Response to reviewers: Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer functions by Yang et al.

Reviewer 2

General comments

The manuscript by Yang and colleagues describes a study whose aim is to test the predictive capabilities of learning models based on different types of predictors such as near-infrared spectroscopy (point data) and large-scale environmental variables with respect to the abundance and diversity of soil fungi.

The study is based on a rather large Australian dataset, comprising several hundred soil samples from different regions of Australia under different land-use types, for which fungal abundance and diversity have already been measured and published in a previous study, and for which near infrared (NIR) spectra (and related environmental data) have also been acquired and published in previous papers.

The authors test different strategies to build their learning models, varying the type of predictors used (spectroscopic data alone, environmental data alone, or a combination of both types of information) and the learning algorithms used (7 different algorithms). The performances of the different learning models to predict the abundance and diversity of soil fungi are tested by cross-validation.

The objective of this manuscript is of great interest to SOIL journal readers: are we able to predict the abundance and diversity of soil fungi by learning models based on simpler-to-obtain predictors in unknown Australian soils?

The article is overall well written and is very well structured. The figures and tables are clear. The learning models seem to have been well constructed by a team that is very well versed in data science techniques and the application of spectroscopy technique to soils.

Authors: We thank the reviewer for reading our manuscript and making an effort to

understand our work. We share the reviewers view that our work is of interest to the readership of SOIL and we are grateful for the overall positive comments.

Specific comments

However, in reading this article, several major shortcomings appeared to me.

Seasonality of the abundance and diversity of soil organisms First, an important point: it seems to me that the seasonal variability of the abundance and diversity of soil organisms in general and soil fungi in particular is a major "unthought" of this manuscript. In this work, no reference to this major determinant of fungal abundance and diversity and no information on the date of sampling and climatic conditions in the month prior to sampling of each soil sample was collected and exploited as a potential predictor by the authors in their learning models. It seems to me that through this "unthought" of seasonallity, the authors give up, without ever informing the reader, an important part of the determinism of their soil fungus-related variables of interest (cf. e.g. reference Kivlin and Hawkes, 2016 which is cited by the authors in their manuscript). In this paper we can read: "The strong seasonal pattern in fungal richness and abundance suggests that fungal studies in tropical forests require temporal sampling to capture the full community. Indeed, the inconsistent correlations of fungi with climate and plant hosts across tropical ecosystems [...] may be due to seasonal variation and spatial heterogeneity across single time point studies" (Kivlin and Hawkes, 2016). In general, how can a soil biodiversity measurement at a time t of a soil can reliably represent the diversity and relative abundance of different species/groups of soil fungi for that soil in its different seasons? Applied to the authors' objective in this manuscript: how can one hope to reliably predict very dynamic soil variables (fungal organism abundance and diversity) with predictor variables that are much less dynamic in the soil or in the environment? At best, one could hope to predict large differences between very contrasting soil pedoclimates: something that Biogeography of soil organisms works have often already identified. However, such Biogeographic works do not constitute

predictive models of soil biodiversity at a given time t. At the very least, it seems to me that this point, which is an important intrinsic limitation of this study (and also of a similar previous study focusing on soil bacteria by Yang et al. 2019 in Soil Biology and Biochemistry; `https://doi.org/10.1016/j.soilbio.2018.11.005`; should be discussed in detail by the authors.

Authors: We thank the reviewer for the comment and suggestion. Of course, we agree with the general comment on the seasonal variability of microorganisms. We will highlight in a revised discussion the shortcomings of any modelling when not accounting for seasonality of the abundance and diversity of soil fungi, also with reference to Kivlin and Hawkes (2016). To make the point clear, we will revise the introduction leading to our aims with something like: '...Fungal abundance and diversity are dynamic in space and time (Duan et al., 2018), often showing that habitat prevalence differs seasonally (Talley et al., 2002).The inconsistent correlations of fungi with climate and plant hosts observed in various ecosystems may be due to seasonal variation and spatial heterogeneity across single time point studies (Kivlin and Hawkes, 2016). Thus, temporal sampling and seasonal climate data are needed to capture the seasonal dynamics of microbial communities. In our study, the development of spectro-transfer functions with vis–NIR spectra and machine learning is based on soil fungal measurements at a single time. Despite this limitation, our approach provides a possible opportunity to infer the continuous distribution of soil fungal communities and diversity more simply and at a smaller cost. Thus it can help to better understand the diversity and biogeography of soil fungi in different habitats.....'. Additionally, for completeness, we could also add information on the climate at the time of sampling and in the month prior to sampling.

Overly optimistic conclusions compared to reported results A second major point concerns the conclusions that the authors draw from their results regarding the predictive capabilities of the learning models they have constructed for the abundance and diversity of soil fungal groups in Australia. These conclusions seem to

me to be overly optimistic compared to the results the authors show in their manuscript. After reading this paper, if I had to answer the question: can we, in unknown Australian soil samples, robustly estimate the abundance and diversity of major soil fungal groups with the learning models constructed in this work, my answer would be: no, and we're a long way off. At most we can identify some major soil and environmental determinants of the abundance of these groups and their diversity, making this a work of Biogeography of soil organisms, much more than a paper that aims to predict these properties in unknown soils to study soil health.

Indeed, in the introduction to their article, the authors mention that "a paucity in the availability of soil microbial data is thought to be one of the main contributors to the uncertainty of soil health assessment and ecosystem management" while in concluding their paper, the authors state that "here, we show that spectro-transfer functions with readily accessible vis–NIR spectra and publicly available soil and environmental data can be developed to estimate soil fungal abundance and diversity" I disagree with the conclusions of the authors: using their models for soil health assessment of a particular Australian soil (for a criteria of soil health linked to soil fungal abundance or diversity for instance) would certainly not help improving the acuracy of this assessment given the predictive performance showed in this manuscript.

Specifically, the models with the best predictive capabilities use a combination of point spectroscopic data and continuous environmental data, and they most often use the deep learning algorithm 1D-CNNs. But despite these (still somewhat obscure) performance differences between the different algorithms, even the "best" results shown by Yang and colleagues are not very encouraging in terms of the actual predictive capabilities of the learning models built on unknown Australian soil samples:

- For the two (very) dominant soil fungus groups in Australia (Ascomycota; Basidiomycota) representing on average 80% of the total fungal abundance; Table 1),

the "best" learning models show R^2 below 0.6. I infer that the learning models constructed in this paper do not robustly quantify the abundance of groups representing 80% of the fungi in Australian soils.

- For the diversity of fungi estimated by the "ACE" indicator, the $R^2$is 0.45, I deduce that the learning models built in this article do not allow to quantify in a robust way the diversity of fungi in Australian soils.

- A few models with slightly better predictive abilities are shown for the abundance of two groups of very low abundance fungi in soils (although it is difficult to quickly judge their performance in the absence of synthetic performance indicators such as the ratio of performance to deviation or the ratio of performance to interquartile distance).

Authors: We agree with the reviewer's concern. Generally, we are some way off very accurate estimates of fungal abundance and diversity with our method. We understand that the large unexplained variance of fungal abundance and diversity were mainly because of the complexity of the living organism itself and large undetermined controlling factors, including seasonal variability (see previous comments and responses). We will make it clear in the revision that in this study we presents a promising method to estimate fungal communities and diversity. We hope that our study will provide food for thought and guide further research towards achieving more robust fungal abundance and diversity predictions, which we think will be possible with our deepening understanding of fungal biology, biogeography, the improved understanding of controlling factors and with the development of new technologies and methods. Nevertheless, the spectro-transfer functions we constructed could fairly accurately (as per the reported statistics) estimate fungal community composition and diversity measured with ITS gene metabarcoding. As we stated in the manuscript, we do not think that our method will replace metabarcoding, however it could complement it when many measurements are needed. Our hypothesis here is that even if some of our spectro-transfer functions are

relatively innacurate (as per our Figure 4), when one has many samples to measure across an area, the estimates will better represent the variability present and as a whole be more accurate than when only taking a few very accurate measurements. We can strengthen the discussion around this point, however, testing of this hypothesis will need to be for a future study. ITS gene metabarcoding analyses are expensive, laborious and require specialised laboratories and methods, while vis–NIR spectroscopy is relatively faster and less expensive, and soil-environmental data are more readily available. Using soil sensing technologies, such as spectroscopy together with molecular approaches could greatly improve the utility of microbial inventory data (Hart et al., 2020). For example, Delgadobaquerizo et al.(2018) mapped the global distribution of the soil bacterial composition based on only 237 bacteria inventory data in the world. The global spectroscopy data is very abundant and reach tens of thousands. With the use of spectro-transfer functions, the bacterial data density will be improved, then benefits the more accurate digital soil mapping of bacteria or fungi. Thus, we will tone down and clarify our conclusions on the predictive capabilities of the learning models constructed.

Other points

Section 2.1: ecosystem types, please clarify the difference between woodland and forest.

Authors: Thank you, we will make the distinction. The term woodland is generally used in Australia to describe ecosystems which contain widely spaced trees, the crowns of which do not touch. Woodland consists of areas with fewer and more scattered trees than forests. In temperate Australia, woodlands are mainly dominated by Eucalyptus species. Temperate woodlands occur predominantly in regions with a mean annual rainfall of between 250–800mm, forming a transitional zone between the higher rainfall forested margins of the continent and the shrub and grasslands of the arid interior.

Section 3.1: In total, more than 60 million quality filtered sequences in the

whole dataset were obtained, with an average of 107 310 sequences per sample. When we clustered the sequences at 97% similarity level 202200 OTUs were detected. Each sample had an average of 666 OTUs : this was already presented in the section 2.2, please remove it from the results section.

Authors: We thank the reviewer for picking this up. We will revise it.

Figure 1b: please add some information to remind reader that this graph shows mean abundances (this graph does not represent errors on the mean, which is huged as shown in Table 1).

Authors: Thanks you, agree and we will revise .

I agree with R1 that interpreting NIR spectra with continuum removed reflectance signals and using savitsky-golay first derivatives of absorbance signals in the modelling work can be misleading. Please interpret the NIR spectra with the signal used for modelling.

Authors: We refer the reviewer to our response to R1 on this matter. We will add the SG1Der spectra too.

I wonder how does the sum of the model predictions for the relative abundance of the 5 main groups of soil fungi behave for the soil sample set used in this study: does this sum get close to 1 for all soil samples?

Authors: Thank you for the question. No, the sum of the model predictions for the relative abundance of the 5 main groups of soil fungi were not close to 1 for all soil samples. There are samples with the sum larger than 1. Our models did nothing to constrain the cumulative abundance.

References

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., et al, 2018. A global atlas of the dominant bacteria found in soil. Science, 359 (6373), 320-325.

Duan, Y, Xie N, Song, Z,et al., 2018. A High-Resolution Time Series Reveals Distinct Seasonal Patterns of Planktonic Fungi at a Temperate Coastal Ocean Site (Beaufort,

North Carolina, USA). Appl Environ Microbiol., 84(21), e00967-18.

Hart, M. M., Cross, A. T., D'Agui, H. M.,et al., 2020.Examining assumptions of soil microbial ecology in the monitoring of ecological restoration, Ecological Solutions and Evidence, 1, e12031.

Kivlin, S. N., Hawkes, C. V., 2016. Tree species, spatial heterogeneity, and seasonality drive soil fungal abundance, richness, and composition in Neotropical rainforests, Environmental Microbiology, 18, 4662–4673, 201.

Talley, S.M., Coley, P.D., Kursar, T.A. 2002. The effects of weather on fungal abundance and richness among 25 communities in the Intermountain West. BMC Ecol., 2, 7.