General comments:

The author has presented a revised version of the manuscript with substantially improved readability and many inaccuracies being cleared out with respect to the original version. It would like to point out again, that probably the lion's share of the work has gone into the code of the R-package and the curation of the database and that this work is of highest value for the community.

On the other hand, I must admit that I am still not convinced by the presentation of the mathematics forming the bases of the climate reconstruction method. This criticism comprises twofold: First, the introduction of Equations (1-5) and (7) suffers from several mathematical inaccuracies – however, I noticed that these in parts trace back to the article presented by Kühl et al. (2002). Since the derivation of these equations is by no means the central aim of this paper, these inaccuracies might be acceptable, given that these equations appear to be correct. Also, it seems that some terms have established as domain specific language and thus may be clearer to the target audience than they are to me. Second, I believe that Equation (6) is in fact not correct, even though its use might generate reasonable results. I have expressed my concerns about Eq. (6) already in my previous review.

If not for this manuscript, I am convinced that putting the probabilistic climate reconstruction method on solid ground mathematically would be a beneficial task for the future. I have attached a pdf that outlines the derivation of Eq.(5) starting from Bayes theorem.

The remainder of the mansucript gives the reader a good overview of the *crest* R package, in terms of its capabilities, requirements and usage. It is well structured and the final example of application really takes the reader / user by the hand.

Specific comments:

l.2 In particular, the methods based on probability density functions (or PDFs) can be used in various environments and with different climate proxies because they rely on elementary calibration data (i.e. modern geolocalised presence data).

I would replace 'the methods' by 'methods'. Maybe 'methods based on probability density functions' are just 'probabilistic methods'.

l.14 It is hoped that crestr will be used to produce the much-needed quantified records from the many regions where climate reconstructions are currently lacking, despite the availability of suitable fossil records.

What is meant 'quantified records'? In my understanding a record is a directly measured time series – so a data processing software could not be used to 'produce a record'? Do you mean 'reconstruction'?

l.15 no paragraph in the abstract

l.19 Over the years, numerous techniques of increasing complexity have been proposed, each one based on a unique set of assumptions regarding the modelling of ecological datasets and their translation into climate reconstructions (e.g. Birks et al. (2010), Chevalier et al. (2020b)).

I assume that with ‘ecological datasets’ your refer to observations. In that case I would say that observations are not being modeled. Of course, sometimes one uses models to draw inference from datasets and probably that is what you mean?

l.26 ...and their accessibility with multiple software solutions.

Saying that an analysis technique is ‘accessible with multiple software solutions’ sounds strange to me. Do you mean, there exist relatively simple software implementations of the techniques?

l.26 However, the limited availability of the necessary calibration datasets beyond the Northern Hemisphere extratropics has often hindered their application in many environments and regions where quantified climate records are needed, despite the existence of suitable fossil records (Chevalier et al., 2020b).

Again, what are quantified climate records? It seems you mean reconstruction – in my understanding records are not reconstructions.

l.32 Because modern occurrence data are generally easier to obtain than modern proxy assemblages, this fundamental difference implies that Indicator species methods can contribute to filling in the reconstruction gaps that exist at the global scale.

Grammar: Modern occurrence data are generally easier to obtain than modern proxy assemblages. This fundamental difference implies that Indicator species methods can contribute to filling in the reconstruction gaps that exist at the global scale.
For the non-paleoecologists: what’s the difference between proxy assemblages and proxy occurrence data?

l.36 Derived from the original work of Kühl et al. (2002) —– who

It seems there is an extra hyphen in the pdf.

l.37 CREST estimates and combines probabilistic proxy-climate relationships to reconstruct past climate parameters from fossil proxy observations.

Maybe you could add: CREST estimates and combines probabilistic proxy-climate relationships **from modern occurrence data** to reconstruct past climate **variables** from fossil proxy observations.

l.45 However, the complexity of collating and formatting the thousands of distinct occurrences required to estimate reliable PDFs limited its practical use.

I understand, that in your context the term PDF carries a very specific meaning. However, in general, this is not the case and a reader that is not used to the specific use of the term PDF as a synonym to your ‘climate response functions’ will probably struggle to understand the above statement. Also, you introduced the abbreviation PDF only in the abstract, but up to this point not in the main text.

l.48 ‘climate records’

l.50 This paper thus introduces ~~a~~ the new multi-platform R package crestr designed to replace the original interface.

l.51 crestr includes the global calibration dataset

It is linked to the dataset but does not include it – strictly spoken.

l.59 As such, the climate reconstructions obtained from CREST can be understood as an ensemble of all data-compatible climate values

Maybe ‘As such, the application of CREST yields a probabilistic quantification of the past climate in view of the data under study as opposed to simpler, less informative ‘most likely’ or ‘best’ climate estimates. While the latter may only capture statistical uncertainties, the former rigorously takes into account the large quantitative uncertainties inherent to analysis of this kind.’

Just to highlight a little bit the fact, that in terms of uncertainty propagation *crest* performs a lot better than the mentioned ‘best-estimate’ methods.

fig.1 Conceptual illustration of the differences between a modelling approach based on the estimation of the full spread of the data with the probabilities spread along the climate gradient (e.g. CREST; dark grey), and a modelling approach focused on the estimation of the ’most likely’ or ’best’ climate value with small statistical errors surrounding it (e.g. MAT or WA-PLS; light grey).

I would say the probabilistic approach is based on the ‘full spread of the data’ but not on the ‘estimation of the full spread of the data’.

l.96 The **individual?** climate responses of all the species identified are estimated as univariate probability density functions (PDFs) for every climate variable.

l.99 The individual species’

l.99 estimation of the empirical mean (m s,c)

An empirical mean is not estimated, but computed from the data. Just delete ‘the estimation of’ .

l.107 Here, the weights are calculated by first sorting the N climate values (all the c i) that compose the modern climate space into bins of equal width (e.g. 2 ◦ C or 50 mm). Then, each climate value c i is given a weight k(c i) defined as the inverse of the relative size of the bin c i it belongs to:

1. As far as I understand N is the number of gridded observations of the modern climate variable c, while N_s is the number occurrences of the species S. According to line 100, the c_i are only those climate values, that coincide with the occurrence of the species. However, for the weights k(c_i), the entire climate space should be taken into account. So the addition (all the

c_i) is not correct. Maybe, the difference between the c_i and **all** climate values form the study area can be clarified somehow.

2. The first sentence defines all bins such that they have equal width. The second sentence refers to different ‘sizes’ of the bins. I assume that the ‘size’ of a bin here means the number of observations falling into one bin. However, typically one would use ‘size’ and ‘width’ interchangeably, so ‘size’ might be misleading, here.

l.117 Once estimated, m s,c and s 2 s,c are used to define a regular, unimodal distribution for the PDF sp (s, c) of species s for climate variable c. Here, we assume that the shape of these species responses should be unimodal and can be either normal:

I know it is a detail, but the term ‘distribution’ is slightly misused in this context. The distribution of a random variable in statistics defines the the probability for the random variable to assume a certain value. Hence, here the PDF **is** a distribution.

Maybe: ‘Assuming unimodality and either normality or log-normality, the estimated m_(s,c) and s^2_(s,c) are used to define the species climate response PDF’s as follows: …’

l.120 delete paragraph

l.125 I have expressed my doubts about this equation already in my previous review. After going through the math once again, I still belief this equation is not correct – even though it might lead to reasonable results. Also in the cited literature, I could not find a convenient derivation of Eq.(6).

l.126 delete paragraph

l.138 Climate c is reconstructed from fossil sample z (z can be an age, depth or any identifyer) by multiplying the PDF tx (t, c) of the T (z) selected taxa:

- Maybe ‘Past climate c that corresponds to a specific age (or depth) z from which a fossil sample is available can be reconstructed…’
- you first assign z := the fossil sample and then state that either z:= age, or z:= depth.
  - what exactly is ‘the fossil sample’? A dataset comprised of abundance data of different taxa on a depth or age axis? This is a little inaccurate.

l.141 delete paragraph

l.151 The summation index should be ‘i’ or sth else and then the depths / ages should be discretized as z_i, but the summation index cannot be a continuous variable. The inner parenthesis in the denominator are not required.

l.152 delete paragraph

l.181 In the gbif4crest database, all the QDGC grid cells were associated with a collection of terrestrial and oceanic environmental variables that can be reconstructed (Fick and Hijmans (2017), Zomer et al. (2008), Locarnini et al. (2019), Zweng et al. (2018), Garcia et al. (2019a), Garcia et al. (2019b), Reynolds et al. (2007), see details in Tables 1 and 2).

Does this mean, that all variables listed in Tab.1 and Tab.2 are reconstructable? If so, I would suggest to express this a bit more clearly and also emphasize this in the table captions.

l.193 For example, the first version of the gbif4crest dataset released in 2018 contained about 17.5 million QDGC entries, while the new version contains approximately 25.3 million entries (~44% increase).

Why QDGC entries and not only entries? The second is a duplication of the first sentence of the paragraph and can be deleted, except the (~44% increase).

l.230 Maybe, it would be helpful to explain, which function call initializes the crestObj in first place already at this stage?

l.257 Does the df have as many columns (+1 for the depth/ age) as fossil taxa are considered for the climate reconstruction?

l.280 In a second time

In a second step

l.281 Mabe here, a sentence like

The pdf for the fossil taxon Stoebe-type is thus comprised of the linear combination of species pdfs according to equation (6) associated with all species that fall into the geni Stoebe and Elytropappus.

would be helpful at this stage, provided that my interpretation is correct.

l.284 Additional taxa can also be added to the PSE file to exclude species known not to be part of a group. For instance, this ‘trick’ could have been used to simplify the climate response of the ‘Asteraceae undiff.’ group by excluding more species from it, even if the pollen grains corresponding to these species have not been observed.

I would suggest to move the ‘even if… ‘ to the first sentence.

l.290 If I understand correctly, if I provide a ‘distributions’ table as an input to the get_modern_data() function, then I do not need the PSE is that correct? Why is the ‘distributions’ table not listed in Fig. 5 in the ‘input’ category? If users decide to use the ‘distributions’ table as input, then they have to provide climate_space data frame as well, is that correct?

l.392 here called rcnstrctn; Fig. 5

here called rcnstrctn and whose structure is displayed in Fig. 5

l.393 Alternatively, the function crest.set_modern_data() could be called instead of crest.get_modern_data() to use personal calibration data instead of the gbif4crest database.

This could already be mentioned in 4.3.3 and 4.3.4.

l.449 Ideally, the climate sampling should be as homogeneous as possible to ensure proper sampling of all the possible climate values, even if the extreme climate values will always be under-represented compared to the median ones. However, deviations from a theoretical one-to-one (or at least proportional) equivalence between climate and occurrence data abundance are not necessarily a bad characteristic.

I know this has not changed much with respect to the previous version of the manuscript, yet I must admit I do not fully understand these sentences.

What exactly is 'the climate sampling'? If c is the climate variable to be reconstructed, I assume, the term refers to the entirety of values for this variable present in the study area, irrespective of the presence or absence of species used for the reconstruction. This relates to my comment with respect to line 107 – a clear distinction between 'the climate space' that comprises all N climate values and the 'the climate sampling' which is comprised only of the N_s climate values accompanied by species occurences (?) would be helpful.

If that interpretation holds true, does the climate sampling contain multiple instances of a climate value c_i if the corresponding grid cell contains multiple independent occurrences of the species?

What is meant by a 'one-to-one equivalence between climate and occurrence data abundance'? Would that be something like, the warmer the clime, the more occurrence data there is in the study region?

l.451 In our case study, the spatial variability represents actual patterns in regional species diversity with the presence of several biodiversity hotspots across the mountainous regions of eastern and southern Africa.

Spatial variability of what?

l.457 (i.e. variables correlated with important variables but do not directly impact the studied proxies; Juggins (2013), Chevalier et al. (2020b))

i.e. variables which are correlated …

Fig.6 number of unique species occurrences

I understand that the author does not want to change the title of the right subplot, however, I would appreciated a lot a clarifying not in the caption, that unambiguously defines the term 'number of unique species occurrences'.

l.508 Here for instance, both Aizoaceae and Chenopodiaceae/Amaranthaceae were excluded because they are not primarily sensitive to temperature in southern Africa.

(from my previous review) To understand this, it would be nice to have them included in the violin plot (Fig.8).

Answer by the author: This exclusion is based on ecological considerations of the taxa, i.e. what is known about them. It is independent of the shape or look of the associated pdf. I think it might in fact be more misleading to plot them because they might not appear widely different from the others.

This does not make sense to me. If these two taxa are not sensitive to temperature, this should be reflected in the data and this be visible in their climate response (or PDF), which in the violin plot should appear more stretched along the y-axis. In fact, the violin plot is advertised a tool to assess the climate sensitivity of the different taxa – the authors answer is not consistent with this role of the violin plot.

Fig.11 Couldn't the strong Loo values for the Ericaceae already be interpreted as some sort of bias in the sense of what you say in line 583?

Large LOO values can arise when the PDFs are biased by unaccounted factors and are, as a result, at odds with the rest of the PDFs.

Outline for a derivation of Equations (1-5) in a Baysian framework

In my previous review, I raised concerns about the mathematical accuracy in the derivation of the equations which constitute the basis for the presented method. I understand, that these equations have already been introduced in a similar way by Kühl et al. (2002) and repeatedly presented by the author himself. Yet, I would still argue, that this article and therewith the credibility of the R package would largely benefit from putting equations (1-7) on more solid ground from a mathematical perspective. While equations (1-5) at least seem to be correct, I have expressed my concerns about equation (6) in my previous review and did not find it to be adressed explicitly in the authors replies. The author's key argument to not carry out a derivation of these equations along the lines I present in my previous review, was that these considerations would require 'absence data' of the species S. The author is correct, that the considerations I presented in my last review relied on the assumption that for any grid cell 'no presence' would mean 'absence' and hence, this is not the way to go. Yet, it is still possible to thoroughly derive the equations (1-5) starting from Bayes theorem without using 'absence data'. I have made another attempt to provide the author with some thoughts on this below.

The author aim to derive a probabilistic climate reconstruction based on palaeoecological data, that is the presence of a given species (here for simplicity I ignore anything beyond species) at a given time. After careful reevaluation of the manuscript and the paper by Kühl et al. (2002), I came to the conclusion, that the probabilstic climate reconstruction as presented here, is in fact still based on Bayes theorem:

$$\underbrace{\rho_{C|s=1}(c|s=1)}_{\text{posterior probability distribution}} = \frac{\overbrace{\rho_{S|c}(S=1|c)}^{\text{likelihood function}}\ \rho_C(c)}{\rho_S(S=1)} \tag{1}$$

where the posterior distribution $\rho_{C|s=1}(c|s=1)$ expresses the probability (or one should say plausibility) for the climate variable C to assume the value c given the presence ($s=1$) of the species S at some point in the past (one could add dependency on t in all probability densities, this is omitted here for sake of readibility). It is reasonable to assume, that the likelihood function in the above equation did not change substantially over time, while the prior distribution of the climate $\rho_C(c)$ definitely. Assuming that we have little (or no) prior knowledge about the past distribution of the climate variable it seems a reasonable conservative approach to base the assessment of past posterior distributions of the climate variable only on the likelihood function.

The authors term the likelihood function $\rho_{S|c}(s=1|c)$ — interpreted as a function of c and evaluated at the value $s=1$ — *climate response of a given species* ($\mathrm{PDF}_{\mathrm{sp}}(c,s)$). This function is indepent from the distribution of the climate variable C and after convenient normalization it can also be interpreted as a probability density function with respect to c.

The authors estimate the likelihood function as follows: First, they consider the set of tuples $\{(s_i,c_i) : s_i = 1\}_{N_s}$ as a sample with N_s members from the modern distribution

of the climate variable c conditioned on the species presence $\rho^*_{C|s}(c|s=1)$. Assuming that this sample is representative one can approximate the distribution as

$$\rho^*_{C|s}(c|s=1) \simeq \frac{1}{N_s}\sum_{i=1}^{N_s}\delta(c-c_i). \tag{2}$$

Next, application of Bayes theorem relates the *climate response function* or PDF_s with the modern conditioned climate distribution that generated the observed sample:

$$\frac{1}{A}\rho_{S|c}(s=1|c) = \frac{\rho^*_{C|s}(c|s=1)\rho_S(s=1)}{\rho^*_C(c)}, \tag{3}$$

where $\rho^*_C(c)$ is the modern distribution of the climate variable C. Computing the mean and the standard deviation from the right hand side allows to approximate the left hand as a normal or log-normal distribution. Eq.(5) in the manuscript can be derived as follows:

$$\begin{aligned}
\mathrm{E}(C)|_{S=1} &= \int \frac{c}{A}\rho^*_{S|c}(s=1|c)\mathrm{d}c \\
&= \int c\frac{\rho^*_{C|s}(c|s=1)\rho_S(s=1)}{\rho^*_C(c)}\mathrm{d}c. \\
&= \int c\frac{\sum_{i=1}^{N_s}\delta(c-c_i)\rho_S(s=1)}{\rho^*_C(c)}\mathrm{d}c \\
&= \sum_{i=1}^{N_s}\frac{c_i\rho_S(s=1)}{\rho^*_C(c_i)} \\
&\simeq \sum_{i=1}^{N_s}\frac{c_i\rho_S(s=1)}{\rho^*_C(c_i)} \\
&\simeq \frac{1}{\sum_i k(c_i)}\sum_{i=1}^{N_s}k(c_i)c_i.
\end{aligned} \tag{4}$$

In the last step, the distribution $\rho^*_C(c)$ was approximated by a coarse grained or local density $k(c)^{-1} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}_{c_j\in\mathrm{bin}(c)}$. Here, N is the number of **all** observations of the climate variable C and not only of the those c_i that coincide with the presence of the species S as indicated by the authors in line 107.

Several concluding remarks

- *It seems that Kühl et al. (2002) make a mistake when they claim they would derive pdf(tmp—taxon). I am fairly certain, that what they do derive is pdf(taxon—tmp), as a function of tmp though.*

- *Furthermore, please note that for considerations presented here, no 'absence data' for the speciees is required. This is because the values c_i can be considered as a representative sample from the distribution of C, conditioned on the presence of the species.*

- *Currently, the crest package reconstructs the past climate only based on the likelihood function - this assessment could easily be supplemented by prior information on past climate, for example from climate modelling studies carried out with EMICS.*