1. General Comments [an initial paragraph or section evaluating the overall quality of the preprint]

This work presents a novel an exciting application of the random forest (RF) reconstruction technique to the estimation of summer vapor pressure (VPD) deficit from an expanded network of European and Eurasian oxygen isotopic data obtained from tree rings. The goal is to place recent observations of temporal changes in summer VPD into a longer term context. I suggest revision of introduction and data and methods sections to provide important missing information on the training dataset, the RF algorithm and its modification and use here, why and how the results are produced, their uncertainty, how the validate against candidate semi-independent reconstructions. Reordering of some elements of Section 2 will improve clarity and logical sequential progression of the work. In the results and discussion section, I suggest revisions to clarify the features of the results that pass validation testing, clearly illustrate the primary features that are interpreted, and integrate discussion of uncertainty into figures and interpretation. In the conclusion, I suggest revisions to more closely align these points with revised results, and suggest that the abstract also be reconsidered once these revisions have been made.

We want to thank the reviewer for the appreciation/suggestions/comments/feedback that will help us improve our manuscript, and for taking the time to read and review our paper. We have reviewed all the comments and suggestions and provided a point-by-point response below. The reviewer's comments are shown in black and the replies are shown in red.

However, we would like to point out that some suggestions and required analyses are not constructive and are beyond the scope of our paper. We think that taking into account all reviewer's suggestions would multiply the length of the manuscript and the number of figures by about five, so we decided not to implement all of them. For the cases where we will not proceed with the suggestions/comments of the reviewer we will give a detailed answer for our decision. A large number of comments in this round of review aim in the direction of changing the content and purpose of our study (e.g., title suggestion that is not related with the initial one). We are aware of these suggestions but we decided to not apply changes, in the revised version of the manuscript, when we consider that they will considerably modify the initial content of the study that already went through several rounds of reviews reviewing process. Please see our point-by-point responses.

2. Specific Comments [individual scientific questions/issues ("specific comments")]

2.0. Title and Abstract

2.0.1. Define "summer" at first use, and clarify everywhere that VPD refers to summer VPD throughout the paper. Elsewhere clarify the season of other targeted and reconstructed variables.

Thank you for this suggestion. We will modify the manuscript accordingly.

2.0.2. Consider revising the title to be more specific to the key motivation "The response of evapotranspiration to anthropogenic warming is of critical importance for the water and carbon cycle." and associated conclusion: "Based on our reconstruction, we show that from the mid-1700s, a trend towards higher [summer] VPD occurred in Central Europe and the Mediterranean region which is related to a simultaneous increase in [summer] temperature and decrease in precipitation." However, also see notes on Data and Methods, Results, and Discussion below, as this statement is not clear from the results and analysis presented, in particular, Fig 6.

We will take into consideration the reviewer's suggestion, but we consider that the presented title is succinct and suggestive of the new summer VPD reconstruction presented here and the presented analyses of summer VPD variability.

2.0.3. For widespread further analysis and impact of the contribution, please also deposit reconstructions in a public long-term archive such as PANGAEA or the NCEI/Paleoclimatology repository, and provide the URL.

The reconstructed summer VPD gridded data, over the last 400 years was deposited on the ZENODO multi-disciplinary open repository and is freely available to everyone.

https://zenodo.org/record/7650420

If the paper will be accepted for publication, we will deposit the date in the aforementioned repositories.

2.0.4. 26 observational sites is a sparse network for the development of a gridded reconstruction over Europe. Add a sentence to the abstract stating the level and the nature of the independently validated skill of the results, specifically summarizing what is shown in Fig 4 and revisions to Figs 5-8. See also notes below on Data and Methods and Results to further refine the content of Figs 4-8.

We made use of the currently available isotope network at the European level, and the used sites represent the maximum possibility to extend the isotope network in Europe. We will add a sentence to the abstract according to the reviewer's suggestion.

2.0.5. With respect to the conclusion described in item 2.0.1, add a sentence to the abstract to explain why the authors think the summer VPD over the region is increasing since the mid-1700s. What changed after this time to cause the observed result? What is the forcing, and is this the expected sign and amplitude in the response? I am not sure I find the development of these results in the manuscript in its present state, but this would be a welcome expansion.

Thank you for this suggestion. In the revised version of our manuscript, we will improve the discussion regarding the increase of the VPD since the mid-1700s.

2.0.6. The results are based on o18 but a growing season VPD effect ought to also be observed in c13 (e.g. Farquhar et al, 1989, https://www.annualreviews.org/doi/abs/10.1146/annurev.pp.40.060189.002443 ; see also

Siegwolf et al 2022, previously cited). If the records exist, do they corroborate the result? If not, why not?

Unfortunately, we do not have access to the δ^{13}C European network.

2.1. Introduction

2.1.1. l. 53-55: "For instance, studies have shown that the [summer?] VPD has been increasing sharply at a global scale since the year 2000 (Simmons et al., 2010; Willett et al., 2014; Yuan et al., 2019). Spatially explicit [summer?] VPD records derived from remote sensing data cover only 55 the last ~50 years and vary in quality, so long-term perspectives of VPD variability are lacking." Revise to cite the remote-sensing and reanalysis-based results more specifically (e.g. Yuan et al 2019 is a remote-sensing based study) so the reader can easily learn more about each of these sources and corresponding results. Plot these records in Fig 6 for comparison with the reconstructions.

We clarified that we mention summer VPD in these statements. Yuan et al. plot observations (HadISDH and CRU) and reanalysis (ERA-Interim and MERRA). We do not understand what remote sensing data the reviewer wants us to use. In any case, we will plot 20CRv3 and ERA5 data in Fig. 6 to check whether 20CRv3 is fine, as it was already shown that the former ERA reanalysis dataset (ERA-Interim) found agreements with CRU observations (Yuan al. 2019). We believe the quality and comparisons between existing datasets are not the point of our study and have been done already, in Yuan et al. 2019 for instance.

2.1.2. l. 55-57: can summer VPD may be estimated from available historical reanalyses and other gridded products? Yes, as described in section 2.4. It would be good to revise this sentence to indicate the potential to do so, and possibly to evaluate the results from publicly available products. Plot the VPD estimate from the 20CRv3 in Fig 6 for comparison with the reconstruction. Do they agree or disagree with results from the more recent reanalyses and remotely-sensed estimates? This would help better establish the validity of the o18 training target.

We aim of this study is to reconstruct the European summer VPD **for the first time,** from a proxy network that is based on 26 series of tree-ring δ18O and covers the period 1600-1994. The reconstruction model was build using the reanalyzes data. The obtained results were compared with different former climate field reconstructions, and the results are presented in Figure 7 and the discussions are presented in section 3.4. Also see our above reply.

2.1.3. l. 62-63: This sentence needs support and development, perhaps its own paragraph: why reconstruct European summer regional VPD? For instance, is it because there is a significant modern regional summer VPD trend that is projected to continue and endanger the health of European forests? In what season is the trend observed? This would also support the argument that we need high resolution records (l. 62-63), at what resolution, and for what season. Consider adding a figure that shows this motivating problem, including the uncertainties in the different sources of information (remotely-sensed, modern reanalysis, historical reanalysis) – this will support the development of the present study in the present region at the reconstructed resolution.

Thank you for this suggestion. We will take into consideration the reviewer's suggestion and will improve the introduction section of the manuscript, but we are sorry to acknowledge what the reviewer asks here would constitute a study on its own that goes quite far from the scope of our work. We will not compute such further analyses that will require several and likely interesting interpretations and new writings, yet not related with our initial work. We will add a sentence to discuss this point.

2.1.4. l. 64-65: pursuant to item 2.1.3, tree-ring records generally record environmental conditions with a specific growing season and possibly antecedent seasons. Revise the sentence to add this point, as it is pertinent to item 2.1.3.

Thank you for this suggestion. We will take into consideration the reviewer's suggestion and will improve the introduction section.

2.1.5. l. 67-70: You might also cite the recent book on stable isotopes in tree rings, [Siegwolf et al, eds, 2022, https://link.springer.com/book/10.1007/978-3-030-92698-4] For this point, I think the same could be said of carbon isotopic composition, based o the same cited references. Explain the model for c13. Could c13 data be used to reconstruct growing season VPD? Why/why not? This helps make the case for use of the o18 dataset, and/or it might provide independent corroborating evidence in support of the results presented.

Thank you for this suggestion. We will cite the suggested book in the new version of the manuscript. In the current paper, we concentrated on reconstructing VPD based on δ18O, while the data δ13C will be targeted for a different paper, and because we did not have access to these data when we designed the study. It would be, however, interesting for future studies to reconstruction VPD fields from c13 and compare the output with our work for instance.

2.1.6. l. 73-76: Explain why variations in the oxygen isotopic composition of atmospheric water are not a confounding factor for summer VPD influences on o18 of tree rings in the European region. For example, a change in subtropical vs temperate or subpolar airmass might masquerade as a change in local evapotranspirative demand. See also item 2.1.5: c13 of tree rings might enable you to distinguish the latter from the former effect.

VPD is a measure of atmospheric dryness and is influenced by temperature and humidity. Tree rings, on the other hand, can record δ18O values that are influenced by a combination of factors, including temperature, humidity, precipitation source, and plant physiological responses. While variations in the oxygen isotopic composition of atmospheric water can influence δ18O values in tree rings, their impact is generally not a major confounding factor when studying the relationship between summer VPD and δ18O of tree rings in the European region due to several reasons, including the dominant source of water for tree growth during the growing season is local precipitation; variations in δ18O of atmospheric water can exhibit short-term fluctuations while the influence of short-term atmospheric variations on tree ring δ18O is relatively limited compared to the broader climate signals that are captured over multiple years; and/or while variations in atmospheric water isotopic composition could influence δ18O values in individual years, the overall tree ring δ18O signal over several years is more likely to reflect the prevailing regional climate conditions, which are driven by factors like temperature and humidity, therefore tree ring-δ18O can be used as a proxy for variations in VPD.

2.1.7. l. 78-83: Revise to precisely define the novelty of this contribution: expansion of the ISONET dataset (l. 111-115; Figs 1,2) and its use to reconstruct VPD using the random forest reconstruction algorithm. To clarify this, expand the sentence here to specifically describe the novelty of this work, and in Figs 1 and 2, indicate the new data series vs the ISONET series with differences in symbols, colors or otherwise as best clearly shown. Or you may choose to cite Table 1 which provides this information.

Thank you for this suggestion. We will modify the manuscript according to the reviewer's suggestion. We will indicate ISONET data. We already discussed the novelty of our work in the introduction.

2.1.8. l. 85-88: Expand the introduction of the RF reconstruction method [but check if this happens sufficiently in Section 2]. Is it particularly skillful for reconstruction of large-scale spatial patterns from sparse observational networks? If so, cite examples of this, as it will help you defend the technique and its application here.

Thank you for this suggestion. The reconstruction consists of several RF models applied for reconstructing each grid point time series, selecting only the relevant tree rings in terms of correlation and optimizing the RF parameter for each. This is thus not, *stricto sensu*, a large-scale pattern reconstruction. In terms of reconstruction skills for climate time series, the RF has shown high quality comparatively to generally used linear Methods (such as Principal Component Regression) (Delcroix et al. 2022, Michel et al. 2020). This will be explained in more details in the manuscript to better defend our approach.

2.1.9. l. 99-101: I think the sentence is saying that the regions of analysis are defined in the IPCC AR6 WG1 report. If this is correct, then please add the particular WG1 Chapter and ideally the figure in which these regions are defined within, and refine the cited reference accordingly.

The aim of this sentence is to indicate that our analyses are made subsequently from a spatiotemporal perspective for three different regions, regions which are defined and used in the Sixth Assessment Report (AR6) of the IPCC, and we cited the necessary references.

2.1.10. l. 104: Since there is validation for extreme year events, potentially add summary of these results and conclusions to the abstract. This could be an important result, speaking to the motivation for the proposed study, especially if it means that European forests might have survived past extreme VPD deficits, but that conditions are different now with the anthropogenic climate change effects superimposed? [check: do not get ahead of authors on this]

Thank you for this suggestion. We tried to keep the abstract short and precise. However, we will take into consideration this suggestion in the revised version of the manuscript.

2.2. Data and Methods

2.2.1. l. 127-132: The different sampling and pooling methods for developing sitewise values implies differences in the observational uncertainties assumed for the RF reconstruction. Also see items 2.1.3 and 2.1.4 regarding the season for which information is encoded in the tree rings. Do you take these differences into account or do you assume the same observational uncertainty in the various series? Are the results sensitive to these assumptions? Demonstrate if and how/why or how not/why not.

The sampling methodology for dendrochronological analyses is an important aspect and a standard methodology was developed and adopted to avoid the uncertainties between different studies (https://doi.org/10.1007/978-94-015-7879-0).

Numerous studies have analyzed the differences between pooling and non-pooling methods for analyzing stable isotopes in tree rings, and the main conclusion is that there is good agreement between different sampling and pooling methods, (doi:10.1016/j.scitotenv.2011.02.010, https://doi.org/10.1007/978-3-030-92698-4).

2.2.2. l. 132-133: I believe ISONET had a ring comparison of precision and accuracy of o18 analyses across the contributing labs [Boettger et al 2007, https://pubs.acs.org/doi/full/10.1021/ac0700023]; did the labs producing the expanded dataset also participate in this study? This might be important if there are differences in the amplitude of variation across labs for working standards across the range of observed o18 values. This is likely to be small compared to intersite differences in VPD amplitude of variation (e.g. Table 1) and observational uncertainty in the wood o18 records themselves, but it might be acknowledged that these interlab differences are assumed negligible.

Recently the ISONET database was published open access and all the details were included in the data set description (https://doi.org/10.5880/GFZ.4.3.2023.001). All the time series form the ISONET are considered homogeneous and can be used together for future analyses. We will indicate that inter-lab differences are assumed negligible

2.2.3. l. 150-154: Clarify and expand: For what purpose have you extracted gridded values from 20CR product? For what region? If for the summer season (which needs to be defined at first use of the term summer), then why are the other seasons of interest? (see also items 2.1.3, 2.1.4). Also please introduce the calculation of VPD from the reanalysis variables (see also items 2.1.3, 2.1.4). --> This item can be resolved by moving section 2.4 here (l. 158-159; section 2.4). Use the 20CRv3 ensemble statistics to estimate the uncertainty in the products and estimations based thereupon (see items 2.1.3 and the potential to add uncertainty into the RF training, item 2.2.6 below).

The propose of the 20CR product data are indicate in the main text line 158-159. We will explain why we use 20CR to make it clearer (because it expands far enough in time to build statistical models). However, we will not modify the whole structure of the manuscript.

2.2.4. l. 159-160: Clarify: Missing data are infilled for which dataset? Or both? if this sentence refers to missing data in the o18 dataset, then it belongs at the end of section 2.1. If in the 20CR data, it belongs in section 2.2. If both, then it might belong where it is, but clarify the infilling procedure is performed on both datasets, and the percentage missing data in each dataset separately.

Thank you for this suggestion. We are fine with this statement as well as its location in the text. We will clarify that these few and marginal gaps refer to the tree ring dataset.

2.2.5. l. 165-166: Revise: these restrictions might be moved to sections 2.1 and 2.2 and reference both Figs 1 and 2, as they arise from the tree-ring o18 network observational restrictions.

Thank you for this suggestion but as we describe the approach in section 2.3, we prefer to keep these details there.

2.2.6. l. 164-172: Revise and expand: Michel et al (2020), the algorithm adapted for the reconstruction, should be specifically cited at l. 95. And the description of the RF algorithm needs to be specifically introduced in section 1, and then the modifications more specifically described here in section 2.3. Generally: How does the RF algorithm work? What distinguishes it from linear approaches? What kind of data requirements or assumptions need to be made? How are reconstruction uncertainties estimated? Specifically, expanding from the modifications described for nesting, are the nests normalized against each other to counteract heterscedasticity associated with changing numbers of observatonal records, and their individual skills?

Thank you for this suggestion. We will give further details of our approach, notably following reviewer's recommendations that are indeed required (see reply to 2.1.8).

How is the 1x1 degree resolution of the reconstruction target established and validated for reconstruction by the sparse observational network of o18 data? How is validation established (are a fraction of data withheld for validation for realizations, and if so how and what fraction?) How are the reconstructions aggregated?

We will also provide these details.

2.2.7. l. 170-172: CE is a measure of resolved variance and mean estimate in an independent period; also consider reporting other skill statistics such as statistics of the ensemble reconstruction validation bias, correlation, root mean square error. Report also statistics for the calibration period to demonstrate the presence or absence of artificial skill (training statistics much better than validation statistics) in the results. [Check: is this done in the results in section 3?]

We already use the CE statistics that we believe it provides enough information. We can provide "CE" (RE in this case) for the calibration periods. However, RF will inevitably fit the training data almost perfectly (which does not inform whether it works well) as these data just finish-up in the node they have been put into during the training. This is very different from linear fits, where indeed some points from the training will be far from the regression line.

2.2.8. l. 173-177: Explain in the introduction (see also items 2.1.2-2.1.4) why these comparisons are useful, as they compare reconstructed VPD to reconstructed temperature, precipitation and PDSI. Why should these different climate variables covary with VPD? As the argument is made in Section 1 that the VPD reconstruction is novel, also explain, how and why does VPD differ from T, P, PDSI? Here also explicitly state if the reconstructions are fully independent from each other. I am guessing that the T, P, PDSI reconstructions do not use o18 data, but if they target and are trained on summer European observations of T, P, PDSI, and the 20CR on which the VPD reconstructions are also based assimilate observed T (but probably not observed P, calculated PDSI), then to what extent do the VPD and T (and possibly P, PDSI) reconstructions correlate because of these common training data (e.g. Fig 7a,b,c)

We already explained what VPD is in the introduction. We cannot know to what extent the correlation is due to the fact statistical models for reconstructions were trained on similar data. We will acknowledge that this may influence the correlations we have found. The paragraph will be improved.

2.2.9. l. 177-180: This is a useful exercise, but explain how the marker years were selected. Produce a similar exercise, but for the same selection criterion applied to the 20CR and modern VPD estimates. This may be useful as a validation of the stationarity of the spatial patterns in extremes, and discussion of the physical mechanisms consistent with the patterns, and/or a discussion of the reconstruction uncertainties and the significances of plotted diagnostics.

In the revised version of the manuscript, we will explain how the extreme years were selected. There is no 20CR data available for the selected years, and the aim of this analysis was to compare our reconstruction with historical data, while the validation with the observational data was presented in a separate section.

2.2.10. Move Section 2.4 into section 2.2, addressing items 2.2.3, and you can address items 2.1.1-2.1.3 with the results.

A note at this level of our revision: We have worked a long time on this paper and had to address several reviews already, as it was under review for another journal before. We also all are working on many other projects. We thank the reviewer for asking us to change most of the content and organization of the paper, but we will certainly not re-arrange everything as it is asked here, including content, analyses, order of sections, changing title, changing abstract. That is our study, and we are only willing to make analyses relevant with its content. We do not believe that we need to explore other topics that each could constitute a research paper on their own without overlapping our current results.

Also the production of the seasonal averages (l. 195-197) needs to be justified by the available data and skillful representation of T and RH for those seasons, and the seasonal representation of the reconstructed seasons by the tree-ring o18 observations. I was confused here by the mention of calculation of non JJA seasons, as if these were input to summer (JJA?) reconstruction target. Please clarify if the only season of 20CRv3 based VPD that is used in the present study is JJA.

Our study is not intended to evaluate RH and T representation in 20CRv3 that we consider to be a good dataset as it is published and maintained. Therefore, we use 20CRv3 for our analyses as other researchers do without being required to evaluate it at each use. We believe this could be an interesting work for other studies, but this is definitely not the topic of our one.

How is site information trained non-locally against the possibility that a site growing season and representation might be, for example, MAM and JJA, but this is used to reconstruct for a grid point for which the growing season response might be the same or different?

The evaluation of proxy data in representing local JJA VPD is already provided in Fig. 3.

Here it will again be useful to explain the RF methodology in more detail (items 2.2.6, 2.2.7) and to produce figures with diagnostics of the training vs validation ensembles. This could also be expanded in discussion: what do we learn from the RF fitting process about the information contained in the tree-ring o18 records? This is potentially another important contribution of the work.

We will never be able to produce all figures asked by the reviewer and it would constitute too far points from the initial scope of our study that we will not change.

2.3. Results.

2.3.1. l. 201-202: Revise associated results and figure 3: produce the correlation map for (a) statistics of the training ensemble and (b) statistics of the validation ensemble: r, bias, RMSE, CE, using the ensemble of the statistics to estimate statistical significance of the results and mask nonsignificant results. Otherwise this presentation conflates training and validation statistics by including the entire 1900-1994 period used for both training and validation (as yet unspecified how; see items 2.2.6, 2.2.7). If they are different, then we might have an artificially better view of the skill in the reconstruction, because by definition, the training results have to be skillful to some extent. If necessary, revise the associated results in subsequent sentences of section 3.1.

Thank you for this suggestion, but we will not make these extensions to figure 3 as we do not see why this would be crucial. Correlations provide enough information and other linear metrics will not bring much more information but only confusion and an overloaded manuscript with several barely relevant information. Of course, these correlations were computed on the same period as RF models' training and testing, it is too small to be shorten further. In a previous round of review, one of the reviewers suggested using this kind of map to show the correlation between every site and the local VPD values, which is a more conscientious way to identify if there are significant correlation.

2.3.2. l. 202-203; Fig 3. Explain how the correlations are produced. Are they for the *nearest* gridpoint of VPD estimated from the 20CR data? Superimpose the reconstruction grid. To the extent that this is not gridded VPD reconstruction, but rather a demonstration that there is correlation between o18 data and local (nearest gridpoint?) computed VPD, this figure could more usefully go into section 2. If expanded to produce a reasonably analogous training and validation period, like is done in the RF reconstruction, the figure could also be expanded to provide: r, bias, RMSE of standardized values, and CE of standardized values, for the validation period. Or these results could possibly be put into an expanded Table 1.

Thank you for this suggestion. We will clarify that correlations are calculated with the closest grid point. We do not see the point of these further analyses, figure 3 already provides enough information to the readers in terms of coherence between our d18O timeseries and local JJA VPD values (thus from the closest grid point). We think superimposing the reconstruction would at best be confusing at this level of the study (this is data description given before the methods explaining how the reconstruction is made); we will not reorganize the whole manuscript just for this non-crucial modification asked for Fig. 3, and we will keep rejecting such suggestions. The main aim of section 3.1 was to demonstrate that there is a strong relationship between δ18O in tree rings and VPD variability, and it is okay to use the δ18O to reconstruct VPD variability over the last 400 years. The validation statistics of the VPD reconstruction are presented in section 3.2.

2.3.3. l. 223-235: Important elements of the methods are given here, but would be valuably moved to section 2 and respond to items 2.2.6 and 2.2.7.

We will move these details to section 2 to respond to items 2.2.6 and 2.2.7.

2.3.4. Add site locations to Fig 4 to help demonstrate the skill local to the sites, and decline in skill away from them. I think the resolution of the target field (1x1 degree) is represented in the pixelation of the colors. Is CE normally distributed, or would it be better to plot the median estimates? Or the median estimate for which the 95th percentile confidence interval does not include the null hypothesis? I am still left without enough explanation of how training and validation is established away from the site-nearest gridpoints, and this should be addressed in Section 2 (items 2.2.6, 2.2.7). I realize much of this may be developed in Michel et al 2020, but for use here and its adaptation, methods and significance estimation should be briefly summarized in section 2 so the paper stands on its own.

We will add site's locations in Fig. 4. We now plot median CE as it is indeed not bounded for the negative values.

This is indeed the case that most of the methods part is explained in Michel et al. (2020). We agree that more details are required here for the paper to be more self-consistent. We will give these additional details.

2.3.5. l. 241-255: To fully understand the results for skillful nonlocal reconstruction (compare: Figs 1 and 2 to Fig 4), we need more information about the nature of nonlocal training and validation.

This will be done through our reply to 2.3.4.

This should be described in section 2 (see items 2.2.6 and 2.2.7). 2.3.6. l. 258-260: The use of +/- 2 standard errors implies that the errors are normally distributed; is this the case? Better might be the 95th percentile confidence interval, if not. See also item 2.3.4.

Errors are normally distributed based on Shapiro-Wilk tests, we will provide a supplementary figure to illustrate it.

For the error calculation, this should be done for the validation and training ensembles separately (see also item 2.3.1 and 2.3.2 notes on Fig 3 as well). The use of RMSE here also suggests support for revising Fig 3 in line with items 2.3.1 and 2.3.2. Revise Fig 5 to produce the same statistics and for the validation ensemble of the statistics. If the use of only validation statistics is already what is shown, at l. 201-202 and l. 258 and elsewhere the 1900-1994 interval is noted, please clarify this.

We will not revise Fig. 3. We also do not see the point of the reviewer more generally.

2.3.7. Fig 5: I am unclear as to how to interpret Fig 5 results. I would revise the figure to show the median RMSD for gridcells for which the 95th percentile CI does not include the null hypothesis, masking all other gridcells inthe domain. Put the o18 record locations into the figures to help make the case that the skill is highest nearest the o18 record locations (also Fig 4; l. 241-255; l. 256-262).

Site information is already present for Figs 1-4 we believe adding it a 5th time is redundant. Fig. 5 has to interpreted as higher values means highest statistical uncertainties.

2.3.8. Section 3.3 and l. 271, 275, 292, elsewhere: define beta and m in Section 2, methods: I believe in all instances, this is a linear regression slope coefficient, is this correct? But clarify if these estimates are regressions on 30 year rolling averages (could they instead be center-weighted 31-year averages, e.g. a Hamming window or similar?) Also give the errors on the slope, and plot the regression fits and the 95th percentile CI error envelopes on Fig 6 for their evalation by the reader.

Thank you for this suggestion but we do not think changing the filter is crucial and we are happy with the one we use. Beta and m will be defined at their first use in the main text. Explaining them in the Methods would just be confusing in our opinion, and we are pretty sure that studies mentioning such basic statistic do not crucially need to define them in their section methods. We do not believe the errors are crucial for readers as we already provide significances of slopes.

2.3.9. Fig 6, l. 303-305: Clarify: are these averages area-weighted, and are these averages with nonsignificant results masked out? From panels (a,b,c), it appears the answer is no, but I would recommend the authors indicate the subregions for which VPD is skillfully reconstructed, and see if the resulting timeseries are different for skill-masked and non-skill masked estimates. From comparison with Fig 4, I would recommend the former in revision. The same masking might need doing for the P, T and PDSI estimates given their reported validation skill scores.

This is a crucial point raised by the reviewer and we thank them for spotting it. Reconstructed data will be masked-out given CE scores as it should have been done initially.

2.3.10. Fig 6, error estimates on time series: based on Figs 1-4, unclear about 5, I am skeptical that the errors on the time series are accurately calculated; they seem too small, given results in Fig 3. Are the nonrandom errors of estimation from the nonlocal estimation used? After all there are only 29 data series and many more gridpoints in each region at 1x1 degree resolution. Are the errors of validation included? Do they vary with the number of o18 series available over time? They probably should, according to Fig 4.

They do but the effect is slight. Errors were computed from each of the intricated reconstructions (thus including training data as in Ortega et al. 2015, Wang et al. 2017, Michel et al. 2020, for instance).

Details of the error estimation need to be added in Section 2 (see also items 2.2.6, 2.2.7).

We will do so but in the results section as the Methods section describes the reconstruction approach and not statistics.

2.3.11. Fig 6: the presentation of timeseries for like regional averages of reconstructed VPD: T, P, PDSI: VPD here are reconstructed anomalies, relative to a historical mean? If not, explain how the mean was estimated (section 2), and give the reconstruction uncertainties for mean estimation. If for anomalies, give the period for which anomalies are relative and give details in Section 2.

Thank you for this suggestion. In the revised version of our manuscript, we will include more information about the average uncertainties from Figure 6. The section 2 will be improved.

2.3.12. Fig 6 and l. 293: put error estimates on the 1652 maximum VPD.

There is already an error bar provided.

2.3.13. l. 290-299: This presentation of results needs introduction in section 1. Are the dates relative to solar activity periods (e.g. Maunder Minimum) and temperature anomalies (e.g. LIA) defined by the reconstructions or by independent estimation? It should be the latter, and these definitions might be usefully described in the introduction, otherwise they present as potentially ad hoc.

Thank you for this suggestion. In the revised version of our manuscript, we will include an explanation of the terms and data used.

2.3.14. Following from item 2.3.10, it is unclear to me that the results support the interpretation at l. 292-299, given the decal variability and questions concerning the significance of linear trends on decadal timescales given reconstruction uncertainty.

There are uncertainties like for any paleoclimate reconstructions and we have provided them. We do not think that this should avoid us to calculate statistics from the reconstruction.

2.3.15. Fig 7 and Fig 6, Section 3.4. For Fig 7 and results presented in the test in section 3.4, are the correlations for specific seasons, or annual averages? Specify. The results in Fig 6 for 30y means suggest little of the skill is at these timescales. Calculate correlations for 31 year centered means, and estimate significances for the reduced degrees of freedom. Also, since the reconstructions are all trained on modern observations, which may be non-independent, it would be good to calculate all correlations for Fig 7 and on decadal timescales in Fig 6 for the pre-training periods, which I believe in all cases might be (please check): before 1850. This is a more conservative validation test and might resolve the apparent inconsistency between Figs 4 and 6 and Fig 7, which I don't understand. These questions might require some revision of the sentence at l. 340-344, and see item 2.3.16.

We do not see where our figures have apparent inconsistencies as it is not explained what these inconsistencies are. We will provide correlations for the pre-training period.

2.3.16. Fig 7, Section 3.4: VPD is in general negative correlated with P and PDSI, and positively correlated with T2M. Explain why, here or in section 4, with reference to section 1 and item 2.2.8, and the physical links between temperature, precipitation, vapor pressure deficit and the Palmer Drought severity index, which includes both temperature and moisture influences.

Thank you for this suggestion. We will include the explanations in the revised version of the manuscript.

2.3.17. Section 3.5, comparison of Figs 6-7: I am having trouble making sense of this section, because the physical links between the variables reconstructed have not been introduced (item 2.3.16), and I am not sure if the changes in the correlation with time and frequency, on different timescales, do not arise from reconstruction non-independencies and uncertainties that vary in space and time (observational density), and frequency (skill?). I am not sure how to suggest specific constructive revisions, except to suggest that the organization by subregion be retained; presentation of the physical linkages (item 2.3.16) be introduced and explained, checking and quantifying uncertainties (items 2.3.15, 2.3.1-2.3.12 and cross-referenced earlier items) be performed, before revisiting and synthesizing the most reliably estimated features and independent comparisons with associated but not-the-same reconstructed variables (Figs 6-7), organized by region as is done here.

Thank you for this suggestion. We will include the explanations in the revised version of the manuscript, and we will improve the section 3.5.

2.3.18. Section 3.6, Fig 8: This section needs more detail on what was done. What does "at the European level" mean, is this an area-weighted average over skillfully reconstructed gridpoints, for what season, for the pre-training interval, and considering the changing uncertainty over time? These details could go into Section 2, such that the discussion of the results can be done here. I think wettest and driest refer to VPD here, but perhaps revise to call them the three lowest and three highest reconstructed VPD years, estimated as the average standardized anomalies over all seasons and all skillfully validated gridcells. Because in panels (a-c) the sign changes across regions, skill masking might be important, and the definition of the extreme years is also important. If the results do not change with these considerations, what is striking to me is the Scandinavian/western European VPD dipole (panels b,c; but not a); and the consistency in the negative VPD anomalies across all Europe in panels (d,e,f). The authors wrote about these features, but maybe could expand the discussion: are there different mechanisms causing these two different results? If so, why? Perhaps this could be an organizing question for the discussion of these results.

Obviously, at the European level means over Europe. We will modify this unclear wording and give more details. We will apply a skill masking here as the reviewer suggests which is indeed relevant. We will include more detail in the revised version of the manuscript and more explanation will be provided. We will improve section 3.6.

2.3.19. Section 3,6, l. 401-406, discussion of the potential for the Laki eruption to have caused the 1785 extreme events, this seems a consistent discussion (but I think should refer to Fig 6, not Fig 5), but it could more comprehensively address the volcanic forcing of VPD, T, P and drought, by showing a composite of the VPD, T, P and drought anomaly responses in the years subsequent to all the strong volcanic events within the time interval of reconstruction.

Thank you for this suggestion. We will correct the reference to the figure. We refer to Laki eruption as we discuss the fact 1785 is an extreme year, which is a rather normal discussion. If we you knew this would lead the reviewer asking us to look at all eruptions for all datasets (nothing to do with our study) we would have not try to explain the result we found in that way. If someone is interested in looking at the VPD (and other variables) response to volcanic eruptions, they are very welcome to use our reconstructions and T, P, PDSI ones to do so. The aim of this section was to discuss the occurrence of the extreme years in the VPD reconstruction in comparison with historical data.

2.4. Limitations of the reconstruction

2.4.1. In general, revising with respect to items 2.2.2, 2.2.3, 2.2.4, 2.2.6, 2.2.7, 2.2.8, 2.2.10, 2.3.1, 2.3.2, 2.3.4-2.3.10, 2.3.15, 2.3.16 will help support and expand the points here with specifics in figures than can be referenced.

Thank you for this suggestion. We will take into consideration the reviewer's suggestion and the point-by-point responses are provided for all mentioned items. See our replies to these points.

2.4.2. Alternatively, to item 2.4.1: the points in Section 2.4 might usefully be made in the figures by the suggested revisions, and used to contextualize the discussion and interpretation of the results, as suggested in earlier items.

Thank you for this suggestion. See our replies to points raised in item 2.4.1.

2.4.3. Section 4 could be replaced by a Discussion section. This could comprise the central argument of the paper that supports the statements made in the Conclusion and the Abstract.

The aim of section 4 is to indicate the limitation of our reconstruction, while the discussions of the results are presented in section 3.

2.5. Conclusion and outlook

2.5.1. l. 444-451: Revisit these statements after assessment of the timeseries with masking for regions of nonsignificant skill, and the assessment of coherence (correlation as a function of the key interpreted periodicities, specifically: seasonal, interannual, and approximately multidecadal.

We will mask regions with non-significant skills. We cannot assess seasonal variability as we focus on JJA VPD.

2.5.2. l. 450-452: "Central Europe and the Mediterranean regions reveal stronger trends of increasing VPD (highest VPD on average and highest VPD variability) which can be explained by a precipitation decrease and a temperature increase." It is unclear to me how this summary statement is supported by the results presented in Figs 4-7, and associated discussion. What is meant by "trends"? The anticorrelation and correlation (or nonlinear associations) between VPD, P, T (and PDSI), and changes in variability, are not apparent from the analysis presented in Fig 6 and Fig 7.

The explanation of the relation between VPD and temperature, precipitation, and drought index will be added in the main text of the revised version of the manuscript, and more explanation regarding the trend and their meaning will be also provided.

2.5.3. l. 452-457: "Our results underline that the European VPD values have increased over the last decades. Based on the obtained long-term perspective, we find that this increasing trend in Europe has not started in 2000, but has already begun a few decades after the Late Maunder Minimum with a simultaneous increase of the temperature in the mid-18th century. In the historical context, however, this long-term trend is unique in magnitude and persistence over the last 400 years. Moreover, the results from our study imply that vegetation in Europe has been subject to an increase in VPD for a longer period of time, but that increase has been significantly amplified by recent climate change, especially in Central Europe and the Mediterranean regions." It is not clear to me that these statements are supported by the results and analysis presented. Clarify and illustrate the results in support, in Figs 4-7, addressing prior questions concerning estimations and uncertainties, or else modify the statements to be more specific and accurate. For example, it is not clear that European VPD values have increased everywhere - not in northern Europe where there is considerable skill in part of the region (Fig 4, Fig 6). And in no region is the recent VPD higher than in previous decades and centuries (Fig 6).

In the revised version of our manuscript, We will revise these statements after the application of the unskilled mask.

2.5.4. l. 457-461, I think the first half of the sentence is supported by the logic of the cited references, but the second half depends on the effect of an increase in VPD on vegetation in Europe since the Late Maunder Minimum, and this is unclear from the results presented in Fig 6.

We will rephrase the sentence in the revised version of the manuscript.

2.5.5. following l. 461, this would be a good place to summarize the expanded results (see items 3.3.18, 2.3.19) from the analysis of extrema in the reconstructions and the mechanisms that might have caused them.

Thank you for this suggestion. We will modify the sentence according to reviewers' suggestion.

2.5.6. l. 462-467: I think this summary of the validation makes the case for interpretation of the results, so it belongs at the beginning of the summary, not at the end. It also needs to be expanded to consider the results of further uncertainty quantification as described in the preceding notes.

Thank you for this suggestion. We will improve the conclusion section in the revised manuscript. We will move this part as suggested by the reviewer.

2.5.7. There is no summary or discussion of the RF reconstruction and diagnostics, but there should be, as this is an important new application of a nonlinear reconstruction technique.

Thank you for this suggestion. We will improve the conclusion section in the revised version of the manuscript.

2.5.8. l. 465: Here it seems that summer VPD is what is constructed - this should be clarified in response to earlier items (e.g. 2.1.3, 2.1.4, 2.2.10).

Thank you for this suggestion. We will clarify this issue in the revised version of the manuscript.

2.5.9. l. 467-469: the logic of this sentence is unclear. How does the gridded reconstruction of European summer VPD minimize statistical uncertainties of historical VPD variability? How does it provide long-term context? (But in Fig 6, if the 20CRv3 VPD and the other estimates of VPD for these regions, it could do so.)

Thank you for this suggestion. We will improve the conclusion section in the revised version of the manuscript.

3. Technical corrections: [compact listing of purely technical corrections at the very end (typing errors, etc.).]

3.1. l. 62: "Therefore, appears the necessity..." --> "Therefore, it appears necessary to ..." and see also item 2.1.3.

We will modify the text accordingly.

3.2. l. 71 and throughout: "tree ring-d18O" --> tree-ring d18O"

We will modify the text accordingly.

3.3. l. 80: "opportunity to develop for the first VPD reconstruction at the European level..." --> "opportunity to develop the first VPD reconstruction at the European level..."

We will modify the text accordingly.

3.4. Section title might be more generalized: "2. Sample sites and used climate data" --> "2. Data and Methods".

Thank you for this suggestion. We will modify the text accordingly.

3.5. l. 403, "favorizing" --> "favoring"

We will modify the text accordingly.

3.6. l. 465: "f" --> "of"

We will modify the text accordingly.

3.7. l. 466: "d18" --> "d18O"

We will modify the text accordingly.