https://doi.org/10.5194/essd-2022-417
Preprint. Discussion started: 16 January 2023

Open Access Earth System Science Data Discussions

ChinaWheatYield30m: A 30-m annual winter wheat yield dataset from 2016 to 2021 in China

Yu Zhao[1,2#], Shaoyu Han[1,3#], Jie Zheng[1], Hanyu Xue[1], Zhenhai Li[1,4], Yang Meng[1,2], Xuguang Li[5], Xiaodong Yang[1], Zhenhong Li[6], Shuhong Cai[5], Guijun Yang[1,6*]

[1] Key Laboratory of Quantitative Remote Sensing in Agriculture of Ministry of Agriculture and Rural Affairs, Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
[2] National Engineering and Technology Center for Information Agriculture, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China
[3] College of Agronomy, Henan Agricultural University, Zhengzhou 450046, China
[4] College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China
[5] Cultivated Land Monitoring and Protection Center of Hebei, Shijiazhuang, 050056, China
[6] School of Geological Engineering and Geomatics, Chang'an University, Xi'an 710054, China

[#] these authors contributed equally as first authors

Correspondence to: Guijun. Yang (guijun.yang@163.com)

Abstract. Generating spatial crop yield information is of great significance for academic research and guiding agricultural policy. Most existing public yield datasets have a coarse spatial resolution. Although these datasets are useful for analyzing regional temporal and spatial change, they cannot deal with spatial heterogeneity, which happens to be the most significant characteristic of the Chinese small-scale farmers' economy. Hence, we generated a 30-m Chinese winter wheat yield dataset (ChinaWheatYield30m) for major winter wheat-producing provinces in China for the period 2016-2021 with a semi-mechanistic model (hierarchical linear model, HLM). The yield prediction model was built by considering the wheat growth status and climatic factors. It can estimate wheat yield with excellent accuracy and low cost using a combination of satellite observations and regional meteorological information (i.e., Landsat 8, Sentinel-2 and ERA5 data from the Google Earth Engine (GEE) platform). The results were validated by using in situ measurements and census statistics and indicated a stable performance of the HLM model based on calibration datasets across China, with r of 0.81** and nRMSE of 12.59%. With regards to validation, the ChinaWheatYield30m dataset was highly consistent with in situ measurement data and census data, indicated by r (nRMSE) of 0.72** (15.34%) and 0.73** (19.41%). The ChinaWheatYield30m is a sophisticated dataset with both high spatial resolution and convictive accuracy, such a dataset will provide basic knowledge of exquisite wheat yield distribution, which can be applied for many purposes including crop production modelling or regional climate evaluation.

1 Introduction

Wheat is the most widely planted crop, supplying a fifth of global food calories and protein (Erenstein et al., 2022). However, wheat production is facing unprecedented challenges in the global context of climate change, such as frequent extreme weather events. Apart from natural factors, socioeconomic events such as the COVID-19 pandemic, regional conflicts, and other global crises can also significantly perturb wheat production (IFPRI, 2022). In China, where needs to feed one-fifth of the world's population on its limited land (FAO, 2020) and food security is crucial, wheat production is an essential agricultural activity. Ensuring stable grain supplies and increasing production are important to the national economy and people's livelihoods (Feng et al., 2020). Therefore, monitoring of crop yields timely is of great significance for regulating import and export decision-making, grain market prices, crop insurance evaluations, smart agriculture applications, and rational allocations of agricultural resources.

In the past decades, remote sensing data from ground-based, aerial-based and satellite-based platforms have received extensive attention for crop yield prediction (Battude et al., 2016; Jiang et al., 2019; Li et al., 2020; Wang et al., 2021). Ground- and aerial-based platforms have high spatial resolution and control, which are advantageous for farm-scale applications. However, their application to large-area yield estimations is too expensive. Satellite-based approaches have been widely used to monitor crop production over large areas in the past few decades, benefitting from capable of acquiring temporally and spatially continuous information (Battude et al., 2016; Huang et al., 2019). With the rapid launch of new satellites carrying various types of sensors, regional yield mapping is becoming more accurate and at higher spatial resolution. The mapping relies on vegetation indices (VIs) that can be derived from visible and near-infrared (NIR) reflectance bands in multispectral optical data, such as the Normalized Difference Vegetation Index (NDVI) (Rouse et al., 1974), the enhanced vegetation index (EVI) (Sims et al., 2008), or the optimized soil adjust vegetation index (OSAVI) (Rondeaux et al., 1996). These VIs have often been used to predict crop yield (Magney et al., 2016; Cao et al., 2021; Zhao et al., 2022). There are many methods to incorporate VIs in yield estimation, such as parametric regressions, deep learning, and data assimilation (Battude et al., 2016; Huang et al., 2019; Li et al., 2020).

Parametric regression models directly establish the relationship between VIs and crop yield, which may be linear or nonlinear (Magney et al., 2016; Li et al., 2020). These parametric regressions are limited to the specific research area and growing season for which they are developed, making it hard to extrapolate them either in the spatial or temporal domains. Non-parametric statistical approaches have been used in recent yield projections research. Notable studies have been done using machine learning (ML) (Cai et al., 2019; Li et al., 2021). An emerging new technique for crop yield estimations is deep learning (Tian et al., 2021) applied to various types of data acquired by satellites and drones (Jiang et al., 2020; Wang et al., 2020). Overall, ML methods need large multidimensional datasets, which can challenge their application (Cao et al., 2021).

Unlike the above-mentioned statistical models, process-based mechanic models simulate crop yield from various inputs, including soil properties, meteorological data as well as crop characters. Examples of such models are the Decision Support System of Agrotechnology Transfer modeling system (DSSAT), the Agricultural Production Systems sIMulator (APSIM) and the Simple Algorithm For Yield (SAFY) and many other crop models (Jones et al., 2003; Keating et al., 2003; Duchemin et al., 2008). These mechanistic models can generate reliable yield estimates (Paudel et al., 2021). Data assimilation provides a way of integrating the monitoring properties of observed data into the predictive and explanatory abilities of crop growth models. Leaf area index (LAI) or biomass are often used as state variables of the DA system to correct a crop growth model behavior and ensure accurate yield predictions (Battude et al., 2016; Kang and Ozdogan, 2019). Yield is a complex trait that is related to numerous factors, including natural drivers (Li et al., 2021), crop variety (Wei et al., 2022; Bailey-Serres et al., 2019), and human factors, majorly consisting of fertilization and irrigation (Jones et al., 2003; Keating et al., 2003; Duchemin et al., 2008). Existing studies demonstrated that only updating one or two state variables is not sufficient to correct a crop growth model and thus cannot improve output predictions (Ines et al., 2013; Huang et al., 2015; Hu et al., 2017; Huang et al., 2019). In addition, uncertainties in the remote sensing monitoring of state variables such as LAI and biomass are also inherited by the DA system (Kang et al., 2019). Although data assimilation techniques allow a formal and well-understood way to combine model predictions with observations, their computational intensity is a problem that tends to be ignored when estimating large-area crop production. Transfer learning techniques can be used to transfer the knowledge learned from a crop growth model to predict wheat yield to effectively improve calculation efficiency (Zhao et al., 2022). A reliable labeled dataset is a prerequisite for the transfer learning method (Zhang et al., 2021). However, building an effective dataset for migration learning over a large region is still challenging.

In addition to traditional crop models and assimilation strategies, there are hybrid models that incorporate the simplicity of a statistical model and the rationality of a mechanistic model and are thus called semi-mechanistic models (Ji et al., 2022). For example, Dong et al. (2020) developed the EC-LUE-GPP model and successfully estimated the wheat yield in Kansas, USA. Li et al. (2020) used the HLM model to estimate interannual yield and showed good performance. Generally, a semi-mechanistic model has great potential in yield estimation, but its application is often limited to a relatively small area, e.g., farm scale, county to city scale, rather than a larger scale. National crop yield datasets, which are of great significance for large-scale agricultural resource allocation, agricultural system model construction, and climate change impact assessment, are produced at coarse spatial resolutions (Table 1), e.g., 0.5°, 10-km, 4-km or 1-km resolution (Monfreda et al., 2008; You et al., 2014; Iizumi and Sakai., 2020; Grogan et al., 2022; Luo et al., 2022; Cheng et al., 2022) and are mostly downscaled based on the statistical yield datasets and other datasets (Monfreda et al., 2008; You et al., 2014; Iizumi and Sakai., 2020; Grogan et al., 2022). This method of yield downscaling may lead to inaccurate yield estimates and incorrect assessments of the impact of climate change. In addition, yield predictions cannot rely on statistical data alone. Luo et al. (2022) and Cheng et al. (2022) developed yield datasets combining coarse-resolution real-time remote sensing data with agricultural statistics, but because 1 km × 1 km plots or 4 km × 4 km farmlands are rare in China, their field application is limited. Although these datasets are useful for analyzing regional temporal and spatial change, they cannot deal with spatial heterogeneity, which happens to be the most significant characteristic of the Chinese small-scale farmers' economy. Therefore, there is an urgent need to construct a high-resolution yield dataset for investigating spatiotemporal patterns of crop production, assessing climate change impacts, and modeling crop growth processes over large spatial extents.

Table 1 Summary of studies on crop yield datasets

References	Species	Method	Resolution	Span	Spatial coverage
Monfreda et al., 2008	175 crops	Dataset summary	10 km	2000	Global
You et al., 2014	20 crops	Global spatial production allocation model	10 km	2000, 2005, 2010	Global
Iizumi & Sakai, 2020	4 crops	Maize, Rice, Wheat and Soybean	43 km	1981-2016	Global
Grogan et al., 2022	26 crops	Gata statistics based on Global Agro-Ecological Zones Version 4 model	10 km	2015	Global
Luo et al., 2022	Wheat	LSTM	4km	1982 - 2020	Global
Cheng et al., 2022	Maize, Wheat	Random Forest	1 km	2001 - 2015	China

In this study, by integrating remote sensing and climate data, we aim to 1) propose a semi-mechanistic model with excellent accuracy and low cost by combining remote sensing observations and regional meteorological information, which can simultaneously overcome inter-annual and cross-regional problems; 2) evaluate model performance by using both validation dataset and the census yield data; 3) generate a high-resolution Chinese winter wheat yield dataset (ChinaWheatYield30m) for the period 2016-2021. This dataset will be useful to further yield-related research and guide related food policies.

2 Data and methods

2.1 Study areas

Our study area consists of the main winter wheat-growing region of China, which includes 12 provinces and municipalities (Figure. 1). Most of the region is in the middle of China and includes temperate-continental monsoon, temperate monsoon, and subtropical monsoon climates. The sown area and production of winter wheat in China accounted for 20.02% and 21.77% of staple food crops in 2021 (National Bureau of Statistics of China, 2021), respectively. Three sample areas were selected for detailed analysis based on their different geographical and climatic conditions.

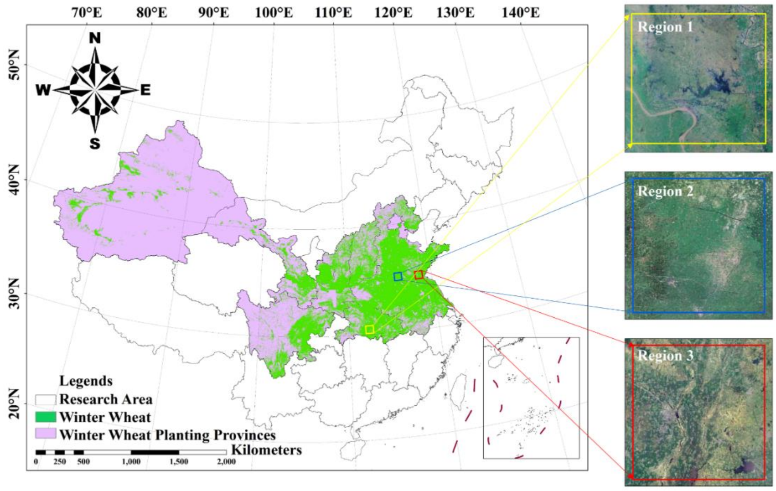

Figure 1. Distribution of winter wheat within the study area and three selected example areas. The wheat planting data is adapted from Dong et al., (2020).

2.2 Data Collection

2.1 The winter wheat land cover data

We used a winter wheat map with a 30-m resolution across the main growing areas of China (Dong et al., 2020). These data produce winter wheat maps from 2016 to 2020, which is the base map of ChinaWheatYield30m production. The yield distribution map of 2021 uses the winter wheat classification map of 2020, and the rest of the yield distribution maps are winter wheat classification maps of that year.

2.2.2 Satellite Imagery Data Acquisition

In this work, we extracted the enhanced vegetation index 2 (EVI2) (Jiang et al., 2008) on the Google Earth Engine (GEE) platform from Landsat 8 and Sentinel-2 images during the 2016-2021 period. These datasets were chosen to increase observation frequency and were used for subsequent phenological extraction and yield estimation. Xu et al. (2020) have shown that Landsat 8 data and Sentinel 2 data have high consistency. The EVI2 is calculated from the reflectance in Red and NIR bands (Eq. (1)):

$$EVI2 = 2.5 * \frac{NIR - Red}{NIR + 2.4 * Red + 1} \tag{1}$$

where NIR and Red represent the Near-Infrared and Red reflectances, respectively, in Landsat 8 or Sentinel-2. The maximum EVI2 (EVI2max) of the winter wheat growing season was used in this paper. It is generally believed that the time of EVI2max corresponds to the heading period, which has been shown to be the best period for remote sensing yield estimation (Luo et al., 2020).

2.2.3 Meteorological data

Meteorological data were important input variable for yield prediction, mainly from March to May, because this period includes most key growth stages of winter wheat (i.e. stem elongation, booting, heading, flowering and filling stages). The meteorological data, including monthly average temperatures (Tem), monthly solar radiation (Rad), and monthly precipitation (Pre), were obtained from the ERA5 dataset provided by the GEE platform (https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_MONTHLY) with a resolution of 0.1° for the sampling site. All three types of meteorological datasets were resampled to a 30-m resolution to ensure data uniformity.

2.2.4 In situ measurement yield data

Georeferenced field-scale yields were obtained by field investigation from 2016 to 2021. During the harvest period, a five-point (1 m2 per point) sampling method was used to destructively sample each winter wheat plot to measure yield. To avoid edge effects, each sample point was at least 2 m away from the edge of the farmland. The harvested grain was threshed and air-dried for yield determination. Then, the final yield was standardized as grain with 14% moisture content. In this paper, the data were randomly split into two dataset, two-thirds of the data were used for modelling, and the remaining data were used for validation.

2.2.5 The province-level statistical data

The province-level yield data were collected from state statistical bureau of the study area from 2016 to 2021 (http://www.stats.gov.cn/tjsj/ndsj/). Most statistical data did not directly record the unit yield data, so the statistical yield data (kg·ha^{-1}) were obtained by converting the total production by the planted area. These data were used for the model validation of the selected research provinces.

Table 2 Details on the datasets used in this study

Data type	Content	Resolution	Span	Data usage	Data sources
Winter wheat land cover data	Classification of winter wheat	30m	2016-2020	Research area	Dong et al., 2020
Satellite data	EVI2max	30m	Winter wheat growing season of each year from 2016 to 2021	Input variables	Landsat8 and Sentinel2 dataset of GEE platform
Meteorological data	Tem Rad Pre	0.1°	March to May of each year from 2016 to 2021	Input variables	ERA5 dataset of GEE platform
In-situ measured yield data	Field-level yield with coordinates	Field-level	2016-2021	Model establishment and evaluation	Field investigation
Census yield data	Province-level yield statistical data	Province-level	2016-2021	Model validation	State statistical bureau
Yield dataset	GlobalWheatYield4km	4km	2016-2020	Dataset comparison	Luo et al., 2022

2.3 Method

2.3.1 Methodology

The hierarchical linear model (HLM) is a simple and efficient method for dealing with nested structures. At present, HLM has been extensively applied to predicting yield, grain protein content, and agronomic traits for inter-annual and transregional (Li et al., 2020; Xu et al., 2020; Li et al., 2022; Zhao et al., 2022). These papers have demonstrated that the HLM method is a stable, reliable and scalable way of solving yield estimation problems. They also demonstrated that, although a linear relationship between EVI2max and crop yield can be established in a particular field of a single year, differences in meteorological factors between regions and years will differentiate this relationship, which is the exact problem that the HLM model was implied to settle. For each province, a set of parameters was generated by using the data collected from the sample fields. The specific yield-predicting models in different provinces using the HLM method in this study involved a two-levels hierarchy. Level 1 of the HLM model was constructed based on the yield and EVI2max:

$$Level\ 1: Yield\ = \beta_{0j} + \beta_{1j} * EVI2_{max}\ + r_{ij} \qquad (2)$$

where β0j and rij represent the intercept and random error, respectively, and β1j represents the slope of the linear model corresponding to EVI2max.

In the HLM, the parameters of β0j and β1 at Level 1 become dependent variables at Level 2. The independent variables of Level 2 are the accumulated meteorological data (Tem, Rad, and Pre) of different growth stages, such that:

$$Level\ 2: \beta_{mj} = \gamma_{m0} + \gamma_{m1} * Tem_{Sm} + \gamma_{m2} * Rad_{Sm} + \gamma_{m3} * Pre_{Sm} + \mu_{mj} \qquad (3)$$

where βmj represents the β0 and β1 from Level 1 of HLM, and γm0 is the intercept. γm1 - γm3 represent slopes of each accumulated meteorological data of different months (m=3, 4, and 5) and μmj is the random error of Level 2 of HLM. Figure 2 shows a schematic of the workflow.

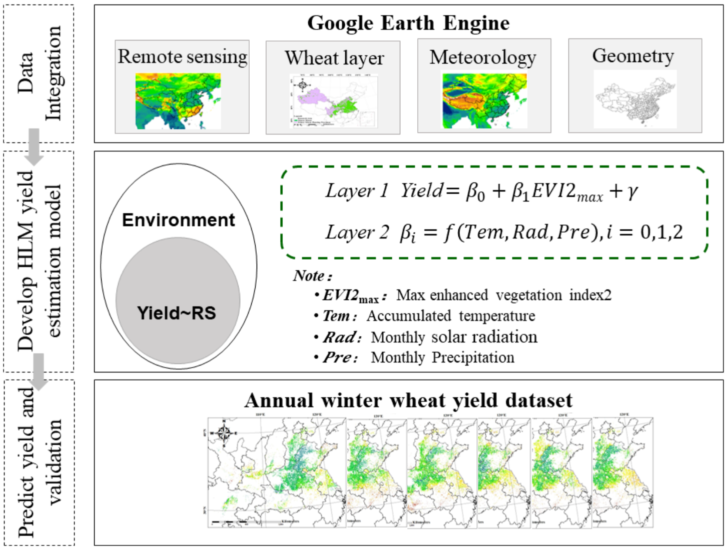
Data Integration
Google Earth Engine
Remote sensing
Wheat layer
Meteorology
Geometry
Develop HLM yield estimation model
Environment
Yield~RS
Layer 1 $Yield = \beta_0 + \beta_1 EVI2_{max} + \gamma$
Layer 2 $\beta_i = f(Tem, Rad, Pre), i = 0,1,2$
Note:
• $EVI2_{max}$: Max enhanced vegetation index2
• Tem: Accumulated temperature
• Rad: Monthly solar radiation
• Pre: Monthly Precipitation
Predict yield and validation
Annual winter wheat yield dataset

Figure 2 Schematic diagram outlining the inputs, major processing steps used, and generated outputs.

2.3.2 Comparison with the Random Forest method and the other yield datasets

Random Forest (RF) is a model with predictive performance commonly used in the current yield estimation literature (Li et al., 2020; Cheng et al., 2022; Luo et al., 2022). RF regression is a classic ensemble machine learning model that establishes multiple unrelated decision trees by randomly extracting samples and features and obtains the prediction results in parallel. Each decision tree can obtain a prediction result through the samples and features extracted, and the regression prediction result of the whole forest can be obtained by averaging the results of all trees (Breiman, 2001).

Given the wide range of RF applications in generating crop yield data, we built a RF prediction model in Matlab and compared its performance with the HLM model. The number of decision trees was set to 200, and the maximum depth of the tree and the number of features were selected for tuning.

We compared our yield production (ChinaWheatYield30m) with an existing 4-km dataset of global wheat yield (GlobalWheatYield4km) (Luo et al., 2022) using in situ data to validate the reliability of our dataset. More specifically, we calculated the r and nRMSE between the in situ measurement yields and the estimates of GlobalWheatYield4km or ChinaWheatYield30m from 2016 to 2021.

2.3.3 Model evaluation

The commonly used correlation coefficient (r) and relative root mean square error (nRMSE) were used to compare the performance of generated models. To estimate the contribution of each input variable of the HLM, we applied an extended Fourier amplitude sensitivity test (Saltelli et al., 1999). The EFAST (Extended Fourier amplitude sensitivity test) was used to determine a sensitivity index (SI) which combined the advantages from both Fourier amplitude sensitivity test and Sobol algorithm. The derived SI quantified how output results were impacted by input variables. The SI of each independent input variable to the yield in different provinces was computed with Simlab (version 2.2.1) software.

3 Results and Discussion

3.1 Exploring the appropriate method and accuracy assessment

The performance of RF and HLM models in situ yield predictions during 2016 – 2021 for each province are shown in Fig. 3. The calibration sets for RF and HLM models have similar performance, with r (nRMSE) ranges of 0.79 - 0.92 (5.78% - 23.37%) and 0.67 - 0.87 (4.87% - 22.06%), respectively. However, in the validation set, the HLM model outperformed RF with the r (nRMSE) range of 0.50 - 0.93 (1.93%-23.00%) and 0.27 - 0.76 (13.44% - 30.86%), respectively. The superior performance of HLM was attributed to its ability to capture the interaction effects among various factors. This interaction explained most of the variation among the provinces, with a sensitive index range of 9.85% - 69.92% (Fig. 4). The sensitive index of input variables to the HLM model is shown in Fig. 4, indicating the contributions of each variable to the HLM model. Overall, in most of the analyzed provinces, EVI2 was the most important variable in the HLM model, with a contribution range of 11.70 % - 63.18% for different provinces. As for the meteorological factors, in general, temperature was the most important factor, whereas radiation and precipitation were less significant. The variables related to accumulated temperature, Tem04 and Tem05, had a high contribution (8.50% - 21.90%) to the HLM model. The results show the importance of weather in April and May, which in our research areas are the key months for the flowering and filling of winter wheat, the critical periods in grain formation when most organic matter is accumulated (Cabas et al., 2010).

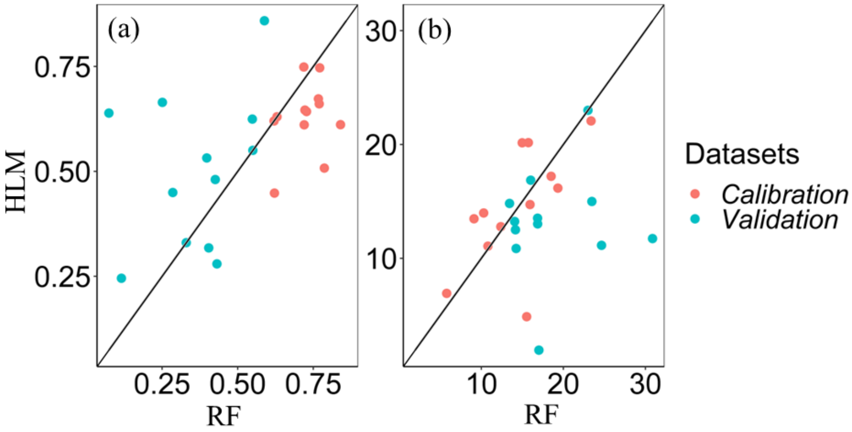

Figure 3. Performance of RF and HLM in yield prediction during 2016 – 2021 across 12 provinces: (a) r and (b) nRMSE.

Therefore, the optimal HLM model was implemented to predict in situ wheat yield using the calibration dataset. The results showed that the predicted yield based on the HLM model is consistent with the measured in situ records (Fig. 5a). We compared the results derived from HLM from 2016 to 2021 with in situ records. The r and nRMSE of the measured and predicted yield of winter wheat were 0.81** and 12.59%, respectively (Fig. 5a). Moreover, the normal distribution of the nRMSE in multiple independent provinces (Fig. 5b) also showed acceptable performance (μ = -0.01, σ = 0.02).

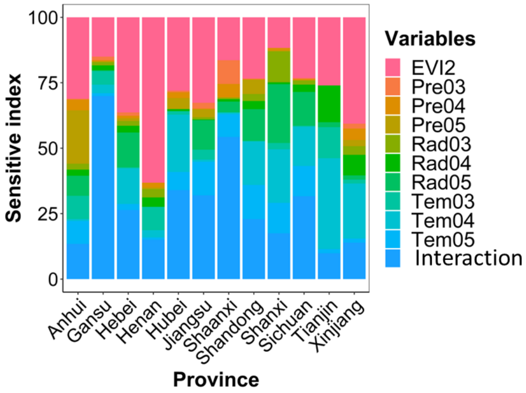

Figure 4 Sensitive index in the trained HLM model for different input variables.

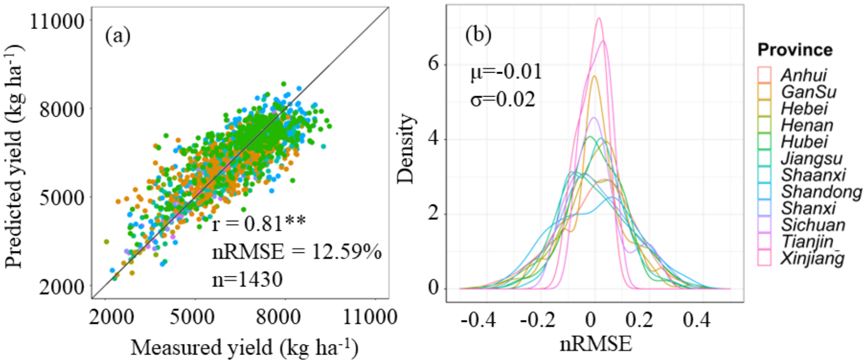

Figure 5. Comparison of measured yield with predicted yield (a) and nRMSE frequency histogram (b) in multiple independent regions.

3.2 Yield data validation using in situ measurements and province-level statistics

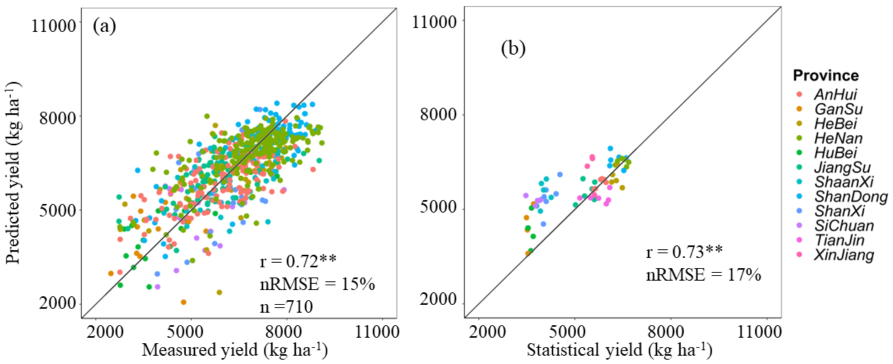

Figure 6. Comparisons between observed and retrieved yield for winter wheat: (a) in situ measurements and (b) province-level statistics.

In situ estimates of wheat yields based on field measurement data were highly consistent with the pixel-level crop yield dataset generated using the HLM model with EVImax and meteorological data (r = 0.72**, nRMSE = 15.34%) (Fig. 6a). In contrast, model performance showed overestimates of wheat crop yield compared with province-level statistical yield (r = 0.73**, nRMSE = 19.41%) (Fig. 6b). Therefore, the field-scale yield prediction dataset has not only high precision at a fine scale, but also performs well on a large scale.

Figure 7 shows the spatial patterns of ChinaWheatYield30m from 2016 to 2021. Generally, the spatial patterns of predicted yields were consistent with in situ measured yields, with large variability from 2273.82 – 10518.82 kg ha^{-1}. We further summarized the province-level statistic yield. The yield averages were highest in Shandong Province (6567.48 kg ha-1), followed by Henan Province (6498.42 kg ha^{-1}) and Hebei Province (6039.39 kg ha^{-1}). By contrast, Jiangsu Province achieved the lowest average yield (4337.05 kg ha^{-1}) (Fig. 6b). Overall, these data are consistent with the census data.

https://doi.org/10.5194/essd-2022-417
Preprint. Discussion started: 16 January 2023

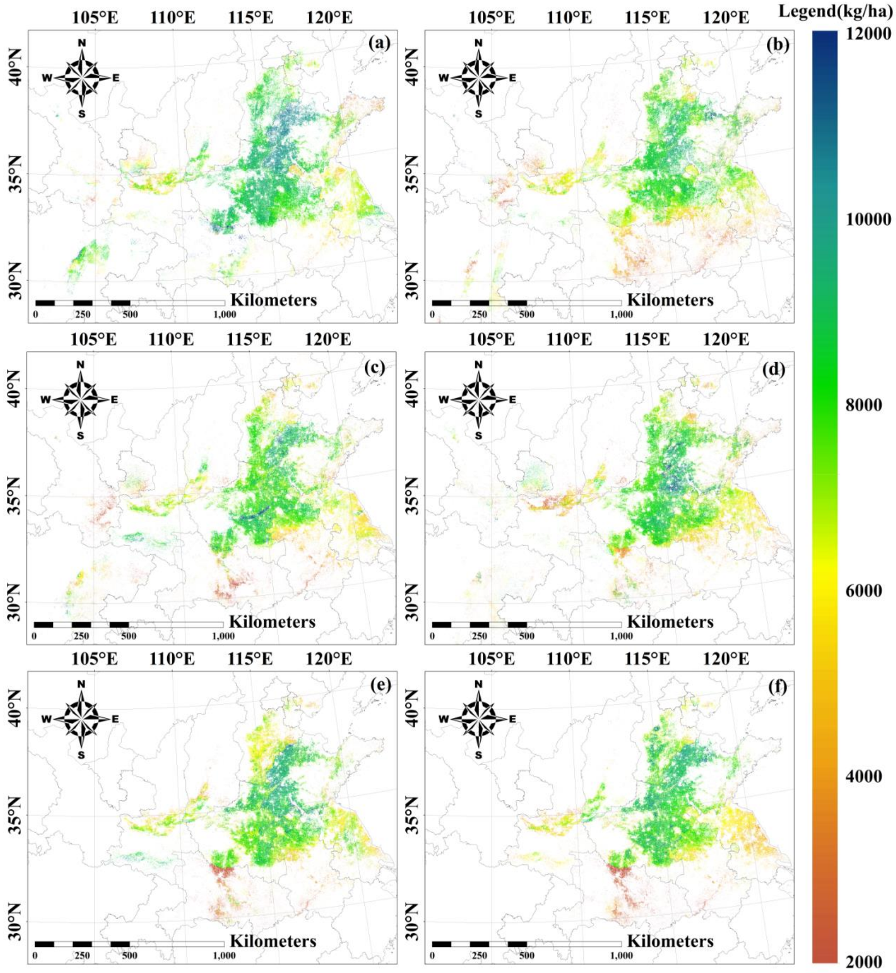

Figure 7. Spatial patterns of annual winter wheat yield during 2016 - 2021.

3.3 Comparing ChinaWheatYield30m with GlobalWheatYield4km

We compared the datasets at the field level using single pixels and through a zonal analysis of three selected research areas. Field-level yield estimates were aggregated to match the ChinaWheatYield30m and GlobalWheatYield4km from 2016 to 2020 and then compared with in situ measurement yields. The yield estimates of ChinaWheatYield30m showed higher consistencies with in situ measurement yields as the scatter points were closer to the 1:1 line than in the case of GlobalWheatYield4km. The results showed that, in different years, ChinaWheatYield30m has a lower nRMSE range (12.40% – 13.84%) compared to GlobalWheatYield4km (20.43% – 33.06%) (Fig. 8).

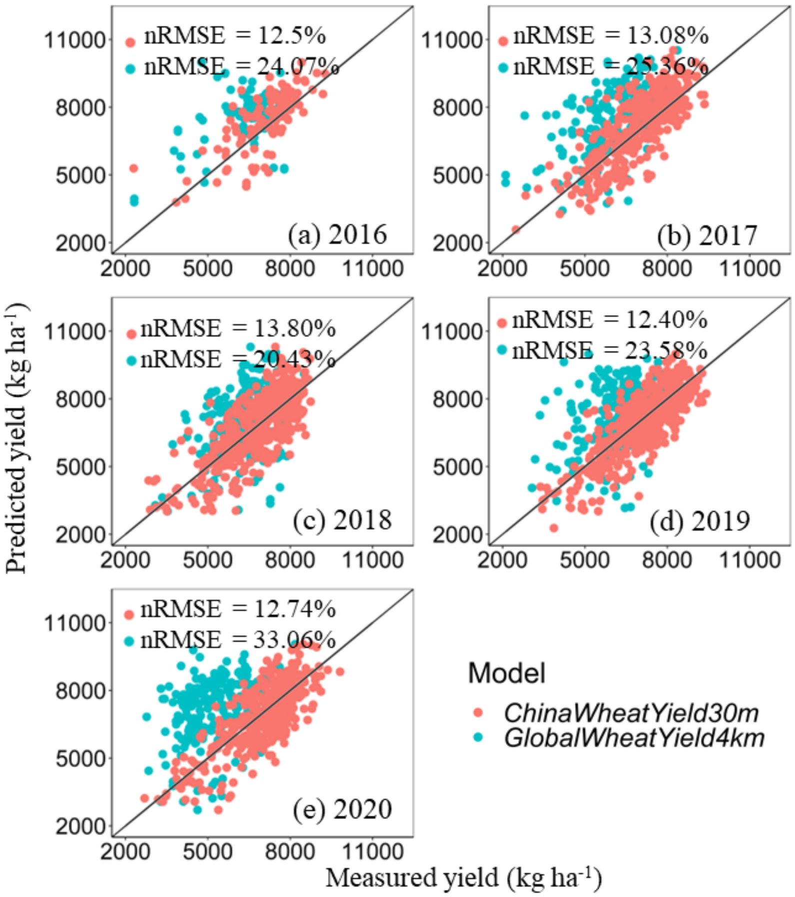

Figure 8 Comparisons between in situ measurement yields and predicted yields of GlobalWheatYield4km or ChinaWheatYield30m for 2016 (a), 2017 (b), 2018 (c), 2019 (d), and 2020 (e).

As for the zonal analysis, winter wheat yield derived from ChinaWheatYield30m also have a close spatial pattern to GlobalWheatYield4km production (Fig. 9 and Table 3). Besides, ChinaWheatYield30m, with a standard deviation of 290.27 – 880.91 kg ha^{-1}, depicts the difference in yield with greater spatial detail compared to the GlobalWheatYield4km standard deviation of 195.46 – 1516.09 kg ha^{-1}. In the selected sample areas, the yield ranges of ChinaWheatYield30m and GlobalWheatYield4km are 2115.95 kg ha^{-1} – 7668.69 kg ha^{-1} and 2653.62 kg ha^{-1} – 10504.50 kg ha^{-1}, respectively. This wide range and minor deviation reveal the advantages of fine-resolution data. Compared with the actual yield records, GlobalWheatYield4km significantly underestimates them, whereas ChinaWheatYield30m is closer to the 1:1 line. In the selected sample areas, the mean yield of ChinaWheatYield30m is generally higher than that of GlobalWheatYield4km because the wheat classification at 30-m resolution is dominated by pure wheat pixels. In contrast, the wheat classification with 4-km resolution has more mixed pixels. For example, buildings and roads cannot be identified in the 4-km classification but result in an underestimation of yield prediction (Fig. 9).

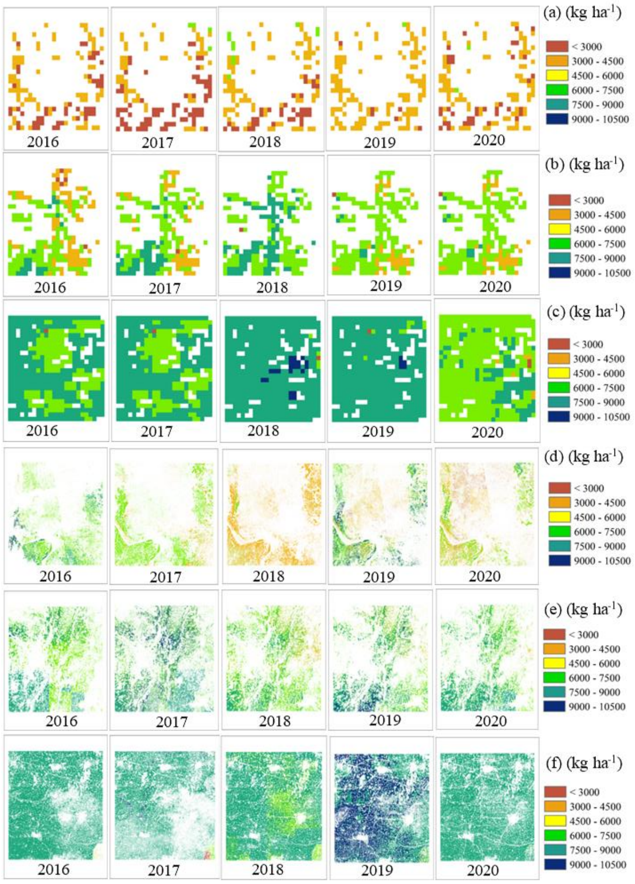

Figure 9 Comparison of spatial patterns between GlobalWheatYield4km (a, b, c) and ChinaWheatYield30m (d, e, f) from 2016 to 2020. The detailed location of the selected example areas (Region 1 and d; Region 2 and e; Region 3) is shown in Figure 1.

Table 3 Statistical analysis of GlobalWheatYield4km and ChinaWheatYield30m

Reg.	Year	GlobalWheatYield4km (kg ha^{-1})				ChinaWheatYield30m (kg ha^{-1})			
		Min	Max	Mean	Std	Min	Max	Mean	Std
1	2016	2215.59	4499.56	3085.45	394.45	3787.87	10504.50	5797.42	711.83
	2017	2115.95	5543.09	3034.08	660.67	3555.29	7470.59	4849.84	374.35
	2018	2461.41	5192.57	3499.66	632.27	3015.54	6231.35	3746.34	422.38
	2019	2802.90	4987.77	3511.21	346.15	2653.62	9978.73	5351.96	1516.09
	2020	2336.31	4584.65	3347.63	505.31	2705.24	7874.20	4238.19	977.18
2	2016	2751.27	6626.20	4807.49	880.91	4257.01	9078.25	6002.05	438.40
	2017	3504.07	7102.54	5349.48	847.29	4997.04	10504.47	6564.42	968.29
	2018	4524.76	6755.62	5880.72	402.58	3818.12	10291.08	6472.96	721.93
	2019	3988.76	6555.61	5551.69	528.77	3198.47	9902.78	6704.211	989.46
	2020	3766.66	6301.66	5069.00	526.35	4352.21	8439.71	6100.75	745.51
3	2016	4388.15	7127.87	6103.27	491.77	3788.11	7554.13	7047.07	321.38
	2017	5000.56	7387.55	6261.93	433.99	5917.13	8266.23	7199.44	214.30
	2018	5637.92	7668.69	6931.35	356.61	4927.40	8384.25	6357.63	378.09
	2019	5589.33	7540.64	6535.69	290.27	5394.00	9980.07	7576.74	652.95
	2020	3861.44	7003.86	5590.34	521.12	5557.38	8186.71	6802.47	195.46

Note: Reg represents Region.

4 Discussion

4.1 Advancements of the 30-m resolution yield dataset

Information on the spatial extent of winter wheat yield is essential for drafting economic and food subsidy policies and rationally allocating resources (FAOSTAT, 2018). To our knowledge, to date there is no fine resolution (30 m) winter wheat yield distribution map. Previous research has generated the winter wheat yield distribution map of some major production areas in China at moderate resolution, e.g., 10-km, 5-arcmin grid, 5-minute grid, 4-km, and 1 km (Monfreda et al., 2008; Fischer et al., 2012; You et al., 2022; Grogan et al., 2022; Luo et al., 2022; Cheng et al., 2022). Moderate-resolution yield maps have a mixed-pixel problem, which may lead to great uncertainties, as mentioned in comparison with the 4-km yield dataset.

Existing wheat yield maps are usually available at the end of the season or based on yield statistics, which limits their application in early field management and government macro-control (Battude et al., 2016; Kang and Ozdogan., 2019). For example, crop growth models strongly depend on daily meteorological data as input; this increases the difficulty in early yield prediction because meteorological data during the season is lacking and long-term meteorological forecasts are unreliable. EVI2max and meteorological data used in this paper can be obtained before May, while wheat in China's main winter wheat production areas is generally harvested in June.

In this study, we generated ChinaWheatYield30m with a 30-m spatial resolution for 2016 – 2021 based on HLM using Landsat- and Sentinel-derived maximum EVI2. ChinaWheatYield30m had the following advantages:

1) The highest resolution of the existing yield data set: 30m, showing more spatial distribution;

2) A stable accuracy at field scale and large regional scale, highly contributing to field management, modeling agricultural systems, drafting agricultural policies;

3) The product has a high real-time performance and can be used to forecast the output in the early period of the year.

Therefore, the proposed method can accurately predict winter wheat yield in real time. The strengths of the HLM model are overcoming inter-annual and regional variations (Li et al., 2020; Xu et al., 2021; Zhao et al., 2022). The results based on field investigation and statistical data show that the method can accurately predict winter wheat yield in the main production areas.

4.2 Uncertainties

Despite the advantages of ChinaWheatYield30m, the dataset also presents some data and model uncertainties.

1) Uncertainties in winter wheat classifications are transferred to the yield predictions. The wheat classification is based on optical remote sensing data and may be affected by meteorological factors such as clouds and rain. In addition, the winter wheat classification data are mainly based on time series, and a similar time series may lead to a wrong classification, which results in uncertainties in regional yield statistics.

2) Optical remote sensing data are another source of uncertainty. In general, maximum EVI2 is obtained at the heading or flowering period (Luo et al., 2020), but due to the irregular availability of usable

Sentinel-2 and Landsat observations, the maximum EVI2 nationwide may correspond to different phenological periods. In future studies, we will attempt to map the yield distribution of wheat using multi-source remote sensing images, including passive remote sensing data.

3) The accessibility of in situ measurement data is also one of the uncertainties in ChinaWheatYield30m. On the one hand, the performance of HLM depends on the quantity and quality of samples. It is more precise when sampling in the quadrat and is often higher than the statistical yield data. It was particularly difficult to collect finer-scale census data with longer time coverage in some areas, such as Xinjiang Province, leading to data gaps in ChinaWheatYield30m. We combined in situ measurements and statistical data to calibrate and validate the ChinaWheatYield30m. However, where sparse observation where available, we could only calibrate the parameters of the mathematical optimization.

5 Data availability

The derived yield dataset for ChinaWheatYield30m during 2016 – 2021 is available at https://doi.org/10.5281/zenodo.7360753 (Zhao et al., 2022). Please be so kind to contact the authors for more detailed information.

6 Conclusions

In the present study, we generated a 30m Chinese winter wheat yield from 2016 to 2021 based on the HLM model, called ChinaWheatYield30m. First, we construct a semi-mechanical model with excellent accuracy and low cost in a combination of RS observations and regional meteorological information for major winter wheat-producing areas in China. The HLM model has stable performance in calibration sets across China, with r of 0.81** and nRMSE of 12.59%, respectively. Next, we validated the predictive performance of in-situ measurement data and statistical data. The ChinaWheatYield30m dataset was highly consistent with in-situ measurement data and statistical data, indicated by r (nRMSE) of 0.72** (15.34%) and 0.73** (19.41%), respectively. Finally, we established a high-resolution yield product for winter wheat in China during 2016 – 2021. Our ChinaWheatYield30m can be applied for many purposes, including further academic research, making economic, food subsidy policies and rationally allocating imperative resources.

Author contributions

YZ, GY and SH designed the research, performed the analysis, and wrote the paper; JZ, HX, YM and XL collected datasets; GY, XY, XX, ZL and SC performed the analysis; GY edited and revised the paper.

Competing interests

The authors declare that they have no known conflict of interest.

Acknowledgements

We would like to thank the editor and the two reviewers for their valuable comments. We would also like to thank all the scientists and students who participated in the field observations.

Financial support

This research has been supported by National Key Research and Development Program of China (2022YFD2001103), Key scientific and technological projects of Heilongjiang province (2021ZXJ05A05), the Platform Construction Funded Program of Beijing Academy of Agriculture and Forestry Sciences (PT2022-24) and Chongqing Technology Innovation and Application Development Special Project (cstc2019jscx-gksbX0092, cstc2021jscx-gksbX0064.

References

Bailey-Serres, J., Parker, J.E., Ainsworth, E.A., Oldroyd, G.E.D., Schroeder, J.I. Genetic strategies for improving crop yields. Nature 2019, 575, 109-118. https://doi.org/10.1038/s41586-019-1679-0.

Battude, M., Al Bitar, A., Morin, D., Cros, J., Huc, M., Marais Sicre, C., Dantec, V., Demarez, V. Estimating maize biomass and yield over large areas using high spatial and temporal resolution sentinel-2 like remote sensing data. Remote Sens. Environ. 2016, 184, 668–681. https://doi.org/10.1016/j.rse.2016.07.030.

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

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric for Meteotol2019, 274, 144-159, https://doi.org/10.1016/j.agrformet.2019.03.010.

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

Cabas, J., Weersink, A., Olale, E. Crop yield response to economic, site and climatic variables. Clim Change, 2010, 101, 599–616. https://doi.org/10.1007/s10584-009-9754-4.

Chen, Y., Zhang, Z., Tao, F. Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data. Eur J Agron 2018, 101, 163-173, https://doi.org/10.1016/j.eja.2018.09.006.

Cheng M., Jiao, X., Shi, L., Penueals, J., Kumar, L., Nie, C., Wu, T., Liu, K., Wu, W., Jin, X. High-resolution crop yield and water productivity dataset generated using random forest and remote sensing. Sci data 2022, 9, 641. https://doi.org/10.1038/s41597-022-01761-0

Dong, J., Fu, Y., Wang, J., Tian, H., Fu, S., Niu, Z., Han, W., Zheng, Y., Huang, J., Yuan, W. Early-season mapping of winter wheat in China based on Landsat and Sentinel images. Earth Syst Sci Data, 2020 12, 3081-3095. https://doi.org/10.5194/essd-12-3081-2020

Dong, J., Lu, H., Wang, Y., Ye, T., Yuan, W. Estimating winter wheat yield based on a light use efficiency model and wheat variety data. ISPRS J Photogramm 2020, 160, 18-32. https://doi.org/10.1016/j.isprsjprs.2019.12.005

Duchemin, B., Maisongrande, P., Boulet, G., Benhadj, I. A simple algorithm for yield estimates: Evaluation for semi-arid irrigated winter wheat monitored with green leaf area index. Environ. Model. Softw. 2008, 23, 876–892. https://doi.org/10.1016/j.envsoft.2007.10.003.

Erenstein, O., Jaleta, M., Mottaleb, K.A., Sonder, K., Donovan, J., Braun, H.-J. Global Trends in Wheat Production, Consumption and Trade. In Wheat Improvement: Food Security in a Changing Climate, Reynolds, M.P., Braun, H.-J., Eds. Springer International Publishing: Cham, 2022, pp. 47-66.

FAO, IFAD, UNICEF, WFP, and WHO: The State of Food Security and Nutrition in the World 2020. ransforming food systems for affordable healthy diets, FAO, Rome, Italy, https://doi.org/10.4060/ca9692en, 2020.

FAOSTAT: Food and Agriculture Organization of the United Nations, FAO Statistical Databases, available at: http://www.fao.org/faostat/en/ (last access: 17 February 2020), 2018.

Feng, P., Wang, B., Liu, D.L., Waters, C., Xiao, D., Shi, L., Yu, Q. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric for Meteotol 2020, 285-286, 107922, https://doi.org/10.1016/j.agrformet.2020.107922.

Grogan, D., Frolking, S., Wisser, D., Prusevich, A., and Glidden, S.: Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015, Sci. Data, 9, 15, https://doi.org/10.1038/s41597-021-01115-2, 2022.

Hu, S., Shi, L., Zha, Y., Williams, M., Lin, L. Simultaneous state-parameter estimation supports the evaluation of data assimilation performance and measurement design for soil-water atmosphere-plant system. J. Hydrol. 2017, 555, 812–831. https://doi.org/10.1016/j.jhydrol.2017.10.061.

Huang, J., Tian, L., Liang, S., Ma, H., Becker-Reshef, I., Huang, Y., Su, W., Zhang, X., Zhu, D., Wu, W. Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. Agric. For. Meteorol. 2015, 204, 106–121. https://doi.org/10.1016/j.agrformet.2015.02.001.

Huang, J.; Ma, H.; Sedano, F.; Lewis, P.; Liang, S.; Wu, Q.; Su, W.; Zhang, X.; Zhu, D. Evaluation of regional estimates of winter wheat yield by assimilating three remotely sensed reflectance datasets into the coupled WOFOST–PROSAIL model. Eur. J. Agron. 2019, 102, 1–13. https://doi.org/10.1016/j.eja.2018.10.008.

Ines, A.V.M., Das, N.N., Hansen, J.W., Njoku, E.G. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 2013, 138, 149–164. https://doi.org/10.1016/j.rse.2013.07.018.

Ji, Z., Pan, Y., Zhu, X., Zhang, D., Wang, J. A generalized model to predict large-scale crop yields integrating satellite-based vegetation index time series and phenology metrics. Ecol Indic 2022, 137, 108759, https://doi.org/10.1016/j.ecolind.2022.108759.

Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., Wang, S., Ying, Y., Lin, T. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level. Glob chang bio 2020, 26, 1754-1766. https://doi.org/10.1111/gcb.14885

Jiang, Z., Huete, A.R., Didan, K., Miura, T., 2008. Development of a two-band enhanced vegetation index without a blue band. Remote Sens. Environ. 112, 3833–3845. https://doi.org/10.1016/j.rse.2008.06.006.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A.,Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T. The DSSAT cropping system model. Eur. J. Agron. 2003, 18, 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7.

June 2021. Washington, DC: International Food Policy Research Institute (IFPRI). https://doi.org/10.2499/9780896294165, 2021.

Kang Y., Ozdogan M. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ 2019, 228, 144-163. https://doi.org/10.1016/j.rse.2019.04.005

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 2003, 18, 267–288. https://doi.org/10.1016/S1161-0301(02)00108-9.

Li, L.,Wang, B., Feng, P., Wang, H., He, Q., Wang, Y., Liu, D.L., Li, Y., He, J., Feng, H. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. Agric for Meteotol 2021, 308-309, 108558, https://doi.org/10.1016/j.agrformet.2021.108558.

Li, Z., Taylor, J., Yang, H., Casa, R., Jin, X., Li, Z., Song, X., Yang, G. A hierarchical interannual wheat yield and grain protein prediction model using spectral vegetative indices and meteorological data. Field Crop Res 2020, 248, 107711, https://doi.org/10.1016/j.fcr.2019.107711.

Li, Z., Taylor, J., Yang, H., Casa, R., Jin, X., Li, Z., Song, X., Yang, G. A hierarchical interannual wheat yield and grain protein prediction model using spectral vegetative indices and meteorological data. Field Crop Res, 248: 107711. https://doi.org/10.1016/j.fcr.2019.107711

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

Luo, Y., Zhang, Z., Cao, J., Zhang, L., Zhang, J., Han, J., Zhuang, H., Cheng, F., Xu, J., Tao, F. GlobalWheatYield4km: a global wheat yield dataset at 4-km resolution during 1982-2020 based on deep learning approaches. Earth Syst Sci Data, 2022. https://doi.org/10.5194/essd-2022-297

Luo, Y., Zhang, Z., Chen, Y., Li, Z., Tao, F. Chinacropphen1km: a high-resolution crop phenological dataset for three staple crops in China during 2000–2015 based on leaf area index (LAI) products. Earth Syst Sci Data, 2020 12(1), 197-214. https://doi.org/10.5194/essd-12-197-2020

Magney, T. S., Eitel, J. U. H., Huggins, D. R., Vierling, L. A. Proximal NDVI derived phenology improves in-season predictions of wheat quantity and quality. Agric for Meteotol2016, 217, 46 – 60. http://dx.doi.org/10.1016/j.agrformet.2015.11.009

Monfreda, C., Ramankutty, N., and Foley, J. A. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000, Glob. Biogeochem. Cy., 22, GB1022, https://doi.org/10.1029/2007GB002947, 2008.

Muñoz Sabater, J., (2019): ERA5-Land monthly averaged data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (<9/10/2022>). https://doi.org/10.24381/cds.68d2bb30

National Bureau of Statistics of China: National statistical yearbook, China Statistics Press http://www.stats.gov.cn/tjsj/ndsj/2021/indexch.htm. (last access: 8 August 2022).

Paudel, D., Boogaard, H., Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N. Machine learning for large-scale crop yield forecasting. Agricultural Systems 2021, 187, 103016, https://doi.org/10.1016/j.agsy.2020.103016.

Rondeaux, G., Steven, M., Baret, F. Optimization of soil-adjusted vegetation indices. Remote Sens. Environ. 1996, 55, 95–107. https://doi.org/10.1016/0034-4257(95)00186-7.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W. Monitoring Vegetation Systems in the Great Plains with ERTS, NASA Special Publication: Washington, DC, USA, 1974, 1, 48-62.

Saltelli, A., Tarantola, S., Chan, P. S. A quantitative model-independent method for global sensitivity analysis of model output, Technometrics. 1999, 41, 39–56. https://doi.org/10.2307/1270993

Sims, D.A., Rahman, A.F., Cordova, V.D., El-Masri, B.Z., Baldocchi, D.D., Bolstad, P.V., Flanagan, L.B., Goldstein, A.H., Hollinger, D.Y., Misson, L. A new model of gross primary productivity for North American ecosystems based solely on the enhanced vegetation index and land surface temperature from MODIS. Remote Sens. Environ. 2008, 112, 1633−1646. https://doi.org/10.1016/j.rse.2007.08.004.

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., Li, H. An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China. Agric for Meteotol2021, 310, 108629, https://doi.org/10.1016/j.agrformet.2021.108629.

Wang, F., Yi, Q., Hu, J., Xie, L.,Yao, X., Xu, T., Zheng, J. Combining spectral and textural information in UAV hyperspectral images to estimate rice grain yield. In j of appl earth obs 2021, 102, 102397. https://doi.org/10.1016/j.jag.2021.102397

Wang, X., Huang, J., Feng, Q., Yin, D. Winter Wheat Yield Prediction at County Level and Uncertainty Analysis in Main Wheat-Producing Regions of China with Deep Learning Approaches. Remote Sen 2020, 12, https://doi.org/10.3390/rs12111744.

Wei, S., Li, X., Lu, Z., Zhang, H., Ye, X., Zhou, Y., Li, J., Yan, Y., Pei, H., Duan, F., Wang, D., Chen, S., Wang, P., Zhang, C., Shang ,L., Zhou, Y., Pan, P., Zhao, M, Huang, J., Bock, R., Qian, Q., Zhou, W. A transcriptional regulator that boosts grain yields and shortens the growth duration of rice. Science 377, eabi8455, https://doi.org/10.1126/science.abi8455.

Xu, X., Teng, C., Zhao, Y., Du, Y., Zhao, C., Yang, G., Jin, X., Song, X., Gu, X., Casa, R., Chen, L., Li, Z.: Prediction of wheat grain protein by coupling multisource remote sensing imagery and ECMWF data. Remote Sensing, 2020, 12: 1349.

You, L. Z., Wood, S., Wood-Sichra, U., and Wu, W. B.: Generating global crop distribution maps: From census to grid, AgrSyst, 127, 53-60, https://doi.org/10.1016/j.agsy.2014.01.002, 2014.

Zhao, Y., Han, S., Meng, Y., Feng, H., Li, Z., Chen, J., Song, X., Zhu, Y., Yang, G. Transfer-Learning-Based Approach for Yield Prediction of Winter Wheat from Planet Data and SAFY Model. Remote Sens. 2022, 14, 5474. https://doi.org/ 10.3390/rs14215474

Zhao, Y., Han, S., Zheng, J., Xue, H., Li Z., Meng, Y., Li, X., Yang, X., Li, Z., Cai, S amd Yang, G. ChinaWheatYield30m: A 30-m annual winter wheat yield dataset from 2016 to 2021 in China. https://doi.org/10.5281/zenodo.7360753

Zhao, Y., Meng, Y., Feng, H., Han, S., Yang, G., Li, Z. Should phenological information be applied to predict agronomic traits across growth stages of winter wheat? Crop J. 2022, 10, 1346–1352. https://doi.org/10.1016/j.cj.2022.08.003.

Zhang, Y., Hui, J., Qin, Q., Sun, Y., Zhang, T., Sun, H., Li, M. Transfer-learning-based approach for leaf chlorophyll content estimation of winter wheat from hyperspectral data. Remote Sens. Environ. 2021, 267, 112724. https://doi.org/10.1016/j.rse.2021.112724.