture of surface water and paddy rice plant, as the growing canopy of the paddy rice reduce the moisture laden

background soil surface area and enhance the reflectance in the NIR channel, leading to increase in EVI signal in comparison to LSWI. The PPPM, as given by Xiao et al. (2005b), hypothesized that a temporary inversion of vegetation index leads to mapping of the flooding and transplanting pixels, and thus, it uses a conditional approach combining LSWI, NDVI and EVI (Equation (3)) for identifying the flooded paddy rice pixel during the transplantation stage. For this study, we used a combination of LSWI and EVI at the transplantation stage. A constant of 0.05 was used in Equation (3), as proposed by Xiao et al. (2005b), for considering the fact that the difference of EVI and LSWI must be less than 0.05 or 5% for a range of different statuses of soil-water-vegetation complex at the time of transplantation of paddy rice ( $EVI - LSWI \leq 0.05$ ). The EVI was chosen over NDVI due to its better sensitivity to the rich biomass content that is generally achieved within 4–5 weeks after transplantation (Huete et al. 2002). We further used a modified form of Equation (3) at the MGT stage, which mathematically remains similar, but the condition that is used here (Equation (4)) is different from Equation (3). The paddy rice pixels can be identified during the transplantation stage by selecting a value greater than 0 – given in Equation (3) – from the time series of the difference of LSWI and EVI. However, as the beginning of the paddy rice during monsoon is associated with the frequent flood, the chance of satisfying the criteria, expressed in Equation (3), by other standing crops is unavoidable. Despite of having differences in the mean vegetation vigore – manifested by average EVI for the whole crop cycle – the crops, other than paddy rice in the monsoon season, which do not need waterlogged situation at the beginning of their phenological cycle, may classified under paddy rice due statisfaction of critetion stated in Equation (3). To deal with this issue and correctly classify the rice pixels, we used the following adjustment in the existing PPPM: a condition of  $[(LSW + 0.05) - EVI] < 0$  for mapping the paddy rice pixels was added towards the MGT stage (Equation (4)). Such condition helps to separate any crop fields submerged due to the presence of residual floodwater by the time of MGT from active paddy rice fields. Hereafter, the suggested improvement over the existing PPPM is called improved PPPM (IPPPM). Given the threshold conditions and prior assumption, i.e., LSWI could not be greater than the EVI during the MGT period, the non-rice, including some of the paddy rice crop pixels, which are flooded, could be discriminated from the active paddy fields when  $[(LSWI + 0.05) - EVI] > 0$ . A schematic representation of the methodology of this study is given in Figure 3.

$$Transplantation : \begin{cases} 1, \{(LSWI + 0.05) - EVI\} \text{ or, } \{(LSWI + 0.05) - NDVI\} > 0 = \text{Paddy rice} \\ 0, \{(LSWI + 0.05) - EVI\} \text{ or, } \{(LSWI + 0.05) - NDVI\} < 0 = \text{Non-paddy rice} \end{cases} \quad (3)$$

$$MGT : \begin{cases} 0, \{(LSWI + 0.05) - EVI\} > 0 = \text{Non-paddy rice} \\ 1, \{(LSWI + 0.05) - EVI\} < 0 = \text{Active Paddy rice} \end{cases} \quad (4)$$

A plot of samples of EVI, LSWI and  $[(LSWI + 0.05) - EVI]$ , meeting these criteria, is given in Figure 4. Time series layers were prepared for  $[(LSWI + 0.05) - EVI]$  from EVI and LSWI, and above-stated conditions were run for all layers to identify the potential paddy rice pixels during the transplantation stage that we considered from 4<sup>th</sup> July to 10<sup>th</sup> September, 2018. The process produced a binary layer of paddy rice and non-paddy rice. This binary layer was later used to mask pixels from time series  $[(LSWI + 0.05) - EVI]$  layer during the MGT (September, 20<sup>th</sup> - October, 10<sup>th</sup>), and the second condition – given in Equation (4) – was used to filter out the candidate pixels which were still in submerged condition, leading to correctly identify the paddy rice pixels. These two sets of date

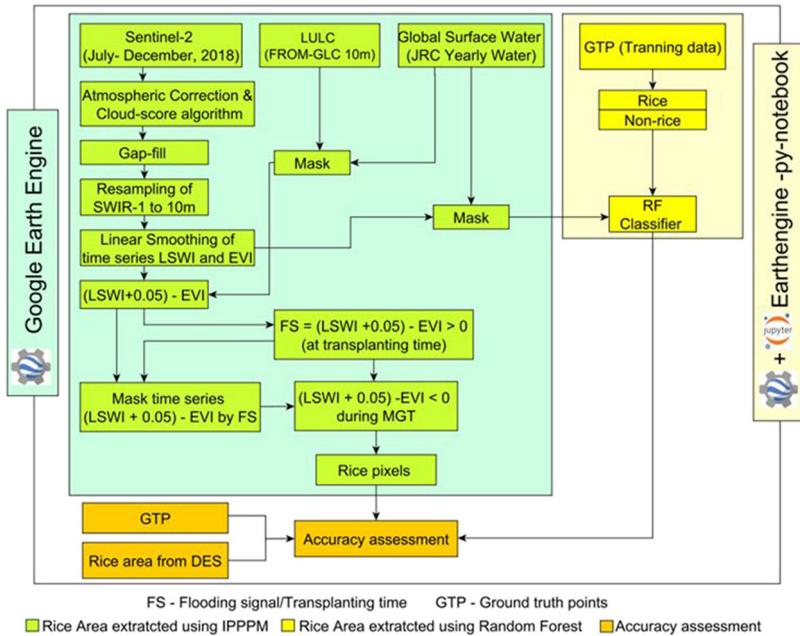


Figure 3. Experimental setup for paddy rice mapping over the LGP during monsoon.

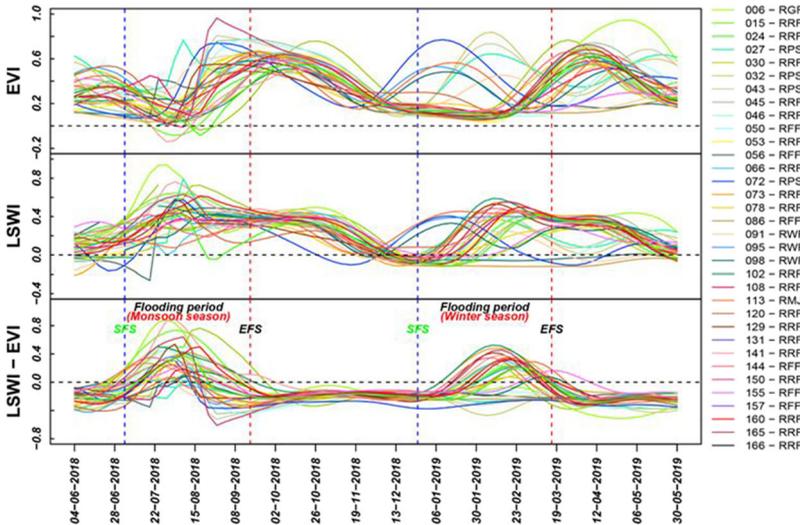


Figure 4. Time series curves of EVI, LSWI and LSWI-EVI, observed over single, double, and triple cropping patterns. For the purpose of demonstrating the variation of time signals over different cropping patterns, the time series length was increased from June 2018 to May 2019. The vertical SFS line represents start of flooding season, and EFS line represents end of flooding season during monsoon season. The cropping patterns on which the ground truth points were taken are given as follows: RGF = Rice-groundnut-fallow, RRF = Rice-rice-fallow, RPS = Rice-potato-sesame, RFF = Rice-fallow-fallow, RWF = Rice-wheat-fallow, RMJ = Rice-mustard-jute.

ranges, i.e., timelines for transplantation and MGT, were used as the average timeline based on the information gathered during the field survey, including a detailed examination of smoothed time series curves of EVI and LSWI. To compare the results of the proposed IPPPM, we estimated the paddy rice area during monsoon using PPPM, as

given in Equation 3 using EVI only. The area coverages of the paddy rice - at the district level - estimated from both these methods were then compared with the MoA reported area statistics under paddy rice.

## 2.8. Random Forest classification

We further compared the results of the proposed IPPPM with the machine learning-based Random Forest (RF) classification. The RF is a nonparametric and distribution-independent supervised classifier (Bargiel 2017). Modified from a single decision tree, the RF is featured by an ensemble of decision trees that used bootstrapping and bagging techniques for training the model (Breiman 2001). Because of its superior performances over other standard classification approaches, it was extensively used in studying the LULC and crop type mapping ( Hao et al. 2015; Li et al. 2015; Belgiu and Csillik 2018; Pelletier et al. 2016; S. Wang et al. 2019; Löw et al. 2018; Rudiyanto and Soh 2019).

The time series images of spectral indices such as LSWI and EVI – during transplantation through harvesting stage (July to December) - were used as input variables in the RF model. A sum of 11 image tiles of Sentinel – 2 were used in training and evaluating the RF classification model. Of these 11 image tiles, 3 tiles – from the central plain of the study area where most of the variations of crop and non-crop categories were observed - were used for training the RF model (Supplement Figure 2). The remaining 8 image tiles were used for testing the classification model. A sum of 72 time series (36 images of LSWI and 36 images of EVI) images per Sentinel-2 image tile were used, and it turned out to be 216 ( $3 \times 72$ ) time series images for training area and 576 ( $8 \times 72$ ) images for test area. A sum of 1,257 GTPs used for the RF classification. 70% (880 GTPs in which 478 and 402 GTPs were labelled as paddy and non-paddy rice targets, respectively) of these GTP were used to extract the samples of spectral indices (i.e., LSWI and EVI) from the training image tiles for building the model. Whereas, the sample observations of rest of the 30% GTPs (187 and 190 GTPs were labelled as paddy rice and non-paddy rice, respectively) were used from testing image tiles for model validation. The RF classification was performed on the Jupyter notebook using Python packages such as geopmap (v0.8.16) (Wu 2020) and Sci-kit learn (v0.24.4).

## 2.9. Feature selection

A total of 72 temporal chronological feature layers of EVI and LSWI were used as input to build the RF classification model. The model was optimized based on the measure of mean decrease in accuracy (MDA) for selecting ideal number of feature layers. We used hyperparameter tuning with 10-fold cross-validation to get the optimum number of features from a set of 72 chronological features. The best-resulted parameters are – number of tree (ntree): 150, maximum depth (max\_depth): 5, maximum features (max\_features): 7, maximum leaf nodes (max\_leaf\_nodes): 10, minimum samples in the leaf (min\_samples\_leaf): 3, and minimum samples for splitting each decision node (min\_samples\_split): 3. We have performed a feature selection step to filter the most contributed features based on MDA. The process turned out with a total of 24 important features, which contributed significantly to the classification.

## 2.10. Validation and comparison

The accuracy of the results from the IPPPM was evaluated in three different ways: (1) an error matrix was calculated between GTP and predicted paddy rice pixels, (2) the area statistics derived from both IPPPM and PPPM models were compared with the reported area under paddy rice from DES, Govt. of India, and (3) the performance of IPPPM algorithm-based classification were compared with RF-derived paddy rice map. It is to be noted that, for assessing the accuracy of the outcome from the IPPPM approach, the GTPs were used only from crop fields. Thus, the total number of GTP used for assessing the accuracy is 953.

Two measures, such as Area Difference (AD) and Percentage of Area Difference (PAD), were used to show the magnitude of error between the Sentinel-2-derived paddy rice area using the IPPPM algorithm and DES-reported paddy rice area (Equation 5-6).

$$AD = A_{Sentinel} - A_{DES} \quad (5)$$

$$PAD = \frac{A_{Sentinel} - A_{DES}}{A_{DES}} \times 100\% \quad (6)$$

where  $A_{Sentinel}$  and  $A_{DES}$  represent the area statistics from Sentinel and DES. Moreover, to compare the model performances with reference to DES statistics, the measures of root mean square error (RMSE) and Chi-square statistics were used.

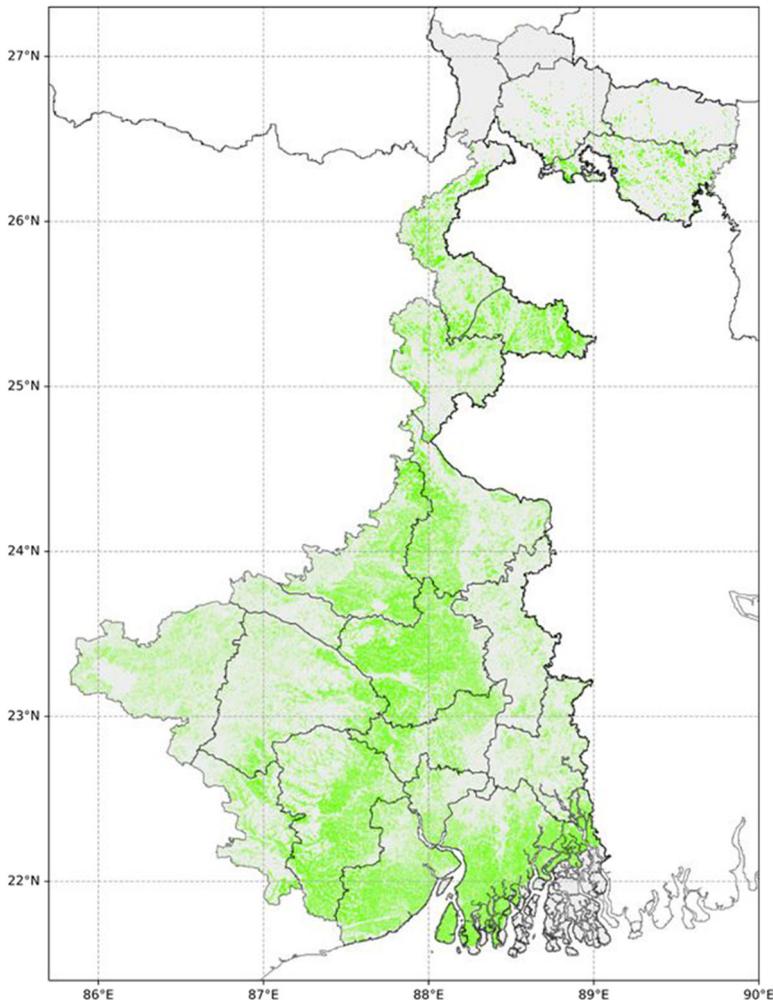
## 3. Results

### 3.1. Spatial distribution of paddy rice area in the lower gangetic plain

The paddy rice area during the monsoon period through the IPPPM approach is measured at around 2.98 million ha (mha), which is about 56.72% of the net sown area. The spatial distribution of mapped paddy rice area, shown in Figure 5, manifests that the territories in the central plain and south-east part of the West Bengal, i.e., Medinipur (East), Medinipur (West), Hooghly, Bardhaman (East), Bardhaman (West), Bankura, Birbhum and South 24-Parganas, respectively, share a significant amount (> 60%) to the NSA (Table 2). In comparison, the territories in the northern part of West Bengal share less than 50% to the NSA.

### 3.2. Accuracy assessment

The overall accuracy estimated from IPPPM is about 85%, including producer's (PA) and user's accuracy (UA) of 97% and 82%, respectively for paddy rice, and PA and UA of 65% and 93% for non-paddy rice (Table 3). The RF model, with 24 best-featuring variables and a mean out-of-bag error of less than 0.05 (Figure 6(a) and (b)), shows an overall accuracy of 88% for separating paddy rice and non-paddy rice fields, including PA and UA of 90.3% and 85.5% for paddy rice, and 86.5% and 91%, respectively for non-paddy rice (Table 3). For all optimized featuring variables in the RF classifier, the contribution of LSWI during September to December overwhelmed the others. The ROCs, as derived from the RF model for both paddy rice and non-paddy rice, show high TPR, including AUC more than 0.9 (Figure 6(c)), suggesting a good level of agreement with the in-situ observations.



**Figure 5.** Spatial distribution of paddy rice fields derived from Sentinel-2 datasets in the lower part of Ganga River Basin during monsoon 2018.

### **3.3. Comparison among PPPM, IPPPM, and RF models with DES statistics**

The paddy rice area, estimated through the PPPM, is about 3.09 mha, while after applying the correction through the IPPPM method, about 2.98 mha paddy rice area is estimated. The difference is around 0.11 mha between the PPPM and IPPPM approaches. The difference is generated from the elimination of paddy fields following Equation (4) due to stagnation of rain water at the MGT stage. At the district level, the estimated paddy rice areas from IPPPM, including the amount of area from the PPPM approach, shows the territories in the south of West Bengal (South 24-Parganas, Medinipur (East), North 24-Parganas, Murshidabad), with average elevation less than 10 m, suffers mostly from water-logging condition due to monsoon rain leading to loss of active paddy rice area (Figure 7). The results also demonstrate that, besides the territories mentioned above, a small amount of paddy fields is affected too either by flood or by water logging due to heavy rain in the rest of the districts. Supplement Figure 3 demonstrates evidence of this water logging due to heavy rain during the MGT stage.

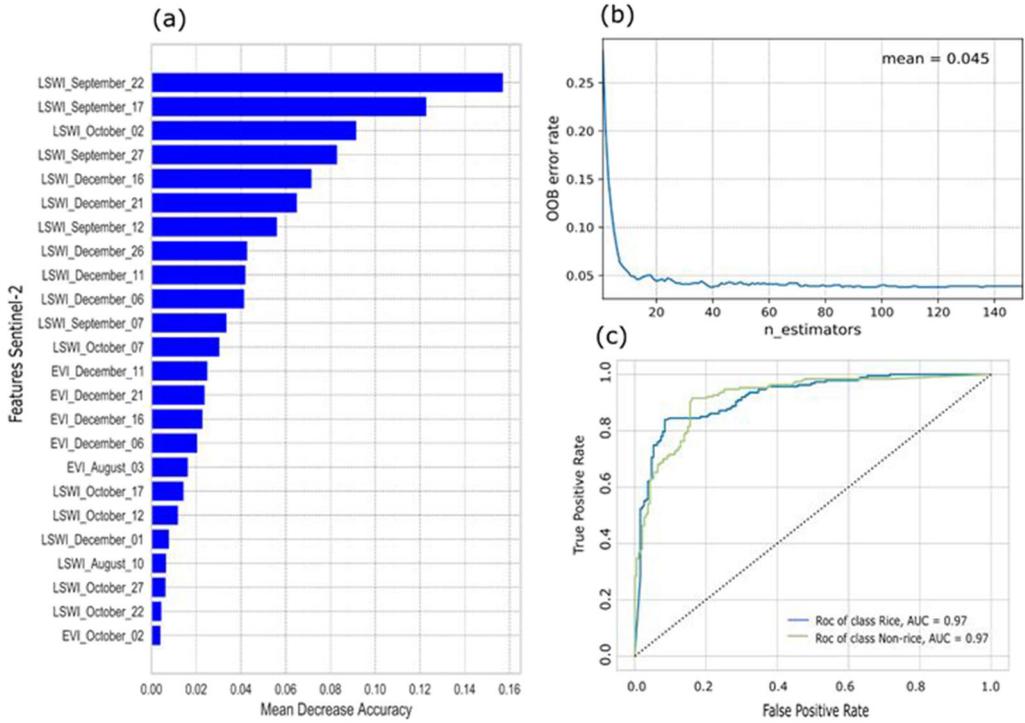
**Table 2.** District wise net sown area, taken from the statistical database of Directorate of Economics and Statistics (DES), Ministry of Agriculture, and percentage of paddy rice area derived from the IPPPM approach.

District	Net sown area (NSA) (1,000ha)	Sentinel-2 Paddy Rice (1,000ha)	% paddy rice area with respected to NSA
Alipurduar	135.27	48.82	36.09
Bankura	339.87	223.2	65.67
Birbhum	326.53	209.38	64.12
Cooch Behar	256.93	149.43	58.16
Dinajpur (North)	273.84	122.23	44.64
Dinajpur (South)	187.49	91.99	49.06
Darjeeling	115.27	14.73	12.78
Hooghli	210.89	130.57	61.91
Howrah	82.22	46.26	56.27
Jalpaiguri	203.77	83.91	41.18
Jhargram	137.87	94.95	68.87
Kalimpong	23.26	0.29	1.25
Malda	233.68	95.08	40.69
Murshidabad	398.34	216.11	54.25
Nadia	293.57	91.18	31.06
North 24-parganas	230.94	89.54	38.77
South 24-parganas	358.44	277.65	77.46
Barddhaman (East)	400.04	318.26	79.56
Barddhaman (West)	55.04	42.09	76.47
Medinipur (East)	286.6	184.46	64.36
Medinipur (West)	383.88	298.13	77.66
Purulia	313.61	147.84	47.14
Total	5247.33	2976.10	56.72

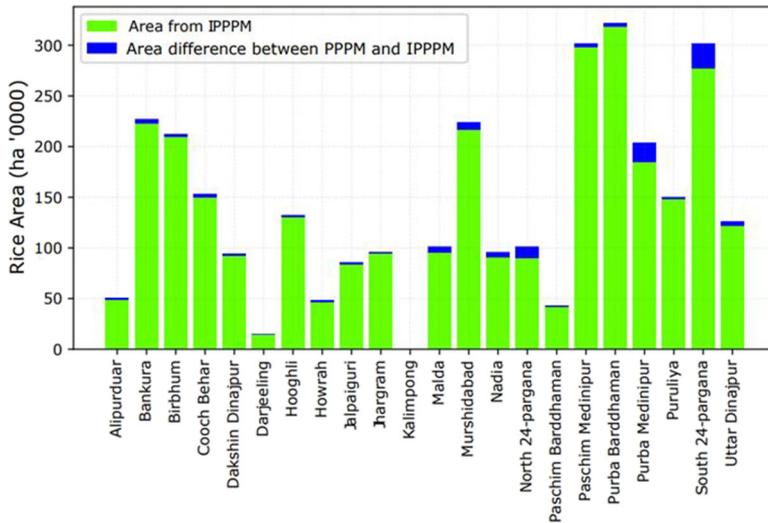
**Table 3.** Accuracy error matrix in the monsoon season for IPPPM, and Random Forest-based classification methods.

	IPPPM			User's Accuracy	Overall Accuracy	Kappa
	Paddy rice	Non-paddy rice	Total			
Paddy rice	592	122	714	0.82	0.85	0.66
Non-paddy rice	15	224	279	0.93		
Total	607	346	953			
Producer's Accuracy	0.97	0.65				
	Random Forest			User's Accuracy	Overall Accuracy	Kappa
	Paddy rice	Non paddy rice	Total			
Paddy rice	160	27	187	0.85	0.88	0.76
Non paddy rice	17	173	190	0.91		
Total	177	200	377			
Producer's Accuracy	0.9	0.86				

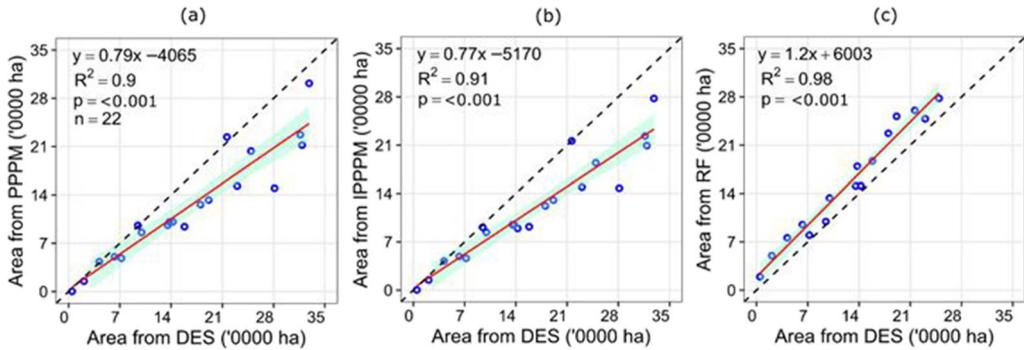
The estimated paddy rice area from the RF model is 4.93 mha, notably higher than the area estimated from both PPPM and IPPPM methods. The degree of correlation between the modelled paddy rice area, from PPPM, IPPPM and RF, and reported area from DES during the monsoon season demonstrated (Figure 8) that RF gives the best estimate with  $R^2$  of 0.98. While the overall performances of PPPM and IPPPM are not significantly different ( $R^2 = 0.90$  for PPPM and 0.91 for IPPPM) as the model comparison through Chi-square ( $\chi^2$ ) test reveals a  $\chi^2$  value of 1.67, which is less than the critical value of 11.59 at the p-value of 0.05 with 21 degrees of freedom. Yet, there is a credible difference in the magnitude of RMSE for PPPM (54885.1 ha) and IPPPM (58854.8 ha) results. The PAD of the estimated paddy rice area from IPPPM and DES shows a difference of about 1.02 mha (Table 4). The PAD at the district level shows many of the northern territories, including some in the south and south-west of the study area, have more than 30% difference. However, the PAD for the territories - belong to the central-plain - exhibit fewer than 25% difference.



**Figure 6.** (a) Variable importance of selected features of spectral indices explained through mean decrease accuracy, (b) out-of-bag error for random forest-based model, (c) receiver operating curve (ROC) for showing accuracy estimated through RF-based model.



**Figure 7.** Extracted paddy rice area by IPPPM approach and the difference of paddy rice area estimated from PPPM and IPPPM methods.



**Figure 8.** Scatterplot showing relationship between statistical data on monsoon paddy rice from Directorate of Economics and Statistics (DES) and (a) PPPM-derived paddy rice area, (b) IPPPM-derived paddy rice area and (c) RF model-based paddy rice area.

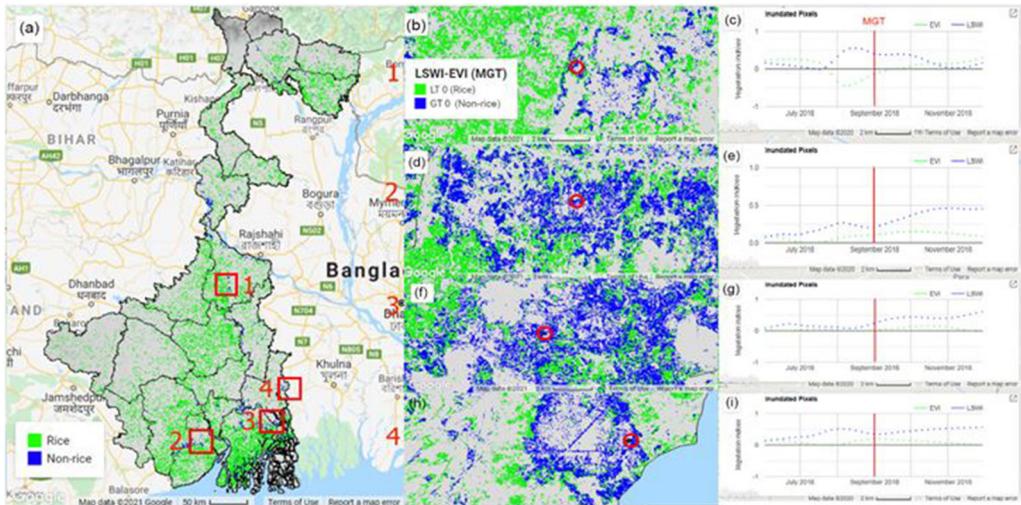
#### 4. Discussion

The PPPM-based algorithm detects the paddy rice pixels work reasonably well at the global scale (Xiao et al. 2005; Xiao et al. 2006). However, it faces some critical issues in identifying the paddy rice areas in monsoon Asia. The major issues related to this are, 1) the time series curve of paddy rice, mapped through LSWI and EVI, during the transplanting stage is similar to wetland or any waterlogged surfaces after precipitation, 2) often, after a short period of the transplantation stage, due to the occurrence of flood, the paddy fields are left untreated due to continued waterlogging leading to loss of production. In dealing with such issues, especially in monsoon dominated areas, the mapping of paddy rice through the PPPM-based approach might fail to estimate actual paddy rice area under active cultivation. In this study, we have modified the existing PPPM approach by allowing an additional criterion during the MGT stage (Equation (4)). The inclusion of the difference of LSWI and EVI at the MGT stage and its conditional criterion has efficiently mapped the active paddy rice area by excluding the submerged pixels. This results in a lower estimation of the total paddy rice area, which remains notably higher in the conventional PPPM approach. To show evidence of such overestimation, four spatial windows were selected, as given in Figure 9. The identified rice pixels through PPPM in those spatial windows are considered paddy rice as they were submerged during the transplantation stage. A notable amount of these pixels is still appeared with a high LSWI value in the MGT, suggesting that they are either destroyed (due to submergence) paddy rice fields or non-paddy rice areas with staggged water (Figure 9(b)–(i)). Further analysis of the precipitation anomaly through CHIRPS data at the MGT during monsoon 2018 suggests that about 74% of the area have faced negative rainfall anomaly, while some of the areas in the northern and southern part of West Bengal – amounting to about 26% of the total area - have faced positive rainfall anomaly (supplement material, Figure 4). Many of these areas, as evidenced in Figure 9, show signs of waterlogging during the MGT stage. Therefore, the occurrence of heavy rain, irrespective of rainfall anomaly, during the MGT stage is associated with the waterlogging condition due to the flat alluvial terrain (supplement material, Figure 3).

The evaluation of the performance of IPPPM with reference to PPPM and standard RF-based classification model, while comparing the results with DES statistics, suggests the RMSE is higher for IPPPM (58854.8 ha) as compared to PPPM (54885.1 ha) and RF (51064.1 ha) since it excludes the submerged pixels due to their potential loss. A high degree of association of estimated paddy rice area (slope ( $\beta_1$ ) = 1.2 with  $R^2 = 0.98$ ) from

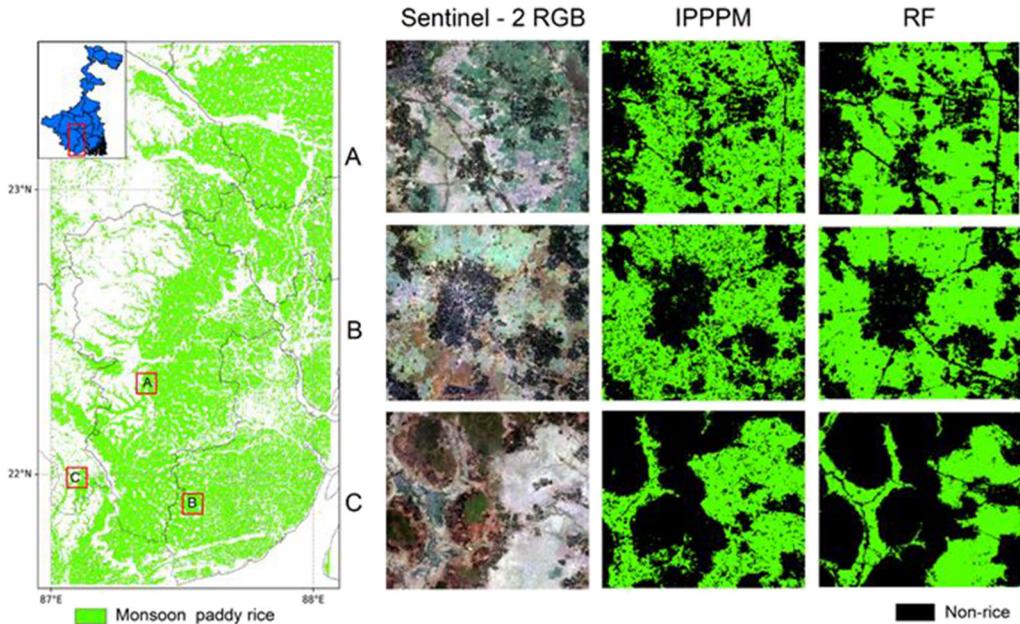
**Table 4.** Comparison of derived paddy rice area (ha) from Sentinel-2 using the PPPM, IPPPM and RF classification models with the area statistics reported by the Directorate of Economics and Statistics (DES) in monsoon season.

District	DES				PPPM (ha '000)				IPPPM (ha '000)				RF (ha '000)			
	(ha '000)	Sentinel-2	Area difference (AD)	Percentage area difference (%)	Sentinel-2	Area difference (AD)	Percentage area difference (%)	Sentinel-2	Area difference (AD)	Percentage area difference (%)	Sentinel-2	Area difference (AD)	Percentage area difference (%)	Sentinel-2	Area difference (AD)	Percentage area difference (%)
Alipurdwar	62.66	50.46	-12.19	-19.46	48.82	-13.83	-22.07	93.51	30.86	49.25						
Bankura	316.5	227.18	-89.32	-28.22	223.2	-93.3	-29.48	416.71	100.2	31.66						
Birbhum	318.97	212.17	-106.8	-33.48	209.38	-109.59	-34.36	373.79	54.83	17.19						
Cooch Behar	230.37	153.05	-77.32	-33.56	149.43	-80.94	-35.14	238.89	8.53	3.7						
Dinajpur (North)	180.08	126.23	-53.85	-29.9	122.23	-57.85	-32.13	262.44	82.36	45.73						
Dinajpur (South)	158.21	94.15	-64.06	-40.49	91.99	-66.22	-41.86	173.25	15.04	9.51						
Darjeeling	21	15.11	-5.89	-28.05	14.73	-6.27	-29.86	49.73	28.73	136.84						
Hooghli	191.23	132.87	-58.36	-30.52	130.57	-60.67	-31.72	250.4	59.17	30.94						
Howrah	72.09	48.44	-23.65	-32.81	46.26	-25.83	-35.83	78.65	6.56	9.1						
Jalpaiguri	99.74	85.86	-13.88	-13.91	83.91	-15.82	-15.87	126.5	26.76	26.83						
Jhargram	135.37	96.11	-39.27	-29.01	94.95	-40.43	-29.86	149.58	14.21	10.5						
Kalimpong	4.75	0.31	-4.44	-93.46	0.29	-4.47	-93.96	19.26	14.51	305.33						
Malda	137.61	100.89	-36.71	-26.68	95.08	-42.53	-30.91	179.38	41.78	30.36						
Murshidabad	216.29	223.85	7.55	3.49	216.11	-0.18	-0.09	244.75	28.45	13.15						
Nadia	94.64	95.94	1.3	1.37	91.18	-3.46	-3.66	96.04	1.4	1.48						
North 24-pargana	142.46	101.08	-41.38	-29.05	89.54	-52.92	-37.15	134	-8.46	-5.94						
South 24-pargana	328.57	301.75	-26.81	-8.16	277.65	-50.91	-15.5	378.57	50	15.22						
Barddhaman (East)	358.5	321.96	-36.54	-10.19	318.26	-40.24	-11.22	463.57	105.08	29.31						
Barddhaman (West)	42	43.26	1.26	3	42.09	0.09	0.21	76.51	34.51	82.17						
Medinipur (East)	249.14	203.4	-45.74	-18.36	184.46	-64.68	-25.96	263.3	14.16	5.68						
Medinipur (West)	360.69	301.59	-59.1	-16.39	298.13	-62.56	-17.35	430.37	69.68	19.32						
Purulia	281.24	149.92	-131.31	-46.69	147.84	-133.39	-47.43	352.05	70.81	25.18						
Total	4002.09	3085.57	-916.52	-22.9	2976.08	-1026.01	-25.64	4851.26	849.17	21.22						



**Figure 9.** The time series spectral indices shown in four selected spatial windows from sowing through the stage of maximum greenness time (MGT). In all these windows the difference of LSWI-EVI at the MGT stage remains positive, suggesting the paddy rice pixels are submerged under water.

the RF-based model, with reference to DES area statistics, suggests RF overestimates the paddy rice area. Such a higher degree of association is also attributed to the large training set with sufficient spatial variation in the rice varieties, including the hyper-tuning parameters optimized during the training stage. The estimated area from PPPM and IPPPM, on the other hand, remains largely underestimated with reference to DES statistics as the estimated slope value for both of them remains below 0.8 (Figure 8a, 8b). One of the reasons for such underestimation is linked to the condition in Equation (4) used to identify the paddy rice pixels. The diurnal cycle of standing water in the paddy rice fields is what makes the fields identifiable through the mathematical condition of PPPM. However, when the satellite passes over an area, the absence of standing water resulting in a negative difference of LSWI and EVI (Equation (4)) shortly after the transplantation stage. Such physical condition resulting in the noted underestimation in the paddy rice area. In conjunction with that, the plausible source of underestimation could be related to the cropland layer used in this study from Sentinel-derived LULC product. Moreover, the inconsistency in season-wise official reporting of the area under paddy rice by DES is contributed to the noted difference. The inclusion of the additional criterion during the MGT stage in the IPPPM approach, which led to the exclusion of the submerged pixels from the target cluster pixels of paddy rice, caused further loss of paddy rice pixels. Nevertheless, it holds a map of the paddy rice area under active cultivation. For PPPM and RF-based output, the traces of such active paddy rice pixels, given the occurrence of flood conditions, are less likely. Figure 10 shows an apparent difference of the mapped paddy rice between IPPPM and RF-model with reference to DES statistics. The comparison of the model outputs shows RF-based classification model performs better than IPPPM model though it needs a large volume of training data with adequate quality. The implication of IPPPM can overcome these constraints through a simple approach of conditional thresholding. A plethora of studies existed on the rice area mapping at different spatial scales in the last decade; nearly all of them used automated to semi-automated methods of machine learning (Nanshan et al. 2021; Chen et al. 2020; Gumma et al. 2020; Htitiou et al. 2019; Zhang et al. 2018), deep learning (Shao et al. 2001; Chen and McNairn



**Figure 10.** A comparative spatial plot of the outcomes from Random Forest model and IPPPM model.

2006), spectral matching (Gumma et al. 2014), pixel and object-based method (Belgiu and Csillik 2018) and unsupervised clustering algorithms (Panigrahy et al. 2010). The majority of these works – for rice area detection from remote sensing images – are band-dependent with a single or a cluster of images. The phenological advancement with changing moisture conditions was deeply undermined in these works. However, in 2005, Xiao et al. (2005b, Xiao et al. 2006) provided a strong foundation for the PPPM approach for paddy rice area detection. This simple mathematical algorithm worked fairly well for monsoon Asia as well as for the global paddy rice area mapping. Yet, it is not capable enough to map active paddy rice areas during the monsoon season in the regions where floods are an integral part of the agricultural system. The submergence of the cropped area due to monsoon rain and the periodic release of stored water in reservoirs causes a range of flood events that vary in severity. The loss of the paddy rice fields due to such events during the post-transplantation through the MGT stages was not considered in the PPPM approach so far. The inclusion of flood effect in the form of prescribed modification during the MGT stage within the PPPM is what makes IPPPM approach effective in mapping the active paddy rice area during the monsoon season; the problem of underestimation with reference to DES statistics remains unavoidable, though. Dong et al. (2016) showed exclusion of flooded pixels during the monsoon season improved the model's accuracy and produced more reliable results. Teluguntla et al. (2015) also applied the PPPM approach on 8-days MODIS-EVI and LSWI over the Krishna River basin (KRB) in south India to map the paddy rice area. Since there is a difference in monsoon rainfall characteristics between KRB and LGP, the prolific flood pixels are absent in the KRB, whereas they are abundant in the LGP. Therefore, the suggested modification is pertinent for near real-time mapping of the active paddy rice area in this part of the LGP. With the availability of the GEE cloud-based computing facility that allows planetary-scale analysis through parallel processing in multiple GPUs, mapping of the near real-time rice field,

especially during the monsoon season, through the IPPPM approach provides up-to-date areal coverage of paddy rice.

## 5. Conclusion

The study fosters the strength of the high-resolution Sentinel – 2 time-series images with five days of repetitive coverage to mapping a high-resolution paddy rice area during the monsoon season at a 10 m scale. Since monsoon in the LGP is featured with frequent flooding, mapping paddy rice area during this season from space-based optical images needs special effort to enhance the accuracy for area estimation. The present study addresses this specific issue by bringing modifications to the existing PPPM approach. The study also deals with the problem of cloud cover that is persistent over the study area during monsoon. After atmospheric correction and screening off the cloud from the scenes, the residual effect was minimized through linear smoothing by selecting the best fit time window. Such smoothing helped to restore the temporal signal of vegetation indices (i.e., EVI and LSWI) which, otherwise, would have lost their identity as crop pixels. The results suggest a wide variation of spatial-scale accuracy from north to the south of the study area with reference to the statistical data. Yet, the validation of the model's outcome – in terms of the point-based accuracy - remains fairly similar to the RF-based result. Based on the outcome, we conclude that there is topographic control that led to the higher amount of offsets with respect to statistical data for territories belong to mountainous and residual-hill regions. The findings also led to the conclusion that the IPPPM approach estimates a lesser amount of paddy rice area in comparison to the conventional PPPM approach due to the exclusion of flooded fields during the MGT stage, though the principle of identification of paddy rice remains the same for both the approaches.

With prolific flood events in the monsoon period, the proposed IPPPM approach - as manifested in the results - is an appropriate method in mapping active fields under paddy rice cultivation in the LGP. For periodic active rain-spells of the south-west monsoon wind, the areas, elsewhere in the world, featuring with poor drainage condition can face similar waterlogging condition for the paddy fields. Therefore, the present experimental setup could be transferred to other areas, preferably to the rice growing areas in the south and south-east Asia where monsoon rainfall plays a vital role for paddy rice cultivation. The findings of this study also suggest, despite having a quality of underestimation in terms of areal coverage, the outcome of the IPPPM method is not significantly differed from other standard approaches, and such an evidence indicates towards the efficacy of this modified approach in mapping active paddy rice at different spatial scale. The GEE-based cloud computing facility, in this connection, provides an opportunity for near real-time mapping of the active paddy rice area. Therefore, the periodic monitoring of the active paddy rice area through the IPPPM approach has a direct linkage to the total production and yield variation over the years while addressing the issues of food security and availability and price of the rice in the market at the same time.

## Acknowledgements

The first author is thankful to the University Grants Commission, Government of India, for financial support (UGC Ref. No. 3265(NET-JULY2016). We acknowledge the Department of Science and Technology, Government of India, for providing funds (SR/FST/ES-I/2017/7/2017) to the Department of Geography, Vidyasagar University. Qi Zhang's collaboration was partially supported by a Microsoft AI for Earth Compute Grant and an American Association of Geographers (AAG) Research Grant.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was work supported by Microsoft Research and University Grants Commission.

## ORCID

Somnath Bar  <http://orcid.org/0000-0003-1679-6130>

Surajit Ghosh  <http://orcid.org/0000-0002-3928-2135>

## Data availability statement

The authors confirm that the data used for preparing this manuscript will be made available upon request.

## Code availability

The codes used for the analysis in this paper can be found in the following link: <https://github.com/ara-binda90/S2-Rice-area-map>

## References

- Akbari V, Simpson M, Maharaj S, Marino A, Bhowmik D, Prabhu GN, Rupavatharam S, Datta A, Kleczkowski A, Sujeetha JARP. 2021. Monitoring aquatic weeds in Indian wetlands using multitemporal remote sensing data with machine learning techniques. In: 2021 IEEE Int Geosci Remote Sens Symp IGARSS; p. 6847–6850. <https://doi.org/10.1109/IGARSS47720.2021.9553207>
- Attri SD, Tyagi A. 2010. Climate profile of India. India Meteorol Dep. :1–129.
- Bargiel D. 2017. Remote Sensing of Environment a new method for crop classification combining time series of radar images and crop phenology information. *Remote Sens Environ.* 198:369–383. <http://dx.doi.org/10.1016/j.rse.2017.06.022>.
- Belgiu M, Csillik O. 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sens Environ.* 204:509–523.
- Breiman L. 2001. Random Forests. *Machine Learn.* 45(1):5–32.
- Chen C, Mcnairn H. 2006. A neural network integrated approach for rice crop monitoring. *Int J Remote Sens.* 27(7):1367–1393. <http://dx.doi.org/10.1080/01431160500421507>.
- Chen N, Yu L, Zhang X, Shen Y, Zeng L, Hu Q, Niyogi D. 2020. Mapping Paddy rice fields by combining Multi-Temporal Vegetation Index and synthetic aperture radar remote sensing data using Google Earth engine machine learning platform. *Remote Sens.* 12(18):2992. <https://www.mdpi.com/2072-4292/12/18/2992/htm>.
- Ding M, Guan Q, Li L, Zhang H, Liu C, Zhang L. 2020. Phenology-based rice paddy mapping using multi-source satellite imagery and a fusion algorithm applied to the Poyang lake plain, Southern China. *Remote Sensing.* 12(6):1022.
- Dong J, Xiao X, Menarguez MA, Zhang G, Qin Y, Thau D, Biradar C, Moore B. 2016. Mapping paddy rice planting area in northeastern Asia with Landsat 8 images, phenology-based algorithm and Google Earth Engine. *Remote Sens Environ.* 185:142–154.
- Elert E. 2014. Rice by the numbers: a good grain. *Nature.* 514(7524):S50–S51.
- FAO. 2018. Rice market monitor. *Food Agric. Organ. United States* 21(1):1–38.
- Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, Husak G, Rowland J, Harrison L, Hoell A, et al. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci Data.* 2(1):1–21. <https://www.nature.com/articles/sdata201566>.
- Gallego FJ. 2004. Remote sensing and land cover area estimation. *Int J Remote Sens.* 25(15):3019–3047.

- Gao BC. 1996. NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens Environ.* 58(3):257–266.
- Gong P, Liu H, Zhang M, Li C, Wang J, Huang H, Clinton N, Ji L, Li W, Bai Y, et al. 2019. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Sci Bull.* 64(6):370–373.
- Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. 2017. Google Earth Engine : planetary-scale geospatial analysis for everyone. *Remote Sens Environ.* 202:18–27.
- Gumma MK, Thenkabail PS, Maunahan A, Islam S, Nelson A. 2014. Mapping seasonal rice cropland extent and area in the high cropping intensity environment of Bangladesh using MODIS 500m data for the year 2010. *ISPRS J Photogramm Remote Sens.* 91:98–113.
- Gumma MK, Thenkabail PS, Teluguntla PG, Oliphant A, Xiong J, Giri C, Pyla V, Dixit S, Whitbread AM. 2020. Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud. *GISci Remote Sens.* 57(3):302–322. <https://doi.org/10.1080/1548160320191690780>.
- Hao P, Zhan Y, Wang L, Niu Z, Shakir M. 2015. Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA. *Remote Sens.* 7(5):5347–5369.
- Hijjoka Y, Lin E, Pereira JJ, Corlett RT, Cui X, Insarov GE, Lasco RD, Lindgren E, Surjan A. 2014. Asia. In: *Climate Change 2014: impacts, adaptation, and vulnerability. Part B: regional aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Htitiou A, Boudhar A, Lebrini Y, Hadria R, Lionboui H, Elmansouri L, Tychon B, Benabdelouahab T. 2019. The performance of random forest classification based on phenological metrics derived from Sentinel-2 and Landsat 8 to map crop cover in an irrigated semi-arid region. *Remote Sens Earth Syst Sci.* 2019 24. 2(4):208–224. <https://link.springer.com/article/10.1007/s41976-019-00023-9>.
- Huete A, Didan K, Miura T, Rodriguez EP, Gao X, Ferreira LG. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens Environ.* 83(1-2):195–213.
- Jensen JR. 2014. *Remote sensing of the environment: an earth resource perspective.* 2nd ed. Harlow: Pearson.
- Jha VC, Bairagya H. 2013. Flood and flood plains of West Bengal, India: a comparative analysis. *Rev Georagauia.* :1–10. <https://periodicoscientificos.ufmt.br/ojs/index.php/geo/article/view/4835>
- Jin Q, Wang C. 2017. A revival of Indian summer monsoon rainfall since 2002. *Nat Clim Change.* 7(8): 587–595..
- Khanal N, Matin MA, Uddin K, Poortinga A. 2020. A comparison of three temporal smoothing algorithms to improve land cover classification: A case study from Nepal. *Remote Sens.* 12(18):5–7.
- Lehane S. 2014. India 's food and water security. *Future Dir Int.* :1–10. <https://apo.org.au/node/39957>
- Li Q, Wang C, Zhang B, Lu L. 2015. Object-based crop classification with Landsat-MODIS enhanced time-series data. *Remote Sens.* 7(12):16091–16107.
- Liu L, Xiao X, Qin Y, Wang J, Xu X, Hu Y, Qiao Z. 2020. Remote Sensing of Environment Mapping cropping intensity in China using time series Landsat and Sentinel- 2 images and Google Earth Engine. *Remote Sens Environ.* 239(June 2019):111624..
- Löw F, Biradar C, Dubovyk O, Fliemann E, Akramkhanov A, Narvaez Vallejo A, Waldner F. 2018. Regional-scale monitoring of cropland intensity and productivity with multi-source satellite image time series. *GISci Remote Sens.* 55(4):539–567.
- Mendelsohn R. 2014. The impact of climate change on agriculture in Asia. *J Integr Agric.* 13(4):660–665.
- Miyan MA. 2015. Droughts in Asian least developed countries: vulnerability and sustainability. *Weather Clim Extrem.* 7:8–23.
- Nanshan Y, Jinwei D, Jianxi H, Guoming D, Zh G, He Y, Tong Y, Yuanyuan D, Xiangming X. 2021. The 10-m crop type maps in Northeast China during 2017 – 2019. *Sci Data.* 8(1):41–12.
- Nguyen TTH, de Bie CAJM, Ali A, Smaling EMA, Chu TH. 2012. Mapping the irrigated rice cropping patterns of the Mekong delta, Vietnam, through hyper-temporal spot NDVI image analysis. *Int J Remote Sens.* 33(2):415–434.
- Oguro Y, Suga Y, Takeuchi S, Ogawa M, Konishi T, Tsuchiya K. 2001. Comparison of SAR and optical sensor data for monitoring of rice plant around Hiroshima. *Adv Sp Res.* 28(1):195–200.
- Pan XZ, Uchida S, Liang Y, Hirano A, Sun B. 2010. Discriminating different landuse types by using multitemporal NDXI in a rice planting area. *Int J Remote Sens.* 31(3):585–596.
- Panigrahy S, Upadhyay G, Ray SS, Parihar JS. 2010. Mapping of Cropping System for the Indo-Gangetic Plain Using Multi-Date SPOT NDVI-VGT Data. *J Indian Soc Remote Sens.* 38(4):627–632.
- Pekel JF, Cottam A, Gorelick N, Belward AS. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature.* 540(7633):418–422. <http://dx.doi.org/10.1038/nature20584>.

- Pelletier C, Valero S, Inglada J, Champion N, Dedieu G. 2016. Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sens Environ.* 187:156–168. <http://dx.doi.org/10.1016/j.rse.2016.10.010>.
- Peng D, Huete AR, Huang J, Wang F, Sun H. 2011. Detection and estimation of mixed paddy rice cropping patterns with MODIS data. *Int J Appl Earth Obs Geoinf.* 13(1):13–23.
- Prashnani M, Singh DK, Joshi R, Ray SS, Neetu. 2014. Understanding crop growing pattern in bardhaman district of West Bengal using multi-date RISAT 1 MRS data. *Int Arch Photogramm Remote Sens Spatial Inf Sci.* XL-8 (8):861–864.
- Qadir A, Mondal P. 2020. Synergistic use of radar and optical satellite data for improved monsoon cropland mapping in India. *Remote Sens.* 12(3):522.
- Rudiyanto M, S, Soh A, S. 2019. Automated near-real-time mapping and monitoring of rice extent, cropping patterns, and growth stages in Southeast Asia Using Sentinel-1 Time Series on a Google Earth engine platform. *Remote Sens.* 11(14):1666. <https://www.mdpi.com/2072-4292/11/14/1666>.
- Schmidt M, Witte C. 2010. Monitoring aquatic weeds in a river system using SPOT 5 satellite imagery. *J Appl Remote Sens.* 4(1):1–17.
- Shao Y, Fan X, Liu H, Xiao J, Ross S, Brisco B, Brown R, Staples G. 2001. Rice monitoring and production estimation using multitemporal RADARSAT. *Remote Sens Environ.* 76(3):310–325.
- Shew AM, Ghosh A. 2019. Identifying dry-season rice-planting patterns in Bangladesh using the landsat archive. *Remote Sens.* 11(10):1235.
- Silva TSF, Costa MPF, Melack JM, Novo EMLM. 2008. Remote sensing of aquatic vegetation: Theory and applications. *Environ Monit Assess.* 140(1-3):131–145.
- Singh LP, Parkash B, Singhi AK. 1998. Evolution of the loewr gangetic plain landforms and soils in West Bengal. *India. Catena.* 33(2):75–104.
- Singh G, Reynolds C, Byrne M, Rosman B. 2020. A remote sensing method to monitor water, aquatic vegetation, and invasive water hyacinth at national extents. *Remote Sens.* 12(24):4021–4024.
- Sun H, Huang J, Huete A, Peng D, Zhang F. 2009. Mapping paddy rice with multi-date moderate-resolution imaging spectroradiometer (MODIS) data in China. *J Zhejiang Univ Sci A.* 10(10):1509–1522.
- Teluguntla P, Ryu D, George B, Walker JP, Malano HM. 2015. Mapping flooded rice paddies using time series of MODIS imagery in the Krishna River Basin, India. *Remote Sens.* 7(7):8858–8882.
- Wang S, Azzari G, Lobell DB. 2019. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. *Remote Sens Environ.* 222(May 2018):303–317..
- Wang J, Xiao X, Qin Y, Dong J, Zhang G, Kou W. 2015. Mapping paddy rice planting area in wheat-rice double-cropped areas through integration of Landsat-8 OLI, MODIS, and PALSAR images. *Sci Rep.* 5(1):1–11. <http://dx.doi.org/10.1038/srep10088>.
- Wu Q. 2020. geemap: A Python package for interactive mapping with Google Earth Engine. *JOSS.* 5(51): 2305.
- Xiao X, Boles S, Frolking S, Li C, Babu JY, Salas W, Moore B. 2006. Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images. *Remote Sens Environ.* 100(1):95–113.
- Xiao X, Boles S, Liu J, Zhuang D, Frolking S, Li C, Salas W, Moore B. 2005. Mapping paddy rice agriculture in southern China using multi-temporal MODIS images. *Remote Sens Environ.* 95(4):480–492.
- Xiao X, Boles S, Liu J, Zhuang D, Frolking S, Li C, Salas W, Moore B. 2005. Mapping paddy rice agriculture in southern China using multi-temporal MODIS images. *Remote Sens Environ.* 95(4):480–492.
- Xu X, Ji X, Jiang J, Yao X, Tian Y, Zhu Y, Cao W, Cao Q, Yang H, Shi Z, et al. 2018. Evaluation of one-class support vector classification for mapping the paddy rice planting area in Jiangsu Province of China from Landsat 8 OLI imagery. *Remote Sens.* 10(4):546.
- Yin Q, Liu M, Cheng J, Ke Y, Chen X. 2019. Mapping paddy rice planting area in northeastern China using spatiotemporal data fusion and phenology-based method. *Remote Sens.* 11(14):1699.
- Yin F, Lewis PE, Gomez-Dans J, Wu Q. 2019. A sensor-invariant atmospheric correction method: application to Sentinel-2/MSI and Landsat 8/OLI. *EarthArXiv.* <https://doi.org/10.31223/osf.io/ps957>
- Zhang M, Lin H, Wang G, Sun H, Fu J. 2018. Mapping paddy rice using a convolutional neural network (CNN) with Landsat 8 Datasets in the Dongting Lake Area, China. *Remote Sens.* 2018, Vol 10, Page 1840. 10(11):1840. <https://www.mdpi.com/2072-4292/10/11/1840/htm>.
- Zhang G, Xiao X, Dong J, Kou W, Jin C, Qin Y, Zhou Y, Wang J, Menarguez MA, Biradar C. 2015. Mapping paddy rice planting areas through time series analysis of MODIS land surface temperature and vegetation index data. *ISPRS J Photogramm Remote Sens.* 106:157–171.
- Zhang G, Xiao X, Dong J, Xin F, Zhang Y, Qin Y, Doughty RB, Moore B. 2020. Fingerprint of rice paddies in spatial-temporal dynamics of atmospheric methane concentration in monsoon Asia. *Nat Commun.* 11(1):554. <http://dx.doi.org/10.1038/s41467-019-14155-5>.

- Zhang G, Xiao X, Biradar CM, Dong J, Qin Y, Menarguez MA, Zhou Y, Zhang Y, Jin C, Wang J, et al. 2017. Science of the Total Environment Spatiotemporal patterns of paddy rice croplands in China and India from 2000 to 2015. *Sci Total Environ.* 579(1):82–92. <http://dx.doi.org/10.1016/j.scitotenv.2016.10.223>.
- Zhen Z, Quackenbush LJ, Stehman SV, Zhang L. 2013. Impact of training and validation sample selection on classification accuracy and accuracy assessment when using reference polygons in object-based classification. *Int J Remote Sens.* 34(19):6914–6930.
- Zhou F, Zhang A, Townley-Smith L. 2013. A data mining approach for evaluation of optimal time-series of MODIS data for land cover mapping at a regional level. *ISPRS J Photogramm Remote Sens.* 84: 114–129.