Decision: Major Revisions

In my review of the previous version, I identified three sets of problems with the manuscript: 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. I believe the authors' responses and revisions have mostly addressed the first set of concerns, and have partly addressed the second set, but have not adequately addressed the third set. Therefore, I recommend further major revisions. The text at the bottom reproduces the original review and the authors' responses and adds my replies (in bold).

The manuscript raises a number of major methodological issues common to much Chinese historical climatology. I believe it essential to address some of these here first in order to help set a precedent for publication in *Climate of the Past*. If, on the one hand, *CP* is to maintain its high standards and, on the other, Chinese historical climatology is become accepted in the global community of (climate) historians, then studies such as this one need to be approached somewhat differently. The task is important because Chinese historical climatology holds considerable promise, including an unparalleled wealth of historical sources and an active community of scholars. Yet that promise remains unfulfilled, because most (climate) historians outside China do not -- indeed, *cannot* -- accept its present methods and results. I therefore apologize for what may seem an unusually long and rigorous review.

Ultimately, the central issue with much Chinese historical climatology -- this manuscript included -- is that it does not study history per se but rather patterns in the data from historical databases compiled by other scholars. The data in those databases and their internal correlations probably reflect actual historical events and causal relationships, but we cannot know to what extent or with what consistency and accuracy, particularly for sources before the late Ming period. Chinese historical climatology studies such as this one rarely include historians who can provide an expert first-hand judgement of the underlying historical sources. Moreover, these studies -- including the present manuscript -- often try to establish causal relations between climatic events and human impacts based on data all drawn from the same historical sources, without adequately considering the ways that biases in reporting and transmission of information might lead to spurious correlations.

What some studies -- including the present manuscript -- do instead, is to try to establish the objectivity of their sources. They assume that if the sources are demonstrably objective by some simple test then what they report can fairly be related as historical fact. Therefore, if patterns emerge in the supposedly factual reporting of two or types of events, then one can extract data about their occurrence and look for patterns to establish actual causal relationships.

This, however, is not a correct method of historical investigation for this purpose. Philosophers of historiography (e.g., Aviezer Tucker) have demonstrated that sound historical research, as a matter of both logic and practice, uses a Bayesian approach to inferring historical conditions and causes from available evidence. In other words, the prior probability of some plausible historical hypothesis about conditions or causation (p(h)) should be updated by weighing, on the one hand, the likelihood of the appearance of the available evidence given the hypothesis (p(o|h)) and, on the other hand, the likelihood that the same evidence would appear regardless of the truth or falsehood of the hypothesis (p(o)). Thus, when researchers make an argument for historical

causation based on a correlation or co-variation among reported incidences of two phenomena, they ought to establish three points: (1) a reasonable prior probability of a historical hypothesis; (2) a high probability of finding the correlation if the hypothesis were true; and (3) a low probability of finding the correlation for any other reason. In this way, researchers may arrive at a high posterior probability (p(h|o)) and thus make a valid historical inference.

The approach taken by Chinese historical climatology -- including the present manuscript -- has problems, first of all, with points (1) and (2) above. That is because it uses a basic inductive approach. It starts with few or no prior hypotheses to test. It simply looks for patterns in the data and then comes up with plausible explanations based only on the variables being considered (usually climatic factors rather than equally plausible factors such as political, economic, demographic or cultural change).

However, the real problem comes when meeting point (3). Normally, (3) might be established by running statistical tests in order to indicate that the correlation or co-variation between the two phenomena -- the proposed cause and proposed effect -- has a low probability of occurring by chance. However, this method is only valid in historical climatology when the reporting of the two phenomena is truly independent and reliable: that is, when they can't be confounded by some other factor, whether in how the phenomena occurred or how the occurrence is reported. If, on the contrary, our knowledge of the (co-)occurrence of the two phenomena depends on reporting and transmission from the same or related historical sources, then there could easily be confounding factors influencing any apparent correlations. Thus, the usual statistical tests no longer suffice to prove point (3) above.

Historical climatologists typically use several methods to confront these problems. They focus on specific historical hypotheses (e.g., the correlation of harvest dates and summer temperatures, or the impact of summer temperatures on grain prices). They create climate indices from documentary evidence but then calibrate and verify these indices with an independent source of information: typically early instrumental measurements or data derived from proxies in natural archives. They make an explicit assumption that any correlation found during the period of overlap between the documentary and instrumental records is more or less constant over time; yet this "principal of stationarity" is applied only to a limited geographical area and type of data (e.g., summer temperatures in France based on grape harvest dates). When making arguments about climate impacts on society, they typically compare their climate data with societal data drawn from historical sources independent of those used to reconstruct climate. All of these steps help to reduce the risk that the ways historical evidence has been recorded, transmitted, and analyzed will generate artefacts in datasets of climate and societal phenomena that lead to spurious correlations and thus false conclusions about causality.

Much current Chinese historical climatology (this manuscript included) does not take such precautions. The datasets on floods and droughts are not calibrated to an independent source of climate information. Both the climatic data, on the one hand, and the harvest data, on the other, come from the same Ming and Qing official sources (principally the *Twenty-Four Histories* and *Qing History Draft*) subject to the same processes of historical reporting and transmission. Therefore, the risk is very high that processes of recording, transmission, interpretation, and recompilation of those historical records has introduced artefacts into the datasets and thus

invalidates supposed causal connections between climate phenomena (in this case, extreme precipitation events) and harvests.

The problem is all the more acute because the basic inductive method described above asks so much of the consistency and independence of the datasets. The authors don't just want to show that drought or flood affected Chinese agriculture in some particular way stated as a prior hypotheses. They want to argue that any patterns they can identify in reported frequencies of floods and/or droughts and reported grades of harvests demonstrates real shifts in the influence of climate. This basic inductive approach not only risks mistaking random fluctuations in the relationships among the data for real shifts in climate impacts; more importantly, it also overlooks the obvious potential for other factors (administrative, geographic, demographic, political, technological) to create patterns in the data, too. Thus, this study, like much Chinese historical climatology, depends too heavily on the homogeneity and consistency of reporting and transmission of information across two millennia of history and millions of square kilometers of territory, across the reigns of scores of emperors and whole new dynasties, over the scope of China's far from homogenous or unified history.

In short, I do not dispute that the data have been properly collected nor do I dispute that the authors have found real patterns in the dataset. I find that the authors have drawn some valid conclusions *about the data*. However, I believe they have not proven their conclusions about the *actual (climate) history* of imperial China.

Nevertheless, I believe that it is worthwhile to publish studies specifically and explicitly about patterns in the data found in Chinese climate and historical databases. First, such studies can help us establish the reliability (or not) and usefulness (or not) of such databases. Second, such studies could form a useful starting point for further historical investigations, even if they cannot point to any firm historical conclusions by themselves.

I would therefore insist on the following further revisions for the publication of this manuscript, and I would strongly encourage their use as standards for similar Chinese historical climatology studies:

1) The title should reflect the fact that this is primarily a study about patterns found in data in historical and climate databases, and only secondarily a study from which we might make inferences about history. In this case, I would accept, for instance, "Patterns in the Reporting of Droughts/Floods and of Harvest Grades in Historical Documents in Eastern China, 801-1910" or "Patterns in Data on Precipitation Extremes and Harvests in Historical Databases for Eastern China 801-1910".

2) The article should clearly distinguish its identification and analysis of data and patterns in that data from any inferences about (climate) history that it makes based on those data and patterns. Those inferences should be more limited, in accordance with the uncertainties and methodological problems outlined above, and should probably appear in the discussion or conclusion of the article. That would also be a good place to discuss whether studies such as Hao et al. 2010 really enable researchers to infer actual historical events and causality from this data.

3) While discussing the data itself and patterns identified in that data, all mentions of historical floods, droughts, and harvest levels should be qualified as "reported" floods or droughts and "reconstructed" harvest levels. Instead of "co-occurrence" of hydrological extremes and poor harvests, the authors should refer to the "co-reporting" of hydrological extremes and poor harvests.

4) Please see the replies below for further specific points.

Previous Review with Author Responses and Reviewer Replies:

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, L23-27)
As this sample suggests, historical documents were usually focused on the events due to no or less precipitation than usual and, thus could be regarded as meteorological drought rather than hydrological or agricultural droughts, although some records also report impacts on the hydrosphere (e.g., rivers drying up for river) or on agriculture (e.g. wilting for crops). Historical documents, therefore, appear to report all droughts equally, rather than only those affecting crops.
Reply: Revision accepted.

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, L27-29)
Similarly, floods recorded in the historical documents could be regarded as more rain or heavier rain than usual, rather than in the context of rivers bursting their banks or tsunamis, although some records also report the impacts of overflowing or bursting lakes and rivers due to more or heavier precipitation.
Reply: Revision accepted.

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, L31-P4, L3)
Grades were classified using 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 the
whole area and all time. These grades were calibrated based on descriptions of duration, intensity,
and area of the drought/flood event during the wet season (usually May to September), and its impact (Table S1). Thus, the season of the drought/flood grade data overlaps with critical agricultural activities and phases of crop growth.
Reply: Revision accepted.

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, L1-3)
In Chinese historical documents, the yearly harvest was usually recorded as a relative level compared to an expected maximum yield, rather than crop yield per hectare, although some records
also report impacts of harvest fluctuation on food availability, tax remissions, livelihoods, and so on.
Reply: If this is the case, then the authors need to be explicit throughout the text that what they have reconstructed is not absolute "harvest levels" but rather interannual variability in local production of staple crops.

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, L30-P4, L2; Table S1) and harvest (P6, L12-15; Table S2). Please refer to the supplementary material which is uploaded additionally as an independent file. Based on these records, Zhang (1996) reconstructed a dataset of annual drought/flood grades at 63
stations from 137 BCE. Each station consisted of a local area of approximately 20 counties with the
same climate. Grades were classified using 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 the whole area and all time. These grades were calibrated based on descriptions
of duration, intensity, and area of the drought/flood event during the wet season (usually May to September), and its impact (Table S1).
In the dataset, yearly harvest levels 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 year-by-year grading of
the documentary records (i.e., grain yield descriptions and related information) were presented by
Su et al. (2014) and summarized by Yin et al. (2015). The classification of the yearly harvest grade
and descriptions recorded in historical documents is shown in Table S2.
Reply: Revision accepted.

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, L30-31). 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, L5-11). 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, L28-31).
Based on these records, Zhang (1996) reconstructed a dataset of annual drought/flood grades at 63
stations from 137 BCE. Each station consisted of a local area of approximately 20 counties with the
same climate.
Reply: This revision is appropriate, but again only if the authors consistently make clear that they are reconstructing interannual variability in local production of staple crops and not an absolute "harvest level." The absolute level of harvests in different regions and eras on decadal or longer scales would still have depended more on changes in political economy, population density, crop strains, and technologies than on weather and climate.

To verify the rationality of this method and criteria, validation was conducted in Hao et al. (2010a),
based on 10 extreme events identified from a series of precipitation observations in each sub-region
according to a threshold of probabilities of 10% and 90% occurrence. In this validation, all or part
of grade 3 stations were deliberately omitted, and only 40% or 60% of stations with disaster or extreme grade were reserved without changing the drought-to-flood ratio within the available data.
The results show that, with one exception, years of extreme drought and extreme flood, identified according to this method and criteria, closely matched those extreme events identified by precipitation data, demonstrating that the method and criteria were reasonable.
However, such social factors should have only limited influence on yearly harvest grade dataset, since the harvest in the documents was reported as a relative level rather than the absolute yield, also the main grain product area, the staple crop, and the cropping system have been relatively stable throughout the study period (Yin et al., 2015).
Reply: It appears that Hao et al. 2010 helps establish that the historical sources underlying the databases were not biased toward reporting floods versus droughts or to reporting them in some regions and not others. That study also appears to establish that these sources did not usually falsely report precipitation extremes. However, Hao et al. 2010 does not establish that the reporting of precipitation extremes was independent of the reporting of poor food production throughout the long and diverse history of today's China. This remains a major shortcoming of the study, which I will return to below.

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, L1-5)
In Chinese historical documents, the yearly harvest was usually recorded as a relative level compared to an expected maximum yield, rather than crop yield per hectare, although some records
also report impacts of harvest fluctuation on food availability, tax remissions, livelihoods, and so on. Therefore these harvest records exclude differences in absolute yield between sub-regions with
different climates, soil fertility and types, crop varieties, etc., as well as difference between historical
periods with changing agricultural centres, farming technologies, staple crops, and so on. (Su et al.,
2014).
Reply: This revision is appropriate, but again only if the authors consistently make clear that they are reconstructing interannual variability in local production of staple crops and not an absolute "harvest level." The absolute level of harvests in different regions and eras

on decadal or longer scales would still have depended more on changes in political economy, population density, crop strains, and technologies than on weather and climate.

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, L25-31)
During this study period, several social factors existed which could have influenced China's total yield, such as 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 introduced from the New World (e.g., peanuts and sweet potatoes), advanced agricultural management technology, and so on. However, such social factors should have only limited influence on yearly harvest grade dataset, since the harvest in the documents was reported
as a relative level rather than the absolute yield, also the main grain product area, the staple crop,
and the cropping system have been relatively stable throughout the study period (Yin et al., 2015).
Reply: Revision accepted.

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, L1-5, L25-31).
In Chinese historical documents, the yearly harvest was usually recorded as a relative level compared to an expected maximum yield, rather than crop yield per hectare, although some records
also report impacts of harvest fluctuation on food availability, tax remissions, livelihoods, and so on. Therefore these harvest records exclude differences in absolute yield between sub-regions with
different climates, soil fertility and types, crop varieties, etc., as well as difference between historical
periods with changing agricultural centres, farming technologies, staple crops, and so on. (Su et al.,
2014).
During this study period, several social factors existed which could have influenced China's total yield, such as 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 introduced from the New World (e.g., peanuts and sweet potatoes), advanced agricultural management technology, and so on. However, such social factors should have only

limited influence on yearly harvest grade dataset, since the harvest in the documents was reported
as a relative level rather than the absolute yield, also the main grain product area, the staple crop,
and the cropping system have been relatively stable throughout the study period (Yin et al., 2015).
Reply: revision accepted

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. (P4, L34-P5, L13)
Extreme drought or extreme flood years were defined in this way, as the probabilities for omitting
drought and flood records were random and unbiased, despite the greater frequency of missing data
in the older records. In other words, if one period had a large number of documents, it was expected
to 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.
To verify the rationality of this method and criteria, validation was conducted in Hao et al. (2010a),
based on 10 extreme events identified from a series of precipitation observations in each sub-region
according to a threshold of probabilities of 10% and 90% occurrence. In this validation, all or part
of grade 3 stations were deliberately omitted, and only 40% or 60% of stations with disaster or

extreme grade were reserved without changing the drought-to-flood ratio within the available data.
The results show that, with one exception, years of extreme drought and extreme flood, identified according to this method and criteria, closely matched those extreme events identified by precipitation data, demonstrating that the method and criteria were reasonable. The reason for the
close match is that precipitation variability in eastern China is dominated by the East Asian Summer
Monsoon (EASM). Therefore, when extreme drought or flood events occur, the precipitation variation for stations within each sub-region usually share similar relative magnitudes.
Reply: It appears that Hao et al. 2010 helps establish that the historical sources underlying the databases were not biased toward reporting floods versus droughts or to reporting them in some regions and not others. It also appears to establish that these sources did not usually falsely report precipitation extremes. However, Hao et al. 2010 does not establish that the reporting of disasters was independent of the reporting of poor food production throughout the long and diverse history of today's China. This remains a major shortcoming of the study, which I will return to below.

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, L18-22; P11, L28-P12, L2)
As found in section 3.2, more occurrence of extreme drought in eastern China led to a significant

increase in the frequency of poor harvests (grade 1+2) when compared with non-extreme years. Since more extreme droughts occurred over eastern China in 920–1300 than in 1310–1880, the harvest in the warm epoch could be expected to be worse than in the cold epoch. However, as Yin et al. (2015, 2016) found, the harvest in warm epoch was better than that in cold epoch. This suggests that the effects of regional extreme drought on the grain harvest differed between warm and cold epochs.

The results show that, during the warm epoch of 920–1300, there was no significant connection between the occurrence of poor harvest and regional extreme drought, although the frequency of poor harvest in extreme drought years was slightly higher than in non-extreme years for each subregion.

In contrast, during the cold epoch of 1310–1880, the frequency of poor harvest in extreme drought years was significantly higher than in non-extreme years, which indicates that the connection between the occurrence of poor harvest and extreme drought was still significant. Moreover, similar characteristics were found for the latter half of the cold period from 1650 to 1880,

which indicates that the shift of harvest grade distribution did not affect the connection between poor harvest and extreme drought/flood during the cold epoch. These results suggest that the warm

period could weaken the impact of extreme drought on poor harvests in historical times.

Reply: I do not find that this approach adequately disambiguates the effects of extreme precipitation and temperature on food production. It's a basic inductive method that simply takes all the data and then comes up with an explanation after the fact based on any observed patterns – patterns that could have emerged entirely by chance, or by some confounding third variable (e.g., population movements or changes in political economy that influenced agricultural vulnerabilities), or due to some artefact in the way events were reported or records kept. I'd be much more satisfied if there were some way to model the effects of both temperature variations and extreme precipitation on food production levels so that the authors could weigh relative contributions of each.

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, L1-5; 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 an expected maximum yield, rather than crop yield per hectare, although some records

also report impacts of harvest fluctuation on food availability, tax remissions, livelihoods, and so on. Therefore these harvest records exclude differences in absolute yield between sub-regions with

different climates, soil fertility and types, crop varieties, etc., as well as difference between historical

periods with changing agricultural centres, farming technologies, staple crops, and so on. (Su et al.,

2014).

Reply: The authors' response does not address my concern. Any causal argument about precipitation extremes and low harvests relies on the assumption that both of these phenomena were consistently reported independently of each other. It appears that Hao et al. 2010 helps establish that the historical sources underlying the databases were not biased toward reporting floods versus droughts or to reporting them in some regions and not others. It also appears to establish that these sources did not usually falsely report precipitation extremes. However, Hao et al. 2010 does not establish that the reporting of precipitation extremes was independent of the reporting of poor food production throughout China's long and diverse history. There remains the strong possibility that officials were more likely to report either the occurrence of such extremes when the harvests were poor (e.g., by way of explaining or justifying those poor harvests) or to mention poor harvests when there were meteorological extremes (e.g., in an effort to take advantage of the situation to secure state funds or tax remissions). I do not make this as merely a theoretical argument. Although I am not an expert on imperial China, this is exactly the pattern I have observed in the records of other early modern empires. Moreover, certain features of this study make this problem particularly troublesome for their conclusions. First, the information concerning both types of events (precipitation extremes and poor harvests) comes from the same set of historical records. Second, the records from earlier eras come primarily from Ming and Qing recompilations and not originals. Third, the older records are very incomplete and biased toward extreme events. Fourth, and most important, the authors' basic inductive method—taking all the data and then explaining any observed patterns on the assumption that they are causally related—sets a very high standard for the independence of the different datasets. In other words, if the authors were simply making the case that over the long run we should see an impact of floods and/or droughts on interannual food production, then small changes in political priorities or record-keeping and transmission practices shouldn't make a big difference. However, because the authors are identifying shifts and patterns over time—such as possible changes in the impact of precipitation extremes due to phases of warming and cooling—then any shifts in the independence of reporting the occurrence of extreme precipitation events and of poor harvests are likely show up in their presentation of the data and to be explained as real historical changes in the impact of floods and droughts on food production. In fact, that is exactly what I think has happened in this study, and that is why I am asking for major revisions.

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-
29)
As this sample suggests, historical documents were usually focused on the events due to no or less
precipitation than usual and, thus could be regarded as meteorological drought rather than hydrological or agricultural droughts, although some records also report impacts on the hydrosphere (e.g., rivers drying up for river) or on agriculture (e.g. wilting for crops). Historical documents, therefore, appear to report all droughts equally, rather than only those affecting crops.

Similarly, floods recorded in the historical documents could be regarded as more rain or heavier rain than usual, rather than in the context of rivers bursting their banks or tsunamis, although some
records also report the impacts of overflowing or bursting lakes and rivers due to more or heavier
precipitation.
Reply: (see previous reply)

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, L25-28)
This relationship may have been caused by both reductions in water supply and other indirect pathways. For example, droughts might harm human and domestic animal health or result in migration, leading to a reduced agricultural labour force. In addition, extreme events might reduce
public revenue or increase public expenses, thus increasing political and economic instability, and
further affecting agricultural activities (Zheng et al., 2014a).
Reply: Revision accepted.

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)
Reply: Accepted.

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, L25-28)
This relationship may have been caused by both reductions in water supply and other indirect pathways. For example, droughts might harm human and domestic animal health or result in migration, leading to a reduced agricultural labour force. In addition, extreme events might reduce
public revenue or increase public expenses, thus increasing political and economic instability, and
further affecting agricultural activities (Zheng et al., 2014a).
Reply: Revision accepted.

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, L25-28)

This relationship may have been caused by both reductions in water supply and other indirect pathways. For example, droughts might harm human and domestic animal health or result in migration, leading to a reduced agricultural labour force. In addition, extreme events might reduce
public revenue or increase public expenses, thus increasing political and economic instability, and
further affecting agricultural activities (Zheng et al., 2014a).

Reply: Revision accepted.