Received: 11 November 2021 | Revised: 17 May 2022 | Accepted: 23 May 2022

DOI: 10.1111/ejss.13258

ORIGINAL ARTICLE

European Journal of Soil Science WILEY

Application of mixed-effects modelling and rule-based models to explain copper variation in soil profiles of southern Germany

Bernard Ludwig[1] | **Petra Wölfel**[2] | **Isabel Greenberg**[1] | **Hans-Peter Piepho**[3] | **Peter Spörlein**[2]

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

[2]Bavarian Environment Agency, Department Geological Survey, Unit Soil Mapping, Soil Protection, Hof, Germany

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

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

Funding information
Bavarian State Ministry of the Environment and Consumer Protection; Ministry of Agriculture and Environment Mecklenburg-Western Pomerania

Abstract

Copper (Cu) is an essential element for plants and microorganisms and at larger concentrations a toxic pollutant. A number of factors controlling Cu dynamics have been reported, but information on quantitative relationships is scarce. We aimed to (i) quantitatively describe and predict soil Cu concentrations (Cu_{AR}) in aqua regia considering site-specific effects and effects of pH, soil organic carbon (SOC) and cation exchange capacity (CEC), and (ii) study the suitability of mixed-effects modelling and rule-based models for the analysis of long-term soil monitoring data. Thirteen uncontaminated long-term monitoring soil profiles in southern Germany were analysed. Since there was no measurable trend of increasing Cu_{AR} concentrations with time in the respective depth ranges of the sites, data from different sampling dates were combined and horizon-specific regression analyses including model simplifications were carried out for 10 horizons. Fixed- and mixed-effects models with the site as a random effect were useful for the different horizons and significant contributions (either of main effects or interactions) of SOC, CEC and pH were present for 9, 8 and 7 horizons, respectively. Horizon-specific rule-based cubist models described the Cu_{AR} data similarly well. Validations of cubist models and mixed-effects models for the Cu_{AR} concentrations in A horizons were successful for the given population after random splitting into calibration and validation samples, but not after independent validations with random splitting according to sites. Overall, site, CEC, SOC and pH provide important information for a description of Cu_{AR} concentrations using the different regression approaches.

Highlights

- Information on quantitative relationships for factors controlling Cu dynamics is scarce

- Site, CEC, SOC and pH provide important information for a description of Cu concentrations
- Validations of cubist models and mixed-effects models for A horizons were successful for a closed population of sites

KEYWORDS

heavy metals, independent validation, mixed-effects models, pseudo-independent validation, rule-based cubist models, soil monitoring

1 | INTRODUCTION

Copper is an essential element for all forms of life and at larger concentrations a toxic pollutant for, among others, plants and microorganisms. In all living organisms and in the soil, the most important Cu oxidation state is +II. Soil Cu consists of geogenic Cu and additional anthropogenic inputs (especially organic fertilisers and Cu-containing pesticides) depending on the anthropogenic activity (i.e., agricultural management, mining or industrial activity). Since Cu mobility in soils is very low (the apparent mean transport rate of anthropogenic Cu was estimated to be approximately 1 cm/year (Bigalke et al., 2010; Bloetevogel et al., 2018)), anthropogenic activities may mostly affect surface soil Cu concentrations. Prediction of Cu concentrations in soils is difficult, not only because of different anthropogenic input sources, but also because of a large number of factors controlling Cu dynamics, such as site-specific soil properties (determined by the parent material, mineralogical composition, and soil texture), precipitation/dissolution of sparingly soluble salts, concentrations of soil organic carbon (SOC), pH, cation exchange capacity (CEC), and extractable Fe, Al and Mn concentrations. The effects of these factors have been investigated in detailed analytical studies, which will be summarised below. However, quantitative information on the relationship between Cu concentrations and the controlling factors is scarce.

Site-specific soil properties determined by parent material, mineralogical composition, and soil texture may affect Cu dynamics considerably. It is well established that geogenic Cu concentrations depend on the Cu present in the parent material (EFSA, 2008). The importance of texture for Cu dynamics is evident from the Cu concentrations in aqua regia (Cu_{AR}) for the main texture classes of a data set of 624 German soil monitoring sites: in sandy soils, median Cu_{AR} concentrations were small (7 and 5 mg/kg for arable and grassland soils, respectively) compared to loamy (15 and 21), silty (17 and 23) and clayey (19 and 21) soils. For forest soils, there was also a gradient from larger concentrations in clayey soils to smaller concentrations in all other soils (UBA, 2004). The importance of texture, which is simply a particle size fraction, for Cu dynamics can be explained by the presence of clay minerals and metal hydroxides in the clay-sized fraction in European soils, which are known to be of great importance for Cu dynamics at most sites (Bloetevogel et al., 2018).

Several sparingly soluble Cu salts exist, but these are important only at sites with very large Cu concentrations, especially in sandy soils with low sorption (e.g., precipitation of malachite, cuprite or copper [I] sulfide), carbonate-containing soils (e.g., precipitation of $CuCO_3$) or in Cu deposit sites (e.g., chalcopyrite) (Bloetevogel et al., 2018).

In addition to clay minerals and metal hydroxides, SOC also affects Cu sorption and thus Cu dynamics. In soils, there is a large fraction of organically-bound Cu and soil organic matter plays an important role in Cu retention (Fijałkowski et al., 2012). Cu dynamics may be considerably affected by the pH because of the pH dependency of sorption, complexation, solubilisation, and desorption processes in soils (Caporale & Violante, 2016; Dinic et al., 2019). The CEC has also been reported to be associated with Cu dynamics, but only to a minor extent. This may be due to actual cation exchange reactions, and mostly indirectly due to the typically close positive relationship between CEC and clay concentration of soils. Additional control factors are dithionite-extractable Fe and Al concentrations and oxalate-extractable Mn concentration, and surface complexation models may describe the Cu ion reactions with the metal (hydr)oxides (Groenenberg & Lofts, 2014; Peng et al., 2018).

Soil monitoring (see Barth et al., 2000 for detailed information on its main aims) may contribute to scientific advances for field-scale predictions of Cu_{AR}, since multivariate data sets with Cu_{AR} and several of the predictors given above are available at high spatial and temporal resolution for several sites. For our study, high-resolution data on Cu_{AR}, SOC concentration, pH and CEC for a number of horizons in soil profiles in southern Germany were available. We used this information in

combination with the variable "site", which aggregates various site-specific variables (e.g., parent material and specific mineral composition), for Cu_{AR} regressions.

Different regression approaches exist, each with particular advantages and limitations. The usefulness may depend on the scale of the variables, the presence of collinearity among the predictors, sample size, sampling design, and research aims. Mixed-effects models are powerful modelling tools (Galecki & Burzykowski, 2013) and may be the method of choice in many studies for elucidating potential relationships between a response variable and independent variables in soil science. However, for this approach, requirements are normality of residuals and—if not explicitly implemented otherwise—variance homogeneity (e.g., Welham et al., 2014; but see discussions by Schielzeth et al. (2020) and Knief and Forstmeier (2021)). Such requirements, nevertheless, can be overcome by using generalised linear mixed models (Stroup, 2012).

In machine learning approaches, such as regression trees and rule-based models, the focus is on cross-validation or calibration–validation approaches rather than on residual inspections (Lantz, 2019). Such approaches may be useful especially for large data sets, which provide an adequate training of the algorithms during the calibration. Regression trees are especially useful for complex, nonlinear relationships among independent variables and the response variable. Moreover, in contrast to regression modelling using fixed or mixed effects, no distributional assumptions about the data are made (Lantz, 2019). Rule-based models, such as the cubist model, differ from regression trees not only in the splitting criterion but also because the terminal nodes (i.e., leaves in regression trees, which contain simple averages of the response variable) contain linear regression models (Kuhn & Quinlan, 2021; Lantz, 2019). A problem—compared to mixed-effects modelling—may be that temporal dependencies or a hierarchical (multi-stratum) sampling design or both are not adequately considered.

The objectives of this study were (i) to quantitatively describe and predict soil Cu_{AR} concentrations considering site-specific effects (which are determined by parent material, mineralogical composition and texture) and effects of pH, SOC and CEC, and (ii) to study the suitability of mixed-effects modelling and rule-based models for the data analysis using long-term monitoring soil data. The predictive ability of the two algorithms was tested in five-fold partitions of the data set in calibration–validation approaches with either random splits of the data set or with random splits according to the site for subsequent predictions at either existing sites or new sites, respectively.

2 | MATERIALS AND METHODS

2.1 | Monitoring and soil analyses

Thirteen locations in Bavaria, southern Germany, so-called focus areas, were selected for soil monitoring with a high-resolution timeline, primarily to detect the background noise in short periods of time and thus characterise them (Figure 1). A selection criterion was that most sites should represent the natural background level, that is, that Cu_{AR} concentrations are predominantly geogenic and thus mostly not related to anthropogenic sources. Background levels for inorganic substances include a geogenic component—that is, the substance content of the soil resulting from the parent rock (lithogenic component) and the redistribution (enrichment and depletion) of substances in the soil influenced by pedogenetic processes—and the ubiquitous substance distribution as a result of diffuse inputs into the soil (Bayerisches Landesamt für Umwelt, 2011). Land uses included one extensive grassland, two intensive grasslands, three coniferous forests, four pastures, one deciduous forest, one litter meadow, and a natural fen. The soils of the sites covered a range of parent materials, soil types (three Cambisols (Ca_1 to Ca_3), two Fluvisols (Fl_1 and Fl_2) a Gleysol (Gl), two Histosols (Hi_1 and Hi_2), four Leptosols (Le_1 to Le_4) and a Luvisol (Lu)) and textures (Table 1). Concentrations of Cu_{AR} in the topsoil of the sites were mostly low, except for sites Gl (a silt loam Gleysol) and Fl_2 (a sandy loam Fluvisol), where increased Cu_{AR} concentrations of 66 and 77 mg/kg, respectively, were present (Table 1). The parent material of site Fl_2, the site with the highest Cu_{AR} concentrations, is a Holocene river sediment of the Regnitz. This sediment was deposited shortly after the confluence of the Pegnitz and the Rednitz, and both rivers previously flowed through the cities of Nürnberg and Fürth. Thus, the increased levels of Cu_{AR} are quite common for river sediments in the Nürnberg/Fürth conurbation due to their proximity to industry.

At each site, soils were analysed to approximately 1 m (see Table 1 for specific depths, Figure 2) and soil properties were determined horizon-wise according to the German soil taxonomy (Ad-hoc-Arbeitsgruppe Boden, 2005). Soil sampling took place per date and horizon within the 30 m × 30 m core area at 18 individual points on two so-called rotating sampling axes (Barth et al., 2000) so that samples were always taken at slightly offset points. One composited soil sample was formed from each six points so that there were three standard composited soil samples in total (three times six points = 18 points). In exceptional cases, for example, where surface inhomogeneities were present, fewer than the three standard composited soil samples were available. Since thicker topsoil horizons

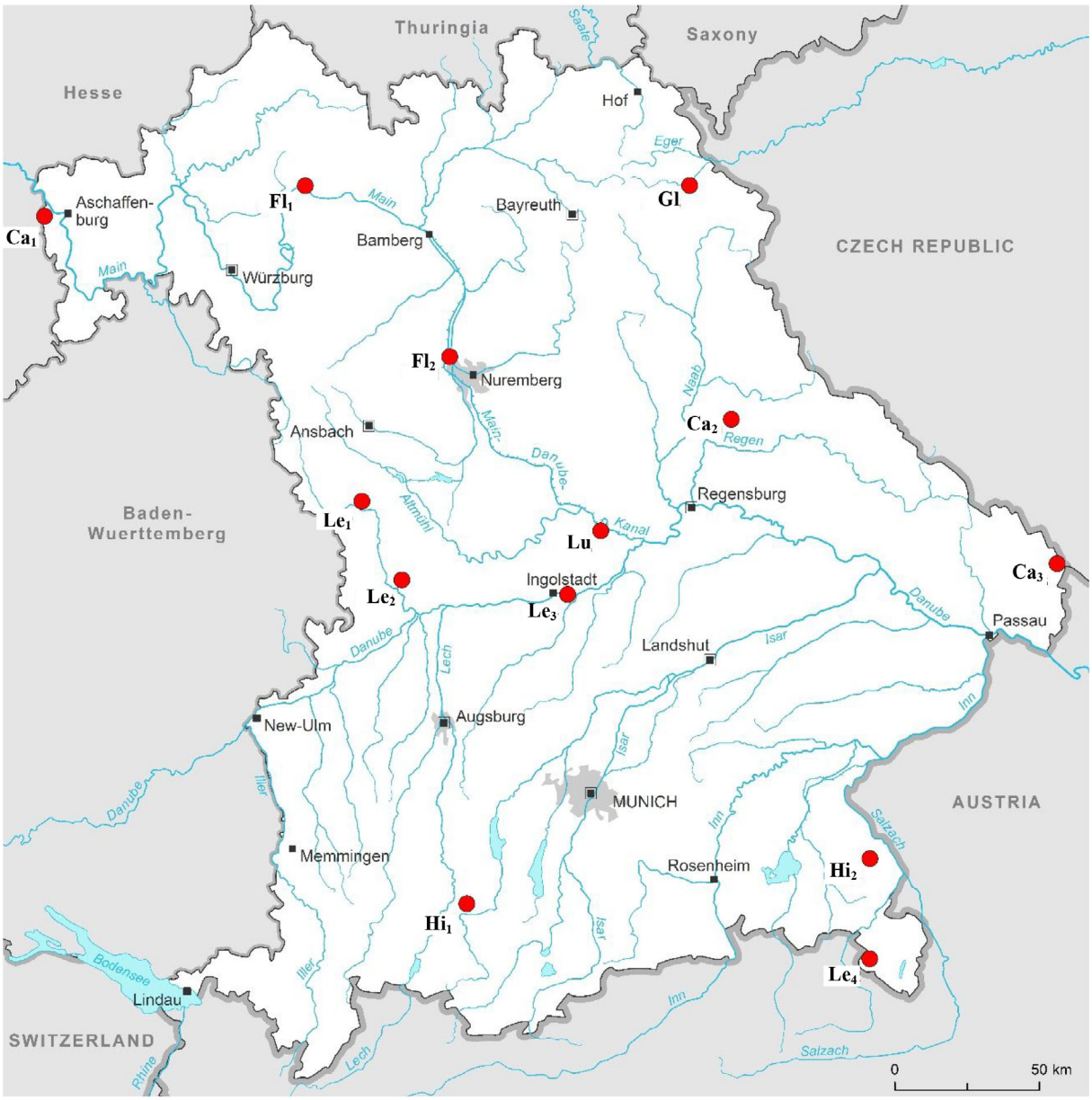

FIGURE 1 Location of the soil monitoring sites in Bavaria, southern Germany. Soil types are three Cambisols (Ca_1-Ca_3), two Fluvisols (Fl_1 and Fl_2), a Gleysol (Gl), two Histolsols (Hi_1 and Hi_2), four Leptosols (Le_1-Le_4) and a Luvisol (Lu)

were sometimes divided into two depth levels, there were up to six composited soil samples per date and horizon. Therefore, the number of replications (consisting of composited soils) for the A horizons ranged from three (standard) to 12 (2 A-horizons, one of which is subdivided) per year with the exception of $n = 1$ at one site in 1987. At the initial sampling date, soil texture, concentrations of SOC and heavy metals as well as pH and CEC were determined following standard methods according to Barth et al. (2000). We determined Cu_{AR} using DIN ISO 11466 (1997), analysed SOC using a CN element analyser (DIN ISO 10694, 1996) and measured pH using a 0.01 M $CaCl_2$ solution (DIN ISO 10390, 1997). The CEC was measured using an unbuffered NH_4Cl solution.

The sites were monitored regularly from 1986 (sites Ca_1, Fl_2, Hi_2, Le_1, Le_3), 1987 (Ca_2, Ca_3, Hi_1, Le_2, Le_4, Lu), 2000 (Gl) or 2001 (Fl_1) onwards. The number of sampling times in the monitoring period was 3 (Gl, Le_4), 5 (Ca_3, Fl_1, Fl_2, Hi_1, Le_1, Le_2, Le_3, Lu) or 6 (Ca_1, Ca_2, Hi_2), and the final sampling dates were in 2013 or 2016.

2.2 | Statistical analyses

2.2.1 | Descriptive statistics

Statistical analyses were performed with R version 4.05 (R Core Team, 2021). Package unikn (Neth and

TABLE 1 Characteristics of the 13 soil monitoring sites

Site, site ID and land use	Parent material	Soil type	Cu_{AR} background value topsoil/ subsoil/bedrock (mg/kg)[a]	Horizon group/ description (cm)	Texture (%) sand, silt, clay
Ca_1 6020 Coniferous forest	young Pleistocene flying sand	Cambisol	10/10/10	Organic/L (−3.4–−3.2) Organic/Of + Oh (−3.2–0) A/IAeh (0–2) AB/IAh-Bv (2–13) B/IBv (13–52) C/IIilCvl (52–100)	 86, 8, 6 84, 9, 7 89, 7, 4 97, 3, 1
Ca_2 6739 Coniferous forest	Cretaceous sand	Cambisol	10/10/15	Organic/L (−3.5–−3) Organic/Of + Oh (−3–0) A/IAeh (0–1) AB/IAh-Bv (1–20) B/IBv (20–40) B/IIBv (40–60) C/IIilCv (60–100)	 86, 10, 4 83, 12, 5 85, 10, 5 77, 14, 10 92, 4, 5
Ca_3 7842 Coniferous forest	loess loam from the Würm period	Cambisol	13/16/28	Organic/L (−4.5–−4) Organic/Of (−4–−1) Organic/Oh (−1–0) A/IAh (0–1) AB/IAl-Bv (1–20) B/IIBtv (20–50) BC/IIIilCv + Bv 1 (50–70) BC/IIIilCv + Bv 2 (70–90) C/IVelCv (90–110)	 14, 70, 16 22, 59, 19 23, 55, 22 27, 36, 37 40, 16, 44 64, 30, 7
Fl_1 5927 Extensive grassland	Holocene river sediment (Main)	Fluvisol	20/16/−	A/IaAh (0–20) M/IaM (20–60) C/IIailCv (60–100)	75, 17, 8 80, 15, 5 97, 2, 1
Fl_2 6431 Intensive grassland	Holocene river sediment (Regnitz)	Fluvisol	77/51/− (increased background value)	A/IaAh (0–20) M/IaM (20–50) M/IIaM (50–85) M/IIIaM (85–110)	66, 27, 7 65, 26, 9 43, 37, 20 91, 5, 4
Gl 5938 Intensive grassland	Holocene brook sediments (Kosseine)	Gleysol	66/41/− (increased background value)	A/IaAh (0–10) G/IGo (10–35) G/IGor (35–70) G/IGr (70–100)	25, 50, 25 38, 48, 15 76, 20, 5 73, 22, 4
Hi_1 8131 Litter meadow	Holocene peat	Histosol	−/−/−/	Peat/IuH (0–20) Peat/IuHw (20–50) Peat/IuHr (50–80) Peat/IInHr1 (80–110)	

(Continues)

TABLE 1 (Continued)

Site, site ID and land use	Parent material	Soil type	Cu_{AR} background value topsoil/ subsoil/bedrock (mg/kg)[a]	Horizon group/ description (cm)	Texture (%) sand, silt, clay
Hi_2 8143 Natural fen	Holocene peat	Histosol	−/−/−/	Peat/IhH1 (0–5) Peat/IhH2 (5–20) Peat/IhH3 (20–30) Peat/IhH4 (30–40) Peat/IhH5 (40–60) Peat/IhH6 (60–80) Peat/IhH7 (80–100)	
Le_1 6929 Pasture land	White Jura	Leptosol	32/51/27	A/IeAh (0–20) AC/IAh-eCv (20–50) AC/IIAh-eCv (50–70) C/IIelCn (70–90)	11, 36, 52 27, 46, 27 48, 34, 19 48, 33, 19
Le_2 7130 Pasture land	Loess loam over White Jura	Leptosol	20/30/27	A/IAh (0–20) AC/IIAh-eCv (20–50) C/IIclCv (50–70)	11, 55, 34 32, 44, 25 41, 40, 19
Le_3 7234 Young deciduous forest	Holocene river sediment (Donau)	Leptosol	29/19/23	A/IaeAh (0–20) C/IaelCv (20–50) C/IIaelCv (50–70) CG/IIaeCv-Go (70–90) G/IIIaGo (90–120)	23, 68, 10 15, 73, 13 21, 60, 19 23, 57, 20 57, 40, 4
Le_4 8442 Pasture land	Holocene brook sediments (Klausbach)	Leptosol	22/40/11	A/IaAh (0–10) AC/IIaAh-eCv (10–20) C/IIIaelCv (20–70) C/IVaelCv (70–100)	36, 53, 12 55, 43, 3 61, 38, 2 86, 12, 3
Lu 7035 Pasture land	Loess loam over White Jura	Luvisol	20/30/27	A/IAh (0–10) AB/IAh-Bv (10–20) BT/IIBv-T (20–40) C/IIIelCv (40–70)	9, 51, 40 14, 52, 34 17, 46, 37 32, 30, 39

Note: –, insufficient data basis for determining the background value ($n < 10$).

[a]Bayerisches Landesamt für Umwelt (2011).

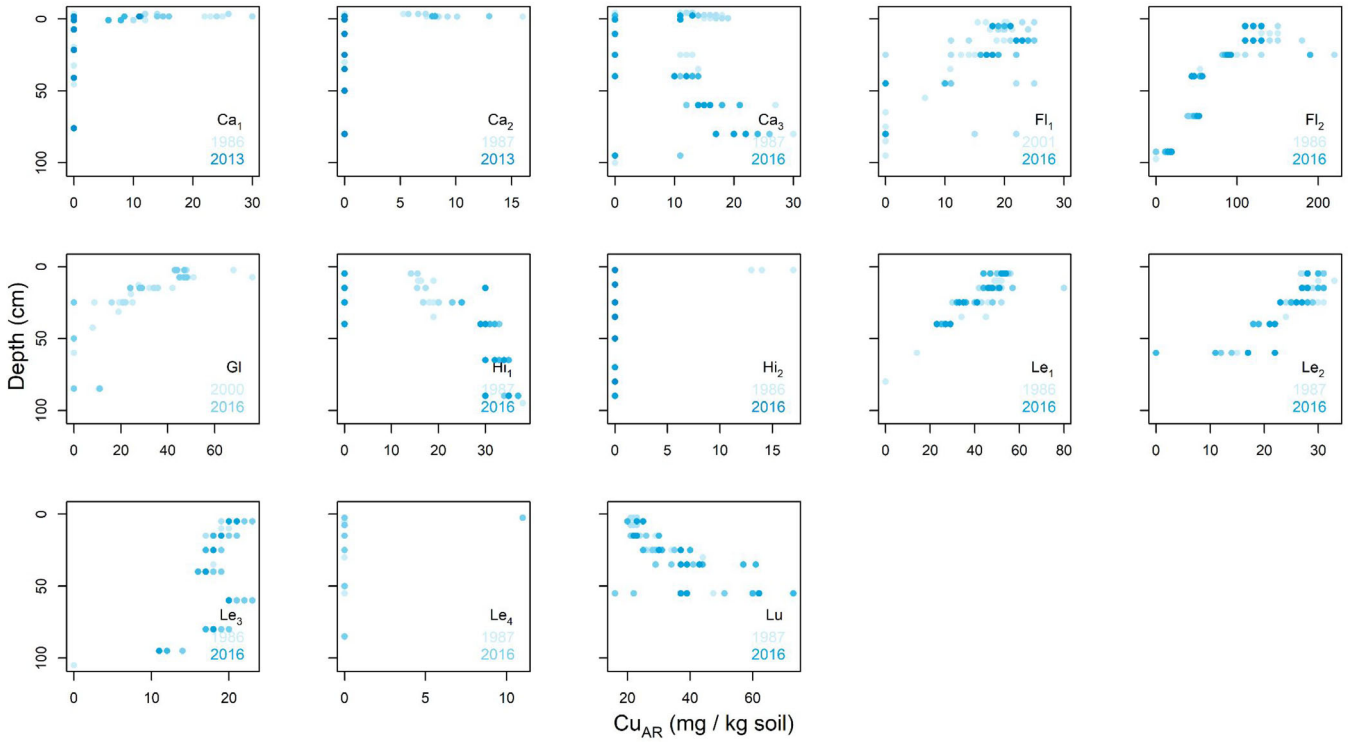

FIGURE 2 Depth-wise changes in the concentrations of Cu in aqua regia (Cu_{AR}) for the thirteen soil monitoring sites. Information on sampling year is included by plotting data points with varying colour intensity, with darker blue indicating more recent sampling. Soil types are three Cambisols (Ca_1-Ca_3), two Fluvisols (Fl_1 and Fl_2), a Gleysol (Gl), two Histolsols (Hi_1 and Hi_2), four Leptosols (Le_1-Le_4) and a Luvisol (Lu)

Gradwohl, 2021) was used to show information on sampling time in the profile plots. Descriptive statistics included boxplots and histograms for Cu_{AR}, calculations of 95%-confidence intervals of differences in means, and scatter plots for Cu_{AR} versus pH, CEC and SOC. 95%-confidence intervals of differences in means of Cu_{AR} concentrations between two sampling dates were calculated for the A horizons—the mineral soil horizon which is affected the most by anthropogenic Cu inputs—of the respective sites for the differences between the first and the last sampling date for those sampling dates where the data were normally distributed. When Cu_{AR} concentrations in the A horizons of the respective sites were not normally distributed for a particular sampling date, we tried logarithmic or Box-Cox-transformations of the Cu_{AR} concentrations. Since transformations were not successful, we calculated the 95%-confidence intervals using the next suitable sampling time. When no normality was achieved or no Cu_{AR} was detected for a given site, 95%-confidence intervals were not calculated.

Inspections of the temporal courses of the Cu concentrations for the different horizons of the sites indicated—in agreement with prior information on Cu sources and loads in southern Germany—that there was no trend of increasing Cu_{AR} concentrations with time and no trend of Cu_{AR} translocation with time. For all A horizons, boxplots and 95%-confidence intervals of differences in means indicate there was no trend of increasing Cu_{AR} concentrations with time (data not shown). An exceptional case was site Fl_2, in which a slightly decreasing trend was noted, but the effect was very close to a difference of zero and there was no trend of increasing Cu_{AR} concentrations in the next horizon (M horizon) for this site.

Since there was no trend of increasing Cu_{AR} concentrations with time, data of different sampling dates were combined (i.e., disregarding the effect of time and assuming that micro-scale variations of SOC concentrations, pH and CEC determine Cu_{AR} concentrations rather than time-dependent effects for these monitoring sites with predominantly geogenic Cu_{AR} concentrations) and horizon-specific data analyses were carried out. Horizons with smaller numbers of observations (BC and CG) were excluded from all analyses. Two different modelling approaches—mixed-effects models and rule-based models—were applied for the following two modelling variants: I. Description of the Cu_{AR} concentrations for the different horizons depending on site, SOC, pH and CEC; and II. Prediction of the Cu_{AR} concentrations for the A horizons—the horizons for which most

TABLE 2 Parameterization and performance of horizon-specific fixed and mixed-effects models for the response variable Cu_{AR} (mg/kg soil)

Horizon, number of sites and sample size n	Final equation[a]	Random components (assumed mean of 0 and variance)	ρ	RMSE
Organic layer 3 sites, $n = 82$	$-134 + 6.55\ SOC + 26.36\ pH + 5.45\ CEC - 0.05\ CEC^2 - 1.73\ SOC * pH - 0.19\ SOC * CEC - 0.57\ pH * CEC + 0.05\ SOC * pH * CEC$	Site ~ N(0, 55) Residual ~ N(0, 22)	0.83	4.38
Peat 2 sites, $n = 167$	$111 - 3.27\ SOC - 50.8\ pH + 3.68\ CEC + 0.02\ CEC^2 + 1.45\ SOC * pH - 0.05\ SOC * CEC - 0.60\ pH * CEC$	Site ~ N(0, 50) Residual ~ N(0, 75)	0.69	8.46
A 11 sites, $n = 258$	$44.7 + 0.72\ SOC - 2.61\ pH - 0.62\ CEC + 0.0066\ CEC^2 + 0.45\ SOC * pH - 0.05\ SOC * CEC$	Site ~ N(0, 1765) Residual ~ N(0, 41)	0.95	6.17
AB 4 sites, $n = 71$	$23.3 - 10.3\ SOC - 4.57\ pH - 1.08\ CEC + 2.49\ SOC * pH + 0.48\ SOC * CEC + 0.24\ pH * CEC - 0.10\ SOC * pH * CEC$	Site ~ N(0, 81) Residual ~ N(0, 1.5)	0.73	1.13
AC 3 sites, $n = 57$	$2.19 + 0.20\ CEC + 1.87\ SOC$	Site ~ N(0, 120) Residual ~ N(0, 14)	0.95	3.57
M 2 sites, $n = 76$	$397.5 - 168.7\ SOC - 64.4\ pH + 32.9\ SOC * pH$	Site ~ N(0, 2118) Residual ~ N(0, 387)	0.96	19.0
B 3 sites, $n = 98$	$76.9 - 2.95\ SOC - 41.7\ pH + 7.71\ CEC + 5.26\ pH^2 - 1.19\ pH * CEC$	Site ~ N(0, 22) Residual ~ N(0, 7.9)	0.80	2.70
BT[b] 1 site, $n = 25$	$12.4 - 1.37\ pH - 0.31\ CEC + 0.05\ pH * CEC$	Residual ~ N(0, $2.9 * 10^{-2}$)	0.76	6.37
G 2 sites, $n = 39$	$-0.55 + 13.23\ SOC$	Site ~ N(0, 0) Residual ~ N(0, 19)	0.93	4.23
C 9 sites, $n = 106$	$0.47 + 20.1\ SOC - 0.30\ CEC + 0.03\ CEC^2 - 0.67\ SOC * CEC$	Site ~ N(0, 97.1) Residual ~ N(0, 28.9)	0.91	5.06

Abbreviations:ρ, Spearman's rank correlation coefficients between measured and estimated values; RMSE, root mean squared error of calibration (mg / kg soil).

[a]The unit for the intercept is mg/kg soil. The units for the regression terms are mg / kg soil multiplied by the respective reciprocals of the units of the variables (1st and 2nd order contributions and interactions, SOC: g/100 g, CEC: cmol(+)/kg).

[b]The response variable was log-transformed.

information was available—for a closed population as well as for new sites. The approaches and parameterizations are discussed below.

I. Description of Cu_{AR} concentrations for the different horizons depending on site, SOC, pH and CEC.

I.1 Fixed-effects and mixed-effects modelling.

A fixed-effects model can be used to describe a response variable y as follows:

$$y = X\beta + e, \quad (1)$$

with y: vector of the response variable, X: design matrix of the independent variables (e.g., measured values) for the fixed effects, β: vector of the fixed effects (e.g., slopes for the fixed effects), and e: vector of the errors.

A mixed-effects model (e.g., Zuur et al., 2009) is formulated as follows:

$$y = X\beta + Z\,u + e, \quad (2)$$

with Z: design matrix for the random effects, and u: vector of the random effects. Random effects are site for the horizon-specific models and additionally the horizon: site nested effect for the general model introduced below.

Mixed-effects modelling was performed using the packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017). Horizon-specific models for Cu_{AR} were created considering the regression terms for the variables SOC concentrations, CEC and pH as fixed effects (see Table 2 for optimised intercept and regression coefficients for each

model) and the structural component site as a random effect (see Table 2 for the residual variance and variance due to the site effect for each model). Fixed effects comprised 1st order and 2nd order polynomial terms as well as the three two-way interactions and the three-way interaction. Concentrations of clay could not be included as a fixed effect, since the concentrations were only available for the site characterizations and not for each sampling.

Model simplifications were carried out as described by Crawley (2012). First, a non-significant three-way interaction was removed, followed by eliminations of non-significant two-way interactions and then removals of non-significant 2nd and 1st order main effects. Thus, we considered fixed effects only in the final models in the case of significant ($p \leq 0.05$) contributions. Non-significant effects of the main effects were included only in the case of a significant interaction or a significant 2nd order contribution of the main effects.

For the BT horizon, where only one site was available, a fixed-effect model was used and the same approach as above was applied, except for the exclusion of the random site effect.

For the mixed-effects models, the estimation procedure for the variance components was restricted maximum likelihood, whereas fixed effects were estimated by generalised least squares and subjected to Wald-type F-tests using the Kenward-Roger method. Residuals were inspected for homoscedasticity and normality.

Two model variants A and B were explored in the mixed-effects modelling which differed in the subsetting of the data.

2.3 | Model variant A: Horizon-specific models

For model variant A, homoscedasticity and normality of residuals were assumed as mandatory conditions for the horizon-specific final models (see e.g., discussions on residual inspection in Welham et al. (2014) and Schielzeth et al. (2020)). In order to achieve the conditions, not only a transformation of the response variable (AC and BT horizons) was required, but also subsetting of the datasets including mathematical handling of the Cu_{AR} concentrations below the detection limit was required. Specifically, sites, where the respective horizons consisted only of Cu_{AR} values below the detection limit, were excluded. For the peat horizon, residuals were non-normal and we assumed that the uncertainty of values below the detection limit contributed to this considerably, given the small overall range of Cu_{AR} values. We, therefore, removed zero Cu_{AR} values for this horizon from the data set. For the A horizon, residuals were non-normal and a log-transformation after the addition of 1 to the response variable was not successful. We thus just modelled the subset of the A horizon for a restricted range and found by a trial-and-error procedure that distributional requirements for the residuals were fulfilled for $Cu_{AR} \leq 40$ mg/kg. For AC, residuals were non-normal and the value 1 was added to the response variable and a log-transformation was then carried out. For M, residuals were non-normal, and we modelled the subset of $Cu_{AR} \leq 150$ mg/kg. For BT, residuals were non-normal, and a log-transformation was carried out as described above. For C, residuals were non-normal, and a transformation was not successful. As was done for the peat horizon, we also removed zero Cu_{AR} values from the data set and limited the data set to a subset of $Cu_{AR} \leq 40$ mg/kg.

2.4 | Model variant B: Horizon-specific and general models

Model variant B used all data available (i.e., no subsetting and no setting of zero Cu_{AR} values as not available) in order to avoid a bias in Cu_{AR} predictions. Schielzeth et al. (2020) reported that estimates of mixed-effects models were usually robust to violations of distributional assumptions of residual and random effects. However, the accuracy of model estimates needs to be inspected and critically discussed. Table 2 shows the final equations for the horizon-specific fixed-effects (horizon BT) and mixed-effects models including the random terms (i.e., residual variance for the fixed effects model, and site and residual variance for the mixed-effects models).

Besides the horizon-specific models described above, also a general model for all horizons was calculated. For the general model, the random effect comprised horizons nested in sites.

I.2 Horizon-specific rule-based cubist models

Horizon-specific modelling was carried out for the entire data set without subsetting (i.e., the same data set as for model variant B above) using the packages Cubist (Kuhn & Quinlan, 2021) and caret (Kuhn, 2021). The cubist models use a boosting–like procedure called committees (Kuhn & Johnson, 2018; Kuhn & Quinlan, 2021). Calibration was carried out using an internal ten-fold cross-validation, where the number of committee models was optimised using the values 1, 10, 50 and 100 (Table 3). The optimal number of committee models was then used for a description of Cu_{AR} concentrations.

2.5 | Model performance parameters

For the fixed-effects model for horizon BT, the coefficient of determination was calculated and is labelled as R^2_f.

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

Horizon	Number of committees	Variable usage in conditions[a]	Variable usage in the model committees[b]	ρ	RMSE
Organic layer	10	Site (42%), pH (30%), SOC (10%)	CEC (78%), SOC (42%), pH (31%)	0.80	4.92
Peat	100	pH (75%), CEC (60%), SOC (31%)	CEC (27%), SOC (16%), pH (16%)	0.82	2.73
A	50	Site (98%), SOC (15%)	pH (70%), CEC (65%), SOC (58%)	0.98	5.06
AB	50	CEC (100%), pH (4%)	CEC (4%), pH (4%)	0.99	1.01
AC	1	CEC (100%)	CEC (100%), SOC (86%)	0.95	3.63
M	1	no conditions	CEC (100%), SOC (100%), pH (100%)	0.95	21.7
B	50	CEC (94%), SOC (27%), pH (21%)	CEC (30%), SOC (14%), pH (13%)	0.94	0.94
BT	10	no conditions	SOC (60%), CEC (50%)	0.72	6.79
G	10	SOC (10%)	SOC (100%), pH (50%), CEC (50%)	0.95	3.86
C	100	CEC (86%), site (40%), SOC (22%), pH (9%)	CEC (49%), SOC (34%), pH (23%)	0.91	3.82

Abbreviations: ρ, Spearman's rank correlation coefficients between measured and estimated values; RMSE, root mean squared error of calibration (mg/kg soil).
[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.

For the mixed-effects models, marginal (R^2_m) and conditional (R^2_c) pseudo-coefficients of determination were calculated, which 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). We used the package MuMIn (Barton, 2020) for the calculations.

For both regression approaches, root mean squared errors (RMSEs) and Spearman rank correlation coefficients ρ between measured and modelled Cu_{AR} values were calculated for both calibration and validation sets, as will be described subsequently.

II. Prediction of the Cu_{AR} concentrations for the A horizons for a closed population and for new sites

For the predictive approaches using calibration–validation procedures, the A horizons were selected for which in total 258 observations from eleven sites were available (Table 2). Two procedures were used and are described below. Model performance parameters were the same as above and calculated for the calibration and validation samples.

II.1 Prediction of the Cu_{AR} concentrations for the A horizons for a closed population

Pseudo-independent calibration–validation is useful for data sets in which the population of interest is available and future predictions will be of interest solely for the sites included in the study. An extension to new sites is not feasible. A five-fold random partitioning of the data was used to split the 258 observations into a calibration and a validation sample ($n = 129$ each in each of the five folds).

For the calibration sample, mixed-effects modelling was carried out as above including model simplification. The optimal model depended on the respective fold of the random partitioning of the data and included a 1st order contribution of SOC (two out of five folds), and site as a random effect.

For the cubist model, the calibration sample was not only used to optimise the number of committee models (values of 1, 10, 50 and 100), but also the number of neighbours (0, 1, 5, and 9) for a subsequent prediction in a grid search. In such a procedure with the number of neighbours greater than 0, model predictions are automatically adjusted using neighbouring points from the calibration (or training) set data (Kuhn & Johnson, 2018; Kuhn & Quinlan, 2021). Optimal values depended on the respective fold of the random partitioning of the data and ranged from 1 to 50 for the committee models and from 0 to 9 for the neighbours. The usage of the variable site in the conditions for the rules ranged from 92% to 100% and for the other variables from 0% to 47% depending on the variable and fold. The variable usage in the committee models also differed in the five folds.

II.2 Prediction of the Cu_{AR} concentrations for the A horizons for new sites

For independent calibration–validation, the eleven sites were randomly assigned to either the calibration or the validation sample in five folds (e.g., Brown et al., 2005; Ludwig et al., 2017).

Mixed-effects modelling was carried out as described above and the optimal models in the five folds differed. The site was again included as a random effect in the calibration, which meant that the effect was zero for the validation with new sites.

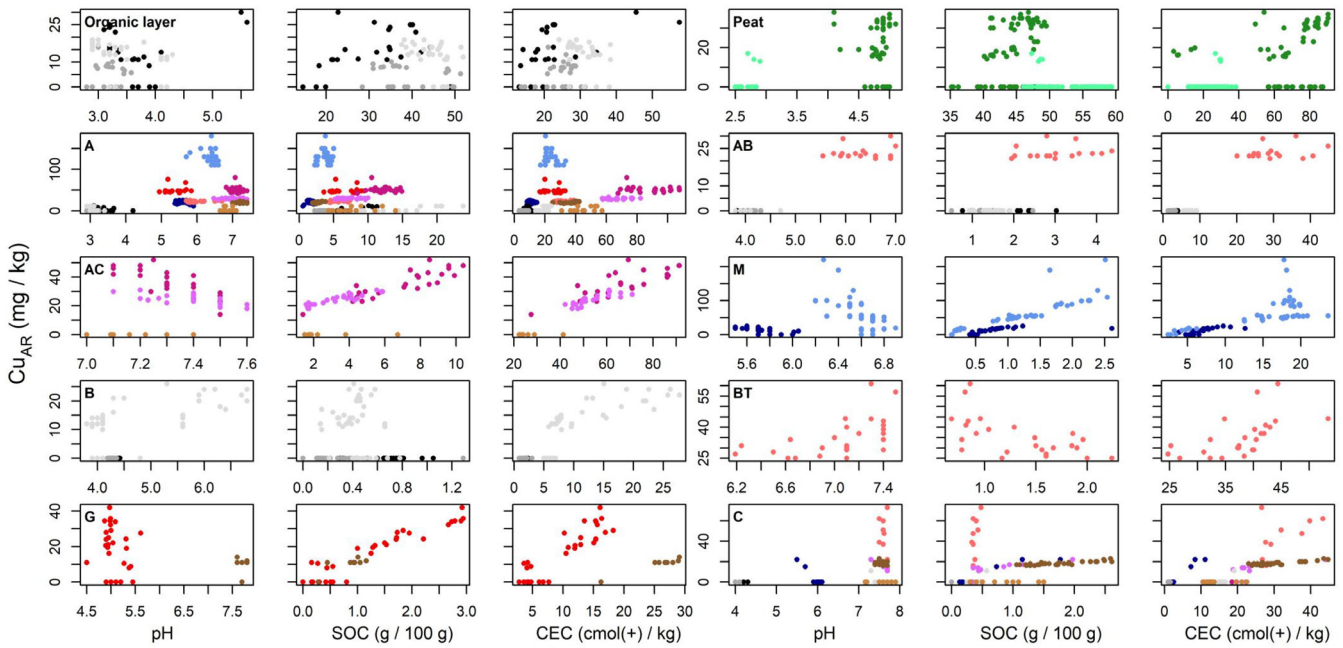

FIGURE 3 Scatter plots for Cu_{AR} and pH, soil organic carbon (SOC) concentration and the cation exchange capacity (CEC). Different colours indicate soils from different sites (Ca_1: Black, Ca_2 and Ca_3: Dark and light grey, Fl_1 and Fl_2: Dark and light blue, Gl: Dark red, Hi_1 and Hi_2: Dark and light green, Le_1 and Le_2: Dark and light pink, Le_3 and Le_4: Dark and light brown, and Lu: Light red)

The cubist model only considered the three variables SOC, pH and CEC. The optimal numbers of committees and neighbours in the five folds ranged from 1 to 100 and from 1 to 9, respectively.

3 | RESULTS

3.1 | Cu_{AR} concentrations in soil profiles and horizons

As expected, since there was no viticulture with the application of Cu pesticides, high organic fertilisation, or mining activity, there was no measurable trend of increasing Cu_{AR} concentrations with time in the respective depth ranges for the thirteen sites. Data were therefore combined for the different sampling times. Depth-wise changes in Cu_{AR} concentrations showed no consistent pattern (Figure 2). For most sites, Cu_{AR} concentrations decreased with depth, but decreases ranged from approximately linear (site Le_2) to approximately exponential. For sites Fl_1, Lu, Le_3 and Ca_3, however, depth-wise changes indicated a large scatter (sites Fl_1 and Ca_3), no consistent change (site Le_3) and even an increase with depth (site Lu). In total, 991 observations were available and Cu_{AR} concentrations ranged from 0 to 220 mg/kg, with most (79%) values <30 mg/kg. The largest Cu_{AR} concentrations were found at site Fl_2 (Figure 2).

Histograms indicated different distributions of Cu_{AR} concentrations in different horizons (not shown). Right-skewness was common for several horizons, which can be generally expected for small concentrations since negative concentrations are not possible. Scatter plots of Cu_{AR} against CEC or SOC showed distinct positive relationships for some horizons (e.g., AC, M and G horizons), but relationships were more complex to non-existent for other horizons (e.g., AB horizons, Figure 3). For pH, large scattering was observed with inconsistent relationships (Figure 3). Spearman correlations ρ for the pair CEC-pH were significant ($p \leq 0.05$) for the horizons peat, A, AB, AC, BT and C and ranged from −0.32 (AC horizons) to 0.83 (A horizons). Typically, one would expect a positive relationship for the pair CEC-pH, since with increasing pH the contribution of variable charges increases. Indeed, the only negative ρ was observed for the AC horizons, where the pH range was narrower than for the other horizons. Spearman correlations ρ for the pair SOC-pH were significant for the horizons peat, AB, AC, B, BT, G and C, and ranged from −0.76 (BT) to 0.39 (C). For the pair SOC-CEC, Spearman correlations ρ were significant for all ten horizons and ranged from −0.54 (BT) to 0.91 (M). For this pair, a positive relationship would be expected, since SOC may contribute to the CEC. The expected positive relationship was observed for all horizons, except for two: peat (sites Hi_1 and Hi_2) and BT (site Lu) horizons.

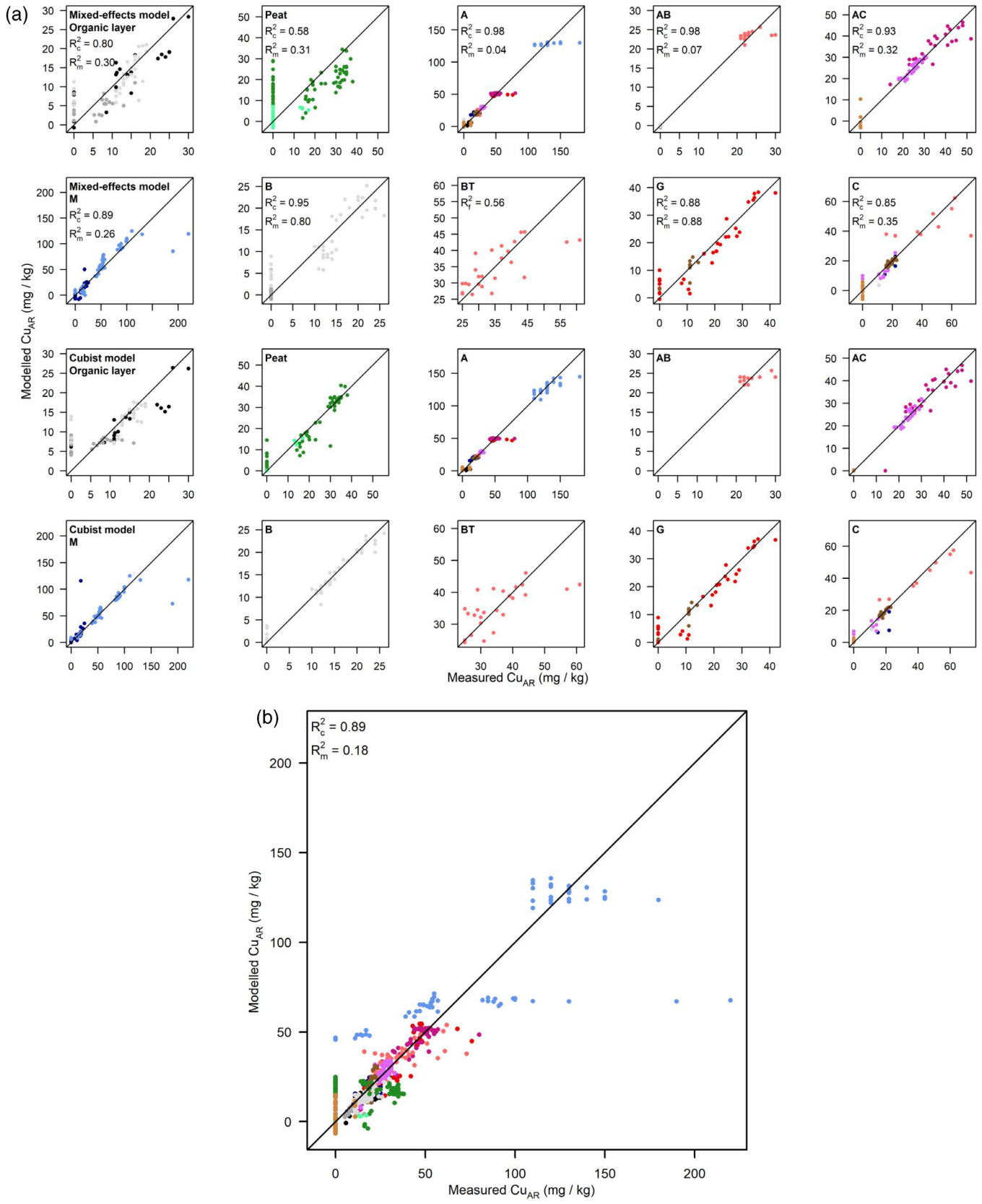

FIGURE 4 Legend on next page.

I. Description of Cu_{AR} concentrations for the different horizons depending on site, SOC, pH and CEC

I.1 Mixed-effects modelling using model variant A

The subset of the total data set with a restricted range of Cu_{AR} (horizons A, M and C) and removed zero Cu_{AR} values (peat and C horizons) was successfully described by the fixed-effects (B and BT horizons) and mixed-effects models with Spearman's rank correlation coefficients ρ between measured and modelled Cu_{AR} concentrations ranging from 0.76 to 0.98 for the different horizons (data not shown). However, the subsetting as described above may result in a severe bias of predictions of Cu_{AR} concentrations for small or large Cu_{AR} concentrations. In order to overcome this bias, variant B was introduced and used for all research objectives.

I.2 Mixed-effects and cubist modelling using variant B

The fixed-effects model for the BT horizon and the mixed-effects models for the other horizons were very useful to describe the Cu_{AR} concentrations as a function of the fixed effects of CEC, SOC and pH. The coefficient of determination (R^2_f) was 0.56 for the BT horizon, and conditional pseudo-coefficients of determination R^2_c ranged from 0.58 (peat horizons) to 0.98 (A horizons) for the mixed-effects models (Figure 4a).

The importance of the random effect of the site according to differences in R^2_m and R^2_c was especially pronounced for the A and AB horizons, and to a lesser extent for the organic layers and peat, AC, M, B, and C horizons (Figure 4a), where most of the variation was explained by site, which consists of the bulked unknown site-specific information (most likely parent material and specific mineral compositions). The variance explained by the random effect of site was especially high for the M horizons (2118 mg^2/kg^2) (where also the residual variance was high) and A horizons (1765 mg^2/kg^2; Table 2). For the other horizons, the importance of site was much smaller or negligible as indicated by similar or identical R^2_m and R^2_c values (Figure 4a) and small or zero variances for the random effect of site (Table 2), and the fixed effects (1st order effects, 2nd order effects and interactions) were very useful for a description of Cu_{AR} concentrations (Table 2, Figure 4a). For these horizons, the prediction of Cu_{AR} concentrations at new sites may be more accurate than for the horizons with high or pronounced site effects.

Relationships for the different horizons varied: complex relationships were obtained for eight of the 10 horizons, where one or several significant interactions were part of the final models (Table 2). In contrast, for the AC and G horizons, Cu_{AR} was described well by a simple linear relationship with SOC (G horizons) or with CEC and SOC (AC horizons, Table 2, Figure 4a).

The general mixed-effects model which included horizons nested in site had R^2_m and R^2_c values of 0.18 and 0.89, respectively. Variances of random effects decreased in the order site (635 mg^2/kg^2) > horizon:site nested effect (147 mg^2/kg^2) > residual error (125.1 mg^2/kg^2) and the fixed effects were SOC, pH, CEC, a squared contribution of SOC, a SOC:CEC interaction and a three-factor interaction. For several horizons and sites, the agreement between measured and modelled Cu_{AR} concentrations was satisfactory, but the increased Cu_{AR} concentrations in the M and A horizons of the site Fl_2 were modelled only poorly (Figure 4b).

Performances of the horizon-specific rule-based cubist models were similarly successful compared to those of the mixed-effects models: RMSEs of the calibrations were lower for six horizons using the cubist models (peat, A, AB, B, G and C) and for four (organic layer, AC, M and BT) using the fixed- and mixed-effects models (Figure 4a, Tables 2 and 3). The importance of the variables differed between mixed-effects models and cubist models. For instance, the site effect was not important for most horizons in the cubist model (Table 3). Also, the simple mixed-effects model for the G horizons, which consisted only of SOC and site, was very different from the cubist model, where all variables were important (Table 3).

II. Prediction of Cu_{AR} concentrations for the A horizons for a closed population and for new sites

For a closed population of A horizons from 11 sites (i.e., pseudo-independent validation), the mixed-effects models and the cubist models were very successful in describing Cu_{AR} concentrations in the calibration samples and a prediction in the validation samples in the five-fold random splittings of the data set. RMSEs in the calibrations ranged from 4.57 to 7.32 mg/kg for the mixed-effects models and from 3.92 to 5.07 mg/kg for the cubist models. Figure 5 shows the calibration models with a median performance in the calibration out of the five folds and the subsequent validation. In the

FIGURE 4 (a) Modelled versus measured Cu_{AR} concentrations resulting from horizon-specific mixed-effects models (top rows) and rule-based cubist models (bottom rows) for the organic layer and peat horizons and the mineral horizons. Coefficients of determination and conditional and marginal pseudo-coefficients of determination are indicated as R^2_f, R^2_c and R^2_m. Different colours indicate soils from different sites (see legend of Figure 3). (b) Modelled versus measured Cu_{AR} concentrations resulting from the general mixed-effects model. Conditional and marginal pseudo-coefficients of determination are indicated as R^2_c and R^2_m. Different colours indicate soils from different sites (see legend of Figure 3)

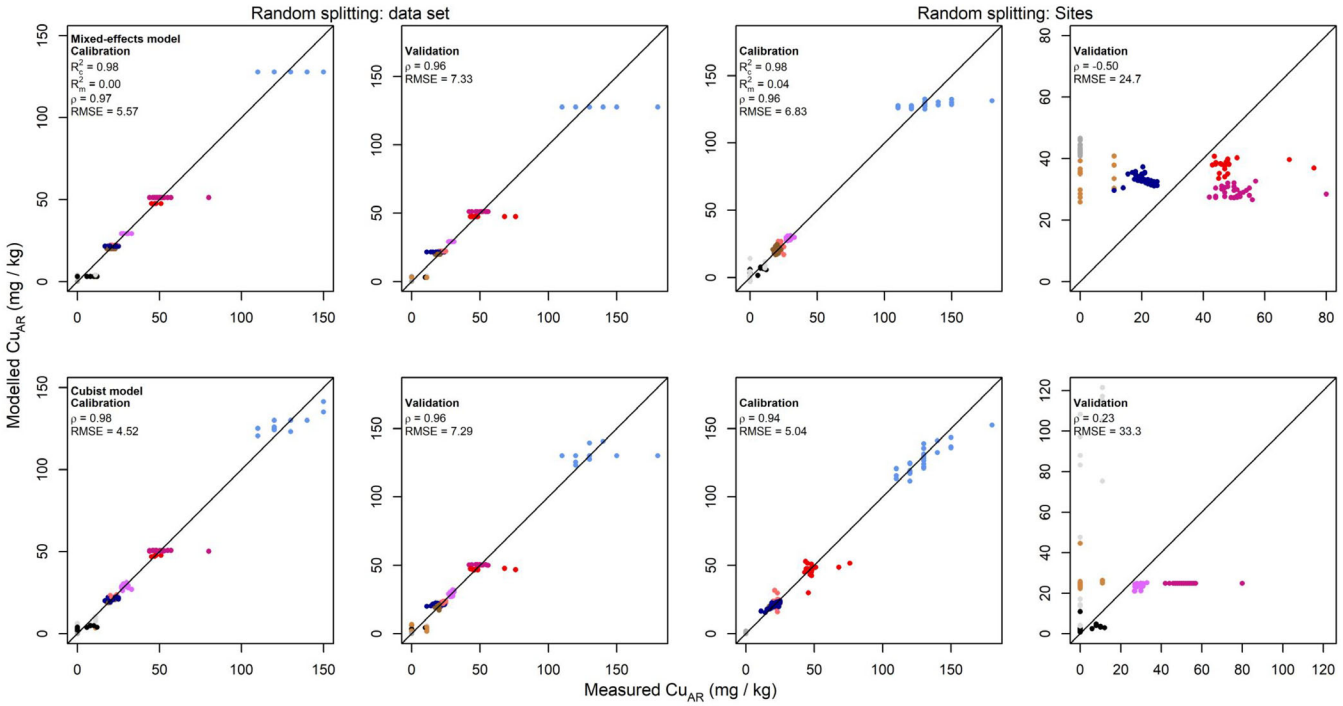

FIGURE 5 Modelled versus measured Cu_{AR} concentrations of the mixed-effects models and rule-based cubist models for the A horizons. The sub-plots on the left refer to pseudo-independent calibration and validation (random splitting of the data set). The sub-plots on the right refer to independent validation (random splitting of the sites). For all cases, calibration and validation plots results referred to those of median performance in the calibrations with respect to RMSE of five-fold random partitions of the data set (left) or of the sites (right). R^2_c, R^2_m, Spearman's rank correlation coefficient ρ and the root mean squared error (RMSE) in mg/kg are given. Different colours indicate soils from different sites (see legend of Figure 3)

validations, RMSEs were slightly higher and ranged from 5.58 to 7.93 mg/kg and from 7.01 to 8.55 mg/kg for the mixed-effects and cubist models, respectively.

The five-fold random splittings of sites into calibration and validation samples (i.e., independent validation) showed very good performances of the mixed-effects and cubist model in the calibrations with RMSEs ranging from 3.90 to 7.86 mg/kg and from 3.73 to 6.18 mg/kg, respectively (see Figure 5 for the calibration models with median performance). In the validations, however, both approaches were unsuccessful as indicated by high RMSEs and Spearman's rank correlation coefficients ρ ranging from −0.67 to 0.52 (mixed-effects models) and from −0.75 to 0.68 (cubist models).

4 | DISCUSSION

I. Cu_{AR} concentrations in soil profiles and horizons

The Cu_{AR} concentrations of the soils from 13 south German sites down to approximately 1 m measured regularly in the period from 1986 to 2016 ($N = 991$ observations) were mostly (79%) < 30 mg/kg. We observed moderately large values for three sites and exceptionally large Cu concentrations of up to 220 mg/kg for one site (site Fl_2). For this site, the large concentrations are due to the parent material (Holocene river sediment of the Regnitz), which contained ubiquitously larger Cu_{AR} background values as a result of proximity to industrial areas. For all sites and horizons, we did not observe any trend of increasing Cu_{AR} concentrations with time and therefore combined data of different sampling dates to study whether micro-scale variations of SOC concentrations, pH and CEC may determine Cu_{AR} concentrations rather than time-dependent effects for these monitoring sites with predominantly geogenic Cu_{AR} concentrations.

The scatter plots showed distinct positive relationships between Cu_{AR} and CEC and SOC for some horizons, but relationships were more complex to non-existent for other horizons. The horizon-specific relationships between Cu_{AR} and the variables CEC, SOC and pH suggested that regression models may be suitable for a description and prediction of Cu_{AR} in the different horizons. However, multicollinearity between the variables pH, CEC and SOC as indicated by high positive or negative Spearman correlations for several horizons may increase uncertainties in parameter estimates

(Wehrens, 2020) and may hamper interpretation of the results (Crawley, 2012).

II. Description of Cu_{AR} concentrations for the different horizons depending on site, SOC, pH and CEC

The fixed-effects model for horizon BT and the mixed-effects models with site as random effect for the other horizons were generally useful with a coefficient of determination R_c^2 of 0.56 and conditional pseudo-coefficients of determination and R_c^2 ranging from 0.58 to 0.98.

Overall, the variables CEC, SOC and pH as well as their interactions were, in many cases, important for successful mixed-effects modelling, but the importance of the variables differed markedly between the horizons. The importance of SOC (present in nine of the 10 horizon-specific fixed- and mixed-effects models) for a description of Cu_{AR} concentrations is supported by the known key role of soil organic matter in Cu retention and the large fraction of organically-bound Cu in soils (Fijałkowski et al., 2012). The importance of CEC (present in eight out of the 10 fixed- and mixed-effects models) can be explained by the typically close positive relationship between CEC and clay concentration of soils and that soils with increased clay concentrations have larger Cu concentrations (UBA, 2004). The contribution of pH (present in seven out of the ten models) can be explained by—among other processes—the pH dependency of Cu adsorption and desorption processes in soils (Caporale & Violante, 2016). The general mixed-effects model which included horizon nested in the site was useful, but again reinforced the site- and horizon-specificity of relationships between Cu_{AR} and the variables CEC, SOC and pH since the random effects horizon nested in site explained the majority of the variation in Cu_{AR} content (R^2_m and R^2_c values of 0.18 and 0.89, respectively).

The horizon-specific, rule-based cubist models were similarly successful compared to the mixed-effects models: for six of the ten horizons (peat, A, AB, B, G and C), they performed better than all other modelling approaches, whereas for the other four horizons (organic layer, AC, M and BT), fixed- and mixed-effects models were slightly more successful. The importance of the different variables for a description of Cu_{AR} concentrations differed for the different horizon-specific approaches. For instance, in contrast to the mixed-effects models, the site effect was not important for most horizons in the cubist models. The main reason for this may be the multicollinearity in the data set of this observational study (Crawley, 2012).

III. Prediction of the Cu_{AR} concentrations for the A horizons for a closed population and for new sites

Five-fold random partitioning of the data indicated that performances of the rule-based cubist model and the mixed-effects model were very promising for the A horizons for a closed population. The five-fold pseudo-independent calibration–validation approach (since all sites are present in both the calibration and validation sample; Brown et al., 2005; Ludwig et al., 2017) resulted in accurate predictions for the validation sample. The results indicate that the approach is useful for estimating Cu_{AR} concentrations in future samplings using the cubist model or the mixed-effects model approach for a fixed set of observational sites.

However, independent validation after five-fold random splittings of sites into calibration and validation samples was not successful for either of the regression approaches, indicating that the three variables SOC, CEC and pH are not sufficient to accurately predict Cu_{AR} concentrations for new sites in southern Germany. Improvements in the modelling may be achieved with specific information on mineralogical compositions.

5 | CONCLUSIONS

A large data set of Cu_{AR} concentrations in different horizons of south German monitoring sites over a period of 30 years was analysed. Mixed-effects and rule-based cubist models described the Cu_{AR} concentrations in the horizons generally well as functions of the site, CEC, SOC and pH.

The importance of the different variables for a description of Cu_{AR} concentrations differed for the different horizon-specific approaches, and—since model performances of mixed-effects and rule-based cubist models were similar—the final equations of the mixed-effects models should be favoured over those of the rule-based models because the hierarchical sampling design is adequately considered in the mixed-effects models.

Validations of cubist models and mixed-effects models for the Cu_{AR} concentrations in A horizons were successful for the given population after random splitting into calibration and validation samples, but not after random splitting according to sites. The two types of validation approaches used in this study (i.e., random splits for subsequent predictions for the existing sites or splits according to the site for subsequent predictions for new sites) highlight the importance of an appropriate data splitting scheme for testing the intended model usage.

AUTHOR CONTRIBUTIONS

Bernard Ludwig: Data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); software (equal); validation (equal); visualization (equal); writing – original draft (lead). **Petra Wölfel:** Data curation (equal); investigation (equal); methodology (equal); writing – original draft (supporting).

Isabel Greenberg: Data curation (equal); methodology (equal); validation (equal); writing – original draft (supporting). **Hans-Peter Piepho:** Methodology (equal); validation (equal); writing – original draft (supporting). **Peter Spörlein:** Conceptualization (equal); data curation (equal); formal analysis (equal); funding acquisition (lead); investigation (equal); methodology (equal); writing – original draft (supporting).

ACKNOWLEDGEMENTS

We are grateful to the Bavarian State Ministry of the Environment and Consumer Protection for supporting the soil monitoring by the Bavarian Environment Agency. The federal-state funding program "Water, Soil and Waste" of the Ministry of Agriculture and Environment Mecklenburg-Western Pomerania contributed to this study. Open Access funding enabled and organized by Projekt DEAL.

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

REFERENCES

Ad-hoc Arbeitsgruppe Boden (2005). *Bodenkundliche Kartieranleitung, (ed: Bundesanstalt für Geowissenschaften und Rohstoffe in Zusammenarbeit mit den Staatlichen Geologischen Diensten)* (5th ed.). Schweizerbart Science Publishers.

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.

Barton, K. (2020). *MuMIn: Multi-model inference. R package version 1.43.17*. Retrieved from 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. https://doi.org/10.18637/jss.v067.i01

Bayerisches Landesamt für Umwelt. (2011). Hintergrundwerte von anorganischen und organischen Schadstoffen in Böden Bayerns - Vollzugshilfe für den vorsorgenden Bodenschutz mit Bodenausgangsgesteinskarte von Bayern 1:500000. *Publikationsshop der Bayerischen Staatsregierung*, 58.

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*, 6801–6813. https://doi.org/10.1016/j.gca.2010.08.044

Bloetevogel, 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. https://doi.org/10.1016/j.chemgeo.2017.11.032

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*, 251–267. https://doi.org/10.1016/j.geoderma.2005.01.001

Caporale, A. G., & Violante, A. (2016). Chemical processes affecting the mobility of heavy metals and metalloids in soil environments. *Current Pollution Reports*, *2*, 15–27. https://doi.org/10.1007/s40726-015-0024-y

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

DIN ISO 10390. 1997. Bodenbeschaffenheit – Bestimmung des pH-Wertes.

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

DIN ISO 11466. 1997. Bodenbeschaffenheit – Extraktion in Königswasser löslicher Spurenelemente.

Dinic, Z., Maksimovic, J., Stanojkovic-Sebic, A., & Pivic, R. (2019). Prediction models for bioavailability of Mn, Cu, Zn, Ni and Pb in soils of Republic of Serbia. *Agronomy*, *9*, 856. https://doi.org/10.3390/agronomy9120856

European Food Safety Authority (EFSA). 2008. Peer Review Report on Copper Compounds. October 1, 2008. 1–414.

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

Galecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R: A step-by-step approach*. Springer. https://doi.org/10.1007/978-1-4614-3900-4

Groenenberg, J. E., & Lofts, S. (2014). The use of assemblage models to describe trace element partitioning, speciation, and fate: A review. *Environmental Toxicology and Chemistry*, *33*, 2181–2196. https://doi.org/10.1002/etc.2642

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*, 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Kuhn, M. 2021. *Caret: Classification and regression training. R package version 6.0-88*. Retrieved from https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2018). *Applied predictive modelling. Springer* (2nd ed.). https://doi.org/10.1007/978-1-4614-6849-3

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

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. https://doi.org/10.1016/j.geoderma.2016.10.022

Nakagawa, S., Johnson, P. C. D., & 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*, 20170213. https://doi.org/10.1098/rsif.2017.0213

Neth, H., & Gradwohl, N. (2021). *unikn: Graphical elements of the University of Konstanz's corporate design*. Social Psychology and Decision Sciences, University of Konstanz, Germany. Computer software (R package version 0.4.0, March 25, 2021). Retrieved from https://CRAN.R-project.org/package=unikn

Peng, L., Liu, P., Feng, X., Wang, Z., Cheng, T., Liang, Y., Lin, Z., & Shi, Z. (2018). Kinetics of heavy metal adsorption and desorption in soil: Developing a unified model based on chemical speciation. *Geochimica et Cosmochimica Acta*, *224*, 282–300. https://doi.org/10.1016/j.gca.2018.01.014

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

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C. T., Reale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*, 1141–1152. https://doi.org/10.1111/2041-210X.13434

Stroup, W. W. (2012). *Generalized linear mixed models: Modern concepts*. CRC Press, Boca Raton.

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

Wehrens, R. (2020). *Chemometrics with R* (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-62027-4

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. https://doi.org/10.1201/b17336

Zuur, A.F., Ieno, E.N., Walker, N.H., Saveliev, A.S, Smith, G. M. 2009. *Mixed effects models and extensions in ecology with R*. Springer.

How to cite this article: Ludwig, B., Wölfel, P., Greenberg, I., Piepho, H.-P., & Spörlein, P. (2022). Application of mixed-effects modelling and rule-based models to explain copper variation in soil profiles of southern Germany. *European Journal of Soil Science*, *73*(3), e13258. https://doi.org/10.1111/ejss.13258