Computers and Electronics in Agriculture 194 (2022) 106790

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Developing machine learning models with multi-source environmental data to predict wheat yield in China

Linchao Li [a,b,c], Bin Wang [b,d,*], Puyu Feng [e], De Li Liu [d,f], Qinsi He [g], Yajie Zhang [b], Yakai Wang [a], Siyi Li [d,g], Xiaoliang Lu [b], Chao Yue [b], Yi Li [h], Jianqiang He [h], Hao Feng [b,h], Guijun Yang [c,i,*], Qiang Yu [b,j]

[a] College of Natural Resources and Environment, Northwest A&F University, Yangling 712100, China
[b] State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Soil and Water Conservation, Northwest A&F University, Yangling 712100, China
[c] 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
[d] NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, NSW 2650, Australia
[e] College of Land Science and Technology, China Agricultural University, Beijing 100193, China
[f] Climate Change Research Centre, University of New South Wales, Sydney, NSW 2052, Australia
[g] School of Life Sciences, Faculty of Science, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia
[h] College of Water Resources and Architectural Engineering, Northwest A&F University, Yangling 712100, China
[i] School of Geological Engineering and Surveying and Mapping, Chang'an University, Xi'an 710054, China
[j] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Keywords:
Yield prediction
Vegetation indices
NIRv
Random forest
Support vector machine
Wheat

ABSTRACT

Crop yield is controlled by different environmental factors. Multi-source data for site-specific soils, climates, and remotely sensed vegetation indices are essential for yield prediction. Algorithms of data-model fusion for crop growth monitoring and yield prediction are complicated and need to be optimized to deal with model uncertainty. This study integrated multi-source environmental variables (e.g., satellite-based vegetation indices, climate data, and soil properties) into random forest (RF) and support vector machine (SVM) models for wheat yield prediction in China. The performance of both RF and SVM models was investigated using different types of vegetation indices associated with other predictors. Relative importance and partial dependence analyses were used to identify the main predictors and their relationships with wheat yield. We found that using remotely sensed vegetation indices improved our model precision, and that near-infrared reflectance of terrestrial vegetation (NIRv) was slightly better than normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) in predicting yield. NIRv was better in detecting climate stress on crops, and could capture more information regarding crop growth and yield formation. Compared with the SVM model, the RF model with NIRv and other covariates had better performance in wheat yield prediction, with R^2 and RMSE being 0.74 and 758 kg/ha respectively. We also found that NIRv from jointing to heading was the most important predictor in determining yield, followed by solar radiation (especially during tillering–heading), relative humidity (during planting–tillering), soil organic carbon, and wind speed (throughout the growing season). In addition, wheat yield exhibited threshold-like responses to most factors based on our RF model. These threshold values can help to better understand how different environmental factors limit wheat yield, which will provide useful information for climate-adaptive crop management. Our findings demonstrated the potential of using NIRv for yield prediction. This approach is broadly applicable to other regions globally using publicly available data.

* Correspondence authors at: NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, NSW 2650, Australia (B. Wang). 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 (G. Yang).
E-mail addresses: bin.a.wang@dpi.nsw.gov.au (B. Wang), yanggj@nercita.org.cn (G. Yang).

https://doi.org/10.1016/j.compag.2022.106790
Received 24 August 2021; Received in revised form 9 February 2022; Accepted 10 February 2022
Available online 20 February 2022

1. Introduction

Food demand is increasing rapidly and has been projected to exceed food production after the mid-21st century due to the growing population and increased living standards (Bajželj et al., 2014; Li et al., 2021; Tilman et al., 2011). Wheat (*Triticum aestivum* L.) is an important staple food (Norouzi et al., 2010). It has been cultivated globally on >220 million ha per year (Shiferaw et al., 2013). China is one of the largest wheat-producing countries in the world, accounting for around 18% of global wheat production (FAO, 2018). With increased food demand, China will need to increase grain production by 36% to feed its own people (Li et al., 2014). However, China's wheat production is largely affected by climatic factors, such as heat and drought stress, even though yield per hectare has significantly increased due to the use of new adapted varieties and advances in agricultural science and technology in the past few decades (Challinor et al., 2010; Tao et al., 2014).

Accurate and timely predicting of crop yield with multi-source environmental data is crucial for ensuring national food security (Cai et al., 2019; Feng et al., 2020). Generally, process-based crop models and statistical models are common methods used for yield prediction. Process-based crop models (e.g., APSIM, CERES, and STICS) can simulate crop growth and yield formation processes, and enable investigation of the interactions between crop yield and environmental conditions (Brown et al., 2018; Feng et al., 2019a; Peng et al., 2020a). However, running crop models is time-consuming at large scales (Cao et al., 2021; Jiang et al., 2020; Peng et al., 2020b), and usually require a mass of data from field observations (e.g., cultivar characteristics, management options, and soil properties) to effectively calibrate models (Leroux et al., 2019; Li et al., 2021). In addition, processes related to extreme climate events (ECE) are greatly simplified in most crop models, resulting in less accurate yield simulations (Li et al., 2019c; Schauberger et al., 2017). Most crop models have shown relatively lower applicability than statistical models at large scales (Huang et al., 2015; Li et al., 2019d).

Compared with process-based crop models, statistical models (e.g., traditional statistical models and machine learning methods) are more efficient, and thus more widely used, in large-scale crop yield estimation (Cao et al., 2021; Peng et al., 2020b). Some studies have used traditional regression models for yield prediction. For instance, Lobell et al. (2007) predicted crop yields in California, USA using multiple regression (linear and quadratic) models during 1980–2003, and reported that using simple equations with 2–3 climate variables could explain more than two-thirds of observed yield variation. However, traditional statistical models (e.g., linear regression models) usually show lower accuracy compared with non-linear regression models. In reality, relationships between crop yields and multi-environmental factors are usually nonlinear (Jeong et al., 2016; Li et al., 2007; Li et al., 2019d). Machine learning (ML) is an advanced statistical technique that can analyze the hierarchical and nonlinear relationships between predictors and response variables (Besalatpour et al., 2014; Feng et al., 2019a; Naimi et al., 2021; Zeraatpisheh et al., 2019; Zeraatpisheh et al., 2021). Recently, many studies have developed statistics-based crop yield prediction models, such as artificial neural network (ANN), least absolute shrinkage and selection operator regression (LASSO), support vector machine (SVM), and random forest (RF) (Anna et al., 2018; Cao et al., 2020; Liakos et al., 2018; Norouzi et al., 2010; Peng et al., 2020b). For instance, Leng and Hall (2020) predicted maize yield variation in 1980–2010 in the US using the traditional linear regression model and RF. Results showed that the RF model (r = 0.93, RMSE = 246 kg/ha) performed better than the linear regression model (r = 0.51, RMSE = 506 kg/ha). Among different ML methods, the decision-tree based RF method has been widely used in different research areas (Rehfeldt et al., 2012; Singh et al., 2017; Zhao et al., 2019), with good performance in estimating crop yields (Han et al., 2020; Maya Gopal and Bhargavi, 2019). Moreover, RF is able to identify the relative importance of each predictor to the response variable.

Multiple-source environmental data have been used to predict crop yields. Integrating more informative and predictive predictors, such as soil properties and vegetation indices (VIs), is important for improving model performance. For instance, Cao et al. (2020) reported that model performance (R^2) was increased by 0.07–0.15 and 0.05–0.16 when satellite-based VIs and soil properties (e.g., soil texture, organic carbon content, pH, etc.), respectively, were included in the model compared with climate variables alone. Moreover, different VIs had different potential for predicting crop yields (Peng et al., 2020b).

Some new VIs had recently been developed, such as sun-induced chlorophyll fluorescence (SIF) products (e.g., Global Ozone Monitoring Experiment-2 (GOME-2), Orbiting Carbon Observatory-2 (OCO-2), and TROPOspheric Monitoring Instrument (TROPOMI)). These new VIs have performed better in monitoring physiological stress and carbon uptake responses than traditional VIs (e.g., normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI)) (Song et al., 2018; Zhang et al., 2014). However, some SIF products (e.g., GOME-2 with 40 km × 40 km) have coarse spatial resolution, and some products (e.g., OCO-2) have sparse and spatially discontinuous measurements (Koehler et al., 2018; Song et al., 2018). For example, TROPOMI can provide SIF data with higher temporal (almost daily) and spatial (7 km × 3.5 km) resolution (Koehler et al., 2018). However, TROPOMI lacks long-term series data, thereby limiting applications in long-term analyses. Recently, Badgley et al. (2017) provided a new gross primary production (GPP) proxy, i.e., near-infrared reflectance of terrestrial vegetation (NIR_V), that can reflect photosynthetic capacity and has strong correlations with SIF (Badgley et al., 2017). Badgley et al. (2019) found that NIR_V accurately predicted photosynthesis at FLUXNET sites and improved estimates of gridded GPP at the global scale. Wang et al. (2020) estimated optimum air temperature for rice GPP in the lower Gangetic plains and delta region using NIR_V, thereby improving the performance of the ORYZA-rice model. However, few studies have combined NIR_V with other environmental variables to predict wheat yield and compared model performance when using different satellite-based VIs (e.g., NDVI and EVI).

In this study, we explored the potential of different VIs to be used for wheat yield prediction, and investigated how major environmental predictors affect yields in China as estimated by machine learning models. The major objectives were to: (1) develop different machine learning models for wheat yield estimation based on different sets of environmental predictors; (2) quantify the relative importance of predictor variables in determining yield; (3) identify how the main predictor variables influence wheat yield in different study regions.

2. Data and methodology

2.1. Study area

The study was focused on the main wheat-growing area in mainland China that is primarily located in northwestern, northern, and subtropical areas of China (Fig. 1). The wheat harvest area in China totals 2.3×10^7 ha (http://data.stats.gov.cn/easyquery.htm?cn=C01). Actual environmental conditions vary in different sub-regions due to the vast territory and complex topography (Piao et al., 2010) that can influence crops differently. Therefore, to better predict wheat yields, we divided the wheat planting area into three sub-regions based on geography and climate conditions (Zhao, 1983). We designated these areas as sub-region I: temperate and warm-temperate northwestern China; sub-region II: warm-temperate humid and sub-humid northern China; sub-region III: subtropical humid central and southern China (Li et al., 2019a). The soil type of most areas in sub-region I and II is clay loam, and sub-region III has loamy clay. Such soil types were classified based on the international standard of soil texture classification (Chen et al., 2020b; Wu and Zhao, 2019).

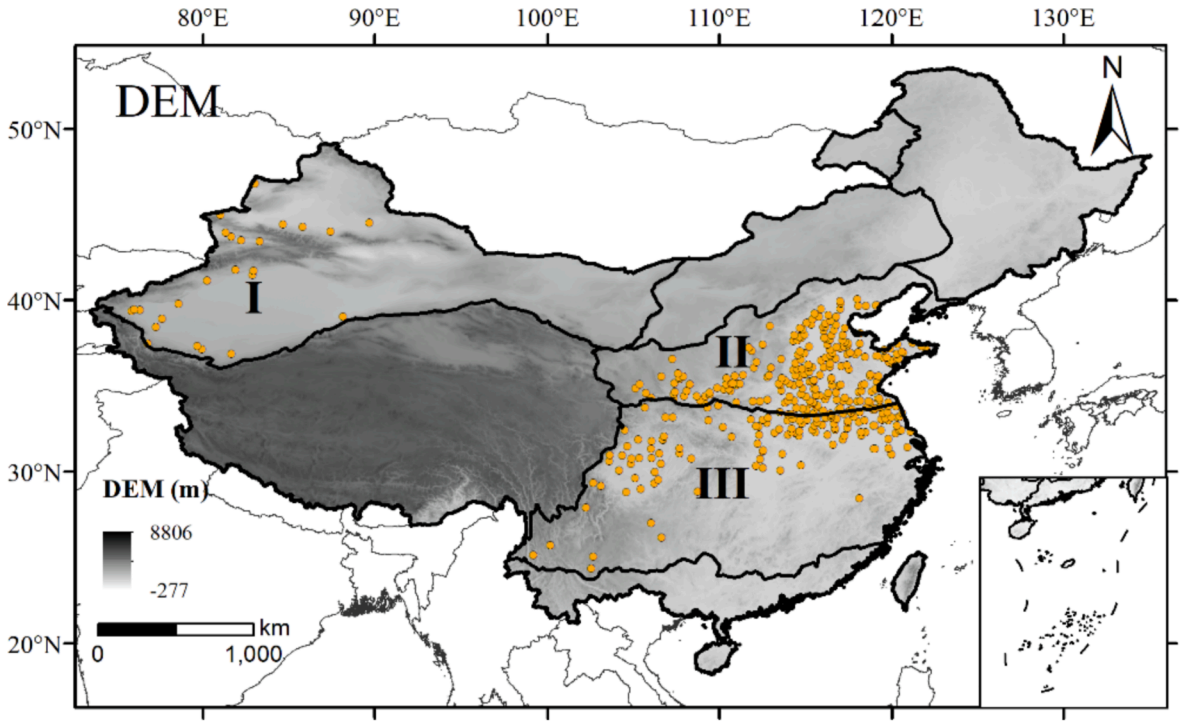

Fig. 1. Map showing the spatial distribution of 373 study sites (orange dots) and elevation (gray shading) for three sub-regions in the wheat-growing area of China.

2.2. Data

2.2.1. Wheat yield data

Wheat yield trial data from 131 sites during 2001–2013 included crop yield and growth period data. These data were collected from the China Meteorological Data Sharing Network (http://data.cma.cn/). Similar wheat yield trial data from 242 sites during 2014–2020 were collected from the national grain crop growth monitoring stations (Li et al., 2021). Management practices used for both sources of data were in keeping with local farmers' practices, and harvest methods were similar. Thus, we combined the two sources of yield data during 2001–2020 (Table S1) to predict wheat yield at the field scale. We excluded outliers of the observed yield data if they were outside of the mean ± 1.5 times the standard deviation. Moreover, since the observed yields were not continuous during 2001–2020, a final total of 1936 sets of field data were used in this study.

The wheat growing season for the trial data was divided into nine stages, including planting, emergence, tillering, overwintering, greening, jointing, heading, milk, and maturity (Chen et al., 2020b) to better explain the influence of predictors on crop yield at different growth stages. In this study, we considered four main growth periods: T1: planting–tillering (Sep–Nov); T2: tillering– jointing (Oct–Mar); T3: jointing–heading (Mar–Apr); T4: heading–maturity (May–Jun).

2.2.2. Satellite and soil property data

Different satellite-based VIs can be used to monitor crop growth conditions. NDVI, EVI, and NIR_V during 2001–2020 were selected as predictors of crop yield in this study. NIR_V is considered as a proxy for photosynthesis, and is calculated as the product of NIR reflectance and NDVI (Badgley et al., 2017). NDVI, EVI, and NIR reflectance were derived from the 16-day global vegetation indices product with a spatial resolution of 500 m (MOD13A1 V6). All VIs were extracted to site-scale values from the Google Earth Engine (GEE) platform, and then aggregated to mean values for each of four growth stages at each site.

Soil properties also play a role in determining crop yields. Our work used three soil features as predictors: soil organic carbon content (SOC), soil bulk density (SBD), and cation exchange capacity of clay (CLAY). These soil property values were collected from a China soil particle-size distribution dataset (http://globalechange.bnu.edu.cn) (Shangguan et al., 2012).

2.2.3. Climate data

The climate-related predictors used in this study included climate variables and ECEs. The climate variables were mean values of temperature (Tmean), precipitation (Prcp), wind speed (Ws), relative humidity (RHum), and sunshine hours (used to calculate solar radiation, Rad) at four growth stages (T1–T4). These data were downloaded from the China Meteorological Data Sharing Network (http://data.cma.cn/). We calculated four kinds of ECEs during the four wheat growth periods to reflect heat, frost, drought, and extreme precipitation (Table 1). Heat and frost ECEs were the number of days with the temperature higher and lower, respectively, than the fixed temperature thresholds (Feng et al., 2019a; Zheng et al., 2012). The Standardized Precipitation and Evapotranspiration Index (SPEI) was used to investigate drought intensity during the four growth periods. SPEI was estimated by standardizing the difference between precipitation and reference crop evapotranspiration (ET_0) (Vicente-Serrano et al., 2010). To better differentiate drought intensity for each growth period, SPEI was calculated at a 1-month timescale. Crop yields are also strongly influenced by precipitation or extreme precipitation. Drizzle and heavy precipitation influence crop yield differently (Lesk et al., 2020). Therefore, to investigate the impact of different precipitation intensities on wheat yield, we designated drizzle as R5 (count of days with precipitation between 0.1 and 5 mm) and heavy precipitation as R20 (count of days with precipitation greater than or equal to 20 mm) (Table 1).

2.3. Modelling methodology

The overall framework of this study is shown in Fig. 2. We first divided the wheat growing season into four periods using the phenology data. Then, we predicted wheat yield using two machine learning methods. The feature importance of each predictor and the nonlinear relationships between predictors and yields were also determined.

Table 1
Environmental variables used in crop yield estimation during 2001–2020.

Type		Term	Definition	Resolution Resolution		Data source
				Temporal	Spatial	
Climate variables	Climate data	Pr (mm)	Total precipitation	daily	site	http://data.cma.cn/
		Tmean (°C)	Mean temperature	daily	site	Same as above
		Rad (MJ m^{-2})	Mean solar radiation*	daily	site	Same as above
		RHum (mm)	Relative humidity	daily	site	Same as above
		Ws (m/s)	Wind speed	daily	site	Same as above
	Extreme climate events	Frost (days)	Number of days with daily Tmin 0 °C	Growth stage	Site	Same as above
		R20 (days)	Count of days for precipitation ≥ 20 mm	Growth stage	Site	Same as above
		Heat (days)	Number of days with daily T_{max} > 28°C	Monthly	site	Same as above
		R5 (days)	Count of days for precipitation 0.1–5 mm	Growth stage	site	Same as above
		Drought	Standardized precipitation and evapotranspiration index (SPEI)	Growth stage	Site	Same as above
		DHW	Number of days with T_{max} > 29; Rhum < 30; Ws > 2	Growth stage	Site	Same as above
Vegetation	Satellite-based vegetation indices	NDVI	Normalized Difference Vegetation Index	16-day	500 m	https://lpdaac.usgs.gov/products/mod13a1v006/
		EVI	Enhance Vegetation Index	16-day	500 m	Same as above
		NIRv	Near-infrared reflectance of terrestrial vegetation	16-day	500 m	Same as above
Soil	Soil properties	SBD	Soil bulk density	\	1 km	Soil particle-size distribution dataset (Shangguan et al., 2012).
		SOC	Soil organic carbon content	\	1 km	Same as above
		CLAY	Cation exchange capacity of clay		1 km	Same as above

* Sunshine hours were used to calculate solar radiation based on equations provided by Allen et al. (1998).

2.3.1. Feature selection

The input data was composed of different variables over four growth periods, resulting in a large dataset that increased the workload and the possibility of overfitting the model due to correlated or unreasonable variables. Therefore, we excluded variables that showed insignificant correlations ($P > 0.05$) with wheat yield. As shown in Figs. S1–S4, the correlation and significance level of climate (climate variable and ECEs) and yield varied with the four growth periods. For instance, during T2, Rad, Tmean, Ws, RHum, Frost, and R5 exhibited strong significant correlations ($P < 0.001$) with crop yield, while the Prcp, DHW (dry and hot wind), and Heat showed insignificant correlations (Fig. S2). This was mainly because ECEs such as DHW and Heat rarely occurred in this period. By contrast, DHW and Heat at T4 had significant correlations with wheat yield (Fig. S4). Precipitation was significantly correlated with yield during T3 and T4, but was insignificantly correlated during T1 and T2. Interestingly, the precipitation-related index (R5) had significant correlations with crop yield throughout the growing season (T1–T4), indicating the necessity of defining ECEs in crop yield estimation. We found that drizzle (R5) had a stronger negative impact on yield than heavy rainfall (R20). The same ECEs may have different influences on wheat yield at different growth stages. For instance, frost had positive correlations with wheat yield during T1–T2, but had negative correlations during T3–T4. Therefore, key growth stages or growth periods should also be considered in predicting crop yields.

Additionally, in order to consider multicollinearity of predictors, we calculated variance inflation factors (VIF) for all predictors in the RF model (Vittinghoff et al., 2011). Predictors with VIF > 10 were removed (Vittinghoff et al., 2011). We found vegetation indices (e.g., NDVI, EVI, and NIR_V) had strong multicollinearity during the same growth period. Therefore, we developed yield prediction models using different vegetation indices (NDVI, EVI, and NIR_V). We developed five kinds of RF models by using different data sets (Table S1): M1 used climate data; M2 used climate + soil data; M3 used climate + soil + NDVI; M4 used climate + soil + EVI; M5 used climate + soil + NIRv data.

2.3.2. Support vector machine

The SVM method is a widely used ML model for regression and classification analysis developed by Cortes and Vapnik (1995). In the SVM model, given a set of observed samples for input and output data, the best fit line is the hyperplane that has the maximum number of points (Besalatpour et al., 2012). Unlike other ML models that try to minimize the error between observed and simulated data, SVM aims to find the best line within the threshold values. In this study, we used the radial basis function (RBF) kernel for SVM. There are two parameters (Cost and Sigma) of the SVM RBF kernel. Cost was set between 2, 4, 8, 16, 32, and 64; Sigma was set between 0.02 and 0.08 at intervals of 0.005. We selected the optimal parameters with three replicates of the three-fold cross validation (Fig. S5 and Table S2).

2.3.3. Random forest model

The RF model is a non-parametric approach based on the ensemble of classification and regression trees (Breiman, 2001). Each tree is built by bootstrap samples, leaving around one-third of all samples for validation. Each tree returns the mean or average prediction to improve the performance of that data set. The RF model can capture the relationship (nonlinear or linear) between yield and predictors (Feng et al., 2020). Recently, the RF method has been widely used for yield estimation (Feng et al., 2019b), and for determining yield response to climate factors (Hoffman et al., 2020). The range of m_{try} (the number of variables randomly sampled as candidates at each split) was set from 1 to 25 with 2 intervals; the range of n_{tree} (the number of trees to grow in the forest) was set from 100 to 900 with 200 intervals. Our study applied the RF model with the optimal values of m_{try} and n_{tree} for each data set (Fig. S6 and Table S2).

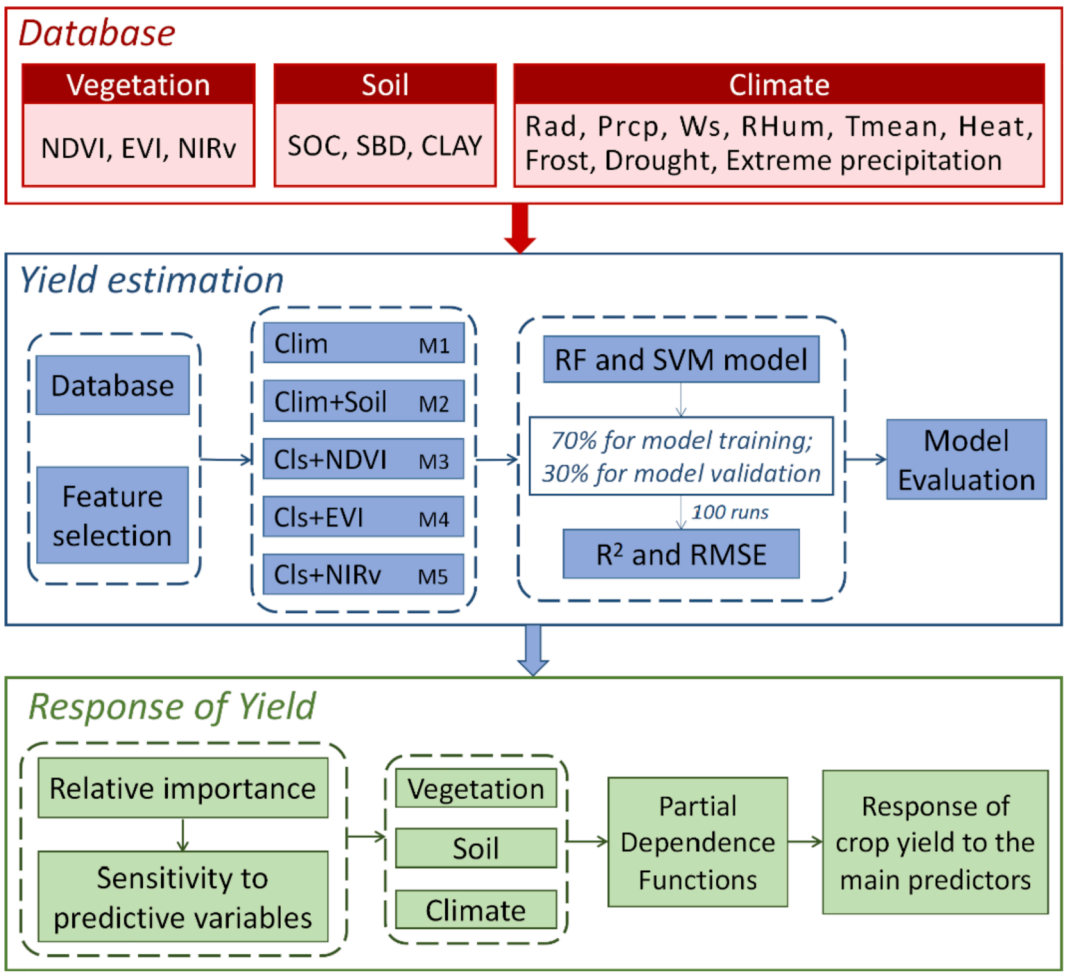

Fig. 2. Framework of the wheat yield prediction model integrating multi-source data with different machine learning techniques. NDVI, Normalized Difference Vegetation Index; EVI, Enhanced Vegetation Index; NIRv, Near-infrared reflectance of terrestrial vegetation; SOC, Soil organic carbon content; SBD, Soil bulk density; CLAY, Cation exchange capacity of clay; Rad, Solar radiation; Prcp, Precipitation; Ws, Wind speed; RHum, Relative humidity; Tmean, mean temperature; DHW, dry and hot wind; ML, machine learning (Random forest and support vector machine); Cls, climate data and soil property data. RF, Random forest; SVM, Support vector machine.

The RF algorithm can also evaluate the importance of each predictor. We first aggregated multi-source data for the four growth periods (T1–T4) at each site for yield prediction. We then applied these predictors in the RF model to predict crop yield. The relative importance of each variable was evaluated by the "%IncMSE" metric. The variables with high relative importance were identified as the main factors that influence wheat yield. We used partial dependence plots from the RF model to estimate the response of crop yield to growth-period predictors.

2.3.4. Model performance assessment

Both RF and SVM were repeatedly run 100 times. Each run used randomly selected values of 70% of the total data for training and 30% of the total data for validation over the entire study area. We used the coefficient of determination (R^2) and root mean square error (RMSE) to assess the performance of the RF and SVM models.

We used the 'randomForest', 'caret', 'e1071', and 'ggplot2' packages in R (version 3.6.1, https://www.r-project.org/) for model development and data analysis.

3. Results

3.1. Model performance

We evaluated the performance of each model across 100 repeated runs. Results showed that for both SVM and RF models, using input data of M3, M4, and M5 had better performance than that of M1 and M2 (Fig. 3), indicating that using more informative predictors would result in higher accuracy of yield estimation. Using M5 data resulted in slightly better performance than M3 and M4, as indicated by the highest mean R^2 values (RF, 0.74; SVM, 0.69) and the lowest mean RMSE values (RF, 758 kg/ha; SVM, 821 kg/ha). These results suggested that including NIRv could slightly improve the performance of wheat yield prediction compared with using EVI and NDVI. Also, we did see that RF showed better performance than SVM in wheat yield prediction regardless of data sets used (Fig. 3).

We used the model with climate + soil + NIRv data as the final model (predictors identified in Table S2) to evaluate the model performance in three sub-regions (Fig. S7). We found that the models' performance varied among the three sub-regions. Generally, the accuracy of wheat prediction was highest in sub-region III ($R^2 = 0.82$, RMSE = 685 kg/ha), followed by sub-region II ($R^2 = 0.66$, RMSE = 761 kg/ha) and sub-region I ($R^2 = 0.1$, RMSE = 890 kg/ha). The RF model showed poor performance in sub-region I mainly due to the insufficient amount of

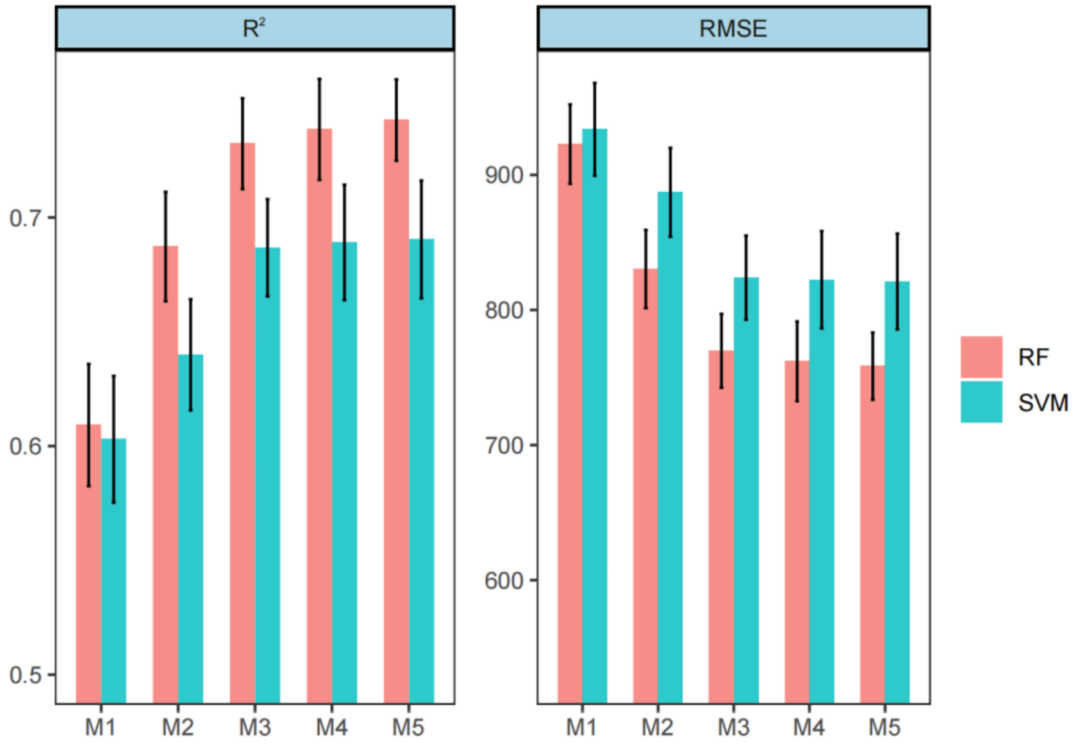

Fig. 3. Model performance of wheat yield prediction using different multi-source data. Each model was evaluated based on 100 repeated runs during 2001–2020. The colored bars represent the mean values of R^2 and RMSE; the error bars represent the standard errors across 100 runs. RF: Random Forest; SVM: Support Vector Machine; M1: climate data only; M2: climate + soil data; M3: climate + soil + NDVI data; M4: climate + soil + EVI data; M5: climate + soil + NIRv data.

data and some agronomic measures (e.g., irrigation) that were not reflected by multi-source environmental data. The RF model showed better performance in sub-region II and III, and the SVM model showed slightly better performance in sub-region I.

3.2. Relative importance of predictors

The relative importance of predictor variables used in models RF_M3, RF_M4, and RF_M5 is shown in Fig. 4. For vegetation indices, the importance of VI was 27% for the NDVI-based model (RF_M3), 29% for the EVI-based model (RF_M4), and 30% for the NIR_V-based model (RF_M5). The importance of soil properties was the same (10%) for all three models. The relative importance of all climate-related predictors was around 60% (RF_M3: 63%, RF_M4: 61%, and RF_M5: 60%). Fig. 5 shows the relative importance of the top 15 variables for each model. The results consistently showed that wheat yield was mainly influenced by VIs (NDVI_3, EVI_3, and NIRv_3) during T3, and Rad during T2 (Rad_2) and T3 (Rad_3)) across the three models (Fig. 5).

3.3. Wheat yield dependence on predictors

3.3.1. The response of yield to vegetation indices

We further investigated how NIRv during three wheat growth periods influenced yield based on the RF_M5 model. The partial

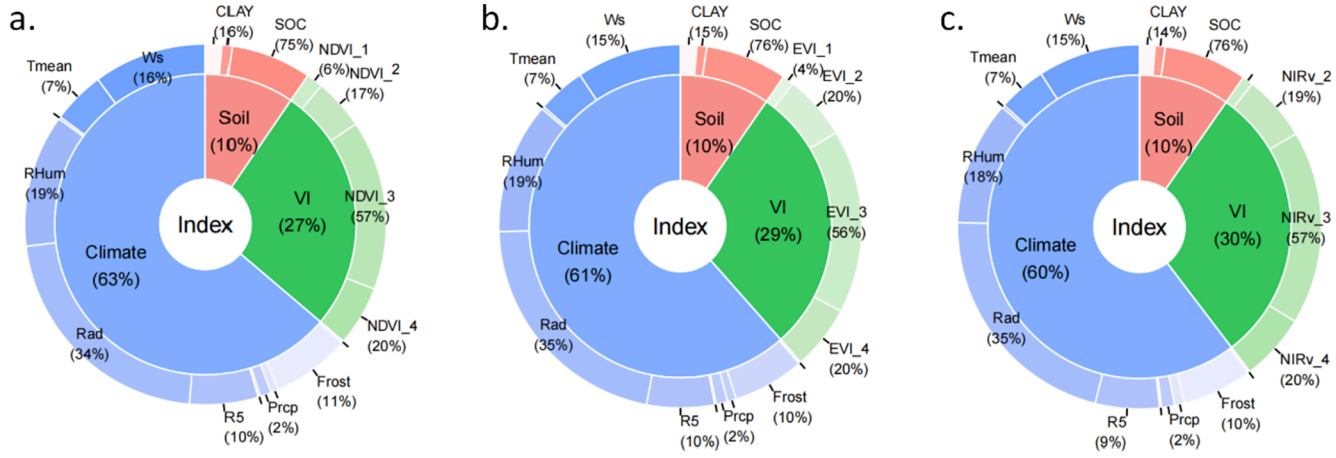

Fig. 4. Proportion of relative importance of different sources of environmental data (climate, soil, and vegetation indices). Each predictor represents the total relative importance during four growth periods. The variables were scaled to sum to 100%. a) relative importance breakdown for the RF_M3 model (using NDVI, soil property, and climate data); b) relative importance breakdown for the RF_M4 model (using EVI, soil property, and climate data); c) relative importance breakdown for the RF_M5 model (using NIRv, soil property, and climate data). Soil: soil properties included CLAY, SOC, and SBD: Climate: climate variables and extreme climate events.

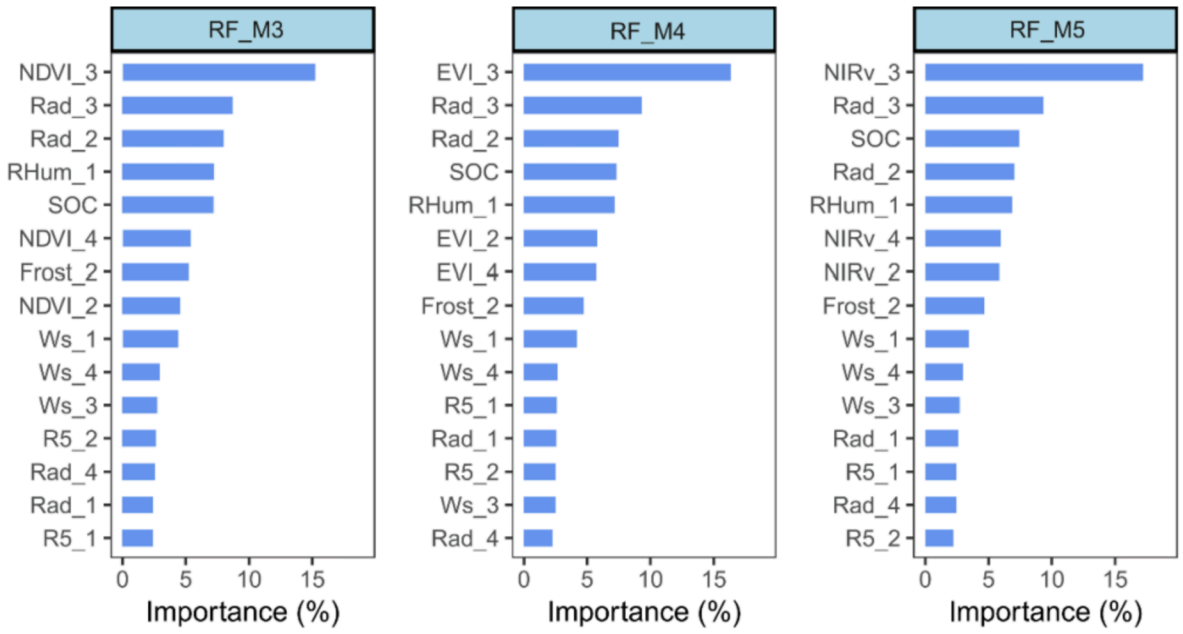

Fig. 5. Relative importance of predictor variables (ranked for the first fifteen) from different RF models. NDVI, Normalized Difference Vegetation Index; EVI, Enhanced Vegetation Index; NIRv, Near-infrared reflectance of terrestrial vegetation; SOC, Soil organic carbon content; Rad, Radiation; Ws, Wind speed; RHum, Relative humidity; R5, Count of days for precipitation 0.1–5 mm; RF, random forest. 1, 2, 3, and 4 represent different growth stages.

dependence plot (PDP) of NIRv revealed the relationship between wheat yield and NIRv_2 (19.1%), NIRv_3 (57.4%), and NIRv_4 (19.9%) (Fig. 6). The response curve of wheat yield vs. NIRv showed a near-linear relationship. For example, when NIRv_2 was in the range of 0.01–0.16, wheat yield would increase greatly as NIRv_2 increased. However, wheat yield became stable when NIRv_2 was over 0.16 (Fig. 6a). When NIRv_3 was in the range of 0.02–0.21, wheat yield would increase greatly as NIRv_3 increased. However, wheat yield became stable when NIRv_3 was over 0.21 (Fig. 6b). Similar results were found for NIRv_4 (Fig. 6c).

The soil property data also greatly contributed to yield estimation. However, those data were not dynamic data (they did not change with time) (Shangguan et al., 2012). Therefore, the PDPs of soil properties were not shown.

3.3.2. *The response of yield to climate variables*

Climate factors strongly influenced wheat yield in different growth periods. The marginal effect of climate predictors (ranked for the first twelve climate predictors), representing the curves of crop yield response to each variable, is shown in Fig. 7. Climate conditions varied in each sub-region (boxplots shown in Fig. 7), indicating that the effect of climate-related factors may vary in different sub-regions. For instance, Rad during T3 (Rad_3) was the main influencing factor in yield prediction, and predicted wheat yields would be highest when Rad_3 was around 16 MJ/m^2 (Fig. 7a). Note that low radiation in sub-region III (Rad_3 < 16 MJ/m^2) was likely to limit wheat yield. Similar results were found for Rad during T2 (Rad_2) (Fig. 7b). Crop yield was also affected by RHum_1 (when RHum_1 > 61%) (Fig. 7c). Frost_2 showed a positive influence on wheat yield when Frost_2 < 90 days. When Frost_2 was>90

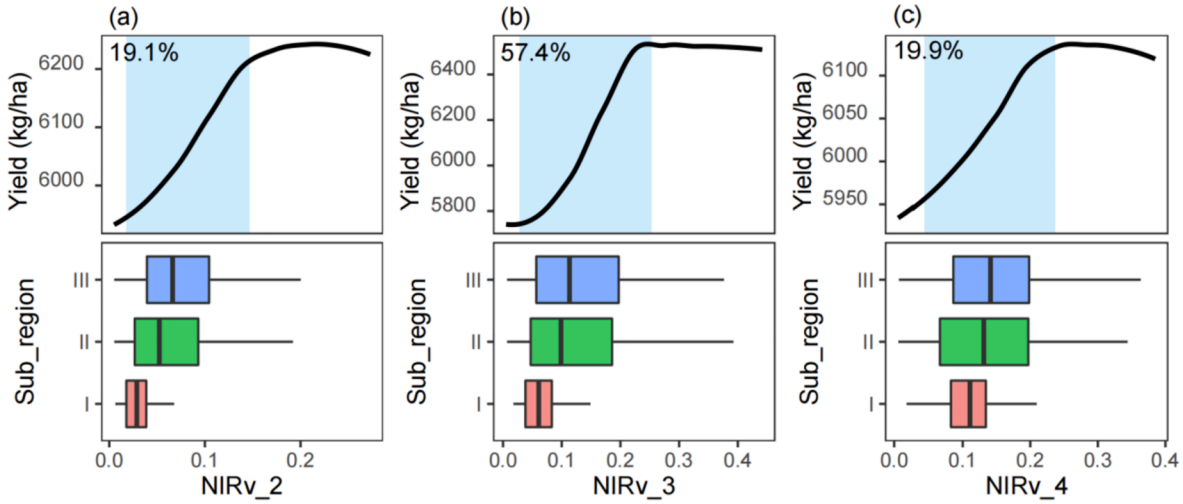

Fig. 6. Partial dependence plot of NIRv with the relative importance value ranked in the first three during four growth stages based on the RF_M5 model. The black lines are smoothed representations of the response with fitted values (model predictions) for the calibration data. The trend of the line, rather than the actual values, describes the nature of the dependence of wheat yield on the predictors. The blue shaded area represents calibration data between the 10th and 90th percentile. The percentage values represent the relative importance of each predictor generated from the random forest model. The box plots indicate the variability and range of NIRv values in different sub-regions.

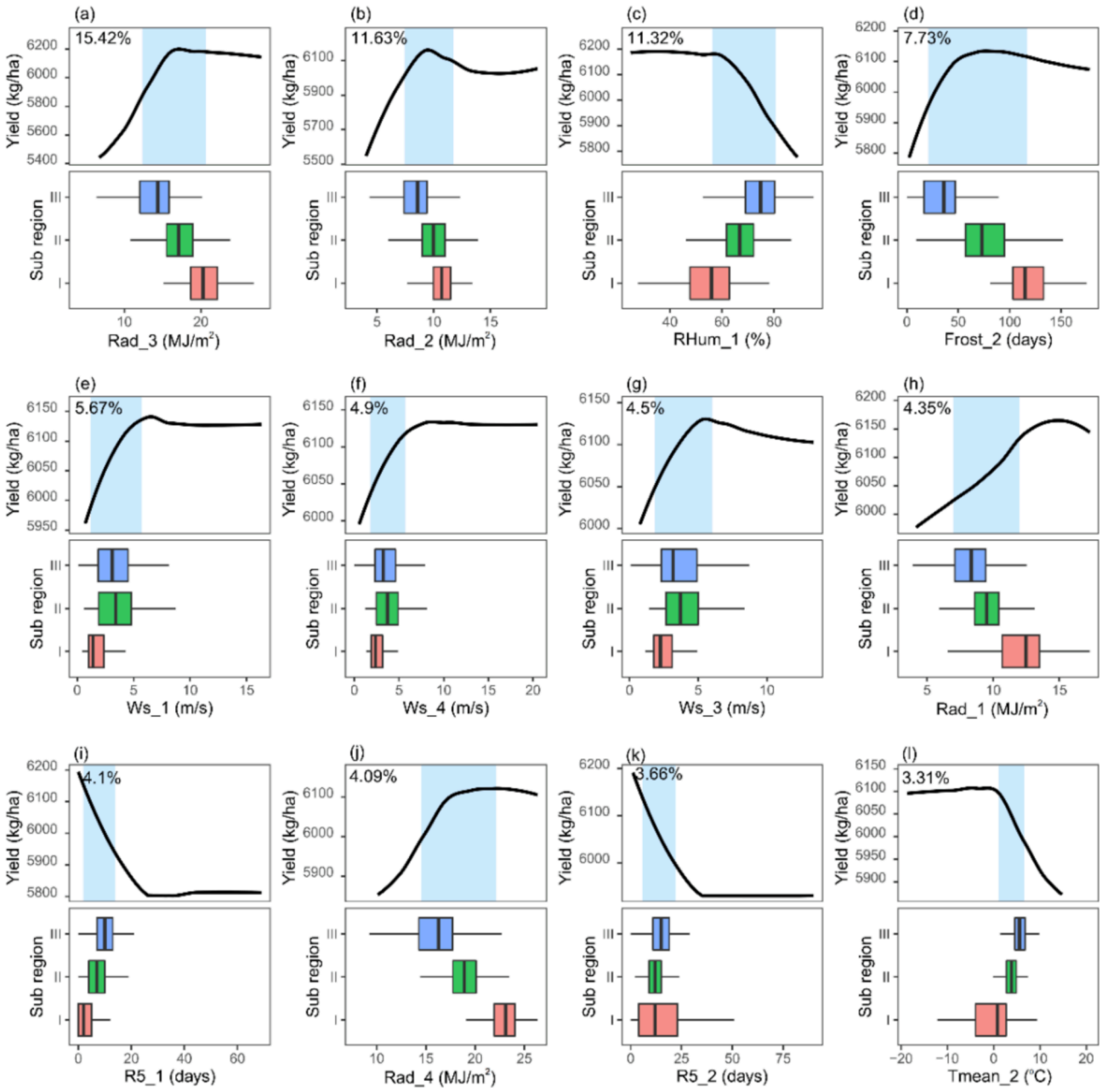

Fig. 7. Partial dependence plots of different climate variables with feature importance ranked for the first twelve climate predictors in different sub-regions using model RF_M5. The black lines are smoothed representations of the response, with fitted values (model predictions) for the calibration data. The trend of the line, rather than the actual values, describes the nature of the dependence of wheat yield on the predictors. The blue shaded area represents calibration data between the 10th and 90th percentile. The percentage values represent the relative importance of each predictor generated from the RF_M5 model. The box plots indicate the variability and range of each climate variable in different sub-regions.

days, wheat yield declined slightly (Fig. 7d). Wind speed was strongly related to wheat yield throughout various growth periods, and showed a positive influence when wind speed was lower than 5–6 m/s (Fig. 7e–g). Wheat yield showed a similar response to Rad during T1 (Fig. 7h) and T4 (Fig. 7j). R5_1 (Fig. 7i), R5_2 (Fig. 7k), and Tmean_2 (when Tmean_2 > 4 °C) (Fig. 7l) showed negative influences on wheat yield. The PDPs of other climate factors are shown in Fig. S8.

4. Discussion

Our results showed that machine learning models with multi-source environmental data can provide reliable wheat yield prediction at the field scale. Previous studies have predicted wheat yield at the county level. For instance, Wang et al. (2020a) predicted winter wheat yield with Convolution Neural Networks (CNN) and Long Short-Term Memory (LSTM) in China, and showed acceptable performance (R^2 was 0.74 and RMSE was 721 kg/ha). Similarly, Han et al. (2020) used RF, SVM, and Gaussian process regression (GPR) to predict county-level yield with R^2 higher than 0.75. In contrast, Cao et al. (2021a) predicted wheat yield with RF and three deep learning methods (CNN, LSTM, and deep neural networks) using multi-source environmental data at county level with $R^2 \geq 0.85$ and RMSE $\leq$ 768 kg/ha, during 2011–2015 across 629 counties, and at field level with R^2 ranging from 0.48 to 0.71 and RMSE from 956 to 1620 kg/ha, during 2011–2013 over 87 sites. Note that predicting crop yield at the field level is more difficult because environmental conditions can be quite variable even across the same county,

and therefore require higher resolution data sets (Feng et al., 2020). Our RF model at a field trial scale yielded similar predictive results compared with previous studies at the county level, mainly due to the longer-term time-series data (2001–2020) used in our study.

We developed different ML models to predict wheat yield in China. We found that the NIRv-based RF model showed slightly better performance in predicting yield than NDVI- and EVI-based models (Fig. 3), indicating that NIRv can capture slightly more information regarding crop growth and yield formation. Moreover, NIRv isolates the soil background from vegetation signals and can distinguish the distribution of photosynthesis with canopy depth (Badgley et al., 2017; Huang et al., 2019; Ryu et al., 2019). Peng et al. (2020b) assessed the potential of different satellite-based VIs (e.g., NDVI, EVI, land surface temperature, NIRv, and SIF (OCO-2, GOME2, and TROPOMI)) for yield prediction of maize and soybean yield in the U.S. They also found that using NIRv could lead to better yield estimation than using other VIs, and using NIRv could lead to similar or even better performance than using SIF products (OCO-2 and TROPOMI). One of the reasons for this conclusion could be that NIRv had strong correlations with GPP, and showed better performance in detecting climate stress than other VIs, such as those associated with drought and heat (Badgley et al., 2019; Wang et al., 2020). NIRv is a MODIS-based index with a higher spatial resolution (<1 km) than SIF products. However, some recently planned geostationary missions aim to produce SIF products with higher temporal-spatial resolution (e.g., GeoCarb and TEMPO), thereby providing great potential for yield prediction with SIF products.

We found that NIRv during T3 was the most important predictor influencing wheat yield estimation (Figs. 4 and 5). This was primarily because the number of tillers and amount of leaf area reach their peak during this time interval (Shao et al., 2013). Thus, crop growth conditions and photosynthesis can be detected by vegetation indices during this period. Our results showed that predicted wheat yields were nearly linearly related to NIRv_3 when NIRv was below a threshold of 0.21, indicating that photosynthetic rate was the main limiting factor for wheat yield. However, wheat yield showed little change once NIRv was over 0.21 (Fig. 6b), indicating potentially asymmetric influences of other factors on wheat yield and photosynthesis, such as light-use efficiency (He et al., 2020; Liu et al., 2017), temperature, or solar radiation factors (Chen et al., 2020a). Moreover, photosynthesis is not always directly related to final crop yield due to feedback influences from interactions of crop growth, development, and environmental conditions (Wu et al., 2019). Therefore, we speculate that crop yield is limited by other factors (e.g., climate conditions and crop nitrogen status) when NIRv is greater than such threshold values.

We found that solar radiation during T3 and T2 was the most important climate variable after NIRv_3 (Fig. 5). This is because solar radiation is important to the photosynthesis process (Hernández-Barrera and Rodríguez-Puebla, 2017), and therefore, response curves are very similar to the yield-NIRv response (Fig. 6, and Fig. 7a–b). Moreover, other climate-related factors (such as drought) had less impact on yield due to management practices (e.g., irrigation). Relative humidity during T1 was also important and negatively affected wheat yield when RHum_1 was>61% (Fig. 7c). One of the reasons for this finding could be high relative humidity increases the risk of physiological disorders and plant diseases (Hand, 1987). It is worth mentioning that SOC was also important in our RF_M5 model because the SOC pool can enhance many ancillary benefits (e.g., increase soil quality and crop productivity), as consistently reported in previous studies (Hammad et al., 2020; Lal, 2006; Majumder et al., 2008). Frost_2 showed a positive impact on wheat yield when Frost_2 was <90 days. This is because wheat needs a low temperature to complete vernalization and has high frost tolerance during the early vegetative period (Bergjord et al., 2008; Xiao et al., 2018). Also, our results showed that light winds during T1, T4, and T3 increased wheat yield (Fig. 7e–g), while drizzle decreased wheat yield (Fig. 7i and k). We speculate that photosynthesis could be promoted by light winds (Ws < 6 m/s), thereby increasing crop biomass. For instance, leaf flutter in response to light winds could lead to better penetration of solar radiation into lower canopy layers, thereby providing more opportunities to maintain greater canopy photosynthesis (Burgess et al., 2019; Roden and Pearcy, 1993). In the case of drizzle, water use efficiency is low for crops when drizzle typically comes with low radiation and temperature, resulting in suitable conditions for foliar fungal pathogens (Harvell et al., 2002; Lesk et al., 2020; Sun and Woods, 1994).

In different sub-regions, the impact of climate-related factors is varied due to different climate conditions. For instance, Rad is one of the main variables influencing crop yield, and its influence should be given greater attention in sub-region III because solar radiation in sub-region III (e.g., Sichuan Basin) is lower than in sub-region I and II (Lau et al., 2007). In addition, frost showed opposite effects in sub-region I and sub-region III (Fig. 7). This is because frost events occur more frequently and frost days are higher than the threshold value in sub-region I, resulting in greater wheat yield losses. Therefore, region-specific characteristics should be considered in climate-related impact studies.

In general, our study quantified how wheat yields respond to climate variables and ECEs, and therefore provides valuable information for reducing the risk of yield loss caused by climate. In our previous studies, we found that drought, heat, and extreme precipitation are likely to become more frequent in the future, resulting in climate conditions more unsuitable for crop growth (Li et al., 2019a; Li et al., 2019b; Yao et al., 2020). Several adaptations to climate change have been assessed by using process-based crop models, such as cultivating new varieties, optimizing sowing date, using irrigation, and managing crop residues (Challinor et al., 2014). Although such adaptations have shown clear contributions of around 7–15% to increasing crop yield (Challinor et al., 2014), the results from process-based crop models are questionable due to poor capabilities under ECEs (Feng et al., 2019a; Li et al., 2019c). Therefore, developing a hybrid model by incorporating machine learning with biophysical models should be pursued to provide more robust climate change impact assessment and to explore the potential of different adaptation options under climate change.

Some limitations of our study should be mentioned. First, statistics-based models are data-driven models. Model performance is largely dependent on the volume of trial data. Second, some satellite-based data, such as combined high-resolution SIF products (e.g., OCO-2 and TROPOMI), can lead to potential improvements in yield estimation (Zhang et al., 2018). Most recently, Camps-Valls et al. (2021) developed a new vegetation index, kernel NDVI (kNDVI) that shows stronger correlations with GPP than NDVI and EVI. Moreover, Dechant et al. (2020) demonstrated that NIRvP (calculated as NIRv × PAR) is a robust proxy of far-red SIF, and showed a stronger correlation with SIF than other indices. Such indices may have great potential for improving crop biomass prediction, and these indices were not considered in this study. Third, despite the fact that machine-learning-based models showed good performance in yield prediction, they lack rational biophysical explanations regarding yield response to environmental conditions (Roberts et al., 2017). In addition, the shift of wheat cultivars in different years was not considered in this study, therefore, attention should be given to changes in cultivars when predicting crop yield. To bridge this gap, our future work will be focused on developing hybrid models by integrating process-based models and machine learning techniques to increase model performance regarding crop yield prediction.

5. Conclusion

We developed machine learning based models to predict wheat yield by incorporating multi-source environmental data (including soil properties, climate, and vegetation indices). We draw the following main conclusions:

(1) RF had better performance than SVM in predicting wheat yield. The RF model using NIRv ($R^2 = 0.74$; RMSE = 758 kg/ha) could lead to slightly better prediction than using EVI ($R^2 = 0.73$;

RMSE = 762 kg/ha) or NDVI (R^2 = 0.73; RMSE = 770 kg/ha). Moreover, vegetation-based indices had the greatest influence on wheat yields compared with other environmental covariates.

(2) Based on RF_M5, we found that NIRv during T3 was the most important predictor for determining crop yield. In addition, Rad (T2 and T3), RHum (T1), SOC, Ws (T1–T4), R5 (T1–T2), and Frost (T2) were identified as the main factors limiting wheat yield. Drought had a relatively small contribution to yield change because irrigations were applied for winter wheat production in this study.

(3) We constructed PDP plots to explain how different predictors at each developmental period affected wheat yield based on RF models. We found that NIRv showed both linear and nonlinear relationships with wheat yield. Wheat yield had threshold-like responses to other environmental variables. These PDP results can help to better understand how factors limit wheat yield.

Our findings demonstrated the potential of using NIRv for yield prediction and our modelling approach is broadly applicable in other regions globally using publicly available data. Our yield prediction model can be enhanced in the future by incorporating process-based crop models or other newly developed vegetation indices.

CRediT authorship contribution statement

Linchao Li: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing-original draft . **Bin Wang:** Conceptualization, Funding acquisition, Methodology, Resources, Validation, Writing-original draft. **Puyu Feng:** Formal analysis, Visualization. **De Li Liu:** Conceptualization, Writing - original draft. **Qinsi He:** Investigation. **Yajie Zhang:** Methodology. **Yakai Wang:** Investigation, Software. **Xiaoliang Lu:** Validation, Writing-review & editing. **Chao Yue:** Validation, Writing-review & editing. **Yi Li:** Project administration, Writing-review & editing. **Jianqiang He:** Writing-review & editing. **Hao Feng:** Writing-review & editing. **Guijun Yang:** Data curation, Formal analysis, Funding acquisition, Project administration, Supervision, Validation. **Qiang Yu:** Funding acquisition, Project administration, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the Natural Science Foundation of China (No. 41961124006, 41730645 and 52079114), the Natural Science Foundation of Qinghai (2021-HZ-811), and the National Key Research and Development Program of China (2019YFE0125300 and 2017YFE0122500). We thank two anonymous reviewers and the editor for their helpful comments to improve the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2022.106790.

References

Anna, C., Salah, S., Brett, W., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151, 61–69.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M, 1998. Crop evapotranspiration-guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper 56. FAO, Roma, Italia.

Badgley, G., Anderegg, L.D.L., Berry, J.A., Field, C.B., 2019. Terrestrial gross primary production: Using NIRV to scale from site to globe. Glob Chang Biol 25 (11), 3731–3740.

Badgley, G., Field, C.B. and Berry, J.A.J.S.A., 2017. Canopy near-infrared reflectance and terrestrial photosynthesis. 3(3): e1602244.

Bajželj, B., Richards, K.S., Allwood, J.M., Smith, P., Dennis, J.S., Curmi, E., Gilligan, C.A., 2014. Importance of food-demand management for climate mitigation. Nat. Clim. Change 4 (10), 924–929.

Bergjord, A.K., Bonesmo, H., Skjelvåg, A.O., 2008. Modelling the course of frost tolerance in winter wheat. Eur. J. Agron. 28 (3), 321–330.

Besalatpour, A.A., Ayoubi, S., Hajabbasi, M.A., Jazi, A.Y., Gharipour, A., 2014. Feature selection using parallel genetic algorithm for the prediction of geometric mean diameter of soil aggregates by machine learning methods. Arid Land Res. Manage. 28 (4), 383–394.

Besalatpour, A., Hajabbasi, M., Ayoubi, S., Gharipour, A., Jazi, A., 2012. Prediction of soil physical properties by optimized support vector machines. Int. Agrophys. 26 (2), 109–115.

Breiman, L., 2001. Random forests. Machine Learn. 45 (1), 5–32.

Brown, J.N., Hochman, Z., Holzworth, D., Horan, H., 2018. Seasonal climate forecasts provide more definitive and accurate crop yield predictions. Agric. For. Meteorol. 260-261, 247–254.

Burgess, A.J., Gibbs, J.A., Murchie, E.H., 2019. A canopy conundrum: can wind-induced movement help to increase crop productivity by relieving photosynthetic limitations? J. Exp. Bot. 70 (9), 2371–2380.

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

Camps-Valls, G., Campos-Taberner, M., Moreno-Martínez, Á., Walther, S., Duveiller, G., Cescatti, A., Mahecha, M.D., Muñoz-Marí, J., García-Haro, F.J., Guanter, L., Jung, M., Gamon, J.A., Reichstein, M., Running, S.W., 2021. A unified vegetation index for quantifying the terrestrial biosphere. Sci. Adv. 7 (9) https://doi.org/10.1126/sciadv.abc7447.

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Han, J., Li, Z., 2020. Identifying the Contributions of Multi-Source Data for Winter Wheat Yield Prediction in China. Remote Sensing 12 (5), 750. https://doi.org/10.3390/rs12050750.

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

Challinor, A.J., Simelton, E.S., Fraser, E.D., Hemming, D. and Collins, M.J.E.R.L., 2010. Increased crop failure due to climate change: assessing adaptation options using models and socio-economic data for wheat in China. 5(3), 034012.

Challinor, A.J., Watson, J., Lobell, D.B., Howden, S.M., Smith, D.R., Chhetri, N., 2014. A meta-analysis of crop yield under climate change and adaptation. Nat. Clim. Change 4 (4), 287–291.

Chen, A., Mao, J., Ricciuto, D., Xiao, J., Frankenberg, C., Li, X., Thornton, P.E., Gu, L., Knapp, A.K., 2020a. Moisture availability mediates the relationship between terrestrial gross primary production and solar-induced chlorophyll fluorescence: Insights from global-scale variations. Glob Chang Biol. 27 (6), 1144–1156.

Chen, X., Li, Y.i., Yao, N., Liu, D.L., Javed, T., Liu, C., Liu, F., 2020b. Impacts of multi-timescale SPEI and SMDI variations on winter wheat yields. Agric. Syst. 185, 102955. https://doi.org/10.1016/j.agsy.2020.102955.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learn. 20 (3), 273–297.

Dechant, B., Ryu, Y., Badgley, G., Köhler, P., Rascher, U., Migliavacca, M., et al., 2020. NIRvP: a robust structural proxy for sun-induced chlorophyll fluorescence and photosynthesis across scales.

FAO, 2018. World Food and Agriculture–Statistical Pocketbook.

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

Feng, P., Wang, B., Liu, D.L., Waters, C., Yu, Q., 2019a. Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. Agric. For. Meteorol. 275, 100–113.

Feng, P., Wang, B., Liu, D.L., Yu, Q., 2019b. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia. Agric. Syst. 173, 303–316.

Hammad, H.M., Khaliq, A., Abbas, F., Farhad, W., Fahad, S., Aslam, M., Shah, G.M., Nasim, W., Mubeen, M., Bakhat, H.F., 2020. Comparative effects of organic and inorganic fertilizers on soil organic carbon and wheat productivity under arid region. Commun. Soil Sci. Plant Anal. 51 (10), 1406–1422.

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., Zhang, J., 2020. Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China. Remote Sensing 12 (2), 236. https://doi.org/10.3390/rs12020236.

Hand, D., 1987. Effects of atmospheric humidity on greenhouse crops. In: Symposium on Biological Aspects of Energy Saving in Protected Cultivation, vol. 229, pp. 143–158.

Harvell, C.D., Mitchell, C.E., Ward, J.R., Altizer, S., Dobson, A.P., Ostfeld, R.S., Samuel, M.D., 2002. Climate warming and disease risks for terrestrial and marine biota. Science 296 (5576), 2158–2162.

He, L., Magney, T., Dutta, D., Yin, Y.i., Köhler, P., Grossmann, K., Stutz, J., Dold, C., Hatfield, J., Guan, K., Peng, B., Frankenberg, C., 2020. From the Ground to Space: Using Solar-Induced Chlorophyll Fluorescence to Estimate Crop Productivity. Geophys. Res. Lett. 47 (7) https://doi.org/10.1029/2020GL087474.

Hernández-Barrera, S., Rodríguez-Puebla, C., 2017. Wheat yield in Spain and associated solar radiation patterns. Int. J. Climatol. 37, 45–58.

L Hoffman, A., R Kemanian, A., E Forest, C., 2020. The response of maize, sorghum, and soybean yield to growing-phase climate revealed with machine learning. Environ. Res. Lett. 15 (9), 094013. https://doi.org/10.1088/1748-9326/ab7b22.

Huang, J., Tian, L., Liang, S., Ma, H., Becker-Reshef, I., Huang, Y., Su, W., Zhang, X., Zhu, D., Wu, W., 2015. 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. 204, 106–121.

Huang, M., Piao, S., Ciais, P., Peñuelas, J., Wang, X., Keenan, T.F., Peng, S., Berry, J.A., Wang, K., Mao, J., Alkama, R., Cescatti, A., Cuntz, M., De Deurwaerder, H., Gao, M., He, Y., Liu, Y., Luo, Y., Myneni, R.B., Niu, S., Shi, X., Yuan, W., Verbeeck, H., Wang, T., Wu, J., Janssens, I.A., 2019. Air temperature optima of vegetation productivity across global biomes. Nat. Ecol. Evol. 3 (5), 772–779.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.-M., Gerber, J.S., Reddy, V.R., Kim, S.-H., Gonzalez-Andujar, J.L., 2016. Random forests for global and regional crop yield predictions. PLoS ONE 11 (6), e0156571.

Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., Wang, S., Ying, Y., Lin, T., 2020. 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. Change Biol. 26 (3), 1754–1766.

Köhler, P., Frankenberg, C., Magney, T.S., Guanter, L., Joiner, J., Landgraf, J., 2018. Global retrievals of solar induced chlorophyll fluorescence with TROPOMI: first results and inter-sensor comparison to OCO-2. Geophys. Res. Lett. 45 (19), 10,456–10,463.

Lal, R., 2006. Enhancing crop yields in the developing countries through restoration of the soil organic carbon pool in agricultural lands. Land Degrad. Dev. 17 (2), 197–209.

Lau, C.C.S., Lam, J.C., Yang, L., 2007. Climate classification and passive solar design implications in China. Energy Convers. Manage. 48 (7), 2006–2015.

Leng, G., Hall, J.W., 2020. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. Environ. Res. Lett. 15 (4), 044027. https://doi.org/10.1088/1748-9326/ab7b24.

Leroux, L., Castets, M., Baron, C., Escorihuela, M.-J., Bégué, A., Lo Seen, D., 2019. Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices. Eur. J. Agron. 108, 11–26.

Lesk, C., Coffel, E., Horton, R., 2020. Net benefits to US soy and maize yields from intensifying hourly rainfall. Nat. Clim. Change 10 (9), 819–822.

Li, A., Liang, S., Wang, A., Qin, J., 2007. Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. Photogramm. Eng. Remote Sens. 73 (10), 1149–1157.

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

Li, L., Yao, N., Li, Y., Liu, D.L., Wang, B., Ayantobo, O.O., 2019a. Future projections of extreme temperature events in different sub-regions of China. Atmos. Res. 217, 150–164.

Li, L., Yao, N., Liu, D.L., Song, S., Lin, H., Chen, X., Li, Y., 2019b. Historical and future projected frequency of extreme precipitation indicators using the optimized cumulative distribution functions in China. J. Hydrol. 579, 124170. https://doi.org/10.1016/j.jhydrol.2019.124170.

Li, Y., Guan, K., Schnitkey, G.D., DeLucia, E., Peng, B., 2019c. Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. Glob Chang Biol 25 (7), 2325–2337.

Li, Y., Guan, K., Yu, A., Peng, B., Zhao, L., Li, B.o., Peng, J., 2019d. Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the U.S. Field Crops Res. 234, 55–65.

Li, Y., Zhang, W., Ma, L., Wu, L., Shen, J., Davies, W.J., Dou, Z., 2014. An analysis of C hina's grain production: looking back and looking forward. Food Energy Security 3 (1), 19–32.

Liakos, K., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine Learning in Agriculture: A Review. Sensors (Basel) 18 (8), 2674. https://doi.org/10.3390/s18082674.

Liu, L., Guan, L., Liu, X., 2017. Directly estimating diurnal changes in GPP for C3 and C4 crops using far-red sun-induced chlorophyll fluorescence. Agric. For. Meteorol. 232, 1–9.

Lobell, D.B., Cahill, K.N., Field, C.B., 2007. Historical effects of temperature and precipitation on California crop yields. Clim. Change 81 (2), 187–203.

Majumder, B., Mandal, B., Bandyopadhyay, P.K., Gangopadhyay, A., Mani, P.K., Kundu, A.L., Mazumdar, D., 2008. Organic amendments influence soil organic carbon pools and rice–wheat productivity. Soil Sci. Soc. Am. J. 72 (3), 775–785.

Maya Gopal, P.S., Bhargavi, R., 2019. Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms. Appl. Artif. Intell. 33 (7), 621–642.

Naimi, S., Ayoubi, S., Demattê, J.A., Zeraatpisheh, M., Amorim, M.T.A., Mello, F.A.D.O., 2021. Spatial prediction of soil surface properties in an arid region using synthetic soil image and machine learning. Geocarto Int. 1–24.

Norouzi, M., Ayoubi, S., Jalalian, A., Khademi, H., Dehghani, A.A., 2010. Predicting rainfed wheat quality and quantity by artificial neural network using terrain and soil characteristics. Acta Agric. Scandinavica Section B-Soil Plant Sci. 60 (4), 341–352.

Peng, B., Guan, K., Tang, J., Ainsworth, E.A., Asseng, S., Bernacchi, C.J., Cooper, M., Delucia, E.H., Elliott, J.W., Ewert, F., Grant, R.F., Gustafson, D.I., Hammer, G.L., Jin, Z., Jones, J.W., Kimm, H., Lawrence, D.M., Li, Y., Lombardozzi, D.L., Marshall-Colon, A., Messina, C.D., Ort, D.R., Schnable, J.C., Vallejos, C.E., Wu, A., Yin, X., Zhou, W., 2020a. Towards a multiscale crop modelling framework for climate change adaptation assessment. Nat Plants 6 (4), 338–348.

Peng, B., Guan, K., Zhou, W., Jiang, C., Frankenberg, C., Sun, Y., He, L., Köhler, P., 2020b. Assessing the benefit of satellite-based Solar-Induced Chlorophyll Fluorescence in crop yield prediction. Int. J. Appl. Earth Obs. Geoinf. 90, 102126. https://doi.org/10.1016/j.jag.2020.102126.

Piao, S., Ciais, P., Huang, Y., Shen, Z., Peng, S., Li, J., Zhou, L., Liu, H., Ma, Y., Ding, Y., Friedlingstein, P., Liu, C., Tan, K., Yu, Y., Zhang, T., Fang, J., 2010. The impacts of climate change on water resources and agriculture in China. Nature 467 (7311), 43–51.

Rehfeldt, G.E., Crookston, N.L., Sáenz-Romero, C., Campbell, E.M., 2012. North American vegetation model for land-use planning in a changing climate: A solution to large classification problems. Ecol. Appl. 22 (1), 119–141.

Roberts, M.J., Braun, N.O., Sinclair, T.R., Lobell, D.B., Schlenker, W., 2017. Comparing and combining process-based crop models and statistical models with some implications for climate change. Environ. Res. Lett. 12 (9), 095010. https://doi.org/10.1088/1748-9326/aa7f33.

Roden, J.S., Pearcy, R.W., 1993. Effect of leaf flutter on the light environment of poplars. Oecologia 93 (2), 201–207.

Ryu, Y., Berry, J.A., Baldocchi, D.D., 2019. What is global photosynthesis? History, uncertainties and opportunities. Remote Sens. Environ. 223, 95–114.

Schauberger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Khabarov, N., Müller, C., Pugh, T.A.M., Rolinski, S., Schaphoff, S., Schmid, E., Wang, X., Schlenker, W., Frieler, K., 2017. Consistent negative response of US crops to high temperatures in observations and crop models. Nat. Commun. 8 (1) https://doi.org/10.1038/ncomms13931.

Shangguan, W., Dai, Y., Liu, B., Ye, A., Yuan, H., 2012. A soil particle-size distribution dataset for regional land and climate modelling in China. Geoderma 171-172, 85–91.

Shao, G.C., Lan, J.J., Yu, S.E., Liu, N., Guo, R.Q., She, D.L., 2013. Photosynthesis and growth of winter wheat in response to waterlogging at different growth stages. Photosynthetica 51 (3), 429–437.

Shiferaw, B., Smale, M., Braun, H.-J., Duveiller, E., Reynolds, M., Muricho, G., 2013. Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. Food Security 5 (3), 291–317.

Singh, B., Sihag, P., Singh, K., 2017. Modelling of impact of water quality on infiltration rate of soil by random forest regression. Model. Earth Syst. Environ. 3 (3), 999–1004.

Song, L., Guanter, L., Guan, K., You, L., Huete, A., Ju, W., Zhang, Y., 2018. Satellite sun-induced chlorophyll fluorescence detects early response of winter wheat to heat stress in the Indian Indo-Gangetic Plains. Glob Chang Biol. 24 (9), 4023–4037.

Sun, D.-W., Woods, J., 1994. Low temperature moisture transfer characteristics of wheat in thin layers. Trans. ASAE 37 (6), 1919–1926.

Tao, F., Zhang, Z., Xiao, D., Zhang, S., Rötter, R.P., Shi, W., Liu, Y., Wang, M., Liu, F., Zhang, H.e., 2014. Responses of wheat growth and yield to climate change in different climate zones of China, 1981–2009. Agric. For. Meteorol. 189-190, 91–104.

Tilman, D., Balzer, C., Hill, J. and Befort, B.L., 2011. From the Cover: Global food demand and the sustainable intensification of agriculture. 108(50), 20260.

Vicente-Serrano, S.M., Beguería, S., López-Moreno, J.I., 2010. A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. J. Clim. 23 (7), 1696–1718.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C., McCulloch, C.E., 2011. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. Springer Science & Business Media.

Wang, X., Wang, S., Li, X., Chen, B., Wang, J., Huang, M., Rahman, A., 2020. Modelling rice yield with temperature optima of rice productivity derived from satellite NIRv in tropical monsoon area. Agric. For. Meteorol. 294, 108135. https://doi.org/10.1016/j.agrformet.2020.108135.

Wu, A., Hammer, G.L., Doherty, A.l., von Caemmerer, S., Farquhar, G.D., 2019. Quantifying impacts of enhancing photosynthesis on crop yield. Nat Plants 5 (4), 380–388.

Wu, K., Zhao, R., 2019. Soil texture classification and its application. Acta Pedol. Sin. 56 (01), 227–241 (in Chinese with English abstract).

Xiao, L., Liu, L., Asseng, S., Xia, Y., Tang, L., Liu, B., Cao, W., Zhu, Y., 2018. Estimating spring frost and its impact on yield across winter wheat in China. Agric. For. Meteorol. 260-261, 154–164.

Yao, N., Li, L., Feng, P., Feng, H., Li Liu, D., Liu, Y., Jiang, K., Hu, X., Li, Y., 2020. Projections of drought characteristics in China based on a standardized precipitation and evapotranspiration index and multiple GCMs. Sci. Total Environ. 704, 135245. https://doi.org/10.1016/j.scitotenv.2019.135245.

Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P., 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338, 445–452.

Zeraatpisheh, M., Ayoubi, S., Mirbagheri, Z., Mosaddeghi, M.R., Xu, M., 2021. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. Geoderma Regional 27, e00440. https://doi.org/10.1016/j.geodrs.2021.e00440.

Zhang, Y., Guanter, L., Berry, J.A., Joiner, J., van der Tol, C., Huete, A., Gitelson, A., Voigt, M., Köhler, P., 2014. Estimation of vegetation photosynthetic capacity from space-based measurements of chlorophyll fluorescence for terrestrial biosphere models. Glob. Change Biol. 20 (12), 3727–3742.

Zhang, Y., Joiner, J., Alemohammad, S.H., Zhou, S., Gentine, P., 2018. A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. Biogeosciences 15 (19), 5779–5800.

Zhao, S., 1983. A new scheme for comprehensive physical regionalization in China. Acta Geograph, Sinica 38 (1), 1–10 (In Chinese with English abstract).
Zhao, X., Yu, B., Liu, Y., Chen, Z., Li, Q., Wang, C., Wu, J., 2019. Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh. Remote Sensing 11 (4), 375.
Zheng, B., Chenu, K., Fernanda Dreccer, M., Chapman, S.C., 2012. Breeding for the future: what are the potential impacts of future frost and heat events on sowing and flowering time requirements for Australian bread wheat (Triticum aestivium) varieties? Glob Chang Biol. 18 (9), 2899–2914.