Response to Reviewer 2

Key
Comments have numbered as Reviewer_number#comment_number (e.g. R2#12).

Italics: original comments by the reviewer
Normal font: response
Bold: changes in the manuscript.

Citations without reference correspond to papers cited in the manuscript. New references are specified here.

R2#1. The paper compares different imputation methods for trait databases not only regarding their accuracy in predicting plant traits and trait-trait relationships but also their ability to preserve trait distribution and multi-trait correlation structures. They use a plant trait database that has complete trait observations for 5 traits in 630 plots together with the auxiliary information and the data are georeferenced. Exhaustive test are done to compare the methods in using only species mean data, adding auxiliary data or geographical information with different levels of gaps and 4 different complementary evaluation method i.e. R2, NRMSE, KGE, differences in the correlation matrices. This I believe is done for the first time in such a thorough way. The study is very relevant for biogeographical and ecological studies that need to use trait databases, which are usually not free of gaps. However, compared to the number of methods and test done, there is not enough explanation and discussion done. I would suggest a re-evaluation of the discussion part. For example try to explain why the selected traits behave differently in the results.

R2#2. I would suggest to add information on the abbreviations of the methods, auxiliary information and traits for all Figures also the figures in the supplementary. The Figure captions must be self-explanatory.

- Done. Here is the caption for Figure 4 as an example:
- ***Figure 4. Trait-specific KGE at increasing missingness levels (10% to 80%) and for different MICE imputations using different combinations of additional predictor sets: species identity (s), climate (c), forest structure (t), sptatial structure (p), lithology (l) and sampling month (m). See Fig. 1 for an overall view of the experimental design and the Methods section for a detailed description of the variables employed in each predictor set. Traits: leaf biomass to sapwood area ratio, BL:AS (t m-2); nitrogen per unit mass, Nmass (%mass); maximum tree height, Hmax (m); leaf mass per area LMA (mg cm-2); wood density, WD, (gm cm-3). Higher values of KGE imply higher performance.***

R2#3. The paper uses different names like environmental information, ecological information/variables, and auxiliary information/variables etc. for the same thing and this is confusing. I suggest to use only auxiliary information or variable throughout the paper and to be consistent, especially since the data listed in the as ecological information (L90) are not all considered ecological exactly.

ABSTRACT

R2#4. I suggest that the authors summarize the results in the order of the questions mentioned at the end of the introduction.

- The structure of this section of the abstract, where we outline the results, is as follows: (i) using averaging to impute trait datasets may work well, on average, but these methods distort trait variability and covariation, (ii) which environmental variables are more useful to produce

plausible imputations, (iii) how do results vary across traits and missingness levels and (iv) describe the application of a best-performing method for the entire dataset. We are aware that the structure suggested by the reviewer would make easier to map the results between different sections of the paper. However, during early stages of manuscript writing, and because of the highly technical nature of the paper, we decided to mix results and discussion. In the case of the abstract, the messages we try to convey are better expressed using this logic, rather than the order of the questions in the Introduction, which are more related to the three different 'exercises' in the manuscript. For these reasons, we prefer to keep the present Abstract structure.

R2#5. L 27: the word globally is misleading.

- We have replaced 'globally' by 'overall'.

INTRODUCTION

This part is well written.
R2#6 L 42: It is not clear what is meant by "primary sources"?

- We refer to individual studies from the primary literature (mainly research papers). For clarity, we have removed 'primary sources and' so the text now reads (section 1):
- **Plant trait databases compiled from multiple individual contributions lack a common design and inevitably result in sparse data matrices (e.g. Jetz et al. 2016).**

R2#7. L 91: why is regKrig not in the second test together with MICE and kNN?

- The second exercise deals with how auxiliary variables can improve imputations, and we chose to do this analysis only for MICE and kNN. Both methods are easy to implement for multivariate datasets in our simulation framework, but regression kriging is not such a flexible tool. For example, we could not include species identity for regKrig because, for less common species, there were not enough data to perform the regKrig imputations at high missingness rates.

R2#8. L91: perhaps would be better to define what you mean by "optimum level" here- best set of auxiliary variables selected in step ii - and then use the phrase hereafter

- We have rewritten the sentence following the reviewer's suggestions (section 1):
- **(iii) to compare the performance of kNN, MICE and RegKrig using optimum levels of ecological information (i.e. the best set of predictors in objective ii);**

METHODS

R2#9. Chapter 2.3: Please use this part to explain "Mean" and Spmean" technique a bit more.

- We think that 'Mean' and 'Spmean' are already well explained when we present the imputation methods (section 2.3):
- **We compared imputation methods with different degrees of complexity. We used two simple approaches to provide baseline imputations; Mean imputation ('Mean') filled missing data using the overall mean value for each trait and species mean imputation ('Spmean') replaced missing values with trait means computed for each species.**

R2#10. L121: Please write the full name for MCAR as well.

- We have rewritten the sentence following the reviewer's suggestion.

- **In this dataset, we randomly deleted measured values at different probability levels (10%, 20%, 30%, 50% and 80%) and independently for each trait, thus the missing data artificially introduced are missing completely at random (MCAR).**

R2#11. L173: what type of "Lithology" data is used? Please include it here.

- Lithology includes three categories: calcareous, non-calcareous and undetermined. We have added this information to the sentence (section 2.4).
- **The auxiliary variables we considered were species identity, a set of climatic variables (mean annual temperature, annual thermal amplitude, both in °C), a set of forest structure variables (total aboveground biomass [T ha^{-1}] and stem density [stems ha^{-1}]), a set of topographical variables (county, elevation [m.a.s.l.], slope [°] and aspect), lithology (calcareous, non-calcareous or undetermined) and sampling month.**

R2#12. L175-178: Why are some auxiliary information added in a factorial design and some sequentially? What about running a test for each trait independently for realizing the most important auxiliary information that explains their variability best and then using those for the MICE and kNN model (also regKrig)?

- We were interested in identifying which combinations of the variables with a major role in explaining trait variability (species identity, climate and forest structure; see Vilà-Cabrera et al 2015), led to improved imputations, and for this reason we included these variables in a factorial design. Other variables which we expected to play a secondary role (topography, lithology, sampling month) were added sequentially. See also response *R1#20 above.*
- We did this exercise mainly for MICE, but also somehow for kNN, because they were easier to implement in our simulation framework and they are both algorithms that are designed to deal with multivariate missing data (as opposed to regression kriging).
- We are not really sure to understand what the reviewer is suggesting here. Does the reviewer suggest to use different information levels depending on each trait instead of using a best-set (e.g. mice_ctsp) for all traits? We think that our approach deals with the role of auxiliary variables in imputation performance at two levels: (1) which combinations of relevant ecological variables improve imputations and (2) which secondary variables can help improve those imputations.

R2#13. Chapter 2.5: I would suggest making clear which statistical method is used for evaluating accuracy, multi-trait correlation structure or bivariate trait relationships.

- Following the reviewer's suggestion, we have added a clarification on the association of NRMSE as a measure of imputation accuracy (section 2.5):
- **For each simulated dataset and trait, we calculated the Normalised Root Mean Square Error (NRMSE) as a measure of accuracy**

- Also, following the reviewer's suggestion we have explicitly added 'multi-trait correlation structure' in the description of Δcormat (section 2.5):
- **The deviations from the original multi-trait correlation structure of the trait dataset were quantified by comparing the correlation matrices of the original and imputed datasets using the following index:**

- As for the impact on bivariate trait relationships, we consider that the explanation is already clear (section 2.5):
- **We also tested the impact of the imputation algorithms on selected bivariate trait relationships: H_{max}–WD and N_{mass}–LMA (log-transformed when necessary); as the**

correlation coefficients (*r*) of these relationships were >0.3 in absolute value and were highly significant in the complete dataset. We quantified the relative difference between the complete and the imputed datasets by calculating:

- KGE is a more complex indicator, which also includes accuracy, and we think that its current explanation is clear enough (section 2.5):
- **KGE jointly assesses correlation, bias and variability between imputed and observed values, and it is therefore a powerful, synthetic indicator of imputation quality in spatially-explicit datasets.**

R2#14. L232: "as the best method globally" – Why globally?

- We mean 'overall', considering all traits and performance metrics, as explained later in the sentence. We have replaced 'globally' by 'overall'.

RESULTS AND DISCUSSION

R2#15. This part is in general not well written. There are a lot of results that looking at them in the paper and its' supplementary brings up many questions but the authors did not discuss them well. It's mostly like a report of results and not much discussion and reasoning of the results. Please discuss the performance discrepancies of the methods for different traits, and for the different evaluation criteria (trait-trait correlation, trait distribution, etc.).

- We have changed the 'Results and discussion' section extensively, adding more statistical tests and rearranging parts of the text. See our replies to comments *R2#16, R2#21 and R2#24.*

R2#16. First paragraph of chapter 3.1: Fig. 2 and Fig. S7 shows different results for different traits. How do you interpret this? Please discuss.

- We have now rewritten this entire paragraph and we place more emphasis on particular results of some traits (Nmass, Hmax, LMA, see section 3.1).
- **In general, 'mice' and 'kNN' imputations resulted in more accurate imputations in terms of NRMSE than 'Mean' at low missingness rates (10%). However, at moderate and high missingness both 'mice' and 'kNN' were comparable or outperformed by 'Mean', and specially by 'OrdKrig' (Fig. 2, Fig. S7). 'OrdKrig' was the best-performing method, in terms of NRMSE, at missingness ≥ 50% (P <0.05), although for three traits its performance was indistinguishable from that of 'Mean' imputations (Nmass, Hmax, LMA; P>0.05). Even if 'Mean' imputations imply the rather naive assumption that species identity may be unknown in a given dataset, it is nonetheless useful to compare 'Mean' imputations against 'mice' and 'kNN', which use the full trait matrix for prediction. In this case, trait covariation did not improve imputations at high missingness. Recent assessments also report that the performance of MICE and kNN notably declines when missingness is ≥ 30% (Penone et al. 2014; Taugourdeau et al. 2014). Therefore, our results for 'OrdKrig', compared to those for 'mice' and 'kNN', show that spatial structure, rather than trait covariation, may provide more accurate trait imputations when gaps are frequent (Fig. 2, Fig. S7, S8).**

- However, it should be taken into account that the focus of the paragraph is rather to look at generalities, not at differences between traits. We will discuss this issue in more detail in the new version when looking at the role of auxiliary variables in improving imputations (and how

these improvements depend specifically on the trait and type of auxiliary variable). For an example, see our reply to comment *R2#24.*

R2#17. L239, L246, L250: Instead of Supplementary S5, please indicate Fig. S number. Please do so for the rest of the paper as well.

- Done.

R2#18. L246 Please remove "however".

- Done.

R2#19. L279: "stand structure" was mentioned before (L25, L90, L146, etc.) as "forest structure". Please choose one and be consistent.

- We have replaced 'stand structure' by 'forest structure' (three appearances throughout the text)

R2#20. Chapter 3.2: Why the comparison of adding different levels of auxiliary information was only used for MICE and not for kNN or RegKrig?

- We actually did this exercise mainly for MICE, but also to some extent for kNN, because we ran kNN using species (kNN_s) and the best predictor set (kNN_ctsp). We used these two methods because they were easier to implement in our simulation framework and they are both algorithms that are designed to deal with multivariate missing data (as opposed to regression kriging). Moreover, the study already includes many different aspects (imputation methods, auxiliary information used in the imputations, missingness, multiple traits) and we think that our approach already shows, for a subset of imputation methods, the role of different auxiliary variables in improving the imputations, for different traits.

R2#21. In addition, please discuss why mice_ctsp was chosen while according to Fig. 4 mice_sc would be sufficient for Nmass, LMA and WD and mice_st for Hmax and only BL:AS needs mice_cstp (extra p).

- We agree with the reviewer in that, for some traits, imputation performance does not improve in mice_ctsp compared to some mice applications with less auxiliary variables. However, we had to choose a number of auxiliary variables for which performance was maximised for all traits. This is the reason why we chose mice_ctsp. We have now specified this in the text (section 3.2):
- **Our results collectively suggest that, apart from species identity, different types of ecological information, particularly forest structure and topography, may improve statistical imputation schemes. In contrast, the role of climate, lithology and sampling month in improving imputations was comparatively minor. However, we selected 'mice_ctsp' as the method that performed best for all traits, because adding climate did not reduce imputation performance and not including 'topography' would worsen BL:AS imputations.**
- Note also that using different levels of auxiliary information for different traits would complicate substantially the comparison of different methods (the main objective of the paper) and the interpretation of the results.

R2#22. L288: Please replace A with S in "Fig. A7 and A12".

- Done

R2#23. L288-289: what about adding only species and forest structure to RegKrig? Climate and topography should be already accounted for using spatial information in RegKrig.

- We could not include species identity for regKrig because, for less common species, there were not enough data to perform the regKrig imputations at high missingness rates. This is also one of the reasons to use methods that are designed to handle multivariate missing datasets, such as MICE or kNN.

R2#24. L292-296: Please discuss and reason the differences of the results for the different traits.

- Please note that our main objective is to compare different methods and different levels of environmental information for a diverse set of widely used traits. Discussing in detail differences in performance between traits would go much beyond the scope of the paper and would tend to blur our main, more general conclusions, which are already complex as a result of the comprehensive nature of the paper, in which we test many different aspects related to imputation performance. Nevertheless, we will try to discuss specific differences among traits when particularly relevant in the context of our study. For example, we have discussed the patterns we observe here with those by Vilà-Cabrera et al 2015, here in section 3.2:
- **Introducing auxiliary variables as predictors improved MICE performance substantially but these improvements were dependent on the specific predictor set and trait (Fig. 4, Fig. S10). Species identity increased KGE for all traits (Fig. 4) and it was the major predictor for Nmass, LMA and WD, as all MICE applications with species identity performed significantly better than those not including it (Fig. 4; P<0.05). Forest structure notably improved imputations for Hmax and for BL:AS at missingness ≥ 50% (P <0.05). Climate only produced significant increases in KGE (i.e., compare 'mice' with 'mice_c' in Fig. 4) for Hmax and WD (P<0.05). These results are in line with the distinct role of phylogeny and environmental variables as drivers of trait variability recently observed for the same tree species in the IEFC (Laforest-Lapointe et al. 2014; Vilà-Cabrera et al. 2015). After controlling for family (Pinaceae and Fagaceae), environmental variables only explained a substantial fraction of the variability for Hmax, they explained very little variability for LMA and WD and played no role in explaining Nmass (Vilà-Cabrera et al. 2015).**

- And in section 3.3:
- **In terms of KGE, 'mice_ctsp' was the best performing method at 50% missingness for all traits, together with 'Spmean' for Nmass,and LMA and with 'RegKrig' for Hmax (P>0.05 for comparisons between 'mice_ctsp', 'Spmean' and 'RegKrig for these traits). However, at 80% missingness, 'mice_ctsp' only ranked first for BL:AS whereas 'Spmean' showed the highest KGE for Nmass, LMA and WD and 'RegKrig' performed best for Hmax (Fig. 6, Fig. S12). These results are consistent with the prominent role of taxonomic identity in explaining variability in foliar traits and WD and with the higher predictive ability of environmental and spatial information in explaining Hmax (Vilà-Cabrera et al. 2015).**

R2#25. L305: "(data not shown here)". But why not adding a figure similar to Fig. S8, for the comparison of ∆cormat and dataset NRMSE between Spmean, mice_ctsp, kNN_ctsp and RegKrig in the supplementary as well?

- We had in fact this figure in a preliminary draft and we decided to leave it out. We will reduce supplementary materials in the revised version (from 32 pages to ~ 19 pages, as also suggested by reviewer #1) and will only provide the result of the corresponding statistical test,

and we will not add a new figure. Here's the text in its modified version although supplementary materials have not been reduced yet:

- **Kernel density plots and Kolmogorov-Smirnov tests (Fig. S18, S19) showed that MICE produced imputations (especially 'mice_ctsp') most consistent with observed distributions at all missingness levels (Fig. S18, S19). 'Spmean' and 'OrdKrig' imputations modified trait distributions substantially, while 'kNN_ctsp' and 'RegKrig' showed an intermediate performance, but generally far from that of 'mice_ctsp' (Fig. S22, Table S1, S2). 'Spmean' and kriging imputations also yielded larger Δcormat values compared to the rest of the methods (P<0.05).**

R2#26. L307-315: I would suggest to make this part chapter 3.4 (related to "Imputing traits for the main forest species in Catalonia 2.6)

IMPLICATIONS
R2#27. This part is well written. I would suggest to add also suggestion for improvement of data collections, e.g. trait collection should be better be accompanied by auxiliary information on coordinates, forest structure, etc.

- Thanks for this suggestion. We have included this idea in the 'Implications' section, where we also mention that databases of vegetation structure could help improve trait imputation:
- **For kNN, MICE and kriging imputations we have highlighted the key role of auxiliary variables as necessary covariates to yield reliable imputations in spatially explicit settings. This result calls for the inclusion of site-specific environmental variables associated with trait data in trait databases. The importance of covariates differed across traits, but, in addition to the expected influence of species, climate and topography in predicting trait values, we also showed a prominent role of forest structure for some traits. The ongoing development of global databases of vegetation structure (e.g. Dengler et al. 2014) will likely enable the incorporation of stand variables in trait imputation approaches using spatial and environmental information (Butler et al. 2017).**

New reference:

- **Butler, E. E., Datta, A., Flores-Moreno, H., Chen, M., Wythers, K. R., Fazayeli, F., Banerjee, A., Atkin, O. K., Kattge, J., Amiaud, B., Blonder, B., Boenisch, G., Bond-Lamberty, B., Brown, K. A., Byun, C., Campetella, G., Cerabolini, B. E. L., Cornelissen, J. H. C., Craine, J. M., Craven, D., Vries, F. T. de, Díaz, S., Domingues, T. F., Forey, E., González-Melo, A., Gross, N., Han, W., Hattingh, W. N., Hickler, T., Jansen, S., Kramer, K., Kraft, N. J. B., Kurokawa, H., Laughlin, D. C., Meir, P., Minden, V., Niinemets, Ü., Onoda, Y., Peñuelas, J., Read, Q., Sack, L., Schamp, B., Soudzilovskaia, N. A., Spasojevic, M. J., Sosinski, E., Thornton, P. E., Valladares, F., Bodegom, P. M. van, Williams, M., Wirth, C. and Reich, P. B.: Mapping local and global variability in plant trait distributions, PNAS, 114(51), E10937–E10946, doi:10.1073/pnas.1708984114, 2017.**

R2#28. L322: "Here we deal here" Please remove the extra "here".

- Done

R2#29. L357: Please change "practices" to "practiced".

- We think this is correct here, as we refer to the noun 'practices'.

SUPPLEMENTARY
R2#30. Please refer to the supplementary with their Fig. numbers and not chapter numbers.

- Done.

R2#31. Please make sure that all supplementary figures are mentioned in the main text. Currently this is not the case.

- Thanks for spotting this. We will check this once we reduce the figures in the supplementary materials.

R2#32. *Fig S7: trait mean (Mean) is missing in the plot.*

- This is because KGE is undefined for mean imputations, as there is no variability in the imputed variables.

R2#32. Fig S2: Please change "palant" to "plant".

- Done

R2#33. Fig S6: Please define "cor.matrix abs.error" in the figure and "correlation matrix error"in the caption. Do you mean ∆cormat? Please be consistent.

- Yes, it should be ∆cormat, thanks for noticing this. We will change it in the revised version.