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

We thank the topical editor and reviewers for their comments. Below we provide detailed responses (in blue text and preceded by **Authors:**), indicating the changes we made in our revision. First, we address the topical editor's comments and then the reviewers.

Topical Editor: Comments to the author

Two reviewers have evaluated the manuscript and suggested several points to improve the manuscript. Thank you for taking up most of these points in your reply and a revised version. Please not that a revised version is generally only due after the paper discussion period and the editorial board decision on the paper.

In the comment to R1 already posted on the site, it seems to me that you have not written clearly what changes you intend to make.

Authors: The changes that we proposed relate to the representation of the spectra in Figure 2, the R^2 used and an improved explanation of the variable importance method. Please see below for details.

I beleive R1 is advocating for parsimonious and understandable models. It is generally advised to run a covariates selection step before interpretation of the models see for example https://doi.org/10.5194/soil-7-217-2021.

Authors: Thank you for the comment. We agree that it is generally sensible to perform a variable selection prior to modelling, particularly when there are hundreds of geographical predictors and when employing the environmental correlation approach in spatial machine learning (like in the example provided by the editor).

Spectroscopic modelling is somewhat different and variable selection tends not to work all that well when developing predictive models—hence the preference for dimensionality reduction methods for multivariate calibrations, such as partial least squares regression (PLSR). Our experience with machine learning is similar in that variable selection prior to modelling tends to produce models that explain less variance that models that use all of the wavelengths. It might be the 'non-specificity' and collinearity of the wavelengths for modelling soil properties. When we were developing the research and during our initial analyses, we did perform a variable selection using different methods to select important wavelengths and the Boruta algorithm performed best. However, we found that the models did not perform as well as when we used all wavelengths. Table 1 below, shows these results for PLSR and CNNs, and shows that models that used all of the wavelengths (full-spectrum+DSM+ENV) accounted for more variance in fungal phyla abundance and diversities than the models with only selected variables (spectrum+DSM+ENV). Please note that we have result for all of the algorithms tested, however, to illustrate our response, we show results only for PLSR and for the more complex CNN approach. Therefore, based on these results, in the manuscript we report only the full spectrum results.

Table 1: Comparison of goodness of estimates by a 10-fold cross validation based on PLSR and 1D-CNNs.

Variables	PLSR		1D-CNNs	
	Full-spectrum + DSM + ENV (R^2)	Boruta selection + DSM + ENV (R^2)	Full-spectrum + DSM + ENV (R^2)	Boruta selection + DSM + ENV (R^2)
Abundance				
Ascomycota	0.37	0.33	0.50	0.45
Basidiomycota	0.42	0.39	0.58	0.52
Mortierellomycota	0.59	0.53	0.71	0.65
Glomeromycota	0.17	0.15	0.48	0.36
Mucoromycota	0.51	0.45	0.73	0.62
Diversity				
ACE	0.32	0.30	0.45	0.41

Generally, the two reviewers suggested more balanced analysis and interpretation. I think that both readers agree that the conclusions are too much optimistic.

Authors: We understand Reviewer 2's comment about the somewhat 'optimistic' conclusions. We have clarified and toned down our conclusions on the predictive capabilities of the spectro-transfer functions. We revised the relevant sections of the discussion as follows:

- We revised the discussion around the estimation with the spectro-transfer functions: "...We show that spectro-transfer functions with readily accessible vis–NIR spectra and publicly available soil and environmental data could variably estimate (with R^2 ranging from 0.45–0.73) soil fungal abundance and diversity measured with ITS gene metabarcoding...". Note that here we report the R^2 range without subjective assessments on the predictability of the models.
- We clarified how these models with varying predictability could complement (not replace) molecular approaches for the assessment, characterization and improved understanding of soil fungal communities. See below, in response to R2.
- We make it clear that our method and results are **encouraging** (not a fait accompli) with potential to help characterise fungal communities and diversity (when used together with the more difficult-to-measure molecular methods). See below, in response to R2.

Authors: To provide a more balanced analysis of our results, we also cite other studies that use the general concept of proxies to achieve rapid estimates of other microbial properties to help diagnosis of soil quality. See below, in response to R2.

Several aspects need some further attention: · the use of a more synthetic performance indicators instead of the R2. I think it is important to give the formula of the R2 you used as it is possible to confuse between 2 formulas. The model efficiency coefficient MEC (Janssen and Heuberger, 1995) is equal to the fraction of the explained variance based on the 1:1 line of predicted versus observed that is defined as 1 minus the ratio between residual sum of squares and total sum of squares. Did you use this one ?

Authors: Yes, we used the Sutcliffe model efficiency and have specified this in the methods. Please see below, in response to R1.

· The use of the soil covariates in the model is derived from the 90m DSM products not from the soil samples, therefore it may decrease the robustness of the pedo-transfer function. Please, could you add a discussion on this aspect.

Authors: We agree that using the estimates from the digital soil mapping products will not be as good as using measured data. However, the idea of the spectro-transfer functions (much like the more conventional PTFs) is to use data that is inexpensive/free and readily available. Using measured data would increase the cost of the approach significantly and in that case, it might make more sense to use the molecular methods directly. We take the general point though, so we added some discussion as follows: "... The soil covariates in the model are derived from digital soil maps and not from measured soil samples. The reason is that using measured data would increase the cost of the approach significantly, making the approach less attractive. We note that the uncertainty in the spectro-transfer estimates caused by using the digital soil map predictors will propagate to the spectro-transfer functions and thereby lowering the precision of the estimates..."

· I am less concerned than R1 about the overlap with previous paper as the modelling of fungi may raise other issues than modelling of bacteria. It is genarally admited that Fungi are for example not as dependent on specific plant species as some bacteria (https://soilquality.org.au/factsheets/soil-bacteria-and-fungi-nsw).

Authors: Thank you, that is also our position.

We are looking forward to a revised version of your manuscript. Yours sincerely,

Nicolas

Revisions based on reviewers comments

Below, we detail the revisions made as per our responses to each reviewer's comments in the discussion.

Revisions based on reviewer1's comments

Authors:

- To prevent confusion between visualization and interpretation of spectra, we included both types of spectra in Figure 2, the continuum removed and the absorbance, first derivative spectra. We revised the relevant sections in the Methods and Results.

- We clarified in manuscript that use used the Nash Sutcliffe model efficiency R^2 as follows: "...We evaluated the estimates using the Nash Sutcliffe model efficiency, other wise known as the coefficient of determination (R^2), which represent the fraction of the explained variance based on the 1:1 line of estimated versus measured values(Janssen and Heuberger, 1995). The R^2 was computed as 1-RSS/TSS, where RSS is the residual sum of squares and TSS is the total sum of squares."

- We revised and improved the description of our implementation of the variable importance as follows: "...To calculate the variable importance of the CNN models, we used permutation variable importance. In our case, we run 1000 permutations and measured the decrease in RMSE after a predictor was permuted (randomly rearranged). The permutation breaks the relationship between the predictor and the response variables, and a reduction in RMSE indicates how much the model depends on the particular predictor. An advantage of this approach is that it can be applied on any estimator and does not require retraining the model (Breiman, 2001; Fisher et al., 2019). In order to compare the importance between different fungal phyla and diversity, we scaled the importance values between 0 and 1."

Revisions based on reviewer2's comments

Authors: In summary, Reviewer 2 has two main concerns with our work: (i) the lack of a discussion on the seasonal variability of microorganisms and (ii) his perceived 'over-optimistic' predictability of the spectro-trasnfer functions. He also has some other points for us to consider. Below, we address each of the comments and suggestions made.

Authors: Addressing the comments on the seasonal variability of microorganisms :

- To clarify, we added discussion on the seasonal variability as follows: "...Fungi vary over space and time (Duan et al., 2018), often showing that their prevalence in different habitats 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 is needed to capture the seasonal dynamics of microbial communities. Our research uses soil fungal measurements at a single point in time. Despite this drawback, our approach allows us to infer the distribution of soil fungal communities and diversity more simply and at a lesser cost, to help better understand the diversity and biogeography of soil fungi in different habitats..."
- In the Methods, we now include general information on seasonality and climate at the time of sampling, as follows: "...In that project, sampling were

undertaken from soil that supports diverse plant communities across Australia. The sampling was carried out during the growing season when hydrothermal conditions are most conducive to typical plant growth. In the higher rainfall forested regions of the continent, the soil samples were collected mostly in spring and summer from September to February. In the shrublands and grasslands of the semi-arid and arid interior, soil samples were collected in spring from September to November. In the transitional zone between the southeast coast and the more arid interior, soil samples were collected in mainly autumn from March to May...“

Authors: To address the comment on the ‘over-optimistic’ predictability of the spectro-transfer functions, we have toned down and clarified our discussion as follows:

- ”...Soil fungi play essential and diverse functional roles in ecosystem. However, they are challenging to investigate due to laborious, time-consuming and costly field sampling, and laboratory analysis. We show that spectro-transfer functions with readily accessible vis–NIR spectra and publicly available soil and environmental data could variably estimate (with R^2 ranging from 0.45-0.73) soil fungal abundance and diversity measured with ITS gene metabarcoding. The general concept of using proxies has been used in other studies to attempt more rapid estimation of microbial properties towards the diagnosis of soil quality. For example, Horrigue et al.(2016) developed a statistical predictive model of soil microbial biomass according to environmental parameters including soil physico-chemical and climatic characteristics across France. Their model ($R^2 = 0.67$) provided a reference value of microbial biomass for a given pedoclimatic condition to enable rapid diagnosis of soil quality across France. Other similar studies exist, for example Griffiths et al. (2016) who focused on the estimation of bacterial community structure and diversity at the Europe scale...”

- “...ITS gene metabarcoding analyses are expensive, laborious and require specialised laboratories and methods, while spectroscopic measurements are faster, less expensive, and soil-environmental data are more readily available. When many measures are needed, for example, to assess, characterise and improve our understanding of soil fungal communities and their associated functions at different scales, the approach could complement molecular techniques (Hart et al., 2020). For instance, to characterise spatial variation (i.e. for mapping), one needs many measurements that would be too expensive with only metabarcoding. In this case, estimates with the spectro-transfer functions (R^2=0.45–0.73) could complement the metabarcoding analysis to represent the variability present better. As a whole, the spatial characterisation will be more accurate than when only taking a few very precise measurements. This is the rationale for the characterisation of soil properties in space and time with sensing (Viscarra Rossel et al., 2011)."

- "...We do not expect that the spectro-transfer method will produce estimates that are as accurate as the more conventional molecular methods, even with further improvements in modelling and better covariates. This is because we understand that the modelling of living organisms is dynamic and hugely complex. Fungi vary over space and time (Duan et al., 2018), often showing that their prevalence in different habitats 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 is needed to capture the seasonal dynamics of microbial communities..."

- "...Our research uses soil fungal measurements at a single point in time and there are likely to be many undetermined controlling factors, including seasonal variability and complex biological interactions. Despite this drawback, our approach allows us to infer the distribution of soil fungal communities and diversity more simply and at a lesser cost, to help better understand the diversity and biogeography of soil fungi in different habitats. Thus, our approach shows promise and could complement molecular methods. We hope that our study will stimulate further research towards achieving more widespread characterisation of fungal abundance and diversity, which will help to deepen our understanding of fungal biology, biogeography and their environmental controls. Different spectra, new sensing technologies and improved methods could also improve the spectro-transfer approach...”

- We removed discussion that refers to soil health because we understand that this is a rather ‘controversial’ topic. Our research here is different and does not necessarily contribute to that discussion.

- We restructured and rewrote parts of the conclusions to emphasise that this work does not aim to provide a replacement method to measure soil fungi, but an encouraging new method that could help to complement the more expensive molecular method. The revised conclusions are: “Our study contributes to the development of methods that could complement, not replace, molecular approaches for characterising and better understanding the diversity and biogeography of soil fungi. We have shown that deep learning spectro-transfer functions are a promising new method for estimating soil fungal communities’ relative abundance and diversity. The optimised 1D-CNNs outperformed the six other machine learning algorithms tested for estimating the relative abundance of fungal phyla and diversity. The spectro-transfer functions (with vis–NIR spectra and soil and environmental data) produced more accurate estimates (R^2 0.45–0.73) than the spectroscopic models (only vis–NIR spectra; R^2 0.36–0.55) and models with only the soil and environmental data (R^2 0.38–0.60). As well as the soil organic and mineral composition, represented by vis–NIR spectra, other edaphic, climatic, and biotic factors including soil nutrients, pH, bulk density, potential evapotranspiration, the soil-water balance and net primary productivity were important predictors in the modelling. "

Authors: Regarding the reviewer's other comments, we revised as follows:

- We clarified the distinction between woodlands and forests in the Method, as follows: "...Woodlands in Australia represent ecosystems which contain widely spaced trees, the crowns of which do not touch. Woodlands consist 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....".

- We removed repetition and only present the relevant text in the Results: "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" .

- We revised the caption of Figure 1 to remind readers that the graph shows mean abundances: "...The mean relative abundances of dominant fungal phyla and unclassified "Others" taxa in five ecosystem types..."

- We added the SG1Der spectra in Fig. 2. We revised the relevant sections in the Methods and Results. Please see above.

- Regarding the comment on the '...absence of synthetic performance indicators such as the ratio of performance to deviation or the ratio of performance to interquartile distance...' We have used evaluation metrics that quantify the error in the model estimates in terms of their inaccuracy (RMSE), bias (ME) and imprecision (SDE) (such that $\text{RMSE}^2 = \text{ME}^2 + \text{SDE}^2$) and the R^2. Reporting RPDs or RPIQs will not help to better characterise or compare the errors and would only be redundant.

References

Janssen, P. and Heuberger, P.: Calibration of process-oriented models, Ecological Modelling, 83, 55-66, 1995.

Breiman, L.: Random forests, Machine learning, 45, 5-32, 2001.

Fisher, A., Rudin, C., and Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an EntireClass of Prediction Models Simultaneously, Journal of Machine Learning Research, 20, 1-81, 2019.

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.