Received: 7 March 2022 | Accepted: 9 September 2022

DOI: 10.1002/jpln.202200075

RESEARCH ARTICLE

Description and prediction of copper contents in soils using different modeling approaches—Results of long-term monitoring of soils of northern Germany

Bernard Ludwig[1] | **Karen Klüver**[2] | **Marek Filipinski**[2] | **Isabel Greenberg**[1] | **Hans-Peter Piepho**[3] | **Eckhard Cordsen**[2]

[1]Department of Environmental Chemistry, University of Kassel, Witzenhausen, Germany

[2]Landesamt für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein (LLUR), Flintbek, Germany

[3]Institute of Crop Science, Biostatistics Unit, University of Hohenheim, Stuttgart, Germany

Correspondence
Bernard Ludwig, Department of Environmental Chemistry, University of Kassel, Nordbahnhofstrasse 1a, 37213 Witzenhausen, Germany.
Email: bludwig@uni-kassel.de

This article has been edited by Jianwei Li.

Abstract

Background: Different regression approaches may be useful to predict dynamics of copper (Cu), an essential element for plants and microorganisms that becomes toxic at increased contents, in soils.

Aim: Our objective was to explore the usefulness of mixed-effects modeling and rule-based models for a description and prediction of Cu contents in aqua regia (Cu_{AR}) in surface soils using site, pH, soil organic carbon (SOC), and the cation exchange capacity (CEC) as predictors.

Methods: Three sites in northern Germany were intensively monitored with respect to Cu_{AR} and SOC contents, pH, and CEC. Data analysis consisted of calibrations using the entire data set and of calibration/validation approaches with and without spiking.

Results: There was no consistent temporal trend, so data could be combined for the subsequent regressions. Calibration using the entire data set and calibration/validation after random splitting (i.e., pseudo-independent validation) were successful for mixed-effects and cubist models, with Spearman's rank correlation coefficients r_s ranging from 0.83 to 0.91 and low root mean squared errors (RMSEs). Both algorithms included SOC, CEC, and pH as essential predictors, whereas site was important only in the mixed-effects models. Three-fold partitioning of the data according to site to create independent validations was again successful for the respective calibrations, but validation results were variable, with r_s ranging from 0.04 to 0.76 and generally high RMSEs. Spiking the calibration samples resulted in generally marked improvements of the validations, with r_s ranging from 0.45 to 0.67 and lower RMSEs.

Conclusions: Overall, the information provided by SOC, pH, and CEC is beneficial for predicting Cu_{AR} contents in a closed population of sites using either mixed-effects or cubist models. However, for a prediction of Cu_{AR} dynamics at new sites in the region, spiking is required.

KEYWORDS
copper, mixed-effects modeling, rule-based modeling, soil monitoring

1 | INTRODUCTION

Copper (Cu), a micronutrient, is essential for plant growth, and adequate Cu concentrations in crop tissues are usually considered to be in the range of 5–30 mg kg^{-1} (Adriano, 2005). At high concentrations in soils, however, Cu is a toxic pollutant for plants, microorganisms and other living organisms (Anatole-Monnier, 2014; Zhou et al., 2011). Its typical oxidation state in soils is +II, but compounds with Cu^{+I} or elementary Cu^0 may also exist. Cu concentrations in soils consist of geogenic and anthropogenic contributions, to which agricultural management, especially organic fertilization and application of Cu-containing pesticides, can be of quantitative importance (Kamermann et al., 2015; Panagos et al., 2018). For European surface soils, Ballabio et al. (2018) reported the highest mean soil Cu concentrations for vineyards (49.3 mg kg^{-1}), followed by olive groves and orchards, which had considerably increased contents relative to the overall average Cu concentration of 16.9 mg kg^{-1}. Cu inputs due to anthropogenic activities affect mostly surface soils, since Cu mobility in soils is very low. For instance, Bigalke et al. (2010) and Blotevogel et al. (2018) estimated the mean transport rate of anthropogenic Cu to be approximately 1 cm $year^{-1}$.

A large number of factors control Cu concentrations and mobility in soils. In the soil solution, Cu^{II} is present as an aqueous complex or organically complexed form. To a minor extent, Cu^{2+} may be reversibly bound as an exchangeable cation. However, the dominant control of Cu mobility in soil is sorption on organic matter and, to a lesser extent, clay, Fe- and Mn-(hydr)oxides, and carbonate (Amelung et al., 2018; Baker & Senft, 1995; Blotevogel et al., 2018; Droz et al., 2021; Groenenberg et al., 2006). Overall, it is well established that geogenic Cu concentrations are determined to a great extent by parent material, mineralogical composition, and soil texture (Amelung et al., 2018; EFSA, 2008; Huschek et al., 2004). The importance of the soil texture for Cu concentrations of European soils may be explained by the contents of clay minerals and the Fe- and Mn-(hydr)oxides in the clay fractions. Additionally, Cu mobility is affected by pH, since higher pH increases sorption sites (e.g., negative charge) for Cu on clay and organic matter, and sand-containing carbonate (Sposito, 2008). The cation exchange capacity (CEC) is also related to Cu in soils within a limited range of parent materials, because of the indirect relationship between CEC and concentrations of clay-size minerals and soil organic carbon (SOC) (Wu et al., 2003).

Different process-based (e.g., Mallants et al., 2017; Michel et al., 2007) and statistical modeling approaches exist for the description and prediction of Cu contents in soils. Statistical approaches have ranged from multiple linear regression (Romić et al., 2004) to general linear models in combination with kriging (Ballabio et al., 2018) and machine learning algorithms such as neural networks, random forests and bagging trees (Droz et al., 2021). The usefulness and limitations of the statistical approaches may depend on the measured variables available for prediction, the scale and resolution of the variables, the presence of collinearity among the variables, sample size, sampling design, and research aims. For many studies, however, especially those that consider temporal dependencies and/or include a hierarchical (multi-stratum) sampling, mixed-effects models may be the method of choice for elucidating potential relationships between a response variable and independent variables (Galecki & Burzykowski, 2013).

Recent developments in machine learning algorithms may also be promising for predictions of Cu contents in soils (Lantz, 2019). For instance, the use of the rule-based model cubist can result in improved predictions by using a boosting-like procedure called committees. Moreover, model predictions can be automatically adjusted using neighboring points from the calibration (or training) set data (Kuhn & Quinlan, 2021). However, model outcomes need to be interpreted cautiously in studies with hierarchical (multi-stratum) sampling designs, since these are not adequately considered in the model.

For the assessment of predictive capabilities of modeling approaches, calibration–validation procedures may be employed. Multiple partitioning of data into calibration and validation sets is recommended for a thorough investigation of the usefulness and limitations of predictive modeling approaches (Cawley & Talbot, 2010). Different calibration–validation procedures may be performed depending on the intended use of the algorithms to be calibrated. For a data set with multi-stratum sampling design, random splitting of the data set only allows the usefulness of Cu predictions to be tested as a function of the predictors for the closed population of sites (pseudo-independent validation), whereas the usefulness of Cu predictions for new sites may be tested after splitting the data set according to sites (see, e.g., Brown et al., 2005), for example, in the field of SOC and spectroscopy in the visible and near infrared range.

For our study, high-resolution data on Cu content in aqua regia (Cu_{AR}), SOC content, pH, and CEC were available for four soil monitoring sites in northern Germany (see Barth et al. (2000) for main aims and additional information). The presence of independent (sites) and dependent data (observations per site) makes application of mixed-effects models promising, and this can additionally be compared with the powerful rule-based cubist algorithm. In our study, the variable "site" was assumed to aggregate various site-specific variables (e.g., parent material and specific mineral composition) for Cu_{AR} regressions. The objective was to explore the usefulness of mixed-effects modeling and rule-based models for a description and prediction of Cu_{AR} contents in surface soils affected by site, pH, SOC, and CEC. We hypothesized that (1) mixed-effects models and rule-based models may be similarly successful in predicting Cu_{AR} contents; (2) both algorithms may point to a similar importance of the independent variables included in the final models; and (3) Cu_{AR} contents may be predicted successfully in a closed population (i.e., after random splitting of the data set) (a) and with decreased accuracy for new sites (b).

2 | MATERIALS AND METHODS

2.1 | Monitoring and soil analyses

In Schleswig-Holstein, there are currently five intensive soil monitoring sites located in the four main soil natural areas: marshland, Vorgeest, Hohe Geest, and the eastern hill country. This study analyzed three of these sites (sites 9, 35, and 36) for which at least 50 complete

TABLE 1 Characteristics of the three intensively monitored sites in northern Germany

Site and land use	Parent material	Soil type	Texture (%) sand, silt, clay
9 (Schuby) arable land	Weichselian sandur sands	Gleyic Podzol	87.5, 8.2, 4.3 in 0–30 cm
35 (Lindhöft 1) grassland	Weichselian periglacial sandy-loamy cover layers over Weichselian boulder clay	Luvisol	63.1, 28.4, 8.5 in 0–5 cm
36 (Lindhöft 2) arable land	Weichselian periglacial sandy-loamy cover layers over Weichselian boulder sand, silt and clay in the moraine	Luvisol	61.8, 27.6, 10.2 in 0–30 cm

observations (Cu_{AR}, SOC, CEC, and pH) are available for the surface soils. The remaining two sites were excluded from the analysis because the amount of data was too limited for meaningful analyses (site 23) or because the chemical characteristics were very different from other sites (site 6 has considerably higher clay concentrations, pH values, and CEC than the other sites; data not shown). The sites include one arable field in Schuby (site 9) and one arable field (site 36) and one grassland (site 35), both located in Lindhöft (Table 1). The soil types include a gleyic Podzol with sandy texture (site 9) and Luvisols with loamy sand texture (sites 35 and 36; Table 1). Additional information on soil monitoring in Schleswig-Holstein is given by Nerger et al. (2011).

The sites were established as basic monitoring sites in 2000 (sites 35 and 36) or 1989 (site 9). From 2003 (sites 35 and 36) to 2005 (site 9) onward, monitoring has been carried out as intensive monitoring, and annual soil measurements (arable sites 9 and 36: 0–30 cm; grassland site 35: 0–5 cm) have been undertaken since 2006. This study focuses on those dates, for which results for all four variables (Cu_{AR} concentrations, pH, SOC, and CEC) were available from the destructive samplings, that is, from 2003 (sites 35 and 36) or 2005 (site 9) until 2019. During this period, Cu_{AR} concentrations were measured frequently, with different numbers of spatial replications per sample date (Figure 1 shows detailed information). In total, 97, 162, and 85 Cu_{AR} concentrations were determined for the sites 9, 35, and 36, respectively, over the observation periods (Figure 1), and the numbers of complete observations for the four variables amounted to 62, 66, and 52 for the three sites 9, 35, and 36.

At the initial sampling date, soil texture, contents of SOC, and heavy metals as well as pH and CEC were determined using standard methods (Barth et al., 2000). Cu_{AR} concentrations were measured using DIN ISO 11466 (1997). SOC concentrations were determined using a CN element analyzer (DIN ISO 10694, 1996). The pH was measured using a volume ratio of soil to water of 1:5, and the CEC was determined using an unbuffered NH_4Cl solution.

2.2 | Statistical analyses

2.2.1 | General statistical analyses: Correlations and 95% confidence intervals

Statistical analyses were carried out with R version 4.05 (R Core Team, 2021). Correlation analyses between Cu_{AR} concentrations and the measured variables pH, SOC, and CEC were carried out using Pearson correlation coefficients r (in cases of bivariate normality) or Spearman rank correlation coefficients r_s. Normality of the variables given above was tested using the Shapiro–Wilk test.

Boxplots were plotted for Cu_{AR} concentrations and two 95% confidence intervals of differences in means of Cu_{AR} concentrations between two sampling dates were calculated for each site to study whether there was any long-term trend in Cu_{AR} concentrations over time. For each 95% confidence interval of difference in means, Cu_{AR} concentrations of both sampling dates were inspected for normality using the Shapiro–Wilk test, which requires a minimum sample size of 3. Specifically, 95% confidence intervals of differences in means were calculated for the periods involving the last observation dates (2019 for all three sites) and the years 2008, 2005, and 2006 for sites 9, 35, and 36, respectively (Figure 1). Earlier observation dates were not considered due to a lack of normality (2006 at site 9) or sample sizes less than 3 (Figure 1). Additionally, we also calculated 95% confidence intervals of difference in means for the periods involving the penultimate observation dates (2018) and the years 2008, 2005, and 2006 (Figure 1). For all three sites, there was no general trend of increasing or decreasing Cu_{AR} concentrations, as indicated by five out of six 95% confidence intervals. Only for site 36, one 95% confidence interval suggested a slight increase in Cu_{AR}, but the large overall variability of Cu_{AR} concentrations for this site sheds doubts on this interpretation. Since there was no general trend of increasing or decreasing Cu_{AR} concentrations with time, we combined the data of different sampling dates for each site and studied whether the information of site, SOC concentration, pH, and CEC is sufficient to describe and predict Cu_{AR} concentrations in soils. Two different modeling approaches, mixed-effects models and rule-based cubist models, were tested in the following modeling variants: I. Description of the Cu_{AR} concentrations using the entire data set; II. Prediction of the Cu_{AR} concentrations for the closed population of the three sites using random partitioning of the data; III. Prediction of the Cu_{AR} concentration for a new site using a three-fold partitioning of the data according to site; and IV. Prediction of the Cu_{AR} concentrations for a new site using a three-fold partitioning of the data according to site using spiking. The approaches and parameterizations are presented below.

2.2.2 | Description of the Cu_{AR} concentrations using the entire data set

Mixed-effects modeling

For mixed-effects modeling, we used the packages *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017). The predictors SOC,

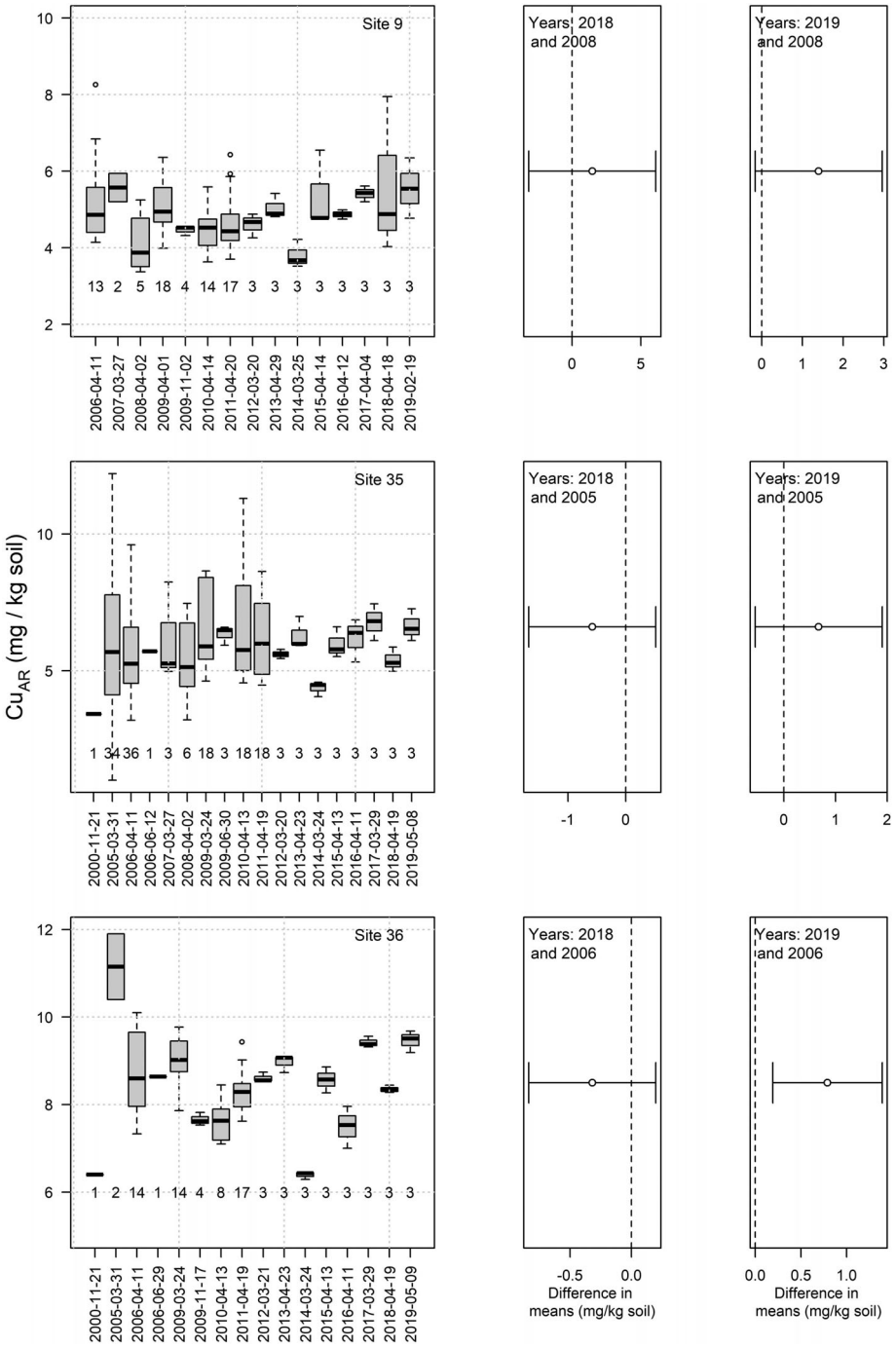

FIGURE 1 Boxplots of the Cu_{AR} concentrations over time for the A horizons of the different sites. Numbers below the boxplots indicate the number of observations per date. Note that 95% confidence intervals for the difference in means of Cu_{AR} concentrations between two sampling dates are shown for cases in which data of the different sampling times were normally distributed. Note that 95% confidence intervals of differences in means which contain 0 are interpreted as nonsignificant differences.

CEC, and pH were included in the regression as fixed effects and the structural component site as a random effect. Besides the first order contributions of the predictors, second order polynomial terms, all two-way interactions and the three-way interaction were considered in the models.

We simplified the model in a stepwise procedure as described by Crawley (2012). First, the three-way interaction was removed if it was nonsignificant. Then, nonsignificant two-way interactions were deleted in a stepwise process and finally nonsignificant second and first order main effects were removed from the model. In summary, fixed effects were only kept in the model in cases of significant ($p \leq 0.05$) contributions. As described by Crawley (2012), nonsignificant effects of the main effects were only included in the case of a significant interaction or a significant second order contribution of the main effects.

TABLE 2 Parameterization and performance of fixed and mixed-effects models[a] for the response variable Cu_{AR} (mg kg^{-1} soil)

Model variant and sample size *n*	Final equation[a]	Random components (assumed mean of 0 and variance)
I. Calibration using the entire data set, $n = 179$	−42.8 + 10.7 SOC + 12.0 pH − 0.42 CEC −0.71 pH^2 + 0.08 CEC^2 −1.68 SOC × pH	Site ~ N(0, 4.8) Residual ~ N(0, 0.42)
II. Calibration after random splitting of the data set, $n = 89$	−12.1 + 8.83 SOC + 1.95 pH + 0.75 CEC −1.37 SOC × pH	Site ~ N(0, 5.5) Residual ~ N(0, 0.44)
IIIa. Calibration using sites 35 and 36, $n = 117$	0.36 + 1.13 SOC + 0.75 CEC	Site ~ N(0, 6.2) Residual N(0, 0.51)
IIIb. Calibration using sites 9 and 35, $n = 127$	−93.0 + 9.61 SOC + 28.8 pH − 0.70 CEC −0.38 SOC^2 −2.21 pH^2 + 0.09 CEC^2 −1.07 SOC × pH	Site ~ N(0, 5.0) Residual ~ N(0, 0.41)
IIIc. Calibration using sites 9 and 36, $n = 114$	−6.89 + 6.13 SOC + 1.41 pH + 0.64 CEC −1.01 SOC × pH	Site ~ N(0, 5.5) Residual ~ N(0, 0.36)
IVa. Calibration using sites 35 and 36, $n = 117 + 4$ spiked observations	−82.5 + 18.5 SOC + 21.1 pH + 0.72 CEC −1.27 pH^2 −2.93 SOC × pH	Site ~ N(0, 4.5) Residual ~ N(0, 0.46)
IVb. Calibration using sites 9 and 35, $n = 127 + 4$ spiked observations	−14.6 + 11.1 SOC + 2.66 pH − 0.58 CEC −0.34 SOC^2 + 0.08 CEC^2 −1.35 SOC × pH	Site ~ N(0, 9.2) Residual ~ N(0, 0.42)
IVc. Calibration using sites 9 and 36, $n = 114 + 4$ spiked observations	−44.9 + 0.07 SOC + 8.08 pH + 7.00 CEC +0.71 SOC^2 + 0.23 CEC^2 −0.69 SOC × CEC − 1.22 pH × CEC	Site ~ N(0, 0.14) Residual ~ N(0, 0.34)

[a]The unit for the intercept is mg kg^{-1} soil. The units for the regression terms are mg kg^{-1} soil multiplied by the respective reciprocals of the units of the variables (first and second order contributions and interactions, SOC: g 100 g^{-1}, CEC: cmol(+) kg^{-1}).
Abbreviations: CEC, cation exchange capacity; SOC, soil organic carbon.

Restricted maximum likelihood was used as estimation procedure for the mixed-effects models, and the Kenward–Roger method was used for the estimation of the denominator degrees of freedom.

Residuals were inspected for homoscedasticity and normality. In total, one extreme value (the maximum Cu_{AR} concentration) had to be removed from the data set to fulfil the distributional requirements for the residuals. Thus, for all modeling approaches, the total number of observations *n* was decreased by 1 to 179. Table 2 shows the final mixed-effects model for variant I.

Rule-based cubist models

For the rule-based cubist modeling, we used packages Cubist (Kuhn & Quinlan, 2021) and caret (Kuhn, 2021). The model was calibrated using 10-fold cross-validation and the number of committee models was optimized using the values 1, 10, 50, and 100. Table 3 shows the optimal number of committee models for a description of Cu_{AR} concentrations.

2.2.3 | Prediction of the Cu_{AR} concentrations for the closed population of the three sites

In this variant, the 179 observations were randomly split into a calibration ($n = 89$) and validation sample ($n = 90$). Since the random split observations of the three sites are present in the calibration and validation sample, the variant is a pseudo-independent calibration–validation. This variant is useful for data sets in which the population of interest is fixed and there is no intention of an extension to new sites. A successful validation indicates a potential for adequate predictions only for the sites included in the study.

Mixed-effects modeling was carried out for the calibration sample as described above including model simplification and residual inspection. The optimal model consisted of the first order contributions of all three fixed effects and additionally of the interaction of SOC and pH, and site as a random effect (Table 2).

For the cubist model, the optimal number of committee models (values of 1, 10, 50, and 100) and the optimal number of neighbors (0, 1, 5, and 9) was obtained in a grid search in the cross-validation approach. Table 3 shows the optimal values for this variant.

2.2.4 | Prediction of the Cu_{AR} concentrations for a new site using a three-fold partitioning of the data according to site

A three-fold partitioning of the data according to site was used for independent calibration–validation (e.g., Brown et al., 2005; Ludwig et al., 2017). In each partition, one site was used as the validation sample. We carried out mixed-effects modeling as described above. The final optimal models differed depending on the sites used in the calibrations. Notably, residual inspection was not successful for variant IIIa, where the residuals were not normally distributed. In all three variants, site was again included as a random effect in the

TABLE 3 Parameterization and performance of rule-based cubist models for the response variable Cu_{AR} (mg kg^{-1} soil)

Model variant and sample size *n*	Number of committees	Number of neighbors from the calibration set data used for the validation	Variable usage in conditions[a]	Variable usage in the model committees[b]
I. Calibration using the entire data set, $n = 179$	10	not applicable	SOC (51%)	SOC (97%), pH (84%), CEC (56%)
II. Calibration after random splitting of the data set, $n = 89$	10	9	SOC (50%)	SOC (100%), pH (23%), CEC (90%)
IIIa. Calibration using sites 35 and 36, $n = 117$	1	0	SOC (100%)	SOC (100%), pH (50%), CEC (100%)
IIIb. Calibration using sites 9 and 35, $n = 127$	10	9	SOC (50%)	SOC (70%), pH (55%), CEC (95%), site (20%)
IIIc. Calibration using sites 9 and 36, $n = 114$	100	0	SOC (40%), pH (11%)	SOC (43%), pH (30%), CEC (53%)
IVa. Calibration using sites 35 and 36, $n = 117$ + 4 spiked observations	10	0	SOC (50%)	SOC (90%), pH (53%), CEC (100%)
IVb. Calibration using sites 9 and 35, $n = 127$ + 4 spiked observations	50	0	Site (48%), SOC (2%)	SOC (35%), pH (27%), CEC (75%), site (38%)
IVc. Calibration using sites 9 and 36, $n = 114 +$ 4 spiked observations	1	9	SOC (100%)	SOC (47%), pH (53%), CEC (100%)

[a]Sum of percentages can be <100% since not all committee models contain conditions or >100% since different variables may contribute to a single condition.
[b]Sum of percentages can be <100% since rules may just contain numbers and no variables or >100% since different variables may contribute to a single rule.
Abbreviations: CEC, cation exchange capacity; SOC, soil organic carbon.

calibration, which meant that the effect was zero for the validation with new sites.

For the cubist models, the grid search for the optimal numbers of committee models and neighbors was carried out as described above.

2.2.5 | Prediction of the Cu_{AR} concentrations for a new site using a three-fold partitioning of the data according to site using spiking

Spiking, which is common in soil spectroscopy (see, e.g., Stenberg et al., 2010), was used in variant IV to provide a small amount of information from the validation sample (site) in the calibration sample. The underlying hypothesis is that with a small analytical extra effort (random sampling of only four soils from a new site and wet-chemistry analysis of the variables Cu_{AR}, pH, SOC, and CEC), the bias in the predictions for the new site may be markedly reduced and accuracy of predictions increased.

The number of observations moved from the validation to the calibration sample was restricted to $n = 4$. The $n = 4$ observations were randomly selected but were identical for the mixed-effects and cubist modeling. Mixed-effects modeling and application of the cubist model were carried out as described above.

2.2.6 | Model performance parameters

Marginal (R^2_m) and conditional (R^2_c) pseudo-coefficients of determination were calculated for the mixed-effects models. They account for the variance explained by fixed effects (R^2_m) and by both fixed and random effects (R^2_c) (Nakagawa et al., 2017). The package *MuMIn* (Barton, 2020) was used for the calculations.

Additionally, we calculated root mean squared errors (RMSEs) and Spearman rank correlation coefficients r_s between measured and modeled Cu_{AR} concentrations (Cu_{AR} concentrations were not normally distributed) for both approaches and all modeling variants. RMSEs are useful for model comparisons. For the interpretation of r_s, an example of a conventional approach to interpreting a correlation coefficient summarized by Schober et al. (2018) can be used, with negligible (correlation coefficient in the range 0.00–0.10), weak (0.10–0.39), moderate (0.40–0.69), strong (0.70–0.89), and very strong correlation (0.90–1.00). However, several authors including Schober et al. (2018) also emphasize that cutoff points are arbitrary and should be used judiciously.

3 | RESULTS

3.1 | Cu_{AR} concentrations in soils

In the three intensively monitored surface soils in northern Germany, there was no general measurable trend of changing Cu_{AR} concentrations with time (Figure 1). Boxplots showed large scatter of the data from replicate samplings for some sampling dates, especially in the early periods of the monitoring. At later stages of the monitoring, when the number of field replicates was reduced, variabilities also decreased. Although medians differed over time, especially for site 36, there was no consistent trend. Note that 95% confidence intervals of differences in means indicated no significant changes over time for five out of the six tested differences in time (i.e., a difference of 0 was part of the 95% confidence intervals). Only for one case at site 36, a slight significant

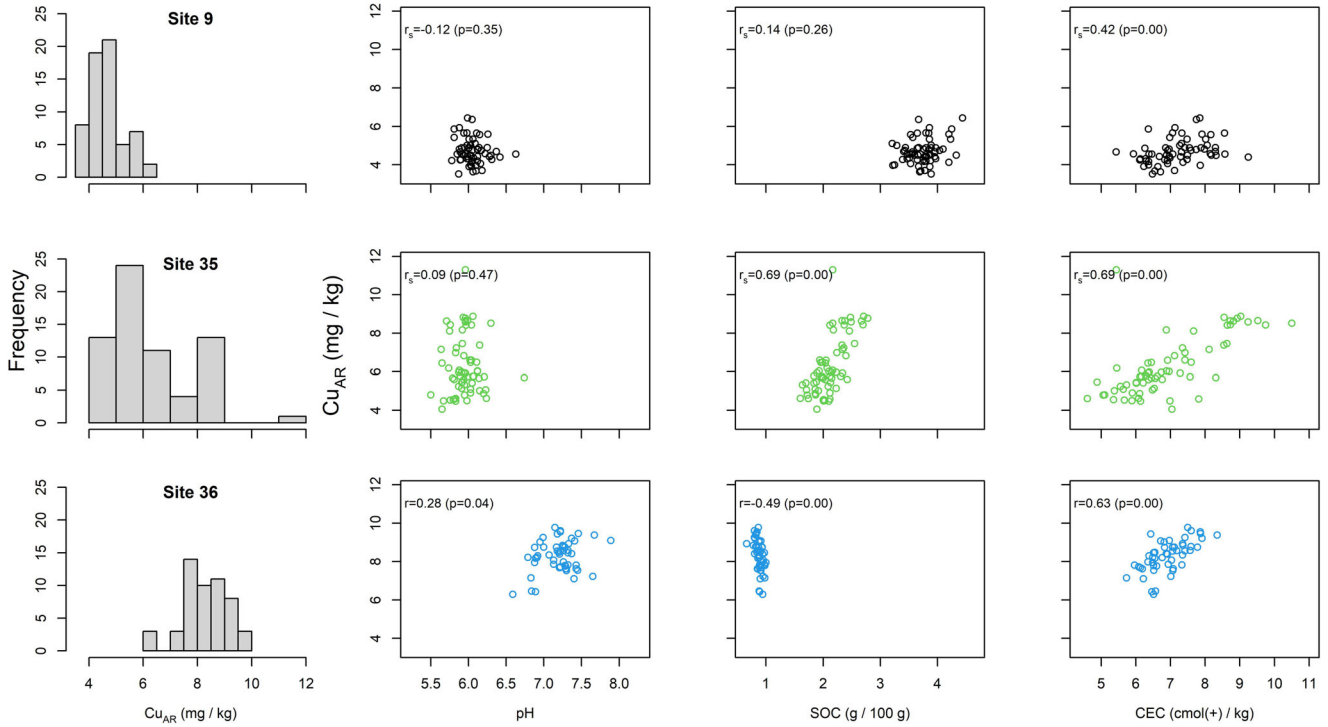

FIGURE 2 Histograms for Cu_{AR} for the surface soils of the three intensively monitored sites. Scatter plots for Cu_{AR} and pH, soil organic carbon (SOC) content, and the cation exchange capacity (CEC) are also shown. Different colors indicate soils from different sites. Spearman rank correlation coefficients r_s and Pearson correlation coefficients r are also shown, with p-values indicating the significance of the correlations.

increase was noted, but no general trend was observed (Figure 1). The absence of a general trend can be explained by the uses of the soils as arable land or grassland without excessive application of Cu pesticides or high organic fertilization. Data were therefore combined for the different sampling times.

Cu_{AR} concentrations in surface soils were approximately normally distributed at site 36 and right-skewed at the other sites (Figure 2). Scatter plots showed positive, negative, and no relationships between Cu_{AR} concentration and the measured variables (Figure 2). Spearman rank correlation coefficients r_s (sites 9 and 35) and Pearson correlation coefficients r (site 36) indicated significant moderate positive relationships between Cu_{AR} and CEC ($r_s = 0.42$ and 0.69 for sites 9 and 35 and $r = 0.63$ for site 36; Figure 2). A significant weak positive relationship between Cu_{AR} and pH was found only for site 36 ($r = 0.28$; Figure 2). For Cu_{AR} and SOC, there was a significant moderate positive relationship ($r_s = 0.69$) for site 35, but a significant negative one ($r = -0.49$) for site 36 (Figure 2).

3.2 | Variant I. Description of Cu_{AR} concentrations depending on site, SOC, pH and CEC

A mixed-effects model was very useful to describe the Cu_{AR} concentrations in the surface soils of the three sites as a function of the fixed effects of CEC, SOC, and pH, and the random effect of site (Figure 3). The difference between the conditional pseudo-coefficient of determination R^2_c (0.93) and marginal R^2_m (0.14) indicated the importance of site for the successful modeling (see also the large variance of the site effect [4.8 mg^2/kg^2] relative to the residual variance [0.42 mg^2/kg^2; Table 2]). For all fixed effects, relationships with Cu_{AR} had positive (SOC, pH, CEC^2) and negative (CEC, pH^2, SOC × pH) contributions (Table 2), indicating a change of the direction of the effects on Cu_{AR} with increasing values of the fixed effects.

The cubist model was as successful as the mixed-effects model in describing the Cu_{AR} concentrations (Figure 3). In the model, site was not required (Table 3)

3.3 | Variant II. Prediction of Cu_{AR} concentrations for a closed population

In variant II, where the data set was randomly split into a calibration and validation sample (i.e., pseudo-independent validation), the mixed-effects model and the cubist model were similarly successful as in variant I in calibration (Figure 3). However, final models differed due to the reduced information available in variant II compared to variant I. In the validation, mixed-effects modeling slightly outperformed the cubist modeling as indicated by the slightly lower RMSE and slightly higher r_s of the mixed-effects modeling (Figure 3).

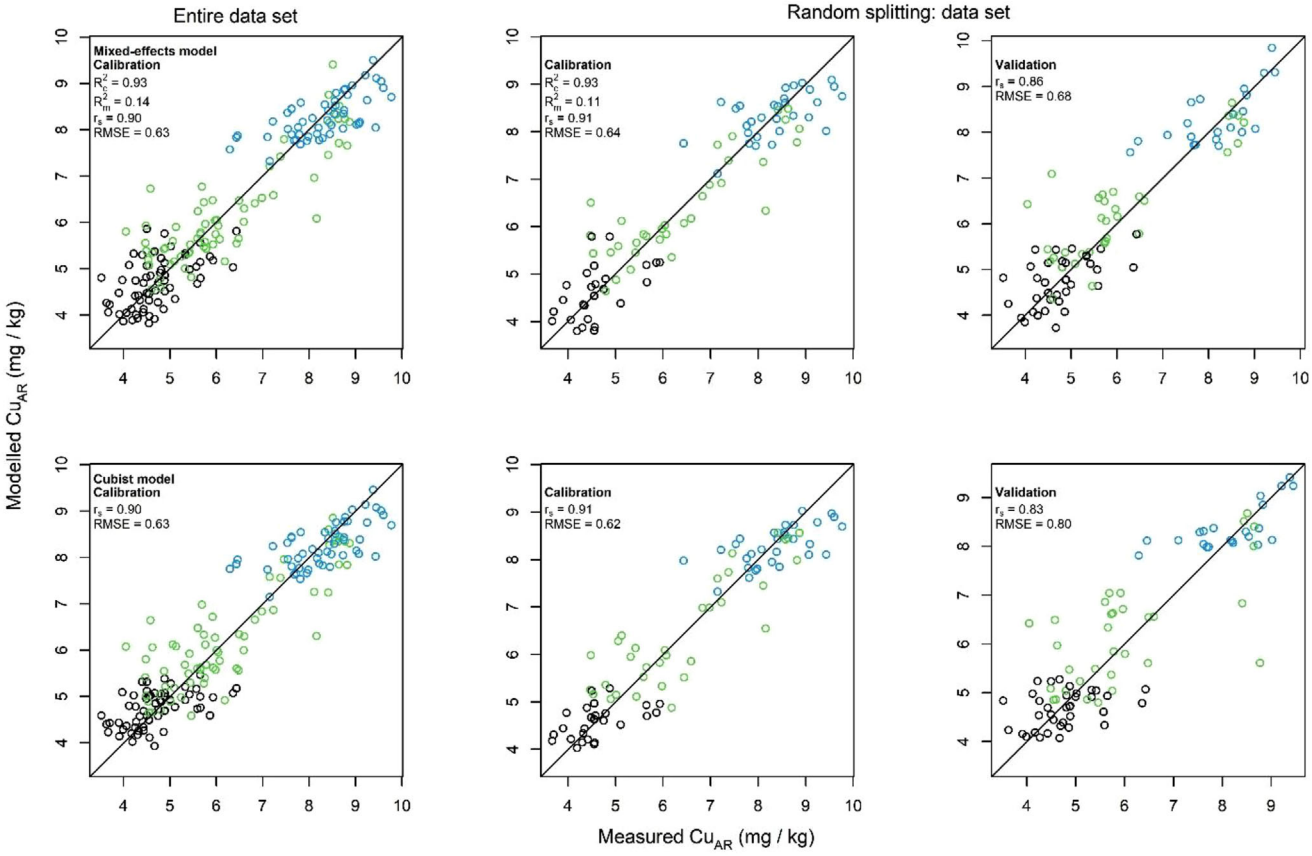

FIGURE 3 Modeled versus measured Cu_{AR} concentrations resulting from mixed-effects models (top) and rule-based cubist models (bottom) using the entire data set (left, variant I) or after random-splitting of the data set (variant II, middle: calibration, right: validation). Conditional and marginal pseudo-coefficients of determination are indicated as R^2_c and R^2_m. Additionally, Spearman's rank correlation coefficient r_s and the root mean squared error (RMSE) in mg kg^{-1} are given. Different colors indicate soils from different sites.

3.4 | Variants III and IV. Prediction of the Cu_{AR} concentrations for a new site using a three-fold partitioning of the data according to site with and without spiking

In variant III, we tested the possibility of independent predictions of Cu_{AR} concentrations for a new site using a calibration equation based on two different sites. Again, mixed-effects modeling was very successful in the calibration with R^2_c of 0.94 for all three partitions of the sites (variant IIIa–IIIc; Table 2, Figure 4) and RMSEs ranging from 0.58 to 0.70 mg kg^{-1} (Figure 4). Independent validations, however, were generally unsuccessful, with high RMSEs in variant IIIa (validation using site 9) and IIIb (validation using site 36). In contrast, for variant IIIc, validation was successful (r_s = 0.76, RMSE = 0.82 mg kg^{-1}; Figure 4). Cubist modeling showed the same pattern: all three calibrations in the variants IIIa–IIIb were successful, but independent validations were unsuccessful in variant IIIa and IIIb. Moreover, validation in variant IIIc was less successful compared to mixed-effects modeling for that variant due to bias in the estimates (Figure 4).

In variant IV (Iva–IVc), spiking was carried out using only four observations from the respective validation site (thus, in all variants, four observations were shifted from the validation samples to the calibration samples). For mixed-effects models, these shifts had only minor effects on the performance data (R^2_c, r_s and RMSE) in the three calibrations (Figure 5), but considerably affected the underlying equations (Table 2). Independent validations after calibration with spiking showed marked reductions in the RMSEs in variants IIIa and IIIb, whereas for variant IIIc, a slight increase was noted (Figure 5). Cubist modeling was also very successful in the three calibrations. In the independent validations, RMSEs decreased markedly in all three variants Iva–IVc compared to IIIa–IIIc, which shows the benefits of spiking with only a small number of observational units from the new site.

4 | DISCUSSIONS

4.1 | Cu_{AR} concentrations in soils and control variables

The Cu_{AR} concentrations in the surface soils of the intensively monitored sites in northern Germany ranged from 1.0 to 12.2 mg kg^{-1} in

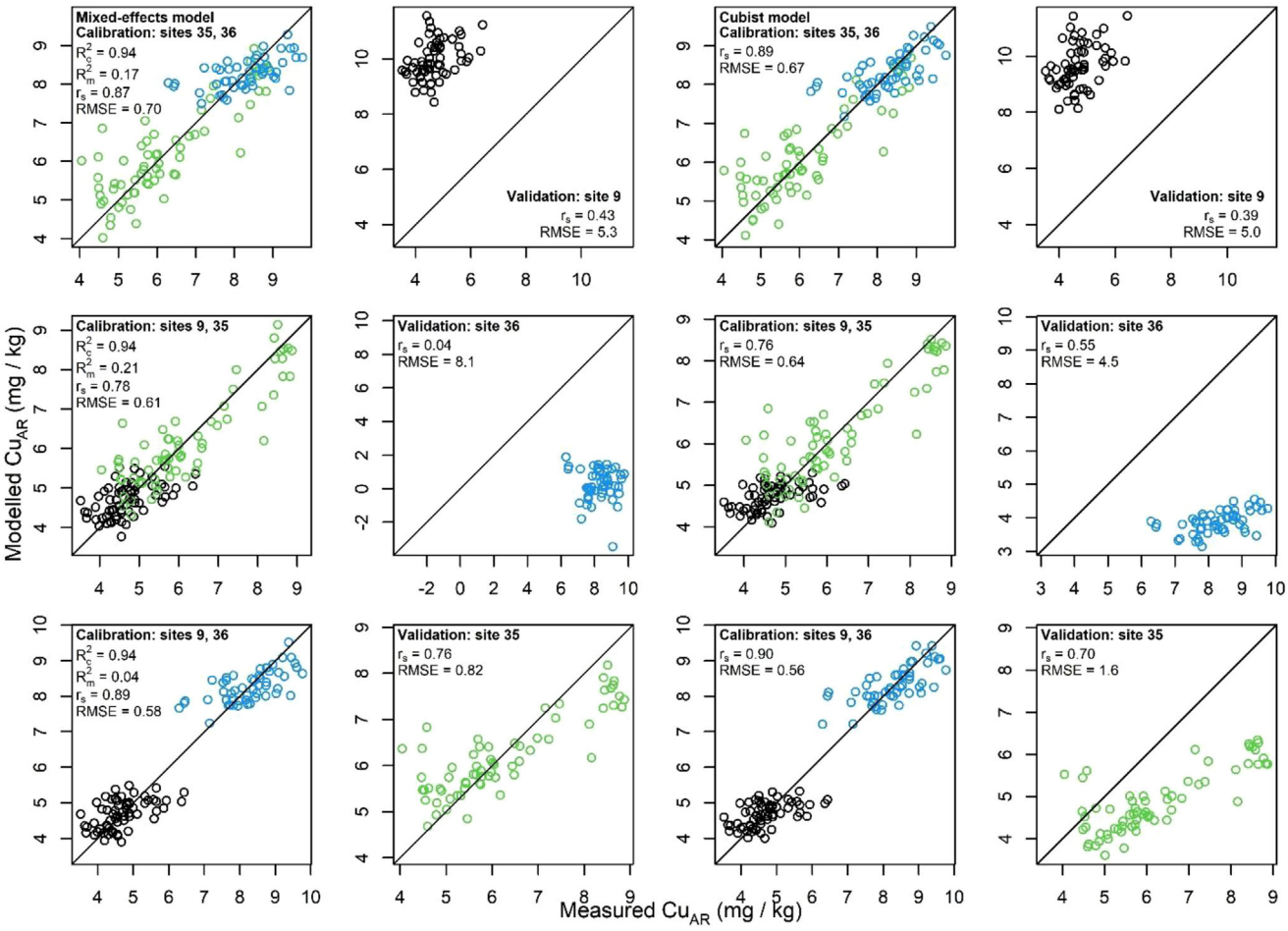

FIGURE 4 Modeled versus measured Cu_{AR} concentrations of the mixed-effects models (first two columns) and rule-based cubist models (last two columns) for three-fold partitioning of the data according to site (variant III). Additionally, R^2_c, R^2_m, Spearman's rank correlation coefficient r_s, and the root mean squared error (RMSE) in mg kg^{-1} are given. Different colors indicate soils from different sites.

the observation period from 2000 to 2019, indicating the absence of the main pollution sources such as high application rates of Cu pesticides (as in viticulture), high organic fertilization, or any metallurgical or mining sources, where much higher concentrations are observed (Reinhold, 2008).

Predictions of Cu_{AR} concentrations using CEC, SOC concentrations, and pH appeared to be promising, not only because of their relationships with Cu_{AR} for the three monitoring sites (significant correlation coefficients for the pairs Cu_{AR}/CEC [three sites], Cu_{AR}/SOC [two sites], and Cu_{AR}/pH [one site], see above), but also due to their known role in Cu dynamics. The importance of CEC for the Cu_{AR} concentrations may be mainly indirect, since CEC and clay contents are typically closely related (e.g., Amelung et al., 2018) and soil with increased clay contents have been reported to have higher Cu contents (Huschek et al., 2004). SOC may be useful for predicting Cu_{AR} concentrations because of the key role of soil organic matter in Cu retention and a reported large fraction of organically bound Cu in soils (Amelung et al., 2018; Fijałkowski et al., 2012). The contribution of pH as a control variable for Cu_{AR} concentrations appears plausible because of the pH dependency of Cu adsorption and desorption processes in soils (Caporale & Violante, 2016; Sposito, 2008).

4.2 | Variant I. Description of Cu_{AR} concentrations depending on site, SOC, pH, and CEC

Description of Cu_{AR} concentrations using a mixed-effects model was very successful for the entire data set consisting of the three sites, indicating the usefulness of the variables site, SOC, pH, and CEC. Mixed-effects models are typically powerful modeling tools (Galecki & Burzykowski, 2013) and are the method of choice in many regression studies in soil science, since they are able to adequately consider hierarchical sampling designs as in this monitoring study with three independent sites and large numbers of observations per site.

The rule-based cubist model was as successful as the mixed-effects model, as indicated by the same Spearman correlation coefficient r_s of 0.90 which indicates a strong correlation between measured and modeled Cu_{AR} concentrations and RMSE of 0.63 mg kg^{-1}. Thus, hypothesis

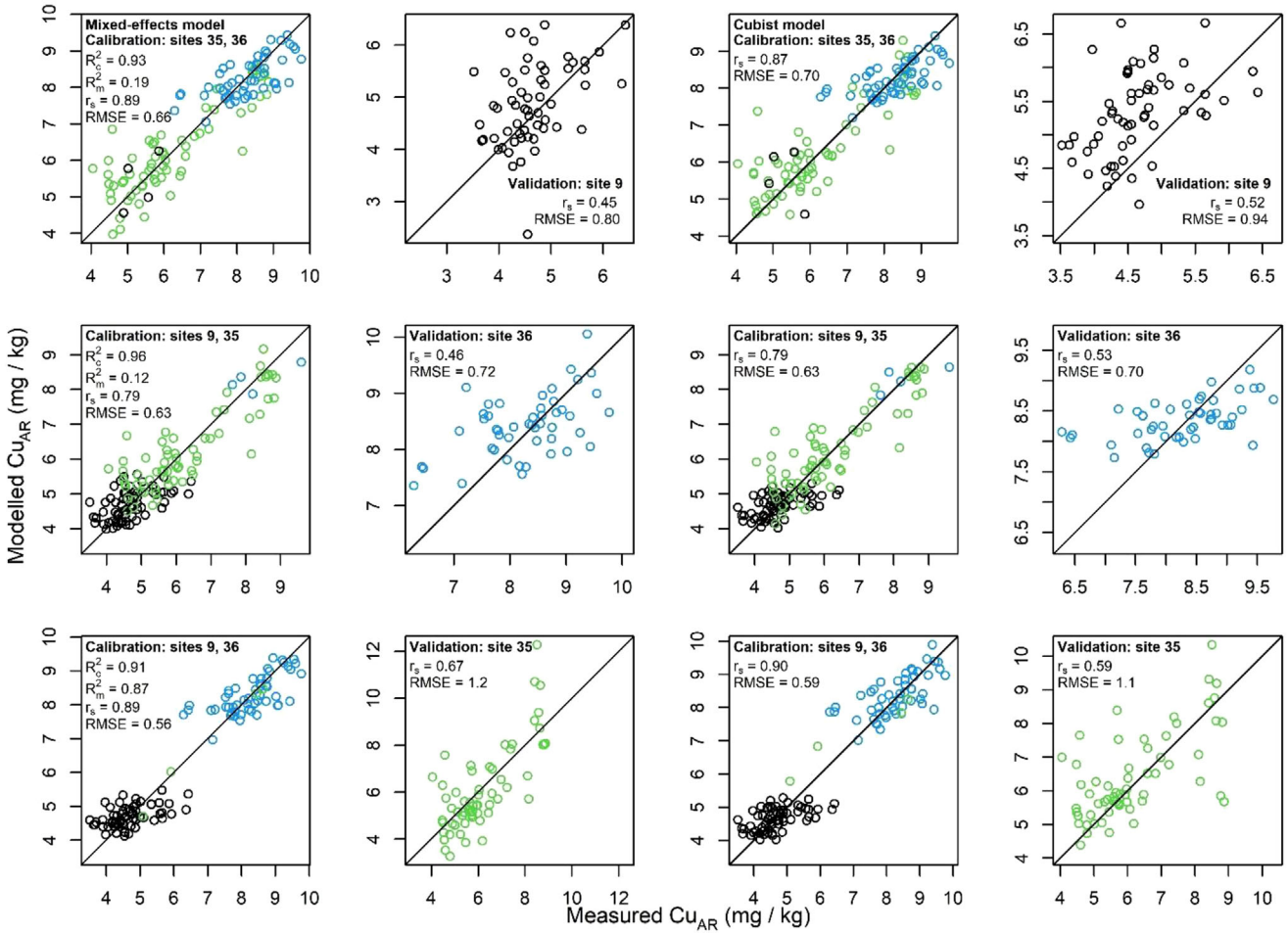

FIGURE 5 Modeled versus measured Cu_{AR} concentrations of the mixed-effects models (first two columns) and rule-based cubist models (last two columns) for three-fold partitioning of the data according to site with spiking (variant IV). Additionally, R^2_c, R^2_m, Spearman's rank correlation coefficient r_s, and the root mean squared error (RMSE) in mg kg^{-1} are given. Different colors indicate soils from different sites.

1 was supported: mixed-effects models and rule-based models were similarly successful in predicting Cu_{AR} contents. The very good performance of the cubist model is due to the use of rules in the cubist models, which can be simplified or pruned in a way that observations are covered by multiple rules and the optional boosting-like procedure called committees. For variants II–IV, which included validation predictions, this modeling approach also provides the possibility to adjust predictions generated by the model rules using nearby points from the calibration sample to create successful models (Kuhn & Johnson, 2018).

We hypothesized that both algorithms may point to a similar importance of the independent variables included in the final models. However, this hypothesis was not confirmed. For instance, in contrast to the mixed-effect model, the site effect was not important in variant I for the cubist model. A main reason for a different importance of independent variables for the two algorithms may be the multicollinearity in the data set (Wehrens, 2020; Welham et al., 2014), which is typical in observational studies. For instance, the comparison of the scatter plots of Cu_{AR} versus SOC and Cu_{AR} versus CEC for site 35 indicates a high correlation between SOC and CEC ($r_s = 0.66$).

4.3 | Variant II. Prediction of Cu_{AR} concentrations for a closed population

In agreement with our hypothesis 3a, Cu_{AR} concentrations were predicted successfully in a closed population (i.e., after random splitting of the data set) using either a mixed-effects model or a cubist model. Thus, in future samplings at the three sites, Cu_{AR} concentrations can likely be predicted with high accuracy as a function of site, SOC, CEC, and pH in this fixed set of observational sites. However, the random split of the data set into a calibration and validation sample is only a pseudo-independent calibration–validation approach, since the three sites are present in both the calibration and validation sample (see, e.g., Brown et al. [2005] and Ludwig et al. [2017]).

4.4 | Variants III and IV. Prediction of the Cu_{AR} concentrations for a new site using a three-fold partitioning of the data according to site with and without spiking

We hypothesized that Cu_{AR} concentrations may be successfully predicted for a new site (with an assumed reduced accuracy), and tested this in a three-fold calibration–validation approach. For both algorithms, mixed-effects, and cubist models, this hypothesis was not supported in two out of the three folds, indicating the importance of the variable site (which aggregates, among others factors, information on parent material and specific mineral composition) for predictions. The best performance was observed when validation was carried out for site 35 with Cu_{AR} concentrations in the intermediate range, that is, when the sites with smallest (site 9) and highest Cu_{AR} concentrations (site 36) were used in the calibration. Thus, models were most useful when the calibration set included all variability present in the validation set.

Variants Iva–IVc indicated that the use of only four observations of the new site in the calibrations, that is, spiking, was sufficient to markedly improve the quality of the subsequent predictions as indicated by the generally (with one exception) considerable decreases of the RMSEs after spiking. Thus, only little analytical extra-effort (four samples in this study) is required to extend an existing calibration for a new site.

ACKNOWLEDGMENT

The *federal-state funding program "Water, Soil and Waste"* of the *Ministry of Agriculture and Environment Mecklenburg-Western Pomerania* contributed to this study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Bernard Ludwig https://orcid.org/0000-0001-8900-6190
Isabel Greenberg https://orcid.org/0000-0002-4762-8474
Hans-Peter Piepho https://orcid.org/0000-0001-7813-2992

REFERENCES

Adriano, D. C. (2005). *Trace elements in terrestrial environments:* Biogeochemistry, bioavailability, and risks of metals. Springer.

Amelung, W., Blume, H. P., Fleige, H., Horn, R., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K., & Wilke, B. M. (2018). *Scheffer/Schachtschabel—Lehrbuch der Bodenkunde, 17. Auflage*. Springer.

Anatole-Monnier, L. (2014). *Effets de la contamination cuprique des sols viticoles sur la sensibilité de la vigne à un cortège de bio-agresseurs*. Ecologie, Environnement. Université de Bordeaux.

Baker, D. E., & Senft, J. P. (1995). Copper. In B. J. Alloway (Ed.), *Heavy metals in soils* (2nd ed., pp. 177–205). Blackie Academic and Professional.

Ballabio, C., Panagos, P., Lugato, E., Huang, J.-H., Orgiazzi, A., Jones, A., Fernández-Ugalde, O., Borrelli, P., & Montanarella, L. (2018). Copper distribution in European topsoils: An assessment based on LUCAS soil survey. *Science of The Total Environment*, *636*, 282–298.

Barth, N., Brandtner, W., Cordsen, E., Dann, T., Emmerich, K. -. H., Feldhaus, D., Kleefisch, B., Schilling, B., & Utermann, J. (2000). Boden-Dauerbeobachtung. Einrichtung und Betrieb von Boden-Dauerbeobachtungsflächen. In D. Rosenkranz, G. Bachmann, W. König, & G. Einsele (Eds.), *Bodenschutz—Ergänzbares Handbuch der Maßnahmen und Empfehlungen für Schutz, Pflege und Sanierung von Böden, Landschaft und Grundwasser* (pp. 10150–11127). Erich Schmidt Verlag.

Barton, K. (2020). MuMIn: Multi-model inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bigalke, M., Weyer, S., Kobza, J., & Wilcke, W. (2010). Stable Cu and Zn isotope ratios as tracers of sources and transport of Cu and Zn in contaminated soil. *Geochimica et Cosmochimica Acta*, *74*(23), 6801–6813.

Blotevogel, S., Oliva, P., Sobanska, S., Viers, J., Vezin, H., Audry, S., Prunier, J., Darrozes, J., Orgogozo, L., Courjault-Radé, P., & Schreck, E. (2018). The fate of Cu pesticides in vineyard soils: A case study using δ^{65}Cu isotope ratios and EPR analysis. *Chemical Geology*, *477*, 35–46.

Brown, D. J., Bricklemyer, R. S., & Miller, P. R. (2005). Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. *Geoderma*, *129*(3–4), 251–267.

Caporale, A. G., & Violante, A. (2016). Chemical processes affecting the mobility of heavy metals and metalloids in soil environments. *Current Pollution Reports*, *2*(1), 15–27.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107.

Crawley, M. J. (2012). *The R book* (2nd Ed.). Wiley.

Core Team, R. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.r-project.org/

DIN ISO 10694 (1996). *Bodenbeschaffenheit—Bestimmung von organischem Kohlenstoff und Gesamtkohlenstoff nach trockener Verbrennung (Elementaranalyse)*. Beuth.

DIN ISO 11466 (1997). *Bodenbeschaffenheit—Extraktion in Königswasser löslicher Spurenelemente*. Beuth.

Droz, B., Payraudeau, S., Martín, J. A. R., Tóth, G., Panagos, P., Montanarella, L., Borrelli, P., & Imfeld, G. (2021). Copper content and export in European vineyard soils influenced by climate and soil properties. *Environmental Science & Technology*, *55*(11), 7327–7334.

EFSA (2008). Peer review report on copper compounds. 1–414. European Food Safety Authority.

Fijałkowski, K., Kacprzak, M., Grobelak, A., & Placek, A. (2012). The influence of selected soil parameters on the mobility of heavy metals in soils. *Inżynieria i Ochrona Środowiska*, *15*, 81–92.

Galecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R: A step-by-step approach*. Springer.

Groenenberg, J. E., Römkens, P. F. A. M., & de Vries, W. (2006). Prediction of the long term accumulation and leaching of copper in Dutch agricultural soils: A risk assessment study. Alterra-rapport 1278. Alterra.

Huschek, G., Krengel, D., Kayser, M., Bauriegel, A., & Burger, H. (2004). Länderübergreifende Auswertung von Daten der Boden-Dauerbeobachtung der Länder. Forschungsbericht 201 71 244. Umweltbundesamt.

Kamermann, D., Groh, H., & Höper, H. (2015). *Schwermetallein- und -austräge niedersächsischer Boden-Dauerbeobachtungsflächen*. Landesamt für Bergbau, Energie und Geologie.

Kuhn, M. (2021). caret: Classification and regression training. R package version 6.0–88. https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2018). *Applied predictive modelling*. Springer.

Kuhn, M., & Quinlan, R. (2021). Cubist: Rule- and instance-based regression modeling. R package version 0.3.0. https://CRAN.R-project.org/package=Cubist

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Lantz, B. (2019). *Machine learning with R*. Packt Publishing.

Ludwig, B., Vormstein, S., Niebuhr, J., Heinze, S., Marschner, B., & Vohland, M. (2017). Estimation accuracies of near infrared spectroscopy for general soil properties and enzyme activities for two forest sites along three transects. *Geoderma*, *288*, 37–46.

Mallants, D., Šimůnek, J., van Genuchten, M. T., & Jacques, D. (2017). Simulating the fate and transport of coal seam gas chemicals in variably-saturated soils using HYDRUS. *Water*, *9*(6), 385. https://doi.org/10.3390/w9060385

Michel, K., Roose, M., & Ludwig, B. (2007). Comparison of different approaches for modelling heavy metal transport in acidic soils. *Geoderma*, *140*(1–2), 207–214.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nerger, R., Schimming, C. - G., & Fohrer, N. (2011). Boden-Dauerbeobachtung Schleswig-Holstein: Auswertung der Projektergebnisse im Hinblick auf Aussagen zu Veränderungen von Böden, Aussagefähigkeit und Optimierung der eingesetzten Untersuchungsverfahren, Flintbek. Landesamtes für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein.

Panagos, P., Ballabio, C., Lugato, E., Jones, A., Borrelli, P., Scarpa, S., Orgiazzi, A., & Montanarella, L. (2018). Potential sources of anthropogenic copper inputs to European agricultural soils. *Sustainability*, *10*(7), 2380. https://doi.org/10.3390/su10072380

Reinhold, J. (2008). Ursachenforschung und Limitierungsstrategien für zunehmende Kupfergehalte in Bioabfällen. Forschungsbericht 204 33 321. Umweltbundesamt.

Romić, M., Romić, D., & Ondrašek, G. (2004). Heavy metals accumulation in topsoils from the wine-growing regions part 2. Relationships between soil properties and extractable copper contents. *Agriculturae Conspectus Scientificus*, *69*(2–3), 35–41.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768.

Sposito, G. (2008). *The chemistry of soils* (2nd ed.). Oxford University Press.

Stenberg, B., Rossel, R. A. V., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*, *107*, 163–215.

Wehrens, R. (2020). *Chemometrics with R* (2nd ed.). Springer.

Welham, S. J., Gezan, S. A., Clark, S. J., & Mead, A. (2014). *Statistical methods in biology. Design and analysis of experiments and regression*. CRC Press.

Wu, J., Norvell, W. A., Hopkins, D. G., Smith, D. B., Ulmer, M. G., & Welch, R. M. (2003). Improved prediction and mapping of soil copper by kriging with auxiliary data for cation-exchange capacity. *Soil Science Society of America Journal*, *67*(3), 919–927.

Zhou, X., He, Z., Liang, Z., Stoffella, P. J., Fan, J., Yang, Y., & Powell, C. A. (2011). Long-term use of copper-containing fungicide affects microbial properties of citrus grove soils. *Soil Science Society of America Journal*, *75*(3), 898–906.

How to cite this article: Ludwig, B., Klüver, K., Filipinski, M., Greenberg, I., Piepho, H.-P., & Cordsen, E. (2022). Description and prediction of copper contents in soils using different modeling approaches—Results of long-term monitoring of soils of northern Germany. *Journal of Plant Nutrition and Soil Science*, *185*, 876–887. https://doi.org/10.1002/jpln.202200075