Response to the referee comments on essd-2026-295

We thank the referees for their careful review and constructive comments and suggestions. We have addressed all comments and revised the manuscript accordingly. Each referee comment is followed by our point-by-point response.

Referee1:

1) The reconstruction methodology is reasonable, but, as applied here, it has clear drawbacks, namely, it cannot produce uncertainty bounds. So the reconstruction shown in Figure 7 is just one time series without any uncertainties attached. This is not the state-of-the-art. Whereas I am aware that estimating uncertainties is to some extent subjective, as it may involve different sources of uncertainty, the authors should provide uncertainty bounds. For instance, the uncertainties may be related to the choice of analogue, the subsampling of the proxy network, the use of some model pools and not others, etc. The goal should not be to produce uncertainty ranges that take into account all possible sources of uncertainty - this would probably be too complex- but the method should be able to answer the question, for instance, of whether the temperature minimum in 1800 could be 0.4, or 0.5, or 0.6 instead of 0.3 K. The reader should have a sense of the degree of reliability the authors place on their reconstruction, given this methodology.

Perhaps one way to address this point is to compare the reconstructions with observations from the 20th-21st centuries. The authors show that the reconstruction underestimates trends and variability, so these biases and uncertainties should also be reflected in the 600-year reconstruction. The series of annual errors has a spectrum in the frequency space, so one way could be to extrapolate those amplitudes to the multidecadal and centennial timescales.

These are just suggestions, but this point must be somehow addressed.

Response: Thank you for highlighting the need to quantify the reliability of the reconstruction. In the revised manuscript, we added a 1,000-member ensemble

analysis to quantify the uncertainty arising from choosing different, similarly well-matched analogues.

For each reconstruction year, we first identified all analogues with a positive and significant correlation with the proxy records (*correlation coefficient > 0* and *p<0.05*). These analogues were ranked by matching distance, and up to the ten best-matched analogues were retained. Then we excluded those whose matching distance exceeded 110% of the smallest distance for that year. One analogue was randomly selected from the remaining analogues, with selection probabilities weighted by matching distance as follows:

$$P_{i,Y} = \frac{e^{-\alpha(\frac{d_{i,Y}}{d_{best,Y}}-1)}}{\sum_{j=1}^{K_Y} e^{-\alpha(\frac{d_{j,Y}}{d_{best,Y}}-1)}}, \quad i = 1,\ldots,K_Y$$

where $P_{i,Y}$ is the probability of selecting the *i-th* retained analogue for reconstruction year Y; $d_{i,Y}$ is the matching distance; $d_{best,Y}$ is the smallest matching distance for that year; K_Y is the number of analogues remaining after all filtering steps, $1 \le K_Y \le 10$; and α controls how rapidly the selection probability decreases as the relative matching distance increases and was set to 20 in this study. The denominator sums only over the K_Y retained analogues for that year. According to this equation, analogues closer to the best analogue received higher sampling probabilities. If only one analogue remained after applying the distance threshold, it was selected for all ensemble members.

The temperature field of the selected analogue was used as the reconstruction for that year. Repeating this selection for every year from 1400 to 2000 produced one complete reconstruction, referred to as an ensemble member. We generated 1000 such reconstructions to form an ensemble and calculated the 5th and 95th percentiles for each year to define the analogue-selection uncertainty range. This range represents the

uncertainty caused by the non-uniqueness of closely matched analogues rather than a formal confidence interval.

We have clarified this scope in the revised Section 3.3 Assumptions on uncertainty and 4.1.1 Regional mean temperature variations. The annual uncertainty range is now shown as blue shading in the revised Figure 7.

2) That the method underestimates the variability, also in the instrumental period, is somewhat surprising, since the method selects just one analogue. Selecting several analogues and calculating the mean would indeed reduce variance, but selecting only the best analogue and assuming that the climate models are perfect should yield the observed variance, not an underestimated one. The reason is that the analogue method just resamples the simulated data.
The simulated data and the observations are all standardised to unit variance, but in the last step the analogues are re-scaled to the original variability of the (model or CRU?) data. If the variance of the model data is smaller than that of the observations, the reconstructed variability is underestimated. This should be checked.
Response: Thank you for raising this important point. We agree that selecting a single analogue avoids the variance reduction associated with averaging several analogues. However, this does not necessarily preserve the variability of the analogue pool. To identify the sources of variability loss in this study, we examined the standardization and rescaling steps, as well as the distribution of the selected analogues.

First, we examined whether the variability loss originated from the model simulations or from the standardization and rescaling steps. During 1902-1977, the standard deviation of CRU regional mean temperature was 0.27 °C, whereas the median standard deviation across the 37 original model simulations was 0.36 °C, equivalent to 131.4% of the CRU value. The original simulations therefore had sufficient temperature variability. Each model temperature field was then standardized using its own mean and standard deviation for 1902-1977. After this step, the standard

deviation of regional mean temperature in the model simulations remained 94.3% of that in the standardized CRU data. Finally, the selected standardized analogue fields were converted back to degrees Celsius using the CRU mean and standard deviation, rather than the model values. This rescaling had only a small effect on the variability of the reconstruction. These results indicate that the variability loss cannot be largely explained by insufficient variability in the original simulations or by the standardization and rescaling steps.

Second, the selected analogues show much less variation in regional mean temperature than the complete analogue pool. For the 601-year reconstruction, the standard deviation of the regional temperature anomalies of all 8,070 analogues was 0.81 in standardized units, compared with 0.38 for the 601 selected analogues. The 5th and 95th percentiles were -1.37 and 1.33 for the complete pool, but -0.84 and 0.48 for the selected analogues, indicating that strong cold and warm anomalies were less represented in the selected sequence. Selecting one analogue for each year therefore does not ensure that the reconstruction retains the variability of the complete pool, because each selection prioritizes the closest match to the proxy pattern. These results show that temperature variability was substantially reduced after analogue selection.

Then we performed two additional diagnostics to investigate the causes of variability loss. First, we examined how the selected regional temperature anomalies relate to the proxy data. The weighted mean of the proxy anomalies correlated with the selected analogue's regional temperature anomaly at $r=0.70$ ($p<0.05$), whereas the weighted root mean square (RMS) of the proxy anomalies correlated with the absolute regional temperature anomaly at $r=0.36$. These results suggest that the selected analogues follow changes between warmer and colder conditions more closely than changes in anomaly magnitude.

We also tested whether selecting stronger regional anomalies among similarly matched analogues could increase the reconstructed variability. We constructed two

sensitivity reconstructions using the five and ten best-matched analogues for each year, respectively. An alternative analogue was eligible only if its matching distance was no greater than 110% of the top-1 distance, its regional anomaly had the same sign as the top-1 anomaly, and its absolute regional anomaly exceeded that of top-1 by at least 0.20 standardized units. When multiple analogues met these criteria, we selected the one with the largest absolute regional anomaly; otherwise, we retained top-1 if none qualified. Replacements occurred in 252 and 329 of the 601 years in the top-5 and top-10 tests, respectively. The standard deviation over 1400-2000 consequently increased from 0.38 in the original top-1 reconstruction to 0.48 and 0.53 in the two sensitivity reconstructions. Thus, stronger regional anomalies were available for many years without substantially increasing the matching distance, but they were not selected by the original minimum-distance criterion.

In summary, these diagnostics indicate that analogue selection is a major source of the reduced temperature variability. Although selecting a single analogue avoids smoothing caused by averaging multiple analogues, minimizing the matching distance to the proxy pattern does not ensure that the selected regional temperature anomaly has the appropriate magnitude. The sensitivity tests show that alternatives with stronger regional anomalies are available for many years at similar matching distances. However, the resulting increase in variability does not by itself establish that these alternatives provide more accurate reconstructions. We have added these findings and discussed the underestimation of temperature variability as a limitation of the reconstruction in Section 5 of the revised manuscript and in Section 3 of Supplementary Materials.

3) In addition to this underestimation in the observational period, an additional underestimation is expected, since most of the simulated data stem from historical (1850-present) simulations, and thus do not contain analogues that go beyond the observational period. How does the reconstruction change when only the past 1000 simulations are included in the analogue pool?

So, points 2 and 3 together could lead to a too-flat reconstruction, as shown in Figure 7.

Response: Thanks for the specific suggestion about the composition of the analogue pool may influence the reconstructed variability. We further set sensitivity experiments based on three analogue pools: (1) the original pool containing 8,070 analogues from 37 models; (2) a pool of 2,328 analogues covering 851-2014 from ACCESS-ESM1-5 and MRI-ESM2-0; and (3) a past1000-only pool of 1,998 analogues covering 851-1849 from the same two models.

Our additional experiments show that analogue-pool composition affects both annual and multidecadal temperature variability in the reconstruction. The long 2-model and Past1000-only reconstructions retain more of the observed variability than the original reconstruction and show larger multidecadal variability before 1850. The differences between the two pools using the same models further indicate that the temporal coverage of the analogue pool influences the reconstructed amplitude. We have added these results and a discussion of their implications to Section 5 along with a new figure (Figure 10).

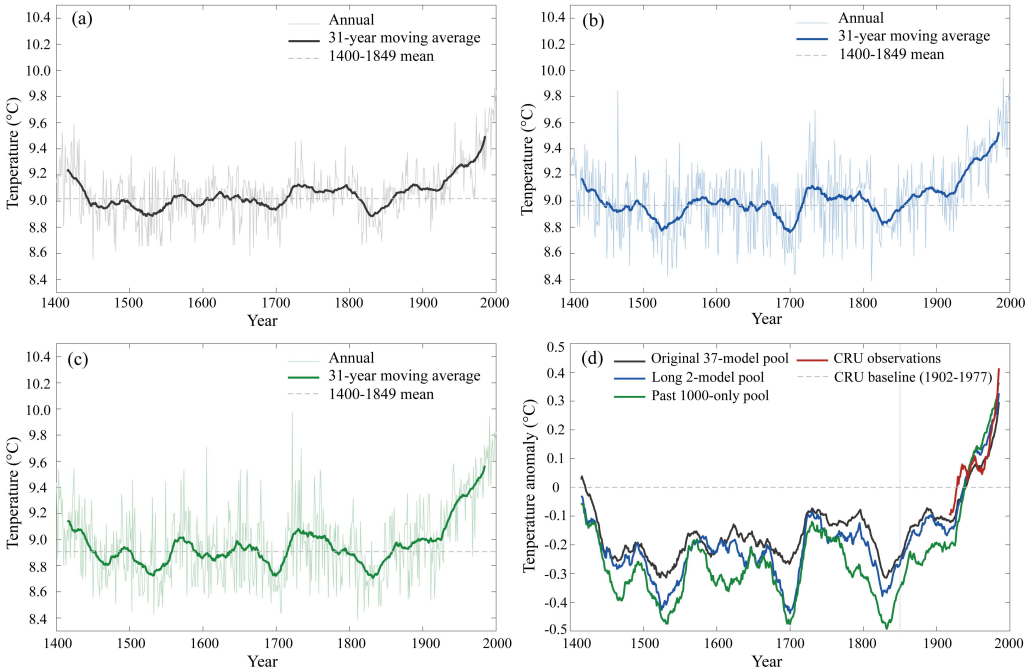


Figure 10: Temperature reconstructions from three analogue pools (1400–2000): (a) original 37-model pool; (b) long 2-model pool; (c) Past1000-only pool; (d) 31-year moving averages compared with CRU

observations.

Particular points

4) Table S1 should specify the type of tree-ring data (ring-width, wood density,..early wood density, etc). The same applies to the ice-core data.

Response: Thank you for this helpful suggestion. We have revised Table S1 to specify the measured variable for each proxy record. We have also added the elevation information for each proxy site to provide a more complete description of the proxy network. Please see Section 1 and Table S1 in Supplementary Materials for details.

Referee2:

(1) The authors attempt to develop a 600-year gridded temperature dataset, and this long temporal coverage is the main feature and selling point. I suggest that in the Introduction they first present the existing mainstream gridded temperature datasets for East Asia, then clearly explain why a temperature dataset with such a long time series is needed and what value such a long record can offer.

Response: Thank you for this constructive suggestion. We have revised the Introduction to first introduce widely used gridded temperature datasets covering East Asia, including CRU TS, Berkeley Earth, and ERA5, and to explain their temporal coverage and applications. These products provide essential information on recent temperature variations, but their limited temporal coverage restricts the assessment of variability over several centuries.

We then explain why a longer record is needed. A temperature dataset spanning 1400-2000 provides a longer baseline for assessing multi-decadal and centennial variability and placing recent warming in the context of earlier temperature fluctuations. Its gridded structure also allows readers to examine regional differences and spatial patterns of past temperature variations across East Asia, beyond those represented by a single regional mean series. Therefore, the revised Introduction

connects the temporal limitations of existing datasets with the specific value of the long gridded temperature record developed in this study.

(2) I did not fully understand this point: why is the annual mean defined from previous October to current September (as stated in "In this study, the monthly data were calculated to annual means (previous October to current September) to match the temporal resolution of the proxy network")? Are all the proxy records treated in the same way? Please clarify.

Response: Thank you for asking us to clarify this point. We chose the previous October-current September window because most of our proxy records are tree-ring records, and tree growth can be influenced by both the preceding autumn and winter and the current growing season. Because these records are used to select analogues, this window was chosen to better align the observed and simulated temperatures with the temperature signal recorded by the tree-ring proxies.

This time window was only applied to the monthly instrumental and model temperature data. The proxy records, including tree-ring, ice-core, and documentary records, retained their original annual values and year labels. For example, a proxy value dated to 1902 was matched with the mean temperature from October 1901 to September 1902.

We have revised Sections 2.1 and 3.2 to distinguish the annual resolution of the proxy records from the averaging window used to define the temperature target.

(3) Please check the wording and symbols in lines 146 and 171.

Response: Thanks for your suggestion. We have revised the wording and clarified the symbols in the passages corresponding to lines 146 and 171, including the paired values used to calculate the correlation and the explanation of notation in Equation (6).

(4) Please check Equation (2) and Equation (6).

Response: Thank you for pointing this out. We have checked the equations and corrected the notation for the inverse sine function. We have also clarified that latitude and longitude are expressed in radians. Please see Equation (2) and (6) in the revised manuscript.

(5) Section 4.1 is the core of the whole paper, and I think it should be further strengthened. Since the CRU data were used in the reconstruction process, using the same CRU data for validation can only partially address the issue. I suggest that the authors find additional independent datasets. These could be divided into two categories: one covering the recent decades, and another covering several centuries (if such data exist), and then perform a thorough comparative validation.

Response: Thank you for this constructive suggestion. We added comparisons with additional datasets in Section 4.1.1 and showed them in Figure 4. The expanded evaluation includes instrumental and reanalysis products covering recent decades. Furthermore, the comparisons with published temperature reconstructions covering several centuries are shown in Section 5 (Discussion).

For the recent period, in addition to the original comparison with CRU, we compared the EA reconstruction with Berkeley Earth monthly land temperatures (BEST-LAND) and ERA5 monthly 2 m temperatures. Both products were processed using the same study region and annual temperature definition as our reconstruction. Temperature anomalies were calculated relative to the 1961-1990 mean of each series. The four series show similar temperature variations over their respective overlapping periods (Figure 4). Specifically, during 1902-2000, the reconstruction was correlated with CRU ($r=0.740$, $p<0.05$) and BEST-LAND ($r=0.758$, $p<0.05$). During 1941-2000, it was correlated with CRU ($r=0.695$, $p<0.05$), BEST-LAND ($r=0.692$, $p<0.05$), and ERA5 ($r=0.644$, $p<0.05$). The results here provide a brief overview of the comparison. Further details are presented in Section 4.1.1 of the revised manuscript. In a word, the additional comparisons with BEST-LAND and ERA5 provide further support for the ability of the reconstruction to represent annual temperature variations in EA.

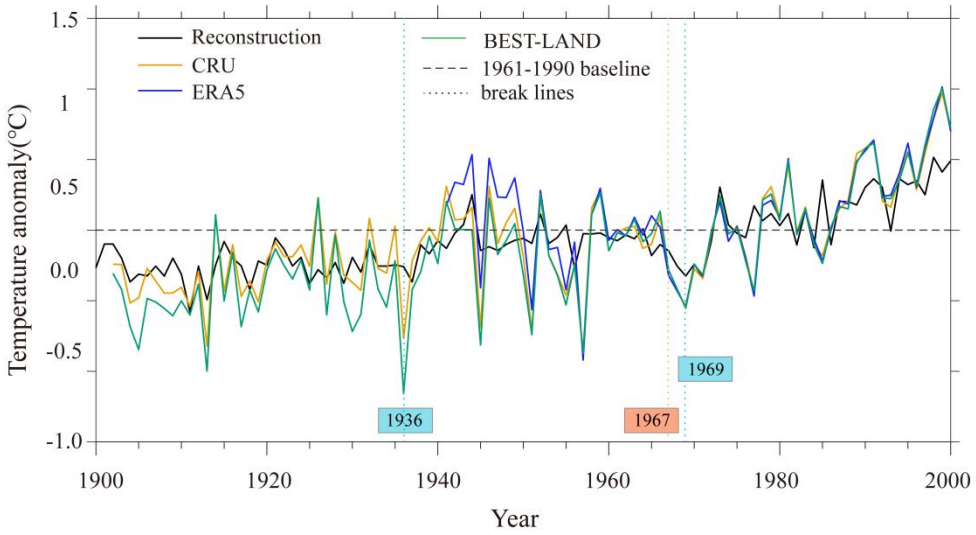


Figure 4: Temporal variations (units: ℃) of EA mean temperature anomalies from the reconstruction, CRU, BEST-LAND, and ERA5. Blue and orange dotted lines indicate the breakpoints identified for 1902-2000 and 1941-2000, respectively.

For the longer-term comparison, we compared our reconstruction with published temperature reconstructions for the globe, the Northern Hemisphere, Asia, and China in the Section5 Discussion. We discussed their similarities and differences in temperature variations and long-term trends.

These comparisons assess complementary aspects of the reconstruction: the recent-period comparisons show that the reconstruction follows the temperature variations shown by multiple reference datasets, while the multi-century comparisons place the East Asia record in the broader context of regional, hemispheric, and global temperature variations.

(6) If possible, please provide an uncertainty analysis, or accompany the dataset with an uncertainty signal (for example, indicating whether the reconstructed value for a given year has high or low confidence). This would be very helpful for readers to interpret the results.

Response: Thank you for this helpful suggestion. To help understand the reconstructed values for each year, we assessed uncertainty in analogue selection by generating an ensemble of 1000 temperature reconstructions for 1400-2000. For each year, eligible analogues were required to have a positive(correlation coefficient > 0)

and statistically significant correlation($p < 0.05$)with the proxy records. We selected up to ten analogues with the smallest matching distances and retained only those with distances no greater than 110% of the minimum distance for that year. From these analogues, we randomly selected one and used its temperature field to reconstruct that year. The probability of selection decreased with increasing matching distance according to the following equation:

$$P_{i,Y} = \frac{e^{-\alpha(\frac{d_{i,Y}}{d_{best,Y}}-1)}}{\sum_{j=1}^{K_Y} e^{-\alpha(\frac{d_{j,Y}}{d_{best,Y}}-1)}}, \quad i = 1,...,K_Y$$

where $P_{i,Y}$ is the probability of selecting the *i-th* retained analogue for reconstruction year Y; $d_{i,Y}$ is the matching distance; $d_{best,Y}$ is the smallest matching distance for that year; K_Y is the number of analogues remaining after all filtering steps, $1 \le K_Y \le 10$; and α controls how rapidly the selection probability decreases as the relative matching distance increases and was set to 20 in this study. The denominator sums only over the K_Y retained analogues for that year. According to this equation, analogues closer to the best analogue received higher sampling probabilities. If only one analogue remained after applying the distance threshold, it was selected for all ensemble members.

Applying this sampling process to all reconstruction years produced one complete ensemble member. We repeated the process 1000 times and used the 5th and 95th percentiles of the resulting regional mean temperatures for each year as the bounds of the analogue-selection uncertainty range. This range describes the spread arising from alternative choices among closely matched analogues. It does not constitute a formal confidence interval.

We have explained what this uncertainty range represents in Section 3.3 of the revised manuscript and displayed the annual range as blue shading in the revised Figure 7.