Reply to the referee comments on "Extreme droughts/floods and their impacts on harvest derived from historical documents in Eastern China during 801–1910" by Zhixin Hao et al

Dear editors and reviewers,

Thank you for your valuable comments and thoughtful suggestions on our manuscript. Following your comments on the manuscript, we made careful revisions, and the point-to-point response of the comments is listed below. We hope these revisions would make this manuscript more acceptable for publication. Please feel free to contact me if you have any questions.

Many thanks again. With best wishes.

Sincerely yours,

Jingyun Zheng

Anonymous Referee #1

Using databases of Song, Ming, and Qing documents, this paper finds that the frequency of reports of extreme droughts (but not always floods) correlates with reductions in harvests as reported in historical sources. On this basis, the authors conclude that there are clear historical periods when droughts reduced harvests, and therefore that these events had significant societal impacts. The sources and methods used in this manuscript appear to be standard in other publications on Chinese historical climatology. In this instance, however, I am not convinced they are adequate to prove the authors' conclusion. The problems concern, first, the author's use of their historical databases; second, the large temporal and spatial scale of the study; and third, the interpretation given to the pattern of correlations found.

Problems in use of databases:

The authors' use of databases of flood and drought events and harvest grades raises numerous questions which must be answered before it is clear whether or not the correlations identified are valid:

1) What kinds of droughts are recorded in the historical sources: meteorological drought? hydrological drought? agricultural drought? or some combination of these? Were observers more likely to report precisely those droughts that affected crops, or did they report all droughts equally?

Accepted and revised. Most of droughts recorded in historical documents were events due to no or less precipitation, and could be regarded as meteorological droughts. Therefore it's reasonable to suggest that they reported all droughts equally rather than those affected crops. (P3, L23-28)

From these samples, it's easy to know that the droughts recorded in the historical documents were usually focused on the events due to no or less precipitation in a period at first, thus could be regarded as meteorological drought rather than hydrological drought or agricultural drought, although some records also reported the impacts on hydrosphere (e.g., dry up for river) or agriculture (e.g. wilting for crops). Therefore, the droughts recorded in the historical documents appear to report all droughts equally rather than those affected crops only.

2) What kinds of floods are recorded in the historical sources: heavy rains? tsunamis? rivers that burst their banks? Does the database control for ongoing problems related to river hydrology? How do major events such as course changes in the Yellow River figure into the measure of flood frequency: as one flood? as many?

Accepted and revised. Similar to the records on drought, flood recorded in historical documents were mostly about more rains or heavy rains. River hydrology events were not taken into account unless it was resulted from more precipitation or heavy rain. (P3, L28-30)

Similarly, the floods recorded in the historical documents could be regarded as more rains or heavy rains rather than the river that burst their banks or tsunamis, although some records also reported the impacts on the overflowing or bursting for lakes or river due to more precipitation or heavy rain.

3) What is the seasonality of the meteorological events recorded in the historical sources? Does the seasonality of floods or droughts necessarily overlap with the seasonality of critical agricultural activities or phases of crop growth?

Accepted and revised. The 63-stations annual drought/flood grades was reconstructed and calibrated with descriptions on duration, intensity, and area of the drought/flood events in wet season, which was usually May to September in the study area and overlaps with the critical agricultural activities and phases of crop growth. (P3, L32-P4, L3)

The grades were classified using the ideal frequency criteria of 10% (grade 1, severe drought), 20% (grade 2, drought), 40% (grade 3, normal), 20% (grade 4, flood), and 10% (grade 5, heavy flood) for whole area and all time, which were calibrated with the descriptions on duration, intensity, and area of the drought/flood events in wet season (usually May to September), and its impact (Table

S1). Thus, the season for the drought/flood grade data overlaps with the critical agricultural activities and phases of crop growth.

4) What is being measured by "harvest"? Yield per seed? Total yield per hectare? Food availability?

Accepted and revised. Harvest in records was a relative concept and represented the ratio of actual yield compared to the possible maximum yield. (P6, L3-5)

In Chinese historical documents, the yearly harvest was usually recorded as a relative level compared to expected maximum yield rather than the individual crop production per hectare, although some records also reported the impacts on food availability, tax remissions, people's livelihoods, etc.

5) Are degrees of flood, drought, and harvest based entirely on narrative descriptions, or are there objective phenological or quantitative measures to help define them?

Accepted and revised. The degrees of flood, drought, and harvest was not based entirely on narrative descriptions. The duration, intensity, and area of these events and their impacts were adopted in calibration, as well. We added two supplementary tables (Table S1, S2) to explain the detailed criteria in grading drought/flood (P3, L31-P4, L2; Table S1) and harvest (P6, L14-17; Table S2). The supplementary materials are attached at the end of this text.

Based on these records, Zhang (1996) reconstructed the dataset of 63-stations annual drought/flood grades from 137 BCE, in which each station was set as a local area consisted of about 20 counties with the same climate. The grades were classified using the ideal frequency criteria of 10% (grade 1, severe drought), 20% (grade 2, drought), 40% (grade 3, normal), 20% (grade 4, flood), and 10% (grade 5, heavy flood) for whole area and all time, which were calibrated with the descriptions on duration, intensity, and area of the drought/flood events in wet season (usually May to September), and its impact (Table S1).

In the dataset, the levels of yearly harvest were classified into 6 grades: 1-Very poor, 2-Poor, 3-Slightly poor, 4-Average, 5-Near bumper, 6-Bumper. The criteria and methods for grading the documentary records (i.e., grain yield descriptions and the related information) year by year were presented by Su et al. (2014) and summarized by Yin et al. (2015), in which the classification of the yearly harvest grade and descriptions recorded in historical documents was shown in Table S2.

Regarding the temporal and spatial scale of the study, I am concerned that it relies on improbable assumptions of continuity and homogeneity in Chinese population, land use, and record keeping. In order to accept as valid any long-term correlations between reported drought or flood frequency and "Chinese" or even "regional" "harvests" I would need the authors to address the following issues:

1) How do the data control for the changing borders of Chinese empires? A priori, I would expect vastly different vulnerabilities and patterns of reporting between the Northern Song and Southern Song periods, simply based on the major geographical shifts in population and wealth between those two dynasties.

Accepted and revised. For droughts and floods, the historical records was transformed into graded data based on 63-stations, each of which was set as a local area consisted of about 20 counties and does not change in different dynasties (P3, L31-32). Although the available graded data was unevenly distributed spatially for different dynasties, it had been proved in the paper of Hao et al. (2010a) that the extreme drought/flood years recognized were mostly robust despite of the percentage of data-missing stations (P5, L5-12). As for harvest, the impact of changing borders on harvest grade should also be limited since the main grain product area had been relatively stable in the study period and the records in documents was about relative harvest rather than absolute yield as suggested by Yin et al. (2015) (P6, L30-33).

Based on these records, Zhang (1996) reconstructed the dataset of 63-stations annual drought/flood grades from 137 BCE, in which each station was set as a local area consisted of about 20 counties with the same climate.

To verify the rationality of this method and criteria, the validation was conducted in the paper of Hao et al. (2010a) based on 10 extreme events identified according to the threshold with probabilities of 10% and 90% occurrences which was based on the series of precipitation observation in each sub-regions. In the validation, all or part of grade 3 stations were omitted deliberately, and only 40% or 60% stations with disaster or extreme grade was reserved without changing the drought-to-flood ratio within the available data. The results showed that the years of extreme drought and extreme flood identified by this method and criteria using grade data well matched with those extreme events identified by the precipitation data, except for one mismatched event of extreme flood that occurred in 1958 in North China Plain, which demonstrated that the method and criteria was reasonable.

However, these social factors should have limited influence on the yearly harvest documents since that the main grain product area, the staple crop, and the cropping system have been relatively *stable in the study period, and besides the records in documents was relative harvest rather than absolute yield (Yin et al., 2015).*

2) How do data on "harvests" control for changes in staple crops, introduction of New World crops including peanuts and sweet potatoes, changing cropping patterns, and the increasing commercial orientation of agriculture?

Accepted and revised. The records for harvests were usually about relative percentage compared to expected maximum yield rather than absolute yield, and thus it should not be influenced by these factors. (P6, L3-7)

In Chinese historical documents, the yearly harvest was usually recorded as a relative level compared to expected maximum yield rather than the individual crop production per hectare, although some records also reported the impacts on food availability, tax remissions, people's livelihoods, etc. Therefore these harvest records excluded the differences on absolute yield between sub-regions with different climate, soil fertility, crop variety, etc., and the difference between historical periods with changing agricultural centre, farming technology, staple crops (Su et al., 2014).

3) How do the data deal with the changing vulnerabilities to climate variability based on changing settlement patterns even within regions (e.g., uplands in the south and southwest colonized by Han settlers during the late Ming and Qing periods)?

Accepted and revised. The graded harvest data represents a nationwide status and since the main grain product area had been relatively stable in the study period, the expansion of agricultural area should have limited influence on the yearly harvest documents. (P6, L27-33)

During the study period, there did exist several social factors that could influence the total yield in China, such as the changing borders of empires, the expansion of agricultural area (e.g., uplands in the south and southwest colonized by Han settlers during the late Ming and Qing periods), the updated crop varieties due to the introduction of New World crops (e.g., peanuts and sweet potatoes), the advanced agricultural management technology, and so on. However, these social factors should have limited influence on the yearly harvest documents since that the main grain product area, the staple crop, and the cropping system have been relatively stable in the study period, and besides the records in documents was relative harvest rather than absolute yield (Yin et al., 2015).

4) Given the very long time period examined here, wouldn't we expect new adaptations to reduce vulnerabilities to predictable climate variability and disasters?

Accepted and revised. This question has also been addressed in the revisions responding to above questions. (P6, L3-7, L27-33).

In Chinese historical documents, the yearly harvest was usually recorded as a relative level compared to expected maximum yield rather than the individual crop production per hectare, although some records also reported the impacts on food availability, tax remissions, people's livelihoods, etc. Therefore these harvest records excluded the differences on absolute yield between sub-regions with different climate, soil fertility, crop variety, etc., and the difference between historical periods with changing agricultural centre, farming technology, staple crops (Su et al., 2014).

During the study period, there did exist several social factors that could influence the total yield in China, such as the changing borders of empires, the expansion of agricultural area (e.g., uplands in the south and southwest colonized by Han settlers during the late Ming and Qing periods), the updated crop varieties due to the introduction of New World crops (e.g., peanuts and sweet potatoes), the advanced agricultural management technology, and so on. However, these social factors should have limited influence on the yearly harvest documents since that the main grain product area, the staple crop, and the cropping system have been relatively stable in the study period, and besides the records in documents was relative harvest rather than absolute yield (Yin et al., 2015).

5) Most importantly, how can we make up for the fact there are simply more records from the Qing period than earlier periods? I don't see that the methods used in this manuscript avoid the problem that more records will create a misimpression of a greater frequency of floods and droughts. The authors propose to ignore reports of "average" conditions in Qing records to make them more comparable to Song and Ming records. However, that would only work if the Song and Ming records still reliably reported all disasters and extremes and only left out "average" conditions. I don't see any reason to make that assumption. Perhaps the authors could experiment with methods of introducing "noise" into the data in order to reflect the events missing from the reports. Or else they could employ a Bayesian method to indicate that the presence or absence of certain descriptions in the records may be used to obtain updated posterior probabilities of actual conditions, without ever assuming that the records provide a complete account of events. In any case, the authors must come up with a way to handle these changes in the documentary record over time if they are to make a convincing case for stable

long-term correlations between floods and droughts and harvests.

Accepted and revised. The method of ignoring "average" conditions is based on the hypothesis that the records on droughts and floods were omitted randomly and unbiased, which suggests that the relative drought-to-flood ratio in the available data would be close to that in actual history. As in the abovementioned revisions, the recognition for extreme drought and flood years was still effective even if 40% or 60% of the available data with disasters was omitted deliberately. (P5, L1-15)

This is because that the probabilities for omitting drought and flood records were random and unbiased, despite of the increasing occurrences of missing data back in time; i.e. if there's one period with a large number of documents, it would be rich in both drought and flood records, and vice versa. Therefore the amount of missing data should not have a significant effect on the relative drought-to-flood ratio within the available data compared with that in actual history. To verify the rationality of this method and criteria, the validation was conducted in the paper of Hao et al. (2010a) based on 10 extreme events identified according to the threshold with probabilities of 10% and 90% occurrences which was based on the series of precipitation observation in each sub-regions. In the validation, all or part of grade 3 stations were omitted deliberately, and only 40% or 60% stations with disaster or extreme grade was reserved without changing the drought-to-flood ratio within the available data. The results showed that the years of extreme drought and extreme flood identified by this method and criteria using grade data well matched with those extreme events identified by the precipitation data, except for one mismatched event of extreme flood that occurred in 1958 in North China Plain, which demonstrated that the method and criteria was reasonable. This is because that the precipitation variability in eastern China is dominated by the east Asian summer monsoon (EASM), thus when an extreme drought or flood event occurs, the precipitation variation for stations within each sub-region usually share a similar relative magnitude.

Third, even if the correlations found in the study are valid, there is a problem with the authors' historical interpretation of them. The correlations discovered here are not between climate and harvests, but rather reports of floods and droughts and reported harvests. The authors assume that the correlations mean that floods and droughts reduced harvests. However, there are a number of potentially confounding variables, which indicate other potential pathways of causality and therefore other historical possibilities:

1. Drought and/or flood might have correlated with other climate variables (such as temperature) that caused harvest failure.

Accepted. As elaborated in previous study, the relationship between temperature and harvest had been investigated by Yin et al. (2015, 2016), which suggested that there would be better harvest in warm climate. And our study, in section 3.2 of the original manuscript, found that more occurrence of extreme drought in eastern China could lead to significant increase of frequency of poor harvest (grade 1+2) compared with non-extreme years. To further examine whether the drought and/or flood are correlated with temperature change, and if so, how the drought and/or flood are correlated with harvest failure under different temperature backgrounds, we presented a study in section 3.3, and found that there were slightly more extreme droughts in the warm period. However, the connection between extreme droughts and poor harvest was not significantly close in the warm epoch, while it was more significant in the cold epoch. These results suggested that warm period could weaken the impact of extreme drought on poor harvest during historical times. (P11, L21-25; P12, L3-10)

As found in section 3.2, more occurrence of extreme drought in eastern China could lead to significant increase of frequency of poor harvest (grade 1+2) compared with non-extreme years. Since there were more extreme droughts occurred over eastern China in 920–1300 compared to 1310–1880, the harvest in warm epoch should be worse than that in cold epoch. However, as found by Yin et al. (2015, 2016), the harvest in warm epoch was better than that in cold epoch. This implicated that there should be different effects of regional extreme drought on grain harvest between warm and cold epochs.

The results showed that, during the warm epoch of 920–1300, the connection between the occurrence of poor harvest and regional extreme drought was not significantly close, though the frequency of poor harvest in extreme drought years was still slightly higher than that in non-extreme years for each sub-region. By contrast, during the cold epoch of 1310–1880, the frequency of poor harvest in extreme drought years was significantly higher than that in non-extreme years, which indicated that the connection between the occurrence of poor harvest and extreme drought was still significant. Moreover, similar characteristics were also found for the latter half of cold period in 1650–1880, which indicated that the shift of harvest grade distribution didn't affect the connection between poor harvest and extreme drought/flood in cold epoch. These results suggested that warm period could weaken the impact of extreme drought on poor harvest during historical times.

2. Drought and/or flood might have increased the likelihood that officials reported problems such as poor harvests and other disasters

Accepted. As expressed in abovementioned revisions (P6, L3-7; Table S2), the records on harvests in historical documents was a relative level and focused directly on cropping in most cases, therefore it is reasonable to suggest that there was no tendency in harvest records.

In Chinese historical documents, the yearly harvest was usually recorded as a relative level compared to expected maximum yield rather than the individual crop production per hectare, although some records also reported the impacts on food availability, tax remissions, people's livelihoods, etc. Therefore these harvest records excluded the differences on absolute yield between sub-regions with different climate, soil fertility, crop variety, etc., and the difference between historical periods with changing agricultural centre, farming technology, staple crops (Su et al., 2014).

3. Harvest failures might have increased the likelihood that officials reported disasters such as droughts and/or floods.

Accepted. As mentioned before, the droughts and floods records in historical documents were usually focused on abnormal precipitation, and appeared to report all extremes equally. (P3, L23-30)

From these samples, it's easy to know that the droughts recorded in the historical documents were usually focused on the events due to no or less precipitation in a period at first, thus could be regarded as meteorological drought rather than hydrological drought and agricultural drought, although some records also reported the impacts on hydrosphere (e.g., dry up for river) or agriculture (e.g. wilting for crops). Therefore, the droughts recorded in the historical documents appear to report all droughts equally rather than those affected crops only. Similarly, the floods recorded in the historical documents could be regarded as more rains or heavy rains rather than the river that burst their banks or tsunamis, although some records also reported the impacts on the overflowing or bursting for lakes or river due to more precipitation or heavy rain.

4. Droughts and/or floods might have harmed human and animal health, reducing labor for harvests.

Accepted and revised. We added these possible pathways for the connection between extreme events and poor harvest in the revised manuscript. (P10, L27-32)

This is because that the relationship between poor harvest and extreme drought might be caused by not only the reductions in water supply, but also other possible indirect pathways. For example, the droughts might harm human and domestic animal health or cause migrations which reduced labor for farming. Moreover, the extreme events might reduce public revenue or increase public expenses which increase the political and economic instability, and further affect the agricultural activities

(Zheng et al., 2014a).

5. Droughts and/or floods might have damaged infrastructure and transportation, leading to food availability decline.

Accepted. Most of the yearly harvest records were direct wording on assessment of crop yield, which could not be influenced by damaged grain transportation. (Table S2)

6. Droughts and/or floods might have driven migrations, creating regional shortages both where agricultural labor emigrated and where people arrived seeking food.

Accepted and revised. This possible pathway has been addressed in revised manuscript along with pathway 4. (P10, L27-32)

This is because that the relationship between poor harvest and extreme drought might be caused by not only the reductions in water supply, but also other possible indirect pathways. For example, the droughts might harm human and domestic animal health or cause migrations which reduced labour for farming. Moreover, the extreme events might reduce public revenue or increase public expenses which increase the political and economic instability, and further affect the agricultural activities (Zheng et al., 2014a).

7. Periods of drought and/or flood might have reduced public revenue and/or increased public expenses, thus increasing the political and economic instability and decreasing food availability. (For instance, it's not clear how much the figures overall are influenced by the very high frequency of disasters and widespread famine during the political turbulence and violence accompanying the collapse of the Ming dynasty.) I am not arguing that any of these scenarios is necessarily the case. Nevertheless, each of these may be influencing the observed correlations.

Accepted and revised. This possible pathway has also been addressed in revised manuscript along with pathway 4 and 6. (P10, L27-32)

This is because that the relationship between poor harvest and extreme drought might be caused by not only the reductions in water supply, but also other possible indirect pathways. For example, the droughts might harm human and domestic animal health or cause migrations which reduced labour for farming. Moreover, the extreme events might reduce public revenue or increase public expenses which increase the political and economic instability, and further affect the agricultural activities (Zheng et al., 2014a)..

In summary, I do not believe that the authors' database and methods currently prove a valid correlation between flood and drought frequency and harvests in imperial China, nor that such a correlation would prove that drought or flood reduced harvest yields. The problem is not that the authors' hypothesis is unreasonable. It is simply that the conditions and data are too heterogeneous over such a large spatial and temporal scale. Any correlations found on such a scale are likely to have arisen from some artefact of the record-keeping or through the influence of some confounding variable, rather than to reflect a real and consistent climatic impact on agriculture.

Nevertheless, I would not like to dismiss this study out of hand. These datasets still have tremendous potential for historical climatology research. Better statistical methods could be devised to deal with changes in the frequency of historical reporting. By bringing trained historians onto such a project, the authors might find ways to handle problems related to historical changes in Chinese population, politics, land use, and economy. I would like to see the authors successfully address such problems in their research

Technical notes:

1). The paper variously sometimes to geographical parts of the country (e.g., "Northeast China") and sometimes to regional designations (e.g., "Jiangnan"). The paper would be clearer if it stuck with regional designations and names of provinces only.

Accepted and revised. The related manuscript has been revised to make sure that same expression were used referring to each sub-regions in the study area (i.e. the North China Plain, the Jiang-Huai area, and the Jiang-Nan area). Although in certain sentences expressions such as "southern China" should still be used for accuracy when referring to specific orientation of China.

2). The paper also needs extensive editing for English language grammar, spelling, and correct syntax. This is not merely a stylistic issue. The meaning of several passages is unclear due to lack of clear and correct English usage.

Accepted and Revised. The revised manuscript is under processing by professional translation company for better English writing. Although it would take some time before it is accomplished. The polished manuscript could be available in about 10 days.

Anonymous Referee #2

This is a nice piece of paper, well-structured with clear scoping and good delineation. It investigates the time evolution of extreme drought/flood events and the correlations of those extreme events with crop harvest in cold/warm epochs. In general, this paper provides very knowledgeable information on the drought/flood and harvest reconstruction method derived from documentary records, very comprehensive literature review in sections 1 and 2.1. However, there are still some points that I would suggest for authors to further improve the scientific quality and literacy of the paper.

Data used for analysis in this study is based on the previous analysis. The rationale for deciding the drought/flood (Zheng et al. 2006 and Hao et al. 2016) and harvest grades (Su et al, 2014; Yin et al, 2015) seemed promising however readers need to trace back those papers for more information on the source and profiles of data. There thus exists an ambiguity about how those records were collected for building the data sets, the characteristics and amounts of records, and data reliability evaluation. At this point, I can only assume that the data are reliable for the following analysis. It can be helpful if the authors provide some basic statistics (as appendix maybe, e.g. number of records per year, min, max etc for variables) of the data profile.

Accepted and revised. Detailed statistics of the data profile had been given in revision (P3, L32-P4, L16; P6, L15-17). In addition, two tables on the methods for building the datasets and one figure on the statistics of data-missing status for extreme events are provided as supplementary materials (Table S1, S2; Fig. S1). These datasets have been commonly used in former studies and proved to be valid (P4, L22-26). The supplementary materials are attached at the end of this text.

The grades were classified using the ideal frequency criteria of 10% (grade 1, severe drought), 20% (grade 2, drought), 40% (grade 3, normal), 20% (grade 4, flood), and 10% (grade 5, heavy flood) for whole area and all time, which were calibrated with the descriptions on duration, intensity, and area of the drought/flood events in wet season (usually May to September), and its impact (Table S1). Thus, the season for the drought/flood grade data overlaps with the critical agricultural activities and phases of crop growth. Although the data for the grade of drought/flood were unevenly distributed spatially for the whole 2000 year, i.e. there were few available data for grade of drought/flood in south China (south of 30 ^{o}N approximately) before CE 760 and even fewer data in south of the Huaihe River (approximately 34 ^{o}N) before CE 300 (Zhang, 1996), the coverage of this dataset has extended to south China since 760 CE and therefore covered the whole study area. There also existed missing data before 1470 as fewer historical documents survived from these earlier times (Zhang, 1996; Hao et al, 2016). The statistics shows that the mean percentage of available *data was 44.1% for 800-1469 and only 20% or even lower in periods around 850 and in 880s-920s, 1230s-1250s, 1360s and 1390s, in which the mean percentage of available data reporting "disasters or extremes" (i.e., grade 1, 2, 4, 5) and "normal" (i.e., grade 3) was 41.8% and 2.3%, respectively (Fig. S1). Moreover, there was no "normal" record in a period with length of 520 years. This means that the most of available grade data recorded disasters or extremes due to the principle of "recording the unusual rather than the normal" in the compilation on Chinese history. In consideration of the ideal frequency criteria in which 40% of all records were defined as "normal" and the other 60% were defined as "disasters and extremes", it could be implicated that about 70% of "disasters and extremes" that actually happened in that period was recorded (41.8% compared with 60% for whole area and all time).*

The criteria and methods for grading the documentary records (i.e., grain yield descriptions and the related information) year by year were presented by Su et al. (2014) and summarized by Yin et al. (2015), in which the classification of the yearly harvest grade and descriptions recorded in historical documents was shown in Table S2.

Therefore, this dataset provides a valuable proxy and has been well used to study characteristics of precipitation changes over eastern China back to the past 2000 years. For example, by using this data set, Zheng et al. (2006) reconstructed a 1500 year regional dry/wet index series for the North China Plain (approximately 34–40 °N), the Jiang-Huai area (approximately 31–34 °N) and the Jiang-Nan area (approximately 25–31 °N).

Based on the data, methods used for analysis is relatively simple. The authors used 50-years moving average (they term it as moving window) to smoothen extreme drought/flood trend, used Wilcoxon rank test to examine/compare median values of every intervals, and used contingency table to examine the effects between extreme drought/flood and harvest and between them and cold/warm periods. For this section, I suggest the authors to add a short paragraph giving readers some concepts about the method structure before going into details.

Accepted and revised. A new paragraph has been added to introduce the method structure in this section. (P7, L7-11)

Four kinds of data processing methods were used in this study, including the moving average method, the Wilcoxon rank sum test, the two-sampled t-test and the contingency table. The first two methods were used to explore the characteristics of occurrences of regional extreme drought/flood and the grain harvest grade from 801 to 1910, while the latter two methods were used to examine the relationship between poor harvest and extreme drought/flood for the whole study period or for specific warm/cold periods.

Also there are some unclear parts:

1).what do you mean by saying 'moving-window of 50 years and step of 10 years (line 24-25, page 5, and figure 3 caption, page 18)'? Please provide explanations.

Accepted and revised. (P7, L14-15)

For example, the smoothed series is made up of means of 801-850, 811-860, 821-870, etc.

2). For the same figure 3 caption, please use real number to replace full confidence, high confidence, medium or low confidence. It is unclear what those mean.

Accepted and revised. (P20, L6-9)

Bars at bottom row of each plate illustrated the confidence levels (probability in being correct, PBC): very high confidence (PBC>90%): dark; high confidence (PBC: 66.7%~90%): 50% shaded dark; medium confidence (PBC: 50%~66.7%): 25% shaded dark; low confidence: 12.5% shaded dark (PBC: 33.3%~50%); and very low confidence: blank (PBC: <33.3%) for each reconstructions at per 50 years.

3). Also, I don't quite understand the sentence in line 26-28 page 5 "This is because the mean of rank series in an interval was equivalent to the frequency of ……by labeling the extreme drought/flood years as 1 and non-extreme years as 0". Please provide more explanations.

Accepted and revised. (P7, L16-19)

By labelling the extreme drought (or flood) years as 1 and non-extreme years as 0, the chronology of extreme drought and flood years would be transformed into a rank series, and the mean of this rank series is equivalent to the frequency of drought (or flood) years. Therefore the intervals with significant more or less drought (or flood) years could be recognized through Wilcoxon rank sum test performed on the rank series.

The research results are clear and straightforward. The drought, flood and harvest trends and their descriptions are clear. However some trends are inconsistent with previous studies. For example, the authors mentioned that "there was an evident jump around 1640s with increase of years of (harvest) grade 4......" (line 12-13 page 7). I admire the following sentences on the discussions of the records in Qing and previous dynasties to clarify the discrepancies of the historical books. However, after removing grade 4 (average harvest), there still existed an

obvious jump of grade 5&6 (bumper) after 1640 which was commonly recognized in previous literatures as having poor harvest and famine in this coldest interval of little ice age. There might be some reasons including suddenly increasing number of local chronicles in Qing dynasty that could dilute the drought magnitude based on your grading method or the 50-years moving average can further smoothen the trend. In a word, it will be extremely valuable if the authors can compare the present analysis with previous studies and provide explanations or new perspectives.

Accepted and revised. Detailed discussion about the abrupt change in frequency of grade 5+6 in harvest has been applied. The jump around 1650s might be resulted from the uncertainty of source data, we added it in the discussion (P13, L15-18). However, the harvest in the whole Qing Dynasty was significantly high indeed, which could also be confirmed by other datasets from independent sources (P11, L29-P12, L1).

In addition, there existed an obvious jump of grade 5+6 (bumper harvest) around 1650s -1660s, yet this period was commonly recognized in previous literatures as having poor harvest and famine in this coldest interval of little ice age. Also, in the periods of 1130~1150, 1210-1270 (i.e., the early and later Southern Song Dynasty), 880-980 (i.e., the later Tang Dynasty and Five Kingdom) and around 1400, there were remarkably more data missing on harvest grade data.

Meanwhile, there existed an evidently high frequency of bumper harvest (grade 5+6) from 1650s to 1810s. This jump could also be confirmed by other datasets from independent sources. For example, according to the harvest reports in Archives of Qing Dynasty, the mean harvest percentage over eastern China for 1730-1820 was even over 70% (i.e. near bumper) (Ge and Wang, 1995), in which it was over 75% in Guangdong Province, southern China for 1707-1800 (Marks, 1998).

Ge, Q., and Wang, W.-C.: Population Pressure, climate change and Taiping Rebellion, Geogr. Res., 14(4), 32–42, 1995. (in Chinese)

Overall, this paper provides new and important insights into the correlations among extreme event, harvest and cold/warm climate. Data statistic is suggested to provide, and since missing data especially for harvest is prominent (35%), it is suggested that authors are more careful to claim their conclusions. Some inconsistency is also found between text and tables, e.g. 49.4% (line 23) and 24.0% (line 24) on page 8 are different from those shown in table 4. Further English editing is strongly suggested to improve high quality writing style of this nice paper.

Accepted and revised. The inconsistency between text and tables has been revised (P10, L15-16) and the revised manuscript has been handed over to professional translation company for better English writing. Although it would take some time before it is accomplished. The polished manuscript could be available in about 10 days.

Table S1: Criteria for calibrating the drought/flood grade and descriptions of drought/flood disasters recorded in historical documents (Zhang, 1996)

Grade	Frequency distribution (%)	Descriptions with meaning in historical documents
1	10	Continuous drought lasting two or more months in wet season (usually May to September) or crossing two seasons with severe intensity and impact over a broad area, such as "villages for hundreds of miles were abandoned"
2	20	Drought lasting two months or more than one month in wet season with visible impacts
3	40	Usual case (such as "rain blended well in seasons") or nothing special to be recorded
4	20	More rains lasting less than two months or more heavy rains in wet season with evident impact
5	10	Continuous more rains lasting two or more months, or extraordinary heavy rains in wet season with severe impacts, such as "driving boats over land"

Table S2: The classification of yearly harvests level derived from the wording recorded in historical documents (Su et al., 2014; Yin et al., 2015)

Levels of yearly harvests	**1**	**2**	**3**	**4**	**5**	**6**
	Very poor	Poor	Slightly poor	Average	Near bumper	Bumper
Harvest rating in percentage	<40%	40%-50%	50%-60%	60-70%	70-80%	>80%
Direct wording on assessment of crop yield	Very poor harvest; Very bad year; Nothing was reaped; No harvest in vast areas	Poor harvest; Poor grain harvest; Bad year; No prospect for harvest; No any bumper crops	Slightly poor harvest; No enough crop yield; No good harvest/year	Normal harvest; Unusual year; Not bad year; Common year	Bumper harvest; Good harvest; Bumper year; Favorable year	Large bumper harvest; Large harvest; A golden year; Bumper harvest of all "five" crops
Impact wording on assessment of food security related to crop harvest directly	People became cannibal; Large famine and no food; Deadly famine year and beggars were everywhere, many people died of hunger	People were starving; No enough food for people; People had pale and anemic complexions; Production on alcohol and related drink were prohibited			Surplus grain stored in the fields; Foods and clothes more than enough	
Related features, e.g., tax remission, people's livelihoods, grain prices and grain storage status	All taxes were remitted; People fled everywhere; Pirates and beggars were rampant; Food storage was exhausted	Taxes were partly remitted; People were not engaged in agriculture; People lived in poverty; People could not support themselves; Most of the people were poor; People were impoverished and had many complaints; The granaries were nearly empty; The whole country was poor and weak; The grain price was five or ten thousand Qian per Dan; The grain price was soaring; Grain became more and more expensive	Farms were laid waste; People had no surplus food storage		People lived a prosperous and contented life; There was enough grain for store	Every family had adequate supplies of food and clothes; People were well-off; The whole country was in safety and good order; The granaries were all full; The grain price was so low that peasants could not earn enough money

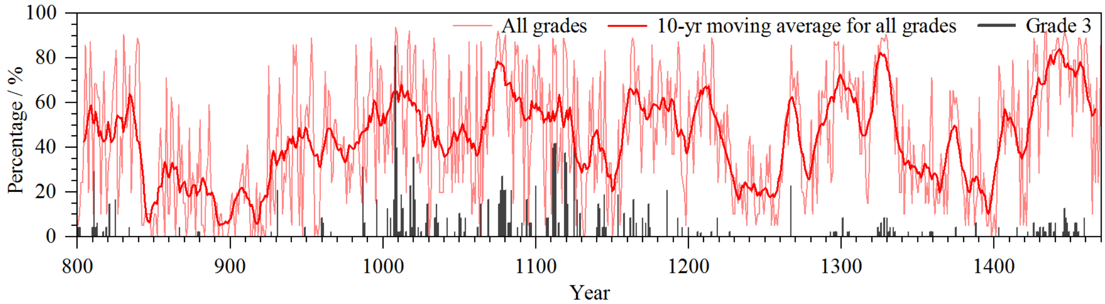

Figure S1: The percentage of available data in the reconstructed 63-stations annual drought/flood grades from 801 to 1470 CE. Pink line indicates the yearly variation for all grades in total and red bold line shows its 10-yr moving average. Black bars indicates the yearly variation for "normal" conditions (i.e. grade 3).