

# Analysis of NDVI Data for Crop Identification and Yield Estimation

Jing Huang, Huimin Wang, Qiang Dai, and Dawei Han

**Abstract**—Crop yield estimation is of great importance to food security. Normalized Difference Vegetation Index (NDVI), as an effective crop monitoring tool, is extensively used in crop yield estimation. However, there are few studies focusing on the aspect of mixed crops grown together. In this study, a correlation-based approach for crop yield estimation is applied to three small counties (Jianshui, Luliang, and Qiubei) in the Nanpan River basin, Yunnan Province of China, and three main crops (paddy rice, winter wheat, and corn) in these areas are selected. Based on the correlation analysis between MODIS–NDVI data and crop yield, the crop planting areas as well as the best periods for a reliable estimation are identified. The best time is found approximately coinciding with the periods of heading, flowering, and filling of the crops. By Akaike’s information criterion, the most fit regression models with extracted NDVI in the corresponding crop planting areas are determined. They work reasonably well in small regions, especially in the areas where crop types are unknown exactly. Further improvements to the regression models are possible by incorporating other physical factors such as soil types and geographical information.

**Index Terms**—Crop identification, normalized difference vegetation index (NDVI), remote sensing, yield estimation.

## I. INTRODUCTION

CROP PRODUCTION plays a vital role in food security and economic development. In the past years, the fluctuation of crop yield in China attracted a great concern in economy and even lead to food crisis of the whole country. For example, a continuous drought in Yunnan Province in 2011 caused approximately 340 million U.S. dollar loss because of

Manuscript received October 30, 2013; revised May 07, 2014; accepted June 23, 2014. Date of publication August 04, 2014; date of current version January 06, 2015. This work was supported in part by the National Social Science Foundation of China under Grant 12&ZD214, in part by the Project of Science and Technology of Yunnan Province under Grant 2010CA013, in part by the Research Fund for the Doctoral Program of Higher Education of China under Grant 20120094110018, in part by the Research Fund of “333” Project on Phase IV of Jiangsu Province under Grant BRA2012131, and in part by the Research Innovation Program for College Graduates of Jiangsu Province under Grant CXLX12\_0259. (*Corresponding author:* Huimin Wang.)

J. Huang is with the State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China, with the Institute of Management Science, Hohai University, Nanjing 210098, China, and also with the Department of Civil Engineering, University of Bristol, Bristol BS8 1TH, U.K. (e-mail: huangjingshow@hotmail.com).

H. Wang is with the State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China, and also with the Institute of Management Science, Hohai University, Nanjing 210098, China (e-mail: hmwang@hhu.edu.cn).

Q. Dai and D. Han are with the Department of Civil Engineering, University of Bristol, Bristol, BS8 1TH, U.K. (e-mail: Q.Dai@bristol.ac.uk; D.Han@bristol.ac.uk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JSTARS.2014.2334332

the substantial crop yield reduction, and further resulted in significant price rises.

Application of remote sensing data on agriculture and crop production has been popular [1]–[4], especially based on the predictive empirical model, because it is possible to estimate crop yield efficiently and quantitatively by such data. The Normalized Difference Vegetation Index (NDVI) data, a product derived from the satellite data, could be used to estimate the vegetation health and monitor changes in vegetation. It is calculated from the normalized total reflectance of the red and near infrared bands, ranging from  $-1.0$  to  $+1.0$  [5]. A higher NDVI indicates more green coverage whereas a lower NDVI signifies the loss of growth and vigor of the crop. Therefore, the NDVI temporal profile rises with the growth of crops typically, reaches the peak level during the productive stage, and declines around the harvest [6]. NDVI data derived from NOAA (National Oceanic and Atmospheric Administration)-Advanced Very High Resolution Radiometer (AVHRR) have been used to forecast crop yield in many countries since 1980s. For instance, monthly global area coverage (GAC) NDVI data were used to estimate crop yield in Mediterranean African countries [7]; decadal average NDVI data were adopted to forecast corn yield in Swaziland [8]. Similar studies have been conducted for various crop types in many other regions, e.g., wheat in Morocco [9] and Italy [10], millet in Senegal [11], and corn in Kenya [12].

NDVI data from the new Moderate Resolution Imagine Spectroradiometer (MODIS) available from 2000 to present have made significant improvements addressing the shortcomings from AVHRR [13], [14]. In recent years, studies have been conducted to set up the relationship between crop yield and NDVI data from MODIS. The spatial accumulative and smoothed MODIS-NDVI data were used to estimate winter wheat production in Shandong Province, China [15]; MODIS data were used to establish an empirical approach for winter wheat in Kansas and Ukraine [16]; and weighted temporal series NDVI from MODIS were applied to forecast sugarcane yield in Kenya [17].

Although previous studies found useful statistical relationships between final crop yield and different forms of NDVI data all around the world, few studies have been conducted in small regions (e.g., at a county level). Moreover, most of the studies were based on specified experimental fields or a region with monotype and known crops, no studies have been carried out in areas with unknown crop types. This study aims to estimate crop yield for different crops grown in a small area by establishing statistical models with MODIS-NDVI data.

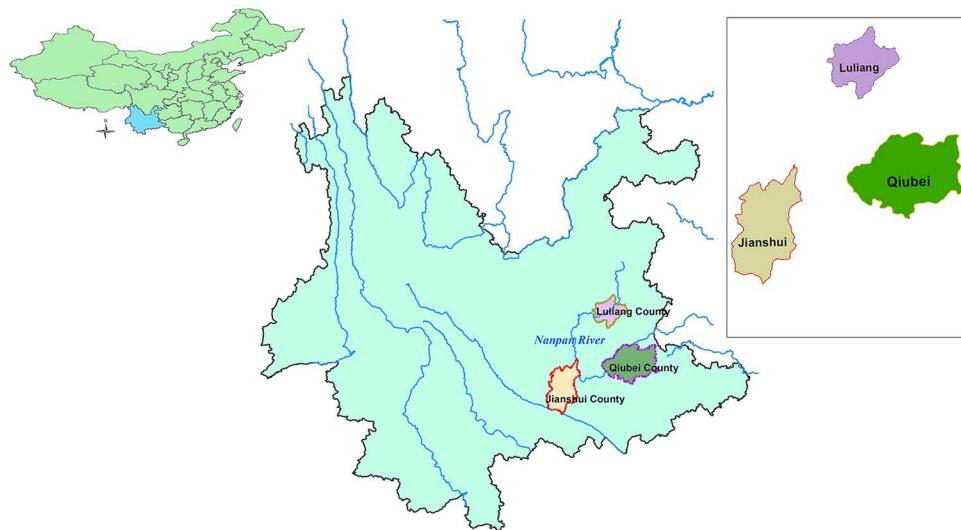


Fig. 1. Three study counties in Nanpan River Basin, Yunnan Province, China.

In particular, the first efforts are devoted to analyzing the correlation between time-series NDVI data and crop yield for each crop type. The research is then directed to the identification of the geographical distribution of these crops and the best period for a reliable crop yield forecast. Next, based on the extracted NDVI data in the selected planting areas during the best forecast periods, regression models are developed to estimate crop yields. The ultimate objective is to put forward a method to forecast crop yield in a small region with available NDVI data in order to assess regional agricultural risk in the future.

## II. MATERIALS AND METHODS

### A. Study Areas

In this study, three counties (Jianshui, Qiubei, and Luliang) are selected as the study areas in Nanpan River basin in Yunnan Province, China (Fig. 1). The whole study area extends northward from  $23^{\circ}13'N$  to  $25^{\circ}17'N$  latitudes and westward from  $102^{\circ}35'E$  to  $104^{\circ}34'E$  longitudes. Jianshui located in the southwest part, covering an area of  $3940 \text{ km}^2$ . The climate is subtropical monsoon with an average temperature of  $18\text{--}20^{\circ}\text{C}$  and an average annual precipitation of  $800\text{--}1000 \text{ mm}$ . The average temperature ranges from about  $12^{\circ}\text{C}$  in January to over  $23^{\circ}\text{C}$  in June and July, and the precipitation is concentrated in May–August, over  $100 \text{ mm}$  every month. Qiubei is situated in the east. The mean annual precipitation is slightly over  $1000 \text{ mm}$ , most of which occurs from May to August. The annual mean temperature is  $16\text{--}17.5^{\circ}\text{C}$ , with the lowest mean value of  $9^{\circ}\text{C}$  and highest approximately of  $22^{\circ}\text{C}$ . Luliang is located in the northern part of the basin with an area of  $2096 \text{ km}^2$ . This area is surrounded by mountains, and the middle is a wide plain. The weather is mild with an average temperature of  $15.5^{\circ}\text{C}$  all year round, and the average annual precipitation of around  $900 \text{ mm}$ . Generally, the weather in Qiubei is relatively mild and moist, whereas the weather in Jianshui is slightly arid. Luliang is in an intermediate location which is neither humid nor dry.

### B. Crop Yield Data and MODIS-NDVI Data

In the study sites, winter wheat (*Triticum aestivum* L.), summer corn (*Zea mays* L.), and paddy rice (*Oryza sativa* L.) are major crop types. The production of the three crops accounts for over two-thirds of whole grain outputs in these sites. Winter wheat grows from mid-October to the end of the following April, and the growing season of summer maize is from the end of winter wheat growth season to September. The seeding of paddy begins from the end of March, and the harvest period is during early October (Fig. 2). Yield data of three main crops in every county (Jianshui, Qiubei, and Luliang) from 2001 to 2009 are obtained from the Yunnan Statistical Yearbook. The yield data of 2000 at the county level cannot be obtained from the Yearbook. However, it is provided by Water Resources and Hydropower Research Institute of Yunnan Province. It is worthy to note that the winter wheat yield data of Luliang in 2000 are inaccurate (discovered during a data quality check). Therefore, they are replaced by the mean winter wheat yield of Yunnan Province, which are obtained from National Bureau of Statistics of China (<http://data.stats.gov.cn>).

The processed global MODIS 16-day (temporal resolution) NDVI data in 1-km resolution for the years 2000–2009 are acquired from the Earth Observing System Data and Information System (EOSDIS) of National Aeronautics and Space Administration (NASA). Two sinusoidal grids that cover the whole study areas are downloaded. The data are composited over 16-day time intervals, so there are 23 periods of NDVI data from January 1 to December 31 in each year. The NDVI data are not complete for the whole period that begins from 18th February, 2000. However, the missing data do not affect paddy rice and maize because they start to be planted after March. For winter wheat, the crop yield in 2000 is related with NDVI both in late 1999 and early 2000. Due to the lack of NDVI data in 1999, yield data of 2000 cannot be used. Therefore, crop yield data during the year 2001–2009 and NDVI data from late 2000 to early 2009 are utilized for winter wheat.

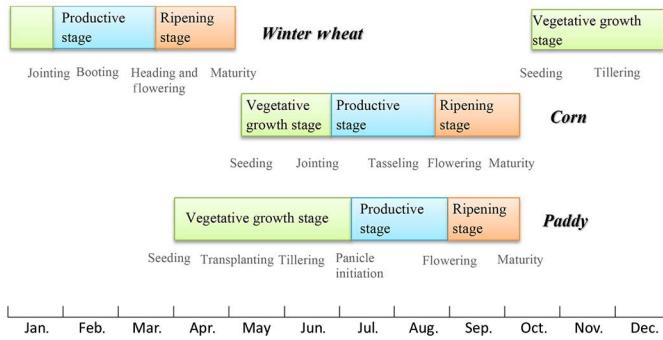


Fig. 2. The growing paths of three crops (winter wheat, corn and paddy).

The basic geographic information data of China are downloaded from the National Fundamental Geographic Information System. They are used to extract MODIS-NDVI images for every county with ArcGIS (Geographic Information System) tool. Besides, the Chinese land cover map is obtained from the Cold and Arid Regions Science Data Centre at Lanzhou [18]. The map is used for the crop land identification which delineates the areas under cultivation in the study sites; it is also used to eliminate the influence of nonagriculture crops on NDVI. Therefore, only those areas of agricultural usage are extracted for further analysis.

### C. Methods

Statistical analysis has been done separately for every crop in each study site using the 16-day NDVI values and crop yield data. Particularly, NDVI data of individual periods in every pixel are used to set up relationships with the yields of paddy rice, winter wheat, and corn, respectively, from year 2000 to 2009.

Based on the calculation of Pearson's correlation coefficient, the data set of  $r$  for each crop in each study site is shown as follows:

$$r^c = \begin{bmatrix} r_{11}^c & \cdots & r_{1t}^c \\ \vdots & \ddots & \vdots \\ r_{p1}^c & \cdots & r_{pt}^c \end{bmatrix} \quad (1)$$

where  $p$  is the number of pixels in each study site (there are 769 pixels in Jianshui, 935 in Luliang, and 1360 in Qiubei);  $t$  is the period of the NDVI data, including 23 periods; and  $c$  stands for the crop types.

After the correlation analysis is applied, three steps should be implemented based on the matrices. First, a threshold value of the correlation coefficient ( $r = 0.6$ ) is set. The efforts are then turned to classify the planting areas of these crops. Based on the  $r$  matrix, the pixel  $p$  in which  $r$  value over the threshold value in any crop growth period is recognized as the planting area for the crop type  $c$ . Such a pixel belongs to set  $P^c$ . The corresponding period  $t$  with the highest  $r$  belongs to set  $T^c$  that is recognized as the best period for crop yield estimation in every pixel. Lastly, for every crop, the mean 16-day NDVI data of all extracted planting areas in every period are calculated in each site as follows:

$$\text{NDVI}_t^c(i) = \frac{\sum_{j=1}^{m^c} \text{NDVI}(j)_t(i)}{m^c} \quad (2)$$

where  $m^c$  is the number of elements in  $P^c$ , which indicates the number of planting pixels of the crop type  $c$ ;  $t$  is the number of period of the NDVI data; and  $i$  is the number of the study year.

Since  $T^c(p)$  is the best estimation period for the pixel  $p$ ,  $N_t^c$  is recorded as the count of pixels for each period  $t$ . Thus, the periods when the values of  $N_t^c$  are relatively high are combined together as the best forecast period for establishing a crop yield estimation model.

The correlation analyses are conducted again using the mean NDVI during the best forecast periods and crop yield. Regression analyses, both linear and nonlinear models, are performed to predict the crop yield from NDVI data by Statistical Product and Service Solutions (SPSS) software. The mean NDVI data during the best forecast periods are taken as the independent variable and the crop yield as the dependent variable. Results from linear, logarithmic, inverse, quadratic, cubic, power, S, growth, and exponential models are compared in order to find the most suitable model. In order to compare the ability and reliability of these regression models, Akaike's information criterion (AIC) is introduced to determine the best fit regression equation (due to the short data records, cross-validation is not appropriate in this study). In 1974, AIC was proposed to statistically identify different models [19]. The AIC value for a given model is a function of its maximized log-likelihood ( $\mathcal{L}$ ) and the number of estimable parameters ( $k$ ), which is defined by

$$\text{AIC} = -2 \log \hat{\mathcal{L}}(\hat{\theta}) + 2k \quad (3)$$

where  $\mathcal{L}$  is the likelihood function,  $\hat{\theta}$  is the maximum likelihood estimate of  $\theta$ , and  $k$  is the number of free parameters in the model. In ordinary least squares (OLS) regression cases,  $\mathcal{L}(\hat{\theta}) = -\frac{n}{2} \log \left( \frac{\text{RSS}}{n} \right)$ , thus

$$\text{AIC} = n \log \left( \frac{\text{RSS}}{n} \right) + 2k \quad (4)$$

where RSS denotes residual sum of squares, which is the estimated residual of the fitted model.

A small sample AIC ( $\text{AIC}_c$ ) is used when sample size ( $n$ ) is small relative to the number of parameters (rule of thumb,  $n/k \leq 40$ ) [20], [21], which is calculated as follows:

$$\text{AIC}_c = \text{AIC} + \frac{2k(k+1)}{n-k-1}. \quad (5)$$

In this study, the model with a minimum  $\text{AIC}_c$  value is chosen as the best one to fit the data.

Moreover, as a geographic information system, ArcGIS tool is applied to provide geographically referenced information, such as location, elevation, and rivers, for crop distribution and correlation distribution.

## III. RESULTS

### A. Correlation Coefficient Analysis

Table I shows the results of correlation coefficients between the mean NDVI and crop yields. All the  $p$  values are below 0.05, which indicate that correlations are significant at 95% confidence level. Through the results, the correlation

TABLE I  
CORRELATION COEFFICIENT, *p*-VALUES, AND THE BEST PERIODS FOR CROP YIELD FORECASTING

Region	Crop Type	<i>r</i>	<i>p</i>	Best periods Date/month (period)
Jianshui	Paddy rice	0.682	0.0298	14/09–15/10 (17–18)
	Winter wheat	0.724	0.0275	19/12–03/01 (23) and 01/01–01/02 (1–2)
	Corn	0.782	0.0075	30/09–31/10 (18–19)
Luliang	Paddy rice	0.677	0.0317	12/07–12/08 (13–14)
	Winter wheat	0.934	0.0001	02/02–05/03 (3–4) and 22/03–06/04(6)
	Corn	0.811	0.0044	25/05–09/06 (10) and 12/07–27/07 (13)
Qiubei	Paddy rice	0.658	0.0388	10/06–25/06 (11) and 14/09–31/10 (17–19)
	Winter wheat	0.746	0.0210	16/10–02/12 (19–21) and 18/02–05/03 (4)
	Corn	0.716	0.0198	25/05–11/7 (10–12) and 30/09–31/10 (18–19)

coefficients vary with crop types and planting locations. In Jianshui, the paddy rice yield is highly correlated ( $r = 0.68$ ) with the mean NDVI. The similar but slightly lower values are found in Luliang and Qiubei. For winter wheat, the correlation coefficients are approximately 0.93 in Luliang, 0.75 in Qiubei, and 0.72 in Jianshui. The correlations between corn yield and the NDVI data are also strong, with the correlation coefficient up to 0.81 in Luliang and 0.78 in Jianshui.

The best periods for making reliable yield forecast are also presented in Table I. For paddy rice, the best period is from mid-September to mid-October in Jianshui, from July 12 to August 12 in Luliang, and period 17–19 together with period 11 in Qiubei. Since winter wheat grows through winter, the crop yield is affected not only by the conditions during planting periods in the previous year but also by that in early of the same year. Therefore, the best period for winter wheat is the combination of last 16 days of the previous year and the first month in the same year in Jianshui. Period 19–21 and period 4 join together as the best period for winter wheat in Qiubei, whereas the period from early February to early April except a period from March 6 to 21 in the same year is the best in Luliang. The corn yield is correlated with the mean NDVI of the period from the end of September to the end of October in Jianshui. Together with the same period in Jianshui, period 10–12 is also vital in Qiubei. In Luliang, period 10 and 13 are the best periods for the corn yield forecasting.

#### B. Crop Identification and Spatial Distribution

Because paddy and corn are grown in rotation with winter wheat, the combination of paddy and winter wheat or that of corn and winter wheat may be planted in the same field. In some pixels, part of the areas is planted with paddy while others are occupied by corn. It is also possible that three crops share one pixel.

Table II illustrates the planting areas for various crop combinations in three counties. In Jianshui, corn is the most widely grown crop. Over 30% of the crop land is exclusive for corn, and another 141 pixels are used for corn together with other crops. Compared with large areas for corn, only

about one quarter of the crop land is planted with paddy rice, of which 54 pixels are especially set for paddy rice. Winter wheat occupies approximately 24% of the crop land including that for rotation and all crops. However, there is still about 30% of the land planted with other crops besides the studied three crops or in fallow. In Qiubei, corn is also the most popular crop, covering over 60% of the crop land. Paddy area and winter wheat area are almost the same, 657 pixels and 640 pixels, respectively. It is of interest to note that exclusive crop lands of three crops only occupy about 27% of the total land in Qiubei. Over 20% the crop land is for all the three crops grown together, which is the largest among three study sites. It means that crops are mix-planted in a substantial area. In Luliang, the land for winter wheat is over half, whereas that of paddy and corn contributes a small part. There is also one-third of the crop land planted with other crops or bare, which is the highest proportion in all the sites. In summary, corn is the most widely planted crop in Jianshui, whereas winter wheat is popular in Luliang.

In Fig. 3, the mean NDVI during 2000–2009 of the selected corn planting areas in Luliang County illustrates the characters of the corn group. There are 79 pixels in which corn is planted exclusively. From those curves, NDVI starts to go up during period 9–10 and peaks around period 14. In period 19, most of the NDVI values are below 0.5.

From Figs. 4 to 6, the crop distribution could be discriminated more clearly. The inclined rectangular pixels are croplands in all the figures, which are extracted from the study in [22]. The triangles signify the land grown corn; the squares are for the land of wheat; and the circles are for the land of paddy. As shown in Fig. 4, the croplands are mainly concentrated in the middle of Jianshui which is a fertile river valley. The crops are rarely grown in the northern part, primarily due to the mountainous terrain. There is also a large planting area in the north-west, where corn is mostly planted. In consideration of the correlation coefficients in these pixels, higher correlation coefficients are found in the middle region for paddy, and winter wheat is more strongly related to NDVI in the areas of high elevations.

There are more croplands in Luliang, occupying nearly half of the total area especially in the middle and southern region

TABLE II  
PLANT AREAS (NUMBER OF PIXELS) AND THE DISTRIBUTION OF DIFFERENT CROP COMBINATIONS IN THREE STUDY SITES

Region	Plant crops	Number of pixels (1 km × 1 km)	Percentage of the total crop land	Comments
Jianshui	Paddy	54	7.02	
	Wheat	73	9.49	
	Corn	238	30.95	
	Paddy and wheat	51	6.63	Paddy areas: 196 pixels
	Corn and wheat	50	6.50	Wheat areas: 203 pixels
	Paddy and corn	62	8.06	
	Paddy, wheat, and corn	29	3.77	Corn areas: 397 pixels
	Other crops or none	212	27.57	
<b>All</b>		<b>769</b>	<b>100</b>	
Luliang	Paddy	44	4.71	
	Wheat	302	32.30	
	Corn	79	8.45	
	Paddy and wheat	59	6.31	Paddy areas: 145 pixels
	Corn and wheat	104	11.12	Wheat areas: 489 pixels
	Paddy and corn	18	1.93	
	Paddy, wheat, and corn	24	2.57	Corn areas: 225 pixels
	Other crops or none	305	32.62	
<b>All</b>		<b>935</b>	<b>100</b>	
Qiubei	Paddy	91	6.69	
	Wheat	119	8.75	
	Corn	157	11.54	
	Paddy and wheat	64	4.71	Paddy areas: 657 pixels
	Corn and wheat	169	12.42	Wheat areas: 640 pixels
	Paddy and corn	214	15.74	
	Paddy, wheat, and corn	288	21.18	Corn areas: 828 pixels
	Other crops or none	258	18.97	
<b>All</b>		<b>1360</b>	<b>100</b>	

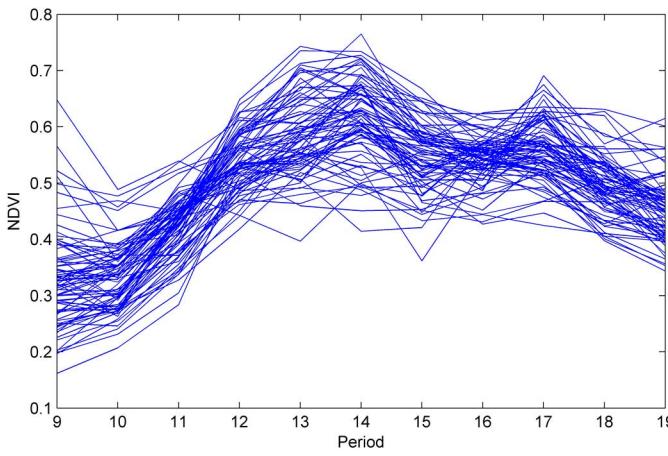


Fig. 3. The mean NDVI of selected corn planting pixels in Luliang County.

[see Fig. 5(a)]. As the most widely planted crop in Luliang, winter wheat is distributed throughout the whole flat area. The corn and paddy crops are scattered across the entire region. However, just a few of them are grown in the same pixel. From Fig. 5(b), the correlation coefficients are shown in every

corn area in Luliang. Higher values are found intensively in the central plain area, mainly in the north-west side of the southern highland and the north-west side of the Nanpan River.

From Fig. 6, the crops are mix-planted in the middle and southern region of Qiubei. The three crops cover similar growing areas. Moreover, most winter wheat and corn share the same area. For the correlation coefficient for every pixel, the values in winter wheat pixels are consistent with those in corn pixels, whereas the values in paddy areas are in the opposite direction to them.

### C. Regression Model of Crop Yield Estimation

In Table III, the results of curve estimation in Luliang are shown, including the coefficient of determination ( $R^2$ ) and equations. There are several models that fit the crop yield data well. Therefore, AIC is introduced to measure the relative qualities of the available statistical models, of which the lower value implies the minimum information loss [23]–[25]. Based on the comparison of  $R^2$  and  $AIC_c$  values, the S model ( $y = e^{a+b/x}$ ) is chosen as the best suitable model for the crop yield estimation.

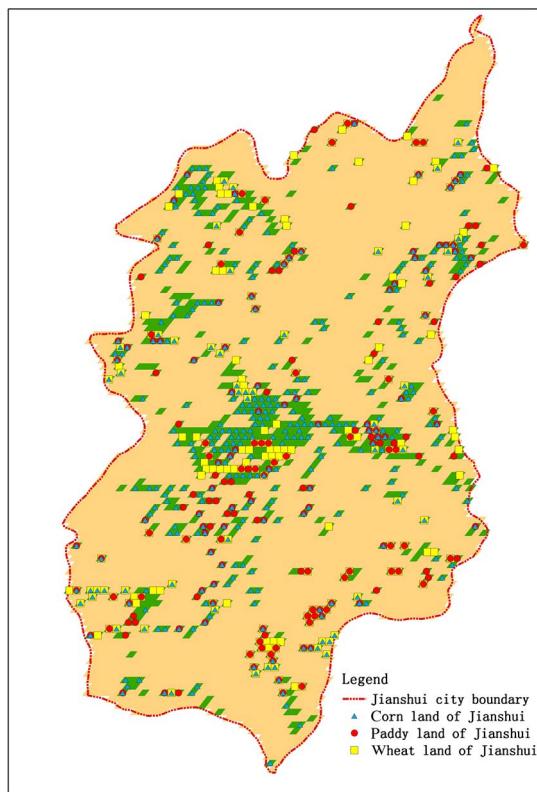


Fig. 4. Crop land distribution in Jianshui.

Since the same curve estimation analyses have been conducted for crops in the other two sites, the suitable regression models are chosen which could be seen in Fig. 7. Actually, the relationship between corn yield and NDVI data could be explained by three models (compound, exponential, and growth) equally well in Jiangsui. Finally, the growth model is selected as the best fit model in consideration of the consistency of the function form. The same situation exists in paddy yield estimation in Qiubei. Compound, exponential, and growth functions performed as well in fitting the paddy yield with NDVI data, of which the growth model is regarded as the most suitable one. In conclusion, the S model and growth model are mainly used for the yield estimation of paddy, winter wheat, and corn in this study. The coefficients of determination ( $R^2$ ) range from 0.427 to 0.484 for paddy, 0.565 to 0.898 for winter wheat, and 0.547 to 0.692 for corn. Fig. 8 illustrates the contrast between the predicted winter wheat yield and the actual one in Luliang; the root mean square error (RMSE) is about 134 kg/ha. In addition, the RMSE is 177.83 kg/ha for paddy, 130.48 kg/ha for winter wheat, 206.59 kg/ha for corn in Jianshui; 217.55 and 262.70 kg/ha for paddy rice and corn, respectively, in Luliang; and 355.26, 110.77, 171.11 kg/ha in Qiubei.

#### IV. DISCUSSION

In our initial study, the correlation between crop yield and 16-day NDVI data in every period was conducted, in an attempt to find a simple way for crop yield estimation.

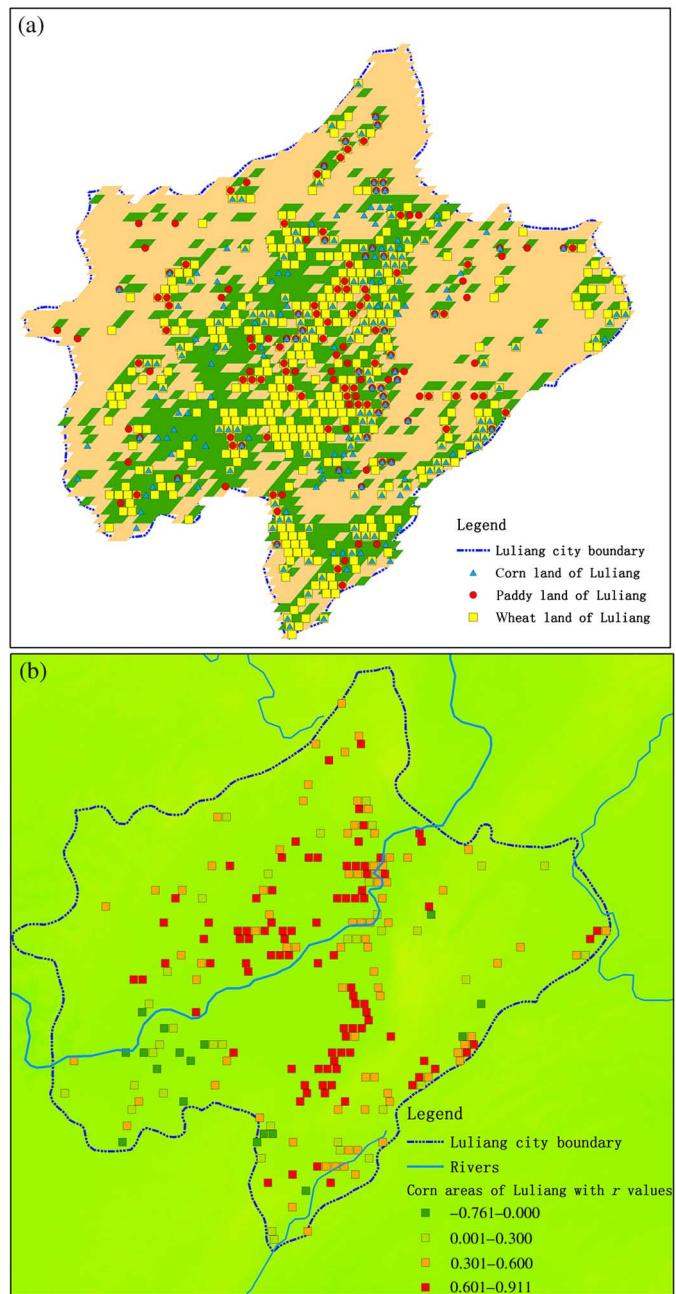


Fig. 5. (a) Crop land distribution in Luliang, (b) the correlation coefficient in every corn area in Luliang.

However, no acceptable or reasonable result was found from the direct correlation analysis. Moreover, a higher correlation coefficient was obtained with the NDVI integrated over several periods. The results were even better than relating the averaged NDVI during whole crop growth stages to crop yield. The reason is that some periods are more important for crop yield, such as the heading to flowering stages for winter wheat.

In practical situations, there are many types of crops growing together in a small region. However, the crop yield is only related to the NDVI data extracted from the corresponding crop planting areas. Therefore, identification of the crop planting area is another crucial part of the estimation model

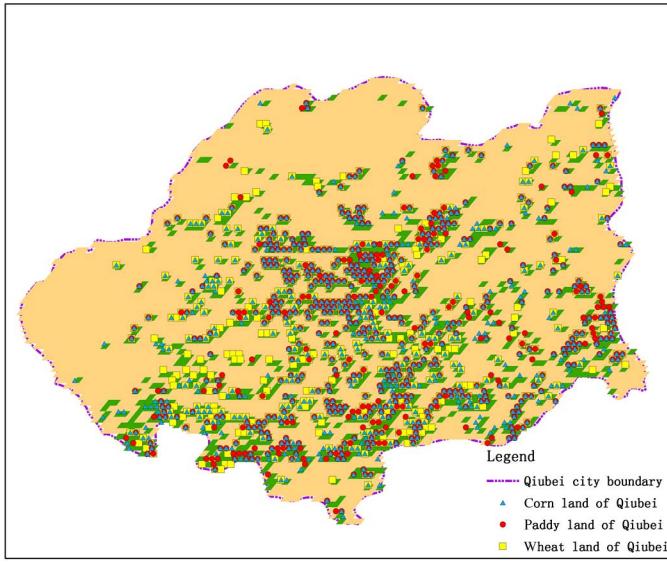


Fig. 6. Crop land distribution in Qiubei.

development and implementation. In the previous studies, the crop type was singular for the study site or a main crop occupied a dominant position hypothetically, so the correlation was established between the NDVI throughout the entire area and the crop yield. For example, the country level NDVI was used to relate winter wheat yield [15]. Such NDVI data, however, could not reflect crop yield effectively in our study because the NDVI data represented the growth condition of all the crops planted in this area. In comparison, a higher correlation in an experimental area was found, where a single crop type was grown [26]. Therefore, the extraction of NDVI for a specific crop land would not only make a significant improvement in crop yield estimation but also lead to a realistic approach. Since only the overall cropland distribution information is available, a statistical method is used to help identify specific crop areas based on the assumption that NDVI data of an area planted with a certain crop are strongly related to the yield of such crop. Therefore, the pixel in which the highest correlation is obtained between 16-day NDVI data in any growth period and the crop yield is recognized as the crop planting area. A previous study has been reported which described a similar statistical method on crop land identification based on computing per-pixel interannual correlations between NDVI and crop yield [7]. However, the study just distinguished the agriculture vegetation from nonagriculture vegetation and still addressed one generic crop type as agriculture vegetation in a whole country. In our study, the planting area identification of multiple crops is tackled. This is more in line with the actual planting situation in the study sites in Yunnan, China. In addition, our study explored 16-day MODIS-NDVI data at 1-km resolution used in a humid climate with more cloudy days (hence a very challenging region for satellite remote sensing), while in that study, the monthly NDVI data from NOAA-AVHRR at 8-km resolution were applied in Mediterranean African Countries where the climate is relatively arid with more clear days. The research in [7] was based on the assumption that those pixels of cropland remain approximately

constant over the study period at a GAC pixel scale. The same hypothesis was proposed in this paper since there is no additional information on the crop areas distribution. However, it may be possible to use a moving window to detect the crop pattern change if there is a long record of data (in a similar way to detecting the trend in climate change).

In the crop land identification, the threshold of the correlation coefficient between every 16-day NDVI data and crop yield was set as 0.6 based on the previous studies. Such a value or above indicates the strong correlation between NDVI data in the corresponding period and crop yield. For example, a correlation coefficient over 0.71 was found for corn in Swaziland [8]; another one up to 0.92 was found for winter wheat in Morocco [9]. However, the threshold may vary with different field experimental scenarios.

The approach used in this study makes the crop identification easier without the use of the time-series NDVI and the crop-growing characters. In our study, the cluster analysis has also been conducted with the NDVI data in order to discriminate the crop types. Particularly, K-means analysis is adopted, and eight partitions are initialized according to different crop combinations. Although the results show the separation of the NDVI time-series data, it is hard to distinguish the paddy areas from the corn areas due to their similar growth paths. In addition, a decision tree method is also tried to keep the crops apart. But, the growth of crops is strongly affected by the planting environments, which results in the swinging values of NDVI data and the unsteady date of crop growth stages. Therefore, the decision tree approach also fails to identify those crops in the study sites.

The best period for crop yield forecasting is associated with the crop growth cycle. For example, according to the NDVI curves of the selected corn planting areas in Luliang, NDVI values begin to rise around the middle of May, peak near the middle of July, and keep falling after early September. Such a pattern is consistent with the features of NDVI during the growth stages of corn completely [27]–[29]. Generally, corn is seeded around early to mid-May. Approximately two months later, it is around the tasseling stage when all the leaves are exposed. Through the productive stage and maturity, it is always harvested in October. Therefore, the approach adopted in this study is proved to be an effective way in crop identification. For paddy rice, NDVI data are highly correlated with its yield in both the heading and maturity stages. The jointing stage also plays a crucial role, especially in Luliang. Generally, those best prediction periods in this paper approximately coincide with the stages of heading, flowering, and filling of the crops, all before the harvest time. The result agrees with those in previous studies [8], [30]–[32]. It is probably due to the fact that these three stages are the most crucial moments to crop yields, in which any water stress would lead to the yield reduction. When studying the relationships between NDVI data and some crops grown in China, the NDVI data correlated well with the winter wheat yield from the jointing stage to the flowering stage in Henan Province [33]. It is interesting to note that the best periods for paddy rice and corn yield estimation are during the filling stage in arid areas (Jianshui) and around the

TABLE III  
RESULTS OF THE CURVE MODELS FOR EVERY CROP IN LULIANG AND CORRESPONDING AKAIKE'S INFORMATION CRITERION (AIC) VALUES

Crop type	Model	Equation	$R^2$	Std. error of the estimation	AIC <sub>c</sub>
Paddy rice	Linear	$Y = 6820.116 + 3093.41 * \text{NDVI}$	0.458	244.1	110.224
	Logarithmic	$Y = 9692.953 + 1976.973 * \ln(\text{NDVI})$	0.461	243.5	110.169
	Inverse	$Y = 10774.696 - 1249.932/\text{NDVI}$	0.462	243.2	110.144
	Compound	$Y = 7013.340 * 1.424^{\text{NDVI}}$	0.456	244.4	110.243
	Power	$Y = 9737.903 * \text{NDVI}^{0.226}$	0.459	243.6	110.183
	S	$Y = e^{9.308 - 0.143/\text{NDVI}}$	<b>0.461</b>	<b>243.2</b>	<b>110.148</b>
	Growth	$Y = e^{8.856 + 0.354 * \text{NDVI}}$	0.456	244.4	110.243
Winter wheat	Exponential	$Y = 7013.340 * e^{0.354 * \text{NDVI}}$	0.456	244.3	110.240
	Linear	$Y = -3723 + 14250 * \text{NDVI}$	0.873	147.0	100.084
	Logarithmic	$Y = 7343.159 + 5846.083 * \ln(\text{NDVI})$	0.878	144.2	99.691
	Inverse	$Y = 7969.733 - 2388.650/\text{NDVI}$	0.881	142.3	99.421
	Compound	$Y = 26647.973 * \text{NDVI}^{2.855}$	0.892	155.2	101.157
	Power	$Y = e^{10.500 - 1.168/\text{NDVI}}$	<b>0.898</b>	<b>150.3</b>	<b>100.519</b>
	S	$Y = e^{4.791 + 6.949 * \text{NDVI}}$	0.885	160.7	101.864
Corn	Growth	$Y = 120.363 * e^{6.949 * \text{NDVI}}$	0.885	160.7	101.864
	Exponential	$Y = 2748.893 + 8154.813 * \text{NDVI}$	0.658	297.9	114.278
	Linear	$Y = 9553.833 + 3914.448 * \ln(\text{NDVI})$	0.665	294.6	114.052
	Logarithmic	$Y = 10562.538 - 1854.818/\text{NDVI}$	0.668	293.3	113.965
	Inverse	$Y = 3688.984 * 3.411^{\text{NDVI}}$	0.679	300.9	114.477
	Compound	$Y = 10273.144 * \text{NDVI}^{0.589}$	0.687	296.4	114.171
	Power	$Y = e^{9.390 - 0.280/\text{NDVI}}$	<b>0.692</b>	<b>293.8</b>	<b>114.001</b>
Soybean	S	$Y = e^{8.213 + 1.227 * \text{NDVI}}$	0.679	300.9	114.478
	Growth	$Y = 3688.984 * e^{1.227 * \text{NDVI}}$	0.679	300.9	114.477

heading stage in moderate regions (Luliang). In mild zones (Qiubei), NDVIs in the heading period and filling stage are of the same importance. For winter wheat, the best period is during the jointing stage in arid regions, which is earlier than that in moderate regions. In mild zones, two stages (tillering and heading) are of equal importance for winter wheat. The probable reason for this situation is that the crops are of great vigor due to the mild weather for a long period from the vegetation growth stage to the ripening stage.

Compared with the linear regression model, other nonlinear models such as exponential regression models and power functions show better performances on crop yield forecasting. The cubic polynomial regression was proved better than linear and exponential regression models [33]. In [34], the NDVI and corn grain yields had a significant exponential relationship. A power function was found suitable for the relationship between NDVI and soybean grain yield appropriately [35]. Considering the determined coefficient of the selected models, the results compare well with the findings in [8], who found that the NDVI explained 51–68% of corn yield variation in Swaziland. Similar results are also found in [36] and [37] for winter wheat in China. Relatively higher determination coefficients, all over 0.6, are found when relating the NDVI to winter wheat yield [38]. The research in [33] reported that over 75% of the yield could be explained by NDVI data. The higher  $R^2$  values in these studies are probably because of the extensive planting of the crops.

For the planting sites, the relatively high  $R^2$  values ranging from 0.461 to 0.811 are found in Luliang, followed by Jianshui ranging from 0.484 to 0.626, and lastly Qiubei ranging from 0.417 to 0.570 (Fig. 7). The different terrains in these sites probably play a key role. Because of the shade from the hill, the values of NDVI are always lower than the original one, and hence the real growth conditions of the crop may not be reflected by the observed NDVI data accurately. Therefore, the uniform correlations between NDVI and crop yield in different sites are mainly due to the topography. Compared with the correlation distribution maps for winter wheat in these three sites, smaller crop areas are found in hill shades in Luliang (see Fig. 9). This is probably the reason for the relatively high correlation between winter wheat yield and the corresponding NDVI data.

Although the proposed models agreed well with the data, the overfitting may easily arise because of the short data records. For this reason, it is not a proper choice to divide the data into calibration and validation data sets and evaluate the fitted models by cross-validation. Instead, AIC is used to find the best models approximate to the unknown data generating process, which is asymptotically equivalent to cross-validation [39]. The AIC approach deals with the tradeoff between the goodness of fit of the model and the complexity of the model, which is also effective for situations with limited data [40]. We will strive to carry out more evaluated processes if longer data could be obtained in the future.

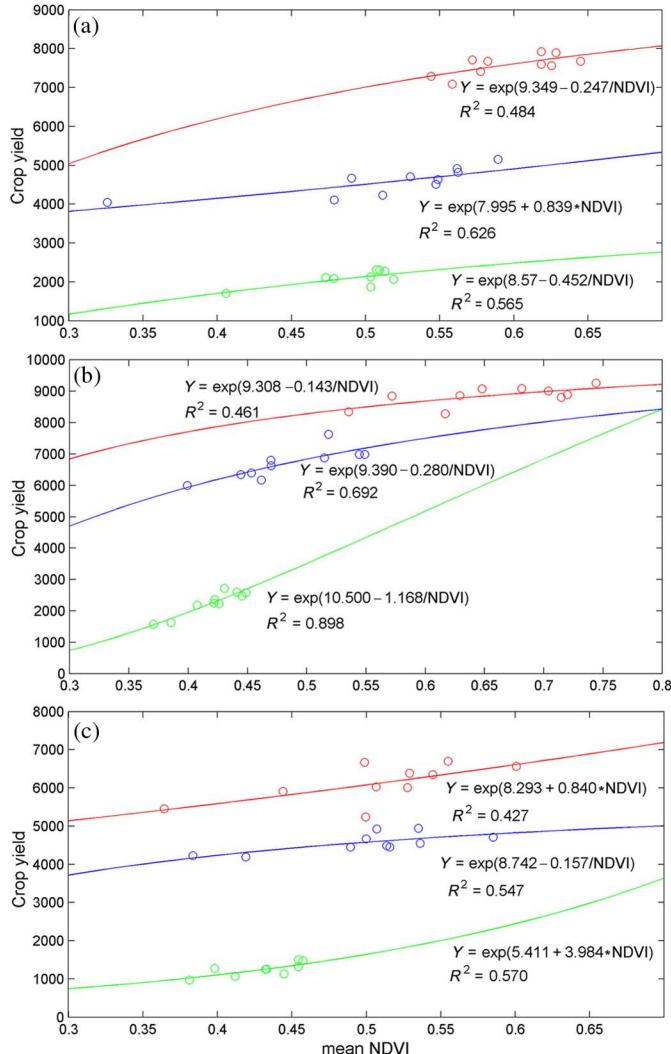


Fig. 7. Regression models for yield (kg/ha) estimation of crops [red lines (upper lines in b/w version) for paddy, blue lines (middle lines in b/w version) for corn, and green lines (lower lines in b/w version) for winter wheat] with the mean NDVI during the best periods (see in Table I) in Jianshui (a), Luliang (b), and Qiubei (c).

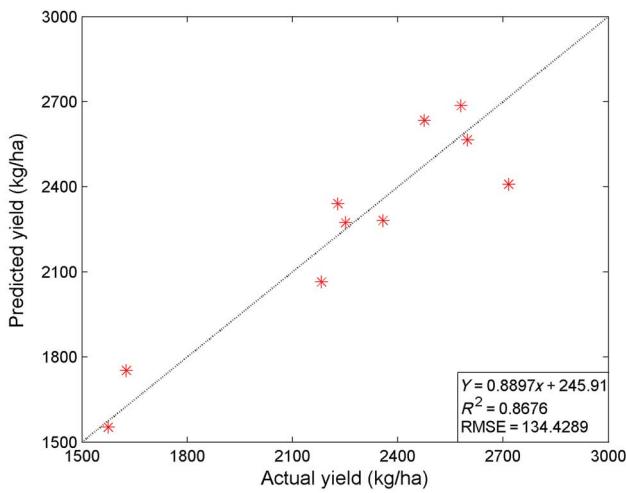


Fig. 8. Scatter plot of the estimated winter wheat yield and actually winter wheat in Luliang.

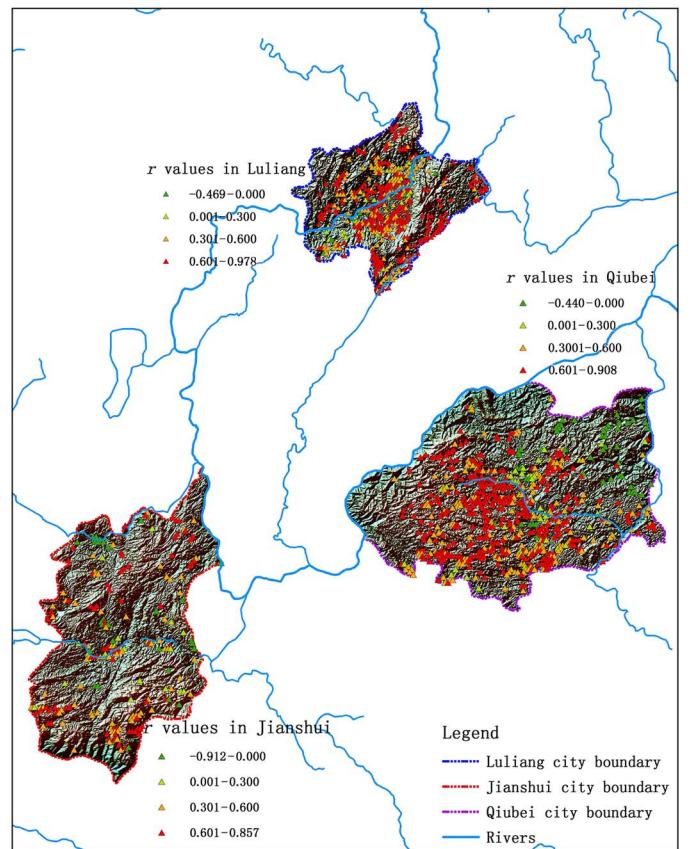


Fig. 9. Correlation coefficients for winter wheat with the hill shade in Jianshui, Luliang and Qiubei.

## V. CONCLUSION

A simple and easily implemented scheme to estimate crop yield in mixed planting regions using time-series MODIS-NDVI data based on geospatial and regression analysis is presented in this paper. With the assumption that the strong correlation between NDVI data and crop yield indicates a high probability of the crop land, the identification method of the crop planting areas is presented through the correlation analysis. Based on the crop classification, the extracted NDVI data from the crop planting areas are conducted for establishing crop estimation models. The approach is applied in three counties (Jianshui, Luliang, and Qiubei) of Yunnan Province in China for three main crops (paddy rice, winter wheat, and corn). The selected models accounted for 42%–48% in Jianshui, 56%–81% in Luliang, and 54%–69% in Qiubei of the variability of paddy rice yield, winter wheat yield, and corn yield, respectively. The RMSE varies from 110.77 to 355.26 kg/ha. Together with the identification of crop planting area, the best periods for yield estimation models are presented, which approximately coincide with the heading, flowering, and filling stages of the crops. In conclusion, the method in this study to estimate the crop yield with the mean NDVI data extracted from the corresponding areas is practical and reasonable. It is also a simple method to study the crop yield in a small region where several crops are planted together.

In contrast with previous studies that dealt with the crops mono-type and known by default, this study conducts crop identification firstly in order to improve the correlation between NDVI data and crop yield. In addition, the S and growth functions are selected based on AIC, which show a significant improvement compared with the linear models. However, only 10 years of data are available and used in this study (because most counties in China have limited data). If a long record of data is available in the future, the models could be validated and updated, and it may be possible to use a moving window to detect the crop pattern change. Although the models proposed in this study show some promising results in crop yield estimation, a more generalized model structure with additional inputs from various physical characteristics may be needed for different crops in different locations. If a wide range of applications of this proposed method could be carried out in various locations, the factors affecting the correlations between crop yield and NDVI, such as air temperature, precipitation, soil moisture, and solar radiation, could be analyzed as model input variables. In brief, with such a generalized model, the crop yield could be estimated in the regions without the historical crop yield records.

#### ACKNOWLEDGMENT

J. Huang acknowledges the support provided by China Scholarship Council during a visit to the University of Bristol.

#### REFERENCES

- [1] P. C. Doraiswamy, S. Moulin, P. W. Cook, and A. Stern, "Crop yield assessment from remote sensing," *Photogramm. Eng. Remote Sens.*, vol. 69, pp. 665–674, Jun. 2003.
- [2] K. N. Chaudhari, R. Tripathy, and N. K. Patel, "Spatial wheat yield prediction using crop simulation model, GIS, remote sensing and ground observed data," *J. Agrometeorol.*, vol. 12, pp. 174–180, Dec. 2010.
- [3] V. K. Boken and C. F. Shaykewich, "Improving an operational wheat yield model using phenological phase-based normalized difference vegetation index," *Int. J. Remote Sens.*, vol. 23, pp. 4155–4168, Sep. 2002.
- [4] M. S. Mkhabela, P. Bullock, S. Raj, S. Wang, and Y. Yang, "Crop yield forecasting on the Canadian Prairies using MODIS NDVI data," *Agric. Forest Meteorol.*, vol. 151, pp. 385–393, Mar. 2011.
- [5] T. Lillesand, R. W. Kiefer, and J. Chipman, *Remote Sensing and Image Interpretation*, 2nd ed. Hoboken, NJ, USA: Wiley, 2008.
- [6] K. Soudani *et al.*, "Ground-based network of NDVI measurements for tracking temporal dynamics of canopy structure and vegetation phenology in different biomes," *Remote Sens. Environ.*, vol. 123, pp. 234–245, Aug. 2012.
- [7] F. Maselli and F. Rembold, "Analysis of GAC NDVI data for cropland identification and yield forecasting in mediterranean African countries," *Photogramm. Eng. Remote Sens.*, vol. 67, pp. 593–602, May 2001.
- [8] M. S. Mkhabela, M. S. Mkhabelab, and N. N. Mashininic, "Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR," *Agric. Forest Meteorol.*, vol. 129, pp. 1–9, Mar. 2005.
- [9] R. Balaghi, B. Tython, H. Eerens, and M. Jlibene, "Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco," *Int. J. Appl. Earth Observ.*, vol. 10, pp. 438–452, Dec. 2008.
- [10] R. Benedetti and P. Rossini, "On the use of NDVI profiles as a tool for agricultural statistics: The case study of wheat yield estimate and forest in Emilia Romagna," *Remote Sens. Environ.*, vol. 45, pp. 311–326, Sep. 1993.
- [11] M. S. Rasmussen, "Operational yield forecast using AVHRR NDVI data: Reduction of environmental and inter-annual variability," *Int. J. Remote Sens.*, vol. 18, pp. 1059–1077, Mar. 1997.
- [12] J. E. Lewis, J. Rowland, and A. Nadeau, "Estimating maize production in Kenya using NDVI: Some statistical considerations," *Int. J. Remote Sens.*, vol. 19, pp. 2609–2617, Sep. 1998.
- [13] C. J. Tucker *et al.*, "An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data," *Int. J. Remote Sens.*, vol. 26, pp. 4485–4498, Oct. 2005.
- [14] A. A. Gitelson and Y. J. Kaufman, "MODIS NDVI optimization to fit the AVHRR data series—Spectral considerations," *Remote Sens. Environ.*, vol. 66, pp. 343–350, Dec. 1998.
- [15] J. Q. Ren, Z. X. Chen, Q. B. Zhou, and H. J. Tang, "Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China," *Int. J. Appl. Earth Observ.*, vol. 10, pp. 403–413, Dec. 2008.
- [16] I. Becker-Reshef, E. Vermote, M. Lindeman, and C. Justice, "A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data," *Remote Sens. Environ.*, vol. 114, pp. 1312–1323, Jun. 2010.
- [17] B. Mulianga, A. Begue, M. Simoes, and P. Todoroff, "Forecasting regional sugarcane yield based on time integral and spatial aggregation of MODIS NDVI," *Remote Sens.*, vol. 5, pp. 2184–2199, May 2013.
- [18] Y. Ran and X. Li. (2012, Jun.). *Multi-Source Integrated Chinese Land Cover Map* [Online]. Available: <http://dx.doi.org/10.3972/westdc.010.2013.db>
- [19] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.
- [20] N. Sugiura, "Further analysis of the data by akaike's information criterion and the finite corrections," *Commun. Statist. Theory Methods*, vol. 7, pp. 13–26, Jan. 1978.
- [21] C. M. Hurvich and C. L. Tsai, "Regression and time-series model selection in small samples," *Biometrika*, vol. 76, pp. 297–307, Jun. 1989.
- [22] Y. Ran, X. Li, L. Lu, and Z. Li, "Large-scale land cover mapping with the integration of multi-source information based on the Dempster–Shafer theory," *Int. J. Geogr. Inf. Sci.*, vol. 26, pp. 169–191, Jan. 2012.
- [23] K. Yamaoka, T. Nakagawa, and T. Uno, "Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations," *J. Pharmacokinet. Biopharm.*, vol. 6, pp. 165–175, Apr. 1978.
- [24] H. Bozdogan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, pp. 345–370, Sep. 1987.
- [25] W. Pan, "Akaike's information criterion in generalized estimating equations," *Biometrics*, vol. 57, pp. 120–125, Mar. 2001.
- [26] L. Wang, Y. L. Bai, Y. L. Lu, H. Wang, and L. P. Yang, "NDVI analysis and yield estimation in winter wheat based on green-seeker," *Acta Agronom. Sin.*, vol. 38, pp. 747–753, Jul. 2012.
- [27] B. D. Wardlow, S. L. Egbert, and J. H. Kastens, "Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. central great plains," *Remote Sens. Environ.*, vol. 108, pp. 290–310, Jun. 2007.
- [28] B. D. Wardlow and S. L. Egbert, "Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the US Central Great Plains," *Remote Sens. Environ.*, vol. 112, pp. 1096–1116, Mar. 2008.
- [29] S. Zhang, Y. Lei, L. Wang, H. Li, and H. Zhao, "Crop classification using MODIS NDVI data denoised by wavelet: A case study in Hebei Plain, China," *Chin. Geogr. Sci.*, vol. 21, pp. 322–333, Jun. 2011.
- [30] M. P. Labus, G. A. Nielsen, R. L. Lawrence, R. Engel, and D. S. Long, "Wheat yield estimates using multi-temporal NDVI satellite imagery," *Int. J. Remote Sens.*, vol. 23, pp. 4169–4180, Oct. 2002.
- [31] J. Wang, P. M. Rich, K. P. Price, and W. D. Kettle, "Relations between NDVI, grassland production, and crop yield in the central great plains," *Geocarto Int.*, vol. 20, pp. 5–11, Sep. 2005.
- [32] L. Salazar, F. Kogan, and L. Roystman, "Use of remote sensing data for estimation of winter wheat yield in the United States," *Int. J. Remote Sens.*, vol. 28, pp. 3795–3811, Aug. 2007.
- [33] D. Jiang, N. B. Wang, X. H. Yang, and J. H. Wang, "Study on the interaction between NDVI profile and the growing status of crops," *Chin. Geogr. Sci.*, vol. 13, pp. 62–65, Mar. 2003.
- [34] R. K. Teal *et al.*, "In-season prediction of corn grain yield potential using normalized difference vegetation index," *Agron. J.*, vol. 98, pp. 1488–1494, Nov. 2006.
- [35] B. L. Ma, L. M. Dwyer, C. Costa, E. R. Cober, and M. J. Morrison, "Early prediction of soybean yield from canopy reflectance measurements," *Agron. J.*, vol. 93, pp. 1227–1234, Nov. 2001.
- [36] W. L. Bai and F. P. Zhang, "Estimation of winter wheat yield in Guanzhong area of shanxi province using SPOT VGT/ NDVI," *Resour. Dev. Market*, vol. 28, pp. 483–485, Jun. 2012.

- [37] S. Hu and X. G. Mo, "Interpreting spatial heterogeneity of crop yield with a process model and remote sensing," *Ecol. Model.*, vol. 222, pp. 2530–2541, Jul. 2011.
- [38] J. Q. Ren, Z. X. Chen, X. M. Yang, X. R. Liu, and Q. B. Zhou, "Regional yield prediction of winter wheat based on retrieval of leaf area index by remote sensing technology," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS'09)*, Cape Town, South Africa, 2009, pp. 374–377.
- [39] N. Murata and H. Park, "Model selection and information criterion," in *Information Theory and Statistical Learning*, F. Emmert-Streib and M. Dehmer, Eds. New York, NY, USA: Springer, 2009, pp. 333–354.
- [40] S. I. Vrieze, "Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)," *Psychol. Methods*, vol. 17, pp. 228–243, Jun. 2012.



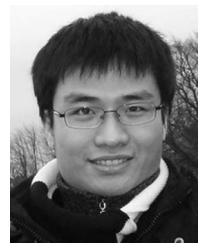
**Jing Huang** was born in Wuxi, Jiangsu, China, in 1986. She received the B.S. degree in information management and information system from the Hohai University, Nanjing, China, in 2009, where she is currently pursuing the Ph.D. degree in management science and engineering.

Her research interests include application of remote sensing and risk assessment and management of flood and drought disaster.



**Huimin Wang** was born in Shanxi, China, in 1963. She received the B.S. degree in mathematics from Shanxi University, Shanxi, China, in 1984, the M.S. degree in mathematics from Chinese Academy of Science, Beijing, China, in 1989, and the Ph.D. degree in management science and engineering from China University of Mining and Technology, Xuzhou, China, in 1997.

She was a Lecturer with the Department of Mathematics, Shanxi University during 1984–1986 and 1989–1993. Since 1999, she has been a Professor with the Department of Management Science and Information Management, School of Business, Hohai University, Nanjing, China. She has undertaken various projects on risk management of extreme flood and drought, water resource management system, water disaster emergency management, etc. She is the author of eight books and more than 200 articles. Her research interests include management science and system engineering, supply chain and optimal control, water resources system operations and management.



**Qiang Dai** received the B.S. degree in geography from Nanjing Normal University, Nanjing, China, in 2009 and the M.S. degree in geography from Sun Yat-sen University, Guangzhou, China, in 2011. He is currently pursuing the Ph.D. degree in civil engineering at the University of Bristol, Bristol, U.K.

From 2011 to 2014, he was a Teaching Assistant with the Faculty of Engineering, University of Bristol, Bristol, U.K. His research interests include rainfall forecast based on weather radar data and numerical weather prediction, in particular, to the uncertainty in the measurement of rainfall using weather radar.



**Dawei Han** was born in Tianjin City, China, in 1961. He received the B.Eng. and M.Sc. degrees in water conservancy from North China University of Water Conservancy and Electric Power, Zhengzhou, China, and the Ph.D. degree in radar hydrology from the University of Salford, Salford, U.K., in 1982, 1984, 1991, respectively.

He is currently a Professor of Hydroinformatics with the Department of Civil Engineering, University of Bristol, Bristol, U.K. He has carried out various projects on weather radar rainfall and numerical weather prediction to aid flood risk assessment, downscaling of global circulation model for climate change, etc. He has published over 150 peer-reviewed journal and conference papers. His research interests include hydroinformatics, real-time flood forecasting, flood risk management, remote sensing and geographic information system, natural hazards, and water resources management.