*Dear Editor*

*Please find enclosed our revised manuscript and the detailed responses to the reviewers. Unfortunately, the responses turned out to be a bit complicated, because all comments by referee #1 and #3 were based on the original version of the manuscript that was submitted to Biogeosciences on September 26th 2014. After the initial review we included the first comments of the three referees and changed our manuscript accordingly. This revised version was then published as a discussion paper in "Biogeosciences Discussion" on January the 9th 2015. Therefore many comments by the referees were already addressed at an earlier stage. Obviously, there was a communication problem so that these referees used the older version for their detailed reviews. However, we tried to document now all the previous changes that were already included and of course also the additional changes that appeared to be necessary. Inevitably, our responses to all points raised by the reviewers ended up being quite long.*

*Before we address the individual issues raised by the reviewers, we would like to point out that we had never intended with this manuscript to use NIRS to develop a mechanistic understanding of the Hedley-P fractions, nor did we want to create a globally valid model to predict the different Hedly-P with NIRS. Some of the comments received seem to indicate that some reviewers might have thought that this was our intention or would have liked the manuscript to achieve just that. That might have triggered some of the comments, which we find difficult to address in the context of our study.*

*Below we have reproduced the greatly appreciated reviewers comments and inserted our responses in italics.*

**Referee 1**

General Comments:

1)

While the authors clearly state that phosphate groups are not detectable by NIRS, it remains unclear whether or not spectra arise from ester bonds or other bonds associated with organic P fractions. Similarly, the reader is not able to follow which other soil properties might be linked to P fractions in terms of NIR spectra and how this could be explained in a more mechanistic way. Based on existing applications (that are able to focus on C-H vs. C-O or C-OH) one could at least come up with a very rough concept.

*Response:*

*To address the issue of a possible relationship between selected spectral regions representing certain types of bonds and P fractions, we included the following paragraph in our Discussion section 4.2 Calibration of organic and inorganic P fractions*

*In our study, we were not able to identify spectral regions to be specific for a P signal as was found in other studies (Malley et al., 2004). Therefore we had also assessed, if focusing on typical NIR spectral regions for C-H, N-H and O-H bonds could influence NIRS model quality. The organic residual which is connected with the phosphate molecule could be dominated by CH, NH, OH bonds or a mixture of them. For this purpose we compared NIRS models based on optimized spectral regions (automated procedure by OPUS software), on the whole*

*spectral range and on specific spectral regions, which are known to represent C-H, N-H and O-H bonds (Conzen, 2005). We found that in all cases, the OPUS-software optimized spectral selection yielded superior models followed by models covering the whole spectral area. Models for selected bonds were in all cases of substantially lower quality, and were thus not presented in detail. The best results based on $r^2$ and RPD were obtained for O-H bonds for the Po-$HCO_3$ and P-$HCL_{conc}$ fractions. This was followed by models focusing on C-H bonds and. The lowest quality models were obtained for models focusing on N-H bonds.*

2)

As NIRS is intended to reduce the number of chemical analyses (still necessary for calibration), it would be highly useful to have an estimate of the mean error (in mass P/mass sample) associated with the predicted concentrations of P fractions of the validation subset (not included in model establishment). This would be comparable to common precision/trueness parameters used for quality assurance in wet chemistry analyses.

*Response:*

*We agree that error estimates are helpful and important additional information. Therefore we included an example for standard errors for the P NaOH fraction in the wet chemical analysis in the text. However, since this information was not related to our original questions, we did not present this for all fractions. A table (Table S1) including all standard errors as well as mean values and standard deviation of all P fractions for a repeatedly measured soil sample was additionally placed in the supplement.*

3) Link between P compounds and spectra; standards to be analyzed (e.g. monoesters, diesters etc.)

*Response:*

*A chemical characterization of the P-fractions, in particularly to distinguish between the various organic P forms with NIR spectroscopy, was never in the focus of our study. NIR spectroscopy is not able to distinguish between phosphate monoesters and phosphate diesters. For this purpose, more suitable methods like the NMR spectroscopy are available (Condron et al., 2005). Our approach was developed to predict the P content of the Hedley P fraction directly from the solid sample and not the characterization of extracts.*

Specific comments:

*The numbers of pages and lines correspond with those in the original version of the manuscript that was submitted to Biogeosciences on September 26th 2014*

| **page/line** | **referee comment** | **our comment** |
|---|---|---|
| P1 L23-25 | Not only R2 is relevant, but also whether or not the regressions were significant. I assume not all regressions were significant. If so, please state the proportion of significant regressions as well | All regressions were significant. |
| P1 L26 | “homogeneity” in terms of? Range of soil properties? Range of P concentrations? Soil types? Specify! | Specified: soil properties |

| | | |
|---|---|---|
| P2 L13 | This is controversially discussed, please add constraints of estimates and other views as well. | We deleted the controversial peak date of 2030 and referenced the publication by Edixhoven et al. 2013. that is critical of the P peak hypothesis. We decided not to add constraints of estimates and other views, since this point only served to provide some background and motivation for the study. |
| P2 L20 | “diminish” because of? Timber harvest? Erosion? Be more specific and add evidence provided by other studies. | We added: through processes such as erosion and timber harvest |
| P2 L22 | The initial idea of the fate of P during ecosystem development and pedogenesis dates back to 1976 (Walker, T.W., and J.K. Syers. 1976. Fate of phosphorus during pedogenesis. Geoderma 15: 1-19.). Should be acknowledged here as well. | The reference to Walker et al. (1976) was added. |
| P4 L32 | As you state hypotheses, these will either be verified or falsified. This is not possible for Hypothesis 1 unless you define criteria associated with “sufficiently well”. Based on which criteria and thresholds do you rate a prediction as “good” or “not sufficient”? | We added the sentence to hypothesis 1: The criteria by which the quality of NIRS models is quantified, will be introduced in the Material and Methods section |
| P5 L1-3 | Again more specific: “quality” in terms of? | See our comment to P4 L32 |
| P5 L21-35 | I would like to see quantitative measures of the selection procedure. What criteria did you use to come up with “typical brown earths” as the final subset (apart from the fact that n = 84 is near to the 100 samples required for model development)? You state no correlation between total P and 25 individual P fractions or other soil properties such as total C, N and pH. But how could correlations aid in selecting subsets? Furthermore, your statements “less heterogeneous” (l. 28) and “still heterogeneous” lack a quantitative evaluation. What is the criterion for heterogeneous versus homogeneous data sets? | Typical Brown earth is a soil type of the German soil taxonomy. The German Soil taxonomy is based on expert assessment and not on quantitative measures like for example the World Reference Base for soil resources. The classification of the soils used in the BZE survey was done by soil experts of the state forest research stations.<br>Homogeneity in our case is related to a low degree of variation in soil properties and in particularly of organic P compounds which are supposed to be better predictable. |
| P6 L5 | On the preceding page, please add approximate area covered by the BZE data set. Furthermore, add mean distance between two sites for the Chinese data set (maybe also for the BZE data set). | The area covered by the BZE samples was a region of approximately 200 km width and 700 km length starting in the southwestern part of Germany and reached up to a line Hannover/Berlin as a northern border. The BZE plots were part of the German Forest Soil inventory net, based on a grid size of 8 x 8 km.<br>The Information was added in the material and method section<br><br>The Chinese data set does not consist of two sites! As has been stated in the |

| | | |
|---|---|---|
| | | manuscript, soils were sampled in one large Nature reserve. Close proximity means on the same slope as the stated three Study plots, up to 100 m distance. The mean distance between all 27 study plots is 3.40 km, with Min = 0.04 km and Max = 8.98 km. This information was added. |
| P6 L8-11 | I do not understand the procedure here: the three (four) topmost diagnostic horizons were located deeper than 47 cm? Or did you select those diagnostic horizons only that did not duplicate the depth increments mentioned before? Please clarify | The sentence was rephrased and clarified. |
| P6 L11 | The tree cluster samples were taken as replicate samples whereas (as far as I understood) all samples described before represent composite soil samples. Please add a critical remark concerning this difference (e.g. pseudoreplicates). | We added: “Each of the four samples of tree clusters, were also composite samples from three cores each. Each composite sample represents different conditions within the cluster; they were collected at the base of individual trees belonging to different or the same tree species and in the center of a triangle between these trees. We cannot rule out a spatial correlation between these samples.” |
| P6 L28-P7 L2 | Add a critical remark on how different sample preparation procedures might affect the relationship between spectra and wet chemical extraction procedures. | We added a reference for NIRS dependency on sieved/ground soil samples. |
| P9 L6-7 | You state functional groups, but show bonds: O-H no functional group (-OH); C-H/-CH3 or –COOH or...; N-H/-NH2. Would you like to refer to the bonds? If so, would this include C for N and O as well (C-N-H; C-O-)? Without such information it is difficult to guess how NIRS could be adapted for P fractions. | We replaced the expression “functional groups” with “bonds”, since the bonds were stimulated not functional groups. |
| P10 L15-17 | Contradiction to pre-selection of typical brown earth (5/21-35). You state that you tested different groupings including soil type. This would not be possible if you pre-selected “typical brown earth” only!? Finally, after the confusing statements on inclusion or exclusion of data subsets (starting five pages before!), the reader is relieved to find the reasoning…(10/23-30). These should precede any statements on in-/exclusion of data to ease readability. Please restructure this section accordingly and rephrase if necessary. | We did not pre-select typical brown earths. The selection of “typical brown earth” was the result a selection process that is now documented in the methods section. |
| P10 L17-20 | Above you stated that NIRS measurements of P are possible BECAUSE of correlations with soil | We rephrased this section. |

| | | |
|---|---|---|
| | organic matter properties. As organic P forms part of SOM, I do not understand what is meant by “original properties of soil P”. Please clarify. | |
| P11 L9 | For readers not familiar with model evaluation please explain how to interpret the RPD. 11/12-13 implies that high RPDs are desirable but why should one aim at high standard errors of prediction used as numerator in the ratio calculation? | This was indeed an error in the depicted formula. Standard error of prediction was of course the denominator.<br>For the explanation of the RPD values, several references were included. |
| P11 L28-P12 L11 | Comparison with variables (pH, C, N) used to classify the data sets as heterogeneous/homogeneous? | The BZE brown earth samples showed a smaller variation of these variables as the total BZE sample set. The BEF China sample set showed the smallest variation for these variables. |
| P12 L7-9 | Please add the proportion of the organic NaOH P fraction relative to total P to enable the reader to judge the relevance of these high Po concentrations. | The proportion of Po NaOH fraction relative to total P was added (BZE = 29%; BZE BB = 31%; BEF = 37%) |
| P12 L16 | State range of $R^2$ and RPD for models of the fractions the at least. | A range of all $R^2$ (0.08-0.68) and RPD (1.04-1.74) values for all global models was added. |
| P13 L26-P14 L7 | You did not state it explicitly in the methods (add information 10/19), but here as well as in Fig. 7 you mention the Spearman Rank correlation coefficient as independent variable. For continuous and metric C or N and P concentrations the Pearson’s correlation coefficient is to be preferred. What was the reasoning for choosing a non-parametric coefficient? Irrespectively, the two variables used for the regression are differently detailed: i) the goodness of fit represents the percentage of data of data that can be predicted by the calibration model; ii) any correlation coefficient will yield the “strength” of the relationship between two variables be it an approximation of the slope of a regression (Pearson) or the relative position if ordering the data from low to high values (Spearman). However, a correlation coefficient of 1 does not mean that the data can be predicted well, because these coefficients are not necessarily related to the scatter in the data. For example, a correlation coefficient of 1 could arise despite the that fact that data points scatter greatly along the 1:1 line. Therefore, no meaningful interpretation can be derived from Figure 7. If the authors used a regression between concentrations of Ct or Nt and P concentrations, the resulting | According to Referee #1 and #3 we replaced figure 7. Now it shows the relationship between goodness of fit values for NIRS models and the relationships between soil C and N and P in different fractions. |

| | | |
|---|---|---|
| | R2 might be used as an independent variable in Figure 7. Delete this paragraph and rewrite it according to the new results. There are already six figures in this manuscript, therefore, a list of results without a figure is sufficient. | |
| P14 L24-26 | Stated at this prominent position (concluding sentence of a paragraph) I would like to see some details of this quality check (coefficient of variation or mean difference between repeatedly analyzed samples or similar) without displaying all the data. | We included as example the standard errors of repeated measured samples of the P NaOH soluble fractions. Values for all fractions are presented in the supplementary material. |
| P15 L28-30 | Given the fact that the preceding sentences repeat information provided in the introduction already and thus, do not lead to an increased knowledge before and after conducting the measurements, I would like to read an educated guess how the different P compounds could influence the spectra. Why should monoesters result in spectra different from that of diesters? | See response to referee #1 general comment 1 |
| P16 L1-6 | I do not agree that this conclusion can be derived from the results because Fig.7 does not allow for a meaningful interpretation (see comment on Fig. 7). | We replaced figure 7. Now it shows the relationship between goodness of fit values for NIRS models and the relationships between soil C and N and P in different fractions. We believe that these relationships can be interpreted to support this conclusion. |
| P16 L6-9 | I am lost now: at several places throughout the manuscript, it is stated that the P-O bond cannot be characterized by NIRS and that P compounds must be detected indirectly based on other soil properties with organic matter being the most promising proxy because of the influence of functional groups/bonds in organic molecules (e.g.,9/10-11). If this is true, what did lead to “sufficiently good” predictions of P fractions and pools in your study? | See response to referee #1 general comment 1 |
| P16 L23-25 | Not all studies listed above stated increasing prediction quality with increasing heterogeneity. Order preceding list of studies accordingly and evaluate which studies agree/disagree with your findings and, most importantly, why there are similarities/differences | The order was correct. Only the conclusion was not clearly described. We rephrased this section and similarities and differences are clearly stated |
| P16 L29-P17 L2 | The BZE brown earth model deviates only slightly from the BZE model. Did this improvement lead to a higher class assigned to model quality in any case? If not, please tune down the statement on improvement of the model. | The improvement in model quality led only in one case to a higher quality class. We had clarified this already in our previously revised manuscript version which was submitted for interactive discussion, after the initial quick reports,. |
| P17 L3-5 | Without any chemical information on the | See response to general comment 3 |

| | | |
|---|---|---|
| | link between P compounds and NIR spectra, the reader is not able to follow this paragraph. How might spectra be related to P compounds? See comments on chemical structures above. | |

Technical comments

| page/line | referee comment | our comment |
|---|---|---|
| P1 L1 | “near-infrared”; “phosphorus fractions | Changed |
| P1 L15 | “fractionation of…into fractions”; awkward phrasing, please rephrase | We changed “fractionation” to “analysis” |
| P1 L27 | “Meaningful models” | Some of the models obtained are useful for NIRS modelling and therefore are rather “useful” than just “meaningful” |
| P2 L2 | “useful” might depend on the view point. Please phrase more specifically what is meant by “useful” (e.g. match between NIRS data and results of chemical extraction). | Changed to usable |
| P2 L25 | “monitoring the” (delete “of”) | Changed |
| P2 L30 | hyphen in “plant-available P”; check throughout manuscript | Changed |
| P3 L6 | “dynamic” | Changed |
| P3 L8 | (relevance…) “has been” | Changed |
| P3 L14 | “Hedley fractionation” (without hyphen) | Changed |
| P3 L16 | red dot? | Changed |
| P3 L16 | “Hedley P” (without hyphen); check throughput manuscript (e.g. 4/29) | Changed |
| P3 L17 | “less expensive” or “cheaper” | Changed to less expansive |
| P3 L25 | “2010).” | Changed |
| P3 L26 | “Furthermore,” | Changed |
| P3 L27 | “which commonly constitute the major portion” | Changed |
| P3 L28 | As it is phrased now, the first part of the sentence is contrary to the second part. The spectral information cannot be complex/heterogeneous and uniform at the same time. What does “its” refer to? Better state an own subject for the first part of the sentence. | “of chemical and physical soil parameters” was added therefore clarifying that first part is referring to chemical and physical soil parameters and the second part to spectral information, which can be different. |
| P4 L7 | “<2mm” | Changed |
| P4 L17 | It would be logical if high variation in chemical composition was associated with high spectral variation. If this was the case, please rephrase (“chemical composition associated with high spectral variation”). | Changed accordingly |
| P4 L26 | “soil P” (no hyphen) | Changed |
| P4 L28-30 | awkward phrasing; merge to one sentence. | Second sentence was changed, therefore there is no need for merging them. |
| P5 L16 | “grouped by soil type” | Changed |

| | | |
|---|---|---|
| P5 L26 | “and, “ | Changed |
| P6 L1 | “research project” | Changed |
| P6 L25 | “measured in” | Changed |
| P6 L31 | “< 2mm” | Changed |
| P6 L32 | “the determination of P fractionation in soil.” | The comment makes no sense. Either „for P fractionation“ or “the determination of P fractions“ the second option was included |
| P7 L8 | “authors discussed” | Changed |
| P7 L17 | tense: “considered”, “used” | Changed |
| P7 L27 | “2008).” | Changed |
| P7 L29 | consistent hyphens | Changed |
| P7 L31 | insert Po in parantheses | Changed |
| P8 L2 | “the resin” | Changed |
| P8 L22-23 | “the Hedley fractionation method” | Changed |
| P8 L33 | “did not” | Changed |
| P9 L3 | “bending, and” | Changed |
| P10 L6 | “This was carried” | Changed |
| P10 L6-10 | I do not understand the last part of the sentence (“second to min and max values”)? Split sentence and rephrase. | We split the sentence |
| P10 L11 | “optimize”? Be consistent throughout manuscript (e.g. characterize 1/16) | We rephrased section and deleted “optimize” |
| P10 L31 | “Set 3” (incl. space); | Changed |
| P11 L2 | “Set 2”, “Sets 1 and 2”; | Changed |
| P11 L4 | “Set 4”; 11/7: “Set 4” | Changed |
| P11 L11 | “was discussed” | Changed |
| P12 L11 | Accuracy by definition includes precision and trueness of measurements (the opposite for inaccuracy, of course). I cannot see how low concentrations fit in either of these meanings. Maybe you refer to the limit of detection or similar? Rephrase. | Was rephrased |
| P13 L13 | “(Fig. 5)” (space) | Changed |
| P13 L29 | “(Fig. 7)” | Changed |
| P14 L21-22 | awkward wording (“can make it difficult”), rephrase. | Rephrased. Deleted “can make it difficult” |
| P15 L2-16 | Pure description without interpretation, move to (method)/results section. | We skipped the descriptive part in this section but kept the parts relevant for the discussion. |
| P17 L10 | Add references on knowledge vs. knowledge gaps concerning inorganic P. | We changed this section and added the sentence “In contrast, the inorganic P forms represented in the distinct P fractions are more specific in their chemical nature and well known (Stevenson and Cole, 1999, Tiessen and Moir, 2008)”. |
| P17 L12-14 | Awkward sentence, rephrase. | Rephrased |
| Table 1 | Replace comma by dots (2 times); reduce decimal places (one) for skewness and curtosis. | Done |
| Table 2 | “carbon”; “nitrogen”; Be consistent with Table 1: one decimal place only. | Done |
| Table 3 | “parameters” | Changed |

| | | |
|---|---|---|
| Figure 1 | too many figures in manuscript; this procedure is described well and easy to understand in the method section. Delete Figure 1. | We think that this figure is helpful for readers not familiar with the Hedley method. It is also helpful for understanding how the P pools are combined. Since the other reviewers did not ask to remove this figure, we kept it in the manuscript. |
| Figure 3 | This might represent one basis for a quantitatively-driven data subset selection. However, neither the method details nor results were described (Which variables are included in the PCA?; Which procedure was used to create the n-dimensional space [e.g. varimax rotation]?; How many principal components were derived?; Which proportion of variance was explained by the two components displayed in Figure 3). Why should this principAL (please change spelling in Figure 3 and caption accordingly) be preliminary as stated in the methods section? | The spelling in the figure and caption was changed as suggested.<br>The PCA is preliminary as stated in the methods section, since it is only calculated on the basis of the spectra. No other variables were included. This is described in the Method section. The procedure is an automatic function within the software. Five principal components were derived. The number of components was added to the caption of Figure 3.<br>According to our knowledge, it is not possible to calculate the proportion of variance with the software used, since this function is only designed to define outliers. |
| Figure 4 and 5 | How could negative concentrations be predicted? The model should set these to zero!? | The model results are mathematical calculations and therefore can become negative. This is an indicator of insufficient quality of a model. With better model quality such negative values disappear, which could be observed in figure 06 with the validation results of the BEF samples. If we delete the negative values, the results appear better than they are. Therefore we decided to keep the negative values. |
| Fig. 4 | I find it strange that the calibration and validation figures both use measured P concentrations as independent variables. For the validation data set, the modelled values were not derived from P concentrations of the wet chemistry protocol (“measured P”) but directly from the NIR spectra. Therefore, the modelled P concentrations should represent independent values plotted at the x axis. | Done |
| Fig. 4/Table 3 | Redundant data display; either as a figure or a table, but not both. | We skipped the calibration results in figure 4 since they are indeed also available in table 3 |

**Referee 2**

Referee 2 is referring to the manuscript published in Biogeoscience Discussion 12, 555–592, 2015

Comment 1.

The selection subsample sets and the procedures used in the calibration/validation or cross-validation need to be much better explained and justified. The authors described four different subsample sets used for calibration (p. 568). Here much more information and justification is required. Important issues are for each of the four subsample sets - what were the sample numbers? - which depth ranges were considered? - the samples are considered to be representative for which population? did the authors make sure that no pseudoreplicates (i.e. in case of calibration/validation: samples from one site were NOT in the calibration and validation data set or i.e. in case of cross-validation: the authors did NOT carry out a leave-one-out-cross-validation and made sure that samples from one site were not in different groups) were present and thus no overoptimistic results? The authors are urged to follow the recommendations by Brown et al. (2005, Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma 129, 251–267).

*Response: (Since Rev. 2 commented on the version of the manuscript published in the Biogeosciences Discussions, we refer in our responses to the page numbers of this document)*

*In our revised manuscript, we clarified and described the NIRS calibration procedure in more detail. We acknowledge the restrictions and concerns described in Brown et al. (2005). We used only two samples per site, but from different depths, in case of the BZE samples. Samples from different depths within one profile differed strongly in many soil properties, in particular with regard to their NIRS detectable organic compounds as well as their P content. One out of approx. 250 (all BZE samples), respectively one out of approx. 78 (pure brown earth), which may deviate strongly in soil properties and P content, can have only a minor influence on model quality. Therefore pseudo-replication should be only a minor issue.*

*It was not the aim of this contribution to generate widely applicable NIRS models to predict P in forest soils. At this experimental stage, we tested if it was possible to replace the very labor intensive wet chemical standard procedure, as we stated in our manuscript. Nevertheless the samples we used, their number, depths of samples and their origin were in our opinion described in detail in our material and method section (p.560ff). Each table contained the number of the samples of the used sample sets, so the numbers should be clear.*

*In fact, we carried out a leave-out-one-cross-validation procedure. Since the cross-validation is de facto a calibration, we did no to use the term validation and used instead the term calibration (p566/13-17) Calibration was performed with cross validation, a common approach for small data sets. Here a defined number of samples, in our case one sample, were step-wise excluded from the calibration process. The rest of the samples were used to predict the excluded samples. This procedure was performed until all samples were excluded once, and the best models to predict all samples were found (Conzen, 2005)).*

Comment 2.
The manuscript has some peculiar statements. The authors wrote: "Since there was no indication of autocorrelation between samples of different depth, we included all samples in our calibration and validation step". I strongly disagree with that statement. Firstly, the authors should study the paper by Brown et al. (2005). Secondly, the authors should give their scale of interest for each data set and should avoid pseudoreplication. I do not see the need for a test of autocorrelation in this study, since the mineralogical background does

affect the spectra. The presence of the same mineralogical background reduces the noise and increased accuracies for the estimations can be expected.

The authors wrote: "Development of robust NIRS-models requires sample populations that cover the whole calibration range with an approximately even distribution of samples across the range of the variable to be predicted. In contrast, populations with normally distributed samples tend to overestimate low values and underestimate high values in model calibration (Williams, 2001)". This may be ok, but the authors still have to give essential information: whenever they present $r^2$ and RPD values (which are calculated from SD and SECV values), they rely on a normal distribution. Thus, skewness and kurtosis should be given for all data sets and constituents, where RPD and $r^2$ are presented and the interpretations are dependent on that additional information.

*Response:*

*We clarified in our revised manuscript the scale of interests. In addition, we clarified in the text our data selection criteria, which should help to avoid misunderstandings. Initially we hoped to be able to create NIRS prediction models which were valid to predict a wide range of forest soils (BZE samples). Since it became soon clear that this was a very challenging endeavor, we reduced our BZE data set to a subset with lower variation in soil properties (BZE "brown earth"). In addition we used samples from the same soil type and a small geographical region (BEF-China samples) to test, if it was possible, to create prediction models for these particular sample sets. In the latter case we cannot rule out a certain spatial correlation among our samples.*

*Our test for autocorrelation among the BZE samples was – in our view - an important indicator that our samples were independent from each other. Additionally we created for our best models of the BEF China data set new models, where we ensured that a selection of CSPs were not included in the calibration process and therefore were independent and no pseudoreplications. We found no substantial differences for both models. Therefore the problem of pseudo replication was only a minor issue within our study.*

*We added as supplementary material tables for all the datasets used with descriptive statistics including skewness and kurtosis for P-content separated in all P-fractions and P-pools which were used in our study (Table S2).*

**Referee 3**

*The numbers of pages and lines refer to the original version of the manuscript that was submitted to Biogeosciences on September 26th 2014*

The paper evaluates the use of NIRS in forest soil phosphorus research. NIRS would make soil P research more cost and time efficient. Up to now, NIRS has not been used to quantify Hedley P fractions in forest soils. Hence, the paper presents a novel and potentially useful application of NIRS. The title reflects the contents of the paper. The authors have to conclude that only some of the Hedley fractions could be quantified by NIRS and that datasets used for NIRS calibration have to fulfill particular prerequisites (e.g., homogeneity of datasets). However, the description of these prerequisites of datasets is confusing and

should be more precise. The methods and assumptions are largely valid, but are not clearly outlined. For example, the selection criteria for the soil sample subsets are not comprehensible. In addition, the description of the NIRS method is too rough. Therefore, reproduction by fellow scientists would not be possible. The results are sufficient to support the interpretations and conclusions, but the phrasing is partly misleading. The authors give proper credit to related work and clearly indicate their own contribution. However, they should add that NIRS is frequently applied in agricultural soil P research to quantify plant-available P. The overall presentation is well structured, but could be clearer; especially the language could be more precise. Some sentences are nested and hard to understand. The number of references is high (approx. 70 references) and could be reduced. However, some few references concerning the use of NIRS in agricultural soil P research could be added.

*Response*

*This general comment by referee 3 is rather a summary of all comments & technical comments that were also listed in a table (below). Therefore we addressed all of these comments specifically below.*

| page/line | referee comment | our comment |
|---|---|---|
| P1 L20/21 | There are different modifications of the Hedley method. Therefore, the particular fractions should be named in the abstract. | The modification is named in parentheses. Since we do not present results for individual fractions in the abstract, this information should be sufficient. |
| 1 L26 | what is meant with “homogeneity of soil sample sets”? -> explain | “Soil properties” was added to qualify the term homogeneity |
| P1 L27 | what is meant with “useful models”? -> explain | Changed to “usable” |
| P2 L4 | how similar do they have to be? In which respect similar? What are the most important properties that have to be similar? -> explain in more detail | We added “e.g. same soil type, one study site.” |
| P2 L17 | describe the hypotheses shortly | Hypotheses have been described |
| P2 L28 - P4 L4 | this paragraph is too long and should be subdivided, e.g., 1. Role of P fractions in tree nutrition<br>2. Usefulness of NIRS | The paragraph was divided (P3L18) |
| P3 L9/10 | to which part of the sentence does the phrase<br>“particularly in forest soils” refer? | It refers to the application of the method. Therefore we rephrased this sentence to make it more clear. |
| P3 L21 | total C and N contents or which fractions? | We refer now to the various C and N forms, that were analyzed within these studies. |
| P3 L26 | NIRS is usually applied to dried and ground samples. Thus, the different liquid and gas status of soils should be of minor importance. | We deleted this sentence and rephrased this section. |
| P4 L2 | describe the “other soil properties being detectable by NIRS” | Other soil properties were specified in the text. |
| P4 L4 | describe what “high quality in spectral datasets” means in the context of NIRS | The sentence was clarified by changing high quality to reliable quality. This is |

| | | |
|---|---|---|
| | (e.g., homogeneity of soil samples (ground vs. sieved); homogeneity of the sample sets (one soil type vs. different soil types); origin of the sample sets (regional vs. global); homogeneity of the soil sample composition (mineral soil samples with low soil organic matter content vs. mineral soil samples with various contents of soil organic matter), …) | further defined in the methods section. |
| P4 L4-25 | rephrase this paragraph | The paragraph had a misleading section in the middle, which was also pointed out by referee #1 (P4L17). We changed this section according to the comment by ref. 1. |
| P4 L16 | "prediction of C content and sample sets" -> is "and" the right word here? If yes, I do not understand the meaning of the sentence. | "and" was changed to "in" |
| P4 L17 | isn't high variation in chemical composition a cause of high spectral variation? Then "or" wouldn't be suitable here. | Sentence was rephrased "chemical composition associated with high spectral variation" |
| P4 L26-28 | there are several studies on NIRS models for different P forms (e.g. microbial P); in agricultural soil P research NIRS is used to quantify different P fractions | Further references have been included in the manuscript. |
| P5 L14/15 | did you select a subset of the BZE dataset? | Yes. Information was added. |
| P5 L25-27 | Explain your selection criteria. If there were "clear correlations" between total P and P fractions, how could that help you to create subsets? | This section was extensively rephrased and the specific section was deleted. |
| P6 L22 and 25 | volume to volume ratio or volume to mass ratio? Better write 2.5ml : 1 ml or 2.5 ml : 1 g | 2.5 ml : 1 g was used |
| P7 L5-10 | this paragraph fits better to the introduction | The paragraph was reduced to one sentence and other sentences were added to the introduction as suggested. |
| P7 L11 | here, you do not write that you used replicate soil samples, but later you write something about replicate soil samples | Information was added |
| P7 L12 | please add type of resin (counterion) | Added: Dowex 1x8 20-50 mesh (Sigma-Aldrich, Taufkirchen, Germany) |
| P7 L14 | please add energy level of ultrasonic treatment | Added: ultrasonic bath RK510H, 35kHz; 23W/l (Bandelin, Berlin, Germany) |
| P7 L23 | write "PO4-P" or "molybdate reactive P"; what kind of photometer did you use (continuous flow, microplate reader,…)? At what wavelength did you measure? | Molybdate reactive P. We measured at 882nm wavelength with the spectrophotometer - Shimadzu UV mini-1240 (information was added) |
| P8 L3 | rephrase; it is not clear from this sentence whether you summed up Pi and Po of the NaOH and S-NaOH fractions or if you summed up Pi of the NaOH and Pi of the S-NaOH fraction as well as Po of the | We clarified this to: we summed up Pi of the NaOH and Pi of the S-NaOH fraction as well as Po of the NaOH and Po of the S-NaOH fraction. |

| | | |
|---|---|---|
| | NaOH and Po of the S-NaOH fraction | |
| P8 L7/8 | Although all acids can act as oxidants, persulfate is by far a stronger oxidant than HCl as it is a source of sulfate radicals. HCl is used for hydrolytic degradation of organic matter, whereas persulfate is a “true” oxidizing agent. Yet, for the degradation of organic P compounds, both treatments might be equally efficient. Please correct your statement and check the literature if others also found no organic P in conc. HCl-extracts. If it is true that Po in 1 M HCl extracts is negligible, why did you measure TP in HCl conc. extracts? | We did not state that we or others did not find organic P in the concentrated HCl fractionation step. What we stated was that the reproduction of inorganic P is very poor, since in the fractionation step with concentrated acid, before adding the persulfate solution, could already degrade organic matter and therefor the distinction between Pi and Po could be not reliable. Therefore we decided to use only TP for the concentrated HCl fraction when developing NIRS models. |
| P8 L9 | what is meant with “satisfying”? | Satisfying was changed to reproducibility within replicates. |
| P9 L3-20 | In part, this has already been mentioned in the introduction, some general remarks may be shortened. | This is the main Material & Methods section on NIRS, which is already quite short. We believe that we cannot shorten it further without losing important content. Since none of the other referees suggested to shorten this section and we already shortened other sections of the M&M section substantially . |
| P9 L6/7 | O-H, C-H and N-H are bounds and not functional groups | Was changed to bonds |
| P9 L11 | NIRS detectable soil properties -> describe them | Soil properties were specified in the text |
| P9 L27 | Why did you not test the second or third derivative? According to Barnes et al. (1989) spectra should be detrended to remove scatter effects. Please consider this. Barnes et al. (1989) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy 43 | Regarding mathematical data pre-processing we specified: “Before statistical analyses, a number of mathematical data pre-processing options were tested. The pre-processing options providing the best results were first derivative, vector normalization, or a combination of these two.” |
| P9 L28 | rewrite this sentence; you did not do these treatments for the PLS | We clarified that the data treatment was done before the PLS |
| P9 L29-31 | please give more details (size of gaps, amount of smoothing) | Information on spacing (1) and range of smoothing points (5, 9,13,17, 21, 25) was added. |
| P9 L31 | cross validation is used to avoid overfitting and to obtain the optimal number of terms in the calibration; why is it a common approach to replace the calibration step by cross validation for small data sets? References? | Because with small datasets and complex sample matrix, too few calibration samples could lead to models which are not robust in validation and application. The advantage of cross-validation is the increase of information since more samples could be integrated for model building. Therefore the cross-validation is an appropriate method to reduce the number of samples which are necessary for model development. References to explain just this were added at the specific position in the manuscript. |
| P10 L4-10 | The criteria for this automated selection | This automated selection is part of the |

| | | |
|---|---|---|
| | do not get clear from this. | software package. Since this is a standard procedure in NIRS software, it is not further explained in the software documentation. |
| P10 L11-17 | move this section to 2.1 Soil samples; did you consider to group samples by parent material? | This section was extensively restructured and the focus changed so it fits into this part of the Material & Methods section. The number of samples originating from soils with the same parent material was too small to develop meaningful NIRS models. |
| P10 L20 | I didn't understand the sentence before I saw the results; you do not mean the relationship between P and soil C and N but the quality of the relationship | As stated above, this section was extensively restructured and the misleading sentence deleted. |
| P10 L26-30 | move this section to 2.1 Soil samples and give the number of samples in each sample set | Number of samples in each sample set was added. The section was not moved to the section 2.1 Soil samples. Even though the sample sets consists of soil samples, it is clearly a description of the dataset used for NIRS modelling and therefore should remain in this section. |
| P11 L 9 | please correct: RDP=ratio of SD to standard error of prediction | Corrected |
| P12 L14ff | You should always write cross-validation instead of calibration | The cross-validation is in fact a calibration since a sample set was used to create a prediction model for unknown samples. All samples of the cross validation are part of the process. The actual validation is performed with independent samples which were at no stage part of the calibration/cross-validation process. To avoid confusion between the actual calibration and the cross-validation we decided to keep the term calibration which was explained in the m&m section. We have used this terminology in a number of previous publications, and this has always been accepted. |
| P12 L16 | do you mean worse than level D when you write "produced no useful calibrations"? | Yes, we clarified this by adding: "(lower than quality level D)" |
| P13 L11-19 | rephrase this paragraph, it is really hard to understand (e.g., Grouping of the Hedley fractions into labile, moderately labile and stable P fractions did result in good models for the BEF-China dataset, while only the stable fractions of the other three datasets (BZE+BEF, BZE, BZE Brown Earth) could well be predicted with NIRS models (Fig. 5).) | Paragraph was rephrased accordingly |
| P13 L19 | useful -> best? | Changed to best |
| P13 L26 | do you mean the levels defined on p 11 with "goodness of fit of calibration models"; in Fig. 7 you use the R2 of the | Please see re-written section on the interpretation of the new Fig. 7. |

| | | |
|---|---|---|
| | calibration model | |
| P13 L28 | "…were best for the Po fractions…<br>" I couldn't find any good relationship in Fig. 7 | We defined $r^2$ values > 0.7 for NIRS models as indicative of useable models. Some of these high values coincided with high values for the relationship between soil C or N and P in Hedley fractions. We agree that we cannot speak about relationships here, but perhaps of indications. The text was modified accordingly. |
| P13 L26-P14 L 7 | It makes no sense to correlate a R2 and a Spearman rank correlation coefficient (rs). rs is a non-parametric measure which may not simply be related to a parametric measure like R2. If your only rationale behind this approach is to test whether NIRS models for P fractions are a result of C-P or N-P relationships, why don't you simply test if your NIRS models for P fractions have a similar predictive power for C and N as for P fractions? | It is not possible to predict C or N with models designed to predict P. Instead we picked up the suggestion of Referee #1 to perform a regression analysis. Please see our comment to Referee #1 (P13 L26 – P14 L7). |
| P14 L7 | the given correlation coefficients are not for the dataset presented in Fig.7, but for a dataset with some fractions removed, right? | We changed figure 7 since Referee #1 and #3 had concerns about the usage of correlation coefficients and $r^2$ of the NIRS models. According to Referee #3 we replaced them with $r^2$ values from regression analysis. Compare response on previous comment (Referee #3 on P13 L26-P14 L7) |
| P14 L12 | you might use the data of your "random quality check" to calculate coefficients of variation for individual fractions | We calculated instead standard errors of the repeated measured samples and displayed them in the text. Compare response on comment Referee #1 on P14 L24-26 |
| P14 L13 | reference method = Hedley fractionation? | Yes, this was further clarified in the text |
| P14 L25/L 26 | "repeatedly analyzed" and "random quality check" -> describe in the material and methods section how many replications you did; did you repeat the analysis or the fractionation? | We included a description of the procedure in the Material and Methods section. |
| P14 L30/31 | the reason for the bad NIRS models might also include other factors | Yes, that's true but values around or below the detection limit might be in this case the most important one. |
| P14 L31 | what is meant with "valid"? | Valid replaced by "meaningful" |
| P15 L13 | what is "a reasonable prediction"? | We rephrased the sentence for more clarity. |
| P15 L29 | "total organic P" is an inadequate term for the sum of Hedley organic P fractions, since not all organic P is extracted during the Hedley procedure | Changed to sum of assigned organic P fractions. |
| P16 L8/9 | Rephrase | Rephrased; Compare response to comment Referee #1 on P16 L1-6ff |

| | | |
|---|---|---|
| P16 L13-16 | Explain why global models are potentially as accurate as more local calibrations. | If sample size in local calibrations are <150, an option would be to increase model quality by combining datasets to a global model. Information was added. |
| P16 L12-23 | This paragraph is a bit confusing, since you compare studies dealing with organic material with studies dealing with soil. Due to numerous reasons (which you partly mentioned) soils are more complex than organic material and to create "global models" for soils is potentially less successful. Please rather refer to studies dealing with soils. For instance, Brunet et al. (2007) also found better predictions for total C when using subsets of soils compared to a "global model". | We changed this paragraph, references and examples of studies dealing with soils were added. |
| P17 L8-10 | Evidence on these questions is limited, but there are for instance combined Hedley fractions/31P NMR studies dealing with these questions. See Negassa and Leinweber (2009) JPNSS 172:305-325 | We rephrased this section to address a comment of Rev#1. Nevertheless it was not the aim of our study to identify organic P forms incorporated in a specific organic Hedley P fraction. |
| P17 L11/12 | even in soils of comparable soil type the variation in P forms within Hedley fractions may be high due to other reasons like tree species -> differing litter quality, climate -> soil humidity -> soil microorganisms | This might be true but in our study our focus was on soil properties. |
| P17 L12 | the development of NIRS models for specific subgroups of soils is probably more promising, but why only create subgroups according to soil types and not parent material? | The number of samples with the same parent material was too small to achieve meaningful NIRS models |
| P17 L12-14 | rewrite the sentence "The possible … individual dataset | Sentence was rephrased |
| Tab.1 and 2 | Dataset "all" is missing | Dataset all did not lead to any useful results and therefore was mainly presented in the text. Furthermore it is a combination of the sample sets BEF and BZE and information can be extracted from the two tables |
| Fig.1 | Check the presentation of the modelled NaOH fractions? What did you combine? | Figure was redesigned to clarify which NaOH fractions were combined for modelling. |

Technical comments

| **page/line** | **referee comment** | **our comment** |
|---|---|---|
| P1 L1 | Near-Infrared -> near-infrared | Changed |
| P1 L1 | Phosphorus -> phosphorus | Changed |
| P1 L15 | P -> phosphorus (P) | Changed |
| P1 L20 | Hedley method -> Hedley sequential | Changed |

| | | |
|---|---|---|
| | extraction method | |
| P2 L10 | Phosphorus -> Phosphorus (P)” | Changed |
| P2 L16 | phosphorus-limitations -> phosphorus limitation | Changed |
| P2 L24 | P-nutrition -> P nutrition | Changed |
| P2 L25 | monitoring of the -> monitoring the | Changed |
| P2 L28 | rephrase “solely total P contents are often measured | Changed |
| P3 L4 | cite the papers of Hedley | Cited: Hedley et al. 1982 |
| P3 L8 | have been -> has been | Changed |
| P3 L10/11 | Here, in contrast to agricultural soils, the slowly cycling P pool contributes -> In contrast to agricultural soils, the slowly cycling P pool in forest soils contributes | Changed |
| P3 L14 | Hedley-fractionation -> Hedley fractionation | Changed |
| P3 L16 | Hedley-P fractions -> Hedley P fractions | Changed |
| P3 L18 | start new paragraph after “may be a promising approach.” | Changed |
| P3 L21 | C or N -> carbon (C) or nitrogen (N) | Changed |
| P3 L24 | bracket in bracket… | Changed |
| P3 L26 | gas -> gases | Changed |
| P3 L29 | of to the USDA -> of the USDA | Changed |
| P3 L31 | “P or” can be deleted L 32 find a more suitable word than “subsequently” (e.g., hence) | Deleted, “subsequently” changed to “therefore” |
| P4 L23/24 | couldn’t “depending on the homogeneity respectively heterogeneity of” be replaced by “for”; would make the sentence shorter and easier to understand | Replaced it with “for” |
| P4 L30 | “to do so” -> could the sentences be rephrased so that “to do so” can be replaced? | To do so -> to assess these P fractions |
| P5 L9-13 | change the order of the two sentences “From each site…” and “Includ-<br>ing 70 sites…” | Changed |
| P5 L22 | delete “aimed to” and change “select” to “selected” | Deleted and changed |
| P5 L24 | add a reference | The section was changed, the references were added in a later part of the Material & Methods section. P10 Chapter 2.3 Near infrared spectroscopy.<br>Compare response to comment referee #3 P9 L31 |
| P6 L1 | Research -> research | Changed |
| P6 L8 | 5-10cm -> 5-10 cm | Changed |
| P6 L14 | delete “and” | Deleted |
| P6 L15 | pH-Values -> pH values | Changed |
| P6 L16 | North Western German Forest Research Institute -> Northwest German Forest Research Station | Changed |
| P6 L17 | rephrase “data was measured according to the Handbuch Forstliche Analytik” | Rephrased: Samples were prepared and measured according to the German Forest science standard |
| P6 L18 | carbon and nitrogen -> C and N | Changed |

| | | |
|---|---|---|
| P6 L19 | 1150 C ? | 1150°C |
| P6 L21 | 2x carbon -> C | Changed |
| P6 L22/25 | rephrase “water solution” | |
| P6 L25 | derivedin -> derived in | Changed |
| P7 L6 | analysis –> analyses | Changed |
| P7 L20 | with the -> after | Changed |
| P7 L23 | Phosphorous -> Phosphorus | Changed |
| P7 L27 | dot is missing | Included dot |
| P7 L29 | remove the different “-“ | Removed |
| P7 L29 | organically bound -> total | Changed |
| P7 L30 | autoclave and the -> autoclave. The | Changed |
| P7 L31 | P(Po) -> P (Po) | Changed |
| P8 L19 | delete the dot | Dot deleted |
| P8 L22 | Hedley Fractionation Method -> Hedley fractionation method | Changed |
| P9 L3 | exited -> excited | Changed |
| P9 L9 | Phosphates and other P compounds -> Phosphates and other inorganic P compounds | Changed |
| P9 L14 | either replace the comma by a dot or fill in “but” or “instead” | Replaced comma by a dot |
| P10 L31 | set3 -> set 3 | Changed |
| P11 L6 | software ) -> software) | Changed |
| P11 L28 | Phosphorus concentrations -> phosphorus contents | Changed |
| P11 L29 | P concentrations -> P contents | Changed |
| P11 L31 | P concentration -> P content | Changed |
| P12 L8 | concentrations -> contents | Changed |
| P12 L11 | within -> below? | Changed to below |
| P12 L11 | Hedley method -> Murphy & Riley (1962) method? | Changed to Murphy and Riley method |
| P12 L13 | 3.2.1 -> 3.2 | Changed |
| P12 L13 | NIRS models by P fractions -> NIRS models for P fractions? | Changed |
| P12 L17 | soils type -> soil type | Changed |
| P12 L19 | in -> with | Changed |
| P12 L20 | in -> with | Changed |
| P12 L20/21 | rephrase: only D level quality or only two fractions? | Rephrased: only D level quality for two of the fractions |
| P12 L23 | concentrations -> contents | Changed |
| P12 L30 | replace “Whereas” by a more suitable word | Deleted whereas. Rephrased sentence structure |
| P13 L10 | 3.2.2 -> 3.3 | Changed |
| P13 L28 | Carbon -> C | Changed |
| P13 L28 | Nitrogen -> N | Changed |
| P14 L10 | NIRS models for Hedley fractions and pools -> NIRS models for P fractions and pools | Changed |
| P14 L27 | minimum -> level of the | Changed |
| P14 L30 | factions -> fractions | Changed |
| P15 L4 | “In addition” is not appropriate here | Deleted In addition |
| P15 L18 | Fractions -> fractions | Changed |
| P15 L19 | rephrase (e.g., Whether P is in organic or inorganic form seemed to be of | Rephrased as recommended |

| | | |
|---|---|---|
| | importance for…") | |
| P15 L21 | models predicting the organic P fractions performed better than for inorganic P fractions throughout -> models predicting the organic P fractions performed better than models predicting the inorganic P fractions | Changed |
| P15 L22-25 | change the order of the two sentences "The superior quality…" and "Similar results…" | Changed sentence order |
| P15 L23 | in which -> because | Changed |
| P15 L27 | Why "Therefore"? | This relates to the prior sentence |
| P15 L30 | to -> by | Changed |
| P16 L3 | and not simply -> and are not simply | Changed |
| P16 L9 | even poorer or non-existent -> even poorer than for organic fractions or non-existent | Changed |
| P17 L8 | To our knowledge -> To our knowledge, | Changed |
| P17 L11 | P-forms -> P forms | Changed |
| P17 L17 | soil P in Hedley fractions of different availability -> soil P Hedley fractions of different availability with NIRS | Changed |
| P17 L30 | represents -> requests | Changed |
| P18 L6 | North Western German Forest Research Institute -> the Northwest German Forest Research Station | Changed |
| Figure 1 | provides -> provide; compound -> compounds | Changed |
| Figure 3 | set4 -> set 4 | Changed |
| Figure 7 | add "triangles = P HCl conc. fractions" | Added |