The Influence of Climate Change on Soybean Yield: Evidence from major cultivation regions in northern China (1980-2020)

Zhe Kong[1] Shitong Sun [2]

[1]School of Environment, Nanjing University, Nanjing, China, 210033
[2]School of Public Administration and Policy, Dalian University of Technology, Dalian, China, 116024,
1448400631@qq.com

Abstract

Climate change has become a global challenge, significantly impacting soybean fields in China. Therefore, we propose a study utilising the long panel data model to investigate the impact in major cultivation regions in northern China from 1980 to 2020. We first propose controlling non-weather factors and capturing the correlation between yield and weather with the long panel model. Then, we assess the significance of different weather variables and eliminate the insignificant ones. Our ultimate goal is to identify optimal climatic conditions to enhance soybean yield in the northern cultivation regions of China and seek to assist various stakeholders in reinforcing existing strategies for climate change adaptation in soybean cultivation.

1 Introduction

When addressing global challenges, climate change stands out as a key player. Huber and Gulledge (2011) argue that global warming has intensified the occurrence of extreme weather events, including prolonged droughts, heatwaves and floods caused by heavy rainfall. Additionally, S. Chen, X. Chen, and Xu (2016) claim that extremely high temperatures seriously threaten crop growth, affecting both yield and quality, thus destabilising the food supply chain. As a result, understanding and mitigating the impact of climate change on global food production and the risk of natural disasters is crucial.

Global warming has significantly impacted China's soybean production. On one hand, S. Chen, X. Chen, and Xu (2016) observed a direct decline in soybean yields. High temperatures and unstable weather patterns affect the growth and development of soybean plants, leading to reduced output. Additionally, regions where soybeans are cultivated in China face challenges such as drought and water scarcity, further jeopardising production. On the other hand, research indicates that the yield of Chinese soybeans is particularly vulnerable to temperature constraints. X. Wang et al. (2015) estimated that the potential loss rate due to temperature constraints could reach as high as 51%.

Studying soybean production in China's northern cultivation regions holds significant importance. First, soybeans are a primary source of edible oil and a crucial livestock feed in China. Second, China heavily relies on imports to meet its domestic soybean demand. In 2021, China imported 99.13 million tons of soybeans, with an import dependency exceeding 85.5% (FAO Faostat-Agriculture, (2021)). Third, the soybean industry plays a profound role in ensuring the safety and well-being of China's 1.3 billion population while exerting a substantial influence on global grain and feed markets. Moreover, the northern cultivation regions, primarily in Inner Mongolia and the three northeastern provinces, constitute the core of China's soybean production, accounting for approximately 60% of the national total. However, Liu et al. (2020) argue that these regions are particularly susceptible to the fluctuations brought about by climate change.

We employ a long panel data model to investigate the impact of climate change on soybean production in China's northern primary cultivation regions from 1980 to 2020. The long panel data model offers several advantages over time series and short panel data models. First, the larger sample size facilitates a better analysis of the heterogeneity among the north's central cultivation provinces and cities, leading to more precise statistical inferences on the relationship between northern climate and soybean yields. Second, the 41-year time span allows the model to capture nonlinear temporal changes, account for time lags, and accommodate seasonal fluctuations more effectively. Elodie and Wolfram (2017) emphasise the importance of explaining and predicting the evolving trends in soybean production over time. Third, owing to its heightened time sensitivity, more extensive data set and more degrees of freedom, the long panel model is more adaptable in handling variable transformations and model specifications.

We employ a long panel data model to synthesise scientific data from various sources. We conduct four regression analyses, utilising maximum, minimum, average, and accumulated temperature as temperature variables in each analysis. The ultimate goal is to compare the results of these four regression analyses and identify the most significant weather variable.

We first employ a stepwise regression to assess the correlations between explanatory and control variables to eliminate the insignificant ones. The location fixed effect is introduced to capture the time-invariant confounding variation that may correlate with the climate. Then, we use three empirical tests to address the issues of autocorrelation, heteroscedasticity, and cross-sectional dependence in the long panel data. Second, we compare Panel-Corrected Standard Errors (PCSE) and Feasible Generalized Least Squares (FGLS) as methods to correct the issues mentioned above, aiming to select the most effective correction method. We also compare the regression results of introducing different temperature indexes to find the most significant one. Third, we use weather variables and yield mainly influenced by these variables to plot curves to determine the optimal weather variable values.

We plan to rely on datasets for temperature, precipitation, and yield. We obtain soybean yield per unit area data from the National Bureau of Statistics of China (1980-2020) and Statistical Yearbooks of the three northeastern provinces and the Inner Mongolia Autonomous Region (1980-2020). We also obtain temperature and precipitation data from the ERA5-Land dataset, publicly available through organisations such as the European Union and the European Centre for Medium-Range Weather Forecasts (ECMWF).

We aim to utilise models to identify the optimal climatic conditions to enhance soybean yield per unit. Extensive literature has explored the impact of climate change on staple crops in China, such as maise, rice, and wheat production. However, much less emphasis has been placed on the impact of climate change on soybean production in China despite the increasing significance of soybeans as a vital agricultural commodity in the country. On the other hand, existing research investigating the relationship between climate and soybean yield often relies on time series analysis or short panel data models. However, there is a lack of comprehensive investigation using long panel models that consider both temporal and spatial dimensions, especially in northern China. Moreover, only some research studies have presented optimal or suboptimal weather conditions

for soybean production under climate challenges. Our objective is to bridge the research gap regarding how climate change affects soybean production in northern China.

The rest of this research proposal is structured as follows. Section 2 further elucidates the underlying logic behind our investigation of optimal climatic conditions for soybean cultivation. Section 3 establishes a conceptual framework that outlines the theoretical models utilised in our study. This framework provides an initial elucidation of the relationships between variables included within the model. Section 4 outlines our data sources and summary statistics and briefly explains our key variables. In section 5, we expound on the model we construct and conduct specific analyses, followed by the presentation of results and discussions in section 6. Lastly, section 7 provides a summary and discussion of conclusions.

2 Literature review

Numerous studies have demonstrated the influence of climate change on major crop yields in China, including maise, rice, and wheat. For instance, Tao et al. (1981) utilised long panel regression models to analyse the period from 1981 to 2009. Their research revealed that changes in temperature, precipitation, and solar radiation resulted in an increase in wheat yield in the north and a decrease in the south. Furthermore, Zhang, Zhu, and Reiner (1981) found that rice yield exhibits a positive correlation with solar radiation. Besides, they observed that the interaction between rainfall and the effectiveness of irrigation water plays a crucial role in determining yield variations. Additionally, the impact of precipitation on crop yields (wheat, rice, maise) is more pronounced in northern China, showing weakening or complex effects in the southern regions. Notably, wheat yield displays the most distinct response to climate warming Liu et al. (2020). Various literature has employed diverse methodologies to systematically and comprehensively analyse these three primary crops.

Although a limited portion of the literature examines the influence of climate change on soybean yield in China, researchers predominantly choose statistical models such as time series models and short panel models. For instance, Jiang et al. (2011) conducted a time series analysis from 1980 to 2008 in Heilongjiang Province. Another study performed a regression analysis on 35 years

of climate change and soybean yield in Heilongjiang Province (Gong et al. (2019)). Furthermore, S. Chen, X. Chen, and Xu (2016) undertook a short panel data analysis covering 2000 to 2009 for corn and soybeans in China. Additionally, Guo et al. (2022) utilised crop models to analyse soybean yield in the three northeastern provinces of China.

However, time series and short panel models have many disadvantages. Time series models address individual changes, potentially overlooking inter-individual variations. Additionally, traditional time series models often need more data processing due to trends and seasonal effects between climate and soybean yield. On the other hand, short panel models have fewer individual observations, resulting in relatively small sample sizes that limit the analysis of individual heterogeneity and the precision of statistical inferences. Moreover, given their limited periods, short panel data might be less sensitive to temporal changes and struggle to capture nonlinear and intricate relationships over time (Elodie and Wolfram (2017)). These models often offer a holistic perspective but lack an in-depth understanding of the interactions among multiple variables. Such approaches may fail to accurately capture the complexities of soybean production, especially in longer timeframes and across regions. Therefore, the decision was made to employ a long-panel model.

Furthermore, prior studies have primarily presented simplistic correlations between climate and yield. Therefore, we introduce a long panel data model to more accurately analyse the impact of long-term climate changes on soybean production. Building upon the non-linear climate-yield relationships observed in previous studies, S. Chen, X. Chen, and Xu (2016) utilised a quadratic form to capture this non-linearity, ultimately confirming an inverted U-shaped relationship between yield and climate variables. This inspired us to employ the quadratic form for capturing the non-linearity between climate and yield, allowing us to identify the optimal or suboptimal states of soybean yield under different climatic conditions. In previous literature examining soybean yield, the weather variables commonly used have been precipitation and temperature. However, we argue that solar radiation sunshine duration, a more robust variable, should be included. As China lacks long-term solar radiation data, we have opted to utilise accumulated sunshine hours as one of our weather variables.

3 Conceptual framework

In our analysis, the per unit soybean yield Y_{it} in province i and year t can be expressed as:

$$Y_{it} = g(A_{it}) + h(B_{it}) + m(C_{it}) + n(D_{it}) + p(K_{it}) + f(E_i) \tag{1}$$

The first three terms $g(A_{it})$, $h(B_{it})$ and $m(C_{it})$ are the per unit yields affected by weather variables. The A_{it} is the temperature variable in province i and year t that should be one of four temperature indexes we select, which are maximum temperature, minimum temperature, average temperature and active accumulated temperature (AAT). We aim to choose the most significant one by comparison. The B_{it} is the total precipitation, and the C_{it} is the accumulated sunshine hours. These three terms should be calculated within the growing seasons of soybeans. We use the $g(A_{it})$, $h(B_{it})$ and $m(C_{it})$ to capture the potential nonlinear effects caused by weather factors which means per unit yield may go up to the peak and then down as the climate change. Our goal is to identify the optimal point that maximises the yield, referred to as the peak mentioned above.

The fourth term $n(D_{it})$ represents the per unit yield affected by the short-term adaptation of farmers D_{it} in province i and year t. The short-term adaptation of farmers is correlated with the weather variables. When confronted with awful weather, farmers will increase their inputs to alleviate the weather effect and maintain crop yields. For instance, farmers may increase the use of fertiliser or the amount of irrigation when the temperature is high or lacking in precipitation. We must take this input into account, or it may cause biased estimation.

Fifthly, we use $p(K_{it})$ to capture the long-term trend of yield, which is mainly associated with the unobserved socioeconomic factor K_{it}. The socioeconomic factor can influence the tendency of yield in the long run through technical and policy ways which have nothing to do with climate.

The last term $f(E_i)$ reflects how E_i, the time-invariant confounding variation in province i, influence the per unit yield. The time-invariant confounding variation can be the regional differences like soil quality that hardly change over time but have a significant impact on yields, and it may correlate with the climate. The time-invariant confounding variation, like soil quality, can decide the baseline of the per-unit yield. If we do not consider this, weather deviations from the mean might be correlated with baseline differences in space that could cause a spurious correlation.

To capture the optimal weather conditions and learn how the yield fluctuates as the climate changes, we use a quadratic form as $\alpha_1 A_t^2 + \alpha_2 A_{it}$. Relying on the coefficient, we can estimate the impacts of climate change.

To obtain accurate estimates of weather variables, we also need to figure out the correlation between the variables. For instance, places with less precipitation tend to have longer sunshine hours, areas with higher temperatures may lead to less precipitation, etc. The strong correlations between the variables may lead to the failure of our estimation.

4 Data

The study aims to utilise a long panel model for quantifying the impact of climate change, assessing the significance of different weather variables, and thereby capturing the optimal climatic conditions. This necessitates understanding how variables like maximum temperature, minimum temperature, average temperature, and accumulated temperature uniquely affect yield. Additionally, comprehending correlations among variables is vital during model establishment and refinement. To achieve this, we compile data from observed historical weather and annual government-published yield data.

4.1 Soybean per unit yield

Soybean per unit yield is the term that more accurately reflects the effects of climate change when the total yield and planted area are strongly influenced by socioeconomic impacts. Total soybean yield divided by total planted area equals the per unit yield. As all four provinces in our research almost plant spring soybean that grows from April to September, the figure we calculate is close to reality without the influence of summer soybean yield.

4.2 Temperature variables

In our model, we compare four temperature indexes and aim to find the most significant one. To achieve this, we will compare the regression results produced by severally introducing these four

indexes to the model. These indexes are maximum temperature, minimum temperature, average temperature, and active accumulated temperature (AAT) during the growing seasons of soybeans. We use AAT to represent the heat requirement of soybeans during the growing seasons. AAT sums up the daily active temperature (daily average temperature ≥10 °C) during the growing seasons of crops, which is a commonly used method in China. Wen et al. (2022) proved the effectiveness of AAT in the northwest region. AAT was calculated in the growing seasons of soybeans from 1980 to 2020 in each province with the following equation:

$$AAT = \Sigma_{d=1}^{n} T_d \, (If \;\; T_d > 10^{\circ}C, T_d = T_d, else \;\; T_d = 0) \tag{2}$$

Where T_d is the daily average temperature, 10°C is the biological zero or development threshold temperature, which means only above this temperature can creatures grow.

4.3 Total precipitation and accumulated sunshine hours

For simplicity, we sum up the precipitation and sunshine hours during the growing seasons of soybeans from 1980 to 2020 in each province (Nishio et al. (2019)). Our model intends to use these two factors to capture the impacts of precipitation and solar radiation on per-unit yield. A possible concern is that accumulated sunshine hours are not equivalent to solar radiation, which may cause estimation bias.

4.4 Short-term adaptation

Short-term adaptation is the adjustment of farmers' input. We use the per unit irrigation and fertiliser as the index of the farmers' input. These two terms are empirically related to climate change, while other inputs, like farm machinery, may have nothing to do with it. We can detect farmers' adjustment through the fluctuation of these two indexes. However, we only get the dataset of these two indexes from 2004 to 2020.

We obtained soybean yield data for 1980-2020 from the China National Bureau of Statistics and the Statistical Yearbooks of the three northeastern provinces and the Inner Mongolia Autonomous Region. Additionally, we source temperature and precipitation data from ERA5-Land, publicly

accessible through organisations like the European Union and the European Centre for Medium-Range Weather Forecasts (ECMWF). Within this dataset, we calculate growing degree days (GDD) by summing temperatures exceeding 10 degrees Celsius during the growing season.

Table 1: Table Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max	CV
Per Unit Yield (Kg/ha)	164	1749.213	497.019	660.000	3495.349	0.284
Total Precipitation (mm)	164	546.475	169.68	225.923	1010.939	0.310
AAT ($°C$)	164	2771.577	202.613	2365.952	3271.301	0.073
Minimum Temperature ($°C$)	164	8.297	1.696	5.080	11.946	0.204
Maximum Temperature ($°C$)	164	26.213	1.398	23.141	29.465	0.053
Average Temperature ($°C$)	164	16.161	1.386	13.578	19.250	0.086
Accumulated Sunshine Hours (h)	164	1187.744	110.735	932.808	1448.127	0.093

Notes: The numbers above are based on observations in the years 1980–2020 from four provinces

As shown in Table 1, the three variables with a larger coefficient of variation are per unit yield, total precipitation and minimum temperature. Per unit yield of soybean varied substantially, ranging from 660 to 3495.349 Kg per hectare, with an average of 1749.213 Kg per hectare, and its standard deviation reaches up to 497.019 Kg per hectare. Total precipitation ranges between 225.923 and 1010.939 mm with an average of 546.475 mm, while minimum temperature changes between 5.080 and 11.946 Celsius with an average of 8.297 Celsius.

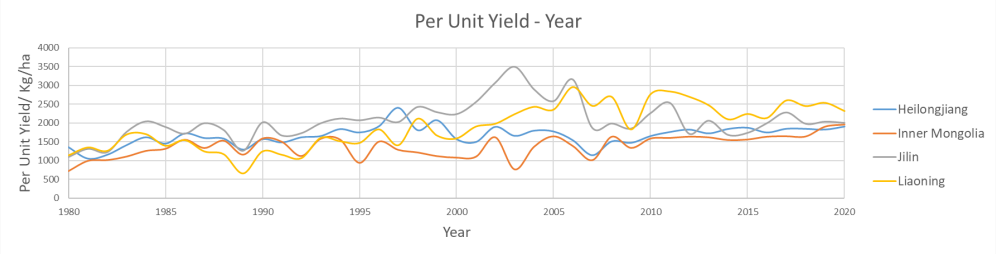

Figure 1. Yield Trends of Soybeans in Four Provinces from 1980 to 2020 [Owner-draw]

Generally, the soybean yield per hectare in the Inner Mongolia Autonomous Region is the lowest. Liaoning Province has a relatively higher soybean yield than other provinces, but its yield shows significant variability. Over the years, soybean production in Heilongjiang Province has remained consistently stable, hovering between 400 to 700 kilograms per hectare. Apart from

Heilongjiang, the soybean yields in the other three provinces are projected to experience an upward trend. The curves representing soybean yield per unit area in the four provinces exhibit fluctuations within a narrow range along a horizontal line, indicating a stable sequence.

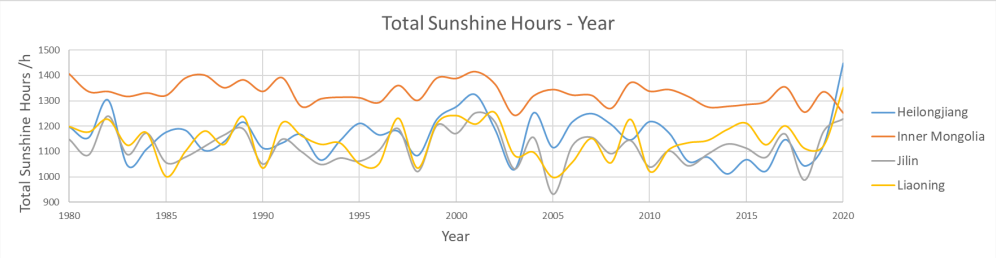

Figure 2. Trend chart of total sunshine duration from 1980 to 2020 in the four provinces [Owner-draw]

The total sunshine hours in the Inner Mongolia Autonomous Region are the highest among the four provinces. The entire sunshine duration in the three northeastern provinces has seen a significant increase after 2018, while Inner Mongolia, on the other hand, is experiencing a declining trend. Besides, the total sunshine hours data curves fluctuate within a narrow range along a horizontal line, indicating stable sequences with minimal amplitude variations.

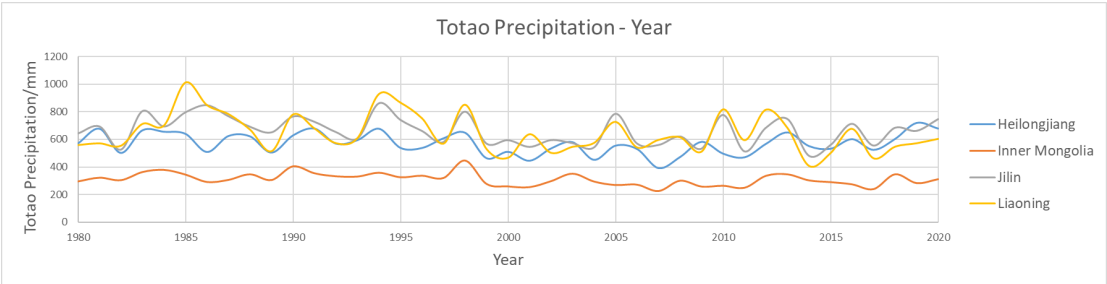

Figure 3. Line chart depicting the total precipitation trends from 1980 to 2020 in the four provinces [Owner-draw]

The curves depicting the total precipitation in the four provinces exhibit a similar pattern of variation. That indicates a stable sequence.

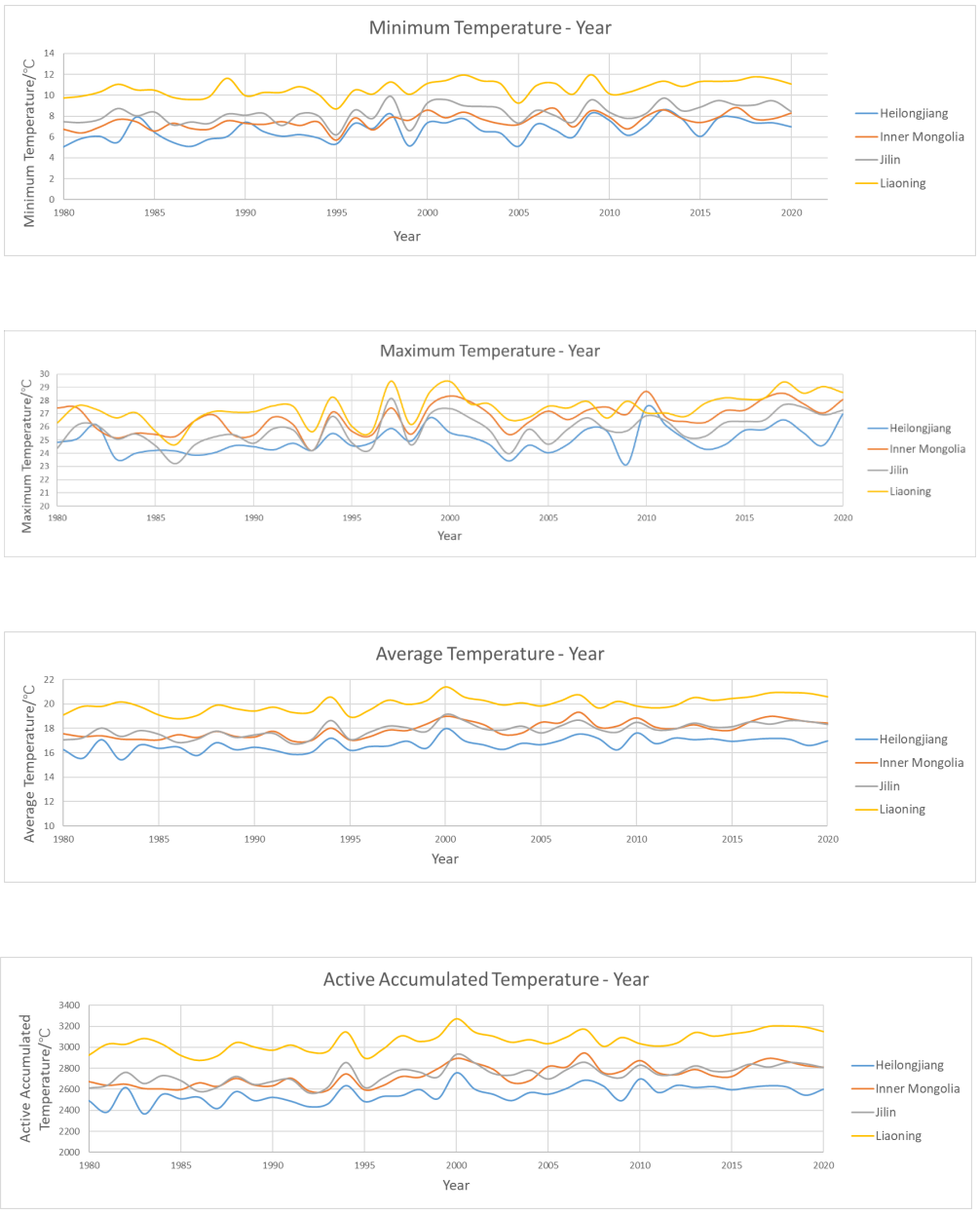

Figure 4. Trend Charts of Minimum, Maximum, Average Temperatures, and Active Accumulated Temperature in Four Provinces [Owner-draw]

The curves representing the minimum, maximum, average, and active accumulated temperature in the four provinces exhibit fluctuations within a narrow range along a horizontal line. The amplitude of these fluctuations shows minimal variation over time, suggesting a stable pattern. From all the plots above, our data is stationary.

5 Empirical approach, methods

Our empirical goal is to find optimal weather conditions. To fulfil this, we use the quadratic form of weather variables (Li and A (2022))) as follows:

$$Y(T_{it}) = \beta_1 T_{it} + \beta_2 T_t^2 \tag{3}$$

$$Y(P_{it}) = \beta_3 P_{it} + \beta_4 P_t^2 \tag{4}$$

$$Y(S_{it}) = \beta_5 S_{it} + \beta_6 S_t^2 \tag{5}$$

Where $Y()$ is per unit yield affected by climate change, T_{it} is one of the maximum temperature, minimum temperature, average temperature and active accumulated temperature (AAT) during the growing seasons of soybean in province i and year t, P_{it} is the total precipitation during the growing seasons of soybean in province i and year t, S_{it} is the accumulated sunshine hours during the growing seasons of soybean in province i and year t

We focus on capturing time-invariant confounding factors, such as soil quality and unrelated socioeconomic trends, for a more accurate estimation. To fulfil this, we incorporate adaptably functional forms like location fixed effects and the quadratic form of time.

In our approach, we apply the long panel data method to estimate the influence of climate change on yield. Actually, this method is a kind of OLS regression called the least squares dummy variable (LSDV) based on the panel dataset. Panel data consists of time series and cross-section data, which allows for more degrees of freedom and contains more information. Relying solely on time-series or cross-sectional data would inevitably lead to an insurmountable issue of neglecting essential time-invariant confounding variations. These variations may correlate with the independent and dependent variables we are interested in. It will undoubtedly lead to biased estimation. The good thing is that the panel data method allows for using fixed effects, which are dummy variables, to absorb all these time-invariant confounding variations. Because of these strengths, the regression coefficient of the panel method can be more precise.

Our long panel data model is similar to a one-way fixed effect model. A location-fixed effect is applied to fix the time-invariant confounding variations and a quadratic form of time to capture the yield trend. In our empirical long panel model, the impact of climate change on yield should be nonlinear and cumulative during the soybean growing seasons as the following (J. Wang et al. (2009)):

$$Y_{it} = \beta_1 T_{it} + \beta_2 T_{it}^2 + \beta_3 P_{it} + \beta_4 P_{it}^2 + \beta_5 S_{it} + \beta_6 S_{it}^2 + I_{it} + \theta_1 t + \theta_2 t^2 + \alpha_i + \epsilon_{it} \tag{6}$$

where I_{it} is the input of farmers, the control variable, t is the value obtained by subtracting 1979 from the year and the year in our research ranges from 1980 to 2020, α_i is the dummy variable that we introduce to the model as the location fixed effect capturing all confounding differences that do not vary over time (e.g., soil quality and other similar factors correlated with climate in different provinces that omitting these factors may bias the coefficients on weather variables), which helps to ensure the resulting weather deviations from the mean are not correlated with the differences in space that could cause a spurious correlation (Elodie and Wolfram (2017)), ϵ_{it} is the error We use the quadratic form of weather variables to capture potential nonlinear effects and $\theta_1 t + \theta_2 t^2$ to capture the trend.

Before conducting the regression, tests are needed for the correlations between independent variables. If significant correlations exist between our variables, we may have to find the principal component or do factor analysis to narrow the set of parameters. If there are no significant correlations, excluding controlled variables will not influence the other variables' coefficients, then these insignificant ones can be omitted to simplify the model.

Stepwise regression is used to analyse the correlation between independent variables. In this regression, we take per unit irrigation and amount of fertiliser as the explained variable and three types of weather variables as the explanatory variable. Firstly, we do a simple regression for each of the explained variables. Secondly, we choose the explanatory variable that contributes the most to the explained variable and use this regression equation as the foundation. Then, we introduce the remaining variables to this equation singly. Every time a new variable is introduced, we must test all the variables in the equation and eliminate those insignificant ones. At last, we will get the optimal regression model and learn whether a correlation exists between variables.

If we simply use this OLS regression method, we will overlook three main problems existing in the long panel model. The First is autocorrelation. The second is heteroscedasticity. And the third is a cross-sectional correlation. The omission of these three problems may cause the significance test to fail.

Our research has three empirical tests to check these three problems. For autocorrelation, we use the Wooldridge test, where the null hypothesis is that there is no first-order autocorrelation in our model and check the P value David (2003). For heteroskedasticity in the residuals of our one-way fixed effect model, we use a modified Wald statistic where our model is assumed to be homoskedasticity and check the P value W (2000), M and Hashem (2015). Lastly, we calculate the Breusch-Pagan statistic for cross-sectional independence in the residuals of our model, where we assume the error to be independent. We predict that the most likely problem should be heteroskedasticity.

Our research plans to use two ways to correct the possible problems and compare the results. The First is to calculate panel-corrected standard error (PCSE) estimates for our model, where OLS estimates the parameters. This method does not change the coefficients obtained by OLS regression but provides a corrected standard error. Because the OLS estimators are unbiased and consistent despite the existence of heteroskedasticity. In this way, we can solve the failure of the significance test and confidence interval caused by heteroskedasticity. The second way is using feasible generalised least squares (FGLS) to fit our model. It is a more efficient way if the error term is correctly handled.

We intend to compare the results of using these two methods and find the better one for later analysis. We also plan to reach the regression results by severally introducing four different temperature indexes to our model, and we can find the most significant one. With the coefficients of weather variables, we can quantitatively calculate the change in per unit yield caused by climate change with Eq. (3)(4) Elodie and Wolfram (2017). We plan to draw the conic with these equations, analyse the developing trend of yield and find the optimal value of three weather variables.

6 Results

6.1 Correlation results

Our first step is to examine the correlation between independent variables and find if there exist statistically significant correlations. We use the SPSS software to do the stepwise regression. We use per unit irrigation and fertiliser as the explained variable and use the four temperature variables, total precipitation and accumulated sunshine hours as the explanatory variable. The stepwise regression finally eliminates all the explanatory variables, indicating no significant correlation between them. As shown in Table 1, the test shows that correlations between those variables are statistically insignificant, but we can still harvest important information from Pearson correlation. For instance, as the increase of AAT and sunshine hours or decrease of precipitation, per unit irrigation may decrease. When AAT or sunshine hours are high, precipitation may be low. These relations between irrigation and weather variables or within weather variables align with what we expect, although it is insignificant.

Table 2: Table Test for the correlation between variables[Owner-draw]

Variable		1	2	3	4	5	6	7	8
1 Irrigation		1							
2 Fertilizer		0.517	1						
3 TMIN		0.263	0.025	1					
4 TMAX		-0.585	-0.195	-0.408	1				
5 TMEAN		0.346	0.185	0.372	0.248	1			
6 AAT		0.334	0.182	0.372	0.261	1.000**	1		
7 Precipitation		-0.488	-0.575	0.156	-0.22	-0.679	-0.673	1	
8 Sunshine hours		0.146	0.695	0.456	-0.18	0.341	0.346	-0.256	1

Note: ** Correlation is significant at the 0.01 level (2-tailed).

Since the input of farmers has no significant correlation with the weather variables, we can simplify Eq. (6) to the following form:

$$Y_{it} = \beta_1 T_{it} + \beta_2 T_{it}^2 + \beta_3 P_{it} + \beta_4 P_{it}^2 + \beta_5 S_{it} + \beta_6 S_{it}^2 + \theta_1 t + \theta_2 t^2 + \alpha_i + \epsilon_{it} \tag{7}$$

Where we eliminate the control variable I_{it}, which is the farmers' input.

6.2 Regression results

As shown in Table 2, the regression result is the best when we use minimum temperature as the temperature index since the weather variables are more significant and their P values are smaller. After comparing all the regression results, we chose the minimum temperature within the soybean growing seasons as the most significant and influential temperature index to the yield. Moreover, we find the P values of precipitation and temperature variables are much smaller than sunshine hours, meaning these two variables are incredibly significant in regression. It is evident that the sunshine hours term is less substantial in all regression results. Hence, we omit the influence of accumulated sunshine hours in the following analysis and only maintain the precipitation and temperature variables.

From our perspective, it is the geographic position of these provinces that led to the minimum temperature having more significant effects. These four provinces all have high latitudes, so the average annual temperature in each area is relatively low. The highest temperature recorded in these four provinces for over forty years was less than thirty degrees Celsius. Hence, the fluctuation in high temperatures only slightly affects per unit yield. In this situation, the impact of low temperature could be more significant, and the minimum temperature fluctuation may cause a more remarkable change in per unit yield.

Moreover, we use PCSE and FGLS separately to correct the heteroskedasticity in each OLS regression model in Table 2. Comparing the PCSE and FGLS methods, we find FGLS to be a more efficient way of correcting the heteroskedasticity. For instance, when choosing minimum temperature as the index, we see the P values of precipitation and temperature are lower in FGLS when neglecting the sunshine hours for their insignificance. Therefore, we consider FGLS a better way to correct heteroskedasticity.

Table 3: Regression Results

	TMin OLS	TMin PCSE	TMin FGLS	TMax OLS	TMax PCSE	TMax FGLS	TMean OLS	TMean PCSE	TMean FGLS	AAT OLS	AAT PCSE	AAT FGLS
P	3.97**	3.97**	4.05***	3.79**	3.79**	3.18**	3.91**	3.91**	3.99***	3.86**	3.86**	3.94***
	(0.024)	(0.011)	(0.005)	(0.037)	(0.025)	(0.031)	(0.031)	(0.016)	(0.008)	(0.034)	(0.018)	(0.009)
SquareP	-0.00**	-0.00**	-0.00***	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**
	(0.023)	(0.013)	(0.009)	(0.026)	(0.021)	(0.029)	(0.027)	(0.016)	(0.010)	(0.029)	(0.017)	(0.011)
TEMP	-443.52**	-443.52**	-428.95***	-390.86	-390.86	272.59	-1433.23**	-1433.23**	-1240.77**	-9.48**	-9.48**	-8.22**
	(0.018)	(0.010)	(0.009)	(0.556)	(0.567)	(0.630)	(0.044)	(0.032)	(0.047)	(0.042)	(0.031)	(0.046)
SquareTEMP	27.61**	27.61**	26.80**	6.94	6.94	-5.22	40.38**	40.38**	35.05**	0.00**	0.00**	0.00**
	(0.017)	(0.014)	(0.013)	(0.583)	(0.594)	(0.630)	(0.037)	(0.028)	(0.043)	(0.036)	(0.028)	(0.043)
S	-9.64*	-9.64*	-6.16	-8.74	-8.74	-6.34	-9.82*	-9.82*	-6.36	-9.82*	-9.82*	-6.36
	(0.089)	(0.066)	(0.193)	(0.129)	(0.107)	(0.143)	(0.085)	(0.063)	(0.187)	(0.085)	(0.063)	(0.187)
SquareS	0.00	0.00*	0.00	0.00	0.00	0.00	0.00	0.00*	0.00	0.00	0.00*	0.00
	(0.115)	(0.082)	(0.217)	(0.152)	(0.117)	(0.163)	(0.110)	(0.079)	(0.212)	(0.110)	(0.079)	(0.212)
t	47.39***	47.39***	38.66***	46.61***	46.61***	34.74***	47.06***	47.06***	37.32***	47.10***	47.10***	37.31***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Squaret	-0.66***	-0.66***	-0.49***	-0.62***	-0.62***	-0.39**	-0.67***	-0.67***	-0.48**	-0.67***	-0.67***	-0.48**
	(0.003)	(0.002)	(0.008)	(0.007)	(0.005)	(0.050)	(0.003)	(0.001)	(0.011)	(0.003)	(0.001)	(0.012)
N	164	164	164	164	164	164	164	164	164	164	164	164

***p*-values in parentheses** $^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Plug the estimated coefficients into the Eq. (2)(3)(4) and we get following equations:

$$Y(T_{it}) = -428.9474T_{it} + 26.8016T_t^2 \tag{8}$$

$$Y(P_{it}) = 4.0479P_{it} - 0.0030P_t^2 \tag{9}$$

Estimated coefficients indicate an inverted U-shaped relationship between precipitation and yield and a U-shaped relationship between minimum temperature and yield. This result is partly opposite to our anticipation, as we expect each relationship to be an inverted U-shape.

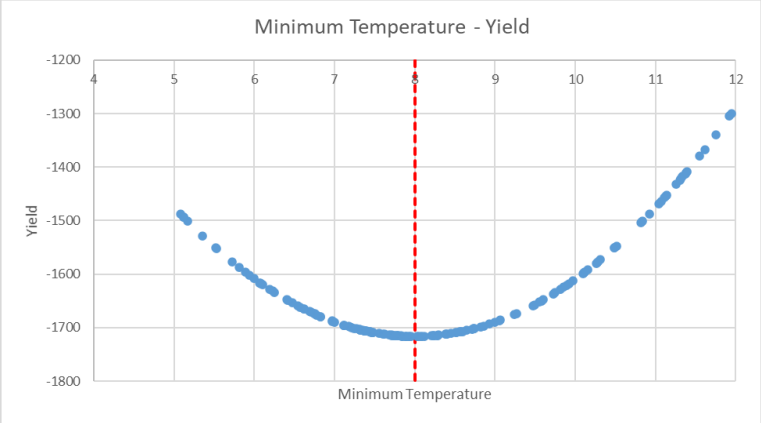

Figure 1: The impact of minimum temperature on yield : The red vertical line is the axis of symmetry of the conic, the horizontal coordinate of which is 8.00[Owner-draw]

Figure 5 shows the relationship between minimum temperature and yield; the sub-optimal temperature might be 8 Celsius. We can see that the yield goes down to the bottom, where the temperature is 8 Celsius and then rises. Undoubtedly, the yield goes up as the minimum temperature increases since low temperature limits the growth of soybeans, but the biggest problem is why the yield first decreases. Possible explanations might be some mistakes in our model that we might change the quadratic form of temperature variables, or it may be because all these minimum temperatures below the optimal 8 Celsius are also below the biological zero of soybeans, under which soybeans will stop growing. As a result, these temperatures are all harmful to the growth of soybeans for whatever reason. Nevertheless, at least we know that raising the minimum temperature above 8 Celsius can increase soybean yield. Therefore, farmers had better take suitable steps to alleviate the damage to soybeans from low temperatures. For instance, building a greenhouse to keep temperatures above 8 Celsius can be a feasible way.

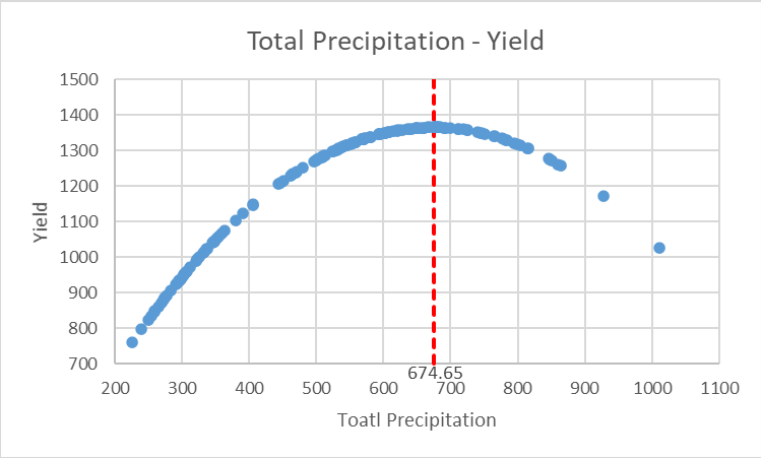

Figure 2: The impact of total precipitation on yield : The red vertical line is the axis of symmetry of the conic, the horizontal coordinate of which is 674.65[Owner-draw]

Figure 6 shows the impact of total precipitation on yield. As we can see, the yield goes up to the peak, where the precipitation is 674.65 mm, and then decreases. Although soybean is a drought-tolerant crop, it also needs much water in growing seasons, so the appropriate increase in precipitation will promote soybeans' growth and yield. However, excessive precipitation will also cause adverse effects on soybean roots, which will stunt growth and reduce yield. According to this, the optimal precipitation during soybean growing seasons is around 674.75mm. We suggest that farmers had better increase irrigation to compensate for the deficiency in precipitation when the weather is dry, and they should build complete agricultural infrastructures such as drainage systems to cope with excessive precipitation.

7 Conclusion

There is growing evidence that climate change has influenced crop growth and agricultural development. To ascertain the optimal climatic conditions for enhancing soybean yield in northern China, we employ a long panel data model to analyse the four provinces over 41 years.

To analyse the impacts of climate change on per unit yield, we propose using a long panel data method with data collected from the National Bureau of Statistics of China (1981-2020), Statistical Yearbooks of the four provinces (1981-2020) and the ERA5-Land dataset. First, we examine the correlation between explanatory variables and eliminate the ones uncorrelated with climate change

we initially tended to control. Location fixed effect is also introduced to capture the time-invariant confounding variation, which may correlate with the climate. Second, we test three main problems common in the panel dataset and use PCSE and FGLS separately to correct the heteroscedasticity we find. We also severally introduce four temperature variables to our model. Third, by comparing the regression results, we choose the minimum temperature as the most significant temperature index and FGLS as the most effective way to correct the problem. Finally, we propose the sub-optimal temperature and optimal precipitation and put forward some possible countermeasures for farmers.

We expect our findings to help farmers find the optimal or suboptimal climate conditions and take timely and appropriate steps to mitigate the adverse effects of climate change and increase per unit soybean yield. Given the growing significance of soybeans in China, we aim to make up for the deficiency in research on how climate change influences the soybean yield in northern China.

Our research still has a few limitations. First, our dataset needs to be bigger so that we can be more detailed and precise, like extending the time frame or subdividing four provinces into counties. Second, the growing seasons of soybeans may differ from different places in different years, so we can only use an approximate time frame. It will tremendously improve accuracy and reliability if we get detailed data for each day in growing seasons. Third, collecting precise data on inputs is challenging work. If we find more accurate data, we might get a deeper insight into the correlation between farmers' input and climate change. Future work will overcome these shortcomings and promote further research.

Our next steps include conducting a sensitivity analysis and quantifying the agricultural economic loss triggered by climate change. We may continue to use per unit yield as the explained variable and calculate the financial loss combined with agricultural product price.

References

Chen, Shuai, Xiaoguang Chen, and Jintao Xu (2016). "Impacts of climate change on agriculture: Evidence from China". In: *Journal of Environmental Economics and Management* 76, pp. 105–124.

David, Drukker (2003). “Testing for Serial Correlation in Linear Panel-Data Models”. In: *The Stata Journal: Promoting communications on statistics and Stata* 3 (2), pp. 168–177. DOI: 10.1177/1536867x0300300206.

Elodie, Blanc and Schlenker Wolfram (2017). “The Use of Panel Models in Assessments of Climate Impacts on Agriculture”. In: *Review of Environmental Economics and Policy* 11 (2), pp. 258–279. DOI: 10.1093/reep/rex016.

Gong, Lijuan et al. (2019). “Optimal Meteorological Indices During the Growing Season of Soybean in Heilongjiang Province”. In: *Soybean Science* 03, pp. 391–398.

Guo, Shibo et al. (2022). “The Possible Effects of Global Warming on Cropping Systems in China XIV. Distribution of High-Stable-Yield Zones and Agro-Meteorological Disasters of Soybean in Northeast China”. In: *Scientia Agricultura Sinica* 09, pp. 1763–1780.

Huber, DG and J Gulledge (2011). “Extreme weather & climate change: understanding the link and managing the risk. Report”. In: *Center for Climate and Energy Solutions. Available at: https://www. c2es. org/document/extreme-weather-and-climate-change*.

Jiang, Lixia et al. (2011). “Impacts of Climate Change on Development and Yield of Soybean over Past 30 Years in Heilongjiang Province”. In: *Soybean Science* 6, pp. 921–926.

Li, Z and Ortiz-Bobea A (2022). “On the timing of relevant weather conditions in agriculture”. In: *Journal of the Agricultural and Applied Economics Association* 1, pp. 180–195.

Liu, Yuan et al. (2020). “The central trend in crop yields under climate change in China: A systematic review”. In: *Science of the Total Environment* 704, p. 135355.

M, Pesaran and Hashem (2015). “Testing Weak Cross-Sectional Dependence in Large Panels’”. In: *Econometric Reviews* 34 (6-10), pp. 1089–1117. DOI: 10.1080/07474938.2014.956623.

Tao, Fulu et al. (1981). “Responses of Wheat Growth and Yield to Climate Change in Different Climate Zones of China”. In: *Agricultural and Forest Meteorology* 189, pp. 91–104. DOI: 10.1016/j.agrformet.2014.01.013.

W, Greene (2000). “Econometric Analysis”. In.

Wang, Jinxia et al. (2009). “The Impact of Climate Change on China’s Agriculture’”. In: *Agricultural Economics* 40 (3), pp. 323–337. DOI: 10.1111/j.1574-0862.2009.00379.x.

Wang, XY et al. (2015). “Comparison of potential yield and resource utilization efficiency of main food crops in three provinces of Northeast China under climate change”. In: *Ying yong sheng tai xue bao= The journal of applied ecology* 26.10, pp. 3091–3102.

Wen, Huiwen et al. (2022). “Analysis of Relationship between Soybean Relative Maturity Group, Crop Heat Units and ≥10 °C Active Accumulated Temperature”. In: *Agronomy* 12 (6), p. 1444. DOI: 10.3390/agronomy12061444.

Zhang, Tianyi, Jiang Zhu, and Wassmann Reiner (1981). “Responses of Rice Yields to Recent Climate Change in China: An Empirical Assessment Based on Long-Term Observations at Different Spatial Scales”. In: *Agricultural and Forest Meteorology* 150 (7), pp. 1128–1137. DOI: 10.1016/j.agrformet.2010.04.013.