Answers to comments by anonymous referee #3 (RC C2582) of 17 February 2014

M. Nussbaum, A. Papritz, A. Baltensweiler, L. Walthert

March 18, 2014

Many thanks for treating our article with so much care and for giving us detailed feedback. In the sequel we would like to comment on your suggestions.

Validation with independent data

Your main criticism relates to our claim that we validate the precision of the predictions with "independent" data. Your view is that "validation with independent data" should be reserved for a comparison of predictions with extra data (collected by probability sampling, cf. Brus *et al.*, 2011) and that our validation approach does not differ from cross-validation exercises done earlier by other authors. We carefully evaluated these arguments, but in the end only partially agree:

We certainly accept that model assessment is most meaningful when predictions are compared to validation data that are 1) newly and independently collected from the calibration data by a randomized sampling design and 2) are not used in any phase of the model building process. Unfortunately, such a procedure can be rarely used in practice as it requires considerable funding and time and cannot be used for obvious reasons when only legacy data are available.

One is then bound to data splitting strategies and cross-validation for assessing the predictive power of a statistical model. You argue that there is no difference between our data splitting approach and cross-validation as done earlier by Martin *et al.* (2009, 2011) and Meersmans *et al.* (2012). We are sorry to say that we do not share this view: Following Hastie *et al.* (2009, chap. 7) our validation set ($n = 175$) corresponds to a *test set* (used for estimating the generalization error, i.e. the prediction error for new data), and our calibration set ($n = 858$) has the combined role of a *training* (used for parameter estimation) and a *validation* set (used for estimating prediction errors during model selection). When no separate training and validation sets are available then cross-validation is often used for parameter estimation and model selection and for choosing tuning parameters of employed algorithms. This is exactly what we did: We chose the relevant covariates of the regression model and the robustness tuning constant by cross-validation using only the calibration data. Then we estimated the generalization error by applying the final model to the test set (i.e. to our validation data).

The "external validation" reported by Martin *et al.* (2009, 2011) and Meersmans *et al.* (2012) does not have the same significance as our validation results: Martin

et al. (2009, 2011) used cross-validation (with the full data set) for selecting the tuning parameters of their algorithm and kept these parameters then fixed for estimating again by cross-validation the generalization error. Clearly, such a procedure gives a too optimistic estimate of the generalization error. Some information in the cross-validation subset, currently being predicted and used for estimating the generalization error, had been used before for model building. Also the reported "external validation" in Meersmans *et al.* (2012) provided only a distorted estimate of the generalization error because the structure of the regression model and its coefficients were not re-estimated in cross-validation (Meersmans, personal communication). Unfortunately, this is common practice in the geostatistical community, but is clearly not correct (Hastie *et al.*, 2009, sec. 7.10.2). Such a "validation" scheme tends to be overly optimistic when reporting the precision of predictions for new data. According to our knowledge only Mishra *et al.* (2009, 2010, 2012) and Wiesmeier *et al.* (2012) did not use measurements of the validation set for their model calibration. This justifies to explicitly mention these studies and to oppose them to the model validation strategy used by Martin *et al.* (2011) and Meersmans *et al.* (2012) who did the "external validation" (partly) with fixed parameters estimated from the full data set.

Moreover, we would like to comment on the concern that our validation data might not be spatially representative for SOC stocks of Swiss forests because it had been obtained by splitting a legacy data set (and not by independent probability sampling): 134 out of 175 soil profiles of our validation set stem from a survey where the sites were arranged on a 8×8 km grid. Data from these sites should be spatially representative because the grid was placed without consideration of SOC stocks. The additional 38 sites of this survey had to be assigned to the calibration set because there were no other data available for certain parts of the country. Another 38 soil profiles of the validation set stem from regional studies where the sites were arranged on 1×1 km grids. Again, these should be representative for local conditions and should not be influenced by sampling bias. You recommend that we use for validation the data available from the national and two cantonal soil monitoring networks as described in section 3.4 of Nussbaum *et al.* (2012). However, the sites of these networks have been purposively selected similar to the majority of the WSL data and are likely not more representative than the data that we assembled for validation. Furthermore, soil sampling and analysis differed in these surveys from the procedure used by WSL and this likely adds some extra variation when we evaluate the precision of our predictions with this data. We have therefore good reason to believe that the generalization error, estimated with our validation set, provides a fair picture of the prediction error of our model for new data. Finally, although we could have repeated splitting the data randomly into calibration and validation sets (and then building and assessing the model for each such split) we abstained from such a procedure because we wanted to have a single final model that can be used in future applications to predict SOC stocks. Furthermore random splitting would have precluded the creation of a spatially representative validation data set as described above.

Further general comments

We further identified the following issues in the general part of your review, on which we would like to comment:

1. On page C2584 (line 25) of the review you suggested to discuss the added value of robust measures for evaluating the precision of SOC predictions. We see no real need for this as the inclusion would extend the length of the article: We mentioned in the article why we computed BIAS and RMSE of the *relative* prediction errors. We used robust variants of these statistics because non-robust quality measures are possibly influenced by only few observations that are poorly predicted. Robust quality measures provide a better picture how well a method predicts the majority of the observations. A detailed discussion of continuous ranked probability score (CRPS) can be found in Gneiting *et al.* (2007), to which we refer repeatedly in the paper, and we see no need to expand on this.

2. On page C2585 (line 1–10) you inquire about the increase in prediction errors when residual autocorrelation is neglected and predictions are computed from the regression models only. We have dealt with this issue in our answer to the review by Philippe Lagacherie (AC C2824).

3. You suggested to include graphs of the variograms in the article. We see again no need for this as such plots can be easily generated from the estimated variogram parameters listed in Table S5 of the Supplementary Material.

4. We refer now in the introduction of our article to the important review paper by Minasny *et al.* (2013). We were quite happy to see that our study addresses two issues that the authors explicitly mentioned as weaknesses of past studies: Validating SOC predictions with independent data and qualifying the precision of predictions by modelled standard errors.

5. Concerning the advantages of LASSO (least absolute shrinkage and selection operator) we refrain from adding a detailed discussion. First, non-specialist information about this procedure is available (e.g. Hastie *et al.*, 2009, section 3). Second, we used LASSO only as a screening tool to find a preliminary set of covariates. The final set of covariates was selected by cross-validating the robustly fitted geostatistical models.

Specific comments

We accepted most of the suggestions as they improve the clarity of the article. We discuss only the comments where we (partly) disagree or some clarification is needed:

P7081 L23 We have changed the text and replaced "common model" by "the same set of fitted parameters". For a linear model the covariance of the prediction

errors can be easily computed from the covariance matrix of the fitted regression coefficients (see any textbook on linear regression), for ML methods, it is not clear how to compute these covariances irrespective whether there is residual autocorrelation or not (see our answer to comments by referee #1, AC C2823).

P7082 L25 According to Table 2 in Martin *et al.* (2011) SOC stocks are more variable in French forests than on cultivated land. There is not yet harmonized soil data available to check this for Switzerland. Nevertheless, we mention now that forest soil stocks might be more variable and use this as a further justification for a separate analysis of the respective data.

P7083 L11 These percentages are correct. 29 % of the *total* area of Switzerland as opposed to 45.5 % of the *vegetated* area of Switzerland is covered by forests.

P7087 L5–15 This is a fair comment, which points to a weakness of our analysis (in particular of the validation scheme): We re-computed therefore the "median mass of soil particles < 2 mm" assigned to geotechnical map units using only the calibration data ($n = 858$) and re-fitted the model for topsoil SOC stocks (0–30 cm) to this data, however, without repeating the full model building process. Table 1 below lists for the validation set the statistics of the relative prediction errors of the re-calibrated model. Comparison with Table 2 of the article reveals that the statistics hardly changed. We abstain therefore from re-computing the predictions of regional and national SOC stocks as the figures published by Nussbaum *et al.* (2012) and listed in our article that are currently used for Switzerland's GHG inventory.

P7094 L6–9 Gneiting *et al.* (2007) give an exhaustive account on quality measures to validate probabilistic forecasts and we see no need to expand on this.

P7095 L20–25 We too refer to our answer to the comments by Philippe Lagacherie (AC C2824).

P7099 L2–9 We computed BIAS and RMSE of the *relative* prediction errors because this seems a natural choice for a lognormal model, where the coefficient of variation (and not the variance) is constant. Other studies have reported absolute BIASes and RMSEs. The (robust) R^2 is therefore the only criterion available for a cross-study comparison.

Table 1: Statistics of relative prediction errors of soil organic carbon (SOC) stocks in topsoil (0–30 cm depth) for the validation set ($n = 175$). The model was fitted to the data where the covariate "mass of soil particles < 2 mm assigned to geotechnical map units" was computed only from the calibration data set ($n = 858$).

BIAS	RMSE	R^2	robBIAS	robRMSE	robR^2	CRPS
0.135	0.488	0.355	0.063	0.390	0.345	0.221

References

Brus, D. J., Kempen, B., and Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, **62**(3), 394–407.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, **69**(2), 243–268.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer, New York, second edition.

Martin, M. P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K. M., Bourgeon, G., and Arrouays, D. (2009). Optimizing pedotransfer functions for estimating soil bulk density using boosted regression trees. *Soil Science Society of America Journal*, **73**(2), 485 – 493.

Martin, M. P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., and Arrouays, D. (2011). Spatial distribution of soil organic carbon stocks in France. *Biogeosciences*, **8**(5), 1053–1065.

Meersmans, J., Martin, M. P., Lacarce, E., De Baets, S., Jolivet, C., Boulonne, L., Lehmann, S., Saby, N. P. A., Bispo, A., and Arrouays, D. (2012). A high resolution map of French soil organic carbon. *Agronomy for Sustainable Development*, **32**(4), 841–851.

Minasny, B., McBratney, A., Malone, B., and Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in Agronomy*, **118**, 1–47.

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., and Van Meirvenne, M. (2009). Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. *Soil Science Society of America Journal*, **73**(2), 614–621.

Mishra, U., Lai, R., Liu, D., and Van Meirvenne, M. (2010). Predicting the spatial variation of the soil organic carbon pool at a regional scale. *Soil Science Society of America Journal*, **74**(3), 906–914.

Mishra, U., Torn, M. S., Masanet, E., and Ogle, S. M. (2012). Improving regional soil carbon inventories: Combining the IPCC carbon inventory method with regression kriging. *Geoderma*, **189 − 190**(0), 288 – 295.

Nussbaum, M., Papritz, A., Baltensweiler, A., and Walthert, L. (2012). Organic carbon stocks of swiss forest soils. Final report, Institute of Terrestrial Ecosystems, ETH Zürich and Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Zürich and Birmensdorf.

Wiesmeier, M., Spörlein, P., Geuss, U., Hangen, E., Haug, S., Reischl, A., Schilling, B., Lützow, M., and Kögel-Knabner, I. (2012). Soil organic carbon stocks in southeast Germany (Bavaria) as affected by land use, soil type and sampling depth. *Global Change Biology*, **18**(7), 2233–2245.