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 comments of anonymous reviewer #1 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. 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	The area covered by the BZE samples

	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).	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. The Information was added in the material and method section 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	We replaced the expression “functional groups” with “bonds”, since the bonds were stimulated not functional groups.

	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.	
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 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.	We rephrased this section.
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. 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	According to Referee #1 and #3 we replaced figure 7. Now it shows the relationship between goodness of fit

	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 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.	values for NIRS models and the relationships between soil C and N and P in different fractions.
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	See response to referee #1 general comment 1

	result in spectra different from that of diesters?	
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 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.	See response to general comment 3

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	Changed accordingly

	composition associated with high spectral variation”).	
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	Was rephrased

	either of these meanings. Maybe you refer to the limit of detection or similar? Rephrase.	
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	The spelling in the figure and caption was changed as suggested. 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. According to our knowledge, it is not

	change spelling in Figure 3 and caption accordingly) be preliminary as stated in the methods section?	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