Dear editors and reviewers,

We have made a major revision on the manuscript cp-2018-177. We have completely **rewritten sections 1, 2 and 4**. There are also **changes on section 3 and figures (in Figs.6, 7)** for better illustration. We also changed the **title** of the paper to: "Holocene temperature response to external forcing: Assessing the linear response and its spatial and temporal dependence".

We also changed the authors' **affiliations** and **corresponding authors** to:

"

Lingfeng Wan[1, 2], Zhengyu Liu[3, 4], Jian Liu[1, 2, 4], Weiyi Sun[1, 2], Bin Liu[1, 2]

[1]Key Laboratory for Virtual Geographic Environment, Ministry of Education; State Key Laboratory Cultivation Base of Geographical Environment Evolution of Jiangsu Province; Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application; School of Geography Science, Nanjing Normal University, Nanjing, 210023, China.
[2]Jiangsu Provincial Key Laboratory for Numerical Simulation of Large Scale Complex Systems, School of Mathematical Science, Nanjing Normal University, Nanjing, 210023, China.
[3]Atmospheric Science Program, Department of Geography, Ohio State University, Columbus, OH43210, USA.
[4]Open Studio for the Simulation of Ocean-Climate-Isotope, Qingdao National Laboratory for Marine Science and Technology, Qingdao, 266237, China.

Correspondence: Jian Liu (njdllj@126.com); Zhengyu Liu (liu.7022@osu.edu)

"

We also checked and added the **references**.

The **point-by-point reply** to all the three reviews are finished.

We hope our revision is satisfactory to the journal.

Thanks!

Best regards!

All authors of manuscript cp-2018-177

Reply to the Anonymous Referee #1:

Interactive comment on "On the linearity of the temperature response in Holocene: the spatial and temporal dependence" by Lingfeng Wan et al.

Anonymous Referee #1

Received and published: 20 February 2019

In this work, the authors make a first attempt to investigate the linearity of forced Holocene variability in dependence on the temporal and spatial scale – by using global surface temperature fields obtained from the TraCE-21ka paleo-climate CGCM simulation. The topic is interesting and the results seem to indicate that further research into this direction may provide further knowledge, that will be useful for both, (a) the interpretation of paleo records and (b) the attribution and detection field of research. Nonetheless, I have a number of concerns regarding the conceptual approach (see 'general comments') which, I think, need clarification before the conclusions, drawn by the authors, can be thoroughly evaluated. A few specific questions are listed at the end (see 'Specific comments').

Reply: Thank you for the comments. We have modified the manuscript carefully according to your valuable comments.

1 General comments

(1) Throughout this work, it seems that the following two *different* questions are mixed up, which makes it basically impossible to evaluate the conclusions drawn from the results:

Q-1. How linear is the response to external forcing? If we denote the temperature resonse to the full external forcing, $F_{all}(t) = F_1(t) + F_2(t) + F_3(t) + F_4(t)$, by $T_R\big(F_{all}(t)\big)$, the response to the individual forcings by $T_R\big(F_i(t)\big)$ (with $i = 1, \ldots, 4$), and the internal temperature variability of the five model simulations by $T_{I,all}$, $T_{I,1}$, $T_{I,2}$, $T_{I,3}$, $T_{I,4}$, respectively, then the linearity of the response could be defined by the extent to which the statement

$$T_R(F_{all}(t)) = \sum_{i=1}^{4} T_R(F_i(t)) \qquad (1)$$

holds, and the linearity could be measured by the correlation between the forced response on the left and that on the right hand side of the above equation. In the manuscript, however, the correlation is computed (see Section 2.2) from the full 'forced plus internal' temperature variability, i.e. between $T_R\big(F_{all}(t)\big) + T_{I,all}(t)$ and $\sum_{i=1}^{4} T_R\big(F_i(t)\big) + T_{I,i}(t)$, Since this latter correlation is influenced by the signal-to-noise ratio, $\mathrm{Var}(T_R)/\mathrm{Var}(T_I)$, a small correlation does not necessarily indicate the absence of linearity, because it could be that simply the signal-to-noise ratio is small, although the response is still perfectly linear. (One would need ensembles of model simulations for each of the five forcing scenarios, and then use

the ensemble average in order to suppress the internal variability.) Hence, Q-1 cannot be answered by this approach (without additional information), unless one would always obtain correlations close to unity, which would indicate strong linearity.

Q-2. What is the relative importance of externally forced vs. internal variability, assuming the response were linear? To answer this question one could use the correlation computed from the full temperature variability, as done in the manuscript, but one had to assume the linearity which, however, is to be proven by this work, in particular, for different temporal and spatial scales.

Hence, the authors should clarify the above issues, and make explicit which of their results contributes to which one of the above two questions. This will also help to clarify the implications of the conclusions for various research fields.

Reply: We thank the reviewer for this comment. We agree with the reviewer completely on the definition of linearity. We apologize for our ambiguity in the original manuscript. Our focus is really on the slow evolution of temperature in the Holocene that is of comparable time scale to the forcing factors. We have made a major revision. First, we have rewritten all the sections except for section 3. In the revision, we clarified our single realization approach and its potential issues for assessing the linear response (in addition to clarification of the data and methods). Second, we have changed the title to: "Holocene temperature response to external forcing: Assessing the linear response and its spatial and temporal dependence"

(2) Even if we had ensembles available for each of the forcing scenarios, it would still be possible to obtain a large correlation coefficient although the response is only weakly linear (i.e., mostly non-linear), if the individual response to, for example, one of the forcings F_i is much larger than the responses to the remaining forcings, because in this case the full temperature variability might still be dominated by the response to the strong forcing (the non-linear interactions might still be relatively small). Thus, one would need to know the strength (e.g., in terms of variance) of the responses to the various individual forcings.

Reply: This raises an excellent point. The explained variance of each forcing factor is indeed very interesting. We think it deserves special attention. Due to the multiple time scales and the strong regional dependence here, however, a detailed study on the variance of each forced response would require much more analyses than in the current paper; it also tends to mix information with our basic information in the first paper here. Therefore, we will still focus on "if the linear response is valid" and will leave the study on "how much each forcing contribute" to a follow-up paper. As for the potential case of a response dominated by a single forcing, it is reasonable to consider this case as a good linear response to the dominant forcing, because the impact from other forcing factors are negligible anyway.

(3) In the Conclusions section it should be mentioned that, even if strong linearity for the given model simulations were proven, then this conclusion is valid only for the given range of forcing amplitudes as non-linearities may appear for stronger forcings.

Reply: Agreed. Comments are added in section 4. "It should therefore be kept in mind that the assessment could differ for different variables, in different models, for different periods and for different sets of forcing factors." "The assessment will

be also different if a different period is assessed, e.g. the last 21,000 years; with a large amplitude of climate forcing, the linear response may degenerate in the 21,000-year period."

(4) How is it justified to estimate the variance of the internal variability at orbital and millennial time scales by the full variance at centennial and decadal variability (page 7, last paragraph)? That is, why should we have

$$\mathrm{Var}_{orb,mill}(T_{I,all}) \approx \mathrm{Var}_{cent+dec}\left(T_{I,all} + T_R(F_{all})\right)? \qquad (2)$$

Even if we assume that $\mathrm{Var}_{cent+dec}(T_R(F_{all}))$ is small compared to $\mathrm{Var}_{cent+dec}(T_{I,all})$, this does not imply anything about the relation between $\mathrm{Var}_{cent+dec}(T_{I,all})$ and $\mathrm{Var}_{orb,mill}(T_{I,all})$. Maybe it could be helpful to investigate the power spectra of the temperature variability under the various forcing scenarios?

Reply: We apologize for the ambiguity here. Again, this is an approximation based on linear thinking. Since our forcing, orbital, ice sheet, meltwater and CO_2 are of time scales of millennial or longer (we don't have solar variability and volcanic forcing!), we assume that the variability at centennial and decadal time scales are caused mostly by internal variability. This point is clarified now in the revision in subsection 3.3. Since our forcing factors are on millennial and orbital time scales, and the linear response is also largely valid for orbital and millennial variability, we use the variance of the orbital and millennial variability as a crude estimate of the linear response signal. Since there is no centennial and decadal forcing in our model and the response of centennial and decadal variability are not linear response, we use the variance of the sum of the centennial and decadal variability as a rough estimate for internal variability as the linear noise.

2 Specific comments

(5) It would be nice if the reasons for showing the linear error index Le were explicitly discussed, and what the implications of this index are for the linearity. And what is the added value of this index over the correlation coefficient?

Reply: The correlation represents the similarity of the ALL and SUM, but can't evaluate the absolute magnitude of the two responses. Even if two time series is perfectly correlated, their magnitudes can differ by an arbitrary constant. The linear error is to reflect the magnitude of the relative error between the ALL and SUM. More clarifications are added in the text in section 2.2 on this.

(6) Please, be a bit more explicit how the significance levels are computed. For example, how is the AR(1) fit done in case of the correlation, and what is the bootstrap design for the error index?

Reply: The bootstrap is greatly expanded in the rewritten section 2.2.

Reply to the Oliver Bothe (Referee #2):

Interactive comment on "On the linearity of the temperature response in Holocene: the spatial and temporal dependence" by Lingfeng Wan et al.

Oliver Bothe (Referee)
ol.bothe@gmail.com
Received and published: 22 February 2019

General Comments

Wan et al. use the TraCE-21ka simulations in their manuscript "On the linearity of the temperature response in Holocene: the spatial and temporal dependence" to provide an initial assessment of the linearity of the climate response to various assumed forcings. They study this separately for a number of spatial and temporal scales. The idea behind the manuscript can result in a valuable contribution to our understanding of past and future climate changes, to assessing paleo-simulations, and to studying paleo-observational records. I do not have real major concerns but I think various clarifications and additional discussions are necessary before the manuscript could be accepted for publication. These clarifications should be re-evaluated by a round of revisions and therefore I, nevertheless, recommend major revisions.

Reply: Thank you for your comments. We have revised the manuscript according to your valuable comments.

Specific Comments

1: Could the authors please discuss more clearly, why they think that the assumptions on potential linearity hold (see also the major comments by anonymous referee 1). This discussion could also include, how the specific setup of the TraCE simulations hampers or supports the approach. It may help to include in these discussions a priori knowledge/references on forced and internal variability across temporal and spatial scales.

Reply: This point has been discussed in much more detail in the revision. Section 1 and section 4 are written. More clarifications are also given now. See the reply to referee #1, general comment (1).

2: Similarly, I think it is necessary to discuss, at least shortly, how the simulations implement the various forcings and how this may influence the results.

Reply: Thank you for your comments. We have added substantially more details on the simulation in subsection 2.1. We also added a paragraph here on the experimental design and its usefulness for linear response study. "It should be noted that the linear response can't be assessed if the individual forcing experiments are performed with the forcing superimposed one-by-one. In this approach, the four external forcing is added sequentially, for example, first the ice sheet, second the ice sheet plus orbital forcing, third the ice sheet, orbital and GHGs, and finally, applying all four forcing of ice sheet, orbital, GHGs

and melting water. In this experimental design, the full forcing response is by default the response of the sum response after adding the four forcing factors together, and therefore can't be used to test the linear response. It should also be noted that our four individual forcing experiments, although in principle feasible for assessing linear response, are not designed optimally for the study of Holocene climate. This is because, except for the variable forcing, all the other three forcing factors is fixed at the 19ka condition. As such, the mean state is perturbed from the glacial state, not a Holocene state. This may have contributed to some unknown deterioration on the linear response discussed later. Nevertheless, we believe, our major conclusion should hold approximately. This is because, partly, the response is indeed almost linear for orbital and millennial variability as will be shown later."

3: Could the authors please discuss, why correlation coefficients and the linear index are appropriate measures of the linearity of the responses as studied. Could they please also clarify, which information is added by the linear index and how to interpret the index in this context.

Reply: see Reply to specific comment 5 to referee #1. The correlation represents the similarity of the ALL and SUM, but can't evaluate the absolute magnitude of the two responses. Even if two time series is perfectly correlated, their magnitudes can differ by an arbitrary constant. The linear error is to reflect the magnitude of the relative error between the ALL and SUM. More clarifications are added in the text in section 2.2 on this. "However, the correlation does not address the magnitude of the response. Even if S_t and T_t has a perfect correlation r=1, the two time series can still differ by any constant factor in their magnitudes. Therefore, we will also use the linear error index L_e to evaluate the magnitude of the linear response."

4: Is the linearity assumption even valid for time scales where internal model processes are known to dominate. That is, we can be quite certain a priori that the decadal scale will be dominated by internal climate over the last 11k years.

Reply: With our approach of single realization, the linear response assumption will fail if the internal variability is dominant. Then, a large ensemble is needed to suppress internal variability. This is discussed in detail in the reply to referee #1, general comment (1).

5: Could the authors please stress that their conclusions really only hold for the specific setup of the TraCE simulations used.

Reply: Thank you for the reminder. Yes, we have been specifically clear about this in a sentence and discussion in section 4. "The result here represents the first such assessment and is carried out for a single variable (surface temperature) in a single model (CCSM3). It should therefore be kept in mind that the assessment could differ for different variables, in different models, for different periods and for different sets of forcing factors. For example, ... "

6: More generally, I think the manuscript is missing a dedicated and thorough discussion-section.

Reply: Thank you for your comments. The last section has now been rewritten and substantial discussions are added. Section 1 is also written to address many potential ambiguities.

7: Page 1 Line 25ff (P1L25): I do not think the authors show this causality conclusively. Anyway, the SNR plots do not show or are even intended to show, according to the manuscript, why linearity is consistent between millennial and orbital scales, but why linearity is strong in these regions. That is, in my understanding, the manuscript does not support this sentence in this form.

Reply: We think the referee is referring to this sentence: "On the millennial scale, the linear response is still strong in the NH over many regions, albeit weaker than on the orbital scale". The reviewer is correct in that the paper is not to show millennial and orbital are consistent, it is only to show the linear response in different regions for each time scale separately. The comparison is only made by the comparison between the two in Fig.5, instead of SNR in Fig.6. Therefore, we think this statement is valid.

8: P2L22ff: Could the authors please be more specific how this answer can benefit our understanding?

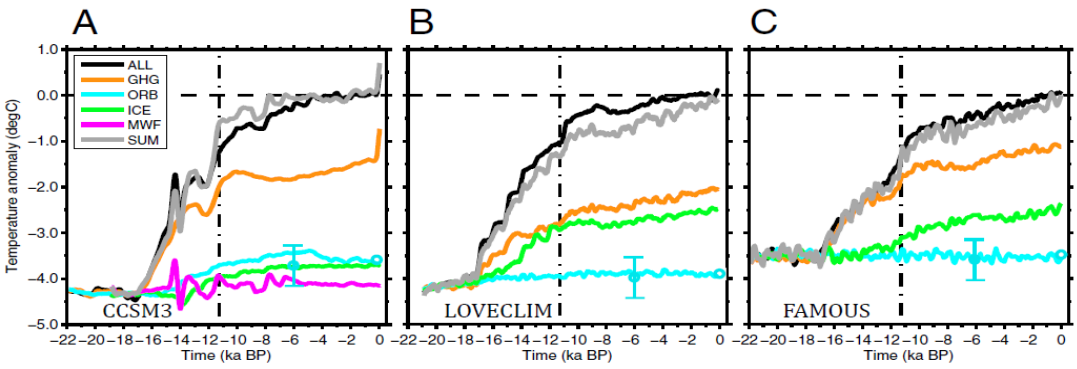
Reply: The linear response is the base for attributing the response to each forcing factors. Indeed, if a regional response is far away from a linear response (that is the sum response is far away from the total response), the attribution of forcing factor becomes not very useful, because there will be large amount of total response that can't be attributed to any forcing. Indeed, our original motivation for this work was to understand a specific climate response in a particular region (Europe + North America) at millennial time scale. But, we realized that we didn't even know if this response is a good linear response to the four individual forcing in the first place. The work is then expanded systematically to all the regions and at all timescales. We have added more explanation on this point in the new section 1: introduction and section 4: summary and discussion.

9: P3L3: I do not think the results by Shakun and colleagues or Marsicek and colleagues allow statements about how reasonable CCSM3's climate sensitivity is.

Reply: We have relaxed the sentence to "suggesting a potentially reasonable climate sensitivity in CCSM3, at global and continental scales." We think this is a reasonable statement.

10: P3L7ff: I think the last part of this sentence requires a reference.

Reply: Thank you for your comments. The reference is Liu et al., 2014, Figure 2A. (We have changed the citation more specifically as: "Figure 2A of Liu et al., 2014"). From this figure we can see in the Holocene period have small change by CO_2 than in the deglacial which warming response is dominated by the response to CO_2. Since the ice sheet retreating is also an important warming effect in the deglaciation in LOVECLIM and FAMOUS, we have also added this factor in the sentence "as it removes the deglacial warming response that is dominated by the response to increased CO_2 and ice sheet retreat".



11: P3L8: I am not sure, I completely understand how the authors perform their binning.

Reply: Thank you for your comments. The binning simply means we use the 100 year mean data as one data point. So, the 11,000 years of data is binned into 110 data points, each representing the average of 100-year. Partly, this binning is to be consistent with Marsicek, et al (2018). We have made some clarification on this and sometimes used "mean" to replace "binning".

12: I find much of the method description on page 4 unclear. For example: I am not fully clear how the authors produce their various time series.

Reply: We apologize for the ambiguity. We have rewritten this part. Basically, the Loess fit is a low-pass filter. The filtered data is therefore described as low-pass filtered data. This paragraph of deriving the various time series have been written.

13: P4L19: Is an AR1 process appropriate or could other models be more appropriate?

Reply: We use the AR(1), instead of white noise, is because we test against the reduced degree of freedom in the low pass the data.

14: P4L23: Could the authors please give more details on their bootstrap. From my point of view, this description does not allow for reproduction of the significance tests.

Reply: Thank you for your comments. Take the 100-yr binned data for the Holocene for example. The ALL run global mean temperature time serial have 110 points of data, each representing a 100-yr bin. For one realization, the order of the data is swapped randomly. Then, the sum is used to compare this realization once to derive one Le. Since the randomly swapped realization is not related to the sum response, one should expect a large error Le. Here, we perform the random realizations for 1,000,000 times. This gives us 1,000,000 values of Le, forming the PDF of Le values. The minimum 5% level is then used as the 95% confidence level. One can reproduce the results using the matlab function bootstrap. An introduction on bootstrap is given in the reference Efron, 1979 or Wikipedia (https://en.wikipedia.org/wiki/Bootstrapping_(statistics)). These detailed explanations are added in the revision.

15: a) I understand that the author's interest is only in the linearity of the response. However, considering Figure 1, I think, the manuscript will benefit if the authors also discuss the visual discrepancy between the SUM and the ALL series. A difference of about 1K between both series between 11K BP and 3K BP is a relevant feature. Indeed, even if the response is largely linear according to the correlation, the potentially smaller nonlinearity appears to be more important here.

b) To add on this, I wonder whether the setup of the simulations really allow for the analyses? However, I am not familiar enough with the setup of the TraCE "single" forcing simulations.

Reply: Thanks for this good point. Indeed, the assessment of linear response depends on the time period and is a "global" measure here. It does not exclude the discrepancy at some times when the nonlinearity or internal variability can be large. Some comments on this point has been added in the discussion on Fig.1. "It should be noted that, the goodness of the linear response is based on the entire period and is meant for the response of the time scale to be studied. Therefore, even for a good linear response at long time scales, the sum response may still differ from the total response significant at some particular time. For example, for the orbital scale response in Fig.1b, even though the linear response is good according to the correlation and L_e, there is a 1^{o}C difference between the sum and total at 11ka and 3ka. Therefore, for the orbital scale response, the linear response mainly refers to the trend-like slow response comparable with orbital scale, instead of response features of shorter time scales." The dependence of the linear response assumption on time period is also discussed in section 4.

More descriptions of the TraCE-21ka simulation setup have been added. The relevance for the assessment of linear response has been discussed in the reply to question 2.

16: P5L28: Could the authors discuss this complexity in more detail, please?

Reply: Thank you for your comments. What we mean is that the goodness of linear response depends on the region and time scale. This sentence has been changed to: "This suggests that the goodness of the linear response depends on both the region and time scale. This further highlight the need to study the linear response at regional scales."

17: P6L5: As far as I can see the authors do not discuss the reason, at least not in depth.

Reply: Thank you for your comments. The reason is discussed in section 3.3. For orbital variability (Fig.5a1-a3), the linear response is strong in most regions in the NH across all three spatial scales, with the correlation coefficients above 0.8. In the SH, the linear response is also strong over the continents, but is poor over the ocean. This leads to the significantly reduced linear response in the SH as discussed in Fig.3a-4a. Since there are more continents in NH than SH and generally speaking, the linear response in continents strong than oceans, the linear response in the NH is better than SH. This is only an explanation from one angle, certainly not a full explanation. We have modified this sentence to: "Part of the reason of the stronger linear response in the NH than over SH will be discussed later"

18: P6L8: Could the authors please be specific, why this should be treated with caution.

Reply: Usually, the linear response becomes better for larger spatial scale, because the large spatial average suppresses noise (internal variability). This case is the opposite. A note is added on this.

19: Considering the centennial time scale: the authors diagnose that the linear response on centennial scales is poor. However, at least in the correlations in Figure 3, there appear to be many regions where linearity still is of modest importance. That is while I agree with the assessment that there is "no strong" linearity, the authors appear to dismiss linearity on centennial scales to easily on page 6.

Reply: Our statement of "no strong" linearity is derived the statement on the centennial variability: "The median linear response on the centennial timescale in either hemisphere across spatial scales ($f>3$, Fig.3c and Fig.4c) is no longer significant, with few correlation coefficients larger than 0.3 and contributing less than 10% of the variance." We are not very clear what the referee is inferring here.

20: P6L32: The description appears to exclude the continent of Australia.

Reply: Thank you for the careful observation. A note is added on Australia.

21: P7L1: a) Could the authors please discuss later on, why the response over the continents should be different from the oceans on these very long time scales. b) Could they please also discuss what the strong internal variability over southern oceans implies for reconstruction efforts.

Reply: Good questions! We don't know the reason. A comment is added. We plan to further explore this in the future.

22: Could the remaining ice sheets and the last freshwater forcing implementations influence the results generally and specifically the poor linearity over North America?

Reply: Again, this is a good question. We plan to further explore the physical mechanism of the linearity response in the future.

23: P7L18: It suggests so for this set of simulations.

Reply: We added "in this model". A general note on the dependence of our results to model, time period, climate variable, et al, is also added in section 4.

24: P7L29: The authors write at a number of instances "North America" but the results differ notably within North America if I interpret the visualizations correctly.

Reply: Thank you for the careful observation. We have changed North America to Canada.

25: P7L31ff, P8L11ff, Figure 7: I am not sure whether these parts add anything to the other analyses.

Reply: This figure is meant to give some intuition of the scatter of the relationship between correlation and SNR.

26: P8L6: I appear to be unable to see the poor SNR over southern North America.

Reply: We clarified it now as “the North America continent outside the central North America”.

27: Figure 1: Are the linear errors really the same in panels (a) and (b)?

Reply: Thank you for your comments. Yes, they are same. I have check it. The Le of Fig.1a is 0.632 and Fig.1b is 0.626. So they have different in the third decimal. But in this paper we only keep two decimal.

Technical Comments

T1: If I understand it correctly, the manuscript will receive language editing by Copernicus if it is accepted. Nevertheless, I think it will help further reviews if the authors check the language everywhere for clarity and grammatical correctness.

Reply: Thank you for your comments. We have gone through the manuscript carefully several times.

T2: Some of the Figures (particulary Figures 3 to 6) are not publication ready. While I assume that Copernicus is going to assist the authors in this if the manuscript is accepted, it probably would shorten the time between submission and final publication if the authors improve on the Figure-quality for the next round of reviews already.

Reply: We have improve the figures.

T3: Could the authors please check that all Figure captions are correct. I was not sure.

Reply: We have checked it again.

T4: P4L20: I think "valid" is not the correct expression, here.

Reply: We didn’t find valid in P4L20, but in P7L20. We have change the wording.

T5: P8L2: Could the authors please skip the exclamation mark.

Reply: Thank you for your comments. All right, fixed it.

T6: Acknowledgements: I think the authors have to acknowledge the repository or the persons which/who produced and provided the data. (I assume this was the Climate Data Gateway at NCAR.)

Reply: Thank you for your comments. All right, fixed it.

Reply to the Anonymous Referee #3:

Interactive comment on "On the linearity of the temperature response in Holocene: the spatial and temporal dependence" by Lingfeng Wan et al.

Anonymous Referee #3

Received and published: 26 February 2019

The linearity of the externally forced temperature evolution during the Holocene is investigated using climate model simulations forced by the total or by individual external forcing factors. In particular, it is tested whether the total forced Holocene temperature variability is a superposition/sum of the individual externally forced temperature responses. Moreover the linearity of the forced temperature response is tested on different spatial and temporal scales. The addressed topic is interesting and important.

Reply: Thank you for your comments. We have revised the manuscript according to your valuable comments.

Major comments:

- please revise the method section. Sometimes it is not clear what was done and why it was done. Please see specific comments below.

Reply: Thank you for your comments. We have revised the method section substantially, adding much more details and clarifications on the model setup, data processing, bootstrap method, et al. see replies to referee #1 and #2.

- the discussion should be more extensive, in particular the limitations of the study (please see the following remarks)

Reply: Thank you for your comments. The final section has been written, with much more complete discussion on the limitation of the study here.

- only a single simulation for each forcing is available. Therefore, a correct definition of external and internal variability is not possible. The internal variability likely differs between the individual simulations and the internal variability is likely not constant during the individual simulations. By summing up the four individual simulations it is not certain that the internal variability cancels out. Moreover, the internal variability might depend on the time and spatial scale. In addition, the ALL-forcing experiment still includes the internal variability. Please make this more clear in the text and discuss.

- an ensemble of Holocene simulations with that model is not available. Therefore, although incorrect, because the internal variability might depend on the forcing, it might be useful to get an estimate of the internal variability of the different time and spatial scales from a long control simulation with the same model.

Reply: Yes. Section 1 and 4 have been written to clarify this issue. Also, see reply to reviewer #1 on the general questions.

- I am wondering if it makes sense to investigate the shorter time and also partly the regional scales if only one ensemble member is available. The signal to noise ratio on the shorter time and regional scales might require a larger ensemble size to make a robust statement? Using a control simulation - please see previous point - an estimate of the signal to noise rate might be possible.

Reply: Agreed. This is only a rough estimation. See the revised section 1 and 4 on the limitations.

- I am wondering if the following definition is useful: "Since our study above shows that the linear response is largely valid for orbital and millennial variability, but not for centennial and decadal variability, we define the variance of the orbital and millennial variability crudely as the linear signals, while define the variance of the sum of the centennial and decadal variability, which is dominated by internal variability, as the linear noise." Please comment.

Reply: Given the single realization we have, there is no precise way of separating signal and noise. In this particular case, since all the four forcing factors are at orbital and millennial time scales, the forced signal should be in these long time scales, and the noise should be at shorter time scales, if linear response is assumed (which is largely confirmed). So, this gives a rational to for our crude estimation of signal and noise. If, for example, we discuss volcanic forcing and solar variability, this separation of signal and noise is no longer effect and an ensemble is necessary. This has been discussed now in the revised section 1 and 4.

- Laepple and Huybers (2014) have shown that "a multiproxy estimate of sea surface temperature variability that is consistent between proxy types and with instrumental estimates but strongly diverges from climate model simulations toward longer timescales. At millennial timescales, model-data discrepancies reach two orders of magnitude in the tropics, indicating substantial problems with models or proxies". Please discuss the implications in the context of the findings

Reply: A good comment. Our conclusion is valid only for this model. If the model internal variability is indeed so much lower than in the observation, the implication of this study to the real world will be limited. This point is added now in section 4. It is an interesting issue to be explored in the future.

- please describe the filtering method in more detail. It is not clear to me what kind of polynomial was used for the LOESS. Moreover, it is not clear whether the authors used several iteration to get more 'robust' estimates. More important, what is the influence of the LOESS-filtering method on the result, in particular on the linearity of the response.

Reply: The filtering is discussed in more detail now. LOESS is used here only as one low pass filter. We use this to be consistent with Marsicek et al (2018). (Our original motivation is to interpret the millennial variability found in Marsicek et al). Marsicek et al also verified the locally weighted regression (Loess) by generalized additive model (GAMM) fit. We test some of our results simply using running mean and the results remain qualitatively similar.

- please describe the method - used to compute the significance of the correlation – in more detail. If I understand the authors correctly, an AR1 process is only fitted to the ALL-forcing simulation on the different time scales. The Monte-Carlo method is then used to produce an ensemble (PDF) of fitted curves. Then the correlations between the fitted curves and the ALL forcing run are computed and the 95% confidence level is determined afterwards. If I understood the authors correctly, I am wondering if this method is sufficient. I would think that an AR1 process has to be fitted to the ALL forcing run and the superposition (sum of the response of the four individual simulations). Then two ensembles - one for the ALL forcing and one ensemble for the superposition – have to be computed using the Monte-Carlo method. The correlations between these two ensembles have to be used to determine the confidence level. Please make also more clear why you choose the AR1 as a benchmark and how robust the parameter of the AR1 process is, in particular for the orbital time scale.

Reply: We think that the randomization on ALL should be sufficient. This is because the key here is to use randomization to destroy the serial relation between ALL and sum. This can be done by randomize either ALL or sum, or both of them. Indeed, we have tested both cases, randomizing one or both time series and confirmed they are the same.

The reviewer is correct in that, strictly speaking, the AR(1) coefficient should be different for each region and should be used for the test of significance. Here, we used the global mean as a common test, mainly for simplicity. Most importantly, our focus here is on the linear response features over the globe, between different regions. Therefore, a common test makes it easy for comparison among different spatial scales and regions. For example, if regional tests are performed, it will be impossible to plot the significance test on the summery figure of Fig.3 and Fig.4 for comparison of different spatial scales. Similarly, it will be hard to compare the value as well as the significance among different regions and spatial scales in Fig.5 and 6. In addition, the global mean AR(1) is meant as a crude representation of most AR(1)'s for different regions. Indeed, except for the orbital scale, the global mean AR(1) is larger than most of the regional AR(1) so that the global mean AR(1) serves as a stricter test. At the orbital scale, the global mean AR(1) is about the middle of the regional AR(1)'s. Finally, we did emphasize that, if one's focus is on a specific region, the regional AR(1) should be used for re-evaluation of the significance. These points are now discussed explicitly in section 2.2.

- it is not clear to me why the authors did not do a spectral analysis of the runs like e.g. wavelet analysis, power spectrum, cross power spectrum ...

Reply: Our study is a first preliminary study. Our interests here is mainly on the linear responses on slow time evolution at the orbital and millennial scales. Given only 11,000 years, it is difficult to derive spectral details with high significance. Nevertheless, we agree it will be interesting to explore the spectral features in the future.

- why was the analysis based on the model grid and not on climate modes using e.g. EOF analysis?

Reply: Fixed region is more practical for using model to interpret the real world proxy. Our original motivation is to interpret the regional climate response over North America and Europe as discussed in Marsicek et al (2018). For overall climate response in the model, it is a good idea to perform this in the EOF space.

Minor comments:

- please be more precise (whole text): please rewrite sentences like 'the linear response is strong' => the response is almost linear; the response is similar to that of a linear system

Reply: Thank you for your comments. We have attempted to clarify these terminologies.

- whole text: I would prefer: forcings => forcing factors

Reply: Done!

- page 3, line 8-9: Please rewrite the sentence

Reply: We have deleted it here and explained the data processing in much more details later in 2.2 as follows:

"To the time scale, we decompose a full 11,000-yr annual temperature time series (from 11 ka to 0 ka) in 100-yr bins (a total of 110 data bins, or points, each representing a 100-yr mean) into three components. The three components are to represent the variability of, roughly, orbital, millennial and centennial timescales. Following Marsicek et al. (2018), we derive the orbital and millennial variability using a low-pass filter called the locally weighted regression fits (Loess fits) (Cleveland, 1979). First, the orbital variability is derived by applying a 6500-yr Loess fit low-pass filter onto the temperature time series, and therefore contains the trend and the slow evolution longer than ~6500 years. Second, we apply a 2500-yr Loess fit low-pass filter onto the temperature time series; then, we derive the millennial variability using this 2500-yr low-pass data subtracting the 6500-yr low-pass data. Finally, centennial variability is derived as the difference between the 100-yr binned temperature time series and the 2500-yr low-pass time series. In addition, we also derive a decadal variability time series. First, we compile the 10-yr bin time series from the original 11,000-yr annual time series (of a total of 1,100 data points, each representing a 10-yr mean). Second, we apply a 100-yr running mean low-pass filter on the time series of the 10-yr binned data. Finally, decadal variability is derived by using the 10-yr binned time series minus its 100-yr running mean time series."

Holocene temperature response to external forcing: Assessing the linear response and its spatial and temporal dependence

Lingfeng Wan[1, 2], Zhengyu Liu[3, 4], Jian Liu[1, 2, 4], Weiyi Sun[1, 2], Bin Liu[1, 2]

[1]Key Laboratory for Virtual Geographic Environment, Ministry of Education; State Key Laboratory Cultivation Base of Geographical Environment Evolution of Jiangsu Province; Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application; School of Geography Science, Nanjing Normal University, Nanjing, 210023, China.
[2]Jiangsu Provincial Key Laboratory for Numerical Simulation of Large Scale Complex Systems, School of Mathematical Science, Nanjing Normal University, Nanjing, 210023, China.
[3]Atmospheric Science Program, Department of Geography, Ohio State University, Columbus, OH43210, USA.
[4]Open Studio for the Simulation of Ocean-Climate-Isotope, Qingdao National Laboratory for Marine Science and Technology, Qingdao, 266237, China.

Correspondence: Jian Liu (njdllj@126.com); Zhengyu Liu (liu.7022@osu.edu)

Abstract. Previous studies show that the evolution of global mean temperature forced by the total forcing is almost the same as the sum of those forced individually by orbital, ice sheet, greenhouse gases and meltwater in the last 21,000 years in three independent climate models, CCSM3, FAMOUS and LOVECLIM. This validity of the linear response is useful because it simplifies the interpretation of the climate evolution. However, it has remained unclear if this linear response is valid on other spatial and temporal scales, and, if valid, in what regions. Here, we using a set of TraCE-21ka climate simulation outputs, the spatial and temporal dependence of the linear response of the surface temperature evolution in the Holocene is assessed using correlation coefficient and a linear error index. The results show that the response of global mean temperature is almost linear on orbital, millennial and centennial scales in the Holocene, but not on decadal scale. The linear response differs significantly between the Northern Hemisphere (NH) and Southern Hemisphere (SH). In the NH, the response is almost linear on millennial scale, while in the SH the response is almost linear on orbital scale. Furthermore, at regional scales, the linear responses differ substantially between the orbital, millennial, centennial and decadal timescales. On the orbital scale, the linear response is dominant for most regions, even in a small area of a mid-size country like Germany. On the millennial scale, the response is still approximately linear in the NH over many regions, albeit less so on the orbital scale. Relatively, the linear response is degenerated somewhat over most regions in the SH. On the centennial and decadal timescales, the response is no longer linear in almost all the regions. The regions where the response is linear on the millennial scale are mostly consistent with those on the orbital scale, notably western Eurasian, North Africa, subtropical North Pacific, tropical Atlantic and Indian Ocean, likely caused a large signal-to-noise ratios over these regions. This finding will be helpful for improving our understanding of the regional climate response to various climate forcing factors in the Holocene.

1 Introduction

Long term temperature evolution in the Pleistocene is often believed, and therefore interpreted, to be driven mainly by several external forcing factors, notably, the orbit forcing, the greenhouse gases (GHGs), the continental ice sheets and the meltwater flux forcing. (Here, we treat the coupled ocean-atmosphere system as our climate system, such that Earth orbital parameters, GHGs, meltwater discharge and continental ice sheet are considered as external forcing). Implicit in this interpretation is the assumption that the response is almost linear to the four forcing factors, that is, the temperature evolution forced by the total forcing combined is approximately the same as the sum of the temperature responses forced individually by the four forcing factors. This linear response, if valid, simplifies the interpretation of the climate evolution dramatically, because each feature of the climate evolution can now be attributed to those on different forcing factors. One example is the global mean temperature evolution of the last 21,000 years (COHMAP members, 1988; Liu et al., 2014). It has been shown that the global mean temperature response is almost linear to the four forcing factors above in three independent climate models (CCSM3, FAMOUS and LOVECLIM, Fig.2 of Liu et al., 2014) with the temperature evolution forced by the total forcing almost the same as the sum of those individually forced by each forcing factor. Furthermore, this deglacial warming response is forced predominantly by the increase of GHGs, and is also contributed by the ice sheet retreat. This linear response, however, has not been assessed quantitatively for the climate evolution in the Holocene. The Holocene period poses a more stringent and interesting test of the linear response to different forcing factors, as it removes the deglacial warming response that is dominated by the response to increased CO_2 and ice sheet retreating (Figure 2A of Liu et al., 2014). An even more interesting and practical question here is, besides the global mean temperature response on the slow orbital time scale, how linear is the response at shorter time scales and more regional spatial scales, throughout the Holocene?

In general, the assessment of the linear response, in principle, can be done in a climate model using a set of specifically designed experiments. It requires experiments forced by the combined forcing as well as each individual forcing. Furthermore, each forcing experiment has to consist of a large number of ensemble members. This follows because a single realization can contain strong internal climate variability on a wide range of timescales (Laepple and Huybers, 2014), from daily variability of synoptic weather storms (Hasselmann, 1976), to interannual variability of El Nino (Cobb et al., 2013), interdecadal climate variability (Delworth and Mann, 2000), all the way to millennial climate variability (Bond et al., 1997). The ensemble mean is therefore necessary for suppressing the internal variability and then generating the forced response to each forcing. The goodness of the linear response can therefore be assessed by comparing the response to the total forcing with the sum of the individual responses. One practical problem with this ensemble approach is, however, the extraordinary computing costs, especially for long experiments in more realistic fully coupled general circulation models. A practical question is therefore: is it possible to obtain a meaningful assessment of the linear response using only a single realization of each forced experiment for the Holocene, such as those in TraCE-21ka experiments (Liu et al., 2014).

Strictly speaking, it is impossible to disentangle the forced response from internal variability in a single realization. This would make the assessment of linear response difficult. However, it is conceivable that, if our interest is the slow climate

evolution of millennial or longer time scales in response to the slow forcing factors such as the orbital forcing, ice sheet forcing, GHGs and meltwater flux, the assessment is still possible, albeit approximately, at least for very large scale variability. This follows because these forcing factors are of long time scales and of large spatial scales, the forced response signal should therefore also be on long time scales and large spatial scales, if the response is approximately linear. An extreme example is the almost linear response in the global temperature of the last 21,000 years as discussed by Liu et al., (2014). In contrast, internal variability in the coupled ocean-atmosphere system tends to be of shorter time scales, decadal to centennial, and of smaller spatial scales, at least in current generation of coupled ocean-atmosphere models. This naturally leads to two questions. First, how linear is the climate response at different spatial and temporal scales, quantitatively? Second, in what regions, the linear response tends to dominate? The answer to these questions should help improving our understanding of regional climate response during the Holocene. A further question is: if the linear approximation is valid, what is the contribution of each forcing factor in different regions and at different time scales. This question will be addressed in a follow-up paper (Wan et al., 2019).

In this paper, we assess the linear response for the Holocene temperature evolution quantitatively, using the 5 forced climate simulations in CCSM3 (Liu et al., 2014), with the focus on the spatial and temporal dependence of the linear response. We will assess the linearity response on the timescales of orbital, millennial, centennial and decadal, and of the spatial scales of global, hemispheric and regional. The data and methodology are given in section 2. The dependence of the linear response on spatial and temporal scales are analysed in section 3. A summary and further discussions are given in section 4.

2 Data and Methods

2.1 Data

The data is from TraCE-21ka (Liu et al., 2009, 2014), which consists of a set of 5 synchronously coupled atmosphere-ocean general circulation model simulations for the last 21,000 years. The simulations are completed using the CCSM3 (Community Climate System Model version 3). The simulation forced by the total forcing (experiment ALL) is forced by realistic continental ice sheets, the GHGs, orbital forcing and melting water fluxes. The ice sheet is changed approximately once every 500 years, according to the ICE-5G reconstruction (Peltier, 2004). The atmospheric GHGs concentration is derived from the reconstruction of Joos and Spahni (2008). The orbital forcing follows that of Berger (1978). The coastlines at the LGM were also taken from the ICE-5G reconstruction and were modified at 13.1ka, 12.9ka, 7.6ka, 6.2ka, after which the transient simulation adopted the present-day coastlines. The meltwater flux follows largely the reconstructed sea level and other paleoclimate information and, in the meantime, reconciles the response of Greenland temperature and AMOC strength in comparison with reconstructions. More information on the details of the experiment and forcing can be seen in He (2011) or the TraCE-21ka website http://www.cgd.ucar.edu/ccr/TraCE/.

The transient simulation under the total climate forcing reproduces many large scale features of the deglacial climate evolution consistent with the observations (Shakun et al., 2012; Marsicek et al., 2018), suggesting a potentially reasonable

climate sensitivity in CCSM3, at global and continental scales. In addition to the all forcing run (ALL), there are four individual forcing runs forced by the orbital forcing (ORB), the continental ice sheets (ICE), the GHGs (GHG) and meltwater forcing (MWF), respectively (Liu et al., 2014, Table 1). In these four experiments, only one forcing varies the same as in experiment ALL, while other forcing/conditions remain the same as at 19ka. Therefore, this set of experiments can be used to study the linear response of the climate to the four forcing factors. Here, we will only examine the surface temperature response in the Holocene (last 11, 000 years).

It should be noted that the linear response can't be assessed if the individual forcing experiments are performed with the forcing superimposed one-by-one. In this approach, the four external forcing is added sequentially, for example, first the ice sheet, second the ice sheet plus orbital forcing, third the ice sheet, orbital and GHGs, and finally, applying all four forcing of ice sheet, orbital, GHGs and melting water. In this experimental design, the full forcing response is by default the response of the sum response after adding the four forcing factors together, and therefore can't be used to test the linear response. It should also be noted that our four individual forcing experiments, although in principle feasible for assessing linear response, are not designed optimally for the study of Holocene climate. This is because, except for the variable forcing, all the other three forcing factors is fixed at the 19ka condition. As such, the mean state is perturbed from the glacial state, not a Holocene state. This may have contributed to some unknown deterioration on the linear response discussed later. Nevertheless, we believe, our major conclusion should hold approximately. This is because, partly, the response is indeed almost linear for orbital and millennial variability as will be shown later.

2.2 Methods

We use two indices to evaluate the linear response: the temporal correlation coefficient r and a normalized linear error index L_e. The correlation coefficient is calculated as

$$r = \frac{\frac{\sum_{t=1}^{n}(S_t-\overline{S_t})\times(T_t-\overline{T_t})}{n}}{\sqrt{\frac{\sum_{t=1}^{n}(S_t-\overline{S_t})^2}{n}}\sqrt{\frac{\sum_{t=1}^{n}(T_t-\overline{T_t})^2}{n}}} = \frac{cov(S_t,T_t)}{\sigma(S_t)\sigma(T_t)} \quad (1)$$

Here, $S_t = \sum_{i=1}^{4} T_i/4$ is the linear sum of the temperature time series T_i of the four single forcing experiments, T_t is the full temperature time series in the ALL run, both at time t, and n is the length of the time series. The overbar represents the time mean. The correlation coefficient represents the similarity of the temporal evolution between the sum response and the ALL response. However, the correlation does not address the magnitude of the response. Indeed, even if S_t and T_t has a perfect correlation r=1, the two time series can still differ by an arbitrary constant in their magnitudes. Therefore, we will also use a normalized linear error index L_e to evaluate the magnitude of the linear response. Here, L_e is defined as the root mean square error (RMSE) of the sum temperature response from the full temperature response divided by the standard deviation of the full temperature response in the ALL run:

$$L_e = \frac{RMSE}{STD} = \frac{\sqrt{\frac{\sum_{t=1}^{n}[(S_t-\overline{S_t})-(T_t-\overline{T_t})]^2}{n}}}{\sqrt{\frac{\sum_{t=1}^{n}[(T_t-\overline{T_t})]^2}{n}}} = \frac{\sigma(S_t-T_t)}{\sigma(T_t)} \quad (2)$$

In general, a large r (close to 1) and a smaller L_e (close to zero) represents a better linear approximation, with r=1 and $L_e = 0$ as a perfect linear response. Therefore, if r is close to 1 and L_e is close to zero, we can conclude that the response is close to linear. With a single realization here, however, our assessment of linear response has limitations. First, if r is sufficiently small and L_e is sufficiently large, we can't confirm the response is linear. But, we can't conclude the response is nonlinear either, because the small r or large L_e can also be contributed by strong internal variability. Second, if the forcing is dominated by shorter time scale variability, say interannual to interdecadal variability, as in the case of volcanic forcing or solar variability, it will be difficult to assess the linear response. This because the time scales of forced response now overlap heavily with strong internal climate variability in the coupled system, and it will be difficult to separate the forced response from internal variability without ensemble mean.

The dependence of the linear response on spatial and temporal scales will be studied by filtering the time series in different scales. For the spatial scale, we will divide the globe into 9 succeeding cases, denoted by 9 division factors: *f=0, 1, 2, 3, 4, 6, 8, 12* and *24* from the largest global scale to the smallest model grid scale. The *f=0* case is for global average while the *f=1* case is for hemispheric average in the NH and SH. Further division will be done within each hemisphere. Note that each hemisphere has $96(lon.) \times 24(lat.)$ grid boxes, with a ratio of 4:1 between longitude span and latitude span. We divide each hemisphere into $f \times f$ sections of equal latitude and longitude spans, with each area containing the same number of $(96/f) \times (24/f)$ grid boxes, maintaining the ratio of 4:1 between longitude span and latitude span. For example, *f=2* is for the 2×2 division, with each area containing 48×12 grid boxes; *f=24* is for the 24×24 division with each area containing 4×1 grid boxes, about the size of $15°(lon.) \times 3.75°(lat.)$, like the size of a mid-size country of Spain or Germany in the mid-latitude.

To the time scale, we decompose a full 11,000-yr annual temperature time series (from 11 ka to 0 ka) in 100-yr bins (a total of 110 data bins, or points, each representing a 100-yr mean) into three components. The three components are to represent the variability of, roughly, orbital, millennial and centennial timescales. Following Marsicek et al. (2018), we derive the orbital and millennial variability using a low-pass filter called the locally weighted regression fits (Loess fits) (Cleveland, 1979). First, the orbital variability is derived by applying a 6500-yr Loess fit low-pass filter onto the temperature time series, and therefore contains the trend and the slow evolution longer than ~6500 years. Second, we apply a 2500-yr Loess fit low-pass filter onto the temperature time series; then, we derive the millennial variability using this 2500-yr low-pass data subtracting the 6500-yr low-pass data. Finally, centennial variability is derived as the difference between the 100-yr binned temperature time series and the 2500-yr low-pass time series. In addition, we also derive a decadal variability time series. First, we compile the 10-yr bin time series from the original 11,000-yr annual time series (of a total of 1,100 data points, each representing a 10-yr mean). Second, we apply a 100-yr running mean low-pass filter on the time series of the 10-yr

binned data. Finally, decadal variability is derived by using the 10-yr binned time series minus its 100-yr running mean time series.

Given the different degrees of freedom especially among variability of different time scales, it is important to test the goodness of the linear response statistically on r and L_e on different time scales differently. The statistical significance of r for a particular timescale is tested using the Monte Carlo method (Kroese, 2011 and 2014; Kastner, 2010; Binder, 1997) with 1,000,000 realizations on the corresponding red noise in the AR(1) model (autoregressive model of order 1). As a reference, the significance level is tested against the global mean temperature series in the ALL run. For the total, orbital, millennial, centennial and decadal temperature time series, the 95% confidence levels are found to be 0.72 (with the AR(1) coefficient 0.96), 0.76 (0.97), 0.65 (0.95), 0.21 (0.31) and 0.19 (0.06), respectively. In this paper, this AR(1) test for global mean temperature is also used as the common significant test for different spatial scales and in different regions as well. This use of a common significance level is for simplicity here. First, the use of different regional AR(1) coefficient will make the comparison of the linear responses among different spatial scales (e.g. Fig.3 and 4) and different regions (Fig.5, 6) difficult. Second, except for the orbital time scale, the AR(1) coefficient for the global mean temperature is larger than most of the regional AR(1) coefficients (not shown), likely caused by the further suppression of internal variability in the global mean. As a result, the global mean AR(1) test serves actually as a more stringent test than the local AR(1) test. At the orbital scale, the global mean AR(1) coefficient is in about the middle of the regional AR(1) coefficients. The uncertainly of using the global mean AR(1) coefficient is therefore about the average of those of regional AR(1) coefficients. Third, and, most importantly, as our first study here, our focus is on the global features of the linear response. The difference among the AR(1) coefficients among different spatial scales and different regions is much smaller than that between different time scales here. Therefore, the global mean AR(1) can still provide an approximate guideline of the significant test properly at different time scales. In later studies, if one's focus is for a specific spatial scale and on a specific region, the regional AR(1) should be used for re-examination of the significance test.

Statistical significance of the L_e of a particular timescale is tested using a bootstrap method (Efron, 1979, 1993) with 1,000,000 realizations on the corresponding time series. Specifically, the bootstrap is done as follows, taking the global mean temperature as example. First, we will derive the L_e from one random realization on the temperature of the ALL run of the 100-binned data (110 points of data, each representing a 100-yr bin). For this random realization, the order of the original temperature time series is swapped randomly. Then, this realization is used as a new ALL response for comparison with the sum response of the four individual experiments to derive a L_e in eqn. (2). Since the random realization distorts the serial correlation time with the sum response, one should expect usually a large error L_e. Second, we repeat this process for 1,000,000 times on 1,000,000 random realizations; this will produce 1,000,000 random values of L_e, forming the PDF of the L_e. Third, the minimum 5% level is then used as the 95% confidence level. Fourth, this bootstrap is repeated on the time series of the original total variability, and the decomposed orbital variability, millennial variability, centennial variability and decadal variability of global mean temperature time series. This gives the 95% confidence levels as 1.23, 1.23, 1.21, 1.24 and 1.36, respectively. This suggests that when the RMSE is less than about 1.2-1.3 times of the total response, the linear

sum is not significantly different from the total response at the 95% confidence level. Finally, as for the correlation test, since our focus here is on the global feature of the linear response, for simplicity, the significance level derived from the global mean temperature is used as the common confidence level for all regional scales.

3 Results

3.1 Linear responses at different temporal scales

The global mean temperature provides a useful example to start the discussion of the dependence of the linear response on timescales. We first examine the linear response of the global mean temperature based on its components of orbital, millennial, centennial and decadal variability (Fig.1). Fig.1a is the total variability of global surface temperature derived from the ALL run and the sum of the four individual forcing experiments. The global temperature response is almost linear on the orbital and millennial scales throughout the Holocene (Fig.1b and 1c). The orbital scale evolution is characterized by a warming trend of about 1℃ from 11ka to ~4.5ka before decreasing slightly afterwards. This feature is captured in the linear sum albeit with a slightly smaller magnitude and an additional local minimum around 3ka (Fig.1b). The total variability is very similar to the orbital variability (r=0.99, Fig.1a vs Fig.1b). The millennial variability shows 5 major peaks around ~9.8ka, 7.8ka, 4.7ka, 3.7ka and 1.8ka. All these peaks seem to be captured in the sum response albeit with a slightly larger amplitude (Fig.1c). For the orbital and millennial variability, the correlation coefficients between the sum and the full responses are r=0.83 and 0.71, respectively, both significant at the 95% confidence level and explaining over 50% of the variance; the linear errors are L_e=0.63 and 0.92, respectively, also significant at the 95% confidence level. It should be noted, however, that the goodness of the linear response is based on the entire period and is meant for the response of the time scale to be studied. Therefore, even for a good linear response at long time scales, the sum response may still differ from the total response significant at some particular time. For example, for the orbital scale response in Fig.1b, even though the linear response is good in terms of r and L_e, there is a 1ºC difference between the sum and total responses at 11ka and 3ka. Therefore, for the orbital scale response, the linear response mainly refers to the trend-like slow response of comparable time scale of the orbital scale, instead of some response features of shorter time scales. Further down the scale, at centennial time scale, the global centennial variability appears also to exhibit a modest linear response (Fig.1d), with r=0.44 and L_e=1.21. But the linear response of the decadal variability is poor (Fig.1e), with r=-0.02 and L_e=1.99, which is not statistically significant at the 95% confidence level. The result of the analysis on global mean temperature is qualitatively consistent with previous hypothesis that the linear response tends to degenerate at shorter temporal scale, because of the smaller forced response signal and the presence of strong internal variability. It shows that, for global temperature, the response is approximately linear at orbital and millennial timescales, but becomes much less so at centennial scales and fails completely at decadal scales.

3.2 Linear responses at different spatial scales

In order to assess the linear response at different regional scales, we first analyse the linear response of the NH and SH (*f=1*). It is interesting that the linear response significantly differs between the NH and SH (Fig.2). Fig.2a and 2f are the total variability of hemispheric surface temperature of the ALL response and sum response for the NH and SH, respectively. Their components on the 4 timescales (i.e. orbital, millennial, centennial and decadal) are shown in the Fig.2b-2e and Fig.2g-2j, respectively. In the NH, the response is almost linear at the millennial scale (r=0.82, L_e=1.01, Fig. 2c), but not so strong on orbital (r=0.55, L_e=0.84, Fig.2b) and centennial (r=0.32, L_e=2.29, Fig.2d) scales, where only L_e is significant on the orbital scale, and only r is significant on the centennial scale. At the decadal scale, the linear response fails completely (r=-0.04, L_e=1.99, Fig.2e). In comparison, in the SH, the linear response is dominant at the orbital scale (r=0.92, L_e=0.43, Fig.2g), but poor on all other timescales, including millennial (r=-0.12, L_e=2.32, Fig.2h), centennial (r=0.14, L_e=3.19, Fig.2i) and decadal (r=0.03, L_e=2.07, Fig. 2j) scales. The linear response of the global mean temperature discussed in Fig.1 therefore seems to be dominated by the SH response on the orbital scale, but by the NH response on the millennial scale. This suggests that the goodness of the linear response depends on both the region and time scale. This further highlight the need to study the linear response at regional scales.

The linear response at different spatial scales and on the orbital, millennial, centennial and decadal timescales are summarized in Fig.3 and Fig.4 in the correlation coefficient and linear error index, respectively. Fig.3a shows the correlation coefficients of the orbital variability in each region for the 9 division factors. The cases of global mean (*f=0*) and hemispheric mean (*f=1*) have been discussed in details before. The correlation coefficients for succeeding division factors (*f=2, 3, …, 24*) show several features. First, as expected, the correlation coefficient tends to decrease towards smaller area (larger *f*). Quantitatively, however, the correlation coefficient does not decrease much, such that even at the smallest area (*f=24*), the correlation in most regions are still above 0.8, statistically significant at 95% confidence level. This suggests that the response at the orbital scale is almost linear over most regions, even at the smallest scale of about a mid-size country like Germany (*f=24*, $15°(lon.) \times 3.75°(lat.)$). Second, the linear response in the NH is slightly better than SH (for *f≥3*), a topic to be returned later. Third, subareas in both hemispheres show comparable linear response across all the spatial scales, with the median correlation all above ~0.8, except that of the NH mean temperature (*f=1).* The linear response of NH is not better than that of regional variability at smaller spatial scales (*f≥2*). This is opposite to the expectation that the linear response becomes more distinct for a larger area, because the average over a larger area tend to suppress internal variability more. This case, however, seems to be a special feature and should be treated with caution. The correlation coefficient therefore shows that, for orbital scale evolution, temperature response is dominated by the linear response over most of the globe, even at regional scales. These features are also consistent with the linear error analysis in Fig.4a.

Millennial variability also shows a weaker linear response for smaller scales, in both the correlation (Fig.3b) and linear error (Fig.4b). Quantitatively, for millennial variability, the response is still approximately linear in the NH over many regions, albeit less so than at the orbital scale. The correlation coefficients remain above 0.6 across most regions even at the smallest

division area (*f=24*), contributing to ~40% of the variance. In contrast to the orbital variability, where regions in both hemispheres show comparable linear response, the millennial variability shows that the response can't be confirmed linear in most regions in the SH, with the median no longer significant at 95% confidence level. Similar to the orbital variability, nevertheless, the responses are more linear in the NH than SH on all the spatial scales.

In contrast to orbital and millennial variability, almost no response can be confirmed linear for the centennial variability. The median linear response is no longer significant on the centennial timescale in either hemisphere across spatial scales (*f>3*, Fig.3c and Fig.4c), with few correlation coefficients larger than 0.3 and contributing less than 10% of the variance. Finally, the decadal variability exhibits absolutely no linear response over any spatial scales in both of the NH and SH (Fig.3d and Fig.4d).

The approximate linear response at the orbital and millennial scales suggest that these two groups of variability are generated predominantly by the external forcing. In contrast, the poor linear response of centennial and decadal variability suggest that these two groups of variability are caused mainly by the internal coupled ocean-atmosphere processes. It should be kept in mind that, in our single realization here, the poor linear response on centennial and decadal variability may also be contributed by nonlinear responses of the climate system. But, given the almost absence of forcing variability at this short time scale in our experiments, we do not think that the nonlinear response is the major cause of the poor linear response here.

3.3 Pattern of the Linear Responses

We now further study the pattern of the linear response. Fig.5 shows the spatial patterns of the correlation coefficients at orbital (Fig.5a1-a3) and millennial (Fig.5b1-b3) scales for three representative spatial scales, *f=3, 6* and *24* (the other factors are not shown because they are similar to the above mentioned three representative spatial scales). For orbital variability (Fig.5a1-a3), the response is almost linear in most regions in the NH in all three spatial scales, with the correlation coefficient above 0.8. In the SH, the response is also almost linear over the continents, except for over Australia, but is not linear over the ocean. This leads to the significantly reduced linear response in the SH as discussed in Fig.3a-4a. This suggests that orbital variability is likely forced predominantly by external forcing over continents. The overall poorer linear response over ocean than land, however, is puzzling. The orbital time scale is so long that one would expect a similar quasi-equilibrium response over both land and ocean. This issue deserves further study in the future. More specifically, at the regional scales, e.g. *f=6* and *24* (Fig.5a2 and 5a3), the response is almost linear over the western half of the Eurasian continent, Northern Africa, Central and South America, most NH oceans and SH tropical oceans, and the Antarctica Continent, as seen in the Fig.5a2 and 5a3 (only those significant correlation coefficients are shown). But the linear response is poor over the North America continent and the eastern Eurasian continent, and the entire Southern Ocean. Similar features can also be seen in the map of linear error (not shown).

For millennial variability, in the NH, the linear response shows a similar feature to that of the orbital variability, but the linear response is poor over almost all the SH. Fig.5b1-b3 show that the response is almost linear in most regions in the NH at the three spatial scales, with the correlation coefficient above 0.6. At the regional scale, e.g. *f=6* and *24* (Fig.5b2 and 5b3),

the response is almost linear over the northwestern Eurasian continent, Northern Africa, northern North America, northern Pacific Ocean, southern North Atlantic Ocean and western Arctic Ocean, as seen in the Fig.5b2 and 5b3 (only those significant correlation coefficients are shown). But the linear response is poor over the southern North America, the eastern and southern Eurasian continent. While in the SH, the linear relationship is poor over almost the entire SH as seen in the correlation map (Fig.5b1-b3). Interestingly, over the NH, the regions of linear response for millennial variability are mostly consistent with the regions of linear response for orbital variability, notably western Eurasian, North Africa, subtropical North Pacific, tropical Atlantic and Indian Ocean. These preferred region of linear response suggests potentially some common mechanisms of the climate response in these regions, in this model.

In order to understand the cause for the preferred regions of linear response, we examine the signal-to-noise ratio. Since our forcing factors are on millennial and orbital time scales, and the linear response is also largely valid for orbital and millennial variability, we use the variance of the orbital and millennial variability as a crude estimate of the linear response signal. Similarly, since there is no centennial and decadal forcing in our model and the response of centennial and decadal variability are not linear response, we use the variance of the sum of the centennial and decadal variability as a rough estimate for internal variability as the linear noise. The ratio of the signal to noise may therefore be able to shed light on the regional preference of linear response. Admittedly, this estimation is crude, limited by the single realization here. Figure 6 shows the signal-to-noise ratio for orbital and millennial variability for the three representative spatial scales (*f=3, 6* and *24*). For orbital variability (Fig.6a1-a3), the signal-to-noise ratio is large (above $10^{(0.6)}$, the log base 10 is taken on the signal to noise ratios) in most regions in the NH. In the SH, the signal-to-noise ratio is also large over the continents, but is small over the ocean. At different regional scales, e.g. *f=6* and *24* (Fig.6a2 and 6a3), the signal-to-noise ratio is large over the western half of the Eurasian continent, Northern Africa, Central and South America, most NH oceans and SH tropical oceans, and the Antarctic Continent. But the signal-to-noise ratio is small over Canada and the eastern Eurasian continent, and the entire Southern Ocean. These spatial features of signal-to-noise ratio on the orbital scale (Fig.6a1-a3) are similar to those in the correlation map (Fig.5a1-a3). The spatial correlation between the map of the signal-over-noise ratio in Fig.6 and the corresponding correlation coefficient in Fig.5 is 0.53, 0.58 and 0.49 for *f=3, 6* and *24* respectively, as seen in the scatter diagram in Fig.7, all significant at 95% confidence level.

For millennial variability, the signal-to-noise ratio also shows a similar feature to that of orbital variability although overall somewhat smaller (note the different color scales). Fig.6b1-b3 show that the signal-to-noise ratio is large in most regions in the NH in all three spatial scales, with the signal-to-noise ratio above $10^{(-1)}$ (the log base 10 is taken on the signal to noise ratios). At the regional scale, e.g. *f=6* and *24* (Fig.6b2 and 6b3), the signal-to-noise ratio is large over the northwestern Eurasian continent, Northern Africa, central North America, northern North Pacific Ocean, Northern Atlantic Ocean and the Arctic Ocean. But the signal-to-noise ratio is small over the North America continent outside the central North America, the South America and the eastern and southern Eurasian continent. While the signal-to-noise ratio is small over almost the entire SH. Over the NH, interestingly, the regions of large signal-to-noise ratio for millennial variability, are mostly consistent with the regions of large signal-to-noise ratio for orbital variability, notably northwestern Eurasian, North Africa,

subtropical North Pacific, Northern Atlantic and tropical Northern Indian Ocean. These features of spatial pattern of signal-to-noise ratio on the millennial scale (Fig.6b1-b3) are similar to those in the correlation map (Fig.5b1-b3). The correlation coefficient between the maps of signal-to-noise ratio and correlation is 0.65, 0.51 and 0.45 for *f=3, 6* and *24*, respectively (Fig.7), again all significant at 95% confidence level.

4 Summary and Discussions

In this paper, the linear response is assessed for the surface temperature response to orbital forcing, GHGs, meltwater discharge and continental ice sheet throughout the Holocene in a coupled GCM (CCSM3). The global mean temperature response is almost linear on the orbital, millennial, and centennial scales throughout the Holocene, but not for decadal variability (Fig.1). Furthermore, the sum response account for over 50% of the total response variance for orbital and millennial variability. Further analysis on regional scale suggests that the response is approximately linear on the orbital and millennial scales for most continental regions over the NH and SH, with the sum response explain over about 50% of the total response variance. However, the linear response is not significant over much of the ocean, especially over the ocean in the SH. There are specific regions where linear response tends to be dominant, notably the western Eurasian continent, North Africa, the central and South America, the Antarctica continent, and the North Pacific. The strong linear response is interpreted as the region of large signal-to-noise ratio. That is, in these regions, either the orbital and millennial response signal is large, or the influence of the centennial and decadal variability noise is small, or both. This suggests that the orbital and millennial variability in these regions are relatively easy to understand. This finding lays a foundation to our further understanding of the impacts of different climate forcing factors on the temperature evolution in the Holocene of orbital and millennial time scales. This understanding is our original motivation of this work. Further work is underway in understanding the contribution of different forcing factors on the temperature evolution (Wan et al., 2019).

The result here represents the first such assessment and is carried out for a single variable (surface temperature) in a single model (CCSM3) for the Holocene. It should therefore be kept in mind that the assessment could differ for different variables, in different models, for different periods and for different sets of forcing factors. For example, if we evaluate the precipitation response in the Holocene in CCSM3, the response is less linear than temperature (not shown); this is expected because precipitation response contains more internal variability and exhibits more nonlinear behaviour than temperature. The assessment will be also different if a different period is assessed, e.g. the last 21,000 years; with a large amplitude of climate forcing, the linear response may degenerate in the 21,000-year period. In addition, the assessment of linear response using only one realization will be difficult to perform for volcanic forcing and solar variability forcing; these forcing factors have short time scales and therefore their impacts will be difficult to separate from internal variability without ensemble experiments. Finally, it is also important to repeat the same assessment here in different models and to establish the robustness of the assessment. It should also be kept in mind that our assessment is implicitly related to the assumption that, at millennial and orbital time scales, internal variability is not strong relative to the forced responses. Although this seems to

be consistent in our model, there is a possibility that internal variability is severely underestimated in the model than in the real world (Laepple and Huybers, 2014). If true, the relevance of our model assessment to the real world will be limited.

Even in the context of this model assessment, much further work remains. Most importantly, the purpose of testing the linear response is for a better understanding of the physical mechanism of the climate response. It is highly desirable to understand why response tends to be linear in some region, but not in other regions. In particular, it is unclear why the linear response is preferred over land than over ocean for orbital and millennial variability. At such a long time scale, one would expect that the upper ocean response has reached quasi-equilibrium and therefore the surface temperature response over land and over ocean should not be too much different. Ultimately, we would like to assess and understand the physical mechanism of the climate evolution in different regions. These work are underway (Wan et al., 2019).

Acknowledgments

This work was jointly supported by the National Key Research and Development Program of China (Grant No. 2016YFA0600401), the National Natural Science Foundation of China (Grant No. 41420104002, 41630527), the Program of Innovative Research Team of Jiangsu Higher Education Institutions of China, and the Priority Academic Program Development of Jiangsu Higher Education Institutions (Grant No. 164320H116). We used the output from the full TraCE-21ka simulation, available at https://www.earthsystemgrid.org/project/trace.html.

References

Berger, A.: Long-term variations of daily insolation and quaternary climate changes. Journal of atmospheric sciences, 35(12), 2362-2367, doi: 10.1175/1520-0469(1978)035<2362:LTVODI>2.0.CO;2, 1978.

Binder, K.: Applications of Monte Carlo methods to statistical physics, Reports on Progress in Physics, 60(5), 487-559, doi: 10.1088/0034-4885/60/5/001, 1997.

Bond, G., Showers, W., Cheseby, M., Lotti, R., Almasi, P., deMenocal, P., Priore, P., Cullen, H., Hajdas, I., Bonani, G.: A Pervasive Millennial-Scale Cycle in North Atlantic Holocene and Glacial Climates, Science, 278(5341), 1257-1266, doi:10.1126/science.278.5341.1257, 1997.

Cleveland, W. S.: Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association, 74(368), 829-826, doi:10.2307/2286407, 1979.

Cobb, K. M., Westphal, N., Sayani, H. R., Watson, J. T., Lorenzo, E. D., Cheng, H., Edwards, R. L., Charles, C. D.: Highly variable El Nino-Southern Oscillation throughout the Holocene, Science, 339(6115), 67-70, doi: 10.1126/science.1228246, 2013.

COHMAP Members.: Climatic Changes in the Last 18,000 Years: Observations and Model Simulations, Science, 241(4869), 1043-1052, doi:10.1126/science.241.4869.1043, 1988.

Delworth, T. L. and Mann, M. E.: Observed and simulated multidecadal variability in the Northern Hemisphere, Climate Dynamics, 16(9), 661-676, doi:10.1007/s003820000075, 2000.

Efron, B. and Tibshirani, R. J. (Eds): An introduction to the bootstrap, Chapman and Hall/CRC, New York, 1993.

Efron, B.: Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics, 7(1), 1-26, doi: 10.1214/aos/1176344552, 1979.

Hasselmann, K.: Stochastic climate models Part I. Theory, Tellus, 28(6), 473-485, doi:10.3402/tellusa.v28i6.11316, 1976.

He, F.: Simulating Transient Climate Evolution of the Last Deglaciation with CCSM3, PhD thesis, 1–177 (Univ. Wisconsin-Madison, 2011).

Joos, F. and Spahni, R.: Rates of change in natural and anthropogenic radiative forcing over the past 20,000 years. Proceedings of the National Academy of Sciences, 105(5), 1425-1430, doi: 10.1073/pnas.0707386105, 2008.

Kastner, M.: Monte Carlo methods in statistical physics: mathematical foundations and strategies, Communications in Nonlinear Science and Numerical Simulation, 15(6), 1589-1602, doi:10.1016/j.cnsns.2009.06.011, 2010.

Kroese, D. P., Brereton, T., Taimre, T., Botev, Z. I.: Why the Monte Carlo method is so important today, Wiley Interdisciplinary Reviews: Computational Statistics, 6(6), 386-392, doi:10.1002/wics.1314, 2014.

Kroese, D. P., Taimre, T., Botev, Z. I. (Eds): Handbook of Monte Carlo Methods, Wiley Series in Probability and Statistics, John Wiley and Sons, New York, doi:10.1002/9781118014967, 2011.

Laepple, T. and Huybers, P.: Ocean surface temperature variability: Large model-data differences at decadal and longer periods. Proceedings of the National Academy of Sciences of the United States of America, 111(47), 16682-7, doi:10.1073/pnas.1412077111, 2014.

Liu, Z., Otto-Bliesner, B. L., He, F., Brady, E. C., Tomas, R., Clark, P. U., Carlson, A. E., Lynch-Stieglitz, J., Curry, W., Brook, E., Erickson, D., Jacob, R., Kutzbach, J., Cheng, J.: Transient Simulation of Last Deglaciation with a New Mechanism for Bølling-Allerød Warming, Science, 325(5938), 310-314, doi:10.1126/science.1171041, 2009.

Liu, Z., Zhu, J., Rosenthal, Y., Zhang, X., Otto-Bliesner, B. L., Timmermann, A., Smith, R. S., Lohmann, G., Zheng, W., Elison, T. O.: The Holocene temperature conundrum, PNAS (USA), 111(34), 3501-5, doi:10.1073/pnas.1407229111, 2014.

Marsicek, J., Shuman, B. N., Bartlein, P. J., Shafer, S. L., Brewer, S.: Reconciling divergent trends and millennial variations in Holocene temperatures, Nature, 554(7690), 92-96, doi:10.1038/nature25464, 2018.

Peltier, W. R.: Global Glacial Isostasy and the Surface of the Ice-Age Earth: The ICE-5G (VM2) Model and GRACE. Annual Review of Earth & Planetary Sciences, 20(32), 111-149, doi: 10.1146/annurev.earth.32.082503.144359, 2004.

Shakun, J. D., Clark, P. U., He, F., Marcott, S. A., Mix, A. C., Liu, Z., Otto-Bliesner, B., Schmittner, A., Bard, E.: Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation, Nature, 484(7392), 49-54, doi:10.1038/nature10915, 2012.

Wan, L., Liu, Z., Liu, J.: Assessing the impact of individual climate forcing on long term temperature evolution in the Holocene. In prep, 2019.

Table 1. TraCE-21ka simulation experiments

No.	experiment	Forcing	time	Resolution(lat×lon)
1	ORB	Orbital forcing	11000	48×96
2	GHG	GHGs forcing	11000	48×96
3	ICE	Ice sheets forcing	11000	48×96
4	MWF	Meltwater forcing	11000	48×96
5	ALL	Orbital + GHGs + ICE sheets + Meltwater forcing	11000	48×96

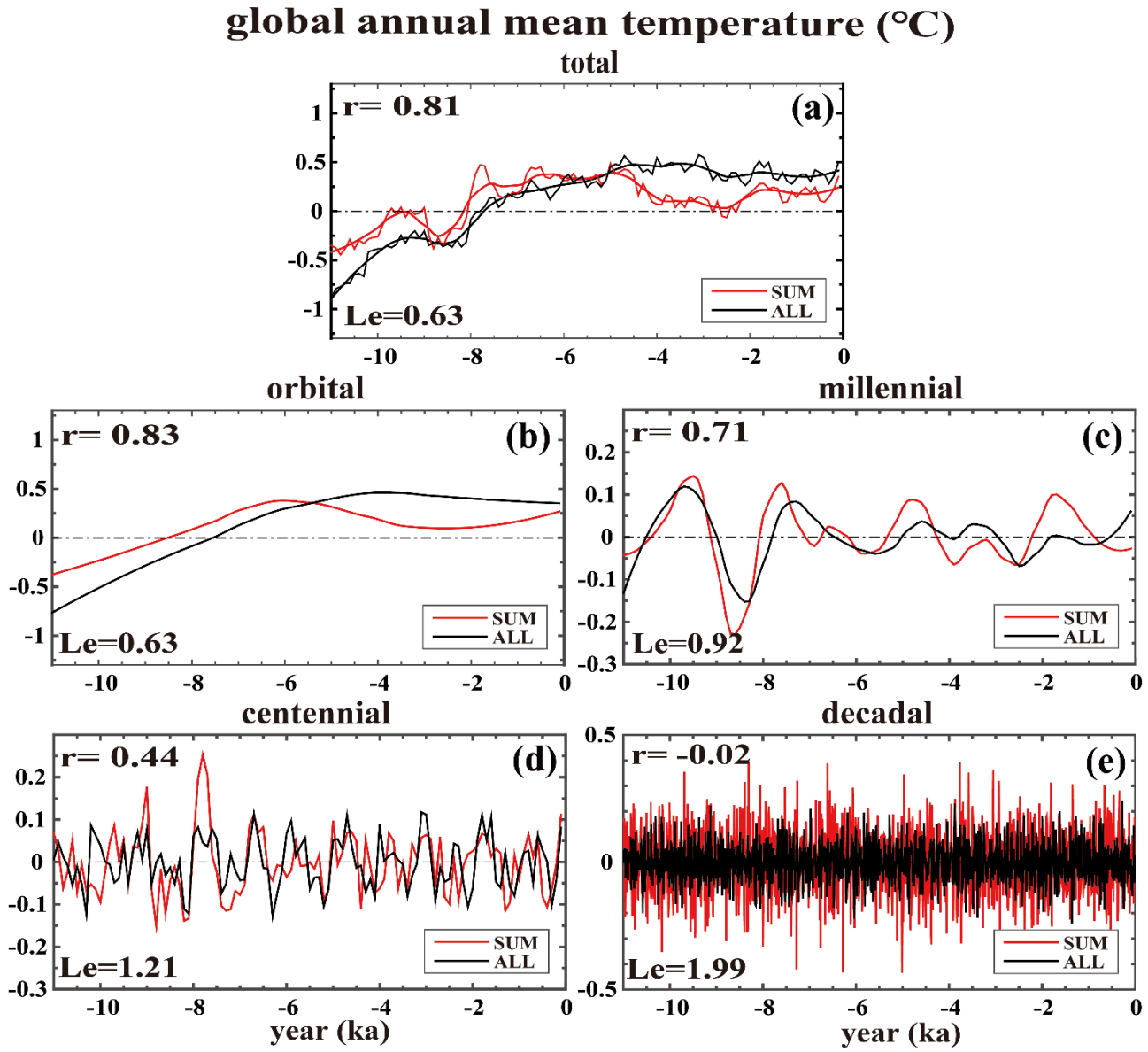


Figure 1: The global annual mean surface temperature time series derived from the ALL run (black) and the SUM (the sum of four single forcing run, red). In the (a), the thin line is the 100-yr binned time series, the thick line is 2500-yr loess fitted time series. The orbital scale variability (b) is represented by the 6500-yr loess fitted series. The millennial variability (c) is represented by the 2500-yr loess fitted data subtracting the 6500-yr loess fitted data. The centennial variability (d) is represented by the 100-yr binned data subtracting the 2500-yr loess fitted data. The decadal variability (e) is represented by the 10-yr binned origin time series subtracting its 100-yr running mean. The correlation coefficient (r) is given at the upper left corner and the linear error (L_e) is given at the lower left corner of each panel. The x axis is year (ka, 0 is AD 1950, negative is before AD 1950), and the y axis is temperature anomaly (°C, relative to 11ka-0ka).

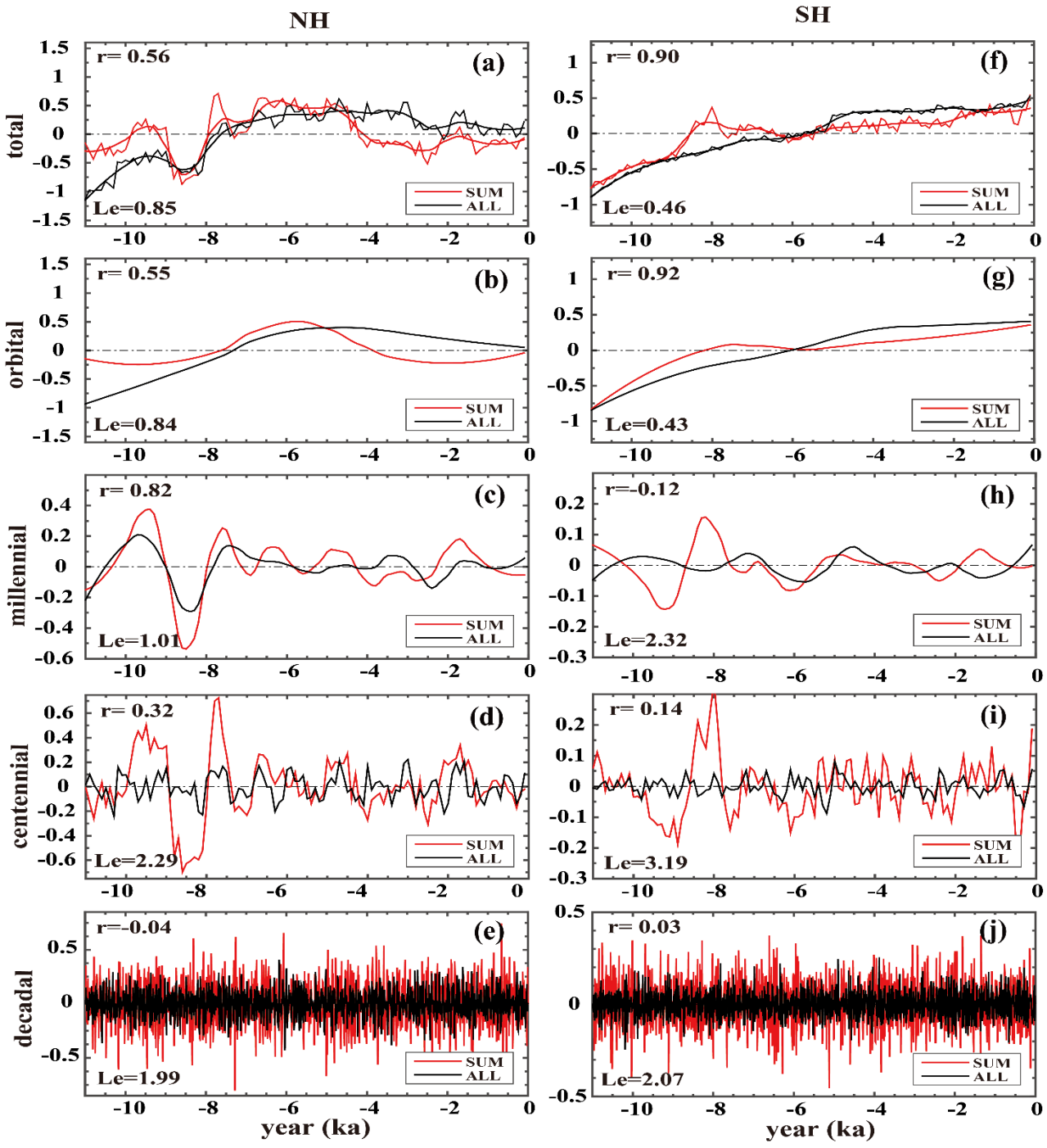


Figure 2: The surface temperature time series derived from the ALL run (black) and the SUM (red). The (a-e) is similar to the Fig.1a-e but for NH, and the (f-j) is similar to the Fig1.a-e but for SH. The x axis is year (ka, 0 is AD 1950, negative is before AD 1950), and the y axis is temperature anomaly (°C, relative to 11ka-0ka).

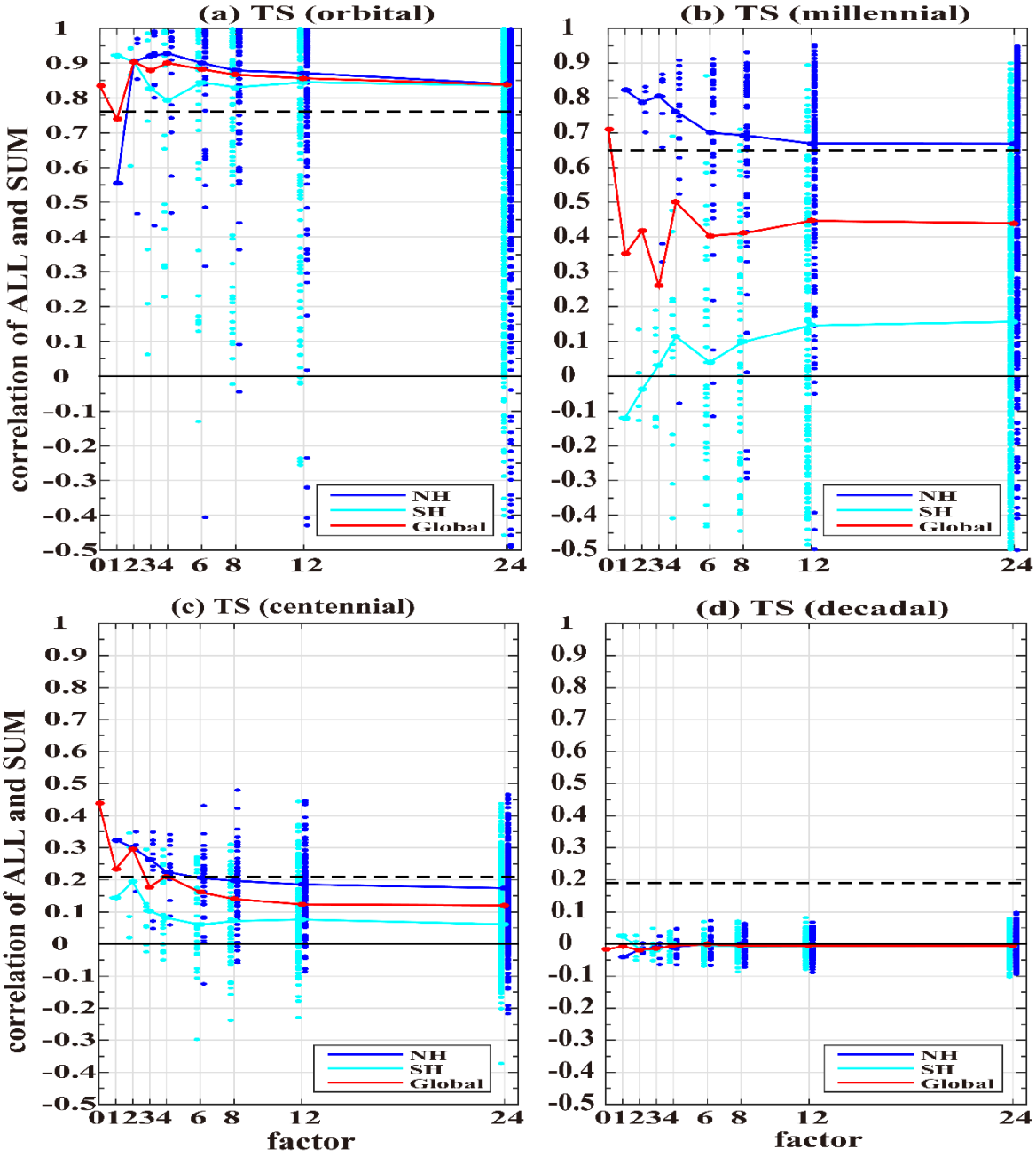


Figure 3: Correlation coefficient of mean surface temperature between the ALL and SUM outputs in different spatial-time scales. a, orbital; b, millennial; c, centennial; d, decadal. The blue dots for NH, cyan dots for SH and red dots for all the correlation coefficient median in the same spatial scale (i.e. factor) of global. The red line is connecting the median dots at all division factors. The blue (cyan) line is connecting the median of all blue (cyan) dots at all division factors. The black thick dashed line is 95% confidence level. The black thin solid line is zero. The x axis is division factor, and y axis is correlation coefficient. There are only about 10 points with the correlation coefficient lower than -0.5 so they are neglected in a-c.

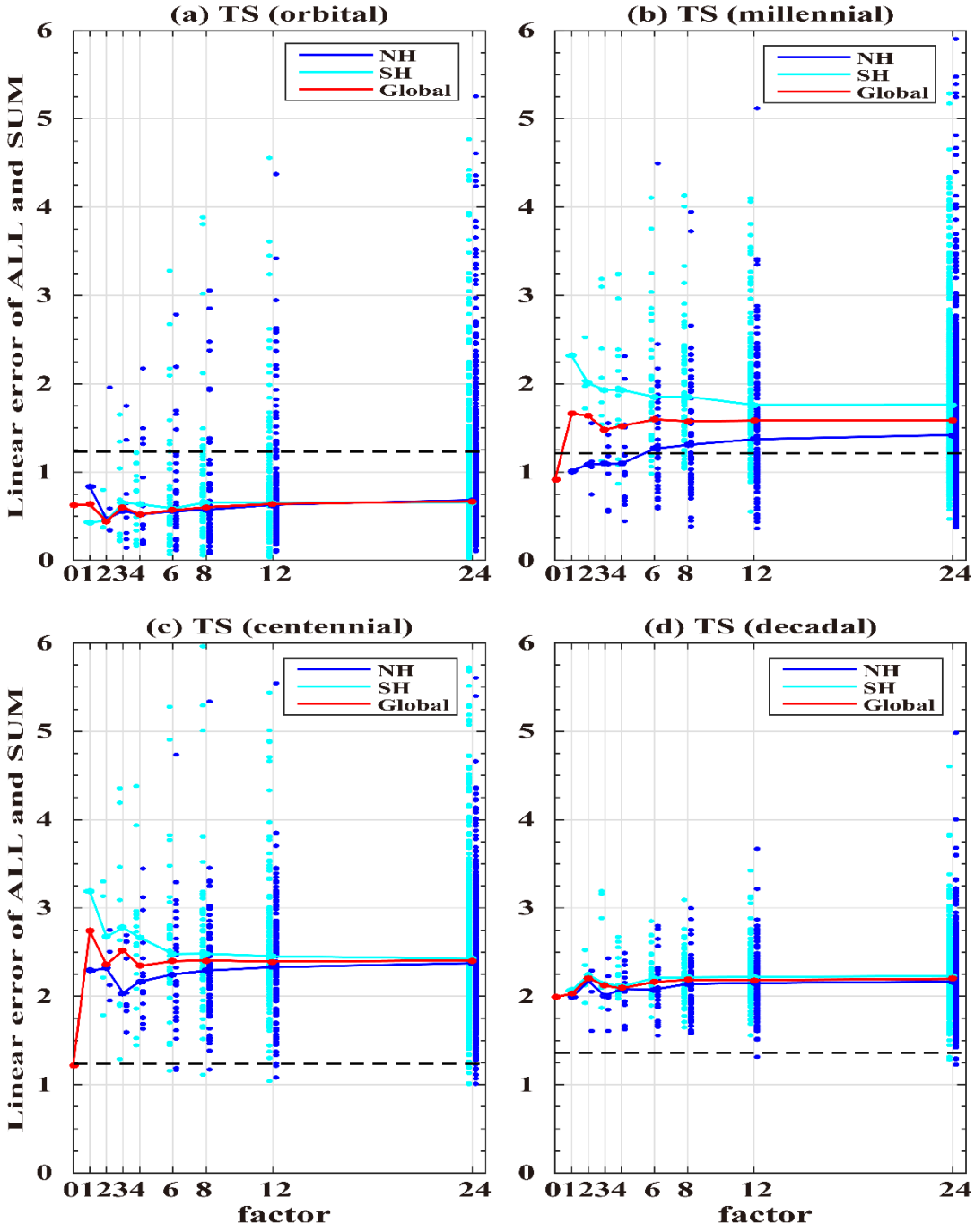


Figure 4: Same as Figure 3, but for linear error (L_e, y axis). The black thick dashed line is 95% confidence level of L_e. There are only about 10 points with the L_e larger than 6 so they are neglected in a-c.

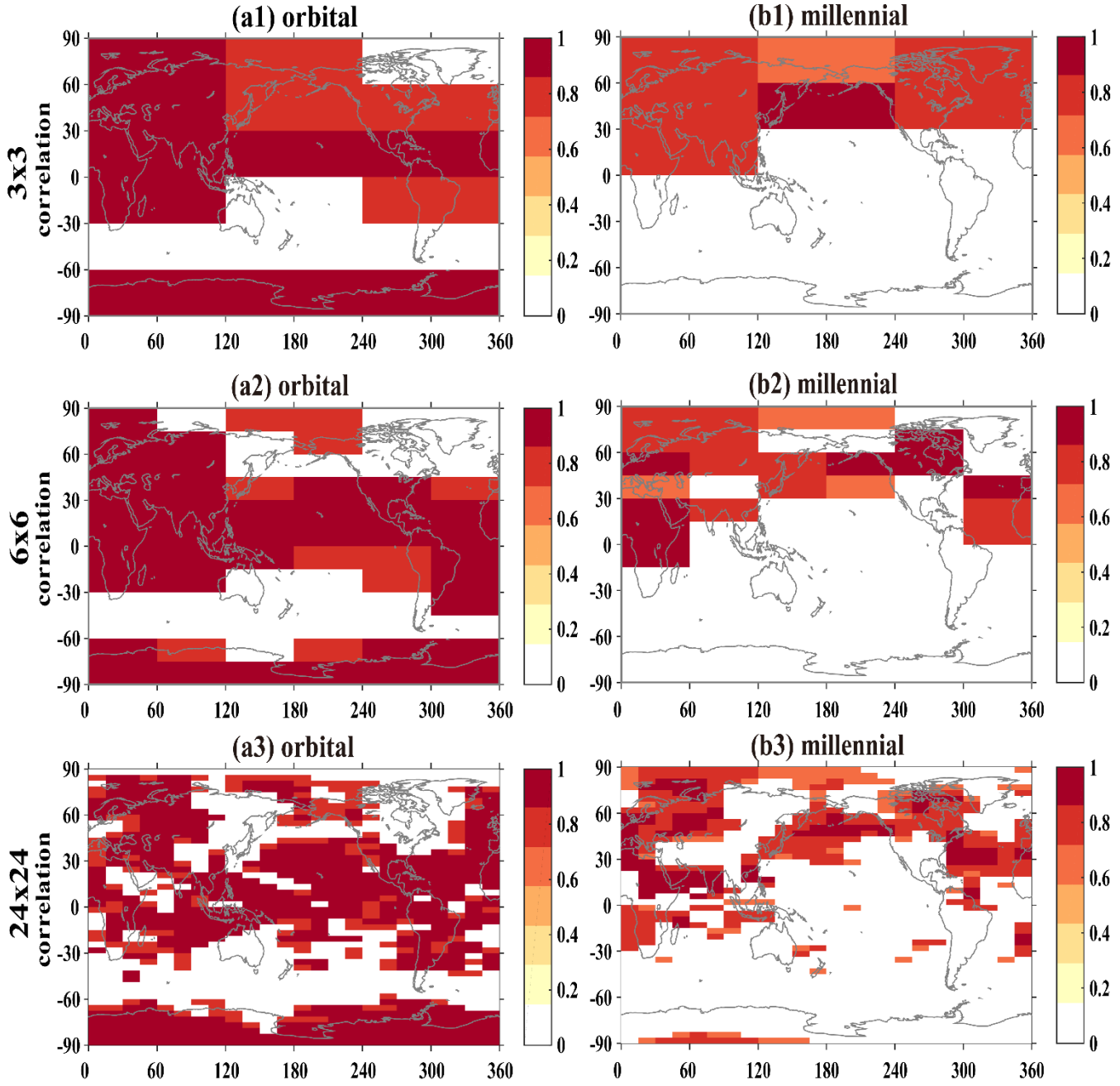


Figure 5: Correlation coefficient of mean surface temperature between the ALL and SUM outputs of the two timescales (a1-a3, orbital; b1-b3, millennial) for three representative spatial scales, *f*=*3*, *6* and *24* (the other factors are not shown because they are similar to the abovementioned three representative spatial scales). Only those regions of significant at 95% confidence level are shaded in colors.

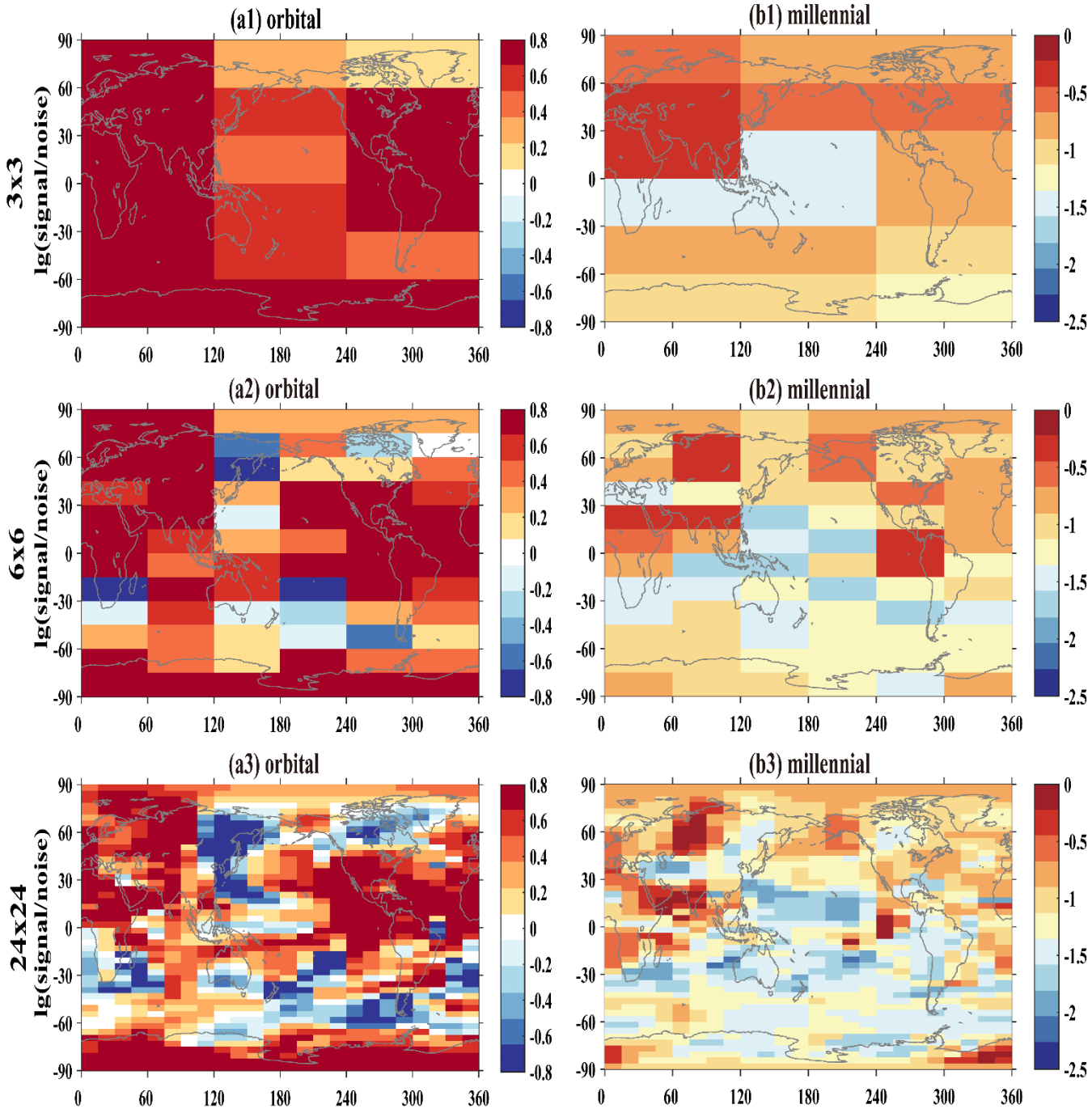


Figure 6: The signal-to-noise ratios on the orbital (a1-a3) and millennial (b1-b3) timescales derived from the ALL run. Here the signal is used by the orbital (millennial) variability variance, and the noise is used by the sum of the centennial and decadal variabilities variance. The numbers of 1, 2 and 3 are for the three representative spatial scales, *f*=3, *6* and *24*, respectively. In order to show the signal clearly, the log base 10 is taken on the signal to noise ratios.

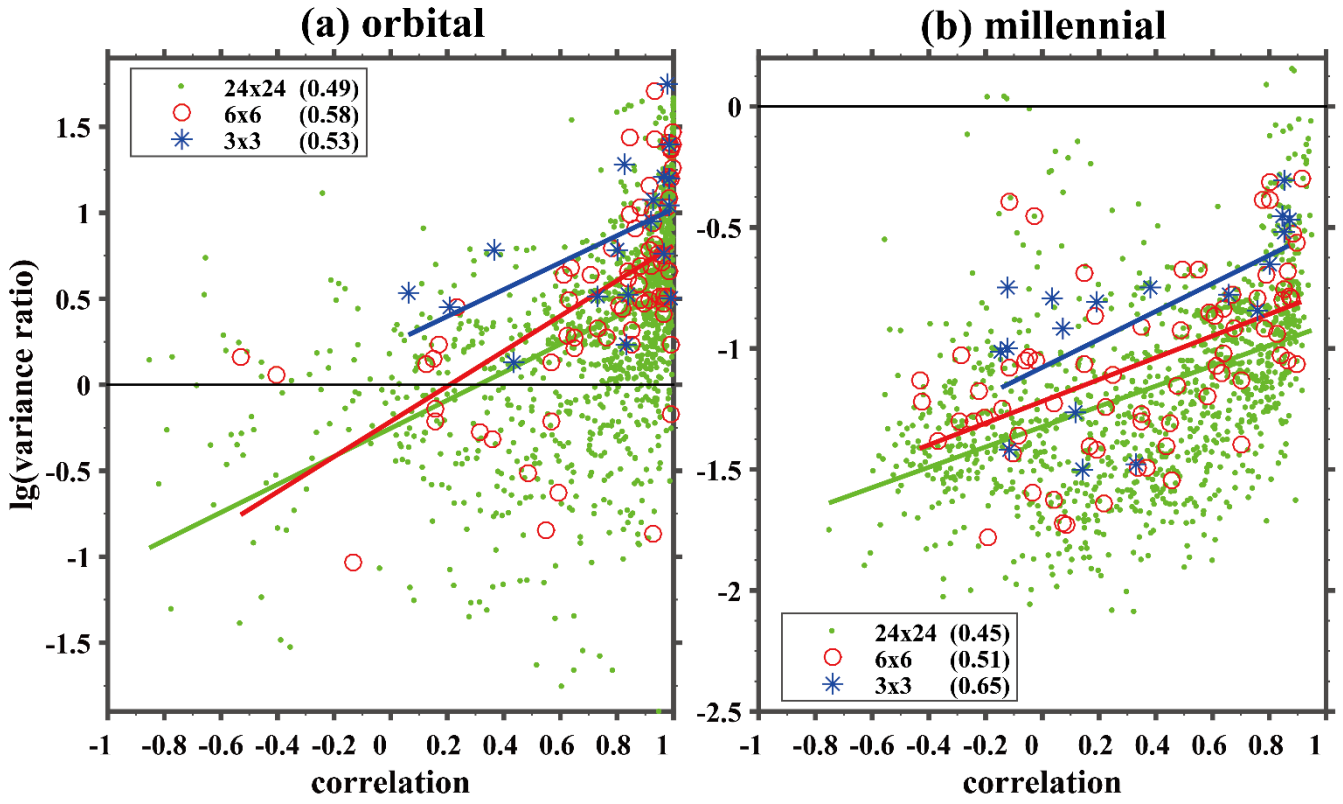


Figure 7: Scatter diagram of the correlation coefficient and the signal-to-noise ratio (variance ratio) on the orbital (a) and millennial (b) time scale. The log base 10 is taken on the variance ratio, as indicated in Figure 6. Only those for the three representative spatial factors (*f=3, 6* and *24*) are shown in the panels. The blue stars and linear fitting line are for factor 3, the red circles and linear fitting line are for factor 6, the green dots and linear fitting line are for factor 24. The fitting coefficients are list in the brackets at the upper (lower) left corners of the panel a (b).

Report #1

Submitted on 17 May 2019

Anonymous Referee #1

I think, the manuscript still suffers from a lack of clarity and conciseness, and the replies to some of the points of my previous review are unsatisfactory or even contradictory. In particular, regarding my previous point

(1) the reply does not explicitly clarify whether the manuscript aims only at my question labelled Q-1 or also Q-2 sometimes, but it seems it is just Q-1, and that the linearity is measured, as before, by the correlation between the SUM and the ALL response, BUT the last sentence of the reply to my previous point

(1) Throughout this work, it seems that the following two *different* questions are mixed up, which makes it basically impossible to evaluate the conclusions drawn from the results:

Q-1. How linear is the response to external forcing? If we denote the temperature resonse to the full external forcing, $F_{all}(t) = F_1(t) + F_2(t) + F_3(t) + F_4(t)$, by $T_R\big(F_{all}(t)\big)$, the response to the individual forcings by $T_R\big(F_i(t)\big)$ (with $i = 1, \ldots, 4$), and the internal temperature variability of the five model simulations by $T_{I,all}$, $T_{I,1}$, $T_{I,2}$, $T_{I,3}$, $T_{I,4}$, respectively, then the linearity of the response could be defined by the extent to which the statement

$$T_R(F_{all}(t)) = \sum_{i=1}^{4} T_R(F_i(t)) \qquad (1)$$

holds, and the linearity could be measured by the correlation between the forced response on the left and that on the right hand side of the above equation. In the manuscript, however, the correlation is computed (see Section 2.2) from the full 'forced plus internal' temperature variability, i.e. between $T_R\big(F_{all}(t)\big) + T_{I,all}(t)$ and $\sum_{i=1}^{4} T_R\big(F_i(t)\big) + T_{I,i}(t)$, Since this latter correlation is influenced by the signal-to-noise ratio, $\mathrm{Var}(T_R)/\mathrm{Var}(T_I)$, a small correlation does not necessarily indicate the absence of linearity, because it could be that simply the signal-to-noise ratio is small, although the response is still perfectly linear. (One would need ensembles of model simulations for each of the five forcing scenarios, and then use the ensemble average in order to suppress the internal variability.) Hence, Q-1 cannot be answered by this approach (without additional information), unless one would always obtain correlations close to unity, which would indicate strong linearity.

Q-2. What is the relative importance of externally forced vs. internal variability, assuming the response were linear? To answer this question one could use the correlation computed from the full temperature variability, as done in the manuscript, but one had to assume the linearity which, however, is to be proven by this work, in particular, for different temporal and spatial scales.

Hence, the authors should clarify the above issues, and make explicit which of their results contributes to which one of the above two questions. This will also help to clarify the implications of the conclusions for various research fields.

Reply: We thank the reviewer for this comment again. We apologize for not directly replied to the question in the last reply. In the last reply, section 1 was almost rewritten, largely to address these two questions. In addition, in the previous subsection 2.2, this issue is addressed extensively again. However, we admit that the two questions are not addressed explicitly. In this new revision, we addressed these two questions explicitly in section 2 by adding a new subsection 2.2 Assessment Strategy. Discussions explicitly on the two questions are also made in section 4. (see the revision with tracking).

(2) contradicts this. Specifically, the reply says that, if the full response is largely dominated by only one forcing such that the other responses are negligible, then this would already imply strong linearity of this response to that single forcing. But if there is only one forcing producing a response, then the employed linearity definition has no meaning anymore! If in that scenario the obtained correlation would be large, it would only mean that the externally forced response dominates over internal variability. But that would be Q-2 again (see above). And it is not obvious to me why the different forcings, which also influence the climate system through different physical mechanisms, should lead to responses of comparable magnitude.

Reply: Thanks for the comment again. We think we understand your point better now and agree with the reviewer on this. We have added a note in section 4 as follows: "It should also be kept in mind that, if the response is dominated by that to a single forcing, the assessment of linear response here becomes one that is more relevant to the question of the forced response vs internal variability, as discussed in Question 2 in subsection 2.2. As a further step, though, one can examine if the magnitude of the total response responds to the magnitude of this single forcing linearly."

(2) Even if we had ensembles available for each of the forcing scenarios, it would still be possible to obtain a large correlation coefficient although the response is only weakly linear (i.e., mostly non-linear), if the individual response to, for example, one of the forcings F_i is much larger than the responses to the remaining forcings, because in this case the full temperature variability might still be dominated by the response to the strong forcing (the non-linear interactions might still be relatively small). Thus, one would need to know the strength (e.g., in terms of variance) of the responses to the various individual forcings.

Reply: See the reply above.

Regarding my previous point

(4) it is still not clear to me why the variance of the internal variability at millennial and orbital time scales, Var(INT.mil.orb), should be (at least roughly) the same as the variance of the internal variability at decadal and centennial time scales, Var(INT.dec.cen). Whereas Var(INT.dec.cen) is the integral of the power spectrum of the internal variability from approximately 1/(2500yrs) to 1/(20yrs), according to the time scale definitions, Var(INT.mil.orb) is the integral from 1/(11000yrs) to 1/(2500yrs). Why should these two integrals yield roughly the same variance? Or do I misunderstand the employed definition of the SNR at millennial and orbital time scales?

Reply: This is a good question. Given the limited time series of a single realization, we do not know the variance of internal variability at slow time scale. In Fig.6-7, we estimate the signal/noise by using decadal as noise. This is only for a heuristics attempt to understand the pattern of linearity in different regions. The fact that the signal/noise ratio seems correlated with the linearity may indicate indeed that internal variability at slow time scale may be proportional to that at shorter time scale. This may be possible if we think of a red noise spectrum where the power kept increasing with the time scale, and, therefore, the power at slow time scale may be proportional to that at faster time scales. But this assumes the slope of the power remains the same which we can't justify. In addition, it should be pointed out that in Fig.6-7, we only need $\mathrm{Var}_{orb,mill}(T_{I,all})$ to be proportional, but not equal, to $\mathrm{Var}_{cent+dec}\left(T_{I,all} + T_R(F_{all})\right)$ across different regions. This is a much relaxed condition.

Of course, ultimately, the question of internal variability of slow time scales can only be solved using ensemble simulations, we think. A comment is added before the discussion of Fig.6 and 7.

(4) How is it justified to estimate the variance of the internal variability at orbital and millennial time scales by the full variance at centennial and decadal variability (page 7, last paragraph)? That is, why should we have

$$\mathrm{Var}_{orb,mill}(T_{I,all}) \approx \mathrm{Var}_{cent+dec}\left(T_{I,all} + T_R(F_{all})\right)? \qquad (2)$$

Even if we assume that $\mathrm{Var}_{cent+dec}(T_R(F_{all}))$ is small compared to $\mathrm{Var}_{cent+dec}(T_{I,all})$, this does not imply anything about the relation between $\mathrm{Var}_{cent+dec}(T_{I,all})$ and $\mathrm{Var}_{orb,mill}(T_{I,all})$. Maybe it could be helpful to investigate the power spectra of the temperature variability under the various forcing scenarios?

Reply: See reply above.

Regarding my previous point

(6) I did not find any answer how the AR(1) fit was done. It does probably make a notable difference whether the fit is done using the lag-1 auto-correlation coefficient, or by fitting an exponential to the auto-correlation function, or by fitting an AR(1) power spectrum.

(6) Please, be a bit more explicit how the significance levels are computed. For example, how is the AR(1) fit done in case of the correlation, and what is the bootstrap design for the error index?

Reply: We thank the reviewer for this comment. The fit is done using the lag-1 auto-correlation coefficient. A comment is added in subsection 2.3.

I suggest to improve the clarity and conciseness of the questions to be answered, of the definitions employed and of the simplifying assumptions made, in order to bring the conclusions drawn on solid ground.

Reply: We thank the reviewer for this comment. We have checked the manuscript twice, again.

Holocene temperature response to external forcing: Assessing the linear response and its spatial and temporal dependence

Lingfeng Wan[1, 2], Zhengyu Liu[3, 4], Jian Liu[1, 2, 4], Weiyi Sun[1, 2], Bin Liu[1, 2]

[1]Key Laboratory for Virtual Geographic Environment, Ministry of Education; State Key Laboratory Cultivation Base of Geographical Environment Evolution of Jiangsu Province; Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application; School of Geography Science, Nanjing Normal University, Nanjing, 210023, China.
[2]Jiangsu Provincial Key Laboratory for Numerical Simulation of Large Scale Complex Systems, School of Mathematical Science, Nanjing Normal University, Nanjing, 210023, China.
[3]Atmospheric Science Program, Department of Geography, Ohio State University, Columbus, OH43210, USA.
[4]Open Studio for the Simulation of Ocean-Climate-Isotope, Qingdao National Laboratory for Marine Science and Technology, Qingdao, 266237, China.

Correspondence: Jian Liu (njdllj@126.com); Zhengyu Liu (liu.7022@osu.edu)

Abstract. Previous studies show that the evolution of global mean temperature forced by the total forcing is almost the same as the sum of those forced individually by orbital, ice sheet, greenhouse gases and meltwater in the last 21,000 years in three independent climate models, CCSM3, FAMOUS and LOVECLIM. This validity of the linear response is useful because it simplifies the interpretation of the climate evolution. However, it has remained unclear if this linear response is valid on other spatial and temporal scales, and, if valid, in what regions. Here, ~~we~~ using a set of TraCE-21ka climate simulation~~s outputs~~, the spatial and temporal dependence of the linear response of the surface temperature evolution in the Holocene is assessed approximately using correlation coefficient and a linear error index. The results show that the response of global mean temperature is almost linear on orbital, millennial and centennial scales in the Holocene, but not on decadal scale. The linear response differs significantly between the Northern Hemisphere (NH) and Southern Hemisphere (SH). In the NH, the response is almost linear on millennial scale, while in the SH the response is almost linear on orbital scale. Furthermore, at regional scales, the linear responses differ substantially between the orbital, millennial, centennial and decadal timescales. On ~~the~~ orbital scale, the linear response is dominant for most regions, even in a small area of a mid-size country like Germany. On ~~the~~ millennial scale, the response is still approximately linear in the NH over many regions~~, albeit less so on the orbital scale~~. Relatively, the linear response is degenerated somewhat over most regions in the SH. On the centennial and decadal timescales, the response is no longer linear in almost all the regions. The regions where the response is linear on the millennial scale are mostly consistent with those on the orbital scale, notably western Eurasian, North Africa, subtropical North Pacific, tropical Atlantic and Indian Ocean, likely caused a large signal-to-noise ratios over these regions. This finding will be helpful for improving our understanding of the regional climate response to various climate forcing factors in the Holocene, especially on orbital and millennial scales.

1 Introduction

Long term temperature evolution in the Pleistocene is often believed, and therefore interpreted, to be driven mainly by several external forcing factors, notably, the orbit forcing, the greenhouse gases (GHGs), the continental ice sheets and the meltwater flux forcing. (Here, we treat the coupled ocean-atmosphere system as our climate system, such that Earth orbital parameters, GHGs, meltwater discharge and continental ice sheet are considered as external forcing). Implicit in this interpretation is often an~~the~~ assumption that the response is almost linear to the four forcing factors, that is, the temperature evolution forced by the total forcing combined is approximately the same as the sum of the temperature responses forced individually by the four forcing factors. This linear response, if valid, simplifies the interpretation of the climate evolution dramatically, because each feature of the climate evolution can now be attributed to those on different forcing factors. One example is the global mean temperature evolution of the last 21,000 years (COHMAP members, 1988; Liu et al., 2014). It has been shown that the global mean temperature response is almost linear to the four forcing factors above in three independent climate models (CCSM3, FAMOUS and LOVECLIM, Fig.2 of Liu et al., 2014) with the temperature evolution forced by the total forcing almost the same as the sum of those individually forced by each forcing factor. Furthermore, this deglacial warming response is forced predominantly by the increase of GHGs, with significant contribution from~~and is also contributed by~~ the ice sheet retreat. This linear response, however, has not been assessed quantitatively for the climate evolution in the Holocene. The Holocene period poses a more stringent and interesting test of the linear response ~~to different forcing factors~~, as it removes the deglacial global warming response that is dominated by that ~~the response~~ to increased CO_2 and ice sheet retreating (Figure 2A of Liu et al., 2014). An even more interesting and practical question here is, besides the global mean temperature response on the slow orbital time scale, how linear is the response at shorter temporal~~ime~~ scales and smaller ~~more regional~~ spatial scales, throughout the Holocene?

In general, the assessment of the linear response, in principle, can be done in a climate model using a set of ~~specifically designed~~ experiments that are~~. It requires experiments~~ forced by the combined forcing as well as each individual forcing. Furthermore, each forcing experiment has to consist of a large number of ensemble members. This follows because a single realization of a coupled ocean-atmosphere model could~~an~~ contain strong internal climate variability on a wide range of timescales (Laepple and Huybers, 2014), from daily variability of synoptic weather storms (Hasselmann, 1976), to interannual variability of El Nino (Cobb et al., 2013), interdecadal climate variability (Delworth and Mann, 2000), all the way to millennial climate variability (Bond et al., 1997). The ensemble mean is therefore necessary for suppressing ~~the~~ internal variability and then generating the truly forced response to each forcing. The goodness of the linear response can therefore be assessed by comparing the response to the total forcing with the sum of the individual responses. One practical problem with this ensemble approach is, however, the extraordinary computing costs, especially for long experiments in more realistic fully coupled general circulation models. A more practical question is therefore: is it possible to obtain a meaningful assessment of the linear response using only a single realization of each forced experiment for the Holocene, such as those in TraCE-21ka experiments (Liu et al., 2014).

Strictly speaking, it is impossible to disentangle the forced response from internal variability in a single realization. This would make the assessment of linear response difficult. However, it is conceivable that, if our interest is the slow climate evolution of millennial or longer time scales in response to the slow forcing factors such as the orbital forcing, ice sheet forcing, GHGs and meltwater flux, the assessment is still possible, albeit approximately, at least for very large scale variability. This follows because these forcing factors are of long time scales and of large spatial scales, the forced response signal should therefore also be on long time scales and large spatial scales, if the response is approximately linear. An extreme example is the almost linear response in the global temperature of the last 21,000 years as discussed by Liu et al., (2014). In contrast, internal variability in the coupled ocean-atmosphere system tends to be of shorter time scales, decadal to centennial, and of smaller spatial scales, at least in current generation of coupled ocean-atmosphere models. This naturally leads to two questions. First, how linear is the climate response at different spatial and temporal scales, quantitatively? Second, in what regions, the linear response tends to dominate? The answer to these questions should help improving our understanding of regional climate response during the Holocene. A further question is: if the linear approximation is valid, what is the contribution of each forcing factor in different regions and at different time scales. This question will be addressed in a follow-up paper (Wan et al., 2019).

In this paper, we assess the linear response for the Holocene temperature evolution quantitatively, using ~~the~~ 5 forced climate simulations in CCSM3 (Liu et al., 2014), with the focus on the spatial and temporal dependence of the linear response. We will assess the linearity response on the timescales of orbital, millennial, centennial and decadal, and of the spatial scales of global, hemispheric and regional. The data and methodology are given in section 2. The dependence of the linear response on spatial and temporal scales are analysed in section 3. A summary and further discussions are given in section 4.

2 Data and Methods

2.1 Data

The data is from TraCE-21ka (Liu et al., 2009, 2014), which consists of a set of 5 synchronously coupled atmosphere-ocean general circulation model simulations for the last 21,000 years. The simulations are completed using the CCSM3 (Community Climate System Model version 3). The simulation forced by the total forcing (experiment ALL) is forced by realistic continental ice sheets, the GHGs, orbital forcing and melting water fluxes. The ice sheet is changed approximately once every 500 years, according to the ICE-5G reconstruction (Peltier, 2004). The atmospheric GHGs concentration is derived from the reconstruction of Joos and Spahni (2008). The orbital forcing follows that of Berger (1978). The coastlines at the LGM were also taken from the ICE-5G reconstruction and were modified at 13.1ka, 12.9ka, 7.6ka, 6.2ka, after which the transient simulation adopted the present-day coastlines. The meltwater flux follows largely the reconstructed sea level and other paleoclimate information and, in the meantime, reconciles the response of Greenland temperature and AMOC strength in comparison with reconstructions. More information on the details of the experiment and forcing can be seen in He (2011) or the TraCE-21ka website http://www.cgd.ucar.edu/ccr/TraCE/.

The transient simulation under the total climate forcing reproduces many large scale features of the deglacial climate evolution consistent with the observations (Shakun et al., 2012; Marsicek et al., 2018), suggesting a potentially reasonable climate sensitivity in CCSM3, at global and continental scales. In addition to the all forcing run (ALL), there are four individual forcing runs forced by the orbital forcing (ORB), the continental ice sheets (ICE), the GHGs (GHG) and meltwater forcing (MWF), respectively (Liu et al., 2014, Table 1). In these four experiments, only one forcing varies the same as in experiment ALL, while other forcing/conditions remain the same as at 19ka. Therefore, this set of experiments can be used to study the linear response of the climate to the four forcing factors. Here, we will only examine the surface temperature response in the Holocene (last 11, 000 years).

2.2. Assessment Strategy

We will use correlation and normalized RMSE to assess the linear response (see next subsection for details). We note, however, our assessment of linear response is approximate. Before introducing the details of the assessment method, it is useful here to make some general comments on the linear response assessment. As pointed out by one reviewer, strictly speaking, the assessment of linear response requires one to answer two questions.

Q1: How linear is the response to external forcing?

Q2: What is the relative importance of externally forced vs. internal variability, assuming the response were linear?

Specifically, for Q1: If we denote the temperature response to the full external forcing by $T_R(F_{all}(t))$, the response to the individual forcings by $T_R(F_i(t))$ (with $i = 1, \ldots, 4$), and the internal temperature variability of the five model simulations by $T_{I,all}, T_{I,1}, T_{I,2}, T_{I,3}, T_{I,4}$, respectively, the linearity of the response could be defined by the extent to which the total forced response equals the sum of the individual response, or

$$T_R(F_{all}(t)) = \sum_{i=1}^{4} T_R(F_i(t)). \quad (1)$$

In our case, we only have a single member for each experiment as

$$T_{all}(t) = T_R(F_{all}(t)) + T_{I,all}(t),$$

and

$$T_i(t) = T_R(F_i(t)) + T_{I,i}(t),$$

and the linearity is assessed from the correlation (and normalized RMSE) between the sum of the individual experiments $\sum_{i=1}^{4} T_i(t)$ and the total forcing experiment $T_{all}(t)$. Therefore, our linearity assessment is contaminated by the noise of internal variability. This can be seen, for example, in the correlation as

$$\left[cor\,\langle T_{all}(t), \sum_{i=1}^{4} T_i(t)\rangle\right]^2 = \frac{[< T_{all}(t), \sum_{i=1}^{4} T_i(t) >< T_{all}(t),\ \sum_{i=1}^{4} T_i(t) >]^2}{< T_{all}(t),\ T_{all}(t) >< \sum_{i=1}^{4} T_i(t),\ \sum_{i=1}^{4} T_i(t) >}$$

$$= \frac{[< T_R(F_{all}(t)),\ \sum_{i=1}^{4} T_R(F_i(t)) >]^2}{\{Var[T_R(F_{all}(t))] + Var[T_{I,all}(t)]\}\{\sum_{i=1}^{4} Var[T_R(F_i(t))] + \sum_{i=1}^{4} Var[T_{I,i}(t)]\}} \qquad (2)$$

~~$= \frac{<T_R(F_{all}(t)),\ \sum_{i=1}^{4} T_R(F_i(t))>}{\{Var[T_R(F_{all}(t))]+Var[T_{I,all}(t)]\}\{\sum_{i=1}^{4} Var[T_R(F_i(t))]+\sum_{i=1}^{4} Var[T_{I,i}(t)]\}}$ (2)~~

Here, < > indicates covariance, and we have assumed that the forced response $T_R(F_*(t))$ and internal variability $T_{I,*}$ (where *=all~~R~~, 1, 2, 3, 4) are independent of each other; in addition, the time series is sufficiently long so sampling errors are negligible. The correlation therefore depends on the single/noise ratio. If the noise (internal variability) is large, the correlation will be much smaller than 1 even if the response is purely linear as in (1). The only way to suppress internal variability is to perform a large number of ensemble simulations for each experiment. Given only one member for each experiment, we have to be content that our linear assessment using (2) is approximate, depending on the signal/noise ratio. ~~–~~Related to this problem of single member experiment, since we can't ~~not~~ distinguish internal variability from forced response clearly, Q2 can't be assessed exactly either.

In spite of these potential issues, with a single member for each experiment, useful information can still be extracted on linear response. Our general hypothesis is that the slow (orbital and millennial) and large (continental and basin) variability is composed mostly of forced signal and the faster (centennial and shorter) and smaller variability is mostly associated with internal variability of noise. In other words, in our set of single member of simulations, the signal over noise ratio is large for slow variability but small for faster variability. Qualitatively, this hypothesis seems reasonable. ~~-~~First, all the four external forcing factors are of slow time scales and large spatial scales; additionally, internal variability is usually weak in the coupled ocean-atmosphere system at slow time scales and large spatial scales. Our focus here is indeed the slow variability and large scale here, so we can roughly treat the slow and large variability in the single realization roughly as the signal and the linearity of the response may be assessed using eqn. (2). Second, again, because our forcing factors are of slow time scales and large spatial scales, higher frequency or small scale variability in the model should not be dominated by the forced variability (unless the response is highly nonlinear!). Therefore, high frequency or small scale variability can be treated roughly as "noise". ~~-~~This is consistent with later assessment that slow variability seems to be approximately linear response while high frequency variability not. Based on this hypothesis, the signal over noise ratio is also estimated using the variance of slow variability as signal and of high frequency variability as noise (as in late Fig.7). It should be noted however that this hypothesis is qualitative in nature. One major purpose of this paper is to give a somewhat more quantitative assessment on this hypothesis. How slow, ~~and~~ how large~~,~~ and~~–~~ how good will be the linear response?

Our experimental design is proper for linear response assessment here. ~~It should be noted that the linear response can't be assessed if~~ Alternatively, in another experimental setting, individual forcing experiments are often superimposed~~the individual forcing experiments are performed with the forcing superimposed~~ sequentially one-by-one~~,~~. ~~In this approach, the four external~~

~~forcing is added sequentially,~~ for example, first the ice sheet, second the ice sheet plus orbital forcing, third the ice sheet, orbital and GHGs, and finally, applying all four forcing of ice sheet, orbital, GHGs and melting water. In t~~In t~~his experimental design, the full forcing response is by default the response of the sum response after adding the four forcing factors together, and therefore can't be used to assess~~test the~~ linearity of the response. Nevertheless, it should be kept in mind that our four individual forcing experiments are not designed optimally for the study of the linear response in the Holocene. This is because, except for the variable forcing, all the other three forcing factors are fixed at the 19ka condition. As such, the mean state is perturbed from the glacial state, instead of a Holocene state. This may have contributed to some unknown deterioration on the linear response discussed later. Nevertheless, we believe, our major conclusion should hold reasonably well. This is because, partly, the response is indeed almost linear for orbital and millennial variability as will be shown later. ~~It should also be noted that our four individual forcing experiments, although in principle feasible for assessing linear response, are not designed optimally for the study of Holocene climate. This is because, except for the variable forcing, all the other three forcing factors is fixed at the 19ka condition. As such, the mean state is perturbed from the glacial state, not a Holocene state. This may have contributed to some unknown deterioration on the linear response discussed later. Nevertheless, we believe, our major conclusion should hold approximately. This is because, partly, the response is indeed almost linear for orbital and millennial variability as will be shown later.~~

2.~~3~~2 Methods

We use two indices to evaluate the linear response: the temporal correlation coefficient r and a normalized linear error index L_e. The correlation coefficient is calculated as

$$r = \frac{\frac{\sum_{t=1}^{n}(S_t-\overline{S_t})\times(T_t-\overline{T_t})}{n}}{\sqrt{\frac{\sum_{t=1}^{n}(S_t-\overline{S_t})^2}{n}}\sqrt{\frac{\sum_{t=1}^{n}(T_t-\overline{T_t})^2}{n}}} = \frac{cov(S_t,T_t)}{\sigma(S_t)\sigma(T_t)} \quad (3\text{~~2~~}1)$$

Here, $S_t = \sum_{i=1}^{4} T_i/4$ is the linear sum of the temperature time series T_i of the four single forcing experiments, T_t is the full temperature time series in the ALL run, both at time t, and n is the length of the time series. The overbar represents the time mean. The correlation coefficient represents the similarity of the temporal evolution between the sum response and the ALL response. However, the correlation does not address the magnitude of the response. Indeed, even if S_t and T_t has a perfect correlation r=1, the two time series can still differ by an arbitrary constant in their magnitudes. Therefore, we will also use a normalized linear error index L_e to evaluate the magnitude of the linear response. Here, L_e is defined as the root mean square error (RMSE) of the sum temperature response from the full temperature response divided by the standard deviation of the full temperature response in the ALL run:

$$L_e = \frac{RMSE}{STD} = \frac{\sqrt{\frac{\sum_{t=1}^{n}[(S_t-\overline{S_t})-(T_t-\overline{T_t})]^2}{n}}}{\sqrt{\frac{\sum_{t=1}^{n}[(T_t-\overline{T_t})]^2}{n}}} = \frac{\sigma(S_t-T_t)}{\sigma(T_t)} \quad (42)$$

In general, a large r (close to 1) and a smaller L_e (close to zero) represents a better linear approximation, with r=1 and $L_e = 0$ as a perfect linear response. Therefore, if r is close to 1 and L_e is close to zero, we can conclude that the response is close to linear. As noted above, w~~W~~ith a single realization here, ~~however,~~ our assessment of linear response has limitations. First, if r is sufficiently small and L_e is sufficiently large, we can't confirm the response is either linear or nonlinear~~we can't confirm the response is linear. But, we can't conclude the response is nonlinear either~~, because the small r or large L_e can also be contributed by strong internal variability. Second, if the forcing is dominated by shorter time scale variability, say interannual to interdecadal variability, as in the case of volcanic forcing or solar variability, it will be difficult to assess the linear response. This because the time scales of forced response now overlap heavily with strong internal climate variability in the coupled system, and it will be difficult to separate the forced response from internal variability without ensemble mean.

But how to assess the goodness of the linear response from the value of r and L_e? We can test the goodness of the linear response statistically on r and L_e.

The statistical significance of r for a particular timescale is tested using the Monte Carlo method (Kroese, 2011 and 2014; Kastner, 2010; Binder, 1997) with 1,000,000 realizations on the corresponding red noise in the AR(1) model (autoregressive model of order 1) which uses the AR(1) coefficient derived from the model to generate time series. The fit is use the lag-1 auto-correlation coefficient.

The statistical significance of the L_e of a particular timescale is tested using a bootstrap method (Efron, 1979, 1993) with 1,000,000 realizations on the corresponding time series. Specifically, the bootstrap is done as follows, taking the global mean temperature as example. First, we will derive the L_e from one random realization on the temperature of the ALL run of the 100-binned data (110 points of data, each representing a 100-yr bin). For this random realization, the order of the original temperature time series is swapped randomly. Then, this realization is used as a new ALL response for comparison with the sum response of the four individual experiments to derive a L_e in eqn. (2). Since the random realization distorts the serial correlation time with the sum response, one should expect usually a large error L_e. Second, we repeat this process for 1,000,000 times on 1,000,000 random realizations; this will produce 1,000,000 random values of L_e, forming the PDF of the L_e. Third, the minimum 5% level is then used as the 95% confidence level.

The dependence of the linear response on spatial and temporal scales will be studied by filtering the time series in different scales. For the spatial scale, we will divide the globe into 9 succeeding cases, denoted by 9 division factors: *f=0, 1, 2, 3, 4, 6, 8, 12* and *24* from the largest global scale to the smallest model grid scale. The *f=0* case is for global average while the *f=1* case is for hemispheric average in the NH and SH. Further division will be done within each hemisphere. Note that each hemisphere has $96(lon.) \times 24(lat.)$ grid boxes, with a ratio of 4:1 between longitude span and latitude span. We divide each hemisphere into $f \times f$ sections of equal latitude and longitude spans, with each area containing the same number of $(96/f) \times (24/f)$ grid boxes, maintaining the ratio of 4:1 between longitude span and latitude span. For example, *f=2* is for the 2×2

division, with each area containing 48 × 12 grid boxes; *f=24* is for the 24 × 24 division with each area containing 4×1 grid boxes, about the size of 15°($lon.$) × 3.75°($lat.$), like the size of a mid-size country of Spain or Germany in the mid-latitude. To the time scale, we decompose a full 11,000-yr annual temperature time series (from 11 ka to 0 ka) in 100-yr bins (a total of 110 data bins, or points, each representing a 100-yr mean) into three components. The three components are to represent the variability of, roughly, orbital, millennial and centennial timescales. Following Marsicek et al. (2018), we derive the orbital and millennial variability using a low-pass filter called the locally weighted regression fits (Loess fits) (Cleveland, 1979). First, the orbital variability is derived by applying a 6500-yr Loess fit low-pass filter onto the temperature time series, and therefore contains the trend and the slow evolution longer than ~6500 years. Second, we apply a 2500-yr Loess fit low-pass filter onto the temperature time series; then, we derive the millennial variability using this 2500-yr low-pass data subtracting the 6500-yr low-pass data. Finally, centennial variability is derived as the difference between the 100-yr binned temperature time series and the 2500-yr low-pass time series. In addition, we also derive a decadal variability time series. First, we compile the 10-yr bin time series from the original 11,000-yr annual time series (of a total of 1,100 data points, each representing a 10-yr mean). Second, we apply a 100-yr running mean low-pass filter on the time series of the 10-yr binned data. Finally, decadal variability is derived by using the 10-yr binned time series minus its 100-yr running mean time series.

Given the different degrees of freedom especially among the filtered variability of different time scales, it is important to test the goodness of the linear response statistically on r and L_e on different time scales differently~~. The statistical significance of r for a particular timescale is tested using the Monte Carlo method (Kroese, 2011 and 2014; Kastner, 2010; Binder, 1997) with 1,000,000 realizations on the corresponding red noise in the AR(1) model (autoregressive model of order 1), which uses the AR(1) coefficient derived from the model to generate time series)~~. As a reference, the significance level is tested against the global mean temperature series in the ALL run. For the total, orbital, millennial, centennial and decadal temperature time series, the 95% confidence levels are found to be 0.72 (with the AR(1) coefficient 0.96), 0.76 (0.97), 0.65 (0.95), 0.21 (0.31) and 0.19 (0.06), respectively.

In this paper, this AR(1) test for global mean temperature is also used as the common significant test for different spatial scales and in different regions as well. This use of a common significance level is for simplicity here. First, the use of different regional AR(1) coefficient for different regions will make the comparison of the linear responses among different spatial scales (e.g. Fig.3 and 4) and different regions (Fig.5, 6) difficult. Second, except for the orbital time scale, the AR(1) coefficient for the global mean temperature is larger than most of the regional AR(1) coefficients (not shown), likely caused by the further suppression of internal variability in the global mean. As a result, the global mean AR(1) test serves actually as a more stringent test than the local AR(1) test. At the orbital scale, the global mean AR(1) coefficient is in about the middle of the regional AR(1) coefficients. The uncertainly of using the global mean AR(1) coefficient is therefore about the average of those of regional AR(1) coefficients. Third, and, most importantly, as our first study here, our focus is on the global features of the linear response. The difference among the AR(1) coefficients among different spatial scales and different regions ~~is~~ are much smaller than that between different time scales here. Therefore, the global mean AR(1) can still provide an approximate

guideline of the significant test properly at different time scales. In later studies, if one's focus is for a specific spatial scale and on a specific region, the regional AR(1) should be used for re-examination of the significance test.

~~Statistical significance of the L_e of a particular timescale is tested using a bootstrap method (Efron, 1979, 1993) with 1,000,000 realizations on the corresponding time series. Specifically, the bootstrap is done as follows, taking the global mean temperature as example. First, we will derive the L_e from one random realization on the temperature of the ALL run of the 100-binned data (110 points of data, each representing a 100 yr bin). For this random realization, the order of the original temperature time series is swapped randomly. Then, this realization is used as a new ALL response for comparison with the sum response of the four individual experiments to derive a L_e in eqn. (2). Since the random realization distorts the serial correlation time with the sum response, one should expect usually a large error L_e. Second, we repeat this process for 1,000,000 times on 1,000,000 random realizations; this will produce 1,000,000 random values of L_e, forming the PDF of the L_e. Third, the minimum 5% level is then used as the 95% confidence level.~~ As a reference, the significance level of L_e is tested against the global mean temperature series in the ALL run. For the total, orbital, millennial, centennial and decadal temperature time series, the 95% confidence levels are found to be 1.23, 1.23, 1.21, 1.24 and 1.36, respectively. ~~Fourth, this bootstrap is repeated on the time series of the original total variability, and the decomposed orbital variability, millennial variability, centennial variability and decadal variability of global mean temperature time series. This gives the 95% confidence levels as 1.23, 1.23, 1.21, 1.24 and 1.36, respectively.~~ This suggests that when the RMSE is less than about 1.2-1.3 times of the total response, the linear sum is not significantly different from the total response at the 95% confidence level. ~~Finally, a~~As for the L_e~~correlation~~ test, since our focus here is on the global feature of the linear response, for simplicity, the significance level derived from the global mean temperature is used as the common confidence level for all regional scales.

3 Results

3.1 Linear responses at different temporal scales

The global mean temperature provides a useful example to start the discussion of the dependence of the linear response on timescales. We first examine the linear response of the global mean temperature based on its components of orbital, millennial, centennial and decadal variability (Fig.1). Fig.1a is the total variability of global surface temperature derived from the ALL run and the sum of the four individual forcing experiments. The global temperature response is almost linear on the orbital and millennial scales throughout the Holocene (Fig.1b and 1c). The orbital scale evolution is characterized by a warming trend of about 1℃ from 11ka to ~4.5ka before decreasing slightly afterwards. This feature is captured in the linear sum albeit with a slightly smaller magnitude and an additional local minimum around 3ka (Fig.1b). The total variability is very similar to the orbital variability (r=0.99, Fig.1a vs Fig.1b). The millennial variability shows 5 major peaks around ~9.8ka, 7.8ka, 4.7ka, 3.7ka and 1.8ka. All these peaks seem to be captured in the sum response albeit with a slightly larger amplitude (Fig.1c). For ~~the~~ orbital and millennial variability, the correlation coefficients between the sum and the full responses are r=0.83 and 0.71,

respectively, both significant at the 95% confidence level and explaining over 50% of the variance; the linear errors are L_e=0.63 and 0.92, respectively, also significant at the 95% confidence level. It should be noted, however, that the goodness of the linear response is based on the entire period and is meant for the response of the time scale to be studied. Therefore, even for a good linear response at long time scales, the sum response may still differ from the total response significant at some particular time. For example, for the orbital scale response in Fig.1b, even though the linear response is good in terms of r and L_e, there is a 1ºC difference between the sum and total responses at 11ka and 3ka. Therefore, for the orbital scale response, the linear response mainly refers to the trend-like slow response of comparable time scale of the orbital scale, instead of some response features of shorter time scales. Further down the scale, at centennial time scale, the global centennial variability appears also to exhibit a modest linear response (Fig.1d), with r=0.44 and L_e=1.21. But the linear response of the decadal variability ~~is~~ becomes poor (Fig.1e), with r=-0.02 and L_e=1.99, which is not statistically significant at the 95% confidence level. The result of the analysis on global mean temperature is qualitatively consistent with previous hypothesis that the linear response tends to degenerate at shorter temporal scale, because of the smaller forced response signal and the presence of strong internal variability. It shows that, for global temperature, the response is approximately linear at orbital and millennial timescales, but becomes much less so at centennial scales and fails completely at decadal scales.

3.2 Linear responses at different spatial scales

In order to assess the linear response at different spatial~~regional~~ scales, we first analyse the linear response o~~n~~f the hemisphere scale for the NH and SH (*f=1*). It is interesting that the linear response significantly differs between the NH and SH (Fig.2). Fig.2a and 2f are the total variability of hemispheric surface temperature of the ALL response and sum response for the NH and SH, respectively. Their components on the 4 timescales (i.e. orbital, millennial, centennial and decadal) are shown in ~~the~~ Figs.2b-2e and Figs.2g-2j, respectively. In the NH, the response is almost linear at the millennial scale (r=0.82, L_e=1.01, Fig. 2c), but not so strong on orbital (r=0.55, L_e=0.84, Fig.2b) and centennial (r=0.32, L_e=2.29, Fig.2d) scales, while~~re~~ only L_e is significant on the orbital scale, and only r is significant on the centennial scale. At the decadal scale, the linear response fails completely (r=-0.04, L_e=1.99, Fig.2e). In comparison, in the SH, the linear response is dominant at the orbital scale (r=0.92, L_e=0.43, Fig.2g), but poor on all other timescales, including millennial (r=-0.12, L_e=2.32, Fig.2h), centennial (r=0.14, L_e=3.19, Fig.2i) and decadal (r=0.03, L_e=2.07, Fig. 2j) scales. The linear response of the global mean temperature discussed in Fig.1 therefore seems to be dominated by the SH response on the orbital scale, but by the NH response on the millennial scale. This suggests that the goodness of the linear response depends on both the region and time scale, ~~. This further~~ highlighting the need to study the linear response at regional scales.

The linear response at different spatial scales and on the orbital, millennial, centennial and decadal timescales are summarized in Fig.3 and Fig.4 in the correlation coefficient and linear error index, respectively. Fig.3a shows the correlation coefficients of the orbital variability in each region for the 9 division factors. The cases of global mean (*f=0*) and hemispheric mean (*f=1*) have been discussed in details before. The correlation coefficients for succeeding division factors (*f=2, 3, …, 24*) show several features. First, as expected, the correlation coefficient tends to decrease towards smaller area (larger *f*). Quantitatively, however,

the correlation coefficient does not decrease much, such that even at the smallest area (*f=24*), the correlation in most regions are still above 0.8, statistically significant at 95% confidence level. This suggests that the response at the orbital scale is almost linear over most regions, even at the smallest scale of about a mid-size country like Germany (*f=24*, $15°(lon.) \times 3.75°(lat.)$). Second, the linear response in the NH is slightly better than SH (for *f≥3*), a topic to be returned later. Third, subareas in both hemispheres show comparable linear response across all the spatial scales, with the median correlation all above ~0.8, except that of the NH mean temperature (*f=1)*. The linear response of NH is not better than thoseat of regional variability at smaller spatial scales (*f≥2*). This is opposite to the expectation that the linear response becomes more distinct for a larger area, because the average over a larger area tend to suppress internal variability more. This case, however, seems to be a special feature and should be treated with caution. The correlation coefficient therefore shows that, for orbital scale evolution, temperature response is dominated by the linear response over most of the globe, even at regional scales. These features are also consistent with the linear error analysis in Fig.4a.

Millennial variability also shows a weaker linear response for smaller scales, in both the correlation (Fig.3b) and linear error (Fig.4b). Quantitatively, for millennial variability, the response is still approximately linear in the NH over many regions, albeit less so than at the orbital scale. The correlation coefficients remain above 0.6 across most regions even at the smallest division area (*f=24*), contributing to ~40% of the variance. In contrast to the orbital variability, where regions in both hemispheres show comparable linear response, the millennial variability shows that the response can't be confirmed linear in most regions in the SH, with the median no longer significant at 95% confidence level. Similar to the orbital variability, nevertheless, the responses are more linear in the NH than SH on all the spatial scales.

In contrast to orbital and millennial variability, almost no response can be confirmed linear for the centennial variability. The median linear response is no longer significant on the centennial timescale in either hemisphere across spatial scales (*f>3*, Fig.3c and Fig.4c), with few correlation coefficients larger than 0.3 and contributing less than 10% of the variance. Finally, the decadal variability exhibits absolutely no linear response over any spatial scales in either both of the NH orand SH (Fig.3d and Fig.4d).

The approximate linear response at the orbital and millennial scales suggest that these two groups of variability are generated predominantly by the external forcing. In contrast, the poor linear response of centennial and decadal variability suggest that these two groups of variability are caused mainly by the internal coupled ocean-atmosphere processes. This is largely consistent with our original hypothesis. It should be kept in mind that, in our single realization here, the poor linear response on centennial and decadal variability may also be contributed by nonlinear responses of the climate system. But, given the almost absence of forcing variability at this short time scale in our experiments, we do not think that the nonlinear response is the major cause of the poor linear response here.

3.3 Pattern of the Linear Responses

We now further study the pattern of the linear response. Fig.5 shows the spatial patterns of the correlation coefficients at orbital (Fig.5a1-a3) and millennial (Fig.5b1-b3) scales for three representative spatial scales, *f=3, 6* and *24* (the other factors are not

shown because they are similar to the above mentioned three representative spatial scales). For orbital variability (Fig.5a1-a3), the response is almost linear in most regions in the NH ~~in~~ on all three spatial scales, with the correlation coefficients above 0.8. In the SH, the response is also almost linear over the continents, except for over Australia, but is not linear over the ocean. This leads to the significantly reduced linear response in the SH as discussed in Fig.3a-4a. This suggests that orbital variability is likely forced predominantly by external forcing over continents. The overall poorer linear response over ocean than land, however, is puzzling. The orbital time scale is so long that one would expect a similar quasi-equilibrium response over both land and ocean surface. This issue deserves further study in the future. More specifically, at the regional scales, e.g. *f=6* and *24* (Fig.5a2 and 5a3), the response is almost linear over the western half of the Eurasian continent, Northern Africa, Central and South America, most NH oceans and SH tropical oceans, and the Antarctica Continent, as seen in the Fig.5a2 and 5a3 (only those significant correlation coefficients are shown). But the linear response is poor over the North America continent and the eastern Eurasian continent, and the entire Southern Ocean. Similar features can also be seen in the map of linear error (not shown).

For millennial variability, in the NH, ~~the~~ linear response shows a similar feature to that of ~~the~~ orbital variability, but the linear response is poor over almost all the SH. Fig.5b1-b3 show that the response is almost linear in most regions in the NH at the three spatial scales, with the correlation coefficient above 0.6. At the regional scale, e.g. *f=6* and *24* (Fig.5b2 and 5b3), the response is almost linear over the northwestern Eurasian continent, Northern Africa, northern North America, northern Pacific Ocean, southern North Atlantic Ocean and western Arctic Ocean, as seen in the Fig.5b2 and 5b3 (only those significant correlation coefficients are shown). But the linear response is poor over the southern North America, the eastern and southern Eurasian continent. While in the SH, the linear relationship is poor over almost the entire SH as seen in the correlation map (Fig.5b1-b3). Interestingly, over the NH, the regions of linear response for millennial variability are mostly consistent with the regions of linear response for orbital variability, notably western Eurasian, North Africa, subtropical North Pacific, tropical Atlantic and Indian Ocean. These preferred region of linear response suggests potentially some common mechanisms of the climate response in these regions, in this model.

In order to understand the cause for the preferred regions of linear response, we examine the signal-to-noise ratio. As discussed in subsection 2.2, ~~Since~~ our forcing factors are on millennial and orbital time scales, and the linear response is also largely valid for orbital and millennial variability. W~~, w~~e will therefore use the variance of the orbital and millennial variability as a crude estimate of the linear response signal. Similarly, since there is no centennial and decadal forcing in our model and the response of centennial and decadal variability are not linear response, we use the variance of the sum of the centennial and decadal variability as a rough estimate for internal variability as the linear noise. Admittedly, this estimation is crude, limited by the single realization here. This signal-to-noise ratio is not directly to address Q2 in subsection 2.2, because the time scales of the signal and noise are different. Instead, it is used as a rough estimation of the relative magnitude of signal-to-noise ratio between different regions, with the assumption that the relative noise level between different regions may be not too sensitive to the time scales. Indeed, t~~T~~he use of ~~ratio of the~~ signal to noise ratio here is ~~may therefore be able~~ to shed some light on the regional preference of linear response. ~~Admittedly, this estimation is crude, limited by the single realization here.~~ Figure 6

shows the signal-to-noise ratio for orbital and millennial variability for the three representative spatial scales (*f=3, 6* and *24*). For orbital variability (Fig.6a1-a3), the signal-to-noise ratio is large (above $10^{(0.6)}$, the log base 10 is taken on the signal to noise ratios) in most regions in the NH. In the SH, the signal-to-noise ratio is also large over the continents, but is small over the ocean. At different regional scales, e.g. *f=6* and *24* (Fig.6a2 and 6a3), the signal-to-noise ratio is large over the western half of the Eurasian continent, Northern Africa, Central and South America, most NH oceans and SH tropical oceans, and the Antarctic Continent. But the signal-to-noise ratio is small over Canada and the eastern Eurasian continent, and the entire Southern Ocean. These spatial features of signal-to-noise ratio on the orbital scale (Fig.6a1-a3) are similar to those in the correlation map (Fig.5a1-a3). The spatial correlation between the map of the signal-over-noise ratio in Fig.6 and the corresponding correlation coefficient in Fig.5 is 0.53, 0.58 and 0.49 for *f=3, 6* and *24* respectively, as seen in the scatter diagram in Fig.7, all significant at 95% confidence level.

For millennial variability, the signal-to-noise ratio also shows a similar feature to that of orbital variability although overall somewhat smaller (note the different color scales). Fig.6b1-b3 show that the signal-to-noise ratio is large in most regions in the NH in all three spatial scales, with the signal-to-noise ratio above $10^{(-1)}$ (the log base 10 is taken on the signal to noise ratios). At the regional scale, e.g. *f=6* and *24* (Fig.6b2 and 6b3), the signal-to-noise ratio is large over the northwestern Eurasian continent, Northern Africa, central North America, northern North Pacific Ocean, Northern Atlantic Ocean and the Arctic Ocean. But the signal-to-noise ratio is small over the North America continent outside the central North America, the South America and the eastern and southern Eurasian continent. While the signal-to-noise ratio is small over almost the entire SH. Over the NH, interestingly, the regions of large signal-to-noise ratio for millennial variability, are mostly consistent with the regions of large signal-to-noise ratio for orbital variability, notably northwestern Eurasian, North Africa, subtropical North Pacific, Northern Atlantic and tropical Northern Indian Ocean. These features of spatial pattern of signal-to-noise ratio on the millennial scale (Fig.6b1-b3) are similar to those in the correlation map (Fig.5b1-b3). The correlation coefficient between the maps of signal-to-noise ratio and correlation is 0.65, 0.51 and 0.45 for *f=3, 6* and *24*, respectively (Fig.7), again all significant at 95% confidence level.

4 Summary and Discussions

In this paper, the linear response is assessed for the surface temperature response to orbital forcing, GHGs, meltwater discharge and continental ice sheet throughout the Holocene in a coupled GCM (CCSM3). The global mean temperature response is almost linear on the orbital, millennial, and even centennial scales throughout the Holocene, but not for decadal variability (Fig.1). Furthermore, the sum response account for over 50% of the total response variance for orbital and millennial variability. Further analysis on regional scale suggests that the response is approximately linear on the orbital and millennial scales for most continental regions over the NH and SH, with the sum response explaining over about 50% of the total response variance. However, the linear response is not significant over much of the ocean, especially over the ocean in the SH. There are specific regions where linear response tends to be dominant, notably the western Eurasian continent, North Africa, the central and

South America, the Antarctica continent, and the North Pacific. The strong linear response is interpreted as the region of large signal-to-noise ratio. That is, in these regions, either the orbital and millennial response signal is large, or the influence of the centennial and decadal variability noise is small, or both. This suggests that the orbital and millennial variability in these regions are relatively easy to understand. This finding lays a foundation to our further understanding of the impacts of different climate forcing factors on the temperature evolution in the Holocene of orbital and millennial time scales. This understanding is our original motivation of this work. Further work is underway in understanding the contribution of different forcing factors on the temperature evolution (Wan et al., 2019).

It should be kept in mind that since there is only one member for each experiment, we can't separate the forced response signal from the internal variability of noise clearly at each time scales. Therefore, we can't address Q1 and Q2 raised in subsection 2.2 accurately. -Instead, our assessment is likely contaminated by internal variability (see discussions in section 1 and 2). In particular, for smaller scale variability, of which internal variability is likely to be strong and the forced signal is likely to be weak, our correlation may underestimate the linearity of the response (see eqn. (2)). Nevertheless, we speculate that our results on large scale variability still remains robust. Furthermore, at regional scales, although the absolute value of linear correlation of the forced response may be underestimated, it is possible that relative between different regions, the linear assessment may still be somewhat valid. These speculations, however, require much further studies, especially with ensemble experiments.

In spite of its limitation, our study~~The result here~~ represents the first systematic ~~such~~ assessment of linear response for the Holocene and can serve as a starting point for further studies in the future.

There are many further issues that need to be studies. Our study here~~and~~ is carried out for a single variable (surface temperature) in a single model (CCSM3) for the Holocene. Yet,~~. It should therefore be kept in mind that~~ the linear response~~assessment~~ could differ for different variables, in different models, for different periods and for different sets of forcing factors. For example, if we evaluate the precipitation response in the Holocene in CCSM3, the response is less linear than temperature (not shown); this is expected because precipitation response contains more internal variability and exhibits more nonlinear behaviour than temperature. The assessment will be also different if a different period is assessed, e.g. the last 21,000 years; with a large amplitude of climate forcing, the linear response may degenerate in the 21,000-year period. In addition, the assessment of linear response using only one realization will be difficult to perform for volcanic forcing and solar variability forcing; these forcing factors have short time scales and therefore their impacts will be difficult to separate from internal variability without ensemble experiments. Finally, it is also important to repeat the same assessment here in different models and to establish the robustness of the assessment. It should also be kept in mind that our assessment is implicitly related to the assumption that, at millennial and orbital time scales, internal variability is not strong relative to the forced responses. Although this seems to be consistent in our model, there is a possibility that internal variability is severely underestimated in the model than in the real world (Laepple and Huybers, 2014). If true, the relevance of our model assessment to the real world will be limited. -It should also be kept in mind that, if the response is dominated by that to a single forcing, the assessment of linear response here becomes one that is more relevant to the question of the forced response vs internal variability, as discussed in Q2 in subsection

2.2. As a further step, though, one can examine if the magnitude of the total response responds to the magnitude of this single forcing linearly.

Even in the context of this model assessment, much further work remains. Most importantly, the purpose of testing the linear response is for a better understanding of the physical mechanism of the climate response. It is highly desirable to understand why response tends to be linear in some region, but not in other regions. In particular, it is unclear why the linear response is preferred over land than over ocean for orbital and millennial variability. At such a long time scale, one would expect that the upper ocean response has reached quasi-equilibrium and therefore the surface temperature response over land and over ocean should not be too much different. Ultimately, we would like to assess and understand the physical mechanism of the climate evolution in different regions. These work are underway (Wan et al., 2019).

Acknowledgments

This work was jointly supported by the National Key Research and Development Program of China (Grant No. 2016YFA0600401), the National Natural Science Foundation of China (Grant No. 41420104002, 41630527), the Program of Innovative Research Team of Jiangsu Higher Education Institutions of China, and the Priority Academic Program Development of Jiangsu Higher Education Institutions (Grant No. 164320H116), and US NSF P2C2. We used the output from the full TraCE-21ka simulation, available at https://www.earthsystemgrid.org/project/trace.html.

References

Berger, A.: Long-term variations of daily insolation and quaternary climate changes. Journal of atmospheric sciences, 35(12), 2362-2367, doi: 10.1175/1520-0469(1978)035<2362:LTVODI>2.0.CO;2, 1978.

Binder, K.: Applications of Monte Carlo methods to statistical physics, Reports on Progress in Physics, 60(5), 487-559, doi: 10.1088/0034-4885/60/5/001, 1997.

Bond, G., Showers, W., Cheseby, M., Lotti, R., Almasi, P., deMenocal, P., Priore, P., Cullen, H., Hajdas, I., Bonani, G.: A Pervasive Millennial-Scale Cycle in North Atlantic Holocene and Glacial Climates, Science, 278(5341), 1257-1266, doi:10.1126/science.278.5341.1257, 1997.

Cleveland, W. S.: Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association, 74(368), 829-826, doi:10.2307/2286407, 1979.

Cobb, K. M., Westphal, N., Sayani, H. R., Watson, J. T., Lorenzo, E. D., Cheng, H., Edwards, R. L., Charles, C. D.: Highly variable El Nino-Southern Oscillation throughout the Holocene, Science, 339(6115), 67-70, doi: 10.1126/science.1228246, 2013.

COHMAP Members.: Climatic Changes in the Last 18,000 Years: Observations and Model Simulations, Science, 241(4869), 1043-1052, doi:10.1126/science.241.4869.1043, 1988.

Delworth, T. L. and Mann, M. E.: Observed and simulated multidecadal variability in the Northern Hemisphere, Climate Dynamics, 16(9), 661-676, doi:10.1007/s003820000075, 2000.

Efron, B. and Tibshirani, R. J. (Eds): An introduction to the bootstrap, Chapman and Hall/CRC, New York, 1993.

Efron, B.: Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics, 7(1), 1-26, doi: 10.1214/aos/1176344552, 1979.

Hasselmann, K.: Stochastic climate models Part I. Theory, Tellus, 28(6), 473-485, doi:10.3402/tellusa.v28i6.11316, 1976.

He, F.: Simulating Transient Climate Evolution of the Last Deglaciation with CCSM3, PhD thesis, 1–177 (Univ. Wisconsin-Madison, 2011).

Joos, F. and Spahni, R.: Rates of change in natural and anthropogenic radiative forcing over the past 20,000 years. Proceedings of the National Academy of Sciences, 105(5), 1425-1430, doi: 10.1073/pnas.0707386105, 2008.

Kastner, M.: Monte Carlo methods in statistical physics: mathematical foundations and strategies, Communications in Nonlinear Science and Numerical Simulation, 15(6), 1589-1602, doi:10.1016/j.cnsns.2009.06.011, 2010.

Kroese, D. P., Brereton, T., Taimre, T., Botev, Z. I.: Why the Monte Carlo method is so important today, Wiley Interdisciplinary Reviews: Computational Statistics, 6(6), 386-392, doi:10.1002/wics.1314, 2014.

Kroese, D. P., Taimre, T., Botev, Z. I. (Eds): Handbook of Monte Carlo Methods, Wiley Series in Probability and Statistics, John Wiley and Sons, New York, doi:10.1002/9781118014967, 2011.

Laepple, T. and Huybers, P.: Ocean surface temperature variability: Large model-data differences at decadal and longer periods. Proceedings of the National Academy of Sciences of the United States of America, 111(47), 16682-7, doi:10.1073/pnas.1412077111, 2014.

Liu, Z., Otto-Bliesner, B. L., He, F., Brady, E. C., Tomas, R., Clark, P. U., Carlson, A. E., Lynch-Stieglitz, J., Curry, W., Brook, E., Erickson, D., Jacob, R., Kutzbach, J., Cheng, J.: Transient Simulation of Last Deglaciation with a New Mechanism for Bølling-Allerød Warming, Science, 325(5938), 310-314, doi:10.1126/science.1171041, 2009.

Liu, Z., Zhu, J., Rosenthal, Y., Zhang, X., Otto-Bliesner, B. L., Timmermann, A., Smith, R. S., Lohmann, G., Zheng, W., Elison, T. O.: The Holocene temperature conundrum, PNAS (USA), 111(34), 3501-5, doi:10.1073/pnas.1407229111, 2014.

Marsicek, J., Shuman, B. N., Bartlein, P. J., Shafer, S. L., Brewer, S.: Reconciling divergent trends and millennial variations in Holocene temperatures, Nature, 554(7690), 92-96, doi:10.1038/nature25464, 2018.

Peltier, W. R.: Global Glacial Isostasy and the Surface of the Ice-Age Earth: The ICE-5G (VM2) Model and GRACE. Annual Review of Earth & Planetary Sciences, 20(32), 111-149, doi: 10.1146/annurev.earth.32.082503.144359, 2004.

Shakun, J. D., Clark, P. U., He, F., Marcott, S. A., Mix, A. C., Liu, Z., Otto-Bliesner, B., Schmittner, A., Bard, E.: Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation, Nature, 484(7392), 49-54, doi:10.1038/nature10915, 2012.

Wan, L., Liu, Z., Liu, J.: Assessing the impact of individual climate forcing on long term temperature evolution in the Holocene. In prep, 2019.

Table 1. TraCE-21ka simulation experiments

No.	experiment	Forcing	time	Resolution(lat×lon)
1	ORB	Orbital forcing	11000	48×96
2	GHG	GHGs forcing	11000	48×96
3	ICE	Ice sheets forcing	11000	48×96
4	MWF	Meltwater forcing	11000	48×96
5	ALL	Orbital + GHGs + ICE sheets + Meltwater forcing	11000	48×96

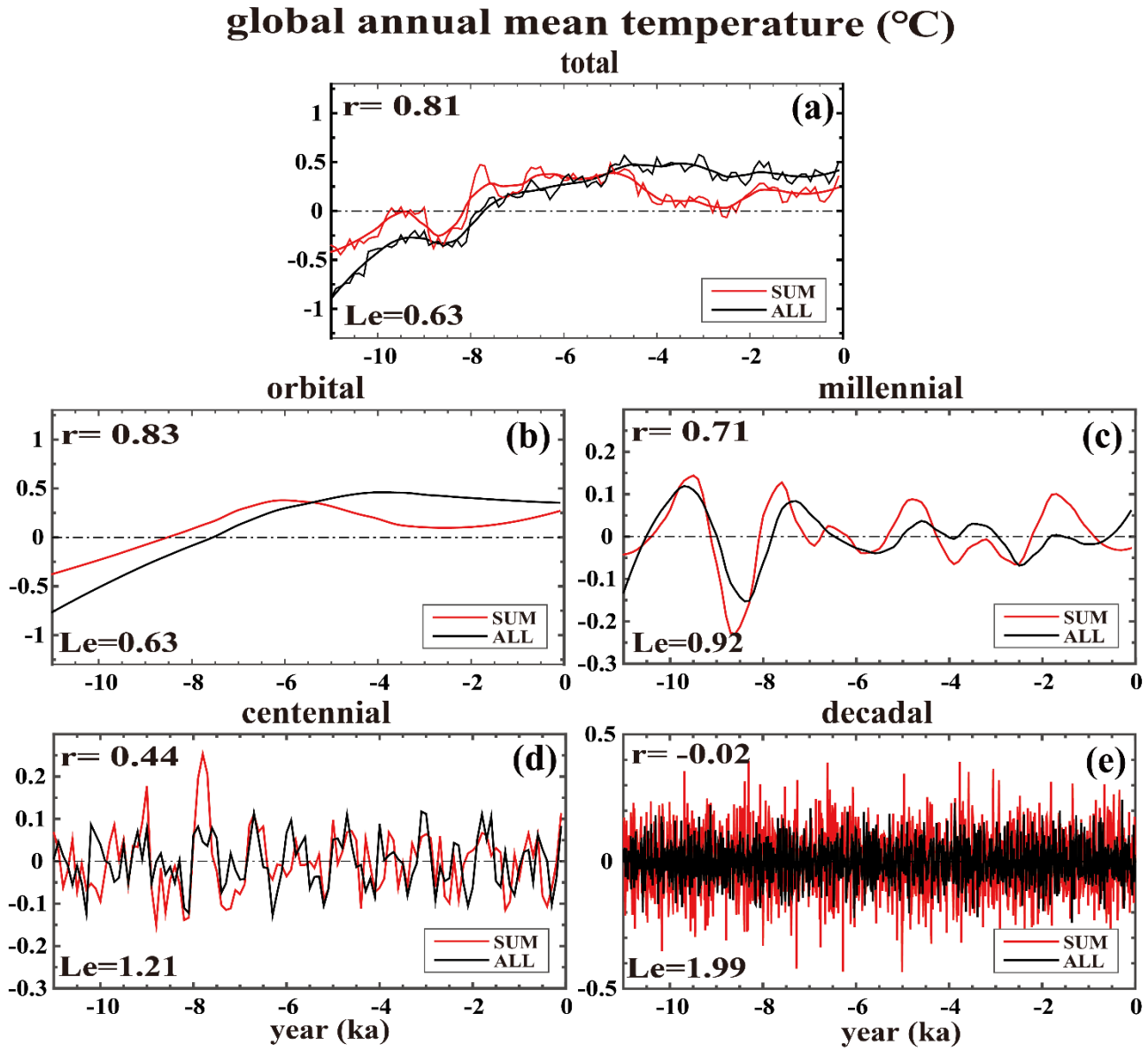


Figure 1: The global annual mean surface temperature time series derived from the ALL run (black) and the SUM (the sum of four single forcing run, red). In the (a), the thin line is the 100-yr binned time series, the thick line is 2500-yr loess fitted time series. The orbital scale variability (b) is represented by the 6500-yr loess fitted series. The millennial variability (c) is represented by the 2500-yr loess fitted data subtracting the 6500-yr loess fitted data. The centennial variability (d) is represented by the 100-yr binned data subtracting the 2500-yr loess fitted data. The decadal variability (e) is represented by the 10-yr binned origin time series subtracting its 100-yr running mean. The correlation coefficient (r) is given at the upper left corner and the linear error (L_e) is given at the lower left corner of each panel. The x axis is year (ka, 0 is AD 1950, negative is before AD 1950), and the y axis is temperature anomaly (°C, relative to 11ka-0ka).

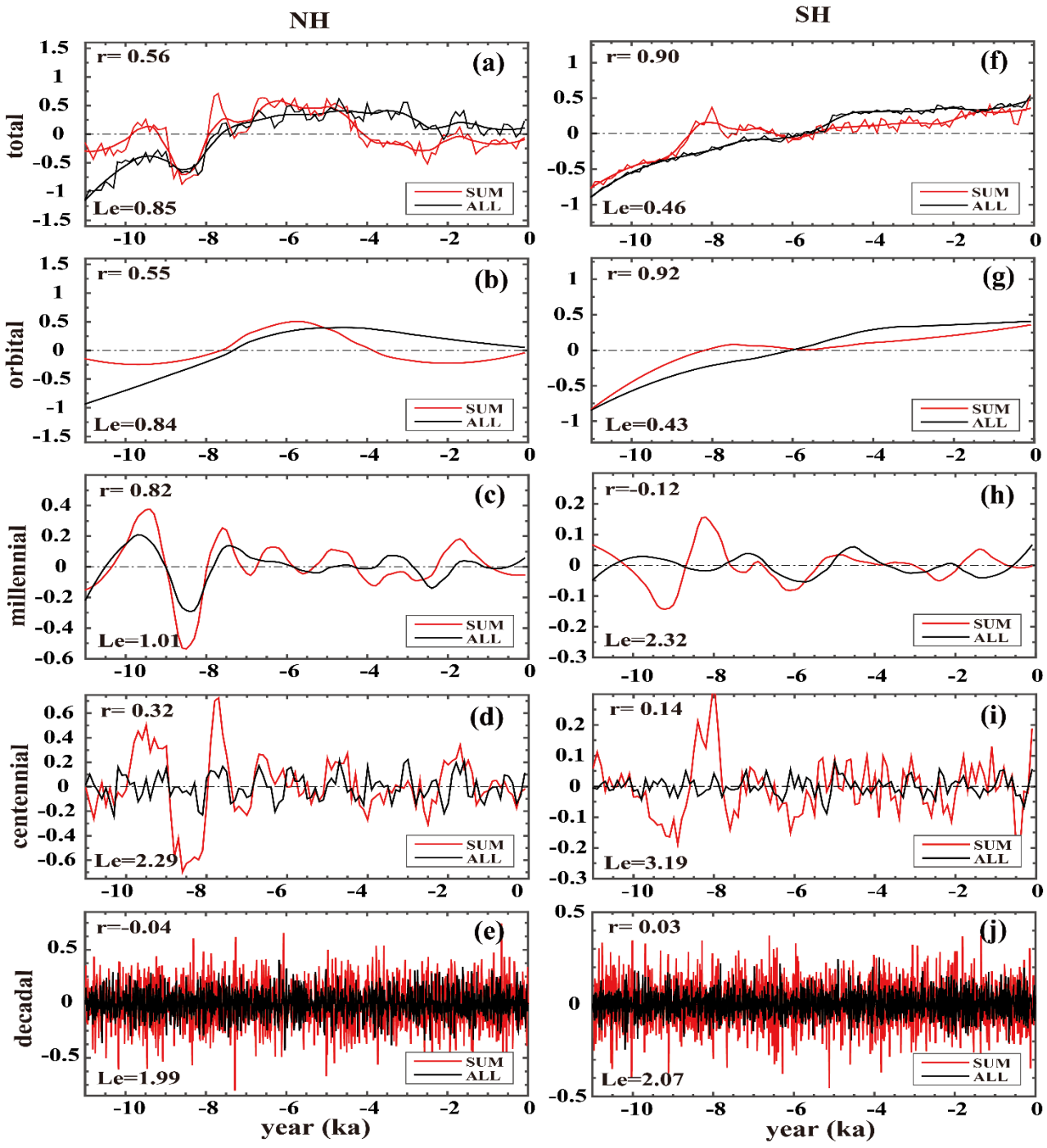


Figure 2: The surface temperature time series derived from the ALL run (black) and the SUM (red). The (a-e) is similar to the Fig.1a-e but for NH, and the (f-j) is similar to the Fig1.a-e but for SH. The x axis is year (ka, 0 is AD 1950, negative is before AD 1950), and the y axis is temperature anomaly (°C, relative to 11ka-0ka).

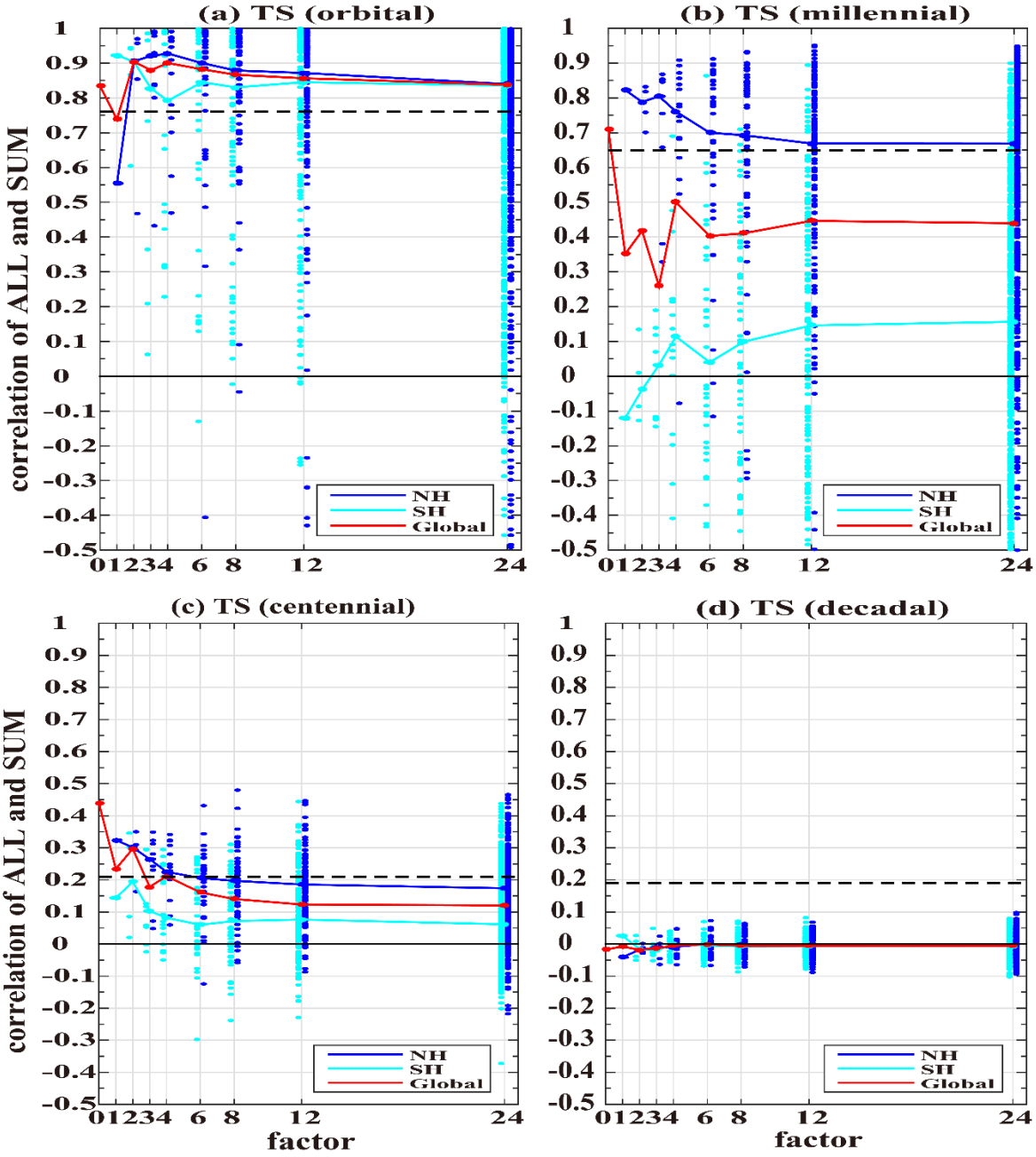


Figure 3: Correlation coefficient of mean surface temperature between the ALL and SUM outputs in different spatial-time scales. a, orbital; b, millennial; c, centennial; d, decadal. The blue dots for NH, cyan dots for SH and red dots for all the correlation coefficient median in the same spatial scale (i.e. factor) of global. The red line is connecting the median dots at all division factors. The blue (cyan) line is connecting the median of all blue (cyan) dots at all division factors. The black thick dashed line is 95% confidence level. The black thin solid line is zero. The x axis is division factor, and y axis is correlation coefficient. There are only about 10 points with the correlation coefficient lower than -0.5 so they are neglected in a-c.

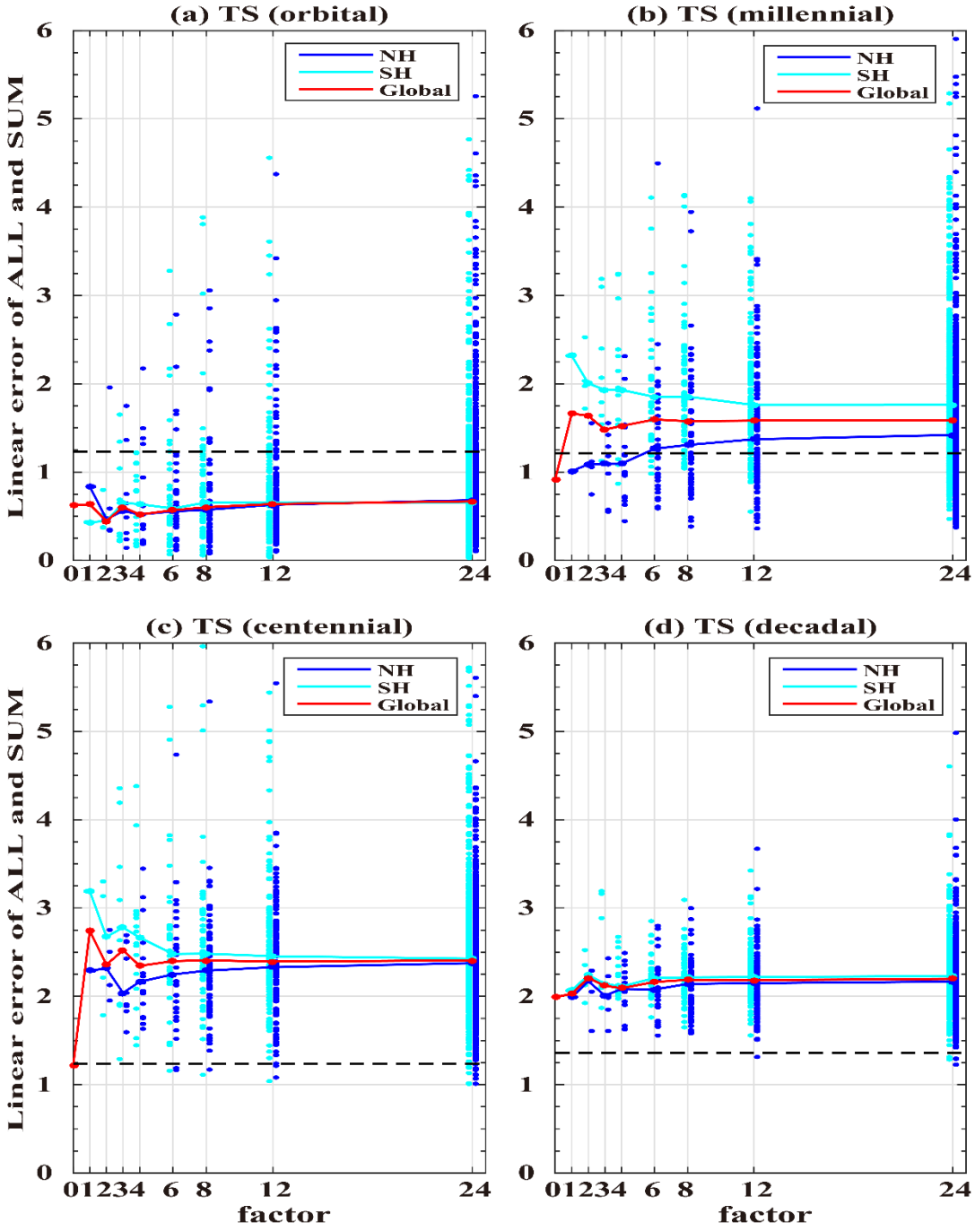


Figure 4: Same as Figure 3, but for linear error (L_e, y axis). The black thick dashed line is 95% confidence level of L_e. There are only about 10 points with the L_e larger than 6 so they are neglected in a-c.

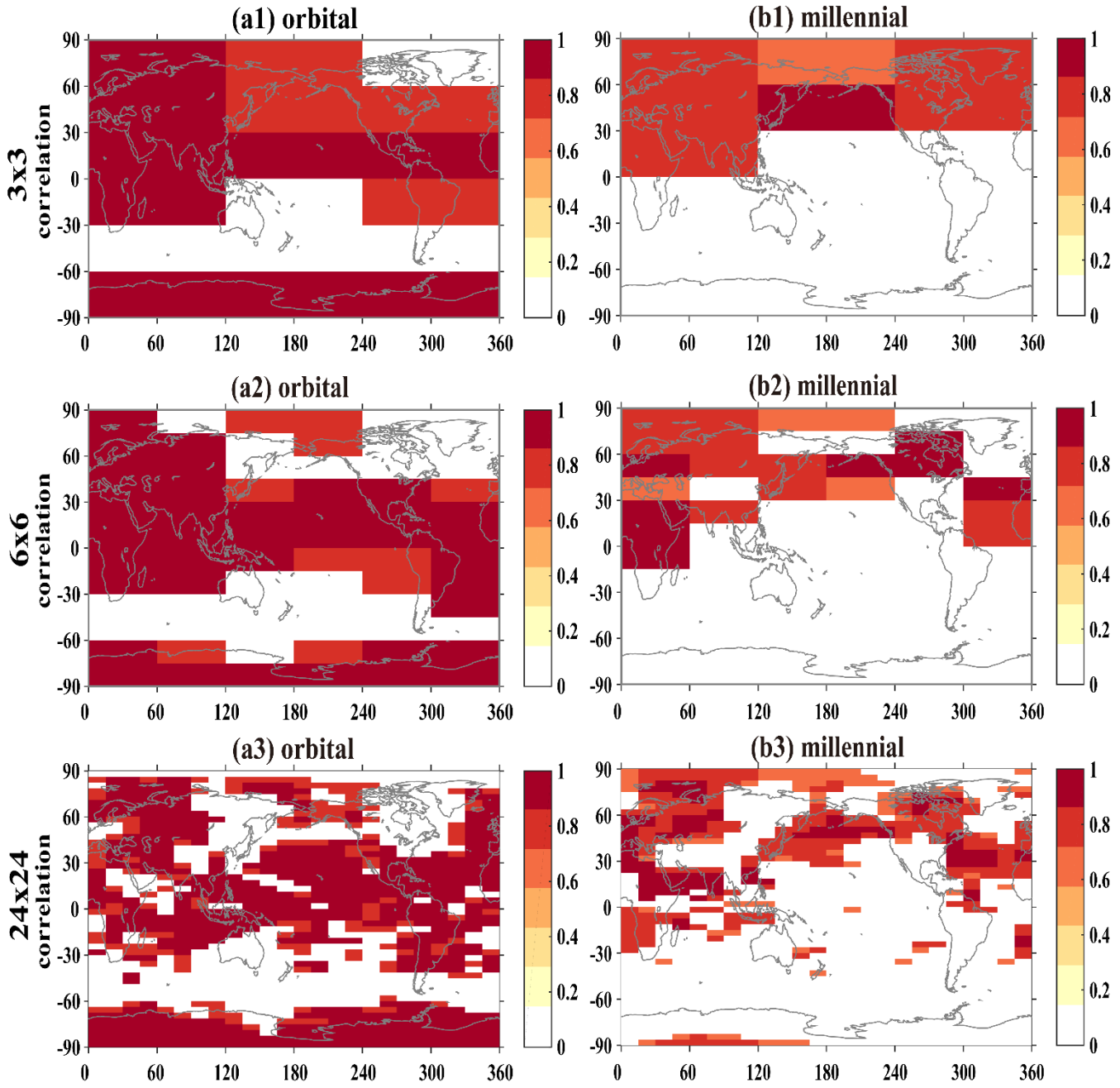


Figure 5: Correlation coefficient of mean surface temperature between the ALL and SUM outputs of the two timescales (a1-a3, orbital; b1-b3, millennial) for three representative spatial scales, *f=3, 6* and *24* (the other factors are not shown because they are similar to the abovementioned three representative spatial scales). Only those regions of significant at 95% confidence level are shaded in colors.

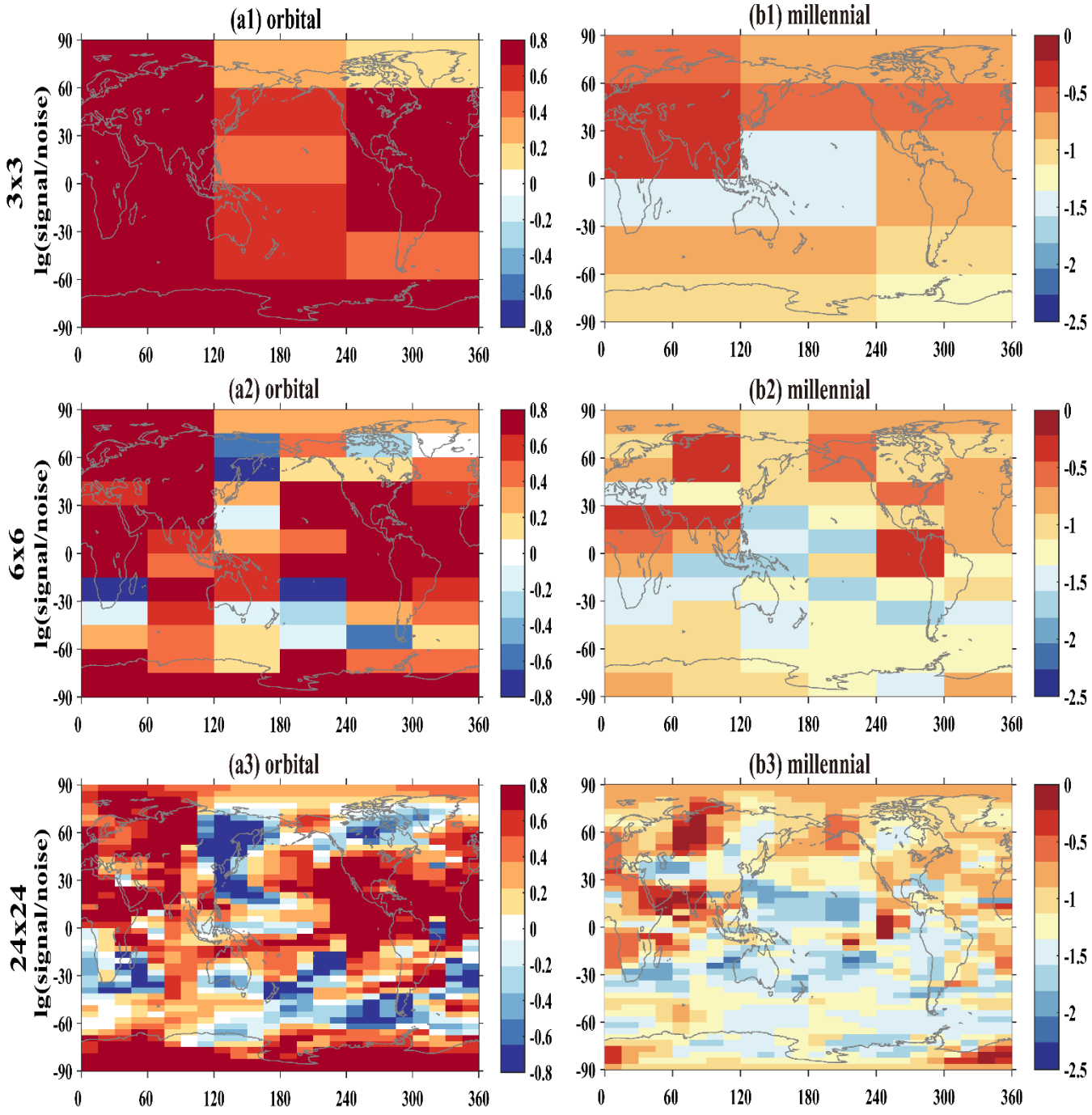


Figure 6: The signal-to-noise ratios on the orbital (a1-a3) and millennial (b1-b3) timescales derived from the ALL run. Here the signal is used by the orbital (millennial) variability variance, and the noise is used by the sum of the centennial and decadal variabilities variance. The numbers of 1, 2 and 3 are for the three representative spatial scales, *f*=*3*, *6* and *24*, respectively. In order to show the signal clearly, the log base 10 is taken on the signal to noise ratios.

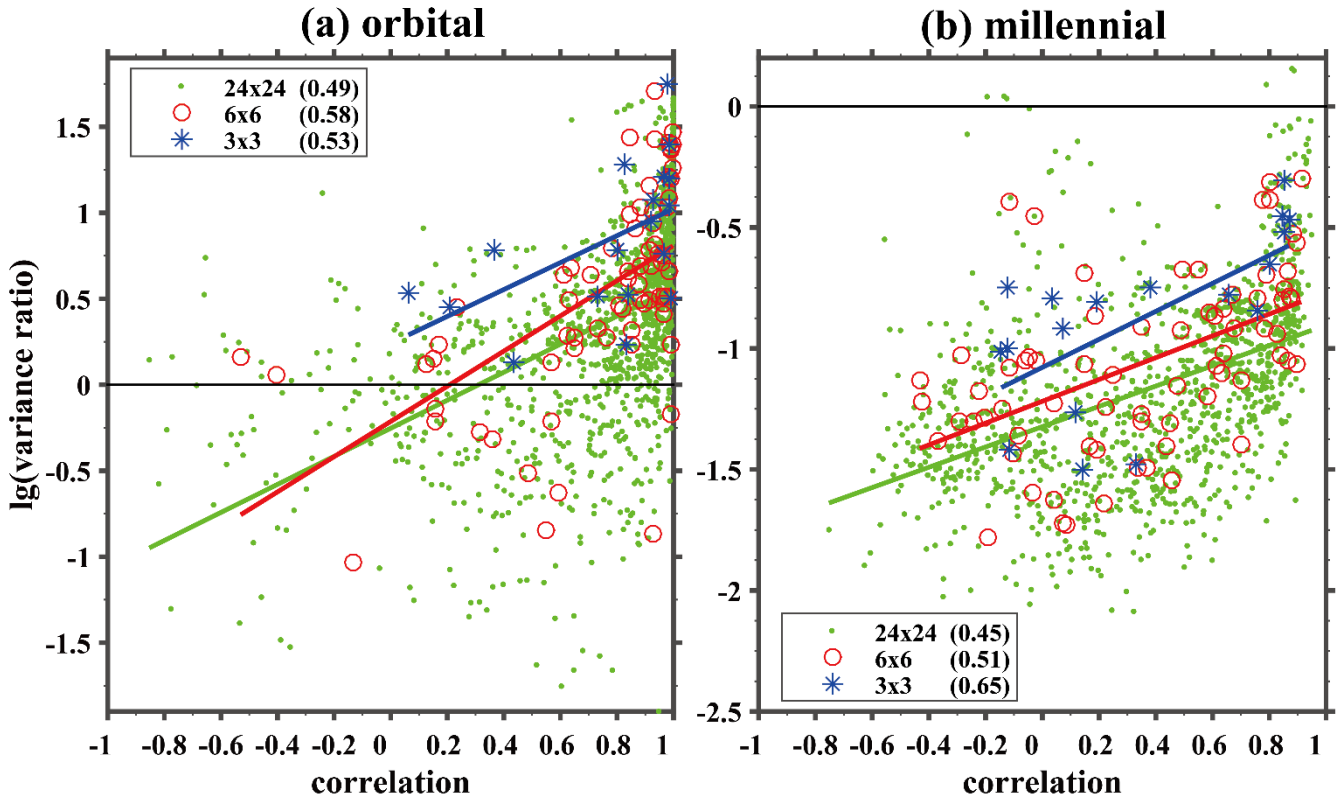


Figure 7: Scatter diagram of the correlation coefficient and the signal-to-noise ratio (variance ratio) on the orbital (a) and millennial (b) time scale. The log base 10 is taken on the variance ratio, as indicated in Figure 6. Only those for the three representative spatial factors (*f=3, 6* and *24*) are shown in the panels. The blue stars and linear fitting line are for factor 3, the red circles and linear fitting line are for factor 6, the green dots and linear fitting line are for factor 24. The fitting coefficients are list in the brackets at the upper (lower) left corners of the panel a (b).