General comments

Borderieux *et al.* study the determining factors of microclimate in mountain forests, and the link between microclimate and plant community composition. The manuscript is super well written, with a clear research question and methods. I am not a specialist of the topic, I would even define myself as an ignorant on that topic, but I enjoyed the reading. I have very few concerns about the article, but I have three general points on which I would like to draw the attention of the authors and the editor. I detail these 3 points below and added few specific and minor comments afterwards.

1. **Sampling plots exhibit very low variability in canopy cover**

My first comment regards the distribution of the loggers over the gradient of canopy cover, which might explain why authors found a very weak effect of the canopy cover on the microclimate. Authors sampled only forest plots with a canopy cover between 79% and 100%, so with very low variability. I am really not an expert in the topic, but by looking at previous papers, it seems that above 75% of canopy cover, the canopy cover does not affect microclimate anymore (Zellweger *et al.* 2019). Thus, in contrast to what authors say it is expected that they do not find any effects of canopy cover when studying plots with very low variability in canopy cover, all above 79%. The article of Zellweger *et al.* (2019) including some authors of the present manuscript, I am really surprised that this point is not discussed, while it could easily explain the results obtained.

Moreover, the proportion of broadleaves versus coniferous species seems an important determinant of canopy effect on microclimate (Díaz-Calafat *et al.* 2023), which varies across elevation, but this is ignored here. Do authors think that this neglected effect could affect their conclusion? I would tend to think that this should be discussed at least.

2. **Residual regressions provide biased estimates**

My second general comment is about how authors analysed the plant community dataset, and mostly about the use of two successive regressions, which for me can be a source of bias. Authors used a univariate linear model to remove the effects of soil pH on species richness, and then used the residuals of that regression as a response variable of another model trying to disentangle the factors influencing community composition. They used this two steps approach, also call residual regression (Freckleton 2002), to deal with collinearity between elevation and soil pH. For me, this approach does not make sense. Using univariate regression to deal with collinearity would rather create omitted variable bias and misestimation of uncertainty rather than solving anything. This point of view is not mine only but have been explicitly detailed by Freckleton (2002) and more recently recalled in the framework of structural causal models (Arif & MacNeil 2023). I explain myself:

If soil pH and elevation are correlated, performing first a model using soil pH as the only predictor will absorb all the effect due to soil pH but also the one related to elevation and correlated with elevation, because the linear model aims to explain the maximum of variance of the response variable with the included predictor. This is what have been termed omitted variable bias, that is the absence of important predictors will bias estimates of studied predictors (Byrnes & Dee 2024; Rinella *et al.* 2020).

To overcome that and include collinearity among predictors in the statistical uncertainty, multivariate regression should be used instead of two steps residual regression (Freckleton 2002). If soil pH and elevation are too colinear to be disentangle and included in the same linear model, then it would be more reasonable to assume that their effects can not be disentangled and to choose one of the two variables, or a composite variable, as a proxy for both, to control for variation in soil pH and elevation together. But I think using a two steps residual regression can only induce bias without solving the problem of disentangling effects of soil pH and elevation.

Thus, to study factor influencing plant species richness I would rather use a multivariate GLM, with a poisson or negative binomial error structure, including soil pH with other predictors, and checking that the variance inflation factor is not too high.

3. **Error propagation between models**

To explain plant species richness and CTI authors used the predicted effects of elevation, topography and canopy cover, converted into deviation in terms of temperature due to that variable (°C) to get comparable effects. However, these predicts have an associated uncertainty, which is not propagated in the model studying plant species richness and CTI. I was wondering if authors tried to do it and if it would affect the results? One way to test the robustness of these results would be to use the raw variables instead of their predicted effects on temperature to see if similar significant effects in terms of direction are found or not.

Specific comments

Lines 40-41: This is purely a personal opinion, but since CI means confidence interval, I would rather write explicitly the interval, [lower, upper], and keep "±" for standard error or standard deviation. Also, I find it nice to indicate the threshold of type I error for the confidence interval (for example $CI_{95\%}$)

Line 74: there is a typo here, with a dot and a comma in a row.

Line 100: change "," to "?"

Line 139: I guess here "product" means output, since product has a mathematical meaning (output of a multiplication), then I would avoid product.

Line 153: I had no idea what the lapse rate was. 2 seconds on google told me I am probably ignorant, but since I am probably not the only one and this concept is key in the paper, it might be worthy to define it very briefly (elevation gradient in air temperature?).

Line 157: I don't really see what authors mean by moderate. The topographic position indices vary from 0 to 0.8 which seems a lot for an index that varies from 0 to 1. But this seems logic because high elevation point have more chance to be closer to a ridge than to the bottom of the valley.

Line 150-161: A supplementary table might be useful to be sure reader get the sampling schema, if I get it well sampling points are distributed like this for each elevation strata:

		Canopy cover (relative to median)	
		below	above
Heat load index	Low (north)	1	1
	High (south)	1	1
Topographic index	Low (valley)		1
	High (ridge)		1
Slope	Flat		1
	steep		1

When representing clearly the sampling schema as a table, we can see that the sampling is quite unbalance between low and high canopy cover. Most of the sampling is spread in plots with high canopy cover, leading to few variations in that predictor of temperature. Do authors think that this could affect their results?

Line 185: "We cleaned the time series with the '*myClim*' R package" this step deserves to be better explained. Did authors remove abnormal values of temperature? If yes how many points were removed? I think, citing only a R package is not enough and require that the reader go to read the help of the package, without even knowing which function has been used.

Line 261 - 269: I think this hierarchical approach can lead to important bias, see my general comment.

Lines 281-282: again, I think that citing a R package is not enough, the method should be describe, at least briefly: is it based on ANOVA or on an iterative process of removing the variable one by one?

Lines 285-290: this long sentence, with a long insert between bracket is hard to follow

Line 292: here I would use variation instead of variability. In my opinion variability is more often use to refer to talk about how far a point is from the mean (dispersion) while variation is more often used to talk about shift in mean over a gradient. Here and after, authors talk about a shift in daily mean temperature across elevation, thus I would use variation instead of variability.

Lines 306-309: Why is there an interaction term included here but not in other models? Did authors perform a model selection for other models that removed the interaction term? I missed the explanations of this difference between the model of Table S3 and models of Table 1, S1 and S2.

Table 1: Are effect size calculated as the range of the predictor multiplied by the estimate? This is not standard and from what I know, it does not allow to get comparable effect sizes among predictors. The standardized effect size is classically calculated as the estimate multiplied by the standard deviation of the predictor.

abbreviations used could be defined in the caption "*…have no units (n. u.)…*"

Lines 380 – 382: this could supported by a statistical test, for example by testing if the skewness change among the three topoclimate categories

Line 385: not sure it is clear what the n refers to, I would write (n = 102 vegetation surveys)

Line 386 - 389: according to figure 4b this sentence is not exactly true. The difference of 300 is between the coldest and the warmest class. Why is there a reference in that sentence, if I am right it is a result of that paper so there is no need to refer to another paper? Or if authors want to do so a bit more context and explaining what the reference says would help.

Line 484 - 486: but here authors used only effects on mean temperature to explain plant community composition, so I find this conclusion a bit weird relative to what they do before. Cf lines 270 – 271: "*We used a linear model to predict the corrected species richness and CTI with the contribution to mean understory temperature of elevation, topoclimate and microclimate as predictors*".

Line 495 - 500: I am really not an expert of that topic, but I think the fact that authors used a canopy covert >75% with very few variation is a possible explanation for the lack of signal.

Fig. 2 caption: "*We restrained the minimal cooling to -1.5°C, however…*", I would write: "We truncated the color scale to -1.5°C to preserve readibility, however…" to be explicit that it is only for the figure, not in the methods.

<u>Table S1:</u> in Table S1 the canopy closure is called tree density, if I did not miss anything it is the same variable. It would be good to homogenize the names.

References

Arif, S. & MacNeil, M.A. (2023). Applying the structural causal model framework for observational causal inference in ecology. *Ecol. Monogr.*, 93, e1554.

Byrnes, J.E.K. & Dee, L.E. (2024). Causal inference with observational data and unobserved confounding variables.

Díaz-Calafat, J., Uria-Diez, J., Brunet, J., De Frenne, P., Vangansbeke, P., Felton, A., *et al.* (2023). From broadleaves to conifers: The effect of tree composition and density on understory microclimate across latitudes. *Agric. For. Meteorol.*, 341, 109684.

Freckleton, R.P. (2002). On the misuse of residuals in ecology: regression of residuals vs. multiple regression. *J. Anim. Ecol.*, 71, 542–545.

Rinella, M.J., Strong, D.J. & Vermeire, L.T. (2020). Omitted variable bias in studies of plant interactions. *Ecology*, 101, e03020.

Zellweger, F., Coomes, D., Lenoir, J., Depauw, L., Maes, S.L., Wulf, M., *et al.* (2019). Seasonal drivers of understorey temperature buffering in temperate deciduous forests across Europe. *Glob. Ecol. Biogeogr.*, 28, 1774–1786.