

OPEN

Using HJ-CCD image and PLS algorithm to estimate the yield of field-grown winter wheat

Peng-Peng Zhang^{1,4}, Xin-Xing Zhou^{1,4}, Zhi-Xiang Wang¹, Wei Mao², Wen-Xi Li², Fei Yun³, Wen-Shan Guo^{1*} & Chang-Wei Tan^{1*}

Remote sensing has been used as an important means of estimating crop production, especially for the estimation of crop yield in the middle and late growth period. In order to further improve the accuracy of estimating winter wheat yield through remote sensing, this study analyzed the quantitative relationship between satellite remote sensing variables obtained from HJ-CCD images and the winter wheat yield, and used the partial least square (PLS) algorithm to construct and validate the multivariate remote sensing models of estimating the yield. The research showed a close relationship between yield and most remote sensing variables. Significant multiple correlations were also recorded between most remote sensing variables. The optimal principal components numbers of PLS models used to estimate yield were 4. Green normalized difference vegetation index (GNDVI), optimized soil-adjusted vegetation index (OSAVI), normalized difference vegetation index (NDVI) and plant senescence reflectance index (PSRI) were sensitive variables for yield remote sensing estimation. Through model development and model validation evaluation, the yield estimation model's coefficients of determination (R^2) were 0.81 and 0.74 respectively. The root mean square error (RMSE) were 693.9 kg ha^{-1} and 786.5 kg ha^{-1} . It showed that the PLS algorithm model estimates the yield better than the linear regression (LR) and principal components analysis (PCA) algorithms. The estimation accuracy was improved by more than 20% than the LR algorithm, and was 13% higher than the PCA algorithm. The results could provide an effective way to improve the estimation accuracy of winter wheat yield by remote sensing, and was conducive to large-area application and promotion.

Scientifically and accurately estimating crop yield is of significant importance for formulating plans for social and economic development, determining agricultural products import and export plans, ensuring national food security, guiding and regulating macroscopic planting structure, as well as improving the management skills of relevant agriculture-related enterprises and farmers¹⁻⁶. With the improvement of spatial, temporal and spectral resolutions of remote sensing data and the significant reduction of cost, currently remote sensing has been widely used in the estimation of production of all kinds of food crops, and it has become a research focus in the interdisciplinary field combining remote sensing and agriculture⁷.

At present, there were many methods and means for estimating crop yield, such as crop yield meteorological forecast, artificial sampling survey, statistical simulation model, remote sensing estimation and so on^{8,9}. Using a Criteria/Wofost simulation model that included the new numerical scheme for soil water balance, some researchers compared field data collected at the university of bologna's experimental farm in 1977–1987 with the median wheat yield, and the predicted value was consistent with the observed value¹⁰. Other researches have suggested that the mars-crop yield forecasting system (M-CYFS) model was more consistent as a predictor of crop yield than meteorological predictors since these predictors summarize the succession of agrometeorological conditions for the yield of the entire growing season¹¹. By using the environmental policy integrated climate (EPIC) crop

¹Jiangsu Key Laboratory of Crop Genetics and Physiology/Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops/Joint International Research Laboratory of Agriculture and Agri-Product Safety of the Ministry of Education of China, Yangzhou University, Yangzhou, 225009, China. ²Station of Land Protection of Yangzhou City, Yangzhou, 225009, China. ³National Tobacco Cultivation and Physiology and Biochemistry Research Centre/Key Laboratory for Tobacco Cultivation of Tobacco Industry, Henan Agricultural University, Zhengzhou, 450002, China. ⁴These authors contributed equally: Peng-Peng Zhang and Xin-Xing Zhou. *email: guows@yzu.edu.cn; tanwei010@126.com

Sample set	Number of samples	Amplitude of variation	Mean	Standard deviation	Standard error
Model development	159	3053.98 ~ 9566.56	5292.51	1314.53	104.25
Model validation	106	4444.82 ~ 9852.93	7115.77	1191.43	115.72

Table 1. Distribution of the yield in the model development and model validation (yield unit: kg ha⁻¹).

growth model and daily standardized precipitation evapotranspiration index (SPEI), a comprehensive method to quantitatively evaluate the effects of drought on crop yield was proposed¹². Among them, the crop yield meteorological forecast was suitable for small-area crops production prediction, but for large-area crops, due to large variations in field meteorological conditions in different wheat areas, the estimation accuracy was low. Manual sampling surveys had higher precision, but were time-consuming and labor intensive, and the cost was too high for large scale application. The statistical simulation model was a non-subjective method for constructing a mathematical relationship model based on historical data and estimating the future trend, but may result in errors due to climate change¹³. Remote sensing has been a high-tech method of obtaining large-area, fast, dynamic and multi-scale information on farmland. It has great application potential in large-scale crop growth monitoring, crop yield estimation, agricultural monitoring and forecasting, and agricultural resources survey^{14–16}. In China and abroad, considerable work has been done on remote sensing estimation of crops, and great progress has been made^{17,18}. Based on moderate-resolution imaging spectroradiometer (MODIS) derived normalized difference vegetation index (NDVI) data, a global agricultural monitoring system for crop monitoring and yield forecasting was built¹⁹. Based on advance very high resolution radiometer (AVHRR) data, the yield prediction model established by principal component analysis (PCA) had an estimated yield error of less than 8%²⁰. The satellite-based vegetation index could be used to predict wheat yield six weeks before the time of harvest in Punjab province of Pakistan²¹. Through the data were obtained on time series remote sensing images fused with high temporal and spatial resolutions, along with grain yield and protein contents at maturity, preliminary harvest was showed that filling and anthesis stages were the best time to estimate wheat yield²². Some research showed that both agroclimate + MODIS-NDVI and agroclimate + MODIS-environmental vegetation index (EVI) performed equally well predicting spring wheat yield at the ecodistrict scale²³. Through using NDVI derived from the data of the MODIS, the method for estimating and forecasting wheat yield in Hungary in the period of 2003–2015 was improved and obtained better prediction results²⁴. In evaluating the influencing factors of wheat yield of the four populations, partial least square (PLS) algorithm could reveal the control factors on wheat yield in the study area and provided a reference tool for analyses in other crops or areas²⁵. However, the satellite remote sensing data used in these researches was relatively short-lived, and the stability of the model simulation effect needed further testing. Some researchers argued that they lacked the spatial detail necessary for studying vegetation phenology in heterogeneous landscapes while MODIS and AVHRR have been the sensors most often used in remote sensing based phenological analysis²⁶. Relevant research summarized yield estimation methods in each region through remote sensing and illustrated the importance of distinguishing between accuracy for spatial and temporal variation²⁷. The data sources of the earth observation satellites were chiefly low spatial resolution MODIS, national oceanic and atmospheric administration (NOAA)/AVHRR images^{23,28}, medium spatial resolution India remote-sensing satellite (IRS-P6), enhanced thematic mapper (ETM), thematic mapper (TM) images, and high spatial resolution Quickbird, SPOT, IKONOS, ALOS foreign images^{29–31}. However, these data sources were expensive, which limited their use in small and medium research units and production management departments^{32,33}. Therefore, it was of great significance to promote the application of image data obtained by satellites developed by China in remote sensing of agricultural conditions. On September 6, 2008, China successfully launched satellites A and B (abbreviated as HJ-CCD) of the “Environment and Disaster Monitoring and Forecasting Small Satellite Constellation System” with independent intellectual property rights. The satellites were equipped with wide-band CCD sensors with spatial resolution of the sensor being 30 m. The resolution was 2 d when satellites A and B were making observations simultaneously, making them an ideal data source for agricultural remote sensing operation.

The objectives of the present study were to investigate the quantitative relationship between the yield and satellite remote sensing variables during flowering period, and developed an effective way to improve the estimation accuracy of winter wheat yield by remote sensing.

Results and analysis

Yield distribution. The amplitude of variation, mean, standard deviation and standard error of the model development and validation were similar (Table 1). At the same time, the model development and model validation samples had desirable consistency.

Quantitative analysis between remote sensing variables and yield. The quantitative analysis of the yield and remote sensing variables of 159 samples in the model development showed that there were significant or extremely significant relationships between the yield and most remote sensing variables (Table 2). The yield was most closely related to structure intensive pigment index (PSRI), followed by green normalized difference vegetation index (GNDVI), the correlation coefficients being -0.69 and -0.65 , respectively. A large proportion of correlations between the yield and vegetation indices were obviously better than single bands. Most remote sensing variables had considerable multiple pairwise correlations, where the correlation coefficients were almost between 0.80 and 1.00. In particular, single-band B₁–B₄ pairwise correlation coefficients were between 0.95 and 0.99, and the pairwise correlation coefficient of most vegetation indices was above 0.90. It indicated that the

	Yield	B ₁	B ₂	B ₃	B ₄	NDVI	SAVI	OSAVI	NRI	GNDVI	SIPI	PSRI	DVI	RVI
Yield	1.00													
B ₁	-0.51	1.00												
B ₂	-0.47	0.99	1.00											
B ₃	-0.33	0.99	0.99	1.00										
B ₄	-0.29	0.96	0.98	0.95	1.00									
NDVI	-0.61	0.86	0.85	0.85	0.97	1.00								
SAVI	-0.49	0.86	0.88	0.85	0.95	0.99	1.00							
OSAVI	-0.48	0.87	0.90	0.84	0.97	0.99	1.00	1.00						
NRI	0.11	0.96	0.93	0.96	0.83	0.80	0.78	0.73	1.00					
GNDVI	-0.65	0.94	0.95	0.93	0.97	0.98	0.95	0.98	0.91	1.00				
SIPI	-0.54	0.94	0.94	0.92	0.97	0.98	0.95	0.95	0.84	0.99	1.00			
PSRI	-0.69	0.95	0.96	0.91	0.93	0.86	0.86	0.87	0.93	0.93	0.97	1.00		
DVI	-0.22	0.81	0.93	0.88	0.97	0.99	0.98	0.97	0.81	0.98	0.98	0.91	1.00	
RVI	-0.23	0.82	0.81	0.80	0.80	0.96	0.98	0.96	0.65	0.91	0.91	0.69	0.97	1.00

Table 2. Correlation between remote sensing variables and winter wheat yield (n = 159). B₁, B₂, B₃ and B₄ denoted spectrum reflectance at blue, green, red and near infrared bands, respectively.

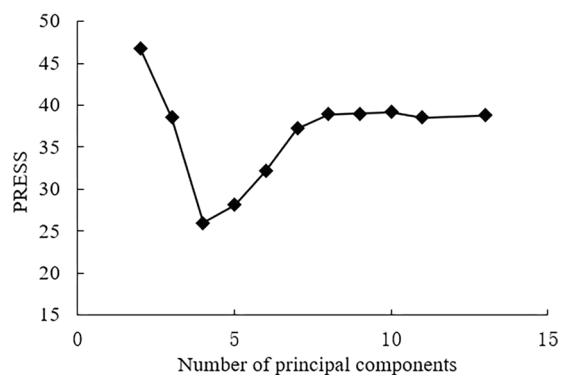


Figure 1. PRESS changes with the principal components.

model established by PLS algorithm was more reasonable than the traditional statistical algorithms and the ordinary least squares method, and might lead to better results.

PLS model. According to the predictive residual error sum of square (PRESS) minimum value, the number of optimal principal components could be determined. Figure 1 showed the variation of PRESS with the number of principal components obtained from the yield's model development. At the beginning, as the number of principal components increased, the yield's PRESS value decreased to a large extent. It has indicated that due to the small number of principal components, the model fitting was extremely inadequate. It meant that the missing fitting phenomenon occurred. When the principal components numbers of the yield's model was 4, the PRESS value (25.96) was the smallest. After that, as the number of principal components increased, the yield's PRESS value increased sharply, until it tended to be saturated. Via this, it was indicated that the over-fitting phenomenon occurred due to too many principal components. Therefore, it was reasonable to select the number of principal components corresponding to the minimum PRESS value. Therefore, the optimal principal components numbers of the yield models based on PLS algorithm was 4.

Based on the PLS algorithm, the yield model had the four vegetation indices with the principal components number of 4, namely, GNDVI, optimized soil-adjusted vegetation index (OSAVI), NDVI and PSRI as independent variables, and the yield was the dependent variable. The yield estimation model was constructed by the yield model development and the HJ-CCD images during the three periods 2016-05-02, 2017-04-24, and 2018-04-26 was:

$$\text{Yield} = 2011.7 \times \text{GNDVI} + 1673.7 \times \text{OSAVI} + 1821.4 \times \text{NDVI} - 2103.8 \times \text{PSRI} + 2810.2 \quad (1)$$

The optimal linear regression equation and its coefficient of determination (R^2) and root mean square error (RMSE) were obtained. Figure 2 showed the evaluation of the yield model's estimation ability. It could be seen from Fig. 2 that the model development samples number was larger than the model validation samples number. The R^2 of the linear equation established by the model development was significantly larger than R^2 of the model validation. The model development RMSE was significantly smaller than the model validation RMSE. It indicated that the prediction model effect of the model development samples was significantly better than

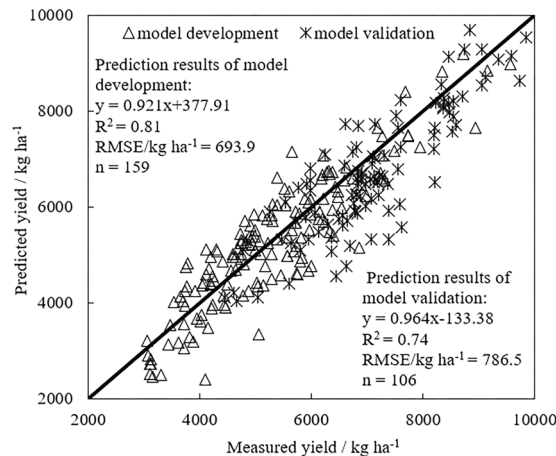


Figure 2. Evaluation of the yield model.

the model validation. Thereby, it has theoretically conformed to the model's estimation law. In addition, The R^2 values between the predicted and measured yield were greater than 0.7 and the RMSE were 693.9 kg ha^{-1} and 786.5 kg ha^{-1} , respectively. These results indicated that the PLS model could be used effectively to estimate the winter wheat yield.

In order to compare with the traditional algorithms, the linear regression (LR) and PCA algorithms were used to establish the yield's estimation model through the model development and model validation samples, respectively. The model of the predicted and measured values of the yield were evaluated by R^2 and RMSE. The specific process was not described here. Table 3 showed the comparison of estimation results with PLS, LR and PCA based on the model development and model validation. It showed that the sample number was the same. The PLS algorithm models' R^2 of the yield were greater than LR and PCA algorithms models, and RMSE were smaller than the LR and PCA algorithms models. This indicated that the PLS algorithm model was better than the LR and PCA algorithms in estimating the yield. The estimation accuracies based on the yield model development and model validation were 19.68% and 25.73% higher than LR algorithm models, respectively, and were 13.49% and 12.86% higher than PCA algorithm models, respectively. The estimation accuracies were above 90%.

According to the above analyses, GNDVI, NDVI, PSRI and OSAVI maps were generated using 2018-04-26 HJ-CCD images. On those the winter wheat planting data was superimposed to remove the non-winter wheat area by one-to-one solution and binarization mask. Based on the administrative boundary vector data, as well as the above PLS model, the spatial distribution map of estimating winter wheat yield in central Jiangsu was produced (Fig. 3). The distribution of the yield was mainly higher than 5250 kg ha^{-1} , of which Yancheng and its surrounding wheat areas was mainly $4500\text{--}6000 \text{ kg ha}^{-1}$ and the northern wheat area of Jiangyan was more than 6000 kg ha^{-1} . The number in the south wheat area rarely appeared above 6000 kg ha^{-1} , and the Yangtze River area was mainly $3750\text{--}5250 \text{ kg ha}^{-1}$, especially the south of the Yangtze River, which was mainly $3750\text{--}4500 \text{ kg ha}^{-1}$. By predicting the results, relevant departments and farmers can formulate corresponding management and trade policies in advance, so as to achieve the effects of graded harvest and quantitative purchasing and storage. Remote sensing technology can be used to monitor crops in a wide range of areas so that agricultural management and farmers can obtain timely crop yield information. This technology saves the cost of manpower and material resources to the greatest extent and has great scientific and production significance.

Discussions

At present, the remote sensing images used in the crop estimation were mainly originated via MODIS, NOAA/AVHRR, etc.^{23,28}. These images had low spatial resolution and were difficult to apply to high-precision winter wheat remote sensing estimation in small areas. On the other hand, the high-resolution images such as Quickbird, SPOT, IKONOS were costly^{30,31}. The medium-resolution TM images had revisiting periods of 16 days, making it difficult to obtain high-quality data in time. This limited continuous crop monitoring and made it inappropriate to predict crop yield³⁴. The HJ-CCD satellites developed by China have been put into use one after another. The quality of the data obtained was continuously improved and was provided free of charge to users. This has created a convenient data platform for remote sensing and estimation of regional crop's quality and productivity³⁵. The experimental area of the present research has been located in the coastal area along the Yangtze River in Jiangsu Province. The whole wheat field has been fragmented and as a result the planting structure was complex. The time resolution of the selected HJ-CCD image was 2d, and the scanning width of the single scene image was 750 km. These characteristics could meet the estimation demands for the actual regional winter wheat. Considering time resolution, spatial resolution and cost, the HJ-CCD image was more appropriate than the data of MODIS, TM, Quickbird, etc.

There was a close relationship between yield and most remote sensing variables. In addition, there were considerable multiple correlations between most remote sensing variables. This made it difficult to establish a higher precision remote sensing estimation model of the yield using traditional algorithms^{6,36–38}. In this study, the PLS algorithm was used to construct the remote sensing estimation model of the yield with GNDVI, OSAVI, NDVI and PSRI as the independent variables. The correlation between the yield and these remote sensing variables was extremely significant. They could be easily extracted and calculated from the HJ-CCD image. The RMSE values of

Algorithm	Number of principal components	Number of samples		R ²		RMSE/kg ha ⁻¹		Accuracy/%	
		Model development	Model validation	Model development	Model validation	Model development	Model validation	Model development	Model validation
PLS	4	159	106	0.81	0.74	693.9	786.5	92.43	90.38
PCA	5	159	106	0.63	0.56	1054.7	1067.3	78.94	77.52
LR	0	159	106	0.57	0.47	1123.6	1342.7	72.75	64.65

Table 3. Comparison of predicted results with PLS, LR and PCA.

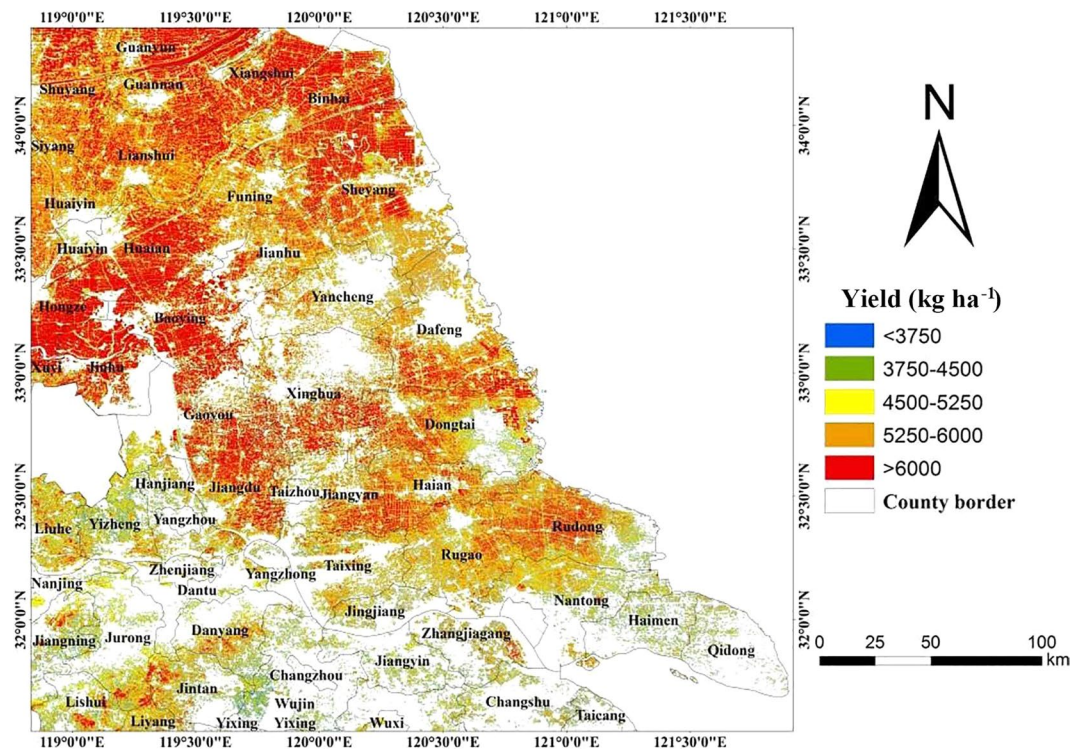


Figure 3. Spatial distribution of winter wheat yield in central Jiangsu region, China.

the yield's estimation model based on these remote sensing variables as the independent variables were lower than the traditional LR and PCA models. The results showed that the PLS model, as a new multivariate analysis method, had a very high adaptability in yield estimation, especially when there were many variables and multiple correlations in the analysis. The PLS algorithm could effectively optimize the dependent variables, and its model was significantly better than LR and PCA algorithms in the yield's estimation of winter wheat. The effects of crop spectral parameters and remote sensing vegetation index on crops are not often single. It is often multiple remote sensing variables acting on a single agronomic index. Therefore, based on the traditional univariate model and linear regression algorithm, it is difficult to make a more accurate prediction. PLS algorithm is highly adaptive to the condition of multivariate and high correlation. The results were consistent with Hanen *et al.*³⁹ and Zhao *et al.*⁴⁰. Based on the traditional linear regression algorithm, it is easy to build a model for monitoring and predicting crop quality, growth and yield, but the prediction accuracy is often not up to the requirements. For the prediction of agricultural indicators under the control of multiple variables, the PLS algorithm is superior to the general linear regression model. The results of this research better than Liu *et al.*⁴¹ and Xue *et al.*⁴². In order to reflect it in a better way, the actual situation of field planting and different varieties were selected in the experiment. Using the data derived from different varieties as test samples, the results were of more general in significance. It was helpful to the promotion and application in actual production. It indicated that it was feasible to use the PLS model to estimate winter wheat yield with high precision. It has, therefore, provided an effective method and technical support for the high-precision estimation of aerospace remote sensing images, and was also conducive to large-area application and promotion.

According to the spatial distribution map for predicting winter wheat yield in Jiangsu province (Fig. 3). Northern Jiangsu, especially in the northwest of Gaoyou and Sheyang county, the yield is higher than other areas. The yield of the middle region of Jiangsu is relatively lower than that of the northern region, which basically maintains around 5000–6000 kg ha⁻¹. But along the river and the region south of the Yangtze river, the yield is generally low. There was large scale wheat cultivation in northern Jiangsu. Local agricultural facilities were well developed, and agricultural production was mainly in the form of farms for planting and management. Therefore, winter wheat planting could be managed uniformly, with good cultivation measures and maximum implementation. Overall agricultural

development and management in the middle area of Jiangsu was slightly worse than that in north of Jiangsu. But the whole structure of agricultural facilities and agricultural management could meet the planting of winter wheat. Therefore, wheat yield presented a general level range. The southern Jiangsu area was mostly metropolis and urban area with less farmland, and there were few areas for wheat cultivation. At the same time, the local farmland was chaotic and scattered, mainly operated by small farmer households. It might result in good cultivation measures and management could not be used effectively. Therefore, the winter wheat yield in southern Jiangsu was relatively low. The predicted results of the spatial distribution map for predicting winter wheat yield in Jiangsu province were basically consistent with the actual situation of winter wheat production. It indicated that it was feasible to use the PLS model to predict winter wheat yield with high precision. It has, therefore, provided an effective method and technical support for the high precision remote sensing prediction of winter wheat yield.

The samples used in the research were relatively concentrated, basically ranging from 4000 kg ha⁻¹ to 9000 kg ha⁻¹. Samples with higher or lower content were relatively few, showing above 9000 kg ha⁻¹ and less than 4000 kg ha⁻¹. There was a lack of samples more than 10063 kg ha⁻¹ and less than 2152 kg ha⁻¹. If the variation of the yield samples was increased, the PLS model would be further optimized and its application range would be further expanded. The remote sensing estimation models of the yield would become more reliable. The results obtained were based only on the HJ-CCD data of the Jiangsu experimental area. Therefore, whether the model would be applicable to other remote sensing sensor data and/or estimate the winter wheat yield in other areas needed further study.

The present study did not compare the PLS algorithm with artificial neural network (ANN)^{43,44}, support vector machines (SVM)⁴⁵, geostatistics⁴⁶, etc. Simultaneously, it also did not take into account the factors affecting winter wheat cultivation. These algorithms and factors actually had a wide range of influence on the estimation results of winter wheat yield and needed further study.

Conclusions

In the present research, a close relationship between yield and most remote sensing variables were found. Significant multiple correlations were also obtained between most remote sensing variables. GNDVI, OSAVI, NDVI and PSRI were sensitive for remotely estimating the yield. Through the model development and model validation evaluation, the estimation model of the yield had R² of 0.81 and 0.74, and the RMSE were 693.9 kg ha⁻¹ and 786.5 kg ha⁻¹. It showed that the PLS algorithm estimated the yield better than the LR and PCA algorithms. The improvements were by more than 20% than the LR algorithm and more than 13% higher than the PCA algorithm. The PLS model provided an effective way to improve the accuracy of estimating winter wheat yield through remote sensing.

Materials and methods

Test design and data acquisition. For the present investigation, data collection was carried out in 5 counties, namely, Taixing, Jiangyan, Yizheng, Xinghua and Dafeng in Jiangsu Province in 2016, the People's Republic of China. There were 15–20 sampling points in each county, totaling 92. The location of each sampling site was determined by using a Juno ST hand-held GPS meter (Trimble Co. USA). The survey mainly included information collection on winter wheat varieties, growth period, population growth and disasters status (mainly pests and diseases). Winter wheat varieties were of medium and weak gluten type, mainly *Yangmai 13*, *Yangmai 15*, *Yangmai 16* and *Yangfumai 2*. These varieties were available in the experimental counties. GPS was used to locate the positions during the mature stage. Samples were taken back to measure the yields in the laboratory.

A total of 3 tests were launched in the experimental counties from 2016–2018 to collect data. The satellite data was HJ-CCD images taken at flowering stage of winter wheat. Data collection for Test 1, 2 and 3 were conducted on May 2, 2016; April 24, 2017 and April 26, 2018, respectively. The sampling points considered for the Test 1–3 were 92, 96 and 67, respectively. Figure 4 showed the distribution of sampling points in 2016, 2017 and 2018.

The yield data measured in Tests 1–3 were arranged in the order of the yield values in the winter wheat grain sample. In order to enhance the stability of the estimation model, the numerical samples of 265 yields were randomly divided into model development and model validation according to a ratio of 3:2, on the premise that the maximum and minimum yields were placed in the modeling sample set.

Image preprocessing. Environment for Visualizing Images (ENVI 5.4) software was used to preprocess satellite images. Firstly, georeferencing process was performed. The specific operation was that the 1:100,000 topographic maps of Jiangsu area were used to roughly correct the satellite image. Thereafter, the GPS control points for ground measuring were used to precisely correct the satellite image. This helped to ensure that the precision of geometric correction was better than one pixel. Atmospheric correction and reflectance conversion were carried out by empirical linear method^{14,47}. According to the analysis of the results, the corresponding single-band value graph was obtained by using a workflow called band math in ENVI. Data of wheat growing areas were obtained by supervised classification. The winter wheat planting data were superimposed and the non-winter wheat area was eliminated by one-to-one solution and binarization mask. By using the administrative boundary vector data and the above PLS model, the spatial distribution map of winter wheat yield in Jiangsu province was produced.

Yield measurement. During the maturity period, 5 plots were selected by five-point sampling method (Fig. 5) in the middle part in the field. The four plots on the periphery formed a rectangle, and they were 10m apart. Each plot was 5m², and each plot grew evenly, which could represent the overall situation of the field. And the field area sampled should be more than 2 hectares. The grains of the five plots were brought back to the laboratory. All the samples from five plots were shelled and weighed separately. After averaging the yields of these 5 plots, the average value was the value of the sampling site and was converted to the value of one hectare, which was the yield.

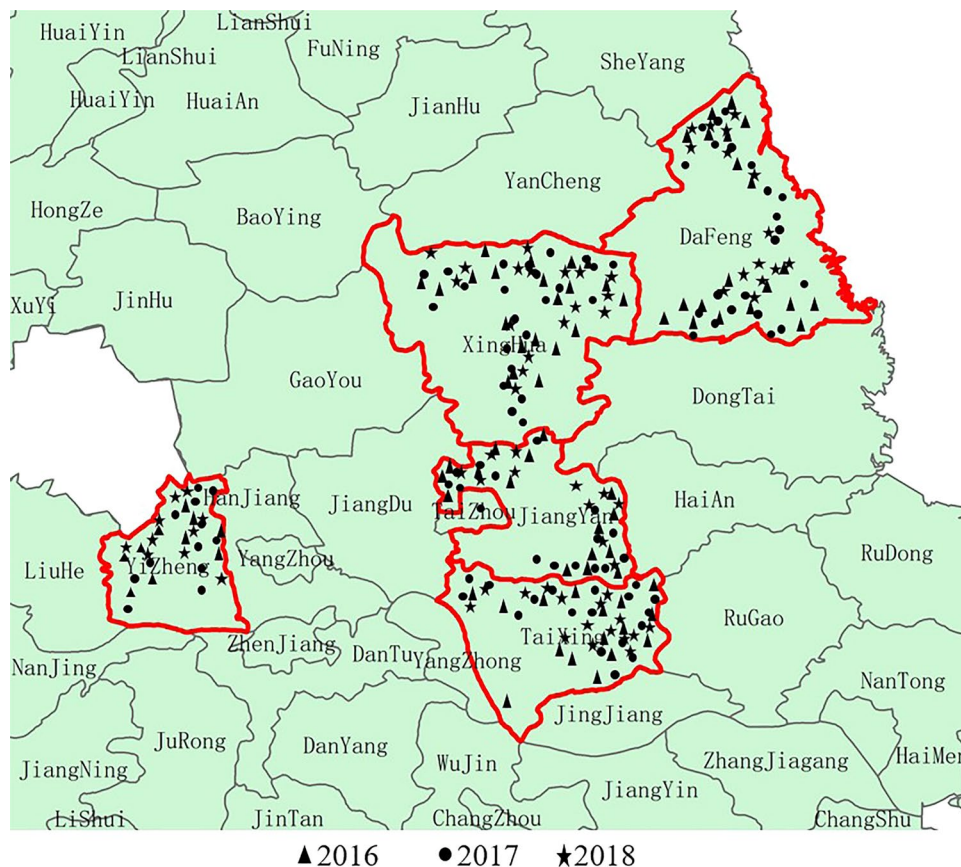
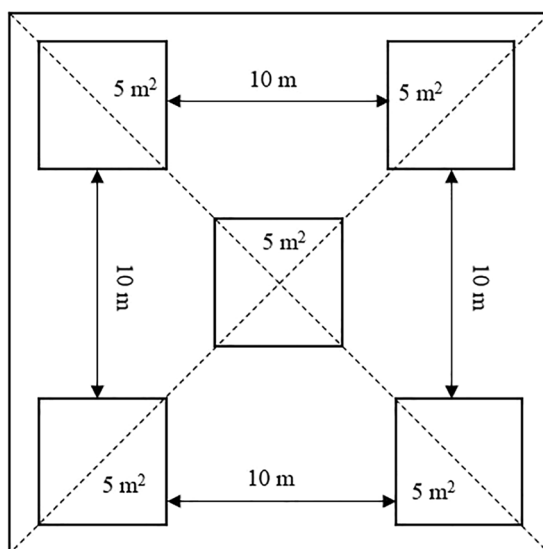


Figure 4. Sampling point information for three consecutive years.



Five-point sampling method

Figure 5. Five-point sampling method.

Satellite remote sensing variables. In combination with the physical significance of spectral indices, selection of model parameters was based on the spectral characteristics of crops and the available literatures in home and abroad. In this study, four HJ-CCD bands and nine common spectral vegetation indices were selected (Table 4) as independent variables for PLS analysis in order to construct remote sensing estimation model of the yield.

Vegetation index	Abbreviation	Algorithm	Source
Normalized difference vegetation index	NDVI	$(B_4 - B_3) / (B_4 + B_3)$	52
Soil-adjusted vegetation index	SAVI	$(B_4 - B_3) / (B_4 + B_3 + 0.5) * 1.5$	53
Optimized soil-adjusted vegetation index	OSAVI	$(B_4 - B_3) / (B_4 + B_3 + 0.16) * 1.16$	54
Nitrogen reflectance index	NRI	$(B_2 - B_3) / (B_2 + B_3)$	55
Green normalized difference vegetation index	GNDVI	$(B_4 - B_2) / (B_4 + B_2)$	56
Structure intensive pigment index	SIPI	$(B_4 - B_1) / (B_4 + B_1)$	57
Plant senescence reflectance index	PSRI	$(B_3 - B_1) / B_4$	55
Difference vegetation index	DVI	$B_4 - B_3$	58
Ratio vegetation index	RVI	B_4 / B_3	59

Table 4. Formulas of remote sensing vegetation indices.

To extract spectral band brightness values of corresponding GPS positioning sampling points, ENVI 5.4 and geographic information system software (ArcGIS 10.2) were used. actual these satellite remote sensing variables were calculated using Excel 2016.

PLS regression. PLS regression was first applied to the field of chemometrics. PLS regression has been considered as a new multivariate analysis method with wide applicability. The PLS regression was concentrated on the characteristics of principal component, linear regression and typical multiple regression analysis. It could effectively solve many problems. Such as, problems that could not be solved by ordinary multiple regression, especially when there were many variables and multiple correlations. In these cases, PLS could effectively decompose and screen the comprehensive variables that were most explanatory to the dependent variables. Therefore, the established model was more reliable than the ordinary regression analysis. The PLS method first extracted a new variable called component as an independent variable, and established a linear combination relationship between the dependent variable and the independent variable. The coefficient was determined by PLS calculation, and then the regression equation of the dependent variable was constructed. The regression model established by the PLS method could be expressed by Eq. (2):

$$y_m = a_{0m} + a_{1m}x_1 + \dots + a_{pm}x_p \quad (m = 1, 2, \dots, p) \quad (2)$$

where x_1, \dots, x_p were linear combinations of remote sensing variables, $a_{0m}, a_{1m}, \dots, a_{pm}$ were parameters of the regression model and could be computed by PLS.

When the model was established by PLS algorithm, the increase of the number of principal components would improve the accuracy of the model. But too many principal components would cause over-fitting and the error would increase. Therefore, it was very important to determine the optimal principal components number of the PLS model. In this study, the sum of squared residuals was calculated by the cross-validation method. The smaller the PRESS value, the stronger the estimation ability of the model. Therefore, the optimal principal components number could be determined according to the minimum value of PRESS. PRESS could be expressed by Eq. (3):

$$PRESS = \sum_{i=1}^k (y_i - \hat{y}_{i,-i})^2 \quad (3)$$

where $y_i, \hat{y}_{i,-i}$ were the measured value corresponding to the i th sample and the estimated value when the i th sample was excluded, and k was the number of validating iterations.

For the basic principles and specific practices of the PLS algorithm and PRESS, please refer to references^{25,48}, which were not described here. Both the PLS and PRESS processes were performed by a self-written MATLAB program. In this study, the yields were estimated based on the PLS algorithm. Then it was compared with the yields' estimation model based on LR and PCA algorithms. For the explanation of LR and PCA algorithms, please refer to the references^{49,50}, which were not described here.

Evaluation of the model. Using the samples of the model development and model validation, the model was evaluated by plotting the 1:1 relationship graph between the predicted and measured values of the yield. The evaluation indices were the R^2 and the RMSE³¹. On one hand, the larger the R^2 , the better the model is. On the other hand, the smaller the RMSE, the stronger the estimation ability of the model is. RMSE and estimation accuracy were calculated using Eqs. (4) and (5), respectively:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

$$Accuracy = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

where y_i and \hat{y}_i represented measured values and predicted values of wheat yields, respectively, and n was the number of samples.

Received: 19 July 2019; Accepted: 9 March 2020;

Published online: 20 March 2020

References

1. Cabas-Monje, J. H., Weersink, A. & Olale, E. Crop yield response to economic, site and climatic variables. *Clim. Change*. **101**, 599–616 (2010).
2. Ittersum, V. & K. M. Crop yields and global food security: will yield increase continue to feed the world? *Eur. Rev. Agric. Econ.* **97**, 191–192 (2016).
3. Newton, A. C., Johnson, S. N. & Gregory, P. J. Implications of climate change for diseases, crop yields and food security. *Euphytica*. **179**, 3–18 (2011).
4. Li, Y. *et al.* Assimilating remote sensing information into a coupled hydrology-crop growth model to estimate regional maize yield in arid regions. *Ecol. Model.* **291**, 15–27 (2014).
5. Tan, C. *et al.* Assessment of Fv/Fm absorbed by wheat canopies employing *in-situ* hyperspectral vegetation indexes. *Sci. Rep.* **8**, 9525 (2018).
6. Basso, B. & Liu, L. Seasonal crop yield forecast: methods, applications, and accuracies. *Adv. Agron.* **154**, 201–255 (2018).
7. Huang, J. & Han, D. Meta-analysis of influential factors on crop yield estimation by Remote Sens. *Int. J. Remote Sens.* **35**, 29 (2014).
8. Müller, C. *et al.* Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications. *Geosc. Model. Dev. Dis.* **10**, 1–39 (2016).
9. Tripathy, R. *et al.* Forecasting wheat yield in Punjab state of India by combining crop simulation model WOFOST and remotely sensed inputs. *Remote Sens. Lett.* **4**, 19–28 (2013).
10. Marletto, V., Ventura, F., Fontana, G. & Tomei, F. Wheat growth simulation and yield prediction with seasonal forecasts and a numerical model. *Remote Sens. Lett.* **147**, 71–79 (2007).
11. Lecerf, R. *et al.* Assessing the information in crop model and meteorological indicators to forecast crop yield over. *Europe. Agric. Syst.* **168**, 191–202 (2019).
12. Wang, Q. *et al.* A comprehensively quantitative method of evaluating the impact of drought on crop yield using daily multi-scale SPEI and crop growth process model. *Int. J. Biometeorol.* **61**, 1–15 (2017).
13. Wit, A. J. *et al.* Using ERA-INTERIM for regional crop yield forecasting in Europe. *Clim. Res.* **44**, 41–53 (2010).
14. Tan, C. *et al.* Estimation of leaf nitrogen concentration in wheat by the combinations of two vegetation indexes using HJ-CCD images. *Int. J. Agric. Biol.* **20**, 1908–1914 (2018).
15. Peña-Arancibia, J. L. *et al.* Assessing irrigated agriculture's surface water and groundwater consumption by combining satellite remote sensing and hydrologic modelling. *Sci. Total Environ.* **542**, 372–382 (2016).
16. Nemani, R. R. Climate-driven increases in global terrestrial net primary production from 1982 to 1999. *Science*. **300**, 1560–1563 (2003).
17. Satir, O. & Berberoglu, S. Crop yield prediction under soil salinity using satellite derived vegetation indices. *Field Crop. Res.* **192**, 134–143 (2016).
18. Tan, C. *et al.* Analysis of different hyperspectral variables for diagnosing leaf nitrogen accumulation in wheat. *Front. Plant Sci.* **9**, 674 (2018).
19. Becker-Reshef, I., Chris, J., Mark, S. & Vermote, E. Monitoring global croplands with coarse resolution earth observations: the Global Agriculture Monitoring (GLAM) Project. *Remote Sens.* **2**, 1589–1609 (2010).
20. Salazar, L., Kogan, F. & Roytman, L. Use of remote sensing data for estimation of winter wheat yield in the United States. *Int. J. Remote Sens.* **28**, 3795–3811 (2007).
21. Dempewolf, J. *et al.* Wheat yield forecasting for Punjab province from vegetation index time series and historic crop statistics. *Remote Sens.* **6**, 9653–9675 (2014).
22. Wang, L. *et al.* Predicting grain yield and protein content in wheat by fusing multi-sensor and multi-temporal remote-sensing images. *Field Crop. Res.* **164**, 178–188 (2014).
23. Kouadio, L. *et al.* Assessing the performance of MODIS NDVI and EVI for seasonal crop yield forecasting at the ecodistrict scale. *Remote Sens.* **6**, 10193–10214 (2014).
24. Bognár, P. *et al.* Yield estimation and forecasting for winter wheat in Hungary using time series of MODIS data. *Int. J. Remote Sens.* **38**, 3394–3414 (2017).
25. Hu, Y. *et al.* Partial least squares regression for determining factors controlling winter wheat yield. *Agron. J.* **110**, 281–292 (2018).
26. Fisher, J. L., Mustard, J. F. & Vadeboncoeur, M. A. Green leaf phenology at Landsat resolution: Scaling from the field to the satellite. *Remote Sens. Environ.* **100**, 265–279 (2006).
27. Azzari, G., Jain, M. & Lobell, D. B. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. *Remote Sens. Environ.* **202**, 129–141 (2017).
28. Boschetti, M., Stroppiana, D., Brivio, P. A. & Bocchi, S. Multi-year monitoring of rice crop phenology through time series analysis of MODIS images. *Int. J. Remote Sens.* **30**, 4643–4662 (2009).
29. Jalan, S. & Sokhi, B. S. Comparison of different pan-sharpening methods for spectral characteristic preservation: multi-temporal CARTOSAT-1 and IRS-P6 LISS-IV imagery. *Int. J. Remote Sens.* **33**, 5629–5643 (2012).
30. Wu, J., Wang, D. & Bauer, M. E. Image-based atmospheric correction of QuickBird imagery of Minnesota cropland. *Remote Sens. Environ.* **99**, 315–325 (2005).
31. Turker, M. & Ozdarici, A. Field-based crop classification using SPOT4, SPOT5, IKONOS and QuickBird imagery for agricultural areas: a comparison study. *Int. J. Remote Sens.* **32**, 9735–9768 (2011).
32. Orynbaikyzy, A., Gessner, U. & Conrad, C. Crop type classification using a combination of optical and radar remote sensing data: a review. *Int. J. Remote Sens.* **40**, 6553–6595 (2019).
33. Das, S. K. & Singh, R. A multiple-frame approach to crop yield estimation from satellite- remotely sensed data. *Int. J. Remote Sens.* **34**, 3803–3819 (2013).
34. Beckschäfer, P. Obtaining rubber plantation age information from very dense Landsat TM & ETM + time series data and pixel-based image compositing. *Remote Sens. Environ.* **96**, 89–100 (2017).
35. Cheng, Z., Meng, J. & Wang, Y. Improving spring maize yield estimation at field scale by assimilating time-Series HJ-CCD data into the WOFOST model using a new method with fast algorithms. *Remote Sens.* **8**, 303 (2016).
36. Khaledian, Y. *et al.* Assessment and monitoring of soil degradation during land use change using multivariate analysis. *Land Degrad. Dev.* **28**, 128–141 (2017).
37. Castaldi, F. *et al.* Estimation of soil properties at the field scale from satellite data: a comparison between spatial and non-spatial techniques. *Eur. J. Soil Sci.* **65**, 842–851 (2015).
38. Mavromatis, T. Spatial resolution effects on crop yield forecasts: An application to rainfed wheat yield in north Greece with CERES-Wheat. *Agric. Syst.* **143**, 38–48 (2016).
39. Hansen, P. M., Jørgensen, J. R. & Thomsen, A. Predicting grain yield and protein content in winter wheat and spring barley using repeated canopy reflectance measurements and partial least squares regression. *J. Agr. Sci.* **139**, 307–318 (2002).
40. Zhao, C. *et al.* Predicting grain protein content of winter wheat using remote sensing data based on nitrogen status and water stress. *Int. J. Appl. Earth Obs.* **7**, 1–9 (2005).
41. Liu, L. *et al.* Predicting winter wheat condition, grain yield and protein content using multi-temporal EnviSat-ASAR and Landsat TM satellite images. *Int. J. Remote Sens.* **27**, 737–753 (2006).

42. Xue, L., Cao, W. & Yang, L. Predicting grain yield and protein content in winter wheat at different N supply levels using canopy reflectance spectra. *Pedosphere*. **17**, 646–653 (2007).
43. Panda, S. S., Ames, D. P. & Suranjan, P. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sens.* **2**, 673–696 (2010).
44. Shabani, A., Ghaffary, K. A., Sepaskhah, A. R. & Kamgar-Haghighi, A. K. Using the artificial neural network to estimate leaf area. *Sci. Hortic.* **216**, 103–110 (2017).
45. Li, D., Yang, F. & Wang, X. Study on ensemble crop information extraction of remote sensing images based on SVM and BPNN. *J. Indian Soc. Remote Sens.* **45**, 1–9 (2016).
46. Pringle, M. J., Marchant, B. P. & Lark, R. M. Analysis of two variants of a spatially distributed crop model, using wavelet transforms and geostatistics. *Agric. Syst.* **98**, 135–146 (2008).
47. Hamm, N., Atkinson, P. M. & Milton, E. J. A per-pixel, non-stationary mixed model for empirical line atmospheric correction in Remote Sens. *Remote Sens. Environ.* **124**, 666–678 (2012).
48. Delaigle, A. & Hall, P. Methodology and theory for partial least squares applied to functional data. *Ann. Stat.* **40**, 322–352 (2012).
49. Eberly, L. E. Correlation and simple linear regression. *Radiology*. **227**, 617–622 (2003).
50. Humberto, B. C., Claassen, M. M. & Presley, D. R. Summer cover crops fix nitrogen, increase crop yield, and improve soil-crop relationships. *Agron. J.* **104**, 137–147 (2012).
51. Tan, C. *et al.* W. Remotely assessing fraction of photosynthetically active radiation (FPAR) for wheat canopies based on hyperspectral vegetation indexes. *Front. Plant Sci.* **9**, 776 (2018).
52. Sarmah, S., Jia, G., Zhang, A. & Singha, M. Assessing seasonal trends and variability of vegetation growth from NDVI3g, MODIS NDVI and EVI over South Asia. *Remote Sens. Lett.* **9**, 1195–1204 (2018).
53. Ren, H. R. & Zhou, G. S. Determination of green aboveground biomass in desert steppe using litter-soil-adjusted vegetation index. *Eur. J. Remote Sens.* **47**, 611–625 (2014).
54. Clevers, J. G. P. W., Kooistra, L. & Brande, M. V. D. Using Sentinel-2 data for retrieving LAI and leaf and canopy chlorophyll content of a potato crop. *Remote Sens.* **9**, 405 (2017).
55. Knox, N. M. *et al.* Nitrogen prediction in grasses: effect of bandwidth and plant material state on absorption feature selection. *Int. J. Remote Sens.* **31**, 691–704 (2010).
56. Cicek, H. *et al.* Using vegetation indices from satellite remote sensing to assess corn and soybean response to controlled tile drainage. *Agric. Water Manage.* **98**, 261–270 (2010).
57. Liu, L. *et al.* Predicting winter wheat condition, grain yield and protein content using multi-temporal EnviSat-ASAR and Landsat TM satellite images. *Int. J. Remote Sens.* **27**, 737–753 (2006).
58. Trombetta, A., Iacobellis, V., Tarantino, E. & Gentile, F. Calibration of the AquaCrop model for winter wheat using MODIS LAI images. *Agric. Water Manage.* **164**, 304–316 (2015).
59. Xie, Q. *et al.* Leaf Area Index Estimation using vegetation indices derived from airborne hyperspectral images in winter wheat. *IEEE J-STARS*. **7**, 3586–3594 (2017).

Acknowledgements

This work was financially supported in part by the National Key Research and Development Program of China (2018YFD0300800), the National Natural Science Foundation of China (41271415, 31771711), the Project Funded by China Postdoctoral Science Foundation (2019M650125), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) and Top-notch academic programs project of Jiangsu higher education institutions (PPZY2015A060). The Jiangsu map vector data was supported by Yangtze River Delta Science Data Center, National Science & Technology Infrastructure of China. Authors are thankful to the anonymous reviewers and editor for their valuable comments in improving the quality of the manuscript.

Author contributions

C.W.T. and W.S.G. conceived the research. C.W.T. and X.X.Z. designed and performed the experiments. C.W.T., P.P.Z., F.Y. and Z.X.W. prepared and revised the manuscript. C.W.T., W.M. and W.X.L. analyzed the data and provided technical. All the authors have reviewed the manuscript and agreed the submission and publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.-S.G. or C.-W.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020