Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2020-316-RC2, 2020

Biogeosciences
Discussions

Interactive comment on "Geochemical zones and environmental gradients for soils from the Central Transantarctic Mountains, Antarctica" *by* Melisa A. Diaz et al.

Natasja van Gestel (Referee)

natasja.van-gestel@ttu.edu

Received and published: 4 November 2020

Review of "Geochemical zones and environmental gradients for soils from the Central Transantarctic Mountains, Antarctica"

This paper is very well written, and it is very interesting. The focus is on predicting spatial patterns of water-soluble salts and the ratio of N to P within 11 distinct ice-free sites along a glacier in Antarctica. Ultimately, the models that best predict those patterns could help find refugia of soil invertebrates who may be sensitive to high salt concentrations. I applaud the incorporation of the data and the R code, so that this research is reproducible.

My major concerns, which are easily addressed, are the statistics. I only devote so much time of describing these in detail, because I found it a most interesting paper that I believe should be published. But the statistical approaches should be sound and match the experimental design and follow the assumptions of linear models (otherwise model parameters cannot be properly interpreted). I would be happy to provide more guidance if needed.

1) Log-transformation need to be done where necessary. First, the authors have not checked their regression models to see how the residuals show a pattern with fitted values. This is important to do as one of the (several) assumptions in a linear model are that residuals are normally distributed and their spread around the regression line should be the same irrespective of the value of x. I highly suggest that the authors use log-transformed values in the data where needed and do some model checking with the plot() function and other model checking procedures. This will certainly help with that aspect. I was surprised to see that in the figure they did use log-values, but then did not use it in their regression. Needless to say, this this also needs to be done in their random forest models. The random forest model they used explained 43% of the variance in total salts. By using log-transformed values that went up to 75%. This also altered the importance ranking of the variables. Irrespective, elevation remained important (at least for Total Salts, I did not check the others), but others switched. 2) Data, such as NP, can have many zeroes. For example, the NP training data set had 116 zeroes of the 189 values. That means that a gaussian distribution of the data set is not followed. Having 0's means also a log-transformation will lead to -Inf, and cannot be used in a model. Solution: consider other family of distributions by using "glm" instead of lm. The "g" stands for generalized, and can thus handle other kinds of distributions. 3) Given that multiple samples were collected at each transect at 11 different locations, some in closer proximity than others, that error structure is not taken into account. For example, two samples from the same transect will likely be more closely related than two samples from different transect. To incorporate this, I propose using a mixed-effects model approach to take into account that multiple samples were obtained

from the same transect. Otherwise: your power is inflated, because your samples are not truly independent from each other (another important assumption in linear models with only fixed effects). 4) For the testing and training data set: rather than randomly sample all 220 observations, randomly sample a proportion of the total transects, say 8 out of 11 transect. Then test it on the remaining 3. That takes into account that the observations within a transect are not entirely independent. 5) For PCA: this is another linear model. I found no information on whether the authors performed any visualization to see if the patterns are linear. Given that the data show non-normal distributions, please do revisit this. Also: is the PCA based on the covariance or the correlation matrix? So, information is missing. 6) Lastly, and I refer to Figure 2: the panels are great, but it also highlights that the authors looked at every possible relationship of the total water-soluble salts, N:P, and ClO4- and ClO3- concentrations. However: the more comparisons are made, the higher the probability of making a type 1 error, unless you make the alpha more stringent. In a scenario like this I would recommend something like a Bonferroni correction.

Minor concerns: Replace 'environmental parameters' with 'environmental variables' or 'environmental conditions'. From a model-perspective, parameters are associated with models, e.g. coefficients are parameters. Variables are the data.

L. 158 It is mean squared error.

Figure 2: âĂć Please add the meaning of the blue, yellow and gray colors. It is evident from the next figure, but having it already here will help the reader. Figures are standalone and should be interpretable without having to look for info elsewhere in the paper.

âĂć Also, technically: if a relationship is not significant, one should not show the best-fit line. However, rather than removing them, I would suggest adding dashed lines instead for those where P-values are greater than 0.05.

Supplementary information: Please add the "library(readxl)" to the R script. Otherwise

users will get an error message: the read_excel() function is used, which is from that package.

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2020-316, 2020.