

Author responses and proposed revision.

We thank the reviewer for their remarks, comments and suggestions. Below, our responses and proposed revisions are in blue text and preceded by **Authors:**.

Reviewer 2 remarks and our responses.

Comment 1 (General): This is a well-written manuscript that addresses a relevant problem, using advanced methods and a solid case study. The methodology is highly complex. Perhaps too complex. I lacked the necessary background to do a thorough review of the methodology. I strongly advise that an expert in Bayesian hierarchical modelling reviews the paper as well.

I did not review the entire paper but do have many comments. These are mostly minor to moderate but there are also some that in my view require a major revision.

Authors: We thank the reviewer for the detailed assessment. The title explicitly identifies compositional spatial modelling and calibrated joint uncertainty propagation as the paper’s focus, and the abstract specifies the use of Bayesian spatial modelling with INLA–SPDE. Given this stated scope and the reviewer’s detailed comments on the Methods, we understand that accepting the review invitation indicates the appropriate expertise to assess the manuscript. The comments identify aspects that need clearer explanation but no specific flaw that would require an additional specialist review. We will therefore improve the accessibility of the Methods by defining each stochastic component before use, explaining the validation and uncertainty propagation more clearly, and addressing the reviewer’s comments.

Comment 2 (L47): I don’t think you can claim that methods are now capable to “produce predictions accurate enough to inform management at sub-hectare resolution”. It all depends on the case at hand (number of observations, predictive power of covariates, degree of residual spatial autocorrelation). We often still find

that prediction uncertainty is large.

Authors: Thank you for this comment. We agree that grid spacing alone does not establish accuracy or management value. We will make the claim conditional on sampling density, covariate performance, residual spatial dependence and calibrated uncertainty.

Comment 3 (L49–51): I do not understand this sentence, how can measurement error be a single point of truth and how can this be used to estimate the trend? I guess the word “error” should not have been there? I also do not understand that you write that a predictive distribution is collapsed into a single residual variance. If uncertainty is ignored there is no predictive distribution and no variance, we only have a point prediction. Note also that DSM often reports maps of quantiles of the predictive distribution, so more than just a variance. And is this paper not also guilty of reducing a predictive distribution to a variance (see L250–251)?

Authors: We agree that the opening sentence is incorrect: the quantity treated as a point value is a spectroscopic prediction, not measurement error itself. We meant that substituting a point spectroscopic prediction for an unobserved laboratory value omits its predictive uncertainty. We also agree that the statement about reducing predictive distributions to a residual variance is too broad, because DSM may report distributional summaries such as quantile maps. In our workflow, NGBoost assumes a conditional Gaussian, which is fully specified by its mean and standard deviation, and both parameters enter the spatial stage: the mean as the trend covariate and the variance through a site-specific heteroscedastic observation weight. The distribution is therefore not collapsed to a single residual variance; instead, we incorporate its parameters through a Gaussian working likelihood. We will rewrite the opening argument, explain this interface and its approximation explicitly, and identify the final map intervals as empirical quantiles of the pooled spatial posterior draws.

Comment 4 (L51–52): This sentence also needs better explanation to be

understandable. What do you mean by “aggregated estimates”? Please be also more consistent throughout the paper in the use of “estimate” and “prediction”, these are not synonyms.

Authors: Thank you. We will define aggregation as the spatial averaging or summation of pixel-level predictions over a stated area. We will use ‘prediction’ for an unobserved soil value or mapped pixel, ‘estimate’ for a fitted parameter or area-level summary, and ‘measurement’ for an analytical observation.

Comment 5 (L59): I guess this is only one order of magnitude given that the cost reduction is factor 10?

Authors: Right. A tenfold reduction is one order of magnitude. We will correct the wording and separate the unit-cost comparison from any study-specific increase in sampling density.

Comment 6 (L100): Replace “by” by “in” (you refer to a publication, not to people)?

Authors: Both are grammatically valid because the citation identifies the authors who described the design. We will nevertheless use ‘described in...’ because it reads more smoothly and removes the ambiguity noted.

Comment 7 (L105): Measurements were composite samples from three cores. This is the only time in the entire paper that mention is made of the spatial “support” of observations and predictions. Given the strong focus on prediction uncertainty and uncertainty propagation, I would have expected that more attention was paid to this. It is well-known that uncertainty is highly dependent on the spatial support of the observations and predictions, with much lower uncertainty at larger spatial supports (e.g. compare point and block kriging variances). Perhaps the case study should not address uncertainty at ‘point’ support but at ‘block’ support? I am not asking that such analysis is done, but the Discussion could make clear that results strongly depend on the support and that users are likely more interested at predictions and

prediction uncertainties at block support than point support.

Authors: We agree that spatial support should be explicit. Our revision will make that clear. Each observation is a composite of three cores spaced 1 m apart, so its support is a small local footprint rather than a point. The maps represent the latent surface at 5 m grid nodes, not 5 m block averages, and we will state this and explain that block averages of the same posterior would have lower variance. The decision-relevant quantities in the paper are already at block support: the land-use comparisons and the farm aggregate are areal means over prediction pixels, computed within each posterior draw, which is why their credible intervals are much narrower than the per-pixel intervals. We will make that distinction explicit and separate latent-surface credible intervals from intervals for a new composite sample.

Comment 8 (L125 & L150): Be careful with the term ‘representative’. Better specify what you mean by this.

Authors: Thank you for this comment. We agree. We will keep ‘representative’ only where we can define the target population and give supporting evidence, and elsewhere we will describe the selection as it was done: conditioned Latin hypercube sampling on the mid-infrared spectra, chosen to span the spectral diversity of the measured population rather than to form a probability sample.

Comment 9 (L150): Is there any overlap between this subset of $n = 200$ samples and the subset of $n = 100$ samples mentioned in L125? And are these $n = 200$ samples evenly distributed over topsoil and subsoil?

Authors: Thank you. We will report the overlap between the two calibration sets and provide numbers for 0–10 and 10–30 cm.

Comment 10 (L166–174): It was not clear to me that data leakage was prevented. Usually one uses a nested CV-approach for that (splitting the dataset in three: calibration, validation and test), but here it seems the dataset was only split in two?

Authors: Nested cross-validation does not require three fixed datasets. It uses inner

folds for tuning and outer folds to estimate performance after model selection. It is not required for leakage-free tuning. In our 10-fold analysis, we predicted each fold from the other nine, and depths from a site remained in the same fold. Table 1 reports these internal model-development results. We then fitted the selected model to the full calibration set. We could add an outer validation in which we define held-out folds before selection. We will do so and clarify the approach in the revision.

Comment 11 (L175–181): Spectroscopic uncertainty was quantified using bootstrapping. The text is very condensed and it was not clear to me how it all worked in detail, but I fear that the procedure used severely underestimates the spectroscopic prediction uncertainty. Bootstrapping means creating datasets that have the same size as the original dataset, using sampling with replacement from the original dataset. If uncertainty is quantified from the variability between bootstrap results, then this uncertainty is only representing sampling uncertainty (i.e., the fact that only a finite dataset is available for model training and prediction). It will not capture other, more important sources of uncertainty, for example the limited capability of spectra to predict a soil property, even with an infinitely large training dataset. Note that the variance estimated by bootstrap sampling converges to zero as the sample size tends to infinity. In summary, a major concern I have is that the bootstrap method only quantifies the sampling uncertainty, which is only a (small) part of the overall prediction uncertainty. I would expect that this comes out from calculating the PICP. Authors use PIT, which I guess is quite similar, but they only do so for the final results (see L371–374). I ask that this is also done for just the spectroscopic uncertainty.

Authors: The bootstrap represents uncertainty from the finite calibration set and the fitted model, and its spread should shrink as the calibration set grows, sure. It does not include the prediction error for a new specimen. Comparing the two columns of Table 1, the mean between-bootstrap standard deviation is between 0.6 % and 46 % of the cross-validated RMSE for the same response, so the bootstrap spread

alone is well below the cross-validated error, as you anticipate. We will add the missing component by drawing a residual for each specimen at each bootstrap, calibrated to that response’s held-out residual distribution, so the propagated spectroscopic distribution includes both the model fit and the irreducible parts. We will then report spectroscopy-only 95 % coverage and mean interval width against the monitoring measurements, before and after the correction. Regarding the diagnostics, PICP gives coverage at a chosen interval, while PIT assesses calibration across the whole predictive distribution, which is why we used PIT for the final posterior; we will report coverage at the spectroscopic stage as you ask. We expect this correction to reallocate variance between stages rather than greatly widen the final intervals, because the omitted component is currently absorbed by the estimated observation variance in the spatial likelihood, which the variance decomposition shows is the largest positive share at every response and depth. We will report whether that expectation holds.

Comment 12 (L175–181): One other key problem is that the prediction errors associated with spectroscopic predictions are likely correlated (i.e., not statistically independent) between samples. It seems this study ignores this completely and assumes that they are all independent. The consequence of this is that spectroscopic uncertainty hardly propagates to DSM, because uncertainties largely cancel out. But is this realistic?

Authors: The bootstrap predictions are not independent between specimens: within each iteration, one fitted model predicts all specimens, so model-fit variation is shared and is not forced to cancel as independent noise during aggregation. The submitted analysis does not, however, represent the specimen-level residual component discussed in R2.11 or covariance within it; if those residuals are positively correlated, uncertainty in aggregated predictions would be underestimated. We will estimate depth-specific variograms from held-out spectroscopic residuals and incorporate their covariance where supported. This follows Viscarra Rossel et al. (2016), who tested

residual variograms before treating spectroscopic errors as independent and found no residual autocorrelation in their dataset.

Comment 13 (L178): Why call this a “discrete predictive distribution”? The variable of interest is continuous-numerical, so we need a continuous probability distribution. I think a better formulation is to write that you consider the 30 predictions a random sample from the continuous conditional distribution of the variable of interest. And is not $B = 30$ a much too small number?

Authors: We agree that the property is continuous. Our phrasing was careless. We will describe the 30 retained values as a Monte Carlo sample from its conditional predictive distribution, not as a discrete distribution. Thirty bootstraps ($B = 30$) is within the adequate range for pooling (In his seminal 1996 paper, Leo Breimman on ‘Bagging Predictors,’ wrote: ‘More than 25 bootstrap replicates is love’s labour lost.’. Note that the quantiles in our analysis come from 15,000 pooled draws.

Comment 14 (L196–198): I realise that reduction of number of covariates using PCA is commonly done but I am sceptic about the validity of that. Variables that are not correlated with others will not contribute much to the first PCs and so will not be retained if you only keep PCs that have at least 10% of the variance. But such variables might well have strong explanatory power, so why delete them without checking? Many machine learning methods have in-built protections against overfitting, so I would delete this step.

Authors: We agree that explained covariate variance is not a criterion for predictive relevance. In our case, we applied PCA only to the strongly correlated multiscale terrain attributes; the proximal-sensing and ratio covariates entered in their original form. We will report the existing cross-validated comparison between the raw-terrain and PCA-terrain candidate sets and retain the better-supported specification.

Comment 15 (L204–211): Again considerable tuning, how did you prevent data/information leakage? This requires that test data were removed before tuning,

but I did not see that mentioned yet.

Authors: Correct, the submitted workflow did not reserve an independent outer test set before covariate selection and hyperparameter tuning. Those steps used cross-validation, so they report internal model-development results. In the revision, we will perform all selection and tuning within each outer training set and evaluate performance only on the corresponding held-out outer fold.

Comment 16 (Section 2.5): I am not an expert in compositional data so could not verify the soundness of this section. I thought the problem was easier and that log-ratio transformation already does the trick. For instance, when applied to soil texture (sand, silt, clay) we can use the ALR transform to reduce the three fractions to two variables, do the modelling and prediction on those two, and later back-transform to the original scale (such as done in Poggio et al. 2021)?

Authors: ILR is a log-ratio transformation and performs exactly the reduction you describe, from three parts to two freely varying coordinates. ALR would also be valid, and the Poggio et al. approach you cite is a reasonable alternative. We chose ILR for two reasons: its coordinates are orthonormal, so an ordinary Gaussian model on them is appropriate without a further correction, and the basis does not privilege one fraction as a denominator. The particular basis we use gives two interpretable coordinates that balance the labile pool against the protected pool and mineral-associated against residual carbon, which the Results then use directly.

Comment 17 (L232): Is it realistic to assume that $\log(\text{TOC})$ and $\log(\text{TIC})$ are independent? And even if this assumption is realistic, I fear that this does not automatically imply that the prediction errors of these two variables will also be independent. Please check.

Authors: Fitting TOC and TIC separately does not require the observed variables to be marginally independent, but it does impose conditional independence on their residual processes. We will estimate the residual covariance from held-out predictions,

include it when deriving TC and TIC/TC, and report the resulting change in their uncertainty.

Comment 18 (L236): It was not clear to me what “multiple-imputation” means. It is only explained much later, in Section 2.7, but perhaps this could be clarified earlier on. As I understand it, it is no more than a Monte Carlo loop over the $B = 30$ bootstrap samples, is this correct?

Authors: Yes, correct. It is a Monte Carlo loop over the 30 complete bootstrap realisations, with the trend and spatial models refitted on each one. We use the multiple-imputation language because each realisation supplies a complete plausible response dataset and because the within- and between-realisation variances are then combined by Rubin’s rules. We will define the procedure where it first appears instead of in Section 2.7, and use it consistently throughout.

Comment 19 (L241–242): Here you assume that spectroscopic prediction errors are statistically independent, which is likely unrealistic. See also an earlier comment. At least this assumption should be made explicit. If I understood correctly, spectroscopic uncertainty is not accounted for in trend modelling with NGBoost, but only in the kriging part through Eq. 3. Perhaps it should indeed not be included in trend modelling (part 2) because you take an approach where you sample from the conditional distribution of the ‘true’ response variable, given spectroscopic uncertainty (i.e., stochastic simulation), but if that is true we also should not account for spectroscopic uncertainty in the spatial modelling (part 3) (i.e., there should be no variance inflation of the ϵ in Eq. 2)? I also did not understand the rationale behind Eq. 3. Is this equation theoretically derived or a heuristic approach?

Authors: We’ll respond to each point in turn. On the independence comment, as in our response to R2.12, the analysis does not assume independent spectroscopic errors across specimens because a single fitted bootstrap model predicts every specimen within an iteration; we will make this explicit (see also above). Regarding where

spectroscopic uncertainty enters, it enters before the trend model rather than only in the spatial part: each bootstrap realisation supplies a new response dataset, and we refit both NGBoost and INLA on it, which is the stochastic simulation. No second count occurs: Equation 3 contains NGBoost’s conditional predictive variance, which is different from the between-bootstrap spectroscopic variance, so inflating the observation variance for trend uncertainty does not reintroduce the spectroscopic component. On the last question, Equation 3 is a first-order errors-in-variables likelihood rather than a derived result. We will clarify this.

Comment 20 (L241): It was not clear to me why INLA-SPDE was used instead of kriging. What is the added value? Maybe this is needed to take a Bayesian approach, but where does the Bayesian come in and is this justified (is the added complexity worth the effort)? Which parameters of the spatial model were treated as stochastic and how large were the variances of their posteriors? If small it is not worth the effort. I also noticed that not much effort was placed in defining priors. Please note I am not an expert in Bayesian hierarchical modelling and could not check the validity of these parts (i.e. Section 2.6.3).

Authors: The ‘added value’ is that INLA–SPDE provides a joint posterior for the trend coefficients, spatial range and variance, residual precision and latent field. Residual kriging on fitted trend and variogram parameters needs uncertainties to be added separately. The stochastic parameters are the intercept, the slope on the trend covariate, the practical range, the marginal standard deviation of the spatial field, the residual precision and the latent field at every mesh node. The variances of the posteriors were not small: at 0–10 cm the posterior range for TOC is 229 m with a 95 % interval of 43 to 646 m, and for TIC it is 9,895 m with an interval of 453 to 59,581 m, so a single plug-in value would misstate the spatial dependence badly for TIC in particular. On priors, we used penalised-complexity priors on the range and the marginal standard deviation, with $\Pr(\text{range} < 50 \text{ m}) = 0.05$ and $\Pr(\sigma > 1) = 0.05$, which shrink the field towards no spatial structure unless the data support it;

we will give that reasoning in the revision.

Comment 21 (L248–252): It was not clear to me whether you used a stochastic or deterministic trend model. L249 mentions that each fit returns a predictive distribution but is this because the input to NGBoost is stochastic (i.e., the 30 bootstrap predictions from the spectroscopic model) or is this because NGBoost accounts for its own prediction uncertainty? If the latter, how is this done? And why is only the standard deviation of the full predictive distribution used? And why is NGBoost’s per-point predictive variance only used to down-weight the fit of the Bayesian spatial model through an increase of the residual variance in Eq. 3 (L282–284)? Can this approach be mathematically derived?

Authors: It comes from both the stochastic and deterministic, but at different levels. Across iterations, the input is stochastic, since each bootstrap realisation changes the response NGBoost is fitted to. Within an iteration, NGBoost is probabilistic: it updates the parameters of a conditional Gaussian by natural-gradient steps, so a single fit returns a mean and a standard deviation at every point. We use only those two moments, and you are right that this is a first-order treatment rather than full propagation of the NGBoost distribution. In INLA the mean is used as the trend covariate, and the standard deviation as the observation weighting at sample locations, so observations with more uncertain trends receive less weight in the spatial fit. We do not sample the NGBoost distribution at prediction pixels. We will describe these more explicitly in the revision.

Comment 22 (L261–262): I think it is bad science to remove outliers only because they are influential and distort the trend. In my view it is only justified to remove outliers if there is sufficient evidence that something went wrong during measurement or processing (e.g. negative SOC) or that the observation location does not belong to the population of interest (e.g. the location is on a road or manure heap).

Authors: Fortunately, we did not simply remove outliers so that the model fit our

data. As described, our procedure did not remove observations merely because their response values were extreme. We calculated Cook’s distance using a ridge-regression surrogate and repeated it over 30 bootstrap resamples. An observation was flagged only when its Cook’s distance exceeded the sample-size-based threshold in more than 50 % of bootstrap resamples. Between 6 and 9 sites per response and depth met this persistence criterion. The phrase ‘removed when they distorted the trend’ does not adequately describe our objective approach.

Comment 23 (L263–265): I do not understand why this prevents overfitting, but this might be because I am not an expert in these methods. Explanation is needed. Maybe it is explained in Section 2.7?

Authors: That statement was specific: the NGBoost covariate for each location is predicted by a model that was not fitted to that location. This prevents an artificially close in-sample trend fit from suppressing the residual and spatial variation; it does not prevent overfitting generally. We will revise the sentence to state this limited purpose and ensure that the revised analysis uses out-of-fold predictions.

Comment 24 (L217): Much of Section 2.6.3 is beyond my knowledge and so I could not check it. I am still not convinced why INLA-SPDE was needed and could not be replaced by the much easier residual kriging. The added value must come from making parameters of the random field stochastic, but is not parameter uncertainty negligibly small in this study? Could you compare your results with those obtained using a simple residual kriging approach?

Authors: We have explained the choice of INLA–SPDE and shown that parameter uncertainty is not negligible in R2.20. We do not propose adding a residual-kriging benchmark because the paper’s objective is to propagate draw-wise uncertainty through the compositional back-transformation and derived carbon quantities, not to compare spatial predictors. Conventional residual kriging would condition on the fitted trend and variogram parameters and therefore would not provide a like-for-like

uncertainty analysis. Instead, we will make this rationale explicit.

Comment 25 (L278): Why not force β_0 to 0 and β_1 to 1? I hope you did allow $w(s)$ to include a nugget (short-distance variability)? Authors will know that short-distance spatial variation is often large for soil properties, so nugget is more than random measurement error. Regrettably, the equation in L296 suggests that this was not included.

Authors: We estimated β_0 and β_1 rather than fixing them because $\beta_0 = 0$ and $\beta_1 = 1$ would assume that the NGBoost trend is already calibrated. We include a nugget as the non-spatial residual $\epsilon(s)$ in Eq. 2 to represent short-distance variation and unresolved measurement and sampling error. The expression at L296 defines only the smooth Matérn field, so it omits the nugget; we will clarify and report this.

Comment 26 (L331): Why replace B by M ?

Authors: M is the number of multiple-imputation iterations, whereas B is the number of spectroscopic bootstrap realisations. Both equal 30 in this analysis because each imputation uses one bootstrap realisation, but they refer to different stages rather than one symbol replacing the other. We will make this distinction explicit when M is introduced.

Comment 27 (L331–332): Here I guess independence between specimens was assumed, which might not be realistic.

Authors: Specimens share the same fitted bootstrap model within an iteration, so independence is not assumed for that component. Residual spatial dependence is separate and is not represented explicitly. As described in R2.12, we will examine horizontal residual variograms separately for the 0–10, 10–30 and integrated 0–30 cm models; the integrated model does not require a cross-depth covariance model, of course.

Comment 28 (L342–347): Text was not entirely clear to me but am I right that you

used a stochastic simulation (instead of prediction) approach to derive quantiles of the back-transformed variables? It seems you did because L334 mentions that you took 500 draws from the posterior distribution. I think this makes sense, and it would be interesting to separate the uncertainty resulting from spectroscopic uncertainty from that of other sources, and I think you do so in Section 2.8. I can see that separating out the spectroscopic uncertainty from the $M \times 500$ draws is easy, but I do not see how you can separate out the other four sources uncertainty (as mentioned in L352) from these draws. Is this clearly explained somewhere in the paper?

Authors: Yes. We use posterior simulations to obtain empirical quantiles after back-transformation. Spectroscopic uncertainty is the between-imputation component. We do not separate the other four components from the pooled back-transformed draws; within each imputation, we calculate them on the transformed scale as the trend variance, spatial-field variance, twice their covariance, and residual variance. This is not sufficiently clear in the submitted manuscript, so we will state this sequence explicitly in Section 2.8 and clarify that the decomposition applies on the transformed scale.

Comment 29 (L364–365): Not clear what CV you refer to here and that data leakage is prevented. As mentioned above, I would assume a nested cross-validation where outer held-out folds intended for performance evaluation are created right at the start, not used in any way during modelling, also not for model selection and hyperparameter tuning.

Authors: The cross-validation at L364–365 generates out-of-fold trend predictions for the spatial model; it is part of model fitting. The CPO/PIT analysis is likewise an internal diagnostic, and neither encloses the earlier feature selection and tuning. As per our response to R2.10, in the revision, we will report the outer held-out folds defined before any selection or tuning.

Comment 30 (L389): Were these 59 sites a subset of the sites previously sampled or

were these a new set of sites? I guess that for this to be a meaningful check these should be newly sampled locations, ideally a probability sample from the study area. L392 suggests this was not the case.

Authors: As stated at lines 389–394 of the submitted ms, these are 59 monitoring sites resampled in a subsequent year at or near the original locations; lines 593–594 specify an interval of approximately 18 months. We agree that a new probability sample would provide a stronger test of spatial generalisation. The 59 revisited sites nevertheless provide a useful test of temporal prediction at monitored locations.

Comment 31 (L424–706): I did not review this part of the paper (I had already spent much time on the review and made many comments).

Authors: We thank the reviewer for the time invested and for the thorough assessment of the Introduction and Methods, which contain the key technical elements of the paper. The title makes the paper’s focus on compositional spatial modelling and calibrated joint uncertainty propagation explicit, and the reviewer’s detailed comments demonstrate careful engagement with these methods.