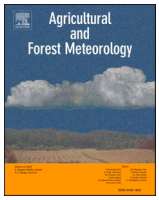
Agricultural and Forest Meteorology 315 (2022) 108810

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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



Weather records from recent years performed better than analogue years when merging with real-time weather measurements for dynamic within-season predictions of rainfed maize yield

Shang Chen [a,b], Liang He [c,*], Wenbiao Dong [a,b], Ruotong Li [a,b], Tengcong Jiang [a,b], Linchao Li [d], Hao Feng [b,d], Kuifeng Zhao [e], Qiang Yu [d,e], Jianqiang He [a,b,e,**]

[a] Key Laboratory for Agricultural Soil and Water Engineering in Arid Area of Ministry of Education, Northwest A&F University, Yangling 712100, China
[b] Institute of Water-Saving Agriculture in Arid Areas of China, Northwest A&F University, Yangling, Shaanxi 712100, China
[c] National Meteorological Center, Beijing 100081, China
[d] State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Water and Soil Conservation, Northwest A&F University, Yangling 712100, China
[e] Key Laboratory of Eco-Environment and Meteorology for the Qinling Mountains and Loess Plateau, Shaanxi Provincial Meteorological Bureau, Xi'an, Shaanxi 710015, China

ARTICLE INFO

Keywords:
Maize
Yield prediction
Weather analogue
Within-season
The Loess Plateau

ABSTRACT

Within-season crop yield prediction with a dynamic crop model can provide valuable references for field management practices and regional food security. However, weather ensembles containing the unknown future weather conditions occurring after prediction dates are essential for such predictions using crop models. Two strategies were established for selecting analogue weather years as the target growing season based on a five-year maize experiment conducted at eight sites in the Loess Plateau of China. The first strategy tried weather data from different lengths of years ahead the planting year. The second strategy used the *k*-nearest neighbor (*k*-NN) algorithm to select analogue weather according to different combinations of weather variables with daily or accumulative values. The results showed that satisfactory predictions could be obtained after maize tasseling (about 50 d prior to maturity). The mean absolute relative error (ARE) and coefficient of variation (CV) of the daily yield predictions after tasseling were 6.6% and 5.7%, respectively, in 2010 at the Yulin site. In the leading-year strategy, the most reliable predictions were obtained by the weather data from the 10 years ahead of planting, with an overall average ARE of 11.7%. In the *k*-NN strategy, the most reliable predictions were obtained by using the analogue weather selected with only accumulative precipitation, with an overall average ARE of 11.5%. Additionally, both of the two optimal strategies improved the original predictions in most cases. However, the *k*-NN strategy was more likely to generate worse predictions in the early part of the growing season. Generally, it was more convenient to use the weather data of 10 leading years before the planting year to represent the unknown weather data after the prediction dates. This strategy provided reliable prediction accuracy without complex programming and requirement for long-term weather records.

1. Introduction

The demands of seasonal crop yield forecasting are increasing in both developed and developing countries (Basso and Liu, 2018). Predictions before harvest are valuable for timely warning of meteorological risks (Hansen et al., 2004), and provide references for decision-making, especially for farms with low disaster resistance ability (Brandes et al., 2016; Chen et al., 2020). Hence, many crop yield prediction methods have been developed across the world (Chipanshi et al., 2015; Feng et al., 2020; Schwalbert et al., 2020). Generally, several methods have been commonly used to predict yield, including statistical regression-based methods, field surveys, and process-based cropping

* Corresponding author.
** Corresponding author at: Key Laboratory for Agricultural Soil and Water Engineering in Arid Area of Ministry of Education, Northwest A&F University, Yangling 712100, China.
E-mail addresses: heliang@cma.gov.cn (L. He), jianqiang_he@nwsuaf.edu.cn (J. He).

https://doi.org/10.1016/j.agrformet.2022.108810
Received 22 August 2021; Received in revised form 17 December 2021; Accepted 3 January 2022
Available online 10 January 2022

system models (crop models).

Statistical methods have been widely used in yield prediction worldwide (Gouache et al., 2015; Lee et al., 2013) in which relationships between crop yield and plant growth status with weather variables (Qian et al., 2009), and remote sensing signals (Cao et al., 2021a, 2021b) have been established. However, the statistical methods ignored the effects of field management practices and environmental conditions on crop growth. Additionally, the remote-sensing-based strategies might be ineffective on cloudy days (Kang et al., 2016) and later growth stages (Haboudane, 2004). Field survey method uses field-measured or questionnaire-surveyed information about plant growth status, and can only provide short lead times for yield prediction (Feng et al., 2020). The prediction accuracy of this method is heavily influenced by the weather conditions after the prediction day and the representativeness of the sampling locations. Additionally, field surveys are also time- and labor-consuming and are not suitable for yield predictions over large areas (Nandram et al., 2014).

Crop models based on biophysical processes and mechanisms can be used to analyze the dynamic interactions of crop genotypes, soil properties, climate conditions, and field practices and have been widely used for crop yield prediction (Lecerf et al., 2018; Morell et al., 2016; Togliatti et al., 2017). However, weather data covering entire growing seasons are usually required before applying crop models for crop growth simulation. Since real weather data can be obtained daily from a local weather station, the challenge in crop-model-based yield predictions lie in the prediction of unknown future weather data after the prediction days (Basso and Liu, 2018; Tollenaar et al., 2017). Average values of multiple historical years have often been used to generate the unknown future weather data on the same day of the target year, and have been further used to predict seasonal yield (Dumont et al., 2014). However, the simulation results have not always been reasonable during the early part of the growing season due to the nonlinear relationship between crop growth and weather conditions (Semenov and Barrow, 1997).

With the development of global/regional circulation models (GCMs/RCMs), median- and long-term climatic predictions have provided new solutions for crop yield forecasts (Baigorria et al., 2008; de Wit et al., 2010; Prakash et al., 2019). However, the original weather data predicted by the GCMs were usually unsuitable for crop growth simulations at both spatial and temporal scales (Bakker et al., 2013; Quiring and Legates, 2008), and yield prediction accuracy diminished with the longer lead times (Dumont et al., 2014). Another way to obtain field-scale weather data is based on stochastic weather generators (WG) (Kilsby et al., 2007; Mavromatis and Hansen, 2001). However, the employment of WGs requires large computing resources, and WGs can perform poorly in the prediction of both precipitation frequency and magnitude (Hartkamp et al., 2003). Additionally, prior distributions of weather variables are required for use of WGs, and therefore new uncertainty can be introduced in the generation of weather data (Bannayan and Hoogenboom, 2008a). At the same time, historical weather data from nearby weather stations provide an alternative method for generating the unknown weather data after the prediction days (Chipanshi et al., 1997). However, yield prediction with the entire records of long-term historical weather is time-consuming to deal with. Additionally, remarkable changes in agricultural production caused by climate changes have been reported worldwide (Harkness et al., 2020; Liu et al., 2020; Tao et al., 2006). Differences in weather conditions might be huge between recent years and the far-distant years. Hence, the entire weather records may not be effective for accurate yield predictions because many years in the weather ensembles may have different weather conditions from what will occur in the target year.

Researchers have given attentions to the selection of historical years with analogue weather, or analogue years. For instance, du Toit and du Toit (2003) selected analogue weather by calculating the index of agreement (D-index) between actual daily observed weather data and historical weather data. The years with the highest fitness were then selected to drive crop models for dynamic within-season yield predictions. Bannayan and Hoogenboom (2008b) used the *k*-nearest neighbor (*k*-NN) algorithm to select the *k* years with most analogue weather conditions based on the Euclidean distance. Chen et al. (2017) modified the *k*-NN algorithm by calculating the Euclidean distance with the seven-day moving-average values of each weather variable. Generally, these studies selected analogue years based on all of the available weather variables (e.g., solar radiation, maximum temperature, minimum temperature, and precipitation). However, Porter et al. (1999) pointed out that the weather variables that mainly influenced final crop yield were air temperature and precipitation. Chen et al. (2020) also reported that rainfed maize yield was mainly affected by the seasonal precipitation in the Loess Plateau of China. Hence, it is necessary to evaluate the method for selecting analogue years for within-season dynamic yield predictions. In addition, crop growth and development are usually determined by the accumulative effects of weather variables. For example, emergence and duration of crop leaves are dependent primarily on heat accumulation or thermal time (Hodges and Evans, 1992). How can we consider the influence of accumulative values of weather variables in the selection of analogue years? This is another knowledge gap that needs to be addressed.

Maize (*Zea mays* L.) is the dominant crop in the Loess Plateau due to the limited water resource and frost-free window, and rainfed maize accounts for more than 80% local arable land (Huang et al., 2011). In this study, rainfed maize yields in the Loess Plateau were dynamically predicted with the DSSAT-CERES-Maize model driven by weather data of analogue years selected by different methods. The main objectives were to (1) assess the accuracy of within-season maize yield forecasts based on local multi-year weather records; (2) establish and evaluate different methods for the selection of the analogue years; and (3) select the most effective methods for within-season predictions of rainfed maize yields in the Loess Plateau.

2. Materials and methods

2.1. Study areas

The Loess Plateau (100°54′–114°33′E and 33°43′–41°16′ N) covers an area of 0.65×10^6 km^2 in northwest China (Fig. 1). The annual mean temperature ranges from 3.6 to 14.3 °C and annual precipitation ranges from 150 to 700 mm from northwest to southeast. Maize is the predominant crop in this region, planted in April and harvested in September. Most precipitation occurs in summer (July, August, and September) in this region, resulting in water stresses in the early maize growing season. The rate of increase in air temperature in this region (0.6 °C decade^{-1}) was greater than the global average increment (0.13 °C decade^{-1}) (IPCC, 2013; Wang et al., 2012), resulting in accelerated phenological development of local maize (He et al., 2015).

2.2. Brief description of the DSSAT-CERES-Maize model

The CERES-Maize model embedded in DSSAT (Decision Support System for Agrotechnology Transfer) was specifically developed to simulate daily crop growth and development of maize, including phenological states, biomass production, and grain yield (Hoogenboom et al., 2017; Jones et al., 1986, 2003). The CERES-Maize model consists of nonlinear, dynamic mathematical functions that describe maize growth and yield formation as well as changes in soil water and nutrient contents at field scale. This model simulates maize growth by considering field practices and is driven by daily weather conditions. Daily potential biomass production is determined by temperature and the interception of photosynthetically active radiation by the plant canopy. All model components are described by a set of parameters. Soil inputs are given as parameters related to physical, chemical, and morphological properties of different soil layers. Crop management information includes crop cultivar; planting date, depth and density; row space;

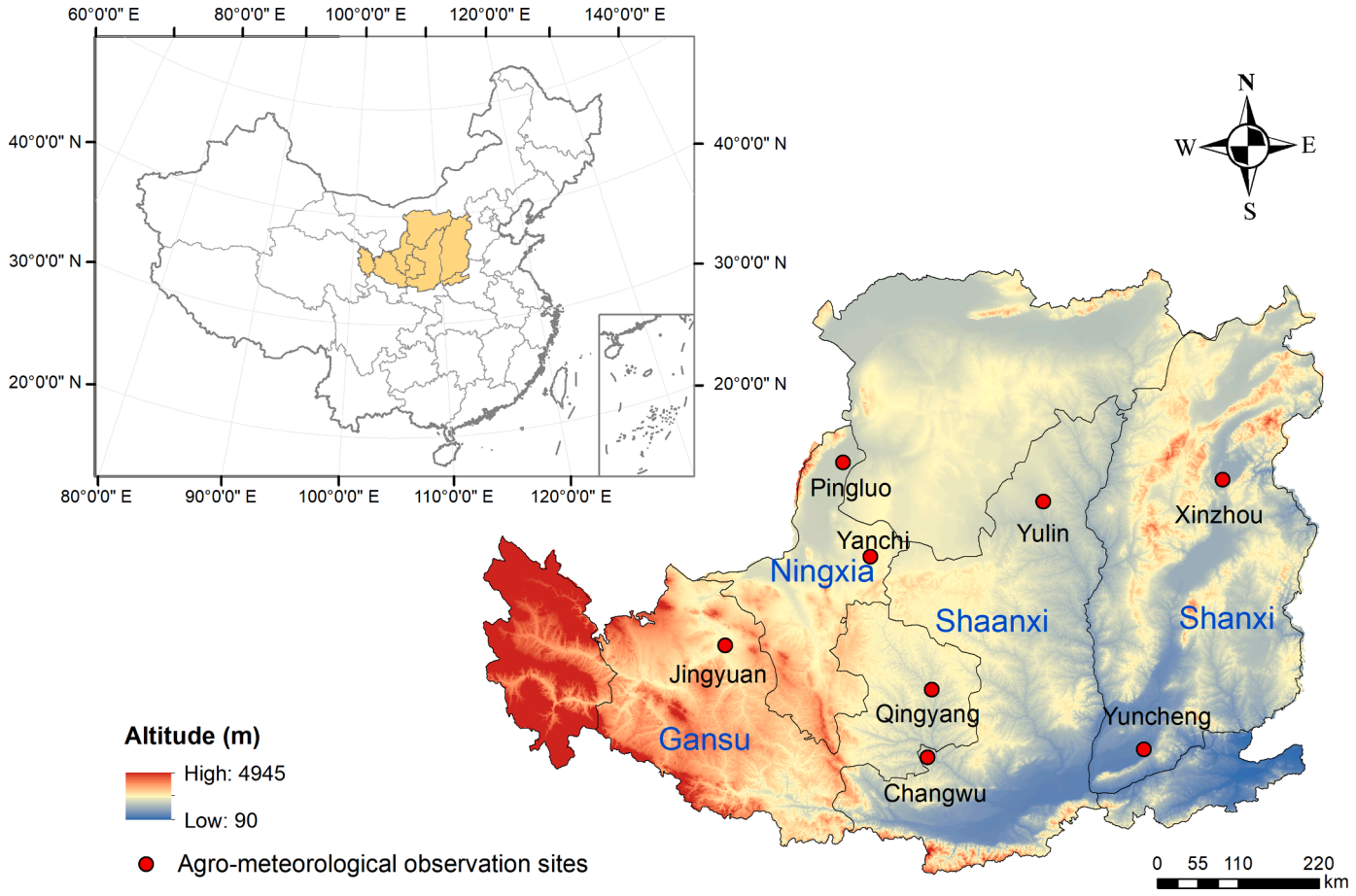


Fig. 1. Locations of the study area and eight agro-meteorological observation sites (Xinzhou and Yuncheng in Shanxi Province; Yulin and Changwu in Shaanxi Province; Qingyang and Jingyuan in Gansu Province; Pingluo and Yanchi in Ningxia Province) in the Loess Plateau, China.

irrigation; fertilizer; and application of organic amendments. A simple water balance algorithm referred to as the "tipping-bucket" approach is used in the CERES-Maize model to calculate yield reduction under water stress. Maize development rates are calculated based on air temperature and photoperiod. Crop physiological features are represented by genetic coefficients associated with cultivar, ecotype, and species.

2.3. Datasets

Four groups of data are generally needed for DSSAT simulations: weather, soil, crop, and management. Daily weather data include daily solar radiation (R_s, MJ m^{-2}), maximum air temperature (T_{max}, °C), minimum air temperature (T_{min}, °C), and precipitation (P, mm). In this study, weather data in 1961–2010 were obtained from the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/). Since solar radiation data were not available for the eight sites, daily cumulative solar radiation was estimated based on daylength and sunshine hours through the Ångström-Prescott formula (Ångström, 1924; Prescott, 1940). Soil profile parameters, including saturated soil moisture, residual soil moisture, and soil hydraulic conductivity, were obtained from the China Soil Hydraulic Parameters Dataset (Dai et al., 2013).

In this study, five-year maize experiments were conducted at eight agro-meteorological observation stations in the Loess Plateau in 2006–2010 (Fig. 1). These stations belong to the Chinese Meteorological Administration (CMA). Management practices at each site, including fertilizer application and weed control, were generally the same as or better than the conventional practices used by local farmers. Plant protection management was undertaken to guarantee optimum growth and avoid weeds and pests. Phenological data included dates of planting, emergence, anthesis, and physiological maturity. Crop cultivar parameters were estimated using the DSSAT-GLUE package (He et al., 2009; Jones et al., 2011a, b) based on field observations of important phenology dates and grain yields of maize. However, several similar cultivars might be grown in different years at each site. In this study, the cultivar grown in the most years from 2006 to 2010 was selected as the representative cultivar for a given site (Table S1). Hence, eight sets of cultivar genetic parameters were estimated using the field-measured anthesis dates, maturity dates, and grain yield in the 2006–2009 growing seasons at each site. The field observations in the 2010 growing season were then used to verify these cultivar parameters at the corresponding sites.

2.4. Sensitivity of maize yield to weather variables in the Loess Plateau

The sensitivity of maize yield to four weather variables (R_s, T_{max}, T_{min}, and P) was evaluated based on simulation results from the CERES-Maize model at the eight sites in the Loess Plateau. First, the experimental files were set up according to the five-year average field conditions at each site (e.g., planting date, seeding density, fertilizer amount, and etc.). Next, mean values for the four daily weather variables were calculated from 1961 to 2010. Then, to evaluate the performance of each weather variable, daily values of the remaining three weather variables were represented by their mean values on the same date. In DSSAT, model runs will stop on the day when of the minimum temperature was greater than the maximum temperature. If minimum or maximum temperature was separately represented by its multi-year mean values, greater minimum temperature might be generated on

some days. Hence, variables T_{min} and T_{max} were treated as one weather factor and replaced simultaneously to avoid contradictions and model running errors. In this way, three kinds of multiple weather files were generated for the eight sites. Finally, these three kinds of weather files were used to run the CERES-Maize model to simulate maize yield from 1961 to 2010 at the eight sites in the Loess Plateau. The variations of simulated yields with different weather scenarios were compared to evaluate the sensitivities of maize yields to these weather variables in the Loess Plateau.

2.5. Yield prediction by merging of real-time weather data and historical weather records

Multi-year historical weather data were used to represent possible future weather scenarios in the target growing season. By incorporating the newly measured daily weather data and historical weather records, the weather series covering the entire growing season could be generated to drive the crop model for within-season yield prediction on each day of the growing season (Chen et al., 2020; Wang et al., 2017). In this study, the unknown weather after the prediction dates was represented by the same-period multiple historical weather records (1961–2010). A total 50 yield predictions were then generated on each day of the growing season. The average value of these 50 predictions was calculated as the final prediction on a given prediction date, and was defined as the original yield prediction in this study. However, this method was time-consuming for maize yield predictions using all of the 50-year weather data. Additionally, weather conditions in some faraway years might be very different from the target year due to the randomness of extreme meteorological events.

Thus, two alternative strategies were established for the selection of analogue years from multiple historical weather records (Fig. 2). The first strategy directly merged real-time weather data with weather data from different historical periods. We set five periods, including 40, 30, 20, 10, and 5 years ahead of the planting year. The second strategy merged the real-time weather data with the weather data of the analogue years. For the second strategy, we also designed two kinds of methods for the selection of analogue years based on the *k*-NN algorithm. The first method was developed based on the similarity of daily

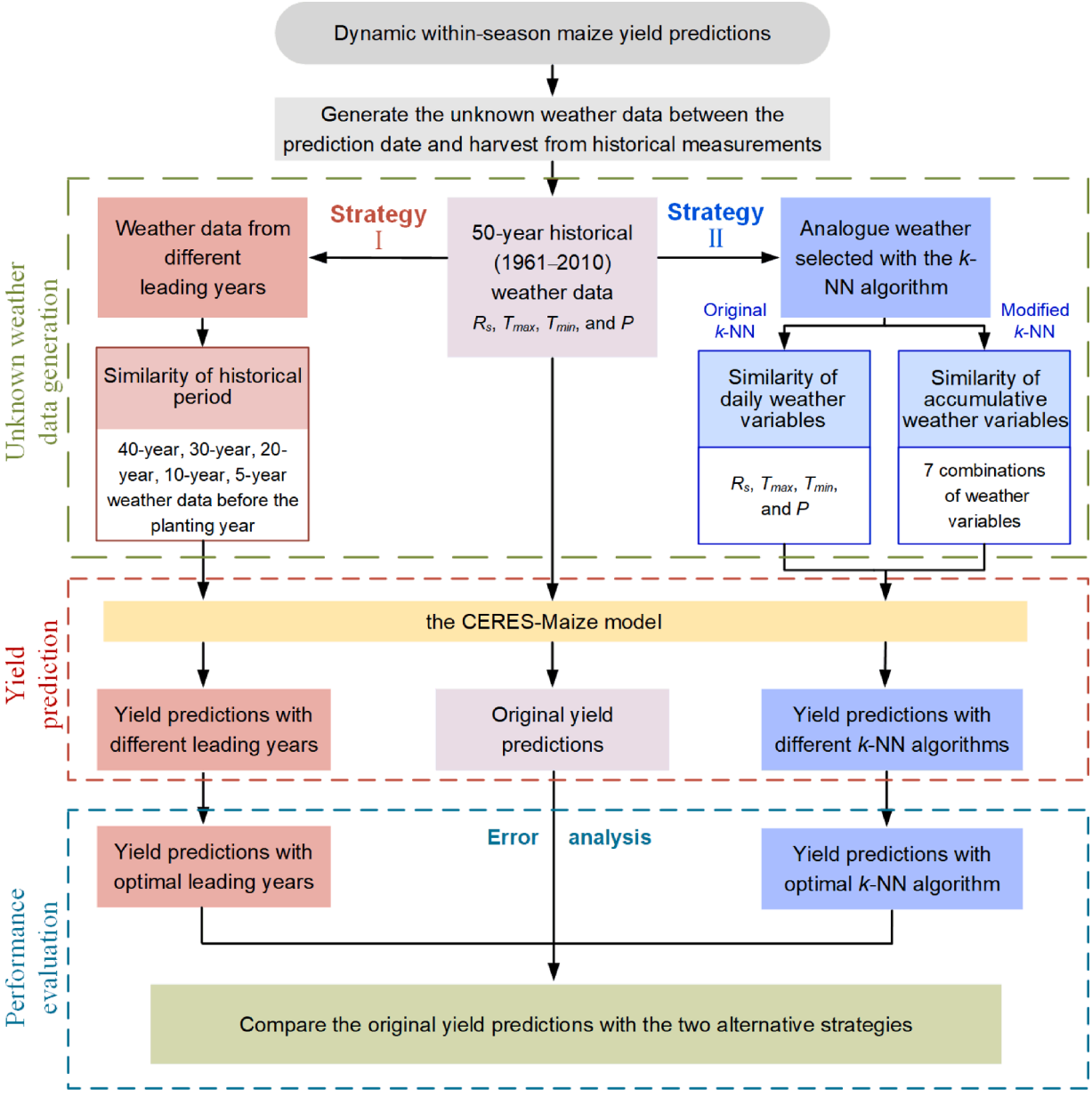


Fig. 2. Flowchart of dynamic within-season predictions of maize yields with different methods for the selections of analogue weather years.

values of R_s, T_{max}, T_{min}, and P. The second method was based on the similarity of different combinations of accumulative values of each weather variable. For daily maize yield predictions with the two strategies, the method in each strategy which obtained the smallest prediction error was recorded on every day. Moreover, the methods were compared for their frequency of smallest errors in maize yield predictions for 40 different growing seasons (eight sites × five years). Finally, the optimal methods with the smallest prediction errors were selected from the two alternative strategies. Yield predictions with the two selected optimal methods were then compared with the original predictions based on the entire 50-historical-year weather records.

2.5.1. Weather data from different leading years (Solution 1)

Wang et al. (2012) reported that there were obvious changes in climate in the Loess Plateau. Hence, we assumed that the weather differences may be larger with the increase of time interval between historical years and the target year. In this study, weather data from different leading years were selected to represent the unknown weather data after the prediction date. Five different lengths of leading times (40, 30, 20, 10, and 5 years) were used to select historical weather data. Taking the 40-leading-year weather data as an example, the unknown weather in the 2006 and 2010 growing seasons was represented by historical weather data in 1966–2005 and 1970–2009, respectively. Maize yields were then dynamically predicted with the five different ensembles of weather data at the eight sites in 2006–2010. The average yield prediction errors of the 40 different growing seasons were calculated to evaluate the prediction performance of these weather ensembles. Finally, the weather ensemble that produced the most effective and stable yield predictions was selected as the optimal leading-year weather.

2.5.2. Weather data from analogue years selected with the k-NN algorithm (Solution 2)

The *k*-NN approach was developed for pattern recognition based on the distance between the target feature vector and the source feature vectors. Salzberg et al. (1991) proposed different indicators to describe the difference between two instances, but the Euclidean distance has been widely used in corresponding studies. The detailed algorithm was reported by Gangopadhyay et al. (2005). To select an analogue-weather year, the pattern of observed weather variables on the *t*-th day was compared with the same variables on the same date in each historical year. The years with the smallest Euclidean distances were selected as the most similar years on *t*-th day. With the increased newly-measured weather data, the weather-analogue years could be selected for each day during the study period. Finally, the *k* years with the highest frequency of analogue were chosen as the most analogue-weather years for the study period. The original *k*-NN algorithm was developed based on the similarity of all measured weather variables on each day. For example, Chen et al. (2017) used the k-NN algorithm to selected analogue-weather years based on the similarities of daily values of four weather variables (T_{max}, T_{min}, R_s, and P). Then, the selected weather data were combined with the real-time measured weather data to generate the complete weather series, which covered the whole growing seasons, to predict maize yields. However, the processes of crop development are dependent primarily on heat accumulation or thermal time (Hodges and Evans, 1992). In addition, crop yields can be linked with phased total water application based on water production functions (Kipkorir et al., 2002; Zhang and Oweis, 1999). Hence, we modified the original *k*-NN algorithm to select analogue years based on the similarities of accumulative values of weather variables. The original and modified *k*-NN algorithms for analogue year selection are described in more detail below.

(1) Original *k*-NN algorithm based on the similarities of daily weather variables

In this algorithm, the Euclidean distance between the target-year weather vector and historical weather vector was computed with four normal weather variables (T_{max}, T_{min}, R_s, and P) on each day of the maize growing season (Eq. (1)).

$$ED_j = \sqrt{\sum_{n=1}^{4} \left(V_{tnj} - V_{hnj}\right)^2} \quad (1)$$

where ED_j is the Euclidean distance on *j*-th day; *n* is the index of weather variables; *t* and *h* are target year and historical year, respectively; V_{tnj} and V_{hng} are the *n*-th weather variable on *j*-th day in the target year and historical year, respectively. Each weather variable was normalized using the min-max normalization method (Eq. (2)) before ED calculation due to different units of these weather variables.

$$V_i = \frac{V_i - V_{min}}{V_{max} - V_{min}} \quad (2)$$

where V_i is the target weather variable on *i* th day; V_{max} and V_{min} are the maximum and minimum values of the related weather variable at the site in 1961–2010.

The year with the smallest ED value was selected as the analogue year for the given day. The analogue year was recorded for each day from the planting day to the prediction day. Then the frequency of each historical year selected as an analogue year was calculated. The *k* historical years with the highest frequency were finally selected as the analogue years for the target year. Finally, the unknown weather data after the prediction dates were represented by the weather data on the same dates in the selected *k* years. The *k* value was set as the same number of the optimal leading years with the highest prediction accuracy in Section 2.4.1.

(1) Modified *k*-NN algorithm based on the similarities of accumulative values of weather variables

In this algorithm, we modified the original *k*-NN algorithm by calculating ED with the accumulative values of weather variables from the planting day to the prediction day (Eq. (3)).

$$EDA_j = \sqrt{\sum_{n=1}^{3} \left(AV_{tnj} - AV_{hnj}\right)^2} \quad (3)$$

where EDA_j is the Euclidean distance of the accumulative weather variables from the planting day to the *j*-th prediction day; AV_{tnj} and AV_{tnj} are the accumulative values of the *n*-th weather variable from the planting day to the *j*-th prediction day in the target year *t* and a historical year *h*, respectively. Readers should be aware that there were only three accumulative weather variables (e.g., accumulative solar radiation, thermal time, and accumulative rainfall) because thermal time (TT) was calculated based on daily maximum and minimum temperatures (Eq. (4)).

$$TT = \frac{(T_{max} - T_{base}) + (T_{min} - T_{base})}{2} \quad (4)$$

where T_{max} and T_{min} are the maximum and minimum daily temperature; T_{base} is the base development temperature above which maize can grow. In this study, T_{base} was set to 8 °C according to Hodges and Evans (1992).

The analogue years selected with the original *k*-NN algorithm were based on the general ED of the daily weather variables (R_s, T_{max}, T_{min}, and P). We selected analogue years based on the ED values of three different combinations of accumulative values of weather variables as follows.

Combination I: One weather variable (three kinds). Analogue years were selected in each growing season individually based on the ED values of accumulative R_s, TT, and P.

Combination II: Two weather variables (three kinds). Analogue years were selected in each growing season based on the general ED values of the accumulative values of double weather variables of $R_s + TT$, $R_s + P$, and $TT + P$.

Combination III: Three weather variables (one kind). Analogue years were selected in each growing season based on the general ED values of accumulative values of $R_s + TT + P$.

Generally, there was only one kind of the original *k*-NN method developed with daily values of weather variables and seven kinds of modified *k*-NN methods developed with different combinations of accumulative values of weather variables. Maize yield predicted with these eight kinds of analogue weather were compared, and the most effective and robust method was selected as the optimal *k*-NN algorithm.

2.5.3. Evaluation of the two kinds of analogue weather selection strategies

Prediction accuracies of maize yield with the two kinds of analogue-weather selection strategies were compared with the original predictions using the entirety of local weather records at the eight sites in the Loess Plateau. Evaluation of these three strategies was conducted only in the 2010 growing season for the sake of brevity.

2.6. Statistical indices

The root mean square error (RMSE, Eq. (5)), absolute relative error (ARE, Eq. (6)), and coefficient of variation (CV, Eq. (7)) were used to evaluate model performance and yield prediction accuracy. The ARE and CV were calculated every day since maize grain yield was predicted daily within growing seasons.

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

$$\mathrm{ARE} = \frac{|O - S|}{O} \times 100\% \tag{6}$$

$$\mathrm{CV} = \frac{SD}{\overline{S}} \times 100\% \tag{7}$$

where S and O are the simulated and observed values of given variables; n is the number of total simulation times; SD and $\overline{S}$ are the standard error and the mean value of daily multi-yield predictions.

3. Results

3.1. Calibration and verification of the CERES-Maize model

Genetic parameters for the eight different maize cultivars were estimated and verified for the simulation of anthesis dates, maturity dates, and grain yields at the eight agro-meteorological observation sites (Fig. 3, Table S1). Generally, both phenology dates and grain yields were well simulated as indicated by all of the data points being close to the 1:1 line. For the simulations of phenology, the RMSE values were 3.5 and 3.7 d for anthesis dates, and 4.9 and 2.5 d for maturity dates for the calibration and verification processes, respectively. Additionally, the verification data points were closer to the 1:1 line than the calibration data points. It can be seen when comparing the calibration and verification datasets that the RMSE values decreased by 2.4 d for the simulations of maturity date and 245 kg ha^{-1} for yield, but slightly increased by 0.2 d for anthesis date. Overall, the results showed that growth and yield of all maize cultivars could be correctly simulated at the eight sites.

3.2. Sensitivity of maize yield to weather variables

Simulated yields responded well to variations of weather variables at the eight sites in the Loess Plateau (Fig. 4). Maize yields simulated with actual historic precipitation were much more dispersed than those with actual solar radiation or temperature. In addition, the average yields simulated with different ensembles of weather variables were less than 7000 kg ha^{-1} at the Xinzhou and Yuncheng in Shanxi Province (Fig. 4c and d). Compared with precipitation, solar radiation and temperature generated smaller variations in yield simulations in the Loess Plateau. Generally, precipitation was the main determinant for maize yield in most areas of the Loess Plateau.

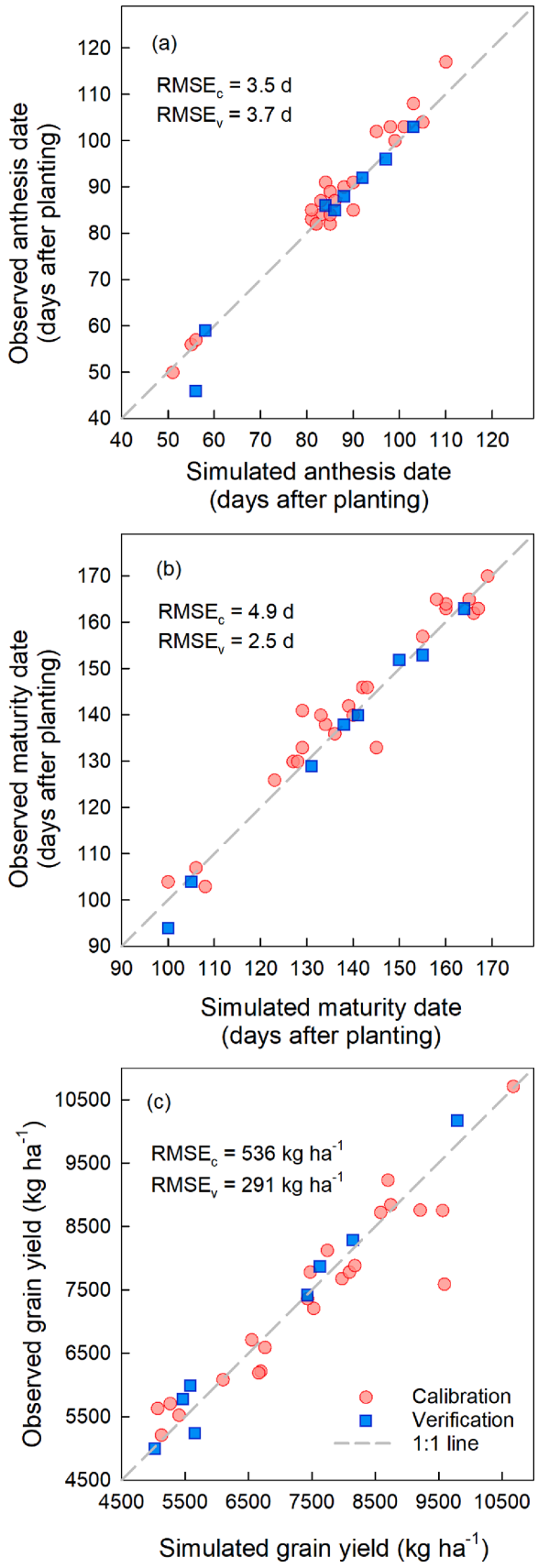
3.3. Dynamic within-season predictions of rainfed maize yield

Maize growth at the Yulin site in Shaanxi Province in the 2010 growing season was taken as an example to show the yield prediction based on the fusion of 50-year historical records with newly measured weather data from a local weather station (Fig. 5). In general, the yield predictions were widely scattered during the early part of the growing season and converged to the actual yield in the later part of the growing season. The prediction uncertainty did not decrease immediately with the introduction of actual weather data in the synthesized weather series. For example, at 10, 30, and 60 days after planting, the predicted yield ranges were 2274–12,633, 3118–12,530, and 2098–12,289 kg ha^{-1}. In contrast, the predicted yields converged rapidly to the actual yield value and remained relatively stable after maize tasseling. For example, the forecasted yield ranges were 3861–12,368, 5994–10,606, and 6546–8139 kg ha^{-1} at 70, 90 and 110 days after planting, respectively. The mean value of the coefficient of variation (CV) of daily predictions was 20.2% from planting to tasseling and 5.7% from tasseling to harvest. Additionally, errors in daily yield prediction showed a similar trend as the prediction uncertainty. The mean value of ARE of daily predictions was 23.8% from planting to tasseling and 6.6% from tasseling to harvest. Both CV and ARE were less than 15% at 85 days after planting (about 50 d before harvest). The final ARE value was 4.0% on the harvest day. In addition, yield predictions were not equally distributed on each day with the most of the values falling in the concentrated range of 7000–11,000 kg ha^{-1}. The concentrated range was 9500–11,000 kg ha^{-1} before tasseling and 7000–8000 kg ha^{-1} after tasseling. Similar results were also obtained for the other seven sites (Fig. S1).

3.4. Yield prediction with weather data from different lengths of leading years

Weather data from five different lengths of leading years were used to predict maize yield at the Yuling site in Shaanxi Province in 2010 (Fig. 6). In general, yield predictions with different weather ensembles varied during the early growing season. Similar yields were predicted with weather data from 40 to 30 years before planting. For the yield predictions made at 50 days after planting, the smallest errors were obtained by the 5-leading-year weather data. However, the prediction errors with this group of weather data increased with time, and ultimately produced the worst yield prediction at 65 days after planting. The results indicated that great uncertainty in yield prediction would be generated with the weather data from five years before the planting year. This result was mainly a product of the limited weather types contained in the five-year weather data. In addition, the differences among yield predictions using weather data from different lengths of leading years declined with the advancement of the growing season, and almost all predictions converged to the same value after maize tasseling.

Yield predictions simulated with weather data from different lengths of leading years were then obtained for all eight sites in 2006–2010 (Figs. 7 and S2). For the simulations using weather data of leading years greater than five years, the maximum values and the 95th-percentile of ARE values were decreased, but the median values were increased when the leading years became closer to the planting year. The smallest median ARE value (11.1%) was produced by the simulations using the 40-leading-year weather data. The smallest 95th-percentiles, 75th-percentiles, and mean ARE values were achieved with simulations using the 5-



(caption on next column)

Fig. 3. Observed and simulated anthesis dates (a), maturity dates (b), and grain yields (c) for the calibration and verification datasets at eight agro-meteorological stations in the Loess Plateau in the 2006–2010 growing seasons. The red-filled circles and blue-filled squares show the calibration and validation datasets, respectively. The gray dashed line is the 1:1 line. RMSE$_c$ and RMSE$_v$ represent the root mean square errors (RMSE) in the calibration and validation datasets, respectively (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

leading-year weather data. Nevertheless, more extreme values were also predicted with 5-leading-year weather data since the largest range of ARE (0–71.2%) was found for the predictions with this weather ensemble, indicating great uncertainty in maize yield predictions with limited weather types (Fig. 7).

The weather data from five years before planting provided the most days (a total of 1566 d) with the minimum ARE in daily yield predictions, followed by the 10-, 40-, 20-, and 30-leading-year weather data (Table 1). Generally, yield predictions with the 30- and 20-leading-year weather data produced similar numbers of days with the minimum prediction ARE. The 40-, 30-, and 20-leading-years weather data separately produced zero days with the minimum prediction ARE in at least four growing seasons (gray-shaded cells in Table 1). In the contrast, the 10- and 5-leading-year weather data produced zero days with the minimum prediction ARE in only one growing season. Hence, both the 10- and 5-leading-year weather data before planting produced more reliable yield predictions. Taking into account the extreme ARE values generated by the 5-leading-year weather data, weather data from 10 leading years before the planting year was selected as the optimal weather ensemble in this study. Furthermore, the *k* value in the *k*-NN algorithm was also set as 10 to select the analogue wheat years for future comparisons between the two alternative strategies.

3.5. *Comparisons of yield predictions with analogue years selected with different k-NN algorithms*

The 10 analogue years selected with different *k*-NN algorithms generated different daily yield prediction errors (Fig. 8). Compared with the analogue years selected with the modified *k*-NN algorithm, the analogue years selected with the original *k*-NN algorithm based on the similarity of daily weather variables produced the largest variation (0–66.3%) and average value (12.3%) of the ARE. Generally, there was no obvious difference among the prediction errors based on the analogue weather selected with the modified *k*-NN algorithm using accumulative weather variables. The average ARE values ranged from 11.5% to 11.8%. For the modified *k*-NN algorithm with single accumulative weather variable (Combination I), the yield predictions with the analogue a weather selected with accumulative precipitation (P) produced the smallest range and average value (11.5%) of ARE. The ARE values were 11.7% and 11.8% for the yield predictions with the analogue selected with the single accumulative solar radiation R_s and thermal time T, respectively. For the combinations with two accumulative weather variables (Combination II), the prediction errors of single R_s and T did not decrease even after combining with the P variable, since the yield prediction errors based on accumulative $R_s + P$ (average ARE of 11.7%) and $T + P$ (average ARE of 11.8%) remained essentially the same (Fig. 7). In contrast, the combination of $R_s + T$ reduced the yield prediction ARE by 0.1% compared with the single R_s or T. For the combination with three weather variables (Combination III), the inclusion of the P variable did not reduce the yield prediction errors since the same ARE value (11.6%) was obtained by both combinations of $R_s + T$ and $R_s + T + P$.

The weather data of analogue years selected based only on accumulative precipitation produced the most days (a total of 1113 d) with the minimum ARE in daily yield predictions (Table 2). However, the number of days with the minimum ARE value decreased with the increasing number of weather variables used in the selection of analogue

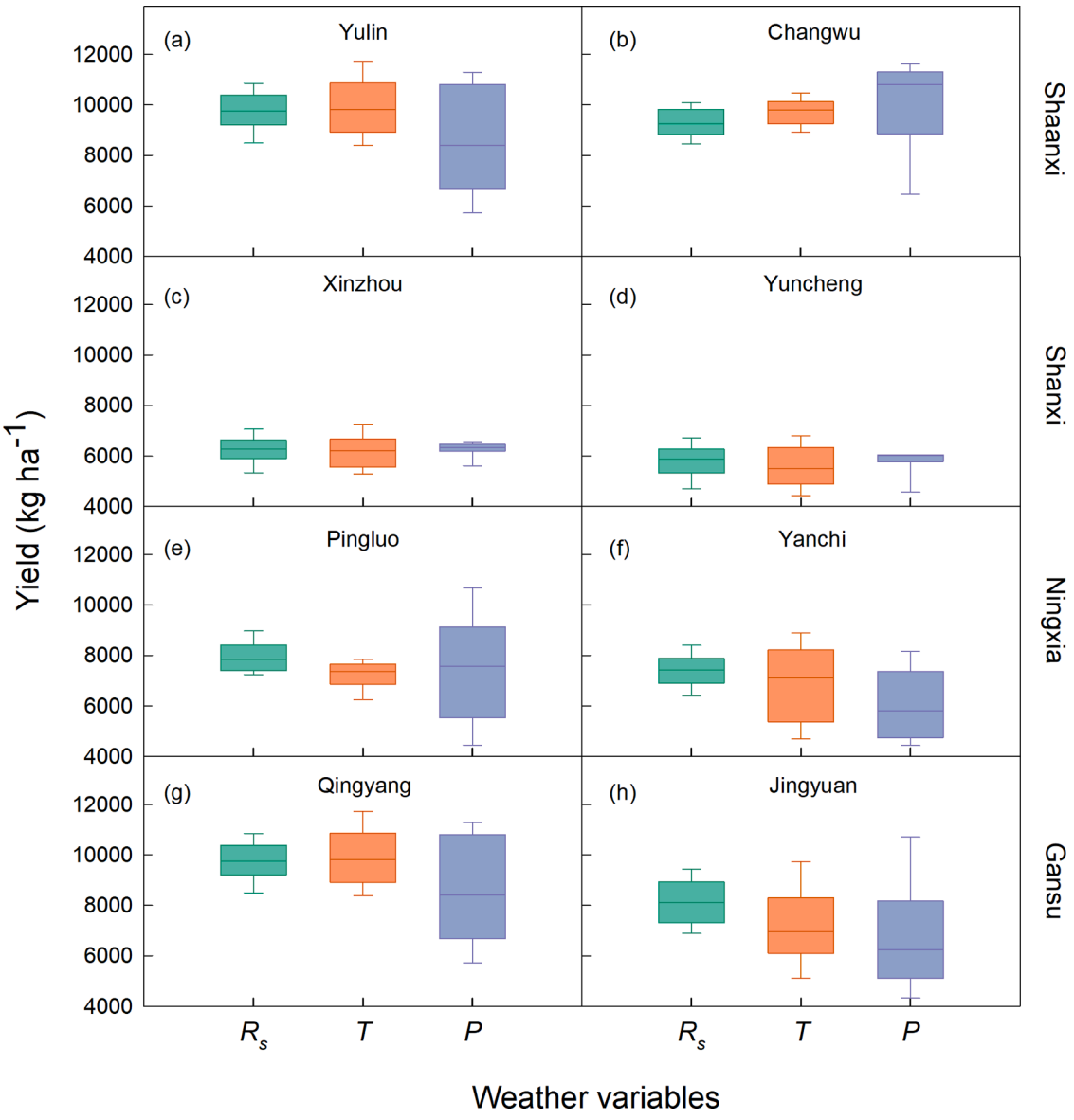


Fig. 4. Variations of maize yields simulated with three different weather variables (R_s, T, and P) at eight agro-meteorological sites (a-h) in the Loess Plateau in 1961–2010. When simulating maize yield response to a given weather variable, the remaining weather variables were replaced by their corresponding multiple mean values. The maximum and minimum temperatures were considered as one variable (T) and replaced together by their means to avoid contradictions. Medians are the horizontal lines within the boxes. The upper and lower edges of the boxes represent the 75th and 25th percentiles, and the whiskers are the 95th and 5th percentiles. The same explanation applies to subsequent box plots.

weather. The numbers of days with the minimum-ARE were greater than 800 d for the single weather variable, 200–600 d for the combinations of two weather variables, and only 135 d for the combination of three weather variables. In addition, the weather variable combinations of $R_s + P$ and $R_s + P + T$ produced zero days with the minimum ARE in daily yield predictions in six and nine different growing seasons (gray-shaded cells in Table 2), respectively. Yield predictions with the original k-NN algorithm produced the second largest number of days with minimum ARE in daily predictions (Table 2), although the original k-NN algorithm generated the largest ARE range and average value (Fig. 8). Generally, the k-NN algorithm based on accumulative precipitation was the most effective and reliable for selecting analogue years to predict maize yield in this study. This finding was consistent with the sensitivity analysis of maize yield to weather variables (Fig. 4) since maize yield was mainly affected by the seasonal precipitation in the Loess Plateau. Finally, the modified k-NN algorithm based on accumulative precipitation was selected as the optimal k-NN strategy ($k = 10$) to select the 10 analogue weather years.

3.6. Comparisons between yield predictions with the two strategies of analogue weather selection

With the weather data separately from the 10-leading historical years before planting (identified as 10 leading years for brevity) and the 10 analogue years selected with the modified k-NN algorithm based on accumulative precipitation (identified as the modified k-NN algorithm for brevity), the corresponding maize yield predictions were then compared with the original yield predictions with 50 year historical weather records (Fig. 9). For maize yield predictions with the 10-leading-year weather data, average ARE values were smaller than the original predictions at six of the sites, only except for Yuncheng in Shanxi Province (Fig. 9d) and Qingyang in Gansu Province (Fig. 9g). Especially, the prediction accuracy based on this weather ensemble was most greatly improved at the Yulin site in Shaanxi Province since the ARE

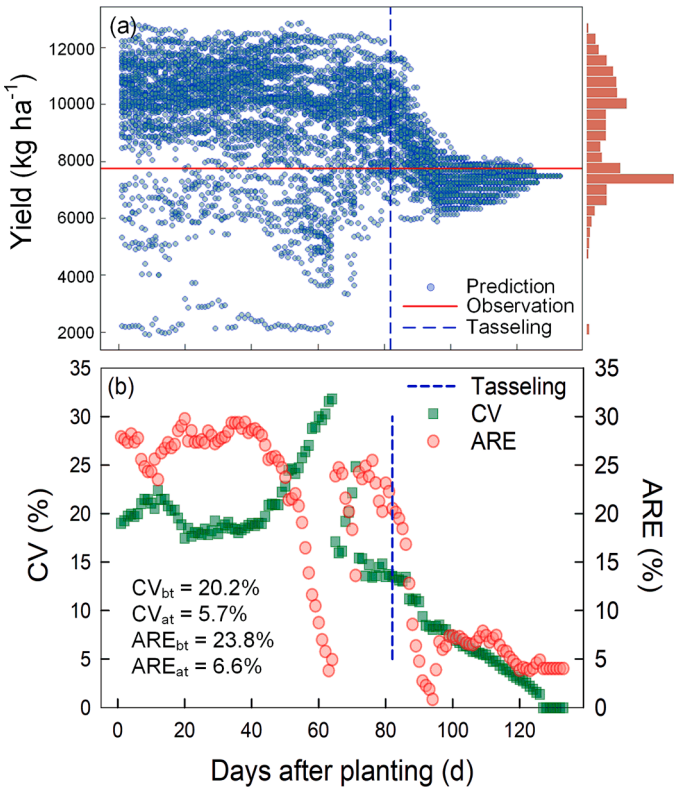


Fig. 5. Dynamic within-season predictions of maize yield (a) and the prediction errors (b) at the Yulin site in Shaanxi Province during the 2010 growing season. The red solid line and blue dashed line represent the observed yield and tasseling date, respectively. The blue dots, green-filled squares, and red-filled circles indicate daily predicted yield, CV (coefficient of variation) values, and ARE (absolute relative error) values. The subscripts of '*bt*' and '*at*' represent the CV and ARE values before tasseling and after tasseling stage of maize, respectively (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

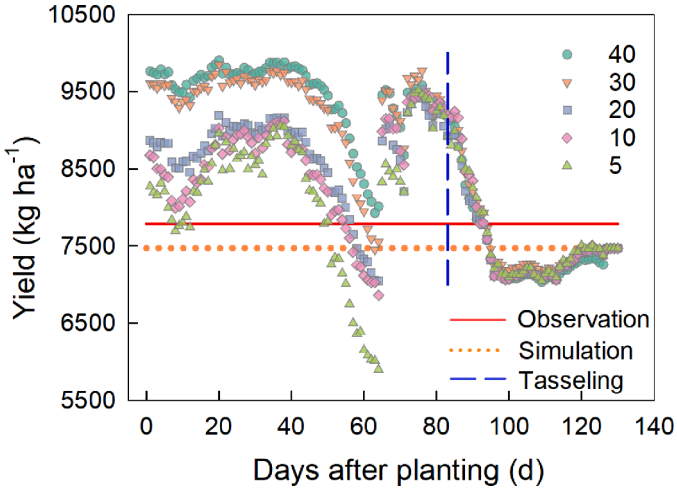


Fig. 6. Dynamic within-season predictions of maize yield at the Yulin site in Shaanxi Province in 2010 with weather data from different lengths of leading years. The red solid line and the orange dotted line represent the observed and simulated maize yields, respectively. The blue dashed line indicates the maize tasseling date. The circles, down-triangles, squares, diamonds, and up-triangles show the maize yield predictions with the weather data from 40, 30, 20, 10, and 5 leading years, respectively, before the planting year (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

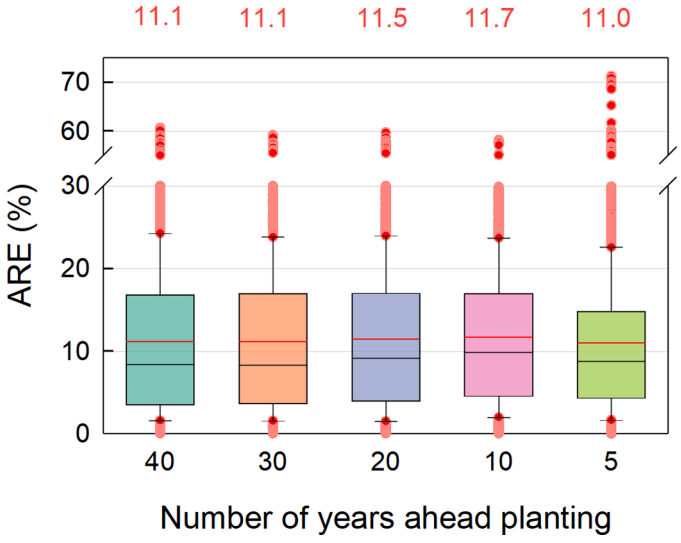


Fig. 7. Distributions of the absolute relative error (ARE) of daily maize yield predictions with weather data from different lengths of leading years at eight sites in the Loess Plateau in 2006–2010. The red-marked values above boxplots show the average values (red line within boxplot). The elements of the boxplots are the same as described in Fig. 4 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

value decreased by 8.4% (Fig. 9a). In addition, large differences were found before maize tasseling between the yields predicted with the entire set of 50 historical years of weather records and the 10-leading-year weather data. For maize yield predictions with the modified *k*-NN algorithm, six sites showed smaller average ARE than the original yield predictions. However, worse yield predictions were obtained at the Yulin site in Shaanxi Province (Fig. 9a) and at the Yanchi site in Ningxia Province (Fig. 9f). Furthermore, the prediction accuracy based on weather data from analogue years was most greatly improved at the Jingyuan site in Gansu Province since the ARE value decreased by 2.5% (Fig. 9h). Generally, the two strategies of analogue weather selection all improved yield prediction accuracies since the ARE values decreased at most sites in the Loess Plateau in 2010. However, maize yield predictions with the 10-leading-year weather data produced smaller ARE values than the predictions with the modified *k*-NN algorithm at five sites. Moreover, larger prediction errors based on the modified *k*-NN algorithm occurred mainly in the early part of the growing seasons (Fig. 9a, c, d, f, and h). Finally, there were no obvious differences among the yield predictions with different weather ensembles after maize tasseling stage.

4. Discussion

4.1. Within-season yield predictions with dynamically merged weather data

In this study, the CERES-Maize model was used in dynamic within-season maize yield predictions at eight agro-meteorological observation sites in the Loess Plateau. Generally, good performance was provided at all of these sites in both model calibration and validation stages in most years. It was noteworthy that one set of cultivar genetic parameters was used to represent the similar cultivars sown at each site since different cultivars were planted in the five-year experiments. Hence, the data point with large simulation errors might result from the variation of sown cultivar.

Weather data series covering the entire growing season is essential for within-season crop yield predictions with crop models. Daily real-time meteorological variables have become readily accessible with the popularity of low-budget, micro-weather stations. Hence, attention has

Table 1
Frequency of days with the minimum prediction ARE (absolute relative error) in daily maize yield predictions based on weather data from different lengths of leading years before planting year at eight sites in the Loess Plateau in 2006–2010. The gray-shaded table cells show zero days with the minimum prediction ARE produced by the corresponding weather data in the growing season.

Province	Site	Year	Prediction days (d) [a]	Leading years before the planting year				
				40	30	20	10	5
Shaanxi	Yulin	2006	120	0	0	3	109	8
		2007	124	0	2	39	71	12
		2008	125	38	15	24	36	12
		2009	140	78	12	14	27	9
		2010	126	31	5	9	41	40
	Changwu	2006	124	23	22	40	28	11
		2007	124	16	6	0	15	87
		2008	131	29	23	7	27	45
		2009	140	39	5	10	59	27
		2010	135	47	0	1	60	27
Shanxi	Xinzhou	2006	131	8	3	9	68	43
		2007	136	15	72	10	4	35
		2008	140	4	3	6	56	71
		2009	129	38	26	36	12	17
		2010	131	9	13	12	20	77
	Yuncheng	2006	96	17	33	46	0	0
		2007	103	41	34	13	1	14
		2008	102	11	31	6	6	48
		2009	103	7	22	21	11	42
		2010	94	14	40	0	1	39
Ningxia	Pingluo	2006	156	33	10	20	67	26
		2007	161	34	32	40	15	40
		2008	167	88	30	3	35	11
		2009	154	100	13	21	11	9
		2010	161	5	6	9	133	8
	Yanchi	2006	145	0	0	2	57	86
		2007	156	8	1	4	81	62
		2008	156	77	24	47	2	6
		2009	152	29	5	93	3	22
		2010	145	11	19	9	43	63
Gansu	Qingyang	2006	126	33	6	0	50	37
		2007	133	11	0	21	56	45
		2008	133	4	57	18	31	23
		2009	140	17	15	21	27	60
		2010	135	25	38	19	5	48
	Jingyuan	2006	144	16	2	0	16	110
		2007	155	31	36	19	16	53
		2008	155	55	40	52	3	5
		2009	163	17	7	20	43	76
		2010	152	0	1	15	24	112
	Total		5443	1059	709	739	1370	1566

Note:
[a]Since different weather ensembles produced almost the same yield predictions in the later part of the maize growing seasons (after tasseling stage), only the days with different yield predictions were selected for the comparisons.

been focused on the generation of reliable weather data between the prediction date and the harvest date. Previous studies have used seasonal weather forecasts (e.g., weather generators, GCMs/RCMs) to represent the unknown future weather (Bakker et al., 2013; de Wit et al., 2010; de Wit and van Diepen, 2007; Mavromatis, 2016). However, crop yield forecasts based on historical weather data tend to be more effective than seasonal weather data predictions (Prakash et al., 2019) because they avoid the errors associated with simplified mechanisms and downscaling processes seen with the use of climate models (Goel and Dash, 2007). In this study, site-specific multi-year weather data were joined with daily measured weather data to dynamically predict maize yields at eight different sites in the Loess Plateau of China in the 2006–2010 growing seasons. The results showed that errors in daily yield predictions, represented by ARE values, were less than 10% at about 50 d before maturity at most sites (Fig. S1), and therefore this method could provide valuable time for decision making with regard to field management practices.

The formation of maize grain yield is determined by the genotype × environment × management interactions. Final maize yield in the Loess Plateau was mainly determined by weather conditions, especially precipitation since no supplemental irrigation was applied during the growing season. Hence, the uncertainty in dynamic within-season predictions of maize yield was assumed to decrease with the increase of actual weather data in the weather data series. However, predicted yields were widely scattered during the early part of the growing seasons until the date of maize tasseling (Figs. 4 and 5). These results are similar to yield predictions for wheat in Europe and New Zealand conducted by Lawless and Semenov (2005). Chen et al. (2020) also reported that

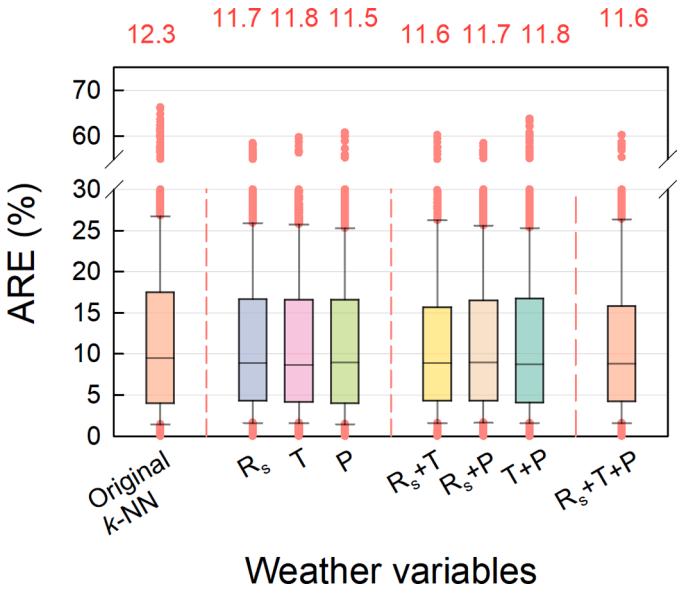


Fig. 8. Absolute relative errors (ARE) of daily maize yield predictions with the weather data of analogue years selected with the original *k*-NN algorithm and the modified *k*-NN algorithm based on different combinations of accumulative weather variables at eight sites in the Loess Plateau in 2006–2010. The red-marked values above boxplots show the average values (red line within boxplot). The x-axis label *Original k-NN* represents the original *k*-NN algorithm that was used to select analogue-weather years based on the similarity of daily weather variables. Labels R_s, *T*, and *P* represent weather variables of solar radiation, temperature, and precipitation, respectively. The elements of boxplots are the same as described in Fig. 4 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

uncertainty in maize yield prediction declined after maize tasseling mainly because of the fully developed canopy established at later vegetative growth that determined the maximum radiation interception capacity. However, the great errors and uncertainties in early part of growing seasons need to be further reduced to improve within-season yield prediction by crop models.

4.2. Performance of the weather data from different lengths of leading years

Obvious changes of meteorological variables have been observed in the Loess Plateau of China over time. The warmer climate has also resulted in obvious changes in crop phenology in this region (He et al., 2015). Wang et al. (2012) reported that annual mean temperature increased by 1.9 °C while annual mean precipitation reduced by 29 mm in the Loess Plateau. Thus, there might be less difference between the weather conditions in the planting year and the several leading years before it. In this study, weather data from five different lengths of leading years (40, 30, 20, 10, and 5 years) were joined with real-time weather measurements to drive the CERES-Maize model to dynamically forecast maize yield in the Loess Plateau. When the leading years became closer to the specified year of interest, yield prediction errors decreased with the corresponding weather data. Hence, the smallest average ARE value was obtained when simulations were done with weather data from five leading years before planting. However, the largest range of ARE was also generated in daily yield predictions with this weather ensemble. This was mainly due to the limited weather patterns contained in the five years. Considering the effectiveness and reliability of yield prediction, we selected the 10-leading years as the optimal leading-year strategy for weather data fusion with current weather for within-season yield predictions. Wang et al. (2017) used the same method to predict cotton yield in Xinjiang in China. They also found that the smallest prediction errors were obtained with the weather data of 10 leading years ahead of planting. However, they only conducted predictions once a month during the 2017 growing season at a single site. The findings in this study provide more reliable evidence for the validity of using weather data from the 10 leading years before planting for within-season yield predictions.

4.3. Performance of the analogue weather selected with the k-NN algorithms

The *k*-NN algorithm has been widely used for selecting analogue weather data due to its simple calculation process for Euclidean distance (ED) (Buishand and Brandsma, 2001; Rajagopalan and Lall, 1999). Bannayan and Hoogenboom (2008a) used the *k*-NN algorithm to develop a tool for daily weather prediction and evaluated it at 16 different sites. They reported that the *k*-NN algorithm was promising for the prediction of solar radiation, maximum and minimum temperatures. However, poor performance was found for precipitation prediction. Additionally, the same amount of precipitation on two different adjacent days might cause large ED in analogue year selections, but had very similar effects on crop growth. Thus, years with temporally close but different distributions of precipitation might be grouped into different weather patterns. However, precipitation events on different but close dates might generate very slight difference in crop growth. In this way, yield predictions with weather data of analogue years selected with the original *k*-NN algorithm could generate large prediction errors (Fig. 8). Chen et al. (2017) modified the original *k*-NN algorithm by using the seven-day moving-average values to represent the daily values of each weather variable (i.e., R_s, T_{max}, T_{min}, and *P*) in the calculation of daily ED. Compared with the original *k*-NN algorithm, the modified *k*-NN algorithm produced smaller prediction errors on most of the prediction dates at three sites in Shaanxi Province in northwest China. However, the modification was time-consuming because more processes were introduced into the selection of analogue years. In this study, we further modified the original *k*-NN algorithm by calculating ED with the accumulative values of relevant weather variables from the planting date to the prediction date. Analogue years were selected based on a single ED value rather than a series of daily ED values, thereby dramatically reducing the prediction errors and consumption time.

In the modified *k*-NN algorithm based on different combinations of accumulative values of weather variables, prediction errors were the smallest with the analogue years selected based only on accumulative precipitation. Yield prediction errors were not reduced by introducing other weather variables in the process of analogue year selection. The results of sensitivity analysis demonstrated that precipitation was the greatest determining meteorological factor influencing maize yield in the Loess Plateau. Greater weight should be given to precipitation in the calculation of ED to improve yield prediction accuracy considered the importance of precipitation in rainfed yield formation. In addition, the *k*-NN method relies on the assumption that the actual weather data observed in the target year could be a replication of weather recorded in the past (Bannayan and Hoogenboom, 2008b). However, this assumption could be invalid in growing seasons with extreme weather events.

4.4. Comparisons between yield predictions with weather data from entire historical records and analogue years

Compared with the original yield predictions obtained using the entire set of historical weather records, both the weather data from 10 leading years before planting and the 10 analogue years selected with the modified *k* -NN algorithm could dramatically reduce the errors and time consumption in within-season predictions of rainfed maize yield in the Loess Plateau. However, we also found that it was more convenient to use the weather data of 10 leading years to represent the unknown weather data after the prediction dates. This strategy provided reliable prediction accuracy without complex programming and requirement for long-term weather records. However, the strategy of using weather data

Table 2
Frequency of days with the minimum prediction ARE based on the weather data from analogue years selected with the modified *k*-nearest neighbor (*k*-NN) algorithm based on different combinations of weather variables at eight sites in the Loess Plateau in 2006–2010. The *k* value in the *k*-NN algorithm was set as 10. The gray-shaded table cells show zero days with the minimum prediction ARE produced by the corresponding weather data in the growing season. Label symbols R_s, T, and P represent weather variables of the accumulative global solar radiation, temperature, and precipitation, respectively.

Province	Site	Year	Prediction days (d) [a]	Combinations of weather variables in the modified *k*-NN [b]							Original *k*-NN [c]
				R_s	T	P	R_s +T	R_s +P	T+P	R_s +T+P	
Shaanxi	Yulin	2006	120	26	22	20	23	15	6	5	3
		2007	124	19	35	16	17	3	23	6	5
		2008	125	16	10	30	19	1	15	3	31
		2009	140	25	31	18	19	5	14	8	20
		2010	126	24	9	28	4	16	4	3	38
	Changwu	2006	124	25	33	15	19	3	7	4	18
		2007	124	20	1	16	5	0	10	1	71
		2008	131	27	17	25	28	1	16	2	15
		2009	140	22	16	30	15	2	13	2	40
		2010	135	2	9	77	3	1	18	0	25
Shanxi	Xinzhou	2006	131	22	52	14	2	0	23	0	18
		2007	136	26	3	80	4	8	6	0	9
		2008	140	10	40	17	16	3	8	6	40
		2009	129	31	7	35	22	12	7	4	11
		2010	96	5	5	34	3	5	21	6	17
	Yuncheng	2006	96	9	5	29	6	1	13	1	32
		2007	103	14	13	42	1	0	5	1	27
		2008	102	18	19	13	2	0	12	3	35
		2009	103	38	5	14	12	6	10	3	15
		2010	94	18	24	19	15	0	5	2	11
Ningxia	Pingluo	2006	156	20	23	25	41	1	2	4	40
		2007	161	23	63	30	16	1	5	0	23
		2008	167	53	33	11	22	11	3	2	32
		2009	154	37	26	26	39	11	5	3	7
		2010	161	22	19	52	30	14	7	2	15
	Yanchi	2006	145	17	30	47	10	3	13	1	24
		2007	156	16	11	22	21	1	9	0	76
		2008	156	23	23	15	12	13	33	5	32
		2009	152	16	32	26	9	14	15	3	37
		2010	145	22	20	23	13	5	15	1	46
Gansu	Qingyang	2006	126	18	36	25	12	1	18	0	16
		2007	133	10	43	15	23	1	23	1	17
		2008	133	23	15	38	15	7	8	0	27
		2009	140	39	21	10	15	0	25	0	30
		2010	135	11	27	43	3	1	12	0	38
	Jingyuan	2006	144	6	19	34	22	8	26	18	11
		2007	155	13	43	26	18	6	16	9	24
		2008	155	31	14	24	27	10	15	8	26
		2009	163	19	33	29	10	10	16	13	33
		2010	152	11	45	20	6	3	50	5	12
	Total		5408	827	932	1113	599	203	552	135	1047

Notes:
[a]Because different weather ensembles produced almost the same yield predictions in the later part of the growing seasons (after maize tasseling stage), only the days with different yield predictions were selected for the comparisons.
[b]The modified *k*-NN algorithm selected the analogue years weather based on seven different combinations of three accumulative weather variables (e.g., accumulative solar radiation, accumulative precipitation, and thermal time).
[c]The original *k*-NN algorithm selected the analogue years based on daily values of four weather variables (e.g., maximum temperature, minimum temperature, solar radiation, and precipitation).

from an optimal number of leading years assumes that there have been no obvious or major changes in climate conditions over time, and that assumption could be invalid in years with extreme weather events. For maize yield predictions with crop models using weather data from analogue years selected with the modified *k*-NN strategy based on accumulative precipitation, worse predictions were mainly generated during the early part of maize growing seasons. This was mainly due to the limited weather observations that could be used to select analogue weather patterns in this period. As time advanced into the growing season, the *k*-NN algorithm provided more reliable predictions. In recent studies, the GCMs/RCMs were also used for crop yield predictions (Jha et al., 2019; Prakash et al., 2019). In future investigations, we want to

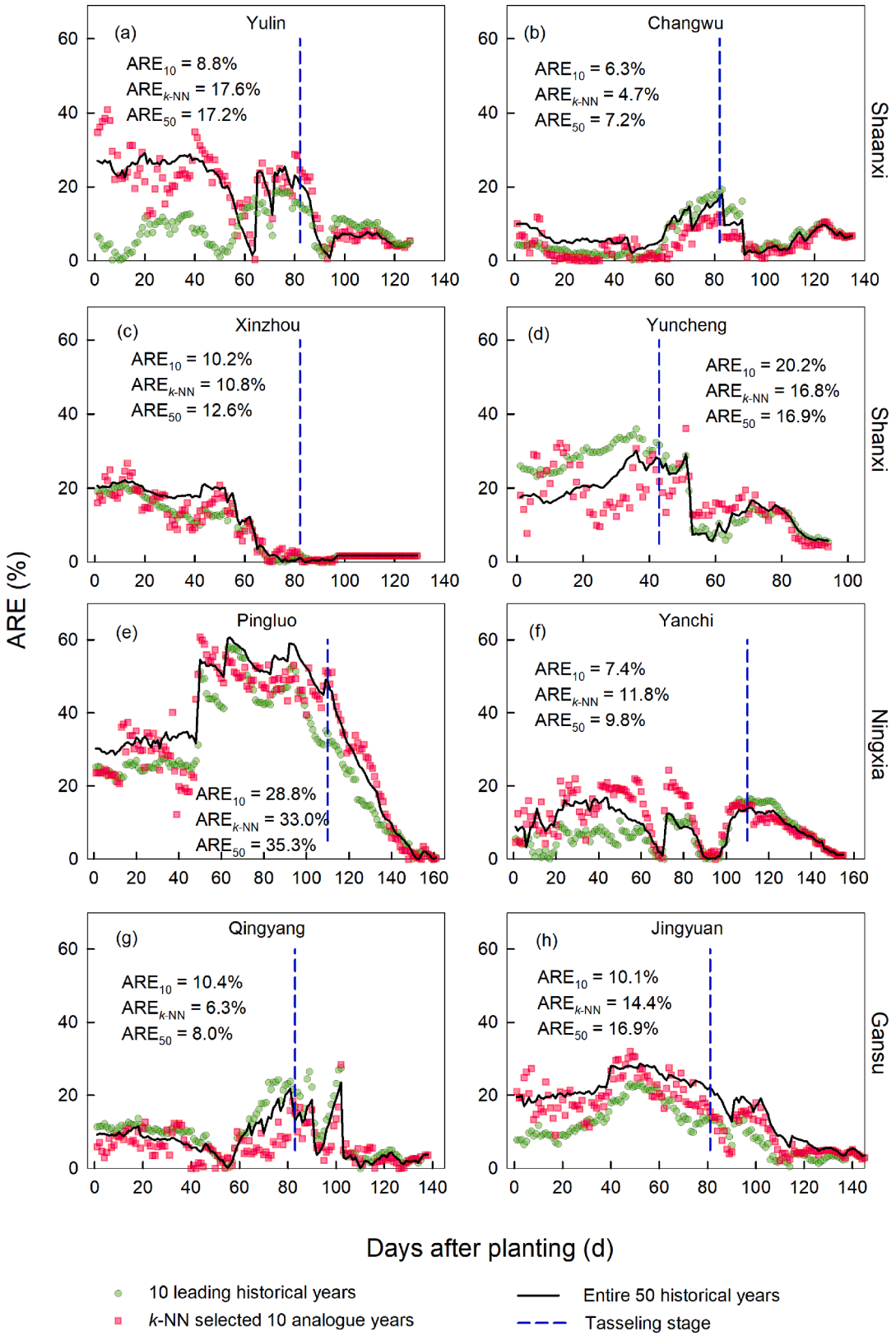


Fig. 9. Absolute relative error (ARE) values of daily maize yield predictions based on two alternative strategies of analogue weather selection and the original predictions at eight sites (a-h) in the Loess Plateau in the 2010 growing season. The green-filled circles show the ARE for yields predicted with the weather data of 10 leading years before the planting year (ARE_{10}). The red-filled squares show the ARE for yields predicted with the weather data of 10 analogue years selected with the modified *k*-NN algorithm based on accumulative precipitation ($ARE_{k\text{-}NN}$). The black solid lines show the ARE for the original predictions with the entire 50 years of historical weather data (ARE_{50}). The blue dashed lines show the dates of maize tasseling stage (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

employ short-term weather forecasts to generate seasonal weather series in a "real-time measurements + short-term prediction + historical records" strategy. In this way, the timeliness and accuracy of within-season yield predictions could be expected to improve further.

5. Conclusions

In most yield prediction studies using crop models, the quality of generated unknown weather data plays an important role in improving or worsening the prediction accuracy. In this study, two alternative solutions were provided for the generation of unknown future weather data after a given yield prediction date. The first solution used weather data from different lengths of leading historical years before the planting year. The second solution used weather data from the analogue years selected with the modified *k*-nearest neighbor (*k*-NN) algorithm based on eight different combinations of accumulative values of four weather variables. For within-season maize yield predictions, lager uncertainties and errors appeared in the early part of maize growing season but decreased obviously after maize tasseling stage (about 50 d before maturity date). Compared with the predictions using the entire 50-year weather data, maize yield predicted with the weather data from 10 leading years before planting provided higher estimation accuracy, with an average ARE of 11.7%. Using the analogue years selected with the modified *k*-NN algorithm based on accumulative values of weather variables produced smaller average ARE values than the original *k*-NN algorithm based on the similarity of daily values of the weather variables involved. The analogue years selected based on accumulative precipitation produced the smallest prediction error, with an average ARE of 11.5%. Both of the two alternative solutions reduced prediction errors and time consumption compared to the original yield predictions using the entire weather records. However, the modified *k*-NN algorithm was more likely to generate worse predictions in the early growing seasons due to the limited weather data used for analogue year selection. Generally, the weather data from 10 leading historical years before the planting year might be a more user-friendly method since it was less complicated and computation time saving.

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.

Acknowledgements

This research was supported by the Natural Science Foundation of China (No. 41961124006, 52079115, 41705095), the Key Research and Development Program of Shaanxi (No. 2019ZDLNY07–03), the Open Project Fund from the Key Laboratory of Eco-Environment and Meteorology for the Qinling Mountains and Loess Plateau, Shaanxi Provincial Meteorological Bureau (No. 2019Z-5), and the "111 Project" (No. B12007) of China. We are also grateful to the editor and anonymous reviewers whose comments and suggestions have greatly improved this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.agrformet.2022.108810.

References

Angstrom, A., 1924. Solar and terrestrial radiation. Report to the international commission for solar research on actinometric nvestigations of solar and atmospheric radiation. Q. J. R. Meteorol. Soc. 50 (210), 121–126.

Baigorria, G.A., Hansen, J.W., Ward, N., Jones, J.W., O'Brien, J.J., 2008. Regional atmospheric circulation and surface temperatures predicting cotton yields in the Southeastern USA. J. Appl. Meteorol. Climatol. 47, 76–91.

Bakker, A.M.R., Bessembinder, J.J.E., de Wit, A.J.W., van den Hurk, B.J.J.M., Hoek, S.B., 2013. Exploring the efficiency of bias corrections of regional climate model output for the assessment of future crop yields in Europe. Reg. Environ. Change 14, 865–877.

Bannayan, M., Hoogenboom, G., 2008a. Predicting realizations of daily weather data for climate forecasts using the non-parametric nearest-neighbour re-sampling technique. Int. J. Climatol. 28 (10), 1357–1368.

Bannayan, M., Hoogenboom, G., 2008b. Weather analogue: a tool for real-time prediction of daily weather data realizations based on a modified k-nearest neighbor approach. Environ. Model. Softw. 23 (6), 703–713.

Basso, B., Liu, L., 2018. Seasonal crop yield forecast: methods, applications, and accuracies. Adv. Agron. 201–255.

Brandes, E., McNunn, G.S., Schulte, L.A., Bonner, I.J., Muth, D.J., Babcock, B.A., Sharma, B., Heaton, E.A., 2016. Subfield profitability analysis reveals an economic case for cropland diversification. Environ. Res. Lett. 11, 014009.

Buishand, T.A., Brandsma, T., 2001. Multisite simulation of daily precipitation and temperature in the Rhine Basin by nearest-neighbor resampling. Water Resour. Res. 37 (11), 2761–2776.

Cao, J., Zhang, Z., Luo, Y., Zhang, L., Zhang, J., Li, Z., Tao, F., 2021a. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. Eur. J. Agron. 123, 126204.

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., Xie, J., 2021b. Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches. Agric. For. Meteorol. 297, 108275.

Chen, S., Dou, Z., Jiang, T., Li, H., Ma, H., Feng, H., Yu, Q., He, J., 2017. Maize yield forecast with DSSAT-CERES-Maize model driven by historical meteorological data of analogue years by clustering algorithm. Trans. Chin. Soc. Agric. Eng. 33 (19), 147–155 (in Chinese with English abstract).

Chen, S., Jiang, T., Ma, H., He, C., Xu, F., Malone, R.W., Feng, H., Yu, Q., Siddique, K.H. M., Dong, Q.g., He, J., 2020. Dynamic within-season irrigation scheduling for maize production in Northwest China: a method based on weather data fusion and yield prediction by DSSAT. Agric. For. Meteorol. 285-286, 107928.

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., Warren, R., Qian, B., Daneshfar, B., Bedard, F., Reichert, G., 2015. Evaluation of the integrated canadian crop yield forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agric. For. Meteorol. 206, 137–150.

Chipanshi, A.C., Ripley, E.A., Lawford, R.G., 1997. Early prediction of spring wheat yields in Saskatchewan from current and historical weather data using the CERES-Wheat model. Agric. For. Meteorol. 84 (3–4), 223–232.

Dai, Y., Wei, S., Duan, Q., Liu, B., Fu, S., Niu, G., 2013. Development of a China dataset of soil hydraulic parameters using pedotransfer functions for land surface modeling. J. Hydrometeorol. 14 (3), 869–887.

de Wit, A., Baruth, B., Boogaard, H., van Diepen, K., van Kraalingen, D., Micale, F., te Roller, J., Supit, I., van den Wijngaart, R., 2010. Using ERA-INTERIM for regional crop yield forecasting in Europe. Clim. Res. 44 (1), 41–53.

de Wit, A.J.W., van Diepen, C.A., 2007. Crop model data assimilation with the ensemble kalman filter for improving regional crop yield forecasts. Agric. For. Meteorol. 146 (1–2), 38–56.

du Toit, A.S. and du Toit, D.L., 2003. Short description of the model statistical package and weather analogue program. modeling temperature response in wheat and maize: Proceedings of the a Workshop. CIMMYT, El Batán, Mexico.

Dumont, B., Leemans, V., Ferrandis, S., Bodson, B., Destain, J.P., Destain, M.F., 2014. Assessing the potential of an algorithm based on mean climatic data to predict wheat yield. Precis. Agric. 15 (3), 255–272.

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.

Gangopadhyay, S., Clark, M., Rajagopalan, B., 2005. Statistical downscaling using k-nearest neighbors. Water Resour. Res. 41, W02024.

Goel, S., Dash, S.K., 2007. Response of model simulated weather parameters to round-off-errors on different systems. Environ. Model. Softw. 22, 1164–1174.

Gouache, D., Bouchon, A.S., Jouanneau, E., Le Bris, X., 2015. Agrometeorological analysis and prediction of wheat yield at the departmental level in France. Agric. For. Meteorol. 209-210, 1–10.

Haboudane, D., 2004. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: modeling and validation in the context of precision agriculture. Remote Sens. Environ. 90 (3), 337–352.

Hansen, J.W., Potgieter, A., Tippett, M.K., 2004. Using a general circulation model to forecast regional wheat yields in northeast Australia. Agric. For. Meteorol. 127 (1–2), 77–92.

Harkness, C., Semenov, M.A., Areal, F., Senapati, N., Trnka, M., Balek, J., Bishop, J., 2020. Adverse weather conditions for UK wheat production under climate change. Agric. For. Meteorol. 282-283, 107862.

Hartkamp, A.D., White, J.W., Hoogenboom, G., 2003. Comparison of three weather generators for crop modeling: a case study for subtropical environments. Agric. Syst. 76 (2), 539–560.

He, J., Dukes, M.D., Jones, J.W., Graham, W.D., Judge, J., 2009. Applying GLUE for estimating CERES-Maize genetic and soil parameters for sweet corn production. Trans. ASABE 52 (6), 1907–1921.

He, L., Asseng, S., Zhao, G., Wu, D., Yang, X., Zhuang, W., Jin, N., Yu, Q., 2015. Impacts of recent climate warming, cultivar changes, and crop management on winter wheat phenology across the loess plateau of China. Agric. For. Meteorol. 200, 135–143.

Hodges, T., Evans, D.W., 1992. Leaf emergence and leaf duration related to thermal time. Agron. J. 84, 724–730.
Hoogenboom, G., Porter, C.H., Shelia, V., Boote, K.J., Singh, U., White, J.W., Hunt, L.A., Ogoshi, R., Lizaso, J.I., Koo, J., Asseng, S., Singels, A., Moreno, L.P. and Jones., J.W., 2017. Decision support system for agrotechnology transfer (DSSAT). version 4.7.
Huang, G., Chen, W., Li, F., 2011. Rainfed Farming Systems in the Loess Plateau of China. Springer, Netherlands, pp. 643–669.
IPCC, 2013. Climate Change 2013. Cambridge, United Kingdom.
Jha, P.K., Athanasiadis, P., Gualdi, S., Trabucco, A., Mereu, V., Shelia, V., Hoogenboom, G., 2019. Evaluating the applicability of using daily forecasts from seasonal prediction systems (SPSs) for agriculture: a case study of Nepal's Terai with the NCEP CFSv2. Theor. Appl. Climatol. 135 (3–4), 1143–1156.
Jones, C.A., Kiniry, J.R., Dyke, P.T., 1986. CERES-Maize: A simulation Model of Maize Growth and Development. A & M University Press.
Jones, J.W., He, J., Boote, K.J., Wilkens, P., Porter, C.H., Hu, Z., Ahuja, L.R., Ma, L., 2011a. Estimating DSSAT cropping system cultivar-specific parameters using Bayesian techniques. Methods of Introducing System Models into Agriculture Research. (Methods Introducing), pp. 365–394.
Jones, J.W., He, J., Boote, K.J., Wilkens, P., Porter, C.H., Hu, Z., 2011b. Estimating DSSAT cropping system cultivar-specific parameters using bayesian techniques. In: Ahuja, L.R., Ma, L. (Eds.), Methods of Introducing System Models into Agricultural Research. Advances in Agricultural Systems Modeling 2. American Society of Agronomy, Madison, WI USA, pp. 365–394.
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., 2003. The DSSAT cropping system model. Eur. J. Agron. 18 (3), 235–265.
Kang, Y., Ozdogan, M., Zipper, S.C., Roman, M.O., Walker, J., Hong, S.Y., Marshall, M., Magliulo, V., Moreno, J., Alonso, L., Miyata, A., Kimball, B., Loheide, S.P., 2016. How universal is the relationship between remotely sensed vegetation indices and crop leaf area index? A global assessment. Remote Sens. (Basel) 8 (7), 597.
Kilsby, C.G., Jones, P.D., Burton, A., Ford, A.C., Fowler, H.J., Harpham, C., James, P., Smith, A., Wilby, R.L., 2007. A daily weather generator for use in climate change studies. Environ. Model. Softw. 22 (12), 1705–1719.
Kipkorir, E.C., Raes, D., Massawe, B., 2002. Seasonal water production functions and yield response factors for maize and onion in Perkerra, Kenya. Agric. Water Manag. 56, 229–240.
Lawless, C., Semenov, M.A., 2005. Assessing lead-time for predicting wheat growth using a crop simulation model. Agric. For. Meteorol. 135 (1–4), 302–313.
Lecerf, R., Ceglar, A., López-Lozano, R., Van Der Velde, M., Baruth, B., 2018. Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe. Agric. Syst. 168, 191–202.
Lee, B.-H., Kenkel, P., Brorsen, B.W., 2013. Pre-harvest forecasting of county wheat yield and wheat quality using weather information. Agric. For. Meteorol. 168, 26–35.
Liu, Y., Tang, L., Qiu, X., Liu, B., Chang, X., Liu, L., Zhang, X., Cao, W., Zhu, Y., 2020. Impacts of 1.5 and 2.0 °C global warming on rice production across China. Agric. For. Meteorol. 284, 107900.
Mavromatis, T., 2016. Spatial resolution effects on crop yield forecasts: an application to rainfed wheat yield in north Greece with CERES-Wheat. Agric. Syst. 143, 38–48.
Mavromatis, T., Hansen, J.W., 2001. Interannual variability characteristics and simulated crop response of four stochastic weather generators. Agric. For. Meteorol. 109 (4), 283–296.
Morell, F.J., Yang, H.S., Cassman, K.G., Wart, J.V., Elmore, R.W., Licht, M., Coulter, J.A., Ciampitti, I.A., Pittelkow, C.M., Brouder, S.M., Thomison, P., Lauer, J., Graham, C., Massey, R., Grassini, P., 2016. Can crop simulation models be used to predict local to regional maize yields and total production in the U.S. Corn Belt? Field Crops Res. 192, 1–12.
Nandram, B., Berg, E., Barboza, W., 2014. A hierarchical Bayesian model for forecasting state-level corn yield. Environ. Ecol. Stat. 21 (3), 507–530.
Porter, J.R., Gawith, M., 1999. Temperatures and the growth and development of wheat: a review. Eur. J. Agron. 10, 23–36.
Prakash, K.J., Athanasiadis, P., Gualdi, S., Trabucco, A., Mereu, V., Shelia, V., Hoogenboom, G., 2019. Using daily data from seasonal forecasts in dynamic crop models for yield prediction: a case study for rice in Nepal's Terai. Agric. For. Meteorol. 265, 349–358.
Prescott, J.A., 1940. Evaporation from a water surface in relation to solar radiation. Trans. Roy. Soc. S. Aust. 46, 114–118.
Qian, B., De Jong, R., Warren, R., Chipanshi, A., Hill, H., 2009. Statistical spring wheat yield forecasting for the Canadian prairie provinces. Agric. For. Meteorol. 149 (6–7), 1022–1031.
Quiring, S.M., Legates, D.R., 2008. Application of CERES-Maize for within-season prediction of rainfed corn yields in Delaware, USA. Agric. For. Meteorol. 148 (6–7), 964–975.
Rajagopalan, B., Lall, U., 1999. A k-nearest-neighbor simulator for daily precipitation and other weather variables. Water Resour. Res. 35 (10), 3089–3101.
Salzberg, S., Delcher, A., Heath, D., Kasif, S., 1991. Best-case for nearest neighbor learning. IEEE Trans. Pattern Anal. Mach. Intell. 17, 599–610.
Schwalbert, R., Amado, T., Nieto, L., Corassa, G., Rice, C., Peralta, N., Schauberger, B., Gornott, C., Ciampitti, I., 2020. Mid-season county-level corn yield forecast for US Corn Belt integrating satellite imagery and weather variables. Crop Sci. 60 (2), 739–750.
Semenov, M.A., Barrow, E.M., 1997. Use of a stochastic weather generator in the development of climate change scenarios. Clim. Change 35 (4), 397–414.
Tao, F., Yokozawa, M., Xu, Y., Hayashi, Y., Zhang, Z., 2006. Climate changes and trends in phenology and yields of field crops in China, 1981–2000. Agric. For. Meteorol. 138 (1–4), 82–92.
Togliatti, K., Archontoulis, S.V., Dietzel, R., Puntel, L., VanLoocke, A., 2017. How does inclusion of weather forecasting impact in-season crop model predictions? Field Crops Res. 214, 261–272.
Tollenaar, M., Fridgen, J., Tyagi, P., Stackhouse, P.W., Kumudini, S., 2017. The contribution of solar brightening to the US maize yield trend. Nat. Clim. Change 7 (4), 275–278.
Wang, Q.X., Fan, X.H., Qin, Z.D., Wang, M.-B., 2012. Change trends of temperature and precipitation in the Loess Plateau Region of China, 1961–2010. Global. Planet. Change 92-93, 138–147.
Wang, X., Pan, X., Wang, S., Hu, L., Guo, Y., Li, X., 2017. Dynamic prediction method for cotton yield based on COSIM model in Xinjiang. Trans. Chin. Soc. Agric. Eng. 33 (8), 160–165.
Zhang, H., Oweis, T., 1999. Water-yield relations and optimal irrigation scheduling of wheat in the Mediterranean region. Agric. Water Manag. 38 (3), 195–211.