

Clim. Past Discuss., referee comment RC2
<https://doi.org/10.5194/cp-2021-153-RC2>, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on cp-2021-153

Patrick Bartlein (Referee)

Referee comment on "*crestr*: an R package to perform probabilistic climate reconstructions from palaeoecological datasets" by Manuel Chevalier, Clim. Past Discuss., <https://doi.org/10.5194/cp-2021-153-RC2>, 2022

General comments:

This paper introduces an R package, *crestr*, for implementing a probability-density-function (PDF) variant of the Mutual Climatic Range (MCR) approach for making climate reconstructions from, for example, fossil-pollen data. The paper is not simply a user manual or vignette, but also discusses some of the conceptual underpinnings of the whole approach, and the philosophy behind some of the methodological choices that necessarily have to be made.

There are a few reorganization or further-explanation issues that need to be sorted out. For example, the motivation for producing the R package in the first place doesn't appear until Section 5 of the paper, and the particular reconstruction approach that is implemented here should be mentioned first among the list of various approaches mentioned in the introductory paragraph. Otherwise, the basic approach, and its implementation using the package is laid out nicely.

One important contribution of the paper is the release of the global data sets, both taxonomic and climatic, that can be used to apply the approach generally. However, there is a tension between attempting to reconstruct as many environmental variables as one can get into a database and reconstructing only those that can be mechanistically related to, say, terrestrial vegetation, which typically are simply growing season warmth, winter cold, and moisture stress. Those that argue for the former approach argue that the environmental variables are all related one way or another, so if you can reconstruct one, you can reconstruct all, while those that argue for the latter approach point out that assumption is nonsense. Likewise, there are probably taxa included in the data base that are completely insensitive to the macroclimatic variables provided, but may wind up contributing to the reconstructions when they provide little real information. From a purely statistical perspective, overfitting is the issue here. I know users don't necessarily have to attempt to reconstruct all of the environmental variables in Tables 1 or 2, and that they can use their own data, and manage the particular variables or taxa that are used, but I think that providing so many variables creates an "attractive nuisance". So, I think it

would be good to caution the users about these issues. No good deed goes unpunished.

Another issue that might be discussed a little is the “no analogue” one. Although usually raised in the context of modern analogue technique (MAT) approaches, it applies here too, as illustrated by Fig. 10, where the reconstruction lies in sort of trough of individual PDFs, and it’s also probably the case that some PDFs don’t overlap at all. If I understand this correctly, all of taxa with PDFs that appear in the figure co-occurred in the sample, but they don’t today (otherwise their PDFs would overlap). This deserves a sentence or two of discussion, perhaps by handing it off to other papers.

One stylistic thing about the paper is the sometimes jarring transitions between text and code blocks. Starting out, there are transitions like “...similar results would be obtained using the following command: (followed by the code block)”, but that format gets abandoned later in the manuscript. I know from experience that Copernicus journals’ choice of a type face for code makes it appear pretty clunky, and sometimes unreadable, which makes setting it off more important for readability in the two-column paper format.

Specific and technical comments:

line 2: “the methods ... are powerful at producing robust results...” I’m not sure that robustness in the usual statistical sense is either evaluated or demonstrated in this paper.

line 3: Not parallel: accessing/curating/the complexity” (action/action/characteristic). The sentence could be made parallel by rewording: “The problem of accessing and curating the necessary calibration data and the complexity of interpretation...”

line 17: “climate drivers” Meaning the climatic controls of the variations in the fossil data, or the controls of the climatic variations themselves?

Line 19: “climate reconstructions”?

line 23: “robust” Again, I don’t think this is the right word. I think that “robustness” is a property of a statistic (e.g. the median) that signals that it will perform well at estimating, in this case, location, no matter what the underlying distribution of the data looks like. I don’t think this notion really applies to a dataset. Maybe “extensive datasets”?

line 25: Before going further, it would be good to alert the reader as to which particular approach this paper implements (i.e. as in Section 5.1.2 of Chevalier et al. 2020), and a

little about how it works. (It seems it's NOT WA, WA-PLS, MAT, etc., but what is it?). I don't think it would be inappropriate at all (in terms of self-citation padding) to use Chevalier et al. 2020 a little more for background.

line 26: By "un-quantified fossil pollen records" do you mean that the data are "qualitative" as opposed to "quantitative" or simply that quantitative reconstructions have yet to be made? Same issue on line 35.

line 27: "built upon the"

line 28: "modern proxy geolocalised occurrence data". I don't know what that means.

line 39: "estimates, weights and propagates..." I'm confused by the "propagates" idea.

line 40: "posterior climate reconstructions" I'm not exactly sure what's Bayesian about this.

line 41: "weighted ensemble" An ensemble in this context is a collection or a group of something. Do you mean a "weighted average of an ensemble of estimates" or "an ensemble average"?

line 46: "can model" Does this mean "can represent"?

line 55: "embedded" Replace with "accompanying"?

Figure 1 caption: Describe the Full (or "true") distribution first.

line 61-70. Repeated material.

line 72: "PDFs are used to transform the information". Transform how? Doesn't the PDF (or histogram) of a particular taxon just represent the frequency of occurrence of a taxon along a climatic gradient?

line 73: "weighted ensemble" again.

line 77: It might be good to remind the reader that, for example, some pollen taxa represent individual species while other only genera (or indistinguishable types, e.g. *Larix* and *Pseudotsuga*). You could refer to section 3.3.2 for a discussion of how the species-to-taxon translation is made.

Line 78: "empirical mean and associated variance" Isn't this step in fact adopting the approach you reject in panel (a) in Fig. 1, especially when a taxon might have a nonsymmetric distribution?

line 85: Right. And see also Liu et al., 2020, Proceedings of the Royal Society A <https://doi.org/10.1098/rspa.2020.0346> and the fxTWAPLS package.

page 5: Some things missing here. What exactly is $k()$? What does i index? N vs N -sub- s ?

line 100: I think this combination-of-PDFs needs to be better explained. Would one want to ever lump all of the species of, say, *Pinus*, into one taxon? Some paleoecological data sources (e.g. pollen) are pretty "blunt" taxonomically, whereas others (e.g. plant macrofossils) are usually identified at the species level.

line 105: "grid cells" This is the first mention of grids/gridding. Does that have something to do with "geolocalized occurrence data"?

lines 105-107: Again, I'm worried about "robustness" which I think is being used more to denote some notion of reliability than in its usual sense. If robustness of the location and scale parameters (eqn. 1 and 2) is really a concern, why not use robust estimators of them?

line 117: "These definitions of sensitive taxa are always specific to a specific region...". Too many specifics. Maybe "the definitions are specific to particular regions"?

line 119: So the estimate is basically a weighted geometric mean of the individual taxon PDFs?

line 135: "own design"?

line 139: It might be good to cite to some of the discussions on fitting SDMs with presence-only data. (See Valavi et al., 2022, Ecol. Mono. <https://doi.org/10.1002/ecm.1486>)

line 148: "upscaled at a spatial 150 resolution"?

Figure 4: Increase the spacing between the bottom two panels to match the others?

Figure 5: Nice figure. The resolution of the image in the discussion .pdf obscures the double-framed boxes, however.

Section 3.2: In a "real-world" example, how large might a crestObj become?

line 236: Hyphenate "species-to-proxy".

line 256: Are the weights completely arbitrary, or should they lie within some particular ranges (e.g. 0-1 or 0-100)?

Section 3.3.4: Describe how the points in the climate-space data frame are associated with the taxon distribution data? Should there be a one-to-one correspondence between the rows?

line 285: "the original CREST software" This raises the question "Why not just use the original CREST software?" which is answered on lines 578-579. I think that the motivation for the development of the R package should be moved up to the introduction.

line 303: "the three input files" Are these files available anywhere? I don't see them in the GitHub repository.

line 320: Should there be some kind of transition between the text and code block? (As on line 349).

line 364: "different parameters" -- "different parameters that control the reconstruction"?

line 374: Again, some kind of transition is needed.

line 387: "the climate values to reconstruct are likely to be in the study area" Does this mean that the range(s) of climate values in the calibration data set should have the same (or larger) amplitude?

line 390: "homogeneous" In geographical space? Climate space? Both would be good I guess.

line 428: "violin plots" Violin plots are cool looking, but they are affected by our tendency to misinterpret/misjudge areas (i.e. Cleveland, W.S., 1993, Visualizing Data). I think viewers tend to notice the blobs as opposed to the profile, which is the important information. I think simply plotting the PDFs as in Fig. 6 of Quick et al. (2021) is more effective.

line 435: "to have objective" -- "to make objective"?

line 460: "to reduce the noise" One pollen type's noise may be another's signal. Pollen types have long-tailed distributions, and so an alternative approach might be to transform the data, with the square-root transformation in particular having some desirable properties.

Figure 10: "the thickness of the lines" I don't see much differentiation in line thickness.

line 505: "full uncertainties" Could a second panel be added to Fig. 9 to show that?

Figure 11: Why are all the values positive? What is the "net effect"?

Code:

I was able to install the package and run the example with only a few issues. The code in the get-started.html vignette and the GitHub README.md is a little different. I was able

to “pur1” the get-started.Rmd R Markdown file without an issue to reproduce the example. It wouldn’t hurt to also provide a pure get-started.R file. The results of the example wound up in an obscure temporary file. Adding “file.path(tempdir())” to the example would help the user to find that folder.

Pat Bartlein