Dear Editor, this is the point-by-point response to Reviewer2's comments and suggestions.

We thank the reviewer for their constructive feedback. We believe the manuscript is significantly improved as a result of these revisions.

The referee's comments are shown in black. Our responses are shown in blue and the added or modified texts are shown in blue italics.

Major comments

1. Usually it is a good idea in these cases to implement "variance stabilization" in chronology development: adjust the time-varying chronology variance for the changing sample size (number of trees or cores). If variance stabilization was used, you should report that in the methods. If not, I suggest doing a quick check to see if it makes a difference to conclusions.

   To address this concern, we have performed a detailed comparative analysis between the reconstruction used in our original manuscript (based on a residual chronology) and an alternative reconstruction. This alternative was developed using a variance stabilization method, which corrects for artifactual changes in variance that arise from fluctuating sample replication over time (implemented as the `chron.stabilized` function in the dplR package).

   The comparative analysis was approached from two perspectives:

   1. **Comparison of the full time series:** The two reconstructions are overwhelmingly similar, exhibiting a Pearson correlation of **r = 0.92**, which indicates they share approximately 85 % of their variance. A paired t-test found no significant difference in their means ($p = 0.087$), and a Kolmogorov-Smirnov test confirmed that their overall probability distributions are statistically indistinguishable ($p = 0.246$).
   2. **Analysis of the frequency and distribution of extremes:** Since a key part of our study's conclusions relies on extreme events, we performed specific statistical tests (McNemar's and Chi-squared tests) to compare the identification of these events. The results were conclusive: **no statistically significant differences were found** either in the classification of extreme years (McNemar's tests, $p = 1.0$ in all cases) or in their temporal distribution across centuries (Chi-squared tests, $p > 0.48$ in all cases).

   Beyond this empirical validation, we also find the reviewer's point insightful and have addressed it with an additional, direct test. The concern is particularly relevant to our chronology, where the combination of historical and modern sampling campaigns results in a stepwise decrease in sample replication in the most recent decades. To specifically test the influence of this structure, we created an independent chronology using only the modern, high-replication samples, thus removing the drop in sample depth. This "modern-only" reconstruction is statistically indistinguishable from our main chronology, confirming that our results are not an artifact of changing replication. This provides direct, data-driven evidence that the main conclusion of our study is robust and not biased by the chronology's structure.

   Taken together, these analyses robustly demonstrate that the main conclusion of our study is not an artifact of the chosen chronology development method. The statistical similarity between both reconstructions, particularly with respect to extreme events, further confirms the validity of our findings.

   We have updated the **Chronology development** subsection of the manuscript (3.1, after current line 227) to include a summary of this sensitivity analysis:

   "Additionally, to ensure the robustness of our findings to the chosen methodology, a comprehensive

sensitivity analysis was performed. An alternative reconstruction was developed using a variance stabilization method (chron.stabilized in dplR; Wigley et al., 1984; Frank et al., 2007), which corrects for variance changes associated with fluctuating sample replication over time. A thorough statistical comparison revealed that both reconstructions are highly similar (Pearson r = 0.92) and exhibit no significant differences in their means (paired t-test, p = 0.087) or their overall probability distributions (Kolmogorov-Smirnov test, p = 0.246). Crucially, a specific analysis of extreme events showed no statistically significant differences in their frequency or temporal distribution (McNemar's and Chi-squared tests, p > 0.48). Given that the study's conclusions are robust to the choice of method, and considering the theoretical suitability of the residual chronology for calibration, the latter was retained for all analyses presented. Full details of this comparative analysis can be found in the Supplementary material (Table S3).”

Added in the Supplementary material:

Table S1: Statistical comparison of the full time series from the residual and variance-stabilized reconstructions over the common period 1505-2024.

Test	*Statistic*	*Value*	*p-value*	*Interpretation*
Pearsons correlation	*r*	*0.920*	*< 0.001*	*Very strong positive correlation between the series.*
Paired t-test	*t (df = 519)*	*1.715*	*0.087*	*No significant difference between the mean values of the two series.*
Kolmogorov-Smirnov	*D*	*0.063*	*0.246*	*The probability distributions of the two series are statistically indistinguishable.*

2. Distinction of “hydroclimate” from “precipitation.” Water availability, which is mentioned in several places in the paper, is a function of not just precipitation but also of evapotranspiration. The authors claim that a strong point of this paper compared with previous works is that it addresses precipitation rather than PDSI or some other type of drought index. This may be a strong point in terms of synoptic climatology and delivery of precipitation to the region, but not necessarily in terms of water resources availability or ecosystem stress. It seems likely that the tree growth of these drought sensitive species would not better reflect the combined stress of low precipitation and high temperature than the stress of low precipitation alone. A natural question is whether the impressive calibration strength (R-squared) for the reconstruction model might not be even more impressive for some sort of drought index that incorporating influence of evapotranspiration. I think the paper could benefit from some comparison statistics. If checking against and index such as SPEI or PDSI, the standard rather than the residual chronology might actually be worth looking at because the strong temperature trend associated with regional warming is a low-frequency signal that could have been removed by autoregressive modeling during standardization. Such additional analysis, if done, would mainly be a discussion point, as readers may wonder why the calibration of a drought-sensitive chronology is not stronger with SPEI, say, than with precipitation.

and

3. I appreciate the rigorous daily climate analysis, which resulted in identification of an optimum window, but wonder about the sampling variability and whether you are making too much of the difference in a annual window and a 320-day window. In the minor comments I bring this up again and suggest adding maybe a sentence about this issue in the discussion. Almost certainly there is no significant different in correlation for the selected 320-day window and the highest 365-day (annual) window.

Regarding major comments 2 and 3, we have reviewed the text and changed “water availability” as it follows:

-Lines 77 and 485: replaced with “weather conditions”.

-Line 504: replaced with “precipitation and droughts”.

We have checked the correlation against SPEI, and added the following (after current line 419):

“A relevant question concerns the choice of the reconstruction target, specifically why precipitation was selected over a more integrated drought index like the Standardized Precipitation-Evapotranspiration Index (SPEI) (Beguería and Vicente-Serrano, 2011). To address this, we conducted a direct comparative analysis to empirically determine the optimal target variable. We performed a series of calibrations testing both our residual and standard chronologies against three potential climate targets: our 320-day precipitation window, a standard 12-month precipitation window, and a 12-month SPEI.

The results, detailed in the Supplementary material (Table S4), demonstrate a clear and consistent hierarchy in model performance. Our selected 320-day window produced a stronger calibration than a standard 12-month window. Crucially, the model with the highest predictive skill was the one used in our study, which calibrates the residual chronology against 320-day precipitation Therefore, the selection of this specific target and predictor is not only based on our initial response function analysis but is also validated by this comprehensive comparative test as the statistically most robust approach for this dataset. It is worth noting, however, that while the 320-day window is empirically superior, the calibration strength of the 12-month window remains high, with only minor differences between the two.”

Added in the Supplementary material:

Table S2: Comparison of calibration statistics for different climate targets and chronology types. The climate window is 320 days ending June 30 for precipitation and 12 months ending June for SPEI and 12-month precipitation. All calibrations are for the period 1951–2022.

Chronology type	***Climate target***	***Calibration Pearson r***	***p-value***
Residual	*Precipitation (320 days, ending June 30)*	*0.749*	*<0.001*
Standard	*Precipitation (320 days, ending June 30)*	*0.746*	*<0.001*
Standard	*Precipitation (365 days, ending June 30)*	*0.741*	*<0.001*
Residual	*Precipitation (365 days, ending June 30)*	*0.738*	*<0.001*
Standard	*SPEI (12-month scale, ending June)*	*0.723*	*<0.001*
Residual	*SPEI (12-month scale, ending June)*	*0.721*	*<0.001*

Minor comments

1. Title: Consider substituting “precipitation” for “hydroclimate,” in the title, because precipitation is what has been reconstructed. Precipitation is just one aspect of hydroclimate. Runoff and streamflow are the sum of net precipitation (P-ET) and are what I think of as key components of hydroclimate, though this is an arguable distinction.

   Done. New title: “*A five-century tree-ring record from Spain reveals recent intensification of western Mediterranean precipitation extremes*”.

2. L77. I disagree that precipitation provides a “more direct measure of past water availability than some drought index. Net precipitation (P-ET) is one possible drought index, and is actually more relevant to water availability the precipitation alone. Of course, precipitation is more directly link weather delivery systems that P-ET, which depends on vegetation and other land surface factors.

   Fair enough.
   -Lines 77 and 485: replaced with *“weather conditions”.*

-Line 504: replaced with *"precipitation and droughts".*

3. L90. "pluviosity" is an overblown word when used here for "precipitation, " which explicitly is what is shown in the climate diagram, and what is measured in a rain gauge.

   Done, now changed to *"precipitation".*

4. L90. On the climate diagram I see May followed by April, not June followed by May, as the months of highest precipitation.

   Thank you. Now corrected to*: "May is the month with the highest precipitation, followed by April".*

5. L92. Looks to me like Feb is a drier month month than Aug. The statement about July and Aug being driest month applies only if just considering summer .

   Ok. Now changed to: *"... showing July and August as the months with the highest mean temperatures, with July also being the driest month".*

6. L114. "Campaigns at…".

   changed "in" to "at".

7. L124. Standardization description needs a bit more information. Was the ratio or difference method used for converting ring widths to indices? Was the site chronology computed as an arithmetic mean or biweight mean of core indices? Was ariance stabilization applied to adjust variance changes in site chronology to time-varying sample size (see Major comments)? Did you compute both standard and residual versions of the chronology, and why did you select the residual version for the reconstruction

   and

8. L 128. How many trees are represented by the 173 series? I'm assuming probably more than one core per sampled tree.

   Minor comments 7 and 8 have been taken into account in order to improve the clarity of the explanations provided regarding the chronology development. In addition to the response to major comment 1, which also covers these topics, current lines 124–129 have now been changed to:

   "After measuring the samples, each individual ring-width series was standardized to remove age/size-related trends and to minimize non-climatic noise. This was performed using the dplR package in R. Following an adaptive approach based on series length, a negative exponential curve or a cubic smoothing spline of variable stiffness was fitted to each series. Tree-ring indices were then calculated as ratios by dividing the raw ring-width measurements by the fitted curve values.

   These individual indices were then combined into a regional chronology using a biweight robust mean, a method that minimizes the influence of outliers. Both a standard and a residual version of the chronology were computed. The residual chronology was selected for the final reconstruction because it is generated via prewhitening, a procedure that fits an autoregressive (AR) model to the standard chronology and removes the statistical autocorrelation inherent in tree growth. (Cook and Kairiukstis, 1990).

   In addition, a sensitivity analysis was conducted to test the effect of an alternative method, variance stabilization, on our results. As detailed in the Supplementary Table S.3, this approach did not improve the calibration skill. Therefore, the residual chronology was retained as it worked as the strongest predictor. After excluding a subset of older series with low signal strength (correlation < 0.3) or that were inaccessible for remeasurement, the final dataset consisted of 173 individual tree-ring series from 103 different trees."

9. L 153. “ with12-” ….insert a space

   Done.

10. L159. “robust” -- I assume the daily window selected is robust to selected segment of the climate-chronology overlap (e.g., approximately same day window if analysis repeated on separate halves of the record)

    Yes, we tested in separate halves of the record (1952–1986 and 1987–2022) and both the windows and ending dates of the windows were similar in the two halves.

11. L 163. “linear transfer function model” --- the statistical reconstruction method could be described more directly as “simple linear regression of the target predictand on the site chronology.”

    Ok. Changed to *“simple linear regression of the target predictand on the site chronology”.*

12. L208 “Such a low AR1 value indicates that the standardization effectively removed most of the tree-ring memory persistence inherent in tree growth, yielding a time series suitable for robust correlation analysis with external environmental variables, such as climate.” Yes, as long as the target predictand also has no autocorrelation. Is that so for precipitation in this region? Also, in regression, analysis of residuals check usually included first order autocorrelation of regression residuals (e.g., by Durbin Watson statistic). It is assumed in regression that there is no significant lag-1 autocorrelation in the regression residuals.

    Thanks. We ran the tests and we have edited the current lines 208–210 in the manuscript, which now say:
    “The first-order autocorrelation (AR1) coefficient for this residual chronology was 0.039. This low value indicates that standardization effectively removed the non-climatic, biological persistence inherent in tree growth. The instrumental precipitation target also exhibited negligible first-order autocorrelation (AR1 = –0.037). This ensures that both predictor and predictand are suitable for correlation analysis. More critically, the residuals of the final linear reconstruction model were tested for autocorrelation using the Durbin-Watson statistic; the test was not significant (DW = 2.14; p = 0.729), confirming that the model meets the assumption of independent errors required for linear regression (Cook and Kairiukstis, 1990)”

13. *Fig 1 caption. “grid cells” – would help the interpretation to indicate the resolution of grid for the precipitation. .*

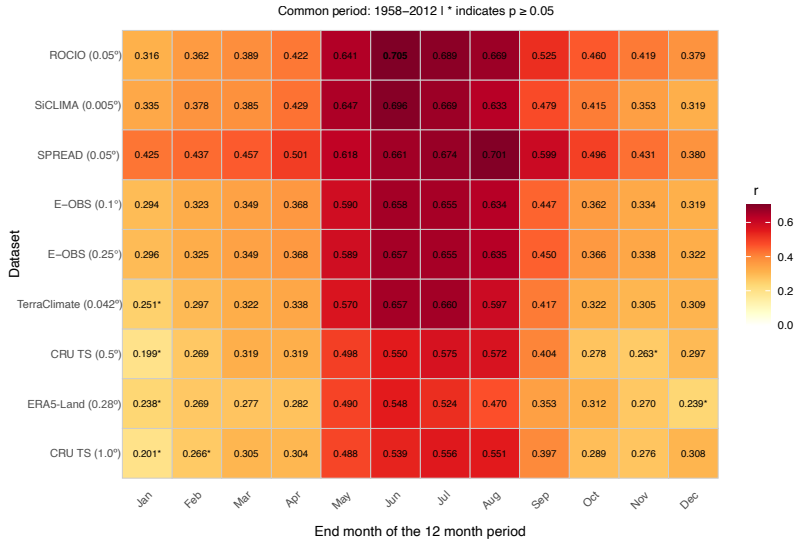
    Modified as it follows:
    “Climate diagram showing monthly precipitation from the ROCIO 5 × 5 km grid (Peral García et al., 2017) and temperature from the Spain02 20 × 20 km grid (Herrera et al., 2016), both calculated as the average value of the 16 grid cells closest to the coordinate 40.3° N, 1.4° W, for the period 1986–2015.”

14. Fig 2 caption. Specify that the “number of samples” is cores or trees.

    Ok. Changed to *“number of cores”.*

15. Fig 3. For consistency, give the grid resolution for all of the datasets in the labels along y axis.

    Ok. Done.



16. Fig 4 caption. The “red dashed line” is not the residual chronology, but the reconstruction based on it.

    Changed to:
    “Figure 4: Calibration of the pine residual chronology against CRU TS and ROCIO previous-year August to current-year June precipitation sums (blue curves) from 1958–2012. The red dashed line represents the chronology-based precipitation reconstruction. Grey shades represent ±1 RMSE.”

17. Fig 5 caption. There seems to be a wide range of day windows with high correlation, or with dark red shading. How much lower is the correlation for the "best" annual (365-day) period that the correlation for selected 320-day window (r=0.749)? Could this just be a result of sampling variability? Perhaps you can add a sentence or two on this in the discussion. In a related question I wondered whether the same 320 day window is identified if use different sub-periods (e.g., first and last halves) of the record.

    We have responded to this as a reply to major comments 2 and 3.