A new strategy for selection of crop cultivated varieties based on yield factors driving and meteorological prediction

Rongsheng Zhu^{1+*}, Kai Sun²⁺, Xuehui Yan², Dawei Xin³, Hongwei Jiang³, Zhanguo Zhang¹, Yang Li¹, Zhenbang Hu³, Zhaoming Qi³, Chunming Yang⁴, Qingshan Chen^{3*}

¹College of Science, Northeast Agricultural University, Harbin, 150030, China; ²College of engineering, Northeast Agricultural University, Harbin, 150030, China; ³ College of Agronomy, Northeast Agricultural University, Harbin, 150030, China; ⁴Agro-Biotechnology Research Institute , Jilin Academy of Agricultural Sciences (JAAS), Changchun, 130300, China.

⁺These authors contributed equally to this work.

*Corresponding author. Phone: 086-0451-55191945; Fax: 086-0451-55190960; E-mail: qshchen@126.com.

Abstract

Meteorological conditions, as an uncontrollable factor, have a direct impact on the growth and yield of crops in the field. It's particularly important to choose the cultivated varieties suitable for environmental change. As for this problem, a new strategy for selection of crop cultivated varieties is proposed in this paper. The experiment selected Harbin pilot (Heilongjiang, China) and Changchun pilot (Jilin, China). In the experiment, the cultivated varieties suitable for forecasting meteorological conditions in 2015 could be provided according to the meteorological data of 2005-2014 years and the yield data of different soybean varieties in Harbin pilot. The same method was also taken in Changchun pilot. The results showed that the recommended varieties of Harbin pilot and Changchun pilot had 70% and 50% probability respectively for selection of soybean cultivated varieties when the yield evaluation trait was 100-seed-weight. As for seed weight per plant, the

possibility of more than 60% of the recommended varieties in two pilots is to provide prejudgment.

Therefore, this method can help to improve the current situation of the impact of uncontrollable meteorological conditions on crop yield and also can be applied to the selection of any other crops cultivated varieties under a certain environment.

Keywords: Meteorological prediction; Phase space reconstruction; Exponential smoothing method; Selection of crop cultivated varieties; Crop yield

1 Introduction

With the development of agricultural science and technology, some management factors such as pesticides, fertilizers and herbicides are basically stable and controllable, however, the effects of environmental factors such as precipitation, temperature and sunshine on crop growth are still in an uncontrollable state¹. In recent decades, due to the changes of natural conditions and the impact of human activities, the global climate has shown a significant warming trend^{2,3}, and agriculture is one of the areas that are sensitive to meteorological conditions⁴. Changes in meteorological conditions will have a significant impact on the ecological environment of agriculture as well as the growth and yield of crops⁵. Currently, many experts have done a lot of research on how meteorological conditions affect crop growth and yield⁶⁻⁸. For example, Chinese experts pointed out that low temperature treatment can slow the growth of soybean plants, delay the growth period, reduce the number of pods and the 100-seed-weight⁹. Mina and others simulated the influence of the future climate on the growth of maize by combining the climate regional model with the crop growth models such as WOFOST and CERES-Maize¹⁰. The key meteorological factors influencing spring soybean sowing to seedling, seedling and blooming period are the average daily temperature, and the key meteorological factor from emergence to maturity is precipitation, however, the key

meteorological factors from flowering to maturity are sunshine hours and precipitation through the method of stepwise regression and correlation of path analysis¹¹. Wang adopted APSIM model to analyze the effects of climate change on winter wheat phenology¹². Pan described the increased demand for water as the plants grew faster. If the drought is encountered during the growth period, the number of shriveled grains and the 100-seed-weight will decrease, which will affect the yield of soybean¹³. Zhou and others introduced the impact of climate change on soil organic carbon and surface soil organic carbon storage and crop yield¹⁴. Chen and others analyzed the influence of meteorological factors on grain production in north China based on APSIM model¹⁵. Dong and others used the APSIM model to analyze the effect of drought on maize yield in northern China¹⁶.

According to the above research, it can be concluded that the environment has a direct impact on the growth and yield of crops. Research pointed out the biological significance of the genetic formula: phenotype = environment + genotype (P=E+G). The phenotype of an organism is the result of the accumulation and interaction of a particular genotype and the environment. But the genotype is built under the repeated effect of the environment, which is the genetic characteristic that the environment has given to the organism for a long time. It also confirms the idea that genotypes come from the environment¹⁷. Yan and He introduced that wheat starch traits were not only influenced by genotype and environmental accumulation factors, but also influenced by the interaction between genotype and environment¹⁸. Suppose the biological phenotype P is the yield in the formula, then yield will be affected by the internal genes of the crops and the external environment. Crop yields are generally affected by factors such as environmental 1 (meteorological factors), environment 2 (soil factors) and internal factors (genetic factors). Generally, it can be expressed as:

$$Y = E + S + G + \varepsilon$$

or

$$Y = f(E, S, G) + \varepsilon$$

Where E is the environment, S is the soil, G is the gene.

According to the expression, the influence factor of yield can be more generally expressed as:

$$Y = E + S + G + E \cdot S + E \cdot G + S \cdot G + E \cdot S \cdot G + \varepsilon$$

For the same place, S can be assumed to be constant, so $Y = f(E, G) + \varepsilon$. If it is assumed that the ε is contained in the environmental factor, then $Y = f(E, G)$. However, both E and G are difficult to describe and the specific functional relationships are difficult to determine as well. Based on the above analysis, it is crucial to select suitable planting varieties according to meteorological conditions to improve crop yield. Therefore, if the weather conditions and crop cultivated varieties are known, an evaluation scheme for the yield will be given.

A new strategy for crop cultivated varieties selection based on meteorological predication is proposed to solve the problem of the influence of meteorological conditions on crop yield in this paper. The strategy adopts the weather prediction model based on phase space reconstruction¹⁹ and exponential smoothing²⁰, and the evaluation classification model based on Euclidean distance²¹, which can achieve the optimal selection of planting varieties according to the target traits. This strategy provides a theoretical basis for the selection of crop cultivated varieties to improve crop yield.

In this paper, Harbin pilot (Heilongjiang Province, China) and Changchun pilot (Jilin Province, China) were selected as the test points. Harbin (N: 44°04'~46°40', E: 125°42'~130°10') and Changchun (N: 43°05'~45°15', E: 124°18'~127°02') are important agricultural bases in China. The two pilots are high yield soybean, which belong to the temperate continental monsoon climate,

with much land and few people, fertile soil and have a long history of planting soybean. However, in recent years, due to the changes of meteorological conditions, the market competitiveness of soybean in northeast China has decreased significantly. Most of the farmers are switching to other crops in the pursuit of higher returns, so the planting area of soybean significantly have been declining year by year.

In this experiment, the soybean varieties suitable for planting in the next year were selected according to the monthly meteorological data of Harbin from 2005 to 2014 and the meteorological data of Changchun from 2008 to 2009 through the use of the new selection strategy proposed in this paper. The experimental results showed that this strategy can predict the meteorological conditions of the next year accurately and can provide the recommended crop cultivated varieties through the classification evaluation model. This strategy also can help to improve the current situation of the impact of uncontrollable meteorological conditions on crop yield and it also has a certain significance for the planting of crops under a certain environment and condition.

2. Methods

2.1 Experiment materials

Compared with other meteorological conditions, precipitation, temperature and sunshine hours have a greater impact on soybean growth, which can be obtained by Shang²². During the growth cycle of soybean, the meteorological conditions of 5-9 months played an important role in the growth of soybean. Therefore, the meteorological data of monthly precipitation (Monthly precipitation refers to the depth of the liquid or solid that has been deposited from the sky to the surface of the ground in a certain month, without evaporation, infiltration, and loss of water.), monthly mean temperature (Monthly mean temperature refers to the average value of the average

temperature of all days in a month.) and monthly sunshine hours (Monthly sunshine hours refers to the summation of the multi day time of the sun's direct rays irradiating the ground within a month.) in Harbin during the past 2005-2015 years were selected as meteorological materials for Harbin pilot, which only included 5-9 months of meteorological data. The meteorological data from 2005 to 2014 are training samples, and the meteorological data of 2015 are testing samples. As for the Changchun pilot, the meteorological data of monthly precipitation, monthly mean temperature and monthly sunshine hours during the past 2008-2010 years were selected as meteorological materials, which also only included 5-9 months of meteorological data. The meteorological data from 2008 to 2009 are training samples, and the meteorological data of 2010 are testing samples. The meteorological data are provided by the statistical yearbook of Heilongjiang province and the statistical yearbook of Jilin province respectively²³. The detailed meteorological data refer to the Supplementary Table S1-S6 online.

The experiment of Harbin pilot was conducted in the experimental field of Northeast agricultural university from 2005 to 2015. A total of 149 soybean germplasm resources were used as the experiment materials, which are provided by Northeast Agricultural University, Harbin, Heilongjiang, China and cover the main varieties at home and abroad.

The experiment of Changchun pilot was conducted in the experimental field of Jilin academy of agricultural sciences from 2008 to 2010, and 274 soybean germplasm resources were selected as the experiment materials, which are provided by the Agricultural Biotechnology Research institute, Jilin Academy of Agricultural Sciences (JAAS). Detailed varieties and phenotype data are shown in Supplemental Table S7 and Table S8 online.

The planting time was at the end of April every year, and the harvest time was in late September

every year. The experiment was conducted in a randomized block group with three rows in each group, and three repeated tests were required in the process of planting. The row distance was 50cm, the total length of the row was 300cm, and the plant spacing of soybean in the two pilots was 6 ± 1 cm. The seedling, weeding and strengthening of the cultivated soil were carried out at seedling stage. In the process of flowering pod, the pod should be prevented from falling off. In the late flowering period, the largest soybean leaf area was guaranteed to be photosynthetic, and the water was replenished in time when the bean grain was about to mature.

2.2 Experiment model

The main experimental model of this paper is the selection model of crop cultivated varieties with single target and single location, such as the 100-seed-weight, seed weight per plant, seed number per plant, seed number per pod, plant height and so on, which can be used as the yield target traits. Plant varieties can be selected according to the requirements of planting by selecting models of crop cultivated varieties.

The crop cultivated varieties selection model for single location and single target trait is as follows:

Suppose:

$y_1, y_2 \dots y_n$ is the target traits of n years m species (such as 100-seed-weight or other traits)

$p_1, p_2 \dots p_n$ is the monthly precipitation of n years (mm)

$t_1, t_2 \dots t_n$ is the monthly mean temperature of n years ($^{\circ}\text{C}$)

$s_1, s_2 \dots s_n$ is the number of sunshine hours of n years (h)

The specific steps of the algorithm are as follows:

Step 1: Predict the average temperature t_{n+1} , monthly precipitation r_{n+1} and monthly sunshine hours s_{n+1} of the $n + 1$ year.

The meteorological conditions of the $n + 1$ year can be obtained based on the prediction model of certain time series, and the results are as follows.

$$p_{n+1} = f_1(p_1, p_2 \dots p_n)$$

$$t_{n+1} = f_2(t_1, t_2 \dots t_n)$$

$$s_{n+1} = f_3(s_1, s_2 \dots s_n)$$

Therefore, we can obtain the meteorological time series:

$$X_1 = (p_1, p_2 \dots p_n, p_{n+1})$$

$$X_2 = (t_1, t_2 \dots t_n, t_{n+1})$$

$$X_3 = (s_1, s_2 \dots s_n, s_{n+1})$$

Step 2: Standardize meteorological time series.

According to the formula below

$$X_i^* = \frac{X_{ij} - \text{mean}(X_i)}{\text{sd}(X_i)} \quad 1 \leq i \leq 3; 1 \leq j \leq n + 1 \quad (1)$$

The results of the standardized time series are as follows:

$$X_1^* = (p'_1, p'_2 \dots p'_n, p'_{n+1})$$

$$X_2^* = (t'_1, t'_2 \dots t'_n, t'_{n+1})$$

$$X_3^* = (s'_1, s'_2 \dots s'_n, s'_{n+1})$$

Step 3: Select a distance (usually the Euclidean distance) as a criterion. The Euclidean distance between the meteorological conditions of then $n + 1$ year and the meteorological conditions of the previous n years is calculated (equation (2)).

$$\begin{cases} d_1 = ((p'_{n+1} - p'_1)^2 + (t'_{n+1} - t'_1)^2 + (s'_{n+1} - s'_1)^2)^{\frac{1}{2}} \\ d_2 = ((p'_{n+1} - p'_2)^2 + (t'_{n+1} - t'_2)^2 + (s'_{n+1} - s'_2)^2)^{\frac{1}{2}} \\ \vdots \\ d_n = ((p'_{n+1} - p'_n)^2 + (t'_{n+1} - t'_n)^2 + (s'_{n+1} - s'_n)^2)^{\frac{1}{2}} \end{cases} \quad (2)$$

Step 4: Order $d_1, d_2 \cdots d_n$ from small to large scale.

$$d_1^* \leq d_2^* \leq \cdots \leq d_n^*$$

and then the Y_k ($1 \leq k \leq n$) year corresponding to the shortest Euclidean distance d_1^* is selected.

Step 5: Sort crop cultivated varieties of Y_k ($1 \leq k \leq n$) year in accordance with the yield target traits. Threshold value α is set according to the need of planting, and the top $\alpha\%$ varieties of the Y_k year are selected to form the recommended varieties set P for the $n + 1$ year.

2.2.1 Meteorological prediction model based on phase space reconstruction

The theory of phase space reconstruction proposed by Packard in 1980, and the theory holds that the evolution of any component in the system is determined by the other components that interact with it. Therefore, the information of these relevant components is implied in the development of any component.

For a nonlinear system, the observed value is recorded as time series $\{x(t_i), i = 1, 2, \cdots, n\}$, and the observed value can be used to form the m dimension vector:

$$X_i(t) = (x(t_i), x(t_i + \tau), \cdots, x(t_i + (m - 1)\tau)) \quad i = 1, 2, \cdots, n$$

Where τ is a delay time, m is embedded dimension, and $X_i(t)$ is a phase point in the constructed space.

Takens proves that the phase space is reconstructed by proper embedding dimension m and delay time τ . The "trajectory" in the embedded space is equivalent to the original system¹⁹. $d \geq 2n + 1$ is a sufficient condition for reconstructing the embedding dimension of phase space, and n is the number of independent parameters in the dynamic system. But Jiang points out that the necessary conditions have not been documented in the literature²⁴. Therefore,

$$X(t + T) = f(X(t)) \quad (t = 1, 2, \cdots, N)$$

Here, $T(T > 0)$ is the prediction model for forward prediction step, and $f(\cdot)$ is the prediction model of reconstruction.

According to the attractor in the fitting phase space, it can be divided into two kinds: the global method and the local method. The global method can get the $f(\cdot)$ by regarding all the points in the trajectory as the fitting objects. Thus, it is feasible in theory to predict the trend of trajectory. However, it is difficult to predict accurately when the phase space structure is more complex. The local method is to take the last point of the phase space trajectory as the center point and the recent locus of the center point as the relevant point, and then the relevant points are proposed to be combined to estimate the direction of the next point. Finally, the predicted values are separated from the coordinates of the predicted trajectories. Compared with the two methods, the application of local method is more extensive²⁵.

2.2.2 Exponential smoothing prediction model

The exponential smoothing method, proposed by American Brown in 1959, is one of the most commonly used methods for analysis of time series, which is mainly applied to the short-term forecast and is a kind of excellent performance, strong adaptability of prediction methods²⁶. This method assumes that the predicted value of the future is related to the known data, and the recent data has a large impact on the predicted value, but the long-term data has less influence on the predicted value. Therefore, the influence shows the trend of decreasing geometrical series. This method is widely used in various fields²⁷⁻²⁹.

The exponential smoothing model consists of an exponential smoothing method, the double exponential smoothing and the cubic exponential smoothing method. The basic formula (equation (3)) of exponential smoothing is:

$$S_t^{(n)} = \alpha S_t^{(n-1)} + (1 - \alpha) S_{t-1}^{(n-1)} \quad (3)$$

Where S_t is the exponential smooth value of the t cycle. $\alpha (0 \leq \alpha \leq 1)$ is a smooth exponent. t is a periodic number.

The simple exponential smoothing method is suitable for the time series analysis with horizontal trend, which can only predict the time series of the short term. If the time series has an upward or downward trend, it is best to predict with the double exponential smoothing. If the time series not only has the obvious rise or decrease trend but also presents the seasonal change, then the cubic exponential smoothing method is selected.

(1) The selection of initial value

In general, the observation value of the first phase can be selected as the initial value when the time series data items are large. If the number of the original numbers is relatively small, the average number of the first several periods (Usually the first three data) can be selected as the initial value.

(2) The selection of smoothness index σ

The smoothness index σ is a proportional factor for the different effects of the new and old data in the prediction, which not only represents the reaction speed of the model to time series data, but also determines the ability of prediction model to repair the error. The larger the value of σ , the greater the impact of the recent observed values, and the more dependent on the recent data, the higher the sensitivity of the model. Conversely, the smaller σ is, the slower the model changes, and the prediction results are more dependent on the early data.

There are three ways to determine the smoothness index σ : Difference-ratio-mean method, Experiential judgment method and Trial calculation method. The general selection criteria are as follows:

- 1) When the observed values show a stable horizontal trend, σ is generally selected between 0.1-0.3. The weight of each period is not very different, which reduces the magnitude of correction. Meanwhile, the new sequence contains more original sequence information, which helps to increase the credibility of prediction results.
- 2) When the observed values fluctuate greatly, σ is generally selected between 0.3-0.5.
- 3) When the observed value fluctuates greatly and presents obvious seasonal variation, the value range of σ is 0.6-0.8, so as to enhance the role of the recent observed value in the prediction process³⁰.

In this paper, σ is selected by Trial calculation method, and the determination of σ is generally evaluated by the minimum *RMSE* (root-mean-square error) (equation (4)) of the predicted results, and then the most reasonable σ value is selected³¹.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n [Y(t) - X(t)]^2} \quad (4)$$

Where n is the number of predicted data; $Y(t)$ represents the predicted values of the t period; $X(t)$ is the actual value of t period.

2.1.3 Evaluation classification model of crop cultivated varieties based on Euclidean distance

Suppose that the monthly precipitation, the monthly mean temperature and the monthly sunshine hours matrix of n years are respectively A, B and C.

$$A = \begin{pmatrix} p_{1,1}^1 & \cdots & p_{1,m}^m \\ p_{2,1}^1 & \cdots & p_{2,m}^m \\ \vdots & \cdots & \vdots \\ p_{n,1}^1 & \cdots & p_{n,m}^m \end{pmatrix} \quad B = \begin{pmatrix} t_{1,1}^1 & \cdots & t_{1,m}^m \\ t_{2,1}^1 & \cdots & t_{2,m}^m \\ \vdots & \cdots & \vdots \\ t_{n,1}^1 & \cdots & t_{n,m}^m \end{pmatrix} \quad C = \begin{pmatrix} s_{1,1}^1 & \cdots & s_{1,m}^m \\ s_{2,1}^1 & \cdots & s_{2,m}^m \\ \vdots & \cdots & \vdots \\ s_{n,1}^1 & \cdots & s_{n,m}^m \end{pmatrix}$$

In the matrix, m is the month of the observed value. According to the meteorological data of the previous n years, the prediction data matrix of the monthly precipitation, monthly average

temperature and monthly sunshine hours of $n + 1$ year can be obtained by using phase space reconstruction and exponential smoothing prediction model. Then a multi-dimensional matrix D with meteorological conditions of observed and predicted values can be obtained.

$$D = \begin{pmatrix} p_{1,1}^1 & \dots & p_{1,m}^m & \dots & p_{1,12}^{12} & t_{1,13}^1 & \dots & t_{1,12+i}^i & \dots & t_{1,24}^{12} & s_{1,25}^1 & \dots & s_{1,24+j}^k & \dots & s_{1,36}^{12} \\ p_{2,1}^1 & \dots & p_{2,m}^m & \dots & p_{2,12}^{12} & t_{2,13}^1 & \dots & t_{2,12+i}^i & \dots & t_{2,24}^m & s_{2,25}^1 & \dots & s_{2,24+j}^k & \dots & s_{2,36}^{12} \\ \vdots & \dots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ p_{n,1}^1 & \dots & p_{n,m}^m & \dots & p_{n,12}^{12} & t_{n,13}^1 & \dots & t_{n,12+i}^i & \dots & t_{n,24}^m & s_{n,25}^1 & \dots & s_{n,24+j}^k & \dots & s_{n,36}^{12} \\ p_{n+1,1}^1 & \dots & p_{n+1,m}^m & \dots & p_{n+1,12}^{12} & t_{n+1,13}^1 & \dots & t_{n+1,12+i}^i & \dots & t_{n+1,24}^m & s_{n+1,25}^1 & \dots & s_{n+1,24+j}^k & \dots & s_{n+1,36}^{12} \end{pmatrix}$$

Where $m(1 \leq m \leq 12)$ is the month of precipitation, $i(1 \leq i \leq 12)$ is the month of average temperature, and $k(1 \leq k \leq 12)$ is the month of sunshine hours.

The steps of the Euclidean distance evaluation classification model are as follows:

Step 1: Matrix D' can be gained after standardizing the column vectors of the multi-dimensional matrix D according to Eq. (1).

$$D' = \begin{pmatrix} p_{1,1}^{1'} & \dots & p_{1,m}^{m'} & \dots & p_{1,12}^{12'} & t_{1,13}^{1'} & \dots & t_{1,12+i}^{i'} & \dots & t_{1,24}^{12'} & s_{1,25}^{1'} & \dots & s_{1,24+j}^{k'} & \dots & s_{1,36}^{12'} \\ p_{2,1}^{1'} & \dots & p_{2,m}^{m'} & \dots & p_{2,12}^{12'} & t_{2,13}^{1'} & \dots & t_{2,12+i}^{i'} & \dots & t_{2,24}^{m'} & s_{2,25}^{1'} & \dots & s_{2,24+j}^{k'} & \dots & s_{2,36}^{12'} \\ \vdots & \dots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ p_{n,1}^{1'} & \dots & p_{n,m}^{m'} & \dots & p_{n,12}^{12'} & t_{n,13}^{1'} & \dots & t_{n,12+i}^{i'} & \dots & t_{n,24}^{m'} & s_{n,25}^{1'} & \dots & s_{n,24+j}^{k'} & \dots & s_{n,36}^{12'} \\ p_{n+1,1}^{1'} & \dots & p_{n+1,m}^{m'} & \dots & p_{n+1,12}^{12'} & t_{n+1,13}^{1'} & \dots & t_{n+1,12+i}^{i'} & \dots & t_{n+1,24}^{m'} & s_{n+1,25}^{1'} & \dots & s_{n+1,24+j}^{k'} & \dots & s_{n+1,36}^{12'} \end{pmatrix}$$

Step 2: Calculate the Euclidean distance between the k rows ($k = 1, 2 \dots n$), and the $n + 1$ row respectively by equation (5).

$$d_f = \left(\sum_{f=1}^k \left[(p_{n+1,m}^{m'} - p_{f,m}^{m'})^2 + (t_{n+1,12+i}^{i'} - t_{f,12+i}^{i'})^2 + (s_{n+1,24+j}^{k'} - s_{n,24+j}^{k'})^2 \right] \right)^{\frac{1}{2}} \quad (5)$$

Step 3: Order $d_1, d_2 \dots d_n$ from small to large scale.

$$d_1^* \leq d_2^* \leq \dots \leq d_n^*$$

Step 4: Select the year Y_k ($1 \leq k \leq n$) that corresponds to the shortest Euclidean distance. In the Y_k ($1 \leq k \leq n$) year, the actual yield of the crop is sorted according to the target traits, and then the suitable threshold α is selected to classify and select the varieties according to the actual

needs.

2.3 Experiment methods

2.3.1 Meteorological prediction method based on phase space reconstruction

In the prediction experiment of precipitation and sunshine, the experimental data materials of the monthly precipitation and monthly sunshine hours of Harbin pilot were converted into line graphs, and the results were shown in Fig. 1. The data line chart of monthly precipitation and monthly sunshine hours in Changchun pilot

Fig. 1 shows that the data of the monthly precipitation and the monthly sunshine hours in Harbin pilot show an obvious nonlinear trend. Therefore, the prediction model based on phase space reconstruction were chosen to predict precipitation and sunshine hours, the concrete steps are as follows:

Step 1: The time series $x(n), n = 1, 2, \dots, N$ can be obtained after zero mean processing according to the data of monthly precipitation and sunshine hours in the period from 2005 to 2014.

Step 2: Determine phase space saturation embedding dimension m and time delay τ . A self-correlation method was used to determine the time delay in phase space reconstruction^{32,33}. The commonly used methods for determining the best embedding dimension m were pseudo nearest neighbor³⁴, G-P method³⁵, singular value decomposition method³⁶, etc. In addition, there were various improved algorithms, such as C-C³⁷ and G-Z algorithms³⁸, etc. To sum up, pseudo nearest neighbor point method is a method with more application. The maximum Lyapunov exponent of the chaotic time series is calculated by using small-data method to prove that the time series of the monthly precipitation and the monthly sunshine time series have chaotic characteristics³⁹.

According to time delay τ and embedding dimension m , time series $x(n), n = 1, 2, \dots, N$ is

converted into the following results.

$$X(1) = [x(1), x(1 + \tau), \dots x(1 + (m - 1)\tau)]$$

$$X(2) = [x(2), x(2 + \tau), \dots x(2 + (m - 1)\tau)]$$

...

$$X(i) = [x(i), x(i + \tau), \dots x(i + (m - 1)\tau)]$$

...

$$X(M) = [x(M), x(M + \tau), \dots x(M + (m - 1)\tau)]$$

Where M is the space point for reconstructing space, $M = N - (m - 1)\tau$.

Suppose: $Y(m, i) = X(i) = [x(i), x(i + \tau), \dots x(i + (m - 1)\tau)]$

$$i = 1, 2, \dots, M$$

Therefore, $\{Y(m, i), i = 1, 2, \dots, M\}$ is the subspace of R^m space, which is the phase space required for the experiment.

Step 3: Search for near neighbors. In the phase space, suppose the current state point is $X(j)$, and the neighborhood point of $X(j)$ is found by calculating the Euclidean distance between the points $X(i)$ $i = 1, 2, \dots, M$ and the $X(j)$.

Step 4: Find the next phase point $X(i_1 + 1), X(i_2 + 1), \dots, X(i_p + 1)$ for all the nearest neighbors.

The coefficient matrix A and B are obtained by fitting the space trajectory in this small neighborhood.

$$\begin{bmatrix} X(i_1 + 1) \\ X(i_2 + 1) \\ \dots \\ X(i_p + 1) \end{bmatrix} = A + B \begin{bmatrix} X(i_1) \\ X(i_2) \\ \dots \\ X(i_p) \end{bmatrix}$$

After obtaining A and B , the trend of trajectories in phase space can be get according to formula

$X(j + 1) = A + BX(j)$, so the predicted value of $x(j + 1 + (m - 1)\tau)$ in monthly precipitation and

monthly sunshine hours is obtained.

2.3.2 Exponential smoothing weather prediction method

In the prediction experiment of average temperature, the average temperature test material data of Harbin pilot was converted into a line chart, as shown in Fig. 2. The line chart of the monthly mean temperature test materials for Changchun pilot is shown in Supplemental Fig. S2 online. Fig. 2 shows that the time series data of the monthly mean temperature show a relatively stable horizontal trend, so an exponential smoothing method was chosen to predict.

An exponential smoothing prediction method is as follows:

Step 1: An exponential smoothing prediction method formula.

It is assumed that the observed values of the time series for the 5-9 months period of 2005-2014 years are $x_i (i = 1, 2 \dots)$. The exponential smoothing value of phase n is s_n^1 , and the corresponding formula is equation (6):

$$s_n^1 = \sigma x_n + (1 - \sigma)s_{n-1}^1 \quad (6)$$

Where x_n is the actual value of phase n . σ is the smoothness index; s_{n-1}^1 is the smooth value of the previous period; s_n^1 is the predicted value of the $n + 1$ period, which is the smoothing value of this period.

Step 2: The selection of initial value.

The number of the observed value of Harbin pilot is 10, which is less than 15. Therefore, the average value of the first three data of the observed values is used as the initial value of the smoothing prediction (equation (7)).

$$S_0^{(1)} = (x_1 + x_2 + x_3)/3 \quad (7)$$

Step 3: The selection of smoothness index σ .

According to Fig 2, it can be concluded that the observed values show a relatively stable horizontal trend, so the smoothness index should be smaller, generally between 0.1 and 0.3. In this paper, 0.1, 0.2 and 0.3 were selected to predict the average monthly temperature respectively, and at the end of the prediction, the σ value corresponding to the minimum *RMSE* can be used as the smoothness index of the monthly average temperature prediction.

2.3.3 Evaluation classification method based on Euclidean distance

Multi-dimensional matrix D can be obtained according to Euclidean distance evaluation classification model and meteorological materials. In this paper, Harbin pilot is taken as an example. Therefore, the first ten lines in the multi-dimensional matrix D represent the actual meteorological conditions of 2005-2014. The eleventh line stands for the forecast weather conditions for 2015. m, i and k range from 5-9, so the matrix D can be expressed as:

$$D = \begin{pmatrix} r_{1,1}^5 & \dots & r_{1,5}^9 & t_{1,6}^5 & \dots & t_{1,10}^9 & s_{1,11}^5 & \dots & s_{1,15}^9 \\ r_{2,1}^5 & \dots & r_{2,5}^9 & t_{2,6}^5 & \dots & t_{2,10}^9 & s_{2,11}^5 & \dots & s_{2,15}^9 \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ r_{10,1}^5 & \dots & r_{10,5}^9 & t_{10,6}^5 & \dots & t_{10,10}^9 & s_{10,11}^5 & \dots & s_{10,15}^9 \\ r_{11,1}^5 & \dots & r_{11,5}^9 & t_{11,6}^5 & \dots & t_{11,10}^9 & s_{11,11}^5 & \dots & s_{11,15}^9 \end{pmatrix}$$

Step 1: Matrix D' can be gained after standardizing the column vectors of the multi-dimensional matrix D .

$$D' = \begin{pmatrix} r_{1,1}^{5'} & \dots & r_{1,5}^{9'} & t_{1,6}^{5'} & \dots & t_{1,10}^{9'} & s_{1,11}^{5'} & \dots & s_{1,15}^{9'} \\ r_{2,1}^{5'} & \dots & r_{2,5}^{9'} & t_{2,6}^{5'} & \dots & t_{2,10}^{9'} & s_{2,11}^{5'} & \dots & s_{2,15}^{9'} \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ r_{10,1}^{5'} & \dots & r_{10,5}^{9'} & t_{10,6}^{5'} & \dots & t_{10,10}^{9'} & s_{10,11}^{5'} & \dots & s_{10,15}^{9'} \\ r_{11,1}^{5'} & \dots & r_{11,5}^{9'} & t_{11,6}^{5'} & \dots & t_{11,10}^{9'} & s_{11,11}^{5'} & \dots & s_{11,15}^{9'} \end{pmatrix}$$

Step 2: The Euclidean distance between the first ten lines of the matrix D' and the eleventh line of the matrix is calculated by equation (5). Then, according to the steps 3 and 4 in the evaluation classification model, the classification selection of the varieties of Harbin pilot was completed. The calculation method of the Changchun pilot was the same.

3 Results

3.1 The Predicted results of meteorological conditions

In the prediction experiment of monthly precipitation, the time delay τ and the embedding dimension m were calculated by self-correlation method and the pseudo nearest neighbor point method in Harbin pilot, so the time delay $\tau = 2$ and the embedding dimension $m = 4$. The maximum Lyapunov index of the chaotic time series was calculated by small-data method, which was 1.6234 and indicated that the monthly precipitation time series was chaotic. The same method was used in Changchun pilot to obtain the time delay τ , the embedding dimension m and the maximum Lyapunov exponent, and the results were 2, 4 and 1.3728 respectively. In the experiment of monthly sunshine hours, the time delay τ and embedding dimension m of Harbin pilot and Changchun pilot were obtained by the self-correlation method and the pseudo nearest neighbor point method respectively. The results were that the time delay $\tau = 2$ and the embedding dimension $m = 6$ in Harbin pilot, the time delay $\tau = 2$ and the embedded dimension $m = 4$ in Changchun pilot.

A prediction model based on an exponential smoothing method was used in the prediction experiment of monthly mean temperature. According to the distribution trend of monthly mean temperature data and the selection criteria of sigma, the smoothness index 0.1, 0.2 and 0.3 were selected for meteorological prediction. The *RMSE* values corresponding to different smoothness index were shown in Table 1 and Fig. 3. The *RMSE* results of Changchun pilot were shown in Supplemental Fig. 3 online. Therefore, according to Fig. 3, the smoothness index corresponding to the minimum *RMSE* value of May to September was selected for meteorological prediction. Therefore, in the experiment of the average temperature forecast from May to September in 2015 in Harbin pilot, the smoothness index was 0.3, 0.2, 0.3, 0.2 and 0.3 respectively (Fig. 4). In

Changchun pilot, the average temperature of May to September in 2010 was predicted by the smoothness index 0.3, 0.2, 0.3, 0.2 and 0.3(Supplemental Fig. 4 online). The prediction meteorological data of the Harbin pilot in 2015 and the Changchun pilot in 2010 were obtained through the above methods (Table 2). As shown in Table 2, the meteorological prediction model can accurately predict the meteorological conditions in the coming year and provide theoretical data support for the implementation of the strategy.

Table 1

The *MSE* value of different smoothness exponent σ in the monthly mean temperature prediction experiment of Harbin pilot and Changchun pilot in May to September.

Pilot	Month	$\sigma=0.1$	$\sigma=0.2$	$\sigma=0.3$
Harbin pilot	5	1.3914	1.2875	1.1781
	6	1.6738	1.5315	2.8954
	7	0.5951	0.5293	0.3863
	8	0.4428	0.3910	0.3918
	9	0.5185	0.4894	0.3913
Changchun pilot	5	1.2763	1.1950	1.0992
	6	1.1816	1.1046	1.0141
	7	0.7611	0.7080	0.6504
	8	0.0954	0.0906	0.0959
	9	0.9684	1.0945	1.0460

Table 2

The forecast data of two pilot meteorological conditions are compared with the actual data. The pilot data of Harbin is the actual meteorological data and forecast meteorological data for 2015. The pilot data of Changchun is the actual meteorological data and forecast meteorological data for 2010.

Pilot	Month	Monthly precipitation(mm)		Monthly mean temperature(°C)		Monthly sunshine hours(h)	
		Actual value	Predicted value	Actual value	Predicted value	Actual value	Predicted value
Harbin pilot	5	77.6	78.2	14.2	15.4	156.9	160.7
	6	77.3	72.4	22.1	22.2	226.7	234.2
	7	52.9	54.1	23.6	23.4	262.9	257.6
	8	110.5	120.7	22.8	22.2	152.8	161.1
	9	24.8	22.1	16.2	15.7	209.2	198.5
Changchun pilot	5	162.0	170.0	16.0	16.9	206.0	212.3
	6	30.5	32.7	23.6	19.6	306.5	313.5
	7	276.3	271.3	23.1	22.4	155.3	149.7
	8	205.9	210.2	21.8	22.2	212.0	209.6
	9	7.9	8.4	16.8	15.9	242.7	248.3

3.2 Evaluation classification results based on Euclidean distance

The multidimensional matrix D is composed of the actual value of the monthly precipitation, the monthly mean temperature and the sunshine hours of the Harbin pilot for 2005-2014 years and the predicted value of the same meteorological conditions in 2015. The multidimensional matrix D consists of only 5-9 months of meteorological data.

Matrix D' was obtained after standardizing the column vectors of the multi-dimensional matrix D . Then the Euclidean distance between 2015 and 2005-2014 years was obtained according to equation (5) (Fig. 5). The matrix D' of Changchun pilot was obtained in the same way, and the Euclidean distance results were shown in Supplemental Fig. 5 online.

Fig. 5 shows that the shortest Euclidean distance in the Harbin pilot is 2014. Therefore, the meteorological conditions of Harbin pilot in 2014 are similar to the meteorological conditions in 2015, so the varieties planted in 2014 can be recommended to be planted in 2015.

The yield evaluation traits in the experiment were 100-seed-weight and seed weight per plant. Therefore, the actual varieties of soybean in 2014 were sorted by 100-seed-weight and seed weight per plant respectively. Then the threshold α was selected according to the actual needs, and the cultivated varieties in 2015 were selected through the sorting data of the actual soybean varieties in 2014. Supplemental Fig. 5 online shows that the shortest Euclidean distance in the Changchun pilot is 2008. Therefore, the meteorological conditions of Changchun pilot 2008 are similar to that of 2010. The cultivated varieties in 2008 were recommended to be planted in 2010. The yield evaluation traits were also 100-seed-weight and seed weight per plant, and the selection of soybean varieties in Changchun in 2010 will be realized through the same method.

According to 100-seed-weight and seed weight per plant, the actual soybean varieties in 2014 and

2015 in Harbin pilot were sorted, and the actual soybean varieties in 2008 and 2010 in the Changchun pilot were also sorted (Table 3; Table 4). Table 3 and Table 4 enumerate only the varieties required for the experiment, and more sorting varieties refer to the Supplemental Table S7 and Table S8.

Table 3

The actual planting varieties of Harbin pilot in 2014 and 2015 and Changchun pilot in 2008 and 2010 were ranked according to 100-seed-weight. 100-seed-weight is an indicator of the size and richness of seeds, which is usually used in large seeds such as corn, soybeans, peanuts, cotton and so on. Expression of varieties: the name of the variety (the community of planting the varieties). For example, He Feng30 (A22) refers to the name of the variety is He Feng30, and the planting community is A22.

Harbin pilot				Changchun pilot			
Recommended varieties for 2015		Actually planted varieties in 2015		Recommended varieties for 2010		Actually planted varieties in 2015	
Varieties	^a 100-seed-weight (g)	Varieties	100-seed-weight (g)	Varieties	100-seed-weight (g)	Varieties	100-seed-weight (g)
142	22.88	141	26.23	He Feng30 (A22)	46.6	Liao Xian 1(A11)	29.2
141	22.81	142	24.69	Liao Xian 1(A11)	31.8	LS201(B122)	26.7
139	22.41	138	24.59	Long Quan (A4)	9.2	Stone soybean 1(B21)	25.9
145	22.26	145	24.38	LS201(B122)	28.9	Ji Yu94 (B23)	25.9
52	22.20	139	23.82	Tong Nong13 (C111)	8.0	Tong Nong13(C111)	25.1
138	21.78	61	23.06	Sui02-336(B17)	26.4	Dong Nong42(A185)	24.9
39	21.77	137	22.94	Sui02-339(B19)	26.3	Huang Dali (C197)	24.3
137	21.02	62	22.75	Kam bean33(C224)	26.3	Ji Yu57(B8)	23.6
46	19.89	12	22.74	Vinton 81(C8)	25.3	Sui02-339(B19)	23.4
61	19.78	125	22.22	Dong Nong42 (A185)	24.7	NMH(C218)	22.7

Table 4

The actual planting varieties of Harbin pilot project in 2014 and 2015 and Changchun in 2008 and 2010 were ranked according to seed weight per plant. Seed weight per plant refers to the total weight of all the seeds harvested from a single plant.

Harbin pilot				Changchun pilot			
Recommended varieties for 2015		Actually planted varieties in 2015		Recommended varieties for 2010		Actually planted varieties in 2015	
Varieties	^a seed weight per plant (g)	Varieties	seed weight per plant (g)	Varieties	seed weight per plant (g)	Varieties	seed weight per plant (g)
142	41.74	130	27.15	Ji Yu71(B124)	25.47	Stone soybean 1(B21)	31.2
141	37.01	143	25.26	Liao Xian 1(A11)	24.53	Ji Yu94(B23)	31.2
99	35.91	145	24.97	Dawson(A189)	24.52	Liao Xian 1(A11)	29.01
44	35.87	142	24.93	Jin Zhou 4-1(C119)	23.34	Ji Yu71(B124)	28.57
145	35.02	141	23.60	Ferric pods(B5)	23.24	Ferric pods(B5)	28.47

Fig. 6a shows the density distribution of yield evaluation traits (100-seed-weight) of the Harbin pilot in 2010. The density distribution of 100-seed-weight in the 2008 year of Changchun pilot is shown in Supplemental Fig. 6a online. The threshold value was selected according to the actual requirements of different pilots during the selection process. For example, suppose that the threshold α of the Harbin pilot was selected 0.07 to carry out the selection of varieties. Therefore, the selection of varieties was carried out at the high-end area (the right side of Fig. 6a). The total number of varieties in Harbin pilot in 2010 was 149, and the top ten varieties were selected according to the threshold. It was assumed that the threshold α was 0.04 for selection of varieties in Changchun pilot. The total number of varieties in Changchun was 274 in 2008, so the top ten varieties were selected. For 100-seed- weight, The final recommended varieties in the Harbin pilot were 142, 141, 139, 145, 52, 138, 39, 137, 46, 61. The final recommended varieties in the Changchun pilot were He Feng30 (A22), Liao Xian 1(A11), Long Quan (A4), LS201(B122), Tong Nong13 (C111), Sui02-336(B17), Sui02-339(B19), Kam bean 33(C224), Vinton 81(C8), Dong Nong42 (A185) (Table 3).

Fig. 6b shows the density distribution of yield evaluation traits (seed weight per plant) of the Harbin pilot in 2010. The density distribution of seed weight per plant in the 2008 year of Changchun pilot is shown in Supplemental Fig. 6b online. For seed weight per plant, the threshold α of the Harbin pilot was selected 0.03 to carry out the selection of varieties. Therefore, the selection of varieties was carried out at the high-end area (the right side of Fig. 6b). The total number of varieties in Harbin pilot in 2010 was 149, and the top five varieties were selected according to the threshold. It was assumed that the threshold α was 0.02 for selection of varieties in Changchun pilot. The total number of varieties in Changchun was 274 in 2008, so the top five varieties were selected. Therefore, the final recommended varieties in the Harbin pilot were 142, 141, 99, 44, 145. The final recommended varieties in the Changchun pilot were Jiyu71(B124), Liao Xian 1(A11), Dawson(A189), Jinzhou 4-1(C119), Ferric pods(B5) (Table 4).

The top ten varieties of 100-seed- weight in 2015 were 141, 142, 138, 145, 139, 61, 137, 62, 12, 125 in Harbin pilot (Table 3). Therefore, it is concluded that there were 7 recommended varieties of soybean in the top ten of the actual ranking in 2015. The top ten varieties of 100-seed-weight in

2010 were Liao Xian 1(A11), LS201(B122), Stone soybean 1(B21), Ji Yu94 (B23), Tong Nong13(C111), Dong Nong42(A185), Huang Dali(C197), Ji Yu57(B8), Sui02-339(B19), NMH(C218) in Changchun pilot (Table 3). It was concluded that there were five recommended varieties of soybean in the top ten of the actual ranking in 2010. Therefore, according to the above analysis ratio, the recommended varieties in Harbin pilot and Changchun pilot could have the possibility of 70% and 50% above, respectively, to provide a prediction for our selection.

The top five varieties of seed weight per plant in 2015 were 130, 143, 145, 142, 141 in Harbin pilot. In Changchun pilot, the top five varieties in 2010 were Stone soybean 1(B21), Ji Yu94(B23), Liao Xian 1(A11), Ji Yu71(B124), Ferric pods(B5) (Table 4). Therefore, it was concluded that the first five soybean varieties in Harbin pilot in 2015 included three recommended varieties, and the result of Changchun pilot was the same as that of it. So as for the seed weight per plant, the possibility of more than 50% of the recommended varieties in Harbin pilot and Changchun pilot was to provide a pre-judgement for the selection of cultivated varieties. The above analysis is of vital importance to the selection of new varieties for selecting new varieties and increasing yield.

4 Discussion

In the study of varieties selection of crops, the multi-objective selection function was established by the multi-objective decision technology which combines the analytic hierarchy process and the utility function method⁴⁰. Liao and Guan completed the combination of computer information technology and agronomy, so the experts' experience knowledge and the methods of solving problem were inherited and solidified of rapeseed, which laid the foundation for the standardization and informatization of rapeseed cultivation and management⁴¹. The methods recommended in the above literatures mostly relied on experiential knowledge rules in knowledge base and a large number of regional experiments to achieve the selection of planting varieties, and the effect of meteorological conditions on crop growth was not considered in the process of seed selection.

Though the paper used knowledge engineering and system modeling methods, the dynamic relationship between varieties selection and varieties, ecological environment and farming system was summarized, and a quantitative knowledge model for suitable varieties selection was established⁴². However, there are few articles on yield, disease resistance and phenotypic quality index, so the selection methods based on yield traits is still worth studying.

The meteorological prediction model based on phase space reconstruction and exponential smoothing method can predict meteorological conditions accurately and provide accurate data support for the selection of soybean varieties. However, some meteorological conditions have a variety of time scales, which makes meteorological conditions pretty complicated. If only one prediction model is used in the trial process, the complexity of the time series itself will lead to the inaccuracy of the prediction results. Therefore, if we want to obtain accurate prediction results, we will choose multiple time series prediction models according to the experiment requirements, such as the combination of phase space reconstruction and neural network.

Although the prediction method of phase space reconstruction and based on exponential smoothing in long term prediction is not able to make full use of multi-scale characteristics and regularities of the variation of the meteorological conditions. Chaotic space theory can make better use of the information contained in the time series to solve the problems that are difficult to solve in the prediction model of general nonlinear meteorological conditions in the short-term prediction, which can predict the meteorological conditions more accurately and meet the requirement of this strategy.

The meteorological data of monthly precipitation, monthly mean temperature and monthly sunshine hours in Harbin during the past 2005-2015 years were selected as meteorological materials for Harbin pilot in this paper. However, the Changchun pilot only used meteorological data from 2008 to 2010. In Harbin pilot, 2013-2015 years of meteorological data can also be used for the application of the strategy, and the concrete experiment steps are as follows.

Step 1: The meteorological data from 2013 to 2014 were selected as training samples, and the meteorological data in 2015 were testing samples. Meteorological data of 2015 were obtained by using the meteorological prediction model in the text (Table 5).

Step 2: According to equation (1), the meteorological data of 2013-2015 years (the meteorological data of 2015 as the prediction data) were standardized. Then, according to the evaluation classification model based on Euclidean distance and equation (5), the Euclidean distance between 2015 and 2013 was 4.78, and the Euclidean distance between 2015 and 2014 was 4.80. Therefore, the meteorological conditions in 2015 were similar to that in 2013, so the soybean varieties in 2013 could be the preferred planting varieties in 2015. The density distribution of 100-

seed-weight of soybean varieties in Harbin pilot in 2013 was shown in Fig. 7.

In this experiment, the threshold α was also selected to be 0.07, and the selection of the varieties could be made in the high-end area (the right side of Fig. 7). According to the Supplemental Table S7, the top ten varieties of 100-seed-weight in 2013 were: 142, 145, 138, 68, 52, 141, 39, 21, 139, 127. In 2015, the top ten varieties of actual soybean were 141, 142, 138, 145, 139, 61, 137, 62, 12, 125. It is easy to conclude that there were five kinds of recommended planting varieties in 2015. The method was to use the meteorological data from 2013 to 2015 for the application of this model. The results showed that even three years of meteorological data also can provide more than 50% possibilities for planting varieties in the coming year, which was also of great practical significance

Table 5

Comparison of forecast data and actual data of meteorological conditions in Harbin in 2015. In this form, the meteorological data in 2013-2014 years are used as experimental materials to predict the meteorological conditions in 2015 through the meteorological prediction model in the text.

Pilot	Month	Monthly precipitation(mm)		Monthly mean temperature(°C)		Monthly sunshine hours(h)	
		Actual value	Predicted value	Actual value	Predicted value	Actual value	Predicted value
Harbin pilot	5	77.6	75.3	14.2	15.5	156.9	150.7
	6	77.3	80.4	22.1	22.4	226.7	230.1
	7	52.9	55.1	23.6	23.5	262.9	258.9
	8	110.5	110.5	22.8	22.1	152.8	158.3
	9	24.8	26.9	16.2	15.6	209.2	213.6

for improving soybean yield. This part only studied the 100-seed-weight, and the evaluation method of seed weight per plant is the same as that of the text.

In this paper, the selection model of single target and single location was discussed, and the model of single target and multi-location, multi-target and single location, and multi-target and multiple locations.

As for the selection model of single target and multi-location, the locations can be recorded as: $P_1, P_2, \dots, P_m$. The meteorological years to be studied can be recorded as $C_1, C_2, \dots, C_s$. Therefore, the environmental factors changed from the previous S to the $P * S$. Then it will be selected according to the model of single target and single location.

As for the selection model of multi-target and single location, the following method can be used.

Suppose there are n species and m traits:

$$\text{Traits: } T_1, T_2, \dots, T_m$$

$$\text{Species: } V_1, V_2, \dots, V_n$$

For example, $s(s \leq n)$ species are selected for the target trait $T_i(i \in \{1, m\})$. The varieties are expressed as follows:

$$T_i V = (V_1^{(i)}, \dots, V_s^{(i)}) \quad V_1^{(i)}, \dots, V_s^{(i)} \in \{V_1, V_2, \dots, V_n\}$$

The soybean varieties are sorted by the size of the main target trait $T_i(i \in \{1, m\})$:

$$(V_1^{(i)} \leq V_1^{(i)} \leq \dots \leq V_s^{(i)})$$

For different traits $T_1, T_2, \dots, T_m$, the sequence vectors of the selected varieties are as follows:

$$\begin{aligned} T_1 V &= (V_1^{(1)}, V_2^{(1)}, \dots, V_s^{(1)}) \\ T_2 V &= (V_1^{(2)}, V_2^{(2)}, \dots, V_s^{(2)}) \\ &\dots \\ T_m V &= (V_1^{(m)}, V_2^{(m)}, \dots, V_s^{(m)}) \end{aligned}$$

$V_{(i)}^{(j)} \in \{V_1, V_2, \dots, V_n\}$, $T_i(V_j)$ represents the value of the i trait of the j variety, $i \in \{1, m\}, j \in \{1, n\}$.

In the selection of multi-target, the required yield main trait can be selected according to the single

target selection strategy, and the minimum value of the remaining traits is then defined according to the actual needs. Suppose the minimum value of the yield main character is set to ∂ , and there is a variety of $V_h^{(2)} (1 \leq h \leq s)$ in T_2V , which does not appear in T_1V . Then the yield main character value of $V_h^{(2)}$ is required. If the yield main character value of the $V_h^{(2)}$ is higher than ∂ , the T_1V is reordered after the $V_h^{(2)}$ is inserted into the T_1V according to the value of the yield main trait of $V_h^{(2)}$. If the yield main character value of $V_h^{(2)}$ is lower than ∂ , it will be discarded directly. According to the above methods, the selection of multi-target traits can be achieved.

As for the selection model of multi-target and multiple locations, the environmental factors of multi-location can be determined according to the model of single target and multi-location. The selection of multi-target can be determined according to the selection model of multi-target and single location.

The classification model of crop cultivated varieties can accurately classify the varieties according to Euclidean distance. The method in the text is carried out after the standardization of the column vector of the meteorological condition matrix D . The method discussed here is to calculate the Euclidean distance between the three meteorological conditions respectively.

Because the Euclidean distance is calculated by the same meteorological conditions, there is no intrinsic difference between the same weather conditions, and standardization is not used in the model. The specific model is as follows:

Suppose the meteorological data matrix for n years is A .

$$A = \begin{pmatrix} x_{11}^i & x_{12}^i & x_{13}^i \\ x_{21}^i & x_{22}^i & x_{23}^i \\ \vdots & \vdots & \vdots \\ x_{121}^i & x_{122}^i & x_{123}^i \end{pmatrix}$$

The $i (1 \leq i \leq n)$ in the matrix means the number of years of the observed value. The three columns in the matrix represent monthly precipitation, monthly mean temperature and monthly sunshine hours respectively. Based on the phase space reconstruction and exponential smoothing prediction model, the matrix B that consists of the predicted data of monthly precipitation, monthly mean temperature and monthly sunshine hours of $n + 1$ year can be obtained.

$$B = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{121} & x_{122} & x_{123} \end{pmatrix}$$

The idea of the evaluation classification model is to get the Euclidean distance d of each column vector of the prediction matrix B and the observation matrix A separately. The approximate degree of the prediction matrix B and the observation matrix A is judged by the size of the Euclidean distance. The smaller the Euclidean distance, the greater the degree of similarity, and then Y_1, Y_2 and Y_3 , which correspond to the shortest Euclidean distances of the three meteorological conditions respectively, are selected. The Euclidean distance can be calculated according to the monthly precipitation calculation formula equation (8), monthly mean temperature formula equation (9) and monthly sunshine hours calculation formula equation (10).

$$d_p = \left(\sum_{m=1}^{12} (x_{m1} - x_{m1}^{i_1})^2 \right)^{\frac{1}{2}} \quad (8)$$

$$d_w = \left(\sum_{m=1}^{12} (x_{m2} - x_{m2}^{i_2})^2 \right)^{\frac{1}{2}} \quad (9)$$

$$d_r = \left(\sum_{m=1}^{12} (x_{m3} - x_{m3}^{i_3})^2 \right)^{\frac{1}{2}} \quad (10)$$

Taking the Harbin pilot as an example, this method can be used to obtain the Euclidean distance between the observation matrix A and the prediction matrix B in Harbin pilot (Fig. 8), and the results of the Changchun pilot are shown in Supplemental Fig. 7 online.

In this method, the 100-seed-weight and seed weight per plant can still be used as yield evaluation traits. Here, only 100-seed-weight is used as an example. Fig. 8 shows that the monthly precipitation and monthly average temperature of the Harbin pilot in 2015 are similar to those in 2014, and the weather condition of the monthly sunshine hours is similar to that in 2007. In 2007 and 2014, the data of the top ten varieties of soybean 100-seed-weight were shown in Table 6. Supplemental Fig. 7 online shows that the monthly precipitation, monthly average temperature and sunshine hours of Changchun in 2010 are similar to those of 2008, and the selection of planting varieties can be realized by the same method.

The weight of different meteorological factors that affect the growth of soybean needs to be considered when selecting the recommended planting varieties by this method. Because of the different areas of planting crops, the weight of meteorological factors will be different. Therefore, the weight of the three meteorological conditions should be allocated according to the meteorological conditions of the planting area, and then the selection of crop cultivated varieties should be carried out. The specific steps are as follows:

Step 1: It is assumed that the total number of crop cultivated varieties planted in Y_1 , Y_2 and Y_3 is A, B and C respectively.

Step 2: The weight of precipitation, average temperature and sunshine hours is $\alpha_1, \alpha_2, \alpha_3$, respectively, according to the meteorological conditions of the actual planting area.

Step 3: A, B and C are sorted according to the yield evaluation traits, and the number of varieties will be selected according to different weights. Therefore, the total number of varieties selected in Y_1, Y_2 and Y_3 is $A \cdot \alpha_1$, $B \cdot \alpha_2$ and $C \cdot \alpha_3$ respectively.

Step 4: The final set of recommended planting variety is $P = A \cdot \alpha_1 \cap B \cdot \alpha_2 \cap C \cdot \alpha_3$.

Table 6

The ranking of 100-seed-weight of soybean varieties in 2007 and 2014 in Harbin pilot. Fig. 8 shows that the monthly precipitation and monthly average temperature of the Harbin pilot in 2015 are similar to those in 2014, and the weather condition of the monthly sunshine hours is similar to that in 2007.

2007		2014	
Varieties	100-seed-weight (g)	Varieties	100-seed-weight (g)
3	23.02	110	31.89
127	23.01	142	22.88
7	22.92	141	22.81
143	22.82	139	22.41
32	22.49	145	22.26
147	22.04	52	22.20
99	21.84	138	21.78
136	20.96	39	21.77
9	22.81	137	21.02
133	21.39	46	19.89

5 Conclusion

In this paper, a new strategy based on meteorological driving for classification and selection of cultivated varieties is presented. The main purpose is to select cultivated varieties based on the prediction of meteorological conditions so as to increase crop yield. The strategy realizes the combination of the meteorological prediction model based on phase space reconstruction and exponential smoothing and the variety evaluation classification model based on Euclidean distance. In the experiment, the strategy accurately grasped the interactive relationship between soybean yield and meteorological conditions, which provided a scientific basis for soybean to make full use of meteorological resources, avoid disadvantages and increase yield. Not only can this strategy complete the selection of soybean varieties, but also it can be extended to other crops' seed selection, so it has broad application prospects for the selection of crop cultivated varieties in the agricultural field.

References

1. Li, C. X., Wang, Y. Z. & Wang, Y. Y. Relationship Between Yield Change of the Main Crops in Heilongjiang and the Climatic Factors. *JOURNAL OF JIXI UNIVERSITY*. (2):55-57 (2015).
2. Zhou, W. K. Impact of Climate Change Impact on Chinese Food Production and its Countermeasures. Nanjing Agricultural University (2012).
3. Qin, D.H. *et al.* Updated Understanding of Climate Change Sciences. *Advances in Climate Change Research*. 3(2):63-73 (2007).
4. Cheng, L., Liu R.H. & Ma, Z.H. Influence of global warming on winter wheat yield in Henan Province. *Chinese Journal of Eco-Agriculture*. 19(4):854-859 (2011).
5. Yu, W.D., Zhao, G.Q. & Chen, H.L. Impacts of Climate Change on Growing Stages of Main Crops in Henan Province. *Chinese Journal of Agrometeorology*. 28(1):9-12 (2007).
6. J.E. Olesen. *et al.* Impacts and adaptation of European crop production systems to climate change. *European Journal of Agronomy*, 34(2):96-112 (2011).
7. Lei, S.L. Effect of global climate change on spring wheat growth in Ningxia. *Chinese Journal of Agrometeorology*. 22(2):33-36 (2001).
8. Qin, Z.H., Tang, H.J. & Li, W.J. Front issues in studying the impacts of climate change on grain

- farming system in China. *Chinese Journal of Agricultural Resources and Regional Planning*. (2015).
9. Zhang, D.R. & Zhang, X.J. Study on injury of soybean. *Soybean science*. (2) (1998).
10. Zhang, M. N. *et al*, Modelling the impacts of future climate change on maize productivity in north-east China. *Journal of Arid Land Resources and Environment*. 26(8):117-123 (2012).
11. Li, X.H. Preliminary study on the relationship between climatic conditions and growth and yield of spring soybean. *Crop research*. (4):15-17 (1998).
12. Wang, J., Wang, E., Feng, L., Yin, H. & Yu, W. Phenological trends of winter wheat in response to varietal and temperature changes in the North China Plain. *Field Crops Research*, 144(6):135-144 (2013).
13. Pan, S.C. Analysis on the relationship between growth and development of soybean and meteorological conditions. *RURAI SCIENCE & THCHNOLOGY*. (6):40-41 (2017).
14. Zhou, W.Q. *et al*. Simulation study of DNDC model based on organic carbon in farmland and crop yield in the Western Songnen Plain. *ENVIRONMENT AND DEVELOPMENT*. 29(2):31-36 (2017).
15. Chen, C., Wang, E., Yu, Q.A. & Zhang, Y.Q. Quantifying the effects of climate trends in the past 43 years (1961-2003) on crop growth and water demand in the North China Plain. *Climatic Change*. 100(3-4):559-578 (2010).
16. Dong, C., Liu, Z. & Yang, X. Effects of different grade drought on grain yield of spring maize in Northern China. *Transactions of the Chinese Society of Agricultural Engineering*. 31(11):157-164 (2015).
17. Zhao, X.P. Biology Significance on G+E=P. *Yinshan Academic Journal (Natural Science Edition)*. (1):132-133 (2003).
18. Yan, J. & He, Z.H. Effects of Genotype, Environment and G× E Interaction on Starch Quality Traits of Wheat Grown in Yellow and Huai River Valley. *Journal of Triticeae Crops*. 21(2): 14-19 (2001).
19. Takens, F. On the numerical determination of the dimension of an attractor. *Lecture Notes in Mathematics*. 1125: 99—106 (1985).

20. Zhang, Z. P. Exponential Smoothing. (China Statistics Press,1996).
21. Wu, Z.L., Shi D.D., Lü, J.J., Yang, Y. & Wang, J. Evaluation of Heavy Rain Disaster in Haihe River Basin by Euclidean Distance Method. *Catastrophology*. 27(3):48-53 (2012).
22. Shang, H.*et al.* Relationship between yield characters of soybean and climatic conditions in Jilin. *Bulletin of Agricultural Science and Technology*. (2):47-49 (2012).
23. State Statistical Bureau. China Statistical Yearbook. (Beijing: China Statistics Press,2017)
24. Jiang, A.H., Zhou, P., Zhang, Y. & Hua, H.X. Research on base vibration of centrifugal pump by phase space reconstruction. *Transactions of the Chinese Society of Agricultural Engineering*. 30(2):56-62 (2014).
25. Zhang, Y.C. & Liu, Y.Z. Monthly precipitation prediction method of neural network based on phase space reconstruction. *Computer Simulation*. 31(1):352-355 (2014).
26. Wood, D.J. & Ormsbee, L.E. Supply identification for distribution systems. *Journal AWWA*, (7): 74-80 (1989).
27. Jr, L.V. An Experiment Comparing Double Exponential Smoothing and Kalman Filter-Based Predictive Tracking Algorithms//IEEE Virtual Reality. *IEEE Computer Society*. 2003:283 (2003).
28. Bejda, P. & Cipra, T., Exponential smoothing based on L-estimation. *Kybernetika -Praha-*. (6):973-993 (2015).
29. Laviola, J.J. Double exponential smoothing:an alternative to Kalman filter-based predictive tracking// The Workshop on Virtual Environments. *ACM*. 199-206 (2003).
30. Wang, C.J. Selection of Smoothing Coefficient via Exponential Smoothing Algorithm. *Journal of North University of China (Natural Science Edition)*. 27(6):558-561 (2006).
31. Shen, X.L., Chen, J.Y. & Wang, K. Forecast of China's Private Car Ownership Based on the Exponential Smoothing. *Journal of East China Jiaotong University*. (1):58-63 (2013).
32. Abarbanel, Henry D. I., Brown, R., Sidorowich, J.J. & Tsimring, Lev. Sh. The analysis of observed chaotic data in physical system. *Rev. Mod. Phy.* 65(4): 1331—1392 (1993).
33. Theiler, J. Estimating fractal dimension. *Journal of the optical society of America A*. 7(6): 1055—1073 (1990).

34. Grassberge, P. & Procaccia, I. Measuring the strangeness of strange attractors. *Physic Review D*. 9(1): 189–208 (1983).
35. Mathew, B.K., Brown, R. & Abarbanel, D.I.H. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physic Review A*. 45(6): 3403–3411 (1992).
36. Broomhead, D.S. & King, G.P., Extracting qualitative dynamics from experimental data. *Physical Review D*. 20(2): 217–236 (1986).
37. Li, X.Z., Yang, W.Y. & Liu, Z.J., C-C Average Method for Estimating the Delay Time and the Delay Time Window of Nonlinear Time Series Embedding. *Journal of Sichuan University* (2007).
38. Gao, J.B. & Zheng, Z.M., Local exponential divergence plot and optimal embedding of a chaotic time series. *Physics Letters A*. 181(2): 153–158 (1993).
39. Rosenstein, M.T., Collins, J.J. & DeLuca, C.J., A practical method for calculating largest Lyapunov exponents from small data. *Physica D*. 65:117-134 (1993).
40. Mao, R.F., Huang, S.Z. & Zhang, T.Y., Preliminary Study on the Multi-target Synthetical Parents Selecting Method for Crop Breeding. *Acta Agriculture Boreali-Sinica*. 8(3):44-50 (1993).
41. Liao, G.P. & Guan, C.Y., Rapeseed Cultivation Management Multimedia Expert System for High Quality, High Yield, and High Profit. *Acta Agronomica Sinica*. 28(1):140-142 (2002).
42. Yan, D.C., Zhu, Y. & Cao, W.X., A knowledge model for selection of suitable variety in rice production. *Journal of Nanjing Agricultural University*. 27(4):20-25 (2004).

Author Contributions

Rongsheng Zhu, Kai Sun put forward the strategy and wrote this paper. Xuehui Yan supplied the meteorological data weather data. Zhaoming Qi and Chunming Yang supplied the yield data of soybean. Dawei Xin, Hongwei Jiang, Zhanguo Zhang, Yang Li, Zhenbang Hu designed the experiment and modified this paper. Rongsheng Zhu, Kai Sun and Qingshan Chen modified this paper.

Additional Information

Competing Interests

The authors declare that they have no competing interests.

Data availability statement

The data involved in this paper are open.

The average temperature line chart of Harbin

The MSE values of different smoothness indices

The Euclidean distance figure of Harbin

a)

b)

The main character density distribution of soybean in Harbin

The Euclidean distance result schematic diagram of Harbin