https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Earth System Science Data
Open Access Discussions

AsiaRiceYield4km: Seasonal Rice Yield in Asia from 1995 to 2015

Huaqing Wu[1,2★], Jing Zhang[1,2★], Zhao Zhang[1,2], Jichong Han[1,2], Juan Cao[1,2], Liangliang Zhang[1,2], Yuchuan Luo[1,2], Qinghang Mei[1,2], Jialu Xu[2], Fulu Tao[3,4]

[1]Key Laboratory of Environmental Change and Natural Disasters, Ministry of Education Beijing Normal University, Beijing 100875, People's Republic of China

[2]School of National Safety and Emergency Management, Beijing Normal University, Beijing 100875 / Zhuhai 519087, People's Republic of China

[3]Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, People's Republic of China

[4]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

★These authors contributed to the work equally and should be regarded as co-first authors.

Correspondence to: Zhao Zhang (zhangzhao@bnu.edu.cn)

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Earth System Science Data
Open Access Discussions

Abstract. Rice is the most important staple food in Asia. However, high-spatiotemporal-resolution rice yield datasets are very limited over a large region. The lack of such products hugely hinders the studies on accurately assessing the impacts of climate change and simulating agricultural production. Based on dynamic rice maps in Asia, we incorporated four predictor categories into three machine learning (ML) models to generate a high-spatial-resolution (4km) rice yield dataset (AsiaRiceYield4km) for main rice seasons from 1995 to 2015. Four predictor categories considered the most comprehensive rice growing conditions and the optimal ML model was determined for each rice season based on an inverse proportional weight method. The results showed that AsiaRiceYield4km has a good accuracy for seasonal rice yield prediction (single rice: R^2 = 0.88, *RMSE* = 920 kg/ha, double rice: R^2 = 0.91, *RMSE* = 554 kg/ha, and triple rice: R^2 = 0.93, *RMSE* = 588 kg/ha). Compared with Spatial Production Allocation Model (SPAM), R^2 of grided rice yields was improved by 0.20 and *RMSE* was reduced by 618 kg/ha on average for single rice. Particularly, constant environmental conditions including longitude, latitude, elevation, and soil properties contributed the most (~45%) to rice yield prediction. As for different growing periods of rice, we found that the predictors in reproductive period had more impacts on rice yield prediction than those of the vegetative period and the whole growing period. AsiaRiceYield4km is a novel high-spatial-resolution gridded rice yield dataset that can fill the unavailability of seasonal yield products across major rice production areas and promote more relevant studies on agricultural sustainability in the world. AsiaRiceYield4km can be downloaded from an open-data repository (DOI: https://doi.org/10.5281/zenodo.6901968; Wu et al., 2022).

1 Introduction

As one major staple crop, rice (*Oryza sativa* L.) provides more than a quarter of calories for about half of the population with only 11% of the arable land on the earth (Maclean et al., 2002; Alexandratos and Bruinsma, 2012; Birla et al., 2017; Qian et al., 2020). Especially, Asia produces and consumes more than 90% of the global rice (Bandumula, 2018), where dominated by poor smallholder farmers. Therefore, information on rice yield in Asia is essentially important to sustain food security and farmers' livelihood (Laborte et al., 2017). In the last half-century, the increased yield has contributed most largely to rice production relative to planting areas (Blomqvist et al., 2020) and will still be a hot spot considering the unprecedented increase of the population and environmental pressure in the future (Kim et al., 2021).

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Besides, Asia rice contains complex cropping system including three rice seasons: single, double and triple (Zhang et al., 2020a). Therefore, it is critically necessary to develop a long-term and explicitly spatiotemporal Asia rice yield dataset to monitor and guide agriculture production.

Previous global-scale crop yield datasets including Harvester Area and Yields of 175 crops (M3Crops) (Monfreda et al., 2008), Spatial Production Allocation Model (SPAM) (You and Wood, 2006; Yu et al., 2020), Global Dataset of Historical Yields of Major Crops (GDHY) (Iizumi et al., 2014; Iizumi and Sakai, 2020), and Global Gridded Crop Model Intercomparison (GGCMI) phase 1 (Müller et al., 2019), have been produced and widely used in many studies (Folberth et al., 2020; Kaltenegger and Winiwarter, 2020; Iizumi et al., 2021; Lin et al., 2021; Liu et al., 2021b). However, due to the different research goals and technical restrictions, their spatial resolutions are relatively coarser (~10km for M3Crops and SPAM; ~55km for GDHY and GGCMI phase 1) and temporal resolutions are mostly annual (Laborte et al., 2017). Fewer datasets are seasonal resolution and cannot cover all rice seasons (Kim et al., 2021). Besides, the time spans are limited (only one year for M3Crops; every five years for SPAM). Although GDHY provides long-term (1981-2016) rice yields, we cannot obtain the interannually spatial dynamics of rice yield because its rice area basemap was fixed for all years only referring M3Crops (around 2000). To the best of our knowledge, a long-term seasonal rice yield dataset with higher spatial resolution and dynamically spatial distribution is currently unavailable for major rice planting regions in the world.

To address the above issues, there is a crucial need to acquire wiser technologies and multi-sources data for rice yield prediction (Chlingaryan et al., 2018; Cao et al., 2020; van Klompenburg et al., 2020; Zhang et al., 2020b; Chen et al., 2022). With the rapid development in remote sensing technology in these years, large-scale and long-term high-spatiotemporal observations could provide ample and timely phenological and growing information about rice seasons. Meanwhile, ground-based data such as climate and soil also provided more key environmental information (Folberth et al., 2016; Zhang et al., 2021). Many publications have expanded our knowledge that combining satellite-derived data and ground environmental information could successfully monitor crop growing states and predict final yields (Huang et al., 2013; Mosleh et al., 2015; Cao et al., 2021; Fernandez-Beltran et al., 2021). On the other side, annual paddy rice areas were mapped (Nelson and Gumma, 2015; Han et al., 2022), but have yet been applied to yield prediction. Thus, integrating multi-sources data can produce seasonal rice yield

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

products with the dynamically spatial distribution and long-term coverage which will provide more spatiotemporal information and consequently will greatly benefit the researchers in the related fields.

Moreover, machine learning (ML) has been increasingly and successfully used in crop yield predictions, such as random forest (RF), extreme gradient boosting (XGBoost) and long short-term memory (LSTM) (Cai et al., 2019; van Klompenburg et al., 2020; Sakamoto, 2020; Luo et al., 2022). Such ML models can overcome the drawbacks of the two traditional methods: process-based crop models and statistical regression methods. Compared with process-based crop models, ML can wisely select input variables according to actual requirements and local geographical environment conditions without inputting complicated parameters (Jeong et al., 2022). Meanwhile, ML is superior to statistical regression methods through solving non-linear problems with higher efficiency and flexibility by complex functions (Chlingaryan et al., 2018). Besides, ML has a good spatial generalization with high computational efficiency. Therefore, ML models combined with multi-sources data potentially provide a good chance for large-scale gridded yield production and their estimates improvement.

Overall, we would propose an explicit method to predict rice yield at large scale based on ML methods integrating multi-sources data. Based on this method, a seasonal 4km resolution rice yield dataset across Asia (AsiaRiceYield4km) from 1995 to 2015 was generated, which has the annually dynamic rice area basemaps. AsiaRiceYield4km will fill the dataset blank, and better support agricultural monitoring systems and the related researches over large scale because of its higher-spatiotemporal resolutions and long-time span (Wu et al., 2022).

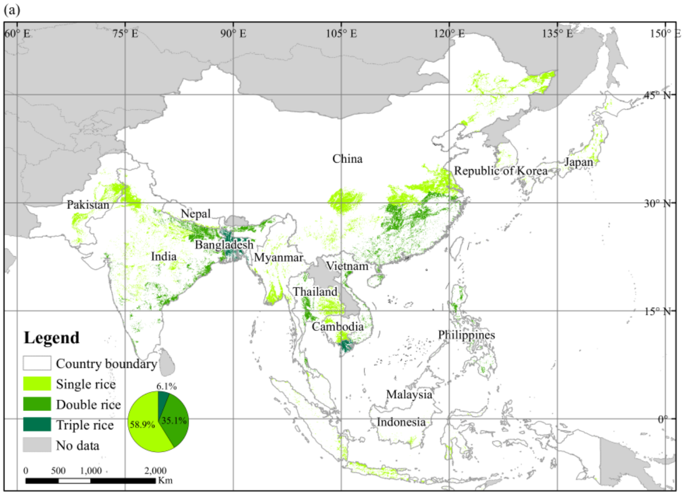
2 Materials and methods

2.1 Study area

Asia is the most important rice-producing area due to its 89% planting area and 91% production in the world (Food and Agriculture Organization of the United Nations, FAOSTAT, 2022). Considering the accessibility of locally census-based rice yield data, 14 main rice-producing countries of Asia were selected and then divided into 27 seasons based on different rice cropping system (single, double, and triple rice), shown in Fig. 1 (see Sect. 2.2.2 below for details).

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

(a)



(b)

Country	Season number	Rice systems
Cambodia	1	Single
China	2	
India	3	
Indonesia	4	
Japan	5	
Malaysia	6	
Myanmar	7	
Nepal	8	
Pakistan	9	
Republic of Korea	10	
Thailand	11	
China	12	Double (Early, Late)
	13	
India	14	
	15	
Philippines	16	
	17	
Thailand	18	
	19	
Vietnam	20	
	21	
Bangladesh	22	Triple (Spring, Autumn, Winter)
	23	
	24	
Vietnam	25	
	26	
	27	

Figure 1: (a) Rice planting areas with different growing seasons in main rice-producing countries of Asia. The green area represents the maximum paddy rice area where paddy rice grew at least for one year during 1995-2015 (Han et al., 2021, 2022). The pie chart represents the area proportion of different rice systems. (b) The season numbers and rice system for each country. Double rice follows the order of early before late (i.e., 12 and 13 represent the early and late season rice of Philippines, respectively) and triple rice follows the order of spring, autumn, and winter (i.e., 25, 26, and 27 represent the spring, autumn and winter season rice of Vietnam, respectively).

2.2 Data

Multi-sources data are collected and used to predict rice yield, including: annual rice area maps, over 45000 yield records at 1400 administrative units, LAI information based on remote sensing, and rice growing environmental conditions including location, time, soil, and climate. Besides, considering the necessity of phenological information, we also produced the gridded key phenological dates according to the LAI data. With exception of yield records from official statistics (Table S1), the other data are gridded (originally information is listed in Table S2) to 4km×4km using the nearest neighbor resampling method in ArcMap 10.2.

2.2.1 Rice area maps

We used the latest public rice distribution maps dataset, an annual dataset (2000 to 2020) of paddy rice area at 500m resolution (APRA500), in this study (Han et al., 2021, 2022). Due to the topography conditions, cloud contamination, and the mixed-pixel effects with fragmented cropland fields, however, APRA500 was somehow underestimated (Han et al., 2022). To reduce this effect, here, we used the rice

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

area union of the three years (current year, last year, and next year) as the rice area of the current year (i.e., the area of 2005 is the union of 2004, 2005 and 2006). Specifically, the union area of 2000, 2001, and 2002 was also applied to the years before 2001 because of the unavailable rice maps.

2.2.2 Seasonal rice yield

RiceAtlas (Laborte et al., 2017), the most comprehensive rice calendar, was widely used in many studies (van Oort and Zwart, 2018; Muehe et al., 2019; Fritz et al., 2019). We determined the rice seasons according to RiceAtlas and the national statistics of each country. The rice seasons have various names in different countries, such as Aman, Aus, and Boro for triple rice of Bangladesh, Rabi, and Kharif for double rice of India. To make them more readable and consistent, we used single rice, double rice (early and late seasons), and triple rice (spring, autumn, and winter seasons) in our study shown in Fig.1 (b). A few rice-cropping seasons (e.g., early season in Cambodia, Malaysia, Myanmar, and Indonesia; and winter season in India) were not considered due to the lack of yield records.

We collected the seasonal rice yield data from FAO and other government websites (Table S1), with around 1400 administrative units from 1995 to 2015. The quality of these data has been checked and some yield outliers were filtered out according to the following rules: (a) exceeding the actual biophysically attainable yields; (b) beyond the averages ± two times variance during 1995-2015 (Zhang et al., 2014; Cao et al., 2020, 2021).

2.2.3 Key phenological dates

Planting, heading, and harvesting dates are three most important phenological dates during rice growing. The whole growing period (WGP) can be divided into two periods according to the three key phenological dates: vegetative period (VEP, planting date to heading date) and reproductive period (REP, heading date to harvesting date).

However, most rice phenology datasets are always at administrative scales without interannual variation, for example, the United States Department of Agriculture (USDA, https://ipad.fas.usda.gov/ogamaps/cropcalendar.aspx, last accessed: 7 April 2022) provided constant and country-scale growing phenological information; RiceAtlas had subnational phenology information but also ignored the annual dynamics (Laborte et al., 2017). Besides, these datasets lack heading dates information of rice. Here, we retrieved the dynamic three key rice phenological dates from remote

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

sensing data in Asia during 1995-2015 at 4km×4km grid scale by inflection-based and threshold-based methods (see Sect. 2.3.1 below for details). The datasets of USDA and RiceAtlas provided a threshold range for phenology and were used to validate our extracted phenological dates.

2.2.4 Location and time

Location information includes longitude (*Lon*), latitude (*Lat*), and elevation (*Ele*). The Global 30-arc-second (1km) gridded Digital Elevation Model (DEM) dataset (1999) from National Oceanic and Atmospheric Administration (NOAA) was used in this study. And the Lon and Lat information were collected from the centroid of each resampled 4km pixel by ArcMap 10.2. The temporal information is represented by the year (1995-2015).

2.2.5 Soil data

Soil properties are important factors controlling rice growing and final yield. Harmonized World Soil Database (HWSD) v1.2 provides key soil properties variables, including Topsoil Sand Fraction (*T_Sand*), Topsoil Silt Fraction (*T_SILT*), Topsoil Clay Fraction (*T_CLAY*), Topsoil Reference Bulk Density, (*T_BULK_DEN*), Topsoil Organic Carbon (*T_OC*), Topsoil pH (H2O) (*T_PH_H2O*) (https://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/, last accessed: 7 April 2022; Wieder et al., 2014).

2.2.6 Climate data

TerraClimate (Abatzoglou et al., 2018), a monthly high spatial resolution (4km) meteorological dataset (http://doi.org/10.7923/G43J3B0R, last accessed: 7 April 2022) from 1995 to 2015, is used in our study. This dataset can provide climate and water balance information for Asian rice (Salvacion, 2022). Climate variables include Palmer Drought Severity Index (PDSI), precipitation accumulated (Pre), downward surface shortwave radiation (Srad), maximum temperature (Tmax), minimum temperature (Tmin), vapor pressure (Vap), and wind speed (Ws).

2.2.7 LAI

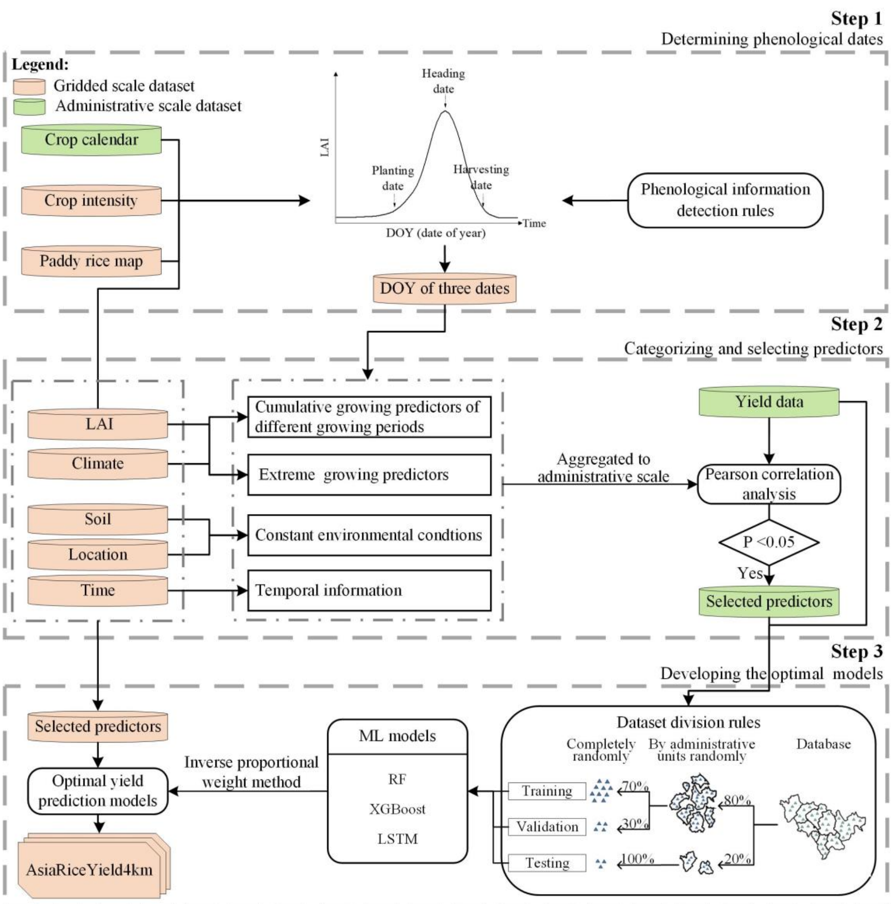
Remote sensing indices have been wildly used in rice yield prediction (Son et al., 2020; Arumugam et al., 2021), but few had been conducted before 2000 (Liu et al., 2021a). To extend the period of gridded

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

yield dataset from 1995 in this study, we adopted the Advanced Very-High-Resolution Radiometer satellite data (AVHRR). LAI can indicate the natural variation of rice phenology and growing status which is significant for yield prediction (Fang et al., 2011; Jin et al., 2013). Thus, Global Land Surface Satellite (GLASS) AVHRR LAI data (http://glass.umd.edu/Download.html, last accessed: 7 April 2022; Xiao et al., 2013, 2016) were used for rice phenological information extraction and yield prediction.

2.3 Methods

We applied three steps to generate AsiaRiceYield4km through incorporating multisource data into three ML method, in which three steps were: determining phenological dates, categorizing and selecting predictors, and developing the optimal models (Fig. 2). Details of each step were in the following sections.



https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Figure 2: Flowchart for generating long-term and high - resolution gridded rice yields by incorporating multisource data into ML models.

2.3.1 Determining phenological dates

Inflection-based method (Chen et al., 2016; Luo et al., 2020) and threshold-based method (Manfron et al., 2017) were used to detect rice phenological dates according to the following (step1 in Fig.2): (1) Planting dates: LAI always keeps a low value for a period of time before the planting date, and increases dramatically after this date (Sakamoto et al., 2005; Chen et al., 2018). Therefore, if there is one point in LAI curve where the first derivative is > 0 after it or its second derivative equals 0, this point is defined as the planting date. (2) Heading dates: the inflection point from VEP to REP (Wang et al., 2018) is characterized by the maximum value of LAI between the planting date and harvesting date (Son et al., 2013). (3) Harvesting dates: the physiological activity will sharply drop during the harvesting period. Therefore, the first inflection point at LAI curves where its first derivative becomes negative is considered as harvesting date. If the phenological dates in some grids cannot be detected by the above rules, the averages of the administrative units where the grids located in are used.

2.3.2 Categorizing and selecting predictors

To provide the comprehensive rice growing information for the ML models, we divided the multisource data into four categories including 50 predictors (Table S3): cumulative growing predictors of different growing periods (CGP), extreme growing predictors (EGP), constant environmental conditions (CEC), and temporal information (TI) (Fig.2 step 2). CGP reflects the continuous growing period for rice and EGP can reflect the impact of extreme events such as drought on rice growth. CGP includes the sum of each climate and LAI variable in different growing periods (VEP, REP, and WGP). EGP consists of the maximum and minimum of each climate and LAI variable. CEC reflects the influence of geographical environment on rice growth. TI can reflect the long-term agronomic technology improvements.

High-dimension predictors often affect the accuracy and computational efficiency of ML methods (LeCun et al., 2015; Zhang et al., 2019). To reduce this effect, Pearson correlation analysis is used to select variables with a significant correlation with yield ($p < 0.05$) (Cao et al., 2021) at each administrative unit. Specifically, the four predictors, *Lon*, *Lat*, *Ele* and *Year*, were considered to have a stable impact on rice yield which were included in all predicted models (Huntington et al., 2020).

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

According to the Sect. 2.2.3, crop growing periods include WGP, VEP, and REP. Considering the covariate-relation between WGP and the rest two periods, the predictor in WGP would be selected if its Pearson *R* was higher than that in the rest two periods, or vice versa.

2.3.3 Developing the optimal models

(1) Dataset division rules

To effectively reducing overfitting effects (Dinh and Aires, 2022), we divided all data into three sets (training, validation, and testing) and use them to optimize ML parameters, select the optimal model, and evaluate its generalization ability, respectively (Ripley, 2007). Firstly, the whole databases were randomly divided into two subsets by the administrative unit (80% for training and validating, and 20% for testing). Then, the training and validation sets are randomly resplit into 70% for training and 30% for validation (Fig.2 step3). Thus, the training, validation, and testing sets contain 56%, 24%, and 20% of the whole dataset, respectively. Such division rules can avoid information leakage from the testing set to the training set (Meroni et al., 2021) and enhance the robustness of the model.

(2) ML models

ML can develop transfer functions based on the associations between predictors and target variables for rice yield prediction (Chlingaryan et al., 2018; Shahhosseini et al., 2020). Three widely used ML models, RF, XGBoost, and LSTM are selected for rice yield prediction. RF is based on the bagging ensemble model, which generates multiple decision trees and gets predictions by the vote of all individual trees (Breiman, 1996, 2001). Besides, extra randomness is introduced to RF at generating trees and searching for the best tree stages (Shahhosseini et al., 2020), which provides more diversity for trees and can generate the overall better prediction model (Zhang et al., 2019). XGBoost uses the optimized gradient boost for decision trees which tries to make weak learners to strong (Chen and Guestrin, 2016). This method adopts an updated strategy to train the predicted model and the updated model can minimize the loss by reducing errors from previous models (Obsie et al., 2020). LSTM is a special recurrent neural network (RNN) that is proposed to overcome the vanishing and exploding gradients problems of RNN (Hochreiter and Schmidhuber, 1997; Sak et al., 2014; Tian et al., 2021). LSTM contains input, hidden and output layers and the hidden layers consist of memory cells (He et al., 2019; Zhang et al., 2019). The hyper-parameters tuning details are shown in Supplementary Methods.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

(3) Model evaluation

The coefficient of determination (R^2) and root-mean-square error ($RMSE$) are adopted to evaluate the performance of each model for each cropping season (Figure 1b).

$$R^2 = 1 - \sum_{i=1}^{n} \left(Y_{ob\ i,j} - Y_{pred\ i,j} \right)^2 / \sum_{i=1}^{n} \left(Y_{ob\ i,j} - \bar{Y}_{ob\ i,j} \right)^2 \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(Y_{pred\ i,j} - Y_{ob\ i,j} \right)^2 / n} \quad (2)$$

where i is the number of the administrative unit, and n is the total number of the administrative units; j is the predicted and observed year. $Y_{ob\ i,j}$ is the observed rice yield from government or FAO websites in the ith administrative unit of j year, $\bar{Y}_{ob\ i,j}$ is the average of observed rice yield in the ith administrative unit of j year, and $Y_{pred\ i,j}$ is the predicted yield in the ith administrative unit of j year.

(4) Select the optimal yield prediction model

In this study, three ML models would generate three different yield prediction results. Previous studies recommend the weighted ensemble method by combining the prediction results of different methods, wishing for a relatively stable result but giving up somehow accuracy (Shahhosseini et al., 2020, 2021). Moreover, many studies also selected the optimal ML model through only comparing the accuracy of validation/testing sets (Zhang et al., 2021; Chen et al., 2022; Luo et al., 2022). Here, to conduct a comprehensive evaluation for different ML models and data sets, we developed an inverse proportional weight (IPW) method to assign weights for training, validation, and testing accuracy to calculate the adjusted accuracy for each ML model (Eq. 3-7). The ML model with the highest adjusted accuracy was selected as the optimal ML model.

$$w_{tr} = p_{tr} / (p_{tr} + p_{va} + p_{te}) \quad (3)$$

$$w_{va} = p_{va} / (p_{tr} + p_{va} + p_{te}) \quad (4)$$

$$w_{te} = p_{tr} / (p_{tr} + p_{va} + p_{te}) \quad (5)$$

$$R_{ad}^2 = R_{tr}^2 \cdot w_{tr} + R_{va}^2 \cdot w_{va} + R_{te}^2 \cdot w_{te} \quad (6)$$

$$RMSE_{ad} = RMSE_{tr} \cdot w_{tr} + RMSE_{va} \cdot w_{va} + RMSE_{te} \cdot w_{te} \quad (7)$$

where tr, va, and te are the abbreviation of training, validation, and testing; p_{tr}, p_{va}, and p_{te} are the inverse proportion for the size of training, validation, and testing sets, which equal 1/0.56, 1/0.24, and 1/0.20, respectively; w_{tr}, w_{va}, w_{te} are the weights of the training, validation, and testing sets. R^2_{ad} and $RMSE_{ad}$

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

represent the adjusted R^2 and *RMSE*. R^2_{tr}, $R^2_{va,}$ nd R^2_{te} are the R^2 of the training, validation, and testing sets; $RMSE_{tr}$, $RMSE_{va}$, and $RMSE_{te}$ are the *RMSE* of training, validation, and testing sets. The ML model with the highest R^2_{ad} and lowest $RMSE_{ad}$ is regarded as the optimal model for each season in Figure 1b.

3 Results

3.1 Performance of the predicted models

After selecting the optimal ML model for each season, we scattered the seasonal training, validation, testing, and adjusted modelling accuracy in Fig. 3. The training R^2 is higher than 0.9 for all seasons (indicated by the green color of all left dots), followed by validation and testing R^2 (average: 0.78 and 0.69). The R^2_{ad} ranged from 0.60 to 0.90 (average: 0.77), with lowest R^2_{ad} in single season of Malaysia and the highest in winter season of Bangladesh (Fig. 3c). As for *RMSE*, the average for training, validation, and testing are 105, 408, and 489 kg/ha, respectively. The $RMSE_{ad}$ ranges from 162 to 817 kg/ha and its average is 396 kg/ha. The highest $RMSE_{ad}$ is single rice of China (Fig. 3d), where rice yields are mostly higher than other countries and might cause more modelling uncertainties. As for double rice systems (Fig. 3b and 3e), there is no significantly statistical differences between their modelling accuracy, with around 0.77 of R_{ad}^2 and 410 kg/ha of $RMSE_{ad}$. Notably, the season with the highest accuracy is single rice in Mymmar (R_{ad}^2: 0.87, $RMSE_{ad}$: 162 kg/ha) and late season in Thailand (R_{ad}^2: 0.90, $RMSE_{ad}$: 259 kg/ha) for single rice and double rice, respectively. As for triple rice, winter season in Bangladesh has the highest R_{ad}^2 (0.90; No. 24 dot in Fig. 3c) and spring season in Vietnam has the lowest $RMSE_{ad}$ (327 kg/ha; No. 25 dot in Fig. 3c). Additionally, 27 optimal models consist of two types of ML models, XGBoost for 15 seasons and RF for 12 seasons, with no LSTM model.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

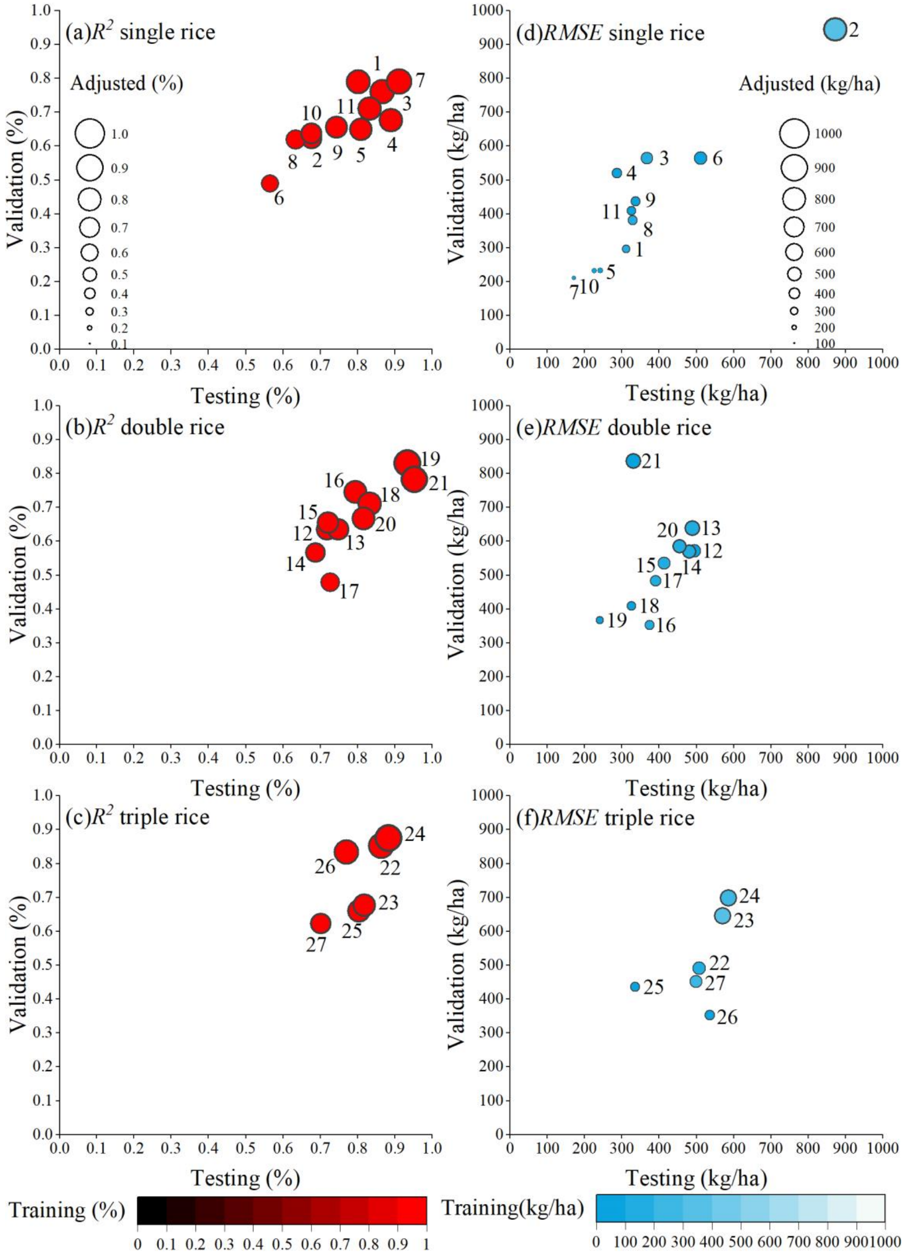


Figure 3: Accuracy (R^2 and *RMSE*) of the predicted yields for seasonal rice in each region. The R^2 (Figc) and *RMSE* (d-f) are present in the left and right panels, respectively. Single, double, and triple rice systems are sorted from top to bottom. The color of dots indicated different training accuracy ranks; testing accuracy at x axis; validation accuracy at y axis; and the size of dots represented the adjusted accuracy. Note: numbers for each dot represent the rice seasons for each region shown in Fig. 1(b) 1-11 represent the single rice; 12, 14, 16, 18 and 20 represent the early season for double rice; 13, 15, 17, 19 and 21 represent the late season for

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

double rice; 22 and 24 represent the spring season, 23 and 26 represent the autumn season and 24 and 27 represent the winter season for double rice.

3.2 The spatial characterizations of AsiaRiceYield4km

Based on the predicted seasonal yields from optimal ML models, we characterize the spatial patterns of averaged yields during 1995-2015. Single rice widely distributes in 11 countries across the whole area, where its yield varies greatly from about 400 to 10000 kg/ha with an average of 5428 kg/ha. Specially, the highest average yield is in China (7384 kg/ha) and the lowest is in India (1889 kg/ha). Such a huge difference might be ascribed to the better irrigation in China (Dawe et al., 2010), while relatively low-level soil fertility, investment, and technology in India (Srivastava and Mahapatra, 2012). Double rice mostly distributes between 30°N~0°, with insignificant differences between early and late yields (early rice ranging from 1041 to 8347, average 4598 kg/ha; late rice ranging from 666 to 7977 kg/ha, average 4539 kg/ha). The highest rice yield is indicated in the east of Asia, while the lowest is in the south of Asia. Triple rice seasons are planted in Bangladesh and Vietnam. The ranges of rice yield for spring, autumn, and winter are from 3034 to 6249, 2690 to 6986, and 2514 to 10870 kg/ha, with the corresponding averages of 4153, 4716, and 7794 kg/ha. Notably, the highest average yield is 8597 kg/ha for winter rice in Bangladesh, due to its high-yielding hybrid varieties and well-managed fieldwork (e.g., fully irrigated, increasing fertilizer, pesticides, and herbicides applications) (Meroni et al., 2021).

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

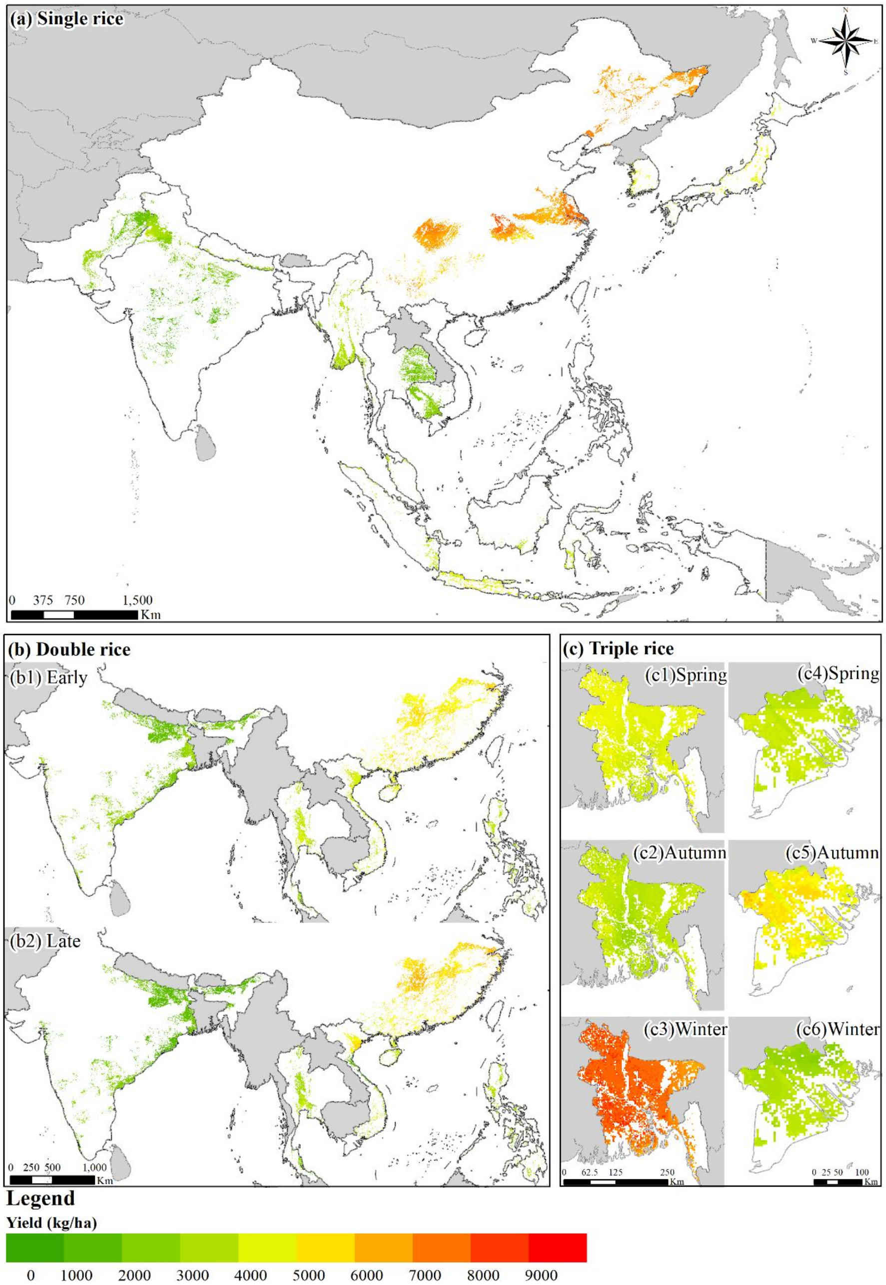


Figure 4: Spatial patterns of the predicted rice yields (the averages during 1995-2015) for different seasons.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

3.3 Comparing AsiaRiceYield4km products with the observations

After aggregating the AsiaRiceYield4km to administrative units, we compared them with the observed yield using R^2 and *RMSE*. The comparisons were separately conducted for single, double, and triple rice in Fig. 5, where the predicted and observed yields were closely around the 1:1 line. The overall R^2 is higher than 0.87 while *RMSE* is lower than 921 kg/ha, suggesting the predicted AsiaRiceYield4km was mostly identical to observations. The accuracy of single rice (R^2: 0.88, *RMSE*: 920 kg/ha) is a bit lower than double (R^2: 0.91, *RMSE*: 554 kg/ha) and triple (R^2: 0.93, *RMSE*: 494 kg/ha) rice mainly because some high-yielding nits were not well predicted for single rice. Moreover, predicted results of late rice show higher accuracy than early rice (R^2: 0.92 > 0.89, *RMSE*: 553 kg/ha < 556 kg/ha), which is consistent with the previous study (Cao et al., 2021). As for triple rice, winter rice has higher accuracy than spring and autumn even though its yield range was the greatest.

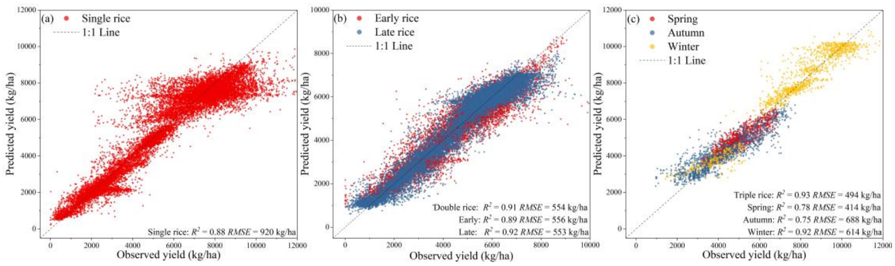


Figure 5: Comparison of AsiaRiceYield4km products with observed yields at administrative units from 1995 to 2015 for (a) single rice, (b) double rice, and (c) triple rice.

3.4 Comparing AsiaRiceYield4km products with SPAM

Due to the limited time coverage and rice seasons information in SPAM, only single rice in 2000, 2005, and 2010 could be compared between AsiaRiceYield4km and SPAM. The spatial distribution of rice yield for AsiaRiceYield4km, SPAM, and observed yield in 2005 were presented in Fig.6 a-c with the zoom-in views of the Indo - Gangetic Plain (IGP) in Pakistan and India (Fig. 6 a1-c1). After aggregating AsiaRiceYield4km and SPAM to administrative units, both were also quantitative compared with the observed yield in Fig. 6d for 2005 and the similar results for 2000 and 2010 were shown in Fig. S1. Compared with SPAM, AsiaRiceYield4km has higher R^2 and lower *RMSE* indicating a better accuracy. More specially, the R^2 of AsiaRiceYield4km are 0.18, 0.23, and 0.20 higher and the corresponding *RMSE* are 570, 692, and 592 kg/ha lower than those of SPAM in 2000, 2005, and 2010, respectively. As for the spatial patterns, moreover, we found AsiaRiceYield4km showed better spatially consistent with the

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

observed yield across the whole area. It can be seen that the yield spatial variation of AsiaRiceYield4km and observed yield is basically identical in IGP, while some administrative unit yields of SPAM have significantly higher than the observed (Fig. 6 a1-c1).

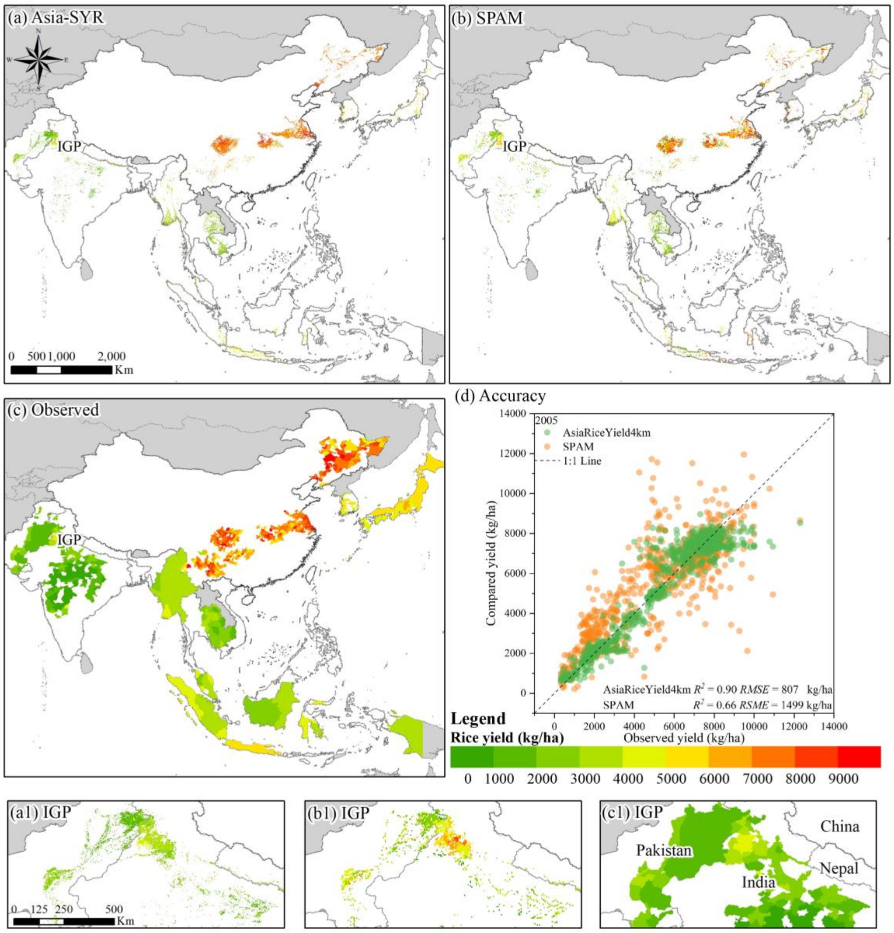


Figure 6: Yield distributions of (a) AsiaRiceYield4km, (b) SPAM, and (c) observed yield in 2005, and (d) the quantitative comparisons with the observed yields in 2005. (a1) to (c1) are the zoom-in views of the Indo - Gangetic Plain (IGP) in Pakistan and India, with (a1) for AsiaRiceYield4km, (b1) for SPAM, and (c1) for the observed yields.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

4 Discussion

4.1 The frequency and importance of the predictors in ML models

.In this study, 50 predictors were used in ML models but their contributions varied greatly. Firstly, only predictors having a significant correlation with yields were selected for ML models for each season with the exception of temporal and spatial predictors (*Year*, *Lon*, *Lat*, and *Ele*) (details in Section 2.3.2). As a result, the selection frequency of temporal and spatial predictors was 27 times and the selection frequency of other predictors ranged from 2 to 25 times (Fig. 7a). Using the selected predictors, ML models then predicted rice yields (Fig. 4) as well as ranked the importance of each predictor (Fig. 7a). Results showed that temporal and spatial predictors had relatively greater importance (>0.05) and the importance of rest predictors was lower than 0.03 (Fig. 7a).

As for different growing stages, REP predictors had greater averaged importance (0.010) in ML models followed by WGP and VEP predictors (0.007 and 0.005), and the averaged selection frequency of WGP and VEP predictors (8.4 and 10.9 times) was much lower than that of REP (14.5 times). Therefore, REP predictors contributed most to yield predictions, which were also consistent with the previous studies (Chang et al., 2005; Nazir et al., 2021). Besides, we also found that EGP predictors (0.014 and 21.3 times) had the greater averaged importance and selection frequency than CGP (0.007 and 11.3times) predictors, indicating that the response of rice yields to extreme growth conditions were stronger.

Fig. 7b further proportioned the importance of the four predictor categories for each seasonal rice. Although the proportioned importance varied for different rice seasons, the overall contribution was highest for CEC predictors (45%) followed by EGP (21%), TI (18%), and CGP predictors (16%). CEC had the greatest proportioned importance for most country which suggested the importance of geographical environment for rice yield prediction. More interestingly, the importance of CEC predictors for Myanmar, Thailand and late season of Vietnam were over 0.8.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Open Access Earth System Science Data Discussions

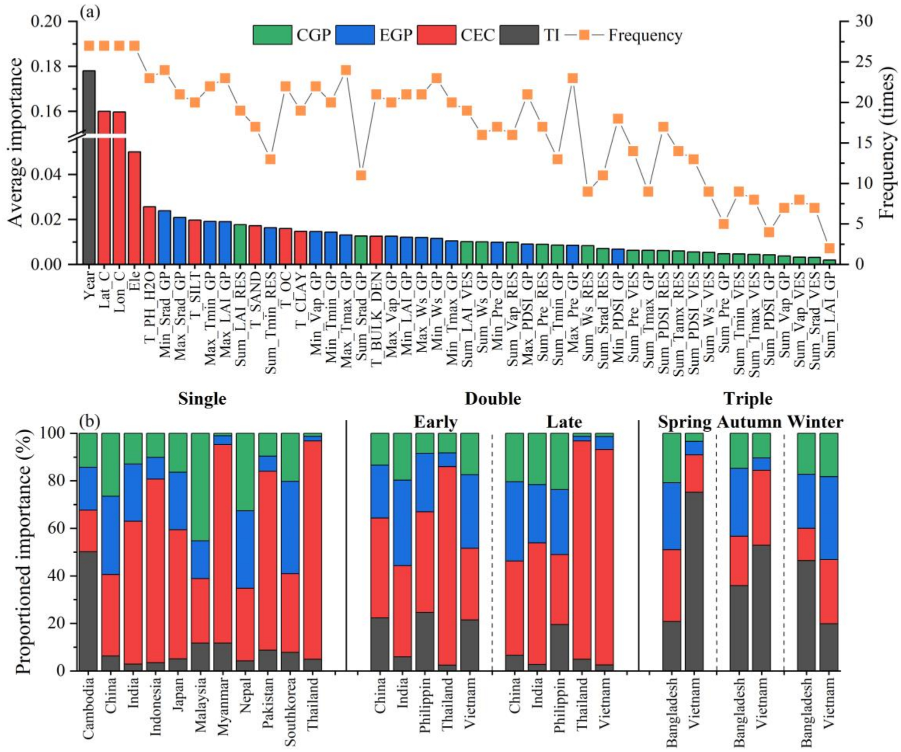


Figure 7: (a) Averaged frequency and importance of each predictor. (b) The proportioned importance of each predictor categories for seasonal rice.

4.2 Improvements in AsiaRiceYield4km

AsiaRiceYield4km is a seasonal rice yield product with high spatiotemporal resolutions and a long time span across the dynamic rice planting areas in main rice-producing countries of Asia. Compared with two public products (M3Crops and SPAM), the spatial resolution of our AsiaRiceYield4km is 4km which is the current highest resolution among all rice yield datasets. Additionally, the product period covers from 1995 to 2015 and includes multi-seasonal rice yields within one year, with more information than most other rice yield datasets. Similarly, our AsiaRiceYield4km considered the annual dynamic change in rice-planting areas and phenological information both at a grid-scale, rather than a constant planting area map and fixed growing period. Such dynamic information assisted us to capture better spatial and temporal variations of rice yields, and consequently greatly improved the accuracy of our product. Moreover, we applied four predictor categories and the optimal ML models to predict seasonal yields. Four predictor categories provided comprehensive rice growth information to ensure accuracy of yield

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

predictions. The optimal predicted models for each rice season are determined by the IPW method. As a weighted ensemble assessment to take a full consideration for training, validation, and testing accuracy, we are sure the IPW method can be more robust and reasonable to select the optimal predicted model for seasonal rice yield in Asia.

4.3 Uncertainty analysis

In this study, we have improved the yield prediction processes to ensure the accuracy of the AsiaRiceYield4km product as possible as we can, however, several factors might negatively impact its results. Due to the limitations of remote sensing technique (i.e., cloud and topography), some paddy rice areas can not be recognized, consequently leading to their map errors (Han et al., 2022). Besides, the spatial resolutions of multi-source data also caused some uncertainties. For example, given that the rice planting areas in Asia are always fragmented (Lowder et al., 2016) but the LAI resolution in this study was somehow coarser (0.05°), the mixed-pixel problem will inevitably influence the accuracy of AsiaRiceYield4km in small size rice-planting areas. Finally, due to the lack of a process-based mechanism, ML is weakly traceable and interpretable for the rice yield variability (Muruganantham et al., 2022). Nevertheless, compared with other public products (Fig. 6), our methods have still generated better seasonal rice yield predictions.

5 Data availability

The seasonal rice yield product for Asia during 1995-2015 (AsiaRiceYield4km) is available at https://doi.org/10.5281/zenodo.6901968 (Wu et al., 2022). We encourage users to independently verify data products before using them.

6 Conclusions

We produced a long-term seasonal rice yield dataset with high spatiotemporal resolution on dynamic paddy rice areas in Asia by multi-source data and ML models. Our AsiaRiceYield4km dataset has higher accuracy compared with other public datasets and shows more spatially consistent with the observed yield. We attributed such improvements to more dynamic information (e.g., rice area and phenological dates), full consideration on rice growing conditions, and the novel IPW method to select the optimal

ML model. Moreover, we found constant environmental conditions contributed the most (~45%) to rice yield prediction than other growing conditions. Predictors in REP had more impacts on yield predictions than those in WGP and VEP. Our dataset can fill the lack of seasonal rice yield datasets and support the studies related to agricultural production and development.

Author contributions.

Conceptualization, Z.Z and F.T.; Data curation, Y.L. and J.H.; Formal analysis: H.W. and J.Z.; Funding acquisition: J.Z., Z.Z. and F.T.; Investigation: J.C., J.H., H.W., J.Z., L.Z., and Y.L.; Methodology, J.Z. and H.W.; Software, H.W. and J.Z.; Supervision, Z.Z., F.T. and J.X.; Validation, J.Z. and J.H.; Visualization: H.W. and J.Z.; Writing – original draft preparation: H.W. and J.Z.; Writing – review & editing: J.Z., Z.Z., and Q.M. All authors have read and agreed to the published version of the manuscript.

Competing interests.

The contact author has declared that neither they nor their co-authors have any competing interests.

Disclaimer.

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements.

We would like to thank the editors and anonymous reviewers for their valuable comments. And we would like to thank the support of the open project of the Key Laboratory of Environmental Change and Natural Disasters, Ministry of Education, Beijing Normal University.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Financial support.

This research was funded by the National Key Research and Development Project of China (2020YFA0608201), China National Postdoctoral Program for Innovative Talents (BX20200064), and National Natural Science Foundation of China (42061144003, 41977405, 42101095).

References

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., and Hegewisch, K. C.: TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015, Sci. Data, 5, 1–12, https://doi.org/10.1038/sdata.2017.191, 2018.

Alexandratos, N. and Bruinsma, J.: World agriculture towards 2030/2050: the 2012 revision, 2012.

Arumugam, P., Chemura, A., Schauberger, B., and Gornott, C.: Remote Sensing Based Yield Estimation of Rice (Oryza Sativa L.) Using Gradient Boosted Regression in India, Remote Sens., 13, 2379, https://doi.org/10.3390/rs13122379, 2021.

Bandumula, N.: Rice production in Asia: Key to global food security, Proc. Natl. Acad. Sci. India Sect. B Biol. Sci., 88, 1323–1328, https://doi.org/10.1007/s40011-017-0867-7, 2018.

Birla, D. S., Malik, K., Sainger, M., Chaudhary, D., Jaiwal, R., and Jaiwal, P. K.: Progress and challenges in improving the nutritional quality of rice (Oryza sativa L.), Crit. Rev. Food Sci. Nutr., 57, 2455–2481, https://doi.org/10.1080/10408398.2015.1084992, 2017.

Blomqvist, L., Yates, L., and Brook, B. W.: Drivers of increasing global crop production: A decomposition analysis, Environ. Res. Lett., 15, 0940b6, https://doi.org/10.1088/1748-9326/ab9e9c, 2020.

Breiman, L.: Bagging predictors, Mach. Learn., 24, 123–140, https://doi.org/10.1007/BF00058655, 1996.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., and You, L.: Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches, Agric. For. Meteorol., 274, 144–159, https://doi.org/10.1016/j.agrformet.2019.03.010, 2019.

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Han, J., and Li, Z.: Identifying the Contributions of Multi-Source Data for Winter Wheat Yield Prediction in China, Remote Sens., 12, 750, https://doi.org/10.3390/rs12050750, 2020.

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., and Xie, J.: Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches, Agric. For. Meteorol., 297, 108275, https://doi.org/10.1016/j.agrformet.2020.108275, 2021.

Chang, K.-W., Shen, Y., and Lo, J.-C.: Predicting rice yield using canopy reflectance measured at booting stage, Agron. J., 97, 872–878, https://doi.org/10.2134/agronj2004.0162, 2005.

Chen, S., Liu, W., Feng, P., Ye, T., Ma, Y., and Zhang, Z.: Improving Spatial Disaggregation of Crop Yield by Incorporating Machine Learning with Multisource Data: A Case Study of Chinese Maize Yield, Remote Sens., 14, 2340, https://doi.org/10.3390/rs14102340, 2022.

Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco, CA, USA, 13-17 August 2016, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Chen, Y., Song, X., Wang, S., Huang, J., and Mansaray, L. R.: Impacts of spatial heterogeneity on crop area mapping in Canada using MODIS data, ISPRS J. Photogramm. Remote Sens., 119, 451–461, https://doi.org/10.1016/j.isprsjprs.2016.07.007, 2016.

Chen, Y., Zhang, Z., Tao, F., Palosuo, T., and Rötter, R. P.: Impacts of heat stress on leaf area index and growth duration of winter wheat in the North China Plain, Field Crops Res., 222, 230–237, https://doi.org/10.1016/j.fcr.2017.06.007, 2018.

Chlingaryan, A., Sukkarieh, S., and Whelan, B.: Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, Comput. Electron. Agric., 151, 61–69, https://doi.org/10.1016/j.compag.2018.05.012, 2018.

Dawe, D., Pandey, S., and Nelson, A.: Emerging trends and spatial patterns of rice production, in: Rice in the global economy: Strategic research and policy issues for food security, edited by: Sushil, P., Derek, B., David, D., Achim, D., Samarendu, M., Scott, R., and Bill, H., International Rice Research Institute (IRRI), Los Baños, Philippines, 15–36, 2010.

Dinh, T. L. A. and Aires, F.: Nested leave-two-out cross-validation for the optimal crop yield model selection, Geosci. Model Dev., 15, 3519–3535, https://doi.org/10.5194/gmd-15-3519-2022, 2022.

Fang, H., Liang, S., and Hoogenboom, G.: Integration of MODIS LAI and vegetation index products with the CSM–CERES–Maize model for corn yield estimation, Int. J. Remote Sens., 32, 1039–1065, https://doi.org/10.1080/01431160903505310, 2011.

Food and Agriculture Organization of the United Nations, FAOSTAT: https://www.fao.org/faostat/en/#data/QCL/visualize, last access: 6 April 2022.

Fernandez-Beltran, R., Baidar, T., Kang, J., and Pla, F.: Rice-yield prediction with multi-temporal sentinel-2 data and 3D CNN: A case study in Nepal, Remote Sens., 13, 1391, https://doi.org/10.3390/rs13071391, 2021.

Folberth, C., Skalský, R., Moltchanova, E., Balkovič, J., Azevedo, L. B., Obersteiner, M., and Van Der Velde, M.: Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations, Nat. Commun., 7, 1–13, https://doi.org/10.1038/ncomms11872, 2016.

Folberth, C., Khabarov, N., Balkovič, J., Skalský, R., Visconti, P., Ciais, P., Janssens, I. A., Peñuelas, J., and Obersteiner, M.: The global cropland-sparing potential of high-yield farming, Nat. Sustain., 3, 281–289, https://doi.org/10.1038/s41893-020-0505-x, 2020.

Fritz, S., See, L., Bayas, J. C. L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., and Crutchfield, J.: A comparison of global agricultural monitoring systems and current gaps, Agric. Syst., 168, 258–272, https://doi.org/10.1016/j.agsy.2018.05.010, 2019.

GLOBE Task Team and others: The Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0., 1999.

Han, J., Zhang, Z., Luo, Y., Cao, J., Zhang, L., Cheng, F., Zhuang, H., and Zhang, J.: APRA500: a 500 m annual paddy rice dataset for monsoon Asia using multisource remote sensing data [data set]. Zenodo., https://doi.org/10.5281/zenodo.5555721, 2021.

Han, J., Zhang, Z., Luo, Y., Cao, J., Zhang, L., Zhuang, H., Cheng, F., Zhang, J., and Tao, F.: Annual paddy rice planting area and cropping intensity datasets and their dynamics in the Asian monsoon region from 2000 to 2020, Agric. Syst., 200, 103437, https://doi.org/10.1016/j.agsy.2022.103437, 2022.

He, T., Xie, C., Liu, Q., Guan, S., and Liu, G.: Evaluation and comparison of random forest and A-LSTM networks for large-scale winter wheat identification, Remote Sens., 11, 1665, https://doi.org/10.3390/rs11141665, 2019.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Huang, J., Wang, X., Li, X., Tian, H., and Pan, Z.: Remotely sensed rice yield prediction using multi-temporal NDVI data derived from NOAA's-AVHRR, PloS ONE, 8, e70816, https://doi.org/10.1371/journal.pone.0070816, 2013.

Huntington, T., Cui, X., Mishra, U., and Scown, C. D.: Machine learning to predict biomass sorghum yields under future climate scenarios, Biofuels Bioprod. Biorefining, 14, 566–577, https://doi.org/10.1002/bbb.2087, 2020.

Iizumi, T. and Sakai, T.: The global dataset of historical yields for major crops 1981–2016, Sci. Data, 7, 97, https://doi.org/10.1038/s41597-020-0433-7, 2020.

Iizumi, T., Yokozawa, M., Sakurai, G., Travasso, M. I., Romanenkov, V., Oettli, P., Newby, T., Ishigooka, Y., and Furuya, J.: Historical changes in global yields: major cereal and legume crops from 1982 to 2006, Glob. Ecol. Biogeogr., 23, 346–357, https://doi.org/10.1111/geb.12120, 2014.

Iizumi, T., Hosokawa, N., and Wagai, R.: Soil carbon-food synergy: sizable contributions of small-scale farmers, CABI Agric. Biosci., 2, 1–15, https://doi.org/10.1186/s43170-021-00063-6, 2021.

Jeong, S., Ko, J., and Yeom, J.-M.: Predicting rice yield at pixel scale through synthetic use of crop and deep learning models with satellite data in South and North Korea, Sci. Total Environ., 802, 149726, https://doi.org/10.1016/j.scitotenv.2021.149726, 2022.

Jin, X., Diao, W., Xiao, C., Wang, F., Chen, B., Wang, K., and Li, S.: Estimation of wheat agronomic parameters using new spectral indices, PloS ONE, 8, e72736, https://doi.org/10.1371/journal.pone.0072736, 2013.

Kaltenegger, K. and Winiwarter, W.: Global gridded nitrogen indicators: influence of crop maps, Glob. Biogeochem. Cycles, 34, e2020GB006634, https://doi.org/10.1029/2020GB006634, 2020.

Kim, K.-H., Doi, Y., Ramankutty, N., and Iizumi, T.: A review of global gridded cropping system data products, Environ. Res. Lett., 16, 093005, https://doi.org/10.1088/1748-9326/ac20f4, 2021.

van Klompenburg, T., Kassahun, A., and Catal, C.: Crop yield prediction using machine learning: A systematic literature review, Comput. Electron. Agric., 177, 105709, https://doi.org/10.1016/j.compag.2020.105709, 2020.

Laborte, A. G., Gutierrez, M. A., Balanza, J. G., Saito, K., Zwart, S. J., Boschetti, M., Murty, M. V. R., Villano, L., Aunario, J. K., Reinke, R., Koo, J., Hijmans, R. J., and Nelson, A.: RiceAtlas, a spatial

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

database of global rice calendars and production, Sci. Data, 4, 170074, https://doi.org/10.1038/sdata.2017.74, 2017.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

Lin, T.-S., Song, Y., Lawrence, P., Kheshgi, H. S., and Jain, A. K.: Worldwide Maize and Soybean Yield Response to Environmental and Management Factors Over the 20th and 21st Centuries, J. Geophys. Res. Biogeosciences, 126, e2021JG006304, https://doi.org/10.1029/2021JG006304, 2021.

Liu, C., Huang, H., and Sun, F.: A Pixel-Based Vegetation Greenness Trend Analysis over the Russian Tundra with All Available Landsat Data from 1984 to 2018, Remote Sens., 13, 4933, https://doi.org/10.3390/rs13234933, 2021a.

Liu, W., Ye, T., Jägermeyr, J., Müller, C., Chen, S., Liu, X., and Shi, P.: Future climate change significantly alters interannual wheat yield variability over half of harvested areas, Environ. Res. Lett., https://doi.org/10.1088/1748-9326/ac1fbb, 2021b.

Lowder, S. K., Skoet, J., and Raney, T.: The number, size, and distribution of farms, smallholder farms, and family farms worldwide, World Dev., 87, 16–29, https://doi.org/10.1016/j.worlddev.2015.10.041, 2016.

Luo, Y., Zhang, Z., Chen, Y., Li, Z., and Tao, F.: ChinaCropPhen1km: a high-resolution crop phenological dataset for three staple crops in China during 2000–2015 based on leaf area index (LAI) products, Earth Syst. Sci. Data, 12, 197–214, https://doi.org/10.5194/essd-12-197-2020, 2020.

Luo, Y., Zhang, Z., Cao, J., Zhang, L., Zhang, J., Han, J., Zhuang, H., Cheng, F., and Tao, F.: Accurately mapping global wheat production system using deep learning algorithms, Int. J. Appl. Earth Obs. Geoinformation, 110, 102823, https://doi.org/10.1016/j.jag.2022.102823, 2022.

Maclean, J. L., Dawe, D. C., Hettel, G. P., and Hettel, G. P. (Eds.): Rice almanac: Source book for the most important economic activity on earth, 3rd ed., CABI Publishing, Wallingford, UK, 2002.

Manfron, G., Delmotte, S., Busetto, L., Hossard, L., Ranghetti, L., Brivio, P. A., and Boschetti, M.: Estimating inter-annual variability in winter wheat sowing dates from satellite time series in Camargue, France, Int. J. Appl. Earth Obs. Geoinformation, 57, 190–201, https://doi.org/10.1016/j.jag.2017.01.001, 2017.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Meroni, M., Waldner, F., Seguini, L., Kerdiles, H., and Rembold, F.: Yield forecasting with machine learning and small data: what gains for grains?, Agric. For. Meteorol., 308, 108555, https://doi.org/10.1016/j.agrformet.2021.108555, 2021.

Monfreda, C., Ramankutty, N., and Foley, J. A.: Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000, Glob. Biogeochem. Cycles, 22, https://doi.org/10.1029/2007GB002947, 2008.

Mosleh, M. K., Hassan, Q. K., and Chowdhury, E. H.: Application of remote sensors in mapping rice area and forecasting its production: A review, Sensors, 15, 769–791, https://doi.org/10.3390/s150100769, 2015.

Muehe, E. M., Wang, T., Kerl, C. F., Planer-Friedrich, B., and Fendorf, S.: Rice production threatened by coupled stresses of climate and soil arsenic, Nat. Commun., 10, 1–10, https://doi.org/10.1038/s41467-019-12946-4, 2019.

Müller, C., Elliott, J., Kelly, D., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Folberth, C., Hoek, S., Izaurralde, R. C., Jones, C. D., Khabarov, N., Lawrence, P., Liu, W., Olin, S., Pugh, T. A. M., Reddy, A., Rosenzweig, C., Ruane, A. C., Sakurai, G., Schmid, E., Skalsky, R., Wang, X., de Wit, A., and Yang, H.: The Global Gridded Crop Model Intercomparison phase 1 simulation dataset, Sci. Data, 6, 50, https://doi.org/10.1038/s41597-019-0023-8, 2019.

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., and Islam, N.: A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing, Remote Sens., 14, 1990, https://doi.org/10.3390/rs14091990, 2022.

Nazir, A., Ullah, S., Saqib, Z. A., Abbas, A., Ali, A., Iqbal, M. S., Hussain, K., Shakir, M., Shah, M., and Butt, M. U.: Estimation and Forecasting of Rice Yield Using Phenology-Based Algorithm and Linear Regression Model on Sentinel-II Satellite Data, Agriculture, 11, 1026, https://doi.org/10.3390/agriculture11101026, 2021.

Nelson, A. and Gumma, M. K.: A map of lowland rice extent in the major rice growing countries of Asia, Los Baños, Philippines, 2015.

Obsie, E. Y., Qu, H., and Drummond, F.: Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms, Comput. Electron. Agric., 178, 105778, https://doi.org/10.1016/j.compag.2020.105778, 2020.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Open Access Earth System Science Data Discussions

van Oort, P. A. and Zwart, S. J.: Impacts of climate change on rice production in Africa and causes of simulated yield changes, Glob. Change Biol., 24, 1029–1045, https://doi.org/10.1111/gcb.13967, 2018.

Qian, H., Huang, S., Chen, J., Wang, L., Hungate, B. A., van Kessel, C., Zhang, J., Deng, A., Jiang, Y., and van Groenigen, K. J.: Lower-than-expected CH4 emissions from rice paddies with rising CO2 concentrations, Glob. Change Biol., 26, 2368–2376, https://doi.org/10.1111/gcb.14984, 2020.

Ripley, B. D.: Pattern recognition and neural networks, Cambridge university press, 2007.

Sak, H., Senior, A., and Beaufays, F.: Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, arXiv [preprint], arXiv:1402.1128, 5 February 2014.

Sakamoto, T.: Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm, ISPRS J. Photogramm. Remote Sens., 160, 208–228, https://doi.org/10.1016/j.isprsjprs.2019.12.012, 2020.

Sakamoto, T., Yokozawa, M., Toritani, H., Shibayama, M., Ishitsuka, N., and Ohno, H.: A crop phenology detection method using time-series MODIS data, Remote Sens. Environ., 96, 366–374, https://doi.org/10.1016/j.rse.2005.03.008, 2005.

Salvacion, A. R.: Chapter 11 - Multiscale drought hazard assessment in the Philippines, in: Computers in Earth and Environmental Sciences, edited by: Pourghasemi, H. R., Elsevier, Amsterdam, Netherlands, 169–179, 2022.

Shahhosseini, M., Hu, G., and Archontoulis, S. V.: Forecasting Corn Yield With Machine Learning Ensembles, Front. Plant Sci., 11, 1120, https://doi.org/10.3389/fpls.2020.01120, 2020.

Shahhosseini, M., Hu, G., Huber, I., and Archontoulis, S. V.: Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt, Sci. Rep., 11, 1606, https://doi.org/10.1038/s41598-020-80820-1, 2021.

Son, N. T., Chen, C. F., Chen, C. R., Chang, L. Y., Duc, H. N., and Nguyen, L. D.: Prediction of rice crop yield using MODIS EVI- LAI data in the Mekong Delta, Vietnam, Int. J. Remote Sens., 34, 7275–7292, https://doi.org/10.1080/01431161.2013.818258, 2013.

Son, N.-T., Chen, C.-F., Chen, C.-R., Guo, H.-Y., Cheng, Y.-S., Chen, S.-L., Lin, H.-S., and Chen, S.-H.: Machine learning approaches for rice crop yield predictions using time-series satellite data in Taiwan, Int. J. Remote Sens., 41, 7868–7888, https://doi.org/10.1080/01431161.2020.1766148, 2020.

Srivastava, V. C. and Mahapatra, I. C.: Advances in Rice Production Technology: Theory and Practice, Agrobios, India, 2012.

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., and Li, H.: An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China, Agric. For. Meteorol., 310, 108629, https://doi.org/10.1016/j.agrformet.2021.108629, 2021.

Wang, C., Zhang, Z., Chen, Y., Tao, F., Zhang, J., and Zhang, W.: Comparing different smoothing methods to detect double-cropping rice phenology based on LAI products–a case study in the Hunan province of China, Int. J. Remote Sens., 39, 6405–6428, https://doi.org/10.1080/01431161.2018.1460504, 2018.

Wieder, W. R., Boehnert, J., Bonan, G. B., and Langseth, M.: Regridded harmonized world soil database v1. 2 ORNL DAAC [data set], https://doi.org/10.3334/ORNLDAAC/1247, 2014.

Wu, H., Zhang, J., Zhang, Z., Han, J., Cao, J., Zhang, L., Luo, Y., Mei, Q., Xu, J., and Tao, F.: AsiaRiceYield4km: Seasonal Rice Yield in Asia from 1995 to 2015 [data set]. Zenodo., https://doi.org/10.5281/zenodo.6901968, 2022.

Xiao, Z., Liang, S., Wang, J., Chen, P., Yin, X., Zhang, L., and Song, J.: Use of general regression neural networks for generating the GLASS leaf area index product from time-series MODIS surface reflectance, IEEE Trans. Geosci. Remote Sens., 52, 209–223, https://doi.org/10.1109/TGRS.2013.2237780, 2013.

Xiao, Z., Liang, S., Wang, J., Xiang, Y., Zhao, X., and Song, J.: Long-time-series global land surface satellite leaf area index product derived from MODIS and AVHRR surface reflectance, IEEE Trans. Geosci. Remote Sens., 54, 5301–5318, https://doi.org/10.1109/TGRS.2016.2560522, 2016.

You, L. and Wood, S.: An entropy approach to spatial disaggregation of agricultural production, Agric. Syst., 90, 329–347, https://doi.org/10.1016/j.agsy.2006.01.008, 2006.

Yu, Q., You, L., Wood-Sichra, U., Ru, Y., Joglekar, A. K., Fritz, S., Xiong, W., Lu, M., Wu, W., and Yang, P.: A cultivated planet in 2010–Part 2: the global gridded agricultural-production maps, Earth Syst. Sci. Data, 12, 3545–3572, https://doi.org/10.5194/essd-12-3545-2020, 2020.

Zhang, G., Xiao, X., Dong, J., Xin, F., Zhang, Y., Qin, Y., Doughty, R. B., and Moore, B.: Fingerprint of rice paddies in spatial–temporal dynamics of atmospheric methane concentration in monsoon Asia, Nat. Commun., 11, 1–11, 2020a.

https://doi.org/10.5194/essd-2022-273
Preprint. Discussion started: 29 August 2022

Zhang, L., Zhang, Z., Luo, Y., Cao, J., and Tao, F.: Combining Optical, Fluorescence, Thermal Satellite, and Environmental Data to Predict County-Level Maize Yield in China Using Machine Learning Approaches, Remote Sens., 12, 21, https://doi.org/10.3390/rs12010021, 2019.

Zhang, L., Zhang, Z., Luo, Y., Cao, J., Xie, R., and Li, S.: Integrating satellite-derived climatic and vegetation indices to predict smallholder maize yield using deep learning, Agric. For. Meteorol., 311, 108666, https://doi.org/10.1016/j.agrformet.2021.108666, 2021.

Zhang, T., Yang, X., Wang, H., Li, Y., and Ye, Q.: Climatic and technological ceilings for C hinese rice stagnation based on yield gaps and yield trend pattern analysis, Glob. Change Biol., 20, 1289–1298, https://doi.org/10.1111/gcb.12428, 2014.

Zhang, Z., Li, Z., Chen, Y., Zhang, L., and Tao, F.: Improving regional wheat yields estimations by multi-step-assimilating of a crop model with multi-source data, Agric. For. Meteorol., 290, 107993, https://doi.org/10.1016/j.agrformet.2020.107993, 2020b.