Below we have reproduced the greatly appreciated comments of anonymous reviewer #2 and inserted our responses in italics.

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).