scientific data

OPEN

DATA DESCRIPTOR

High-resolution crop yield and water productivity dataset generated using random forest and remote sensing

Minghan Cheng[1,2,3], Xiyun Jiao[4], Lei Shi[3], Josep Penuelas[5,6], Lalit Kumar[7], Chenwei Nie[1], Tianao Wu[4], Kaihua Liu[8], Wenbin Wu[9] ✉ & Xiuliang Jin[3,10] ✉

Accurate and high-resolution crop yield and crop water productivity (CWP) datasets are required to understand and predict spatiotemporal variation in agricultural production capacity; however, datasets for maize and wheat, two key staple dryland crops in China, are currently lacking. In this study, we generated and evaluated a long-term data series, at 1-km resolution of crop yield and CWP for maize and wheat across China, based on the multiple remotely sensed indicators and random forest algorithm. Results showed that MOD16 products are an accurate alternative to eddy covariance flux tower data to describe crop evapotranspiration (maize and wheat RMSE: 4.42 and 3.81 mm/8d, respectively) and the proposed yield estimation model showed accuracy at local (maize and wheat rRMSE: 26.81 and 21.80%, respectively) and regional (maize and wheat rRMSE: 15.36 and 17.17%, respectively) scales. Our analyses, which showed spatiotemporal patterns of maize and wheat yields and CWP across China, can be used to optimize agricultural production strategies in the context of maintaining food security.

Background & Summary

Crop water productivity (CWP), calculated as the ratio of crop yield to gross evapotranspiration (ET), is a quantitative indicator of agricultural performance[1] that may be used to assess the impact of agri-environment and crop management strategies on crop growth[2,3]. Thus, accurate measurement of crop yield and ET as components of CWP is important in agricultural production decision-making and management of water resources[4].

Methods that measure ET, such as lysimeter devices[5] and the eddy covariance technique[6], and approaches to its estimation, such as the energy balance Bowen ratio[7] and the Penman-Monteith algorithm[8,9], have tended to be used in point-scale and small area-scale studies[10], while crop yield has generally been measured using quantitative field-based sampling, qualitative farmer or expert estimates, and micrometeorological measurements[1]. Policy-driven management of agricultural production often requires regional-scale, high spatial resolution monitoring of yield and ET; however, conventional methods and approaches to ET measurement and estimation are limited by low levels of efficiency and a lack of suitability for regional scale studies. Thus, remote-sensing technology has been adopted as an alternative data source for regional-scale, high spatial resolution estimates of ET, including in the Surface Energy Balance Algorithm for Land[11,12], the Surface Energy Balance System[13], the Two-source Energy Balance method[14], and improved Penman-Monteith[15,16] and Priestley-Taylor[17] algorithms,

[1]Jiangsu Key Laboratory of Crop Genetics and Physiology/Jiangsu Key Laboratory of Crop Cultivation and Physiology, Agricultural College, Yangzhou University, 225009, Yangzhou, P.R. China. [2]Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, 225009, Yangzhou, P.R. China. [3]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences/Key Laboratory of Crop Physiology and Ecology, Ministry of Agriculture, Beijing, 100081, P.R. China. [4]College of Agricultural Science and Engineering, Hohai University, Nanjing, Jiangsu Province, 210098, P.R. China. [5]CSIC, Global Ecology Unit CREAF-CSIC-UAB, Bellaterra, 08193, Barcelona, Catalonia, Spain. [6]CREAF, Cerdanyola del Vallès, 08193, Barcelona, Catalonia, Spain. [7]EastCoast Geospatial Consultants, Armidale, NSW, 2350, Australia. [8]College of Hydrology and Water Resources, Hohai University, Nanjing, Jiangsu Province, 210098, P.R. China. [9]Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, 100081, Beijing, P.R. China. [10]National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, 572024, Sanya, China. ✉e-mail: wuwenbin@caas.cn; jinxiuliang@caas.cn

where the widely used MOD16 ET product, generated using the improved Penman–Monteith method, has been shown to have good levels of accuracy[18,19].

Estimates of remotely sensed (RS) crop yields derive from data assimilation (DA) in crop models[20–23] or regression analysis of RS indicators (RSIs)[1,24]. In general, the DA approach has been applied over a wide range of crops and land surface and environment conditions[23], for example, Jin, *et al.*[25] assimilated RS data from RADARSAT-2 and HJ-1A/B into an AquaCrop model to estimate wheat yields ($R^2 = 0.42$). However, performance of crop models is limited by complexity and uncertainty of input parameters, such as soil properties, meteorological data, crop cultivars, and management practices, that negatively affect simulation processes and cause larger errors in crop yield estimates[26]. In contrast, approaches that use RSI are based on fitted relationships, which tend to be nonlinear[24,27], between *in-situ* measurements of yield and indicators, such as vegetation indices (VIs), ET, and gross primary productivity (GPP)[28–30]. These approaches have been widely used, due to their simplicity and efficiency; for example, Noland, *et al.*[31] found 81−90% of the variation in alfalfa yields was explained by VIs calculated from multispectral data and Cao, *et al.*[32] found the combination of the enhanced vegetation index (EVI) with deep-learning algorithms accounted for 71% of the variation in winter wheat yields. Machine-learning algorithms are well suited for dealing with nonlinear heteroscedastic problems and are used for efficient data processing and data mining[33,34], and algorithms, such as support vector regression[35], random forest (RF) regression[36], and artificial neural networks[35], have been used successfully to analyze agricultural RS data. For example, Maimaitijiang, *et al.*[35] analyzed multimodal data (canopy texture and structure, spectra and temperature) collected by unmanned aerial vehicles (UAV) using machine-learning algorithms to estimate field-scale soybean yields, while Johansen, *et al.*[37] leveraged multi-spectral UAV data and a RF model to prediction tomato phenotype yield and biomass.

The distribution of water resources across China is heterogenous, with particular areas of scarcity in the northwest[38], and nationally, agricultural production accounts for 60−65% of water consumption[39]. Maize and wheat are staple dryland crops in China, with areas of cultivation of 41.3×10^6 and 23.7×10^6 ha, respectively, in 2019, so the accurate estimation of CWP at high spatial resolution is essential for ensuring sustainable agricultural production and water resource management in the context of maintaining food security. Currently, understanding of CWP of key food security crops in China is lacking, therefore, the aim of this study was to estimate CWP of maize and wheat across China at a high level of spatiotemporal resolution, based on multiple remote sensing indicators and combined ensemble machine learning and RF algorithms. Specifically, our objectives were to: (1) evaluate the accuracy of the MOD16 ET product in the estimation of crop water consumption; (2) test the accuracy of estimates of CWP based on RS-EVI and combined machine learning and RF algorithms; and, (3) quantify spatiotemporal patterns of crop yield and CWP across China.

Methods

Study area. China (3°31′00″–53°33′47″N, 73°29′59.79″–135°2′30″E) covers a land area of approximately 9.6×10^6 km^2 that is largely dominated by temperate climate conditions, with tropical climate conditions prevailing over a smaller relative area. The study area comprised the Qinghai Tibet Plateau (QTP), Huang-Huai-Hai Plain (HHHP), Loess Plateau (LP), Sichuan Basin (SB), Middle-lower Yangtze River Plain (MLYR), Northeast China Plain (NeCP), Yunnan-Guizhou Plateau (YGP), and the Northern arid and semiarid region (NaR) regions of agricultural production, but excluded Southern China (SC) due to the small areas of cultivation of maize and wheat[40] (Fig. 1).

Study parameters and data sources. *Cropland map.* We used cultivation area, yield, and CWP data for maize and wheat from 2001 to 2015. Data for cultivation area of maize and wheat were obtained from the 1-km National Land Cover Dataset (NLCD) (http://www.resdc.cn; Fig. 2) and generally showed an increase over the study period in most regions, where area of maize cultivation was greatest in NeCP and HHHP and area of wheat cultivation was greatest in HHHP.

Input variables. We selected seven indicators of crop yield (GPP; ET; land surface temperature, Ts; leaf area index, LAI; and, soil content of clay, sand, silt) as model inputs to estimate maize and wheat yield. Crop phenology data (annual at 1-km) were obtained from the ChinaCropPhen1km dataset[41,42] that comprises Julian day (day of the year, DOY) of the main crop growth stages: from V3 in maize (the third leaf is fully expanded) to maturity, and from emergency (spring wheat) or green up (winter wheat) to maturity in wheat.

Data for Ts and crop ET, GPP, and LAI were obtained from MOD11A2 Ts products, MOD16A2 ET products, MOD17A2 GPP products, and MOD15A2 LAI products, respectively, for regular 500-m grid cells aggregated to 1 km, to harmonize with the 1-km resolutions of the NLCD and ChinaCropPheno datasets, for the global vegetated land surface at an 8-d composite. Soil clay, silt, and sand content data were obtained from the 1:1 million soil type map and soil profile data were obtained from the Second China Soil Survey[43]; all soil data were at a spatial resolution of 1 km.

In situ crop yield. Crop yield data across the study period at the administrative county level were obtained from the China Rural Statistical Yearbook in the National Bureau of Statistics of China (NBSC, http://www.stats.gov.cn/), with gaps of several years in parts of some counties, and outliers were identified and excluded if they were outside the range of biophysical attainable yields (maize: <500 kg/ha or >15,000 kg/ha; wheat: <500 kg/ha or >13,000 kg/ha), or they were greater or less than 3 SD from the study period average, or derived from counties with >10,000 ha of planting area[32,44,45]. As a result of this filtering process, our dataset comprised 1981 and 2487 records of maize and wheat yields, respectively. Pixel-level crop yield data, derived from the National Meteorological Data Center of China[41], were measured at 12 (in which, a total of 9 sites recorded two year's samples and others only recorded one year's sample) and 23 (in which, a total of 11 sites recorded three year's

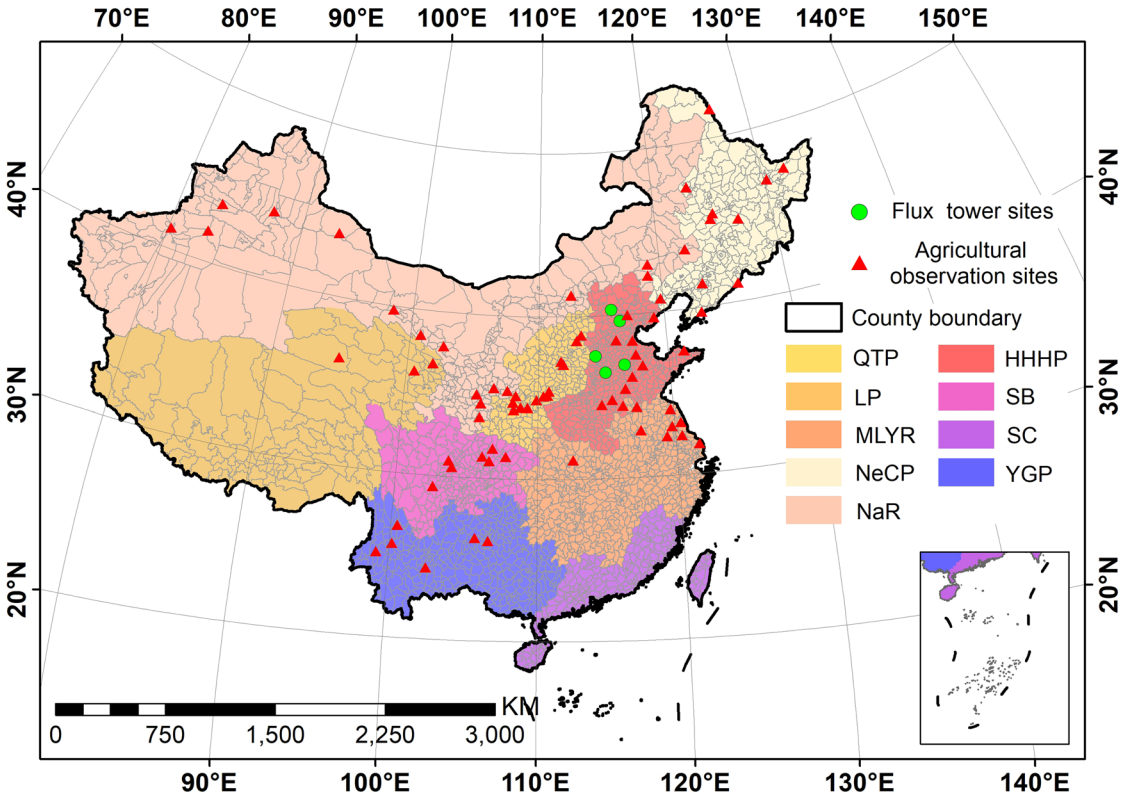

Fig. 1 Study area and study sites by agricultural production region. QTP: Qinghai Tibet Plateau; HHHP: Huang-Huai-Hai Plain; LP: Loess Plateau; SB: Sichuan Basin; MLYR: Middle-lower Yangtze Plain; SC: Southern China; NeCP: Northeast China Plain; YGP: Yunnan-Guizhou Plateau; and, NaR: Northern arid and semiarid region.

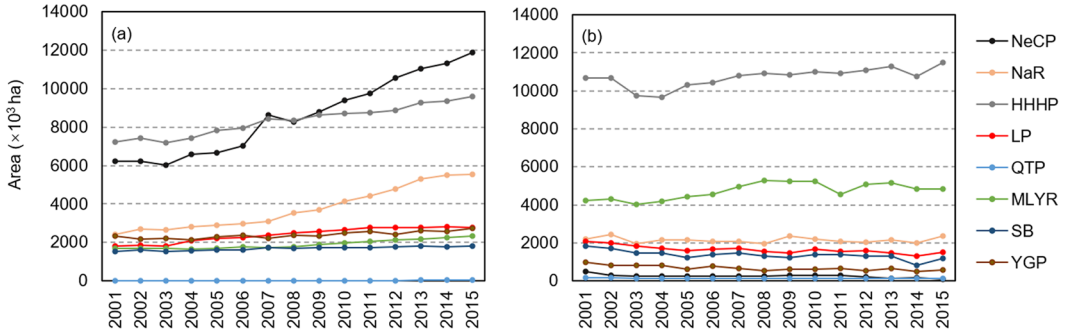

Fig. 2 Cultivation areas of maize (**a**) and wheat (**b**) in China over the period 2001−15. QTP: Qinghai Tibet Plateau; HHHP: Huang-Huai-Hai Plain; LP: Loess Plateau; SB: Sichuan Basin; MLYR: Middle-lower Yangtze Plain; SC: Southern China; NeCP: Northeast China Plain; YGP: Yunnan-Guizhou Plateau; and, NaR: Northern arid and semiarid region.

samples, 6 sites recorded two year's samples and others only recorded one year's sample) study sites for maize and wheat, respectively, and at 42 study sites (only recorded one year's sample) for both crops in a rotation. In summary, a total of 63 maize yield samples and 103 wheat yield samples were available for validation. It should be noted that the crop yield at county level and pixel level were recorded based on the harvested and measured grain yield, in which the maize yield was converted at the moisture of 14% and wheat yield was at 12.5%.

Flux tower observations. We derived EC data from ChinaFLUX recording stations located in maize and wheat crops in Daxing, Guantao, Huailai, Luancheng, and Yucheng for MOD16 ET assessment (Fig. 1), where ET was cumulated over 8-d periods, to harmonize with the MOD16 ET product temporal resolution (8-day composite). Table 1 shows the main information and sources of all data used in this study.

Data type	Temporal resolution	Spatial resolution	Source
Evapotranspiration (ET)	8-day composite	500 m (Aggregated to 1 km)	NASA, MOD16A2 ET product (http://ladsweb.modaps.eosdis.nasa.gov)
Gross primary productivity (GPP)	8-day composite	500 m (Aggregated to 1 km)	NASA, MOD17A2 GPP product (http://ladsweb.modaps.eosdis.nasa.gov)
Surface temperature (Ts)	8-day composite	500 m (Aggregated to 1 km)	NASA, MOD11A2 Ts product (http://ladsweb.modaps.eosdis.nasa.gov)
Leaf area index (LAI)	8-day composite	500 m (Aggregated to 1 km)	NASA, MOD15A2 LAI product (http://ladsweb.modaps.eosdis.nasa.gov)
Soil properties	n/a	1 km	Resource and Environment Science and Data Center, Chinese Academy of Science (http://www.resdc.cn)
Phenology information	Yearly	1 km	ChinaCropPhen1km[41] (https://doi.org/10.6084/m9.figshare.8313530)
Cultivated-land layer	Yearly	1 km	Resource and Environment Science and Data Center, Chinese Academy of Science (http://www.resdc.cn)
Recorded yield (regional-scale)	Yearly	County-level	China Rural Statistical Yearbook, National Bureau of Statistics of China (http://www.stats.gov.cn)
Measured yield	Yearly	Point-scale	National Meteorological Data Center of China (http://data.cma.cn)
Flux tower observed data	Daily (Cumulated to eight days)	Point-scale	ChinaFLUX (http://www.chinaflux.org)

Table 1. Data types, spatiotemporal resolution, and sources.

Estimation of crop water productivity. *Model process of evapotranspiration and yield.* Crop ET was derived from the MOD16 ET product, using an improved Penman-Monteith algorithm[15,16] and crop yields were estimated using the Random Forest (RF) regression algorithm. The steps for generating the crop yield dataset are as follows:

(1) Collecting the input variables: ET, GPP, LAI, Ts and three soil properties datasets. All the variables were resampled to 1 km spatial resolution by using Nearest algorithm[46].
(2) Using the 1 km National Land Cover Dataset (NLCD) to mask the seven input variables.
(3) Using the 1 km ChinaCropPheno dataset to calculate the cumulative value of ET, GPP and Ts and the averaged value of LAI from the V3 stage of maize (emergency or green up stage of wheat) to maturity stage.
(4) Statistic the seven indicators processed in (2) and (3) to county-level to match the annual crop yield from National Bureau of Statistics of China (NBSC).
(5) Using RF to fit the seven indicators in county-level with the crop yield. In which, the 80% of the county-level maize yield samples were randomly selected for training the model estimates of yield, to ensure reliability, and the remaining 20% of samples were used to validate accuracy of the estimates. Model training data should contain maximum and minimum yield values. Given temperature[47], GPP[48], LAI[49], and ET[50] affect crop yield, they were input to the model individually and in combination, with effects of soil clay, sand, and silt content held as constant, to compare levels of accuracy of yield estimates and build the optimal model[46].
(6) After optimal model training for yield estimation had been completed, the input indicators at pixel-level resolution (processed in (2) and (3)) were directly input to generate pixel-level annual crop yield datasets, at a spatial resolution of 1 km. Using the point-scale crop yield data derived from the National Meteorological Data Center of China to assess the generated dataset. See Fig. 3 for workflow of data preprocessing, model construction, and generation of datasets.

Crop water productivity definition. We defined CWP (kg/m^3) of maize and wheat as the ratio of yield to cumulative ET (Eq. 1):

$$CWP = \frac{Yield}{\sum ET} \quad (1)$$

where crop yield (kg/ha) was estimated by the proposed model; cumulative ET (mm) is across the main crop growth stage. In terms of the spatial difference of crop phenology, the cumulative ET was calculated using the ET from V3 stage of maize (emergency or green up stage of wheat) to maturity, which is the main period of crop growth stage. Therefore, it should be the cumulative ET in this study will less than other studies which were calculated in the whole crop growth period[4].

Random forest algorithm. Random Forest (RF) regression algorithm is widely used ensemble learning method by combining multiple decision trees, where each regression tree represents a set of restrictions or conditions on indicators of the target variable; in this study, the variable is county-level crop yield. The RF algorithm begins with subsamples randomly selected from the training set, and then the regression tree is fitted to the subsamples; the final modeled value is the average across all trees. The details of RF can be referred to the study of Breiman[51]. In this study, the two important parameters: tree numbers and the randomly sampled potential variables in each split, were set as 100 and 4 by debugging and referring other studies[52].

The RF algorithm has been shown to be effective in coping with over-fitting[53], performs well in multiple regressions, and has been widely used in the analysis of RS data[32,35,52,54,55].

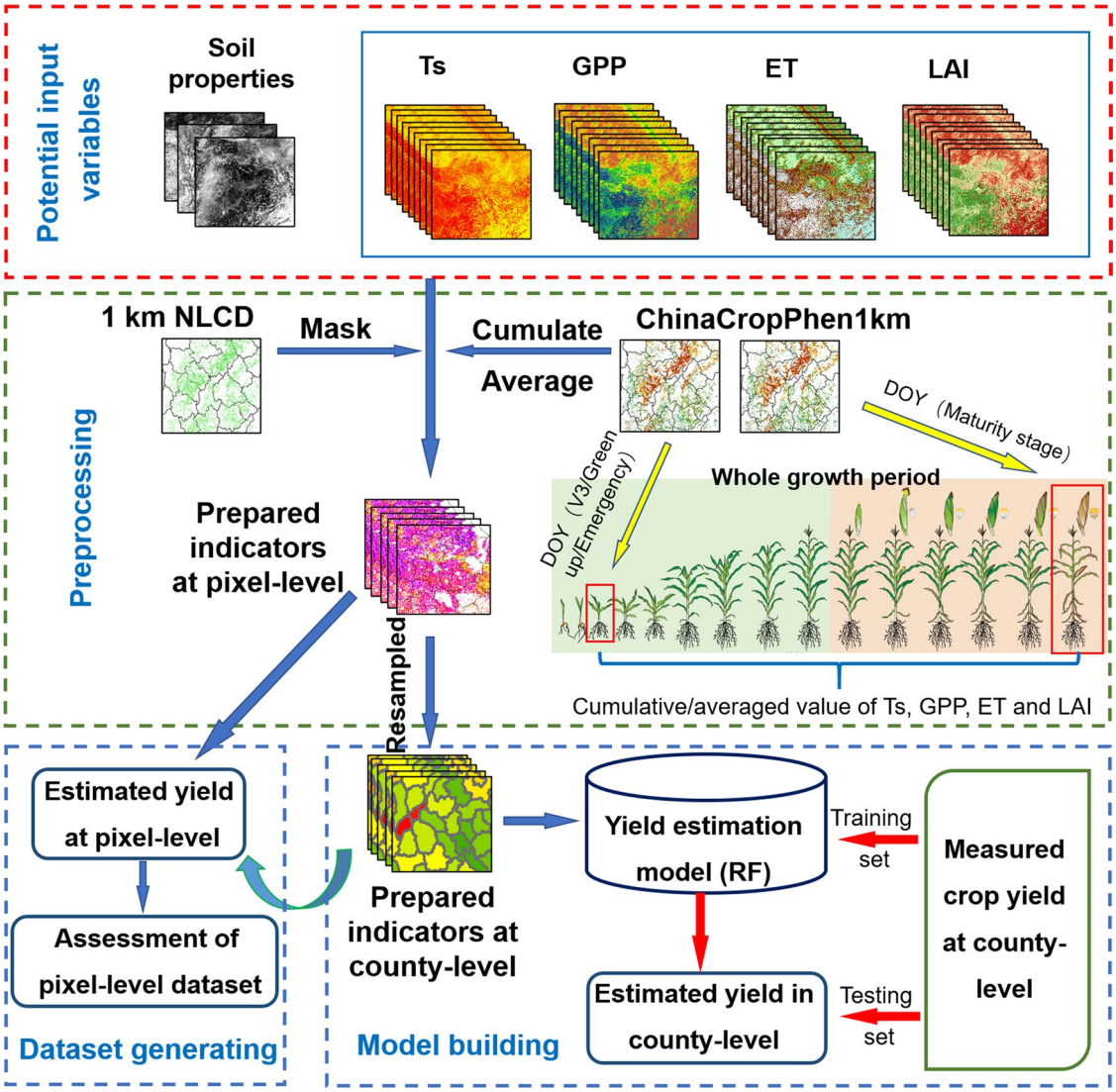

Fig. 3 Schematic of data preprocessing, model construction, and generation of datasets for estimation of maize and wheat yields using RF and yield indicators.

Assessment of model input and output accuracy. *Evapotranspiration dataset.* The EC method of estimating ET measures λET (latent heat flux) from covariance in heat and moisture fluxes, with vertical velocity using rapid response sensors at frequencies typically equal to or greater than 10 Hz, and is regarded as the most effective method for the estimation of ET[10]. The energy balance closure issue, which indicates the sum of sensible heat (H), λET and soil heat flux (G), is not equal to net radiation (Rn), is frequently found in the EC method, so values measured using this system value should be filtered and corrected. Here, data with energy balance closure ratios (ECR, Eq. 2) <80% were not selected for validation[56] and the remaining data with ECR >80% were corrected using the Bowen ratio energy balance correction (Eq. 3)[57].

$$ECR = \frac{H + \lambda ET}{Rn - G} \tag{2}$$

$$\lambda ET_{cor} = \frac{Rn - G}{H + \lambda ET} \times \lambda ET \tag{3}$$

where Rn, G, H and λET are values measured using the EC system, and λET_{cor} is the corrected value. To ensure reliable evaluation, the pixel value at the flux tower location (area: 1 × 1 km) was extracted for comparison with the measured value[19].

Estimated yield. We used county-level empirical yield data in the model for yield estimation, where 20% of the samples (maize N = 396; wheat N = 497) were used for regional-scale validation of crop yields and empirical

pixel-level yield data, obtained from the 12 maize, 23 wheat, and 42 mixed sites, were used to validate estimated yields at the point-scale. Each yield measurement site comprised data recorded over one or multiple years, and overall, our dataset comprised 63 maize and 103 wheat yield samples at the point-scale; pixel values (1 km) of estimated crop yields at these measurement sites were directly compared with their corresponding measured values.

Model performance. We calculated the adjusted coefficient of determination (R^2), root-mean-square error (RMSE), relative root-mean-square error (rRMSE), and mean bias error (MBE), following Jin *et al.* (2020), to quantify model performance:

$$R^2 = 1 - \left(1 - \frac{\sum_{i=1}^{n}(M_i - \overline{M})^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}\right)\frac{n-1}{n-m-1} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_i - O_i)^2} \tag{5}$$

$$rRMSE = \frac{RMSE}{\overline{O}} \times 100\% \tag{6}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(M_i - O_i) \tag{7}$$

where M and O are the estimated and recorded/measured value (ET or yield), respectively, n is the number of samples, and m is the number of variables.

Spatial autocorrelation analysis. Spatial patterns of crop yield are affected by spatiotemporal variations in soil properties, climate, land-use change, diseases, and management practices[58], so heterogeneity and dependency of crop yield may similarly vary spatially, particularly over large areas[35]. While assumptions of location invariance and spatial independence have been applied to yield estimates[59,60], they may lead to inaccurate model estimates without spatial variation and autocorrelation analysis[58]. To cope with this issue, we used Global Moran's I (Moran[61], which ranges from −1 to 1, to examine spatial autocorrelations between model yield estimate errors[35,62] that were calculated as the difference between estimated and measured yields at the county level. Global Moran's I represents the spatial autocorrelation of errors in estimates of yield or the degree of clustering[63] and it has been used widely in the evaluation of model spatial performance[64,65]. In this study, a Global Moran's I of zero indicates a random spatial distribution, while a near zero value indicates that errors in the estimates of yield were randomly distributed, where higher randomness tends to indicate better model performance over space. Global Moran's I was calculated as follows:

$$I = \frac{n \times \sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}(x_i - \overline{x})(x_j - \overline{x})}{S \times \sum_{i=1}^{n}(x_i - \overline{x})} \tag{8}$$

where n is number of counties; ω_{ij} is the weight matrix between counties i and j, with a value of 1 or 0 when the two counties are adjacent or nonadjacent, respectively; x_i and x_j are the difference between estimated yield and recorded yield of counties i and j, respectively; and, S is the sum of ω_{ij}.

Model performance, based on R^2, rRMSE and Moran's I across input single and combined indicators, was tested using one-way analysis of variance (ANOVA) at $P < 0.01$ in SPSS (Version 21, IBM Corp., Armonk, US). Similarly, differences in crop yield and CWP among the eight agricultural production regions were tested using ANOVA.

Data Records

The dataset that was generated using random forest regression and multiple remotely sensing indicators, at a spatial resolution of 1 km and a yearly temporal resolution, which can be used for optimizing agricultural production strategies and water resources management, etc. The crop yield and water productivity dataset for China is distributed under a Creative Commons Attribution 4.0 International license. The dataset is named ChinaCYWP and consists of 15 years of data, with the format of TIF. More information and data are freely available from the Zenodo repository at https://doi.org/10.5281/zenodo.5121842[66].

Technical Validation

Validation of evapotranspiration dataset. Crop rotations at the five EC flux measurement stations comprised maize-wheat rotations, and we used the EC estimates of ET to validate MOD16 estimates of ET (Fig. 4). For maize, MOD16 estimates of ET varied from 4.18 to 27.51 mm/8d ($R^2 = 0.73$; RMSE = 4.42 mm/8d), while for wheat, ET estimates varied from 1.39 to 26.32 mm/8d ($R^2 = 0.74$; RMSE = 3.81 mm/8d). In general, MOD16 estimates of crop ET were lower than observed EC estimates of ET (maize MBE = −0.99 mm/8d; wheat MBE = −0.68 mm/8d).

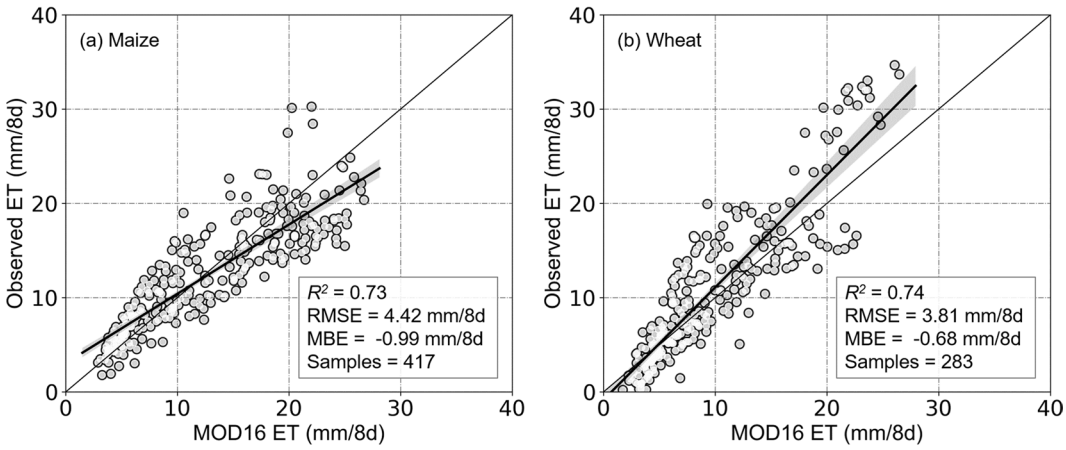

Fig. 4 Validation of MOD16 ET products for (**a**) maize and (**b**) wheat. Note: R^2 indicates coefficient of determination, RMSE indicates root-mean-square error and MBE indicates mean bias error.

In addition to the MOD16 ET product, several other ET products, such as Global Land Evaporation Amsterdam Model, GLEAM[67], Global Land Data Assimilation System, GLDAS[68], and Evapotranspiration-Energy Balance, ET-EB[69] products, generated by different algorithms have been evaluated in previous studies[19,70,71]. Algorithms for the estimation of RS ET tend to be complementary, with contrasting strengths and weaknesses[72]; for example, the spatiotemporal resolution (500 m and 8-d composite) of MOD16 is finer than other ET products, including GLEAM (0.25° and daily), GLDAS (0.25° and monthly), and ET-EB (0.1° and daily), and is more appropriate for the generation of crop yield and CWP data at 1-km spatial resolution. As a result, we found that MOD16 yielded an acceptable level of accuracy for describing the ET of maize and wheat. Previous research has also demonstrated the greater estimate accuracy of MOD16 products, including Velpuri, *et al.*[19], who concluded that accuracy of MOD16 for estimates of cropland flux tower data was greater than that of SSEBop, while Khan, *et al.*[73] similarly found that accuracy of MOD16 in cropland was greater (bias: 0.22 mm/8 d) than that of GLDAS and GLEAM (4.32 and 5.35 mm/8d, respectively). Although validation of flux tower data represent a useful method for ET measurement[10], uncertainties remain, including large error size (10–30%) in eddy covariance flux tower data[70,74] and mismatches between flux tower footprint and RS information caused by effects of wind direction, atmospheric stability, and surface type[75].

Validation of model yield estimates. *Regional-scale.* In general, the accuracy of maize and wheat yield estimates improved with increasing number of input indicators, with four indicators accounting for the greatest amount of variation in yield estimates (maize $R^2 = 0.80$, rRMSE = 15.36%; wheat $R^2 = 0.66$, rRMSE = 17.17%), and while there were no differences in R^2 and rRMSE indicators of model estimates between the two crops ($P < 0.01$), RMSE for maize (1025−1958 kg/ha) was larger than for wheat (845−1166 kg/ha) ($P < 0.01$) (Fig. 5). In general, Moran's I decreased with increasing number of indicators included in the model (i.e., better spatial applicability), where it was lowest for maize with the inclusion of four indicators ($I = 0.16$) and lowest in wheat when ET, LAI, and Ts were included ($I = 0.13$) (Fig. 5).

Overall, inclusion of four indicators led to best estimates of maize ($R^2 = 0.80$; rRMSE = 15.36%) and wheat ($R^2 = 0.66$; rRMSE = 17.17%) yields (Fig. 6). Thus, the pixel-level crop yield dataset was generated using the four indicators.

Point-scale. We found pixel-scale estimates of maize and wheat yields, based on point-scale yield data, were similar (maize: $R^2 = 0.65$, RMSE = 2144.75 kg/ha, rRMSE = 26.81%; wheat: $R^2 = 0.51$, RMSE = 1119.22 kg/ha, rRMSE = 21.80%), while model performance was less accurate than for regional-scale estimates, with underestimates (MBE) of maize and wheat crop yield, compared with empirical data, of −928.91 and −275.10 kg/ha, respectively (Fig. 7).

Summary. Approaches for crop yield estimation based on RS data[29,32,35,37,76,77] tend to use single or multi-phase RS images to describe crop growth status and estimate yield; for example, Maimaitijiang, *et al.*[35] used single-phase UAV images (multi-sensors) at the start of the pod stage of soybean to estimate yield. However, given the status of each stage of the entire growth period may contribute to crop final yield, phenological information, such as that provided by crop growth stage indicators, is likely to be essential for accurate crop yield estimation. Indeed, Guo, *et al.*[78] found the inclusion of phenology and climate data led to more accurate model estimates of rice yield in China ($R^2 = 0.33$ and RMSE = 737 kg/ha). Remotely sensed data for yield estimation tends to be based on VIs, such as in the studies by Cao, *et al.*[32] and Chen, *et al.*[77], who used RS normalized difference vegetation index (NDVI) and a combination of NDVI, enhanced vegetation index (EVI), and soil adjusted vegetation index (SAVI), respectively, to estimate maize yield in China. Although physiological indicators of crop growth, such as GPP and ET, correlate with crop yield[48,50,79], characterization of crop growth status by VIs may be limited, whereas relative indicators of temperature, such as growing degree days and effective accumulated temperature (EAT), have been shown to be associated with crop growth status and yield[80–82]. Of the single

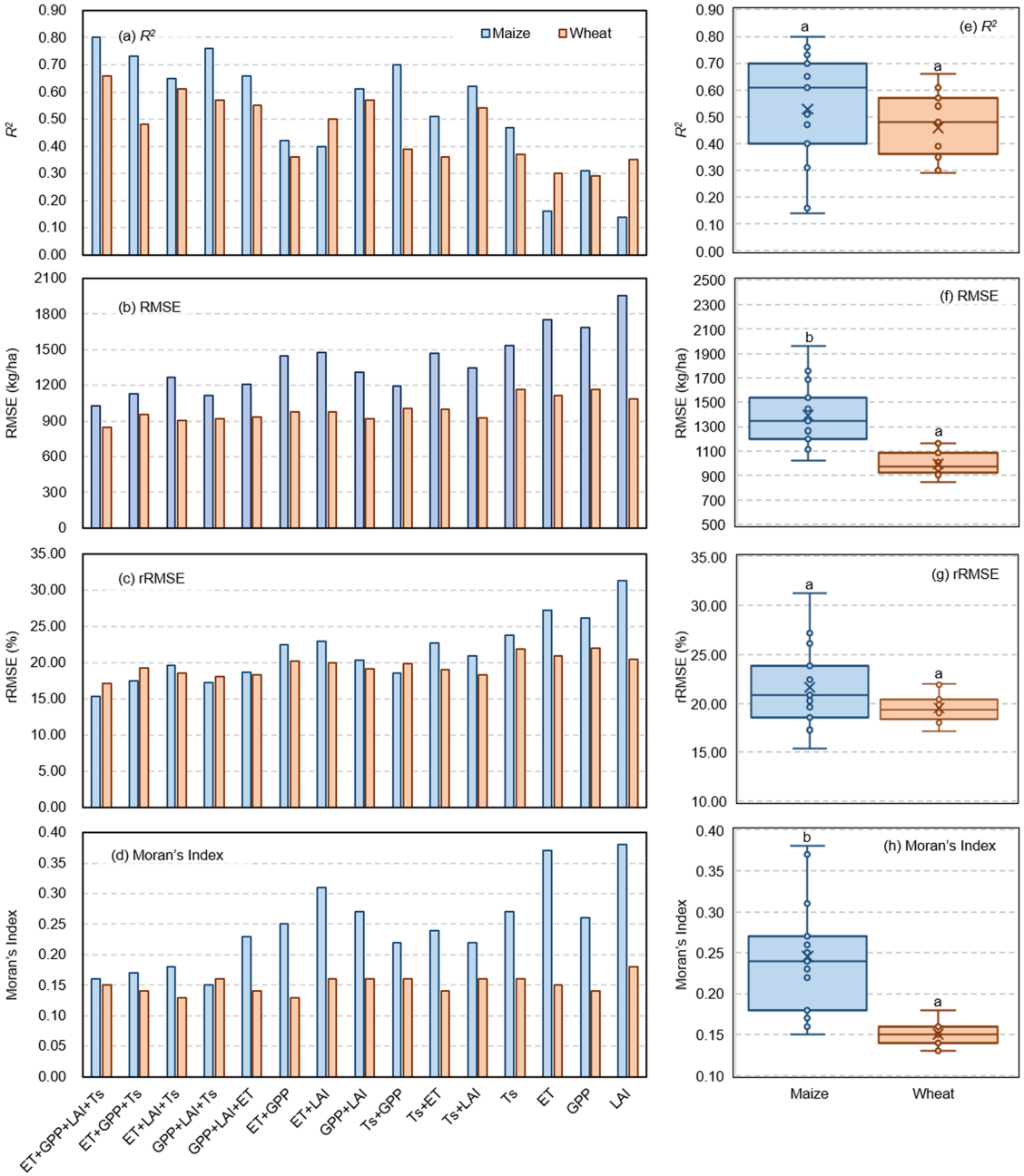

Fig. 5 Model estimates of crop yield based on inclusion of single and combined indicators: histogram of (**a**) R^2, (**b**) RMSE, (**c**) rRMSE, and (**d**) Moran's *I*; distribution of median and range (±95% CI) of (**e**) R^2, (**f**) RMSE, (**g**) rRMSE, and (**h**) Moran's *I*. Different letters indicate differences in accuracy of crop model yield estimates at $P < 0.01$.

indicators used in this study, we found that cumulative Ts, which may be regarded as EAT without threshold filtering, explained most of the variation in maize yield (Fig. 5); in contrast, Maimaitijiang, *et al.*[35] found that Ts were poor predictors of soybean yield, possibly due to the use of single-phase images.

In order to further explore the influence by the accuracy of the input indicators to model performance, a sensitive analysis was conducted by taking the maize yield estimation as an example, i.e., a random error was artificially set in each indicator or multi-indicators, and the changes in performance were analyzed. The sensitive analysis method was referred to Cheng, *et al.*[39] and Long, *et al.*[83]. The results were showed in Fig. 8. In general, the model still performed good ($R^2 > 0.62$ and rRMSE < 20%) when only one indicator had errors, even if a random error between 0 to 40% (−40% to 0) was set. The model results changed the most when the errors were existed in Ts. But these differences among the four indicators was small. However, when the four indicators all had errors, the model performance changed a lot. The R^2 was decreased to 0.30 when random errors of 0 to 40% were existed in the four indicators and rRMSE was increased to 28.12% when random errors of −40 to 0 were existed, which were the worst situation. As reported in previous studies, MODIS products have errors to

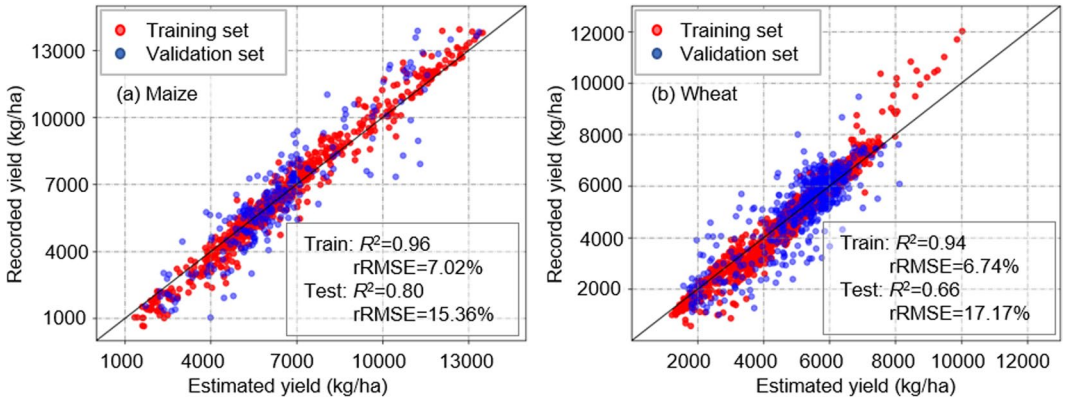

Fig. 6 Regional-scale validation of estimated maize (**a**) and wheat (**b**) yields based on model inclusion of GPP, ET, Ts, and LAI. Note: R^2 indicates coefficient of determination and rRMSE indicates relative root-mean-square error.

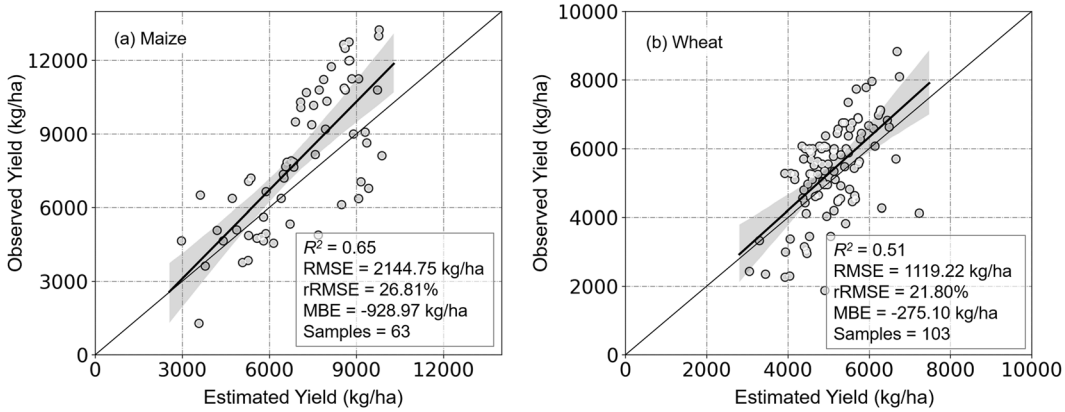

Fig. 7 Point-scale validation of estimated maize (**a**) and wheat (**b**) yields based on model inclusion of GPP, ET, Ts, and LAI.

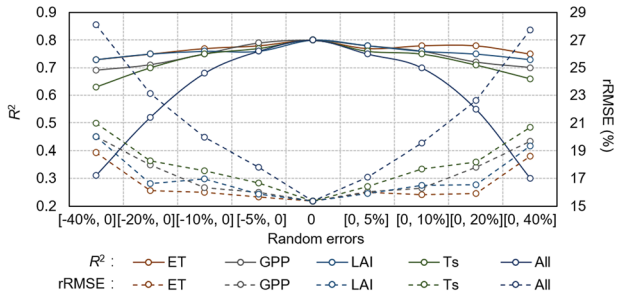

Fig. 8 Sensitive analysis of the effect of input variables to yield prediction.

different extents. For example, MOD16 ET product showed approximately 15–30% errors in China[39]. MOD17 GPP product has been evaluated by Liu, *et al.*[84] and showed R^2 varied from 0.21 to 0.90 in China. Be that as it may, the proposed method still performed an acceptable robustness and tolerance when confronted to the uncertainties of indicators accuracy. Which was likely contributed by the correlations among indicators, i.e., when the information of a specific indicator was loss caused by the accuracy errors, the other indicators which have strong correlation, may fill this information gap.

Overall, our proposed model for estimation of maize and wheat yields performed with good accuracy at county-level (rRMSE: 15.36 and 17.17%, respectively) and pixel-level validation (rRMSE: 26.81 and 21.80%, respectively). These levels of accuracy are comparable to, or greater than previous studies[29,32,77] and, although the accuracy of the yield estimates improved with increasing number of input indicators, we found the accuracy of wheat yield estimates was lower than that for maize, possibly as a result of duplicated information among some indicators. We note a lower performance of model estimates of maize and wheat yield performance at the

pixel-level than county-level, possibly due to model training by county-level yield data and potential differences in data measurement protocols.

Many scholars have made efforts to estimate CWP. Bastiaanssen and Steduto[4] estimated the average value of global maize CWP by using WATPRO model as $2.25 \pm 0.94\,kg/m^3$; Edreira, *et al.*[85] estimated that the CWP of maize in Africa was $1.8\,kg/m^3$ and that in Europe was $2.9\,kg/m^3$ by using meteorological data and crop models. Li, *et al.*[86] estimated the CWP of maize in Hetao irrigated area as $2.59–3.34\,kg/m^3$ by using the AquaCrop model. In comparing, the CWP estimated in this study presented relative higher than others (4.14 ± 1.62 and $4.78 \pm 2.43\,kg/m^3$ for maize and wheat, respectively), three causes were discussed as follows: (1) as proved in Section 4.1, MOD16 presented a certain underestimation of crop ET, in which, MBE was −0.99 mm/8d for maize and −0.68 mm/8d for wheat; (2) the cumulative ET of the crop growth period in this study was calculated using the ET from V3 stage of maize (emergency or green up stage of wheat) to maturity stage, which was shorter than the whole crop growth period The short time period also caused the lower accumulated ET; (3) this study was conducted covering whole China planting area of maize and wheat, including rainfed and spring maize planting area, which lead the lower ET than irrigated area and summer maize planting area[85]. In general, lower ET estimation caused the higher CWP. Despite all this, the CWP dataset generated in this study presented a certain accuracy and comparability of spatial and temporal.

Although we found that maize and wheat ET and yield were good predictors of observed CWP, direct verification of RS CWP is difficult[1], because *in situ* benchmark values for CWP tend not to be available[4]; however, given some calculations of CWP have been based on GPP, rather than crop yield, it is possible to directly evaluate estimates using EC flux tower observations[56,87]. Even though we found separate validation of the two CWP components to be acceptable, the uncertainties from error propagation should not be ignored and we recommend further studies to identify improved methods for the validation of gridded CWP datasets.

Code availability

The codes we developed for crop yield computation and crop yield dataset generation are available at https://doi.org/10.5281/zenodo.6444614[88]. The code was programmed using Python 3.9. In this code, we used the sklearn library for calling machine learning algorithm and GDAL library for raster data reading and processing. Moreover, the band calculation tool of ArcGIS 10.4 was used for crop water productivity dataset generation.

Received: 11 April 2022; Accepted: 6 October 2022;
Published online: 21 October 2022

References

1. Blatchford, M. L., Mannaerts, C. M., Zeng, Y., Nouri, H. & Karimi, P. Status of accuracy in remotely sensed and *in-situ* agricultural water productivity estimates: A review. *Remote Sensing of Environment* **234**, 111413, https://doi.org/10.1016/j.rse.2019.111413 (2019).
2. Geerts, S. & Raes, D. Deficit irrigation as an on-farm strategy to maximize crop water productivity in dry areas. *Agricultural Water Management* **96**, 1275–1284, https://doi.org/10.1016/j.agwat.2009.04.009 (2009).
3. Hellegers, P., Soppe, R., Perry, C. & Bastiaanssen, W. Combining remote sensing and economic analysis to support decisions that affect water productivity. *Irrigation Science* **27**, 243–251, https://doi.org/10.1007/s00271-008-0139-7 (2009).
4. Bastiaanssen, W. G. M. & Steduto, P. The water productivity score (WPS) at global and regional level: Methodology and first results from remote sensing measurements of wheat, rice and maize. *The Science of the total environment* **575**, https://doi.org/10.1016/j.scitotenv.2016.09.032 (2017).
5. Seneviratne, S. I. *et al.* Investigating soil moisture–climate interactions in a changing climate: A review. *Earth Science Reviews* **99**, https://doi.org/10.1016/j.earscirev.2010.02.004 (2010).
6. Hu, X., Shi, L., Lin, L. & Zha, Y. Nonlinear boundaries of land surface temperature–vegetation index space to estimate water deficit index and evaporation fraction. *Agricultural and Forest Meteorology* **279**, https://doi.org/10.1016/j.agrformet.2019.107736 (2019).
7. Bowen, I. S. The Ratio of Heat Losses by Conduction and by Evaporation from any Water Surface. *Physical Review* **27**, 779–787, https://doi.org/10.1103/PhysRev.27.779 (1926).
8. Penman, H. L. Natural evaporation from open water, hare soil and grass. *Proceedings of the Royal Society of London. Series A, Mathematical and physical sciences* **193**, https://doi.org/10.1098/rspa.1948.0037 (1948).
9. Monteith, J. L. Evaporation and environment. The stage and movement of water in living organisms. *Symp.soc.exp.biol.the Company of Biologists* (1965).
10. Wang, K. & Dickinson, R. E. A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. *Reviews of Geophysics* **50**, https://doi.org/10.1029/2011RG000373 (2012).
11. Bastiaanssen, W. G. *et al.* A remote sensing surface energy balance algorithm for land (SEBAL) Part 1: Fomulation. *Journal of hydrology* **212**, 213–229, https://doi.org/10.1016/S0022-1694(98)00253-4 (1998).
12. Bastiaanssen, W. G. M. *et al.* A remote sensing surface energy balance algorithm for land (SEBAL) Part 2. Validation. *Journal of Hydrology* **212**, https://doi.org/10.1016/S0022-1694(98)00254-6 (1998).
13. Su, Z. The Surface Energy Balance System (SEBS) for estimation of turbulent heat fluxes. *Hydrology and Earth System Science* **6**, 85–99, https://doi.org/10.5194/hess-6-85-2002 (2002).
14. Norman, J. M., Kustas, W. P. & Humes, K. S. Source approach for estimating soil and vegetation energy fluxes in observations of directional radiometric surface temperature. *Agricultural and Forest Meteorology* **77**, https://doi.org/10.1016/0168-1923(95)02265-y (1995).
15. Mu, Q., Heinsch, F. A., Zhao, M. & Running, S. W. Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. *Remote Sensing of Environment* **111**, https://doi.org/10.1016/j.rse.2007.04.015 (2007).
16. Mu, Q., Zhao, M. & Running, S. W. Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment* **115**, 1781–1800, https://doi.org/10.1016/j.rse.2011.02.019 (2011).
17. Fisher, J. B., Tu, K. P. & Baldocchi, D. D. Global estimates of the land–atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. *Remote Sensing of Environment* **112**, 901–919, https://doi.org/10.1016/j.rse.2007.06.025 (2008).
18. Kim, H. W., Hwang, K., Mu, Q., Lee, S. O. & Choi, M. Validation of MODIS 16 global terrestrial evapotranspiration products in various climates and land cover types in Asia. *KSCE Journal of Civil Engineering* **16**, https://doi.org/10.1007/s12205-012-0006-1 (2012).
19. Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S. & Verdin, J. P. A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sensing of Environment* **139**, https://doi.org/10.1016/j.rse.2013.07.013 (2013).

20. Jin, X. *et al*. Estimation of water productivity in winter wheat using the AquaCrop model with field hyperspectral data. *Precision Agriculture* **19**, 1–17, https://doi.org/10.1007/s11119-016-9469-2 (2016).
21. Felix, R., Clement, A., Igor, S. & Oscar, R. Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection. *Remote Sensing* **5**, 1704–1733, https://doi.org/10.3390/rs5041704 (2013).
22. Lu, Y. *et al*. Assimilation of soil moisture and canopy cover data improves maize simulation using an under-calibrated crop model. *Agricultural Water Management* **252**, https://doi.org/10.1016/j.agwat.2021.106884 (2021).
23. Jin, X., Kumar, L., Li, Z., Feng, H. & Wang, J. A review of data assimilation of remote sensing and crop models. *European Journal of Agronomy* **92**, https://doi.org/10.1016/j.eja.2017.11.002 (2018).
24. Weiss, M., Jacob, F. & Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment* **236**, https://doi.org/10.1016/j.rse.2019.111402 (2019).
25. Jin, X. *et al*. Winter wheat yield estimation based on multi-source medium resolution optical and radar imaging data and the AquaCrop model using the particle swarm optimization algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing* **126**, 24–37 (2017).
26. Tao, F., Rötter, R. P., Palosuo, T., Díaz-Ambrona, C. G. H. & Schulman, A. H. Contribution of crop model structure, parameters and climate projections to uncertainty in climate change impact assessments. *Global Change Biology* **24**, https://doi.org/10.1111/gcb.14019 (2017).
27. Jin, X. *et al*. A review of data assimilation of remote sensing and crop models. *European Journal of Agronomy* **92**, 141–152, https://doi.org/10.1016/j.eja.2017.11.002 (2018).
28. Anikó, K. *et al*. Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. *Agricultural and Forest Meteorology* **260-261**, 300–320, https://doi.org/10.1016/j.agrformet.2018.06.009 (2018).
29. Wang, Y., Zhang, Z., Feng, L., Du, Q. & Runge, T. Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States. *Remote Sensing* **12**, 1232, https://doi.org/10.3390/rs12081232 (2020).
30. Franz, T. E. *et al*. The role of topography, soil, and remotely sensed vegetation condition towards predicting crop yield. *Field Crops Research* **252**, https://doi.org/10.1016/j.fcr.2020.107788 (2020).
31. Noland, R. L. *et al*. Estimating alfalfa yield and nutritive value using remote sensing and air temperature. *Field Crops Research* **222**, 189–196, https://doi.org/10.1016/j.fcr.2018.01.017 (2018).
32. Cao, J., Zhang, Z., Luo, Y., Zhang, L. & Tao, F. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. *European Journal of Agronomy*, 126204, https://doi.org/10.1016/j.eja.2020.126204 (2021).
33. Jacinta, H. & Kerrie, M. Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review. *Remote Sensing* **10**, 1365, https://doi.org/10.3390/rs10091365 (2018).
34. Jin, X., Liu, S., Baret, F., Hemerlé, M. & Comar, A. Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery. *Remote Sensing of Environment* **198**, 105–114, https://doi.org/10.1016/j.rse.2017.06.007 (2017).
35. Maimaitijiang, M. *et al*. Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment* **237**, 111599, https://doi.org/10.1016/j.rse.2019.111599 (2020).
36. Hossein, A., Mohsen, A., Davoud, A., Salehi, S. H. & Soheil, R. Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI. *IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing* **PP**, 1–15, https://doi.org/10.1109/JSTARS.2018.2823361 (2018).
37. Johansen, K. *et al*. Predicting Biomass and Yield in a Tomato Phenotyping Experiment Using UAV Imagery and Random Forest. *Frontiers in Artificial Intelligence* **3**, 28, https://doi.org/10.3389/frai.2020.00028 (2020).
38. Zhang, L., Ding, X., Shen, Y., Wang, Z. & Wang, X. Spatial Heterogeneity and Influencing Factors of Agricultural Water Use Efficiency in China. *Resources and Environment in the Yangtze Basin* **28**, https://doi.org/10.11870/cjlyzyyhj201904008 (2019).
39. Cheng, M. *et al*. Satellite time series data reveal interannual and seasonal spatiotemporal evapotranspiration patterns in China in response to effect factors. *Agric. Water Manage*. **255**, https://doi.org/10.1016/j.agwat.2021.107046 (2021).
40. Zhou, L. *Comprehensive agricultural regionalization in China*. (Agricultural Press of China, 1985).
41. Luo, Y., Zhang, Z., Chen, Y., Li, Z. & Tao, F. ChinaCropPhen1km: A high-resolution crop phenological dataset for three staple crops in China during 2000-2015 based on LAI products. *Figshare* https://doi.org/10.6084/m9.figshare.8313530.v6 (2019).
42. Luo, Y., Zhang, Z., Chen, Y., Li, Z. & 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 System Science Data* **12**, 197–214, https://doi.org/10.5194/essd-12-197-2020 (2020).
43. Song, D. *Second China Soil Survey*. (Chinese Science Press, 1979).
44. Zhang, T., Yang, X., Wang, H., Li, Y. & Ye, Q. Climatic and technological ceilings for Chinese rice stagnation based on yield gaps and yield trend pattern analysis. *Global Change Biology* **20**, 1289–1298, https://doi.org/10.1111/gcb.12428 (2014).
45. Chen, Y., Zhang, Z. & Tao, F. Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data. *European Journal of Agronomy* **101**, 163–173, https://doi.org/10.1016/j.eja.2018.09.006 (2018).
46. Cheng, M. *et al*. Combining multi-indicators with machine-learning algorithms for maize yield early prediction at the county-level in China. *Agricultural and Forest Meteorology* **323**, https://doi.org/10.1016/j.agrformet.2022.109057 (2022).
47. Amir, J. & Sinclair, T. A model of the temperature and solar-radiation effects on spring wheat growth and yield. *Field Crops Research* **28**, 47–58, https://doi.org/10.1016/0378-4290(91)90073-5 (1991).
48. Prince, S. D., Haskett, J., Steininger, M. & Wright, S. R. Net Primary Production of U.S. Midwest Croplands from Agricultural Harvest Yield Data. *Ecological Applications* **11**, 1194–1205, https://doi.org/10.1890/1051-0761(2001)011[1194:NPPOUS]2.0.CO;2 (2001).
49. Gilardelli, C. *et al*. Downscaling rice yield simulation at sub-field scale using remotely sensed LAI data. *European journal of agronomy* **103**, 108–116, https://doi.org/10.1016/j.eja.2018.12.003 (2019).
50. Shakoor, R., Hassan, M. Y., Raheem, A. & Wu, Y.-K. Wake effect modeling: A review of wind farm layout optimization using Jensen′ s model. *Renewable and Sustainable Energy Reviews* **58**, 1048–1059, https://doi.org/10.1016/j.rser.2015.12.229 (2016).
51. Breiman, L. Random Forests. *Machine Learning* https://doi.org/10.1023/A:1010933404324 (2001).
52. Li, L. *et al*. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. *Agricultural and Forest Meteorology* 308–309, https://doi.org/10.1016/j.agrformet.2021.108558 (2021).
53. Wang, L. A., Zhou, X., Zhu, X., Dong, Z. & Guo, W. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal* **4**, 212–219, https://doi.org/10.1016/j.cj.2016.01.008 (2016).
54. Feng, P. *et al*. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agricultural and Forest Meteorology* **285-286**, 107922, https://doi.org/10.1016/j.agrformet.2020.107922 (2020).
55. Lu, F., Sun, Y. & Hou, F. Using UAV Visible Images to Estimate the Soil Moisture of Steppe. *Water* **12**, 2334, https://doi.org/10.3390/w12092334 (2020).
56. Wang, S. *et al*. High spatial resolution monitoring land surface energy, water and CO2 fluxes from an Unmanned Aerial System. *Remote Sensing of Environment* **229**, 14–31, https://doi.org/10.1016/j.rse.2019.03.040 (2019).
57. Chen, Y. *et al*. Comparison of satellite-based evapotranspiration models over terrestrial ecosystems in China. *Remote Sensing of Environment* **140**, 279–293, https://doi.org/10.1016/j.rse.2013.08.045 (2014).
58. Peralta, N., Assefa, Y., Du, J., Barden, C. & Ciampitti, I. Mid-Season High-Resolution Satellite Imagery for Forecasting Site-Specific Corn Yield. *Remote Sensing* **8**, 848, https://doi.org/10.3390/rs8100848 (2016).
59. Russello, H. Convolutional neural networks for crop yield prediction using satellite images. *IBM Center for Advanced Studies* (2018).
60. You, J., Li, X., Low, M., Lobell, D. & Ermon, S. in *Proceedings of the AAAI Conference on Artificial Intelligence*.

61. Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
62. Imran, M., Stein, A. & Zurita-Milla, R. Using geographically weighted regression kriging for crop yield mapping in West Africa. *International Journal of Geographical Information Systems* **29**, 234–257, https://doi.org/10.1080/13658816.2014.959522 (2015).
63. Harries, K. Extreme spatial variations in crime density in Baltimore County, MD. *Geoforum* **37**, 404–416, https://doi.org/10.1016/j.geoforum.2005.09.004 (2006).
64. Ghulam, A. *et al*. Remote Sensing Based Spatial Statistics to Document Tropical Rainforest Transition Pathways. *Remote Sensing* **7**, 6257–6279, https://doi.org/10.3390/rs70506257 (2015).
65. Maimaitijiang, M., Ghulam, A., Sandoval, J. S. O. & Maimaitiyiming, M. Drivers of land cover and land use changes in St. Louis metropolitan area over the past 40 years characterized by remote sensing and census population data. *International Journal of Applied Earth Observation Geoinformation* **35**, 161–174, https://doi.org/10.1016/j.jag.2014.08.020 (2015).
66. Cheng, M. Long time series (2001-2015) high-resolution crop yield and water productivity dataset of China, *Zenodo*, https://doi.org/10.5281/zenodo.5121842 (2021).
67. Martens, B., Miralles, D. G., Lievens, H., Schalie, R. D. & Verhoest, N. GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development* **10**, https://doi.org/10.5194/gmd-10-1903-2017 (2017).
68. Wang, W., Cui, W., Wang, X. & Chen, X. Evaluation of GLDAS-1 and GLDAS-2 forcing data and Noah model simulations over China at the monthly scale. *Journal of Hydrometeorology* **17**, 2815–2833, https://doi.org/10.1175/JHM-D-15-0191.1 (2016).
69. Chen, X. *et al*. Development of a 10-year (2001–2010) 0.1° data set of land-surface energy balance for mainland China. *Atmospheric Chemistry and Physics* **14**, 14471–14518, https://doi.org/10.5194/acp-14-13097-2014 (2014).
70. Ramoelo, A. *et al*. Validation of Global Evapotranspiration Product (MOD16) using Flux Tower Data in the African Savanna, South Africa. *Remote Sensing* **6**, https://doi.org/10.3390/rs6087406 (2014).
71. Yang, X., Yong, B., Ren, L., Zhang, Y. & Long, D. Multi-scale validation of GLEAM evapotranspiration products over China via ChinaFLUX ET measurements. *International Journal of Remote Sensing* https://doi.org/10.1080/01431161.2017.1346400 (2017).
72. Hu, G., Jia, L. & Menenti, M. Comparison of MOD16 and LSA-SAF MSG evapotranspiration products over Europe for 2011. *Remote Sensing of Environment* **156**, 510–526, https://doi.org/10.1016/j.rse.2014.10.017 (2015).
73. Khan, M. S., Liaqat, U. W., Baik, J. & Choi, M. Stand-alone uncertainty characterization of GLEAM, GLDAS and MOD16 evapotranspiration products using an extended triple collocation approach. *Agricultural and Forest Meteorology* **252**, 256–268, https://doi.org/10.1016/j.agrformet.2018.01.022 (2018).
74. Glenn, E. P. *et al*. Scaling sap flux measurements of grazed and ungrazed shrub communities with fine and coarse-resolution remote sensing. *Ecohydrology* **1**, 316–329, https://doi.org/10.1002/eco.19 (2008).
75. Gamon, J. A. Reviews and Syntheses: optical sampling of the flux tower footprint. *Biogeosciences* **12**, 4509–4523, https://doi.org/10.5194/bg-12-4509-2015 (2015).
76. Cai, Y. *et al*. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology* **274**, 144–159, https://doi.org/10.1016/j.agrformet.2019.03.010 (2019).
77. Chen, X. *et al*. Prediction of Maize Yield at the City Level in China Using Multi-Source Data. *Remote Sensing* **13**, https://doi.org/10.3390/rs13010146 (2021).
78. Guo, Y. *et al*. Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecological Indicators* **120**, 106935, https://doi.org/10.1016/j.ecolind.2020.106935 (2021).
79. Yuan, W. *et al*. Estimating crop yield using a satellite-based light use efficiency model. *Ecological Indicators* **60**, 702–709, https://doi.org/10.1016/j.ecolind.2015.08.013 (2016).
80. Anandhi, A. Growing degree days – Ecosystem indicator for changing diurnal temperatures and their impact on corn growth stages in Kansas. *Ecological Indicators* **61**, 149–158, https://doi.org/10.1016/j.ecolind.2015.08.023 (2016).
81. Wart, J. V. Estimating Crop Yield Potential At National Scales. *Field Crops Research* **143**, 34–43, https://doi.org/10.1016/j.fcr.2012.11.018 (2013).
82. Kang, Y. S. *et al*. Yield prediction and validation of onion (Allium cepa L.) using key variables in narrowband hyperspectral imagery and effective accumulated temperature. *Computers and Electronics in Agriculture* **178**, https://doi.org/10.1016/j.compag.2020.105667 (2020).
83. Long, D., Singh, V. P. & Li, Z.-L. How sensitive is SEBAL to changes in input variables, domain size and satellite sensor? *Journal of Geophysical Research: Atmospheres* **116**, https://doi.org/10.1029/2011jd016542 (2011).
84. Liu, Z., Wang, L. & Wang, S. Comparison of Different GPP Models in China Using MODIS Image and ChinaFLUX Data. *Remote Sensing* **6**, 10215–10231, https://doi.org/10.3390/rs61010215 (2014).
85. Edreira, J., Guilpart, N., Sadras, V., Cassman, K. G. & Grassini, P. Water productivity of rainfed maize and wheat: A local to global perspective. *Agricultural and Forest Meteorology* **259**, 364–373, https://doi.org/10.1016/j.agrformet.2018.05.019 (2018).
86. Li, H. *et al*. *Water Use Characteristics of Maize-Green Manure Intercropping Under Different Nitrogen Application Levels in the Oasis Irrigation Area Scientia Agricultura Sinica* **54**, 2608–2618 (2021).
87. Wang, S., Ibrom, A., Bauer-Gottwein, P. & Garcia, M. Incorporating diffuse radiation into a light use efficiency and evapotranspiration model: An 11-year study in a high latitude deciduous forest. *Agricultural and Forest Meteorology* https://doi.org/10.1016/j.agrformet.2017.10.023 (2018).
88. Cheng, M. High-resolution crop yield and water productivity dataset generated using random forest and remote sensing. *Zenodo* https://doi.org/10.5281/zenodo.6444614 (2022).

Acknowledgements

The study was supported by the National Key Research and Development Program of China (grant 2021YFD1201602), National Natural Science Foundation of China (Grant No. 42071426, 51922072, 51779161, 51009101), and Central Public-interest Scientific Institution Basal Research Fund for Chinese Academy of Agricultural Sciences (Grant Nos. Y2020YJ07), the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences, Hainan Yazhou Bay Seed Lab (B21HJ0221), and Special Fund for Independent Innovation of Agricultural Science and Technology in Jiangsu, China(CX(21)3065). JP was funded by the Fundación Ramón Areces project ELEMENTAL-CLIMATE, the Catalan Government projects (SGR 2017-1005 and AGAUR-2020PANDE00117), and the Spanish Government project (CGL2016-79835-P).

Author contributions

Minghan Cheng: Formal analysis, Writing – original draft, Visualization, Validation. Xiyun Jiao: Conceptualization, Writing – original draft. Lei Shi: Formal analysis, Validation. Josep Penuelas: Writing – revision. Lalit Kumar: Writing – revision. Chenwei Nie: Formal analysis, Visualization, Validation. Tianao Wu: Formal analysis, Validation. Kaihua Liu: Visualization. Wenbin Wu: Conceptualization. Xiuliang Jin: Conceptualization.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.W. or Xiuliang Jin.