

TEXTE 00/2020

Ressortforschungsplan of the Federal Ministry for the
Environment, Nature Conservation and Nuclear Safety

Project No. (FKZ) 3715 67 420 0

Necessary adaptations for a harmonized field-testing procedure and risk assessment of earthworms (terrestrial)

Final report

by

Jörg Römbke, Bernhard Förster, Stephan Jänsch, Florian
Kaiser, Adam Scheffczyk
ECT Oekotoxikologie GmbH, Flörsheim


Martina Roß-Nickoll, Benjamin Daniels, Richard Otter-
manns, Björn Scholz-Starke
RWTH Aachen University, Aachen


On behalf of the German Environment Agency

Imprint

Publisher

Umweltbundesamt
Wörlitzer Platz 1
06844 Dessau-Roßlau
Tel: +49 340-2103-0
Fax: +49 340-2103-2285
buergerservice@uba.de
Internet: www.umweltbundesamt.de

 [/umweltbundesamt.de](https://www.facebook.com/umweltbundesamt.de)

 [/umweltbundesamt](https://twitter.com/umweltbundesamt)

Report performed by:

ECT Oekotoxikologie GmbH
Böttgerstr. 2-14
65439 Flörsheim
Germany

RWTH Aachen University
Worringerweg 1
52074 Aachen
Germany

Report completed in:

June 2020

Edited by:

Section IV 1.3 Pesticides, Ecotoxicology and Environmental Risk Assessment
Silvia Pieper, Pia Kotschik und Susanne Walter-Rohde (Fachbegleitung)

Publication as pdf:

<http://www.umweltbundesamt.de/publikationen>

ISSN 1862-4804

Dessau-Roßlau, June 2020

The responsibility for the content of this publication lies with the author(s)

Abstract: Necessary adaptations for a harmonized field-testing procedure and risk assessment of earthworms (terrestrial)

The purpose of this project was to provide scientifically robust and practical information on the variability of the endpoints assessed in earthworm field studies, the statistical significance of the results and the level of the statistically detectable effects of the chemicals tested - with the aim of developing suggestions for improving the test design. Best-practice studies reveal low power to detect differences between control and test chemical treatment plots. An adapted test design should contain an option to perform regression (EC_x) approaches, which have been suggested as an alternative to the currently performed threshold (NOEC) approach. A pilot field study was performed according to a newly developed combined NOEC- and EC_x-test design with the test chemical carbendazim. The EC_x design leads to more robust conclusions for environmental risk assessment. The calculation of effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure for given data prerequisites. If applicable to the data, the closure principle computational approach test (CPCAT) is the preferred option. The evaluation and interpretation of the data at plot (pooled samples of 1 m² in total used as replicates) and sub-plot level (single samples as replicates of 0.25 m²) should be requested. According to the experiences made during the performance of the pilot study and the results of the statistical analyses, a draft OECD test guideline was developed. As of now, the discussion of the draft test guideline is ongoing.

Kurzbeschreibung: Notwendige Anpassung zur harmonisierten Freiland-Testung und Risikobewertung für Regenwürmer (Terrestrik)

Ziel dieses Projekts war es, wissenschaftlich belastbare und praktische Informationen über die Variabilität der in Feldstudien mit Regenwürmern ermittelten Endpunkte, die statistische Signifikanz der Ergebnisse und die Höhe der sicher statistisch nachweisbaren Auswirkungen der getesteten Chemikalien zu liefern, um Vorschläge für die Verbesserung des Testdesigns zu entwickeln. Best-Practice-Studien zeigen, dass die statistische Trennschärfe zur Erkennung von Unterschieden zwischen Kontroll- und mit Testchemikalien behandelten Parzellen gering ist. Ein angepasstes Testdesign sollte eine Option zur Durchführung von Regressionsansätzen (EC_x) enthalten, die als Alternative zum NOEC-Ansatz vorgeschlagen wurden. Eine Pilotfeldstudie wurde nach einem neu entwickelten kombinierten NOEC- und EC_x-Testdesign mit der Testchemikalie Carbendazim durchgeführt. Das EC_x-Design führt zu belastbareren Aussagen für die Umweltrisikobewertung. Die Berechnung der Wirkungsschwellen (NOEC/LOEC) sollte unter den gegebenen Voraussetzungen mit dem leistungsstärksten Mehrfachtestverfahren durchgeführt werden. Wenn möglich, ist der CPCAT-Ansatz (closure principle computational approach test) die bevorzugte Option. Die Auswertung und Interpretation der Daten auf der Parzellen- (gepoolte Proben von insgesamt 1 m², die als Replikate verwendet wurden) sowie der Probenebene (einzelne Proben von 0,25 m² als Replikate) sollte gefordert werden. Basierend auf den Erfahrungen während der Durchführung der Pilotstudie und den Ergebnissen der statistischen Auswertungen wurde ein OECD-Prüfrichtlinienentwurf formuliert. Die Diskussion über den Prüfrichtlinienentwurf ist derzeit noch nicht abgeschlossen.

Table of content

List of figures	9
List of tables	13
List of abbreviations	15
Summary	17
Zusammenfassung.....	29
1 Introduction.....	42
2 Evaluation of existing data and development of proposals for an optimized design of the earthworm field test (WP1).....	44
2.1 Earthworm field study database – compilation and quality check.....	44
2.2 Data collection: environmental and agricultural variables.....	46
2.3 Field study data: Species composition, variability and MDDs	50
2.4 Development of a pilot study test design.....	54
2.4.1 First proposal	55
2.4.2 Discussion of the pilot study in the ad hoc SETAC GSIG sub-group.....	56
2.4.2.1 Final test design	56
2.4.2.2 Identification of the test chemical concentrations.....	57
3 Experimental investigations and statistical analyses (WP2)	59
3.1 Performance of the pilot field study.....	59
3.1.1 Experimental site	59
3.1.1.1 Characterisation of the experimental site	59
3.1.1.2 Field site history.....	61
3.1.1.3 Installation of experimental plots.....	62
3.1.2 Test chemical, test performance and application	64
3.1.2.1 Test chemical (a.s. carbendazim).....	64
3.1.2.2 Test design and application rates	64
3.1.2.3 Calibration of spray equipment	65
3.1.2.4 Performance of application	65
3.1.2.5 Weather conditions during application	67
3.1.2.6 Irrigation of experimental plots.....	68
3.1.3 Conditions of the experimental site during study duration	68
3.1.3.1 Maintenance of experimental plots	68
3.1.3.2 Weather conditions during the study period	72

3.1.4	Assessment of the earthworm community	72
3.1.4.1	Sampling of earthworms.....	72
3.1.4.2	Identification of earthworm species.....	75
3.1.4.3	Weighing of earthworms	75
3.1.5	Chronology of the study	76
3.1.6	Results of the study	76
3.1.6.1	Species diversity of earthworms.....	76
3.1.6.2	Abundance and biomass of earthworms before application	77
3.1.6.3	Effects of the test chemical	77
3.2	Statistical analysis: field study and database.....	80
3.2.1	State of the art of statistical procedures to analyse ecotoxicological field tests	80
3.2.2	Data description of the pilot study	84
3.2.3	Advanced statistical procedures - database and pilot field study.....	86
3.2.3.1	Analysis of natural variability in earthworm communities.....	86
3.2.3.2	Calculation of effect thresholds (NOEC)	95
3.2.3.3	Calculation of effect concentrations - pilot study	100
3.2.3.4	Community analyses (PRC) – pilot study	104
3.2.4	Design requirements for earthworm field studies -conclusions from statistical procedures	106
3.2.5	Limitations and open questions.....	108
3.3	Derivation of a new test design	109
4	Participation in the OECD process (WP3).....	111
5	Conclusions and outlook	112
6	List of references	115

List of figures

Figure 1:	Exemplary illustration of an earthworm field study test design (random design). The different colours of the boxes represent a control treatment and different concentrations of the tested substance (=treatments). The white dots correspond to the samples (= subplots) collected at each time point of sampling. Four samples (=0.25 m ²) per time of testing are aggregated to one replicate according to the current guideline.....	45
Figure 2:	Correspondence analysis of earthworm species abundance data for field studies of the database (adults only, all sampling time points and treatments)	49
Figure 3:	Distribution of the probability density for the minimum detectable difference (MDD in %) of total earthworm abundance data in the earthworm field study database extracted from ISIS (all sampling dates, empirical MDD between 11% and 100.2%).....	52
Figure 4:	Distribution of the probability density for the minimum detectable difference (MDD in %) of <i>Aporrectodea caliginosa</i> abundance data in the earthworm field study database (all sampling dates, lowest empirical MDD at 15.4%).....	53
Figure 5:	Original proposal for the design of the pilot earthworm field study provided to the involved stakeholders prior to the meeting in Flörsheim.....	55
Figure 6:	Aerial view of the experimental site in Flörsheim Wicker.....	59
Figure 7:	Experimental site on 28 March 2017, i.e. 12 days after glyphosate application	61
Figure 8:	Experimental site on 30 March 2017 after installation of the plots.....	63
Figure 9:	Scheme of the trial area with randomized allocation of the treatment to the plots (squares).....	63
Figure 10:	Application of the test chemical on 11 April 2017	66
Figure 11:	Experimental site on 24 May 2017.....	69
Figure 12:	Experimental site on 12 June 2017	69
Figure 13:	Experimental site on 25 August 2017 prior to mowing.....	70
Figure 14:	Mowing of the experimental site on 25 August 2017 with a string trimmer	70
Figure 15:	Experimental site on 28 August 2017 after mowing	71
Figure 16:	Experimental site on 23 April 2018 during the last earthworm sampling	71
Figure 17:	Buckets containing soil for hand-sorting and watering cans containing AITC solution.....	73
Figure 18:	Hand-sorting and AITC-extraction of earthworms	74
Figure 19:	Sampling vessel containing 70% ethanol and earthworms.....	74

Figure 20:	Total earthworms abundance [ind/m ²] during the pilot field test . C = control; T1 - T6: treatment rates with carbendazim (T1 = 0.6, T2 = 1.8, T3 = 3.2, T4 = 5.8, T5= 10.5, T6 = 31.5 kg a.s./ha).....	79
Figure 21:	Total earthworms biomass [g/m ²] during the pilot field test. C = control; T1 - T6: treatment rates with carbendazim (T1 = 0.6, T2 = 1.8, T3 = 3.2, T4 = 5.8, T5= 10.5, T6 = 31.5 kg a.s./ha).....	80
Figure 22:	Overview of sampled total individuals per taxonomic/morphological group or earthworm species in the performed pilot field study	85
Figure 23:	Distribution of coefficients of variation for control treatments (pilot study and database) on plot level (1.0 m ² for database studies and 1.5 m ² for pilot study) for earthworm biomass and abundance data at all tested times of sampling	86
Figure 24:	Relationship between mean control earthworm abundances and mean coefficients of variation of the control treatments [%] for the identified species and groups of all studies (orange: pilot study, blue: database studies; only selected species are labelled). Illustration at sample (= subplot) level (left, 0.25 x 0.25 m) and at plot level (right, 1.0 x 1.0 m)	89
Figure 25:	Relationship between mean earthworm control biomass and mean coefficients of variation of the control treatments [%] for the identified species and groups of all studies (orange: pilot study, blue: database studies; only selected species are labelled). Illustration at sample (= subplot) level (left, 0.25 x 0.25 m) and at plot level (right, 1.0 x 1.0 m).	90
Figure 26:	Number of required replicates (plots) per treatment in earthworm field tests plotted against the detectable difference (in percent) between treatment and control. Variation of control was set to 32.9%, which is the mean variability of total earthworms in available field studies. Coloured lines: Required replicates for controls using different numbers of test treatments (a); dotted line: Required replicates of test treatments.....	92
Figure 27:	Detectable difference (in percent) of a treatment in earthworm field tests compared to the control depending on the number of required replicates for a given variability of the community (coefficient of variation of the control) at plot level (left) and at sample (= subplot) level (right). A type-II error of 0.2 respectively a test power of 80% was fixed for the sample size simulations. Coloured lines: Required replicates for controls using different numbers of test treatments (a); dotted line: Required replicates of test treatments	93

Figure 28:	Extrapolation analysis - improvement of the detectable difference in earthworm field studies [%] depending on theoretically assumed coefficients of variation. The calculation shown is based on the following fixed parameters, similar to the earthworm pilot field study (type-I error: 0.05; test power: 0.8, number of treatments: 6, number of plots: 6; total number of samples (= subplots): 36. More details in the text.....	94
Figure 29:	Percentage difference between control and treatments for all single species and aggregated earthworm groups abundances in the pilot field study plotted against respective calculated p-values calculated with the CPCAT (blue dots) and Dunnett (orange crosses) method for all sampling time points. Background colours: Scaling of magnitude of effects as suggested in the Scientific Opinion on Soil Organisms (EFSA PPR 2017).....	96
Figure 30:	Percentage difference between control and treatments for total earthworm abundances in database and pilot field study plotted against p-values calculated with the CPCAT (blue dots) and Dunnett (orange crosses) method for all sampling time points. Background colours: Scaling of magnitude of effects as suggested in the Scientific Opinion on Soil Organisms (EFSA PPR 2017).....	97
Figure 31:	Histogram of the classified frequencies (class width: $p=0.1$) for the difference of p-values between Dunnett test and CPCAT method for all tested earthworm species groups, treatments and sampling time points of the pilot field study. Daa = days after application	98
Figure 32:	Percentage of cases where the calculated NOEC of endpoints in the earthworm pilot field study according to CPCAT is higher (yellow), lower (green) or equal (blue) to the NOEC of the Dunnett procedure. Daa = days after application	99
Figure 33:	Dose-response curves of the group "total earthworms" using the endpoint measure total abundance (adults & juveniles) for pilot study data (regression method: Probit). Daa = days after application	101
Figure 34:	Percentage of significant dose-response relationships for all calculated regression procedures with the endpoints of the of the earthworm pilot field study (F-test, threshold $p < 0.05$). Single findings of species were not considered. Daa = days after application	102

Figure 35:	Percentage of detectable EC ₁₀ (left) and EC ₅₀ values (right) of all calculated dose-response curves for endpoints of the earthworm pilot field study data. EC values are detectable in this representation if they lie between concentration levels and can therefore be interpolated on a calculable regression line. The significance of the curve as a pre-test (see above) as well as single findings of species were not considered here. Daa = days after application.	103
Figure 36:	Exemplary illustration of an elaborated curve adaptation method by integrating a so-called hormesis function for the earthworm data set <i>total epilobous adults</i> of the earthworm pilot field study (sampling time: 188 days after application, endpoint: total abundance). A modified four-parametric log-logistic model according to Brain Cousens was used to model an hormesis-like response.....	104
Figure 37:	Principal-Response-Curve (PRC) for species abundance data of the earthworm pilot study. Different treatments (application rates from 0.6 to 31.5 kg carbendazim/ha) have different colours	105
Figure 38:	Principal-Response-Curve (PRC) for species biomass data of the earthworm pilot study. Different treatments (application rates from 0.6 to 31.5 kg carbendazim/ha) have different colours	106
Figure 39:	Scheme of the statistical testing procedure for earthworm field study data when assessing differences between treatments and controls (e.g. for No Observed Effect Concentrations, NOEC; calculation in Mixed Design)	108

List of tables

Table 1:	Number of plots and treatments for the ECx- and the mixed-design in earthworm field tests. More information on the design type in the text above. C control; T 1-x treatments; R reference substance	25
Table 2:	Environmental and agricultural variables, study class and number of sampling time points of selected earthworm field tests from the ISIS Database (UBA)	47
Table 3:	Percentages of sampled earthworm species and assigned ecological and morphological groups.....	51
Table 4:	Number of plots and treatments for the combined NOEC- and ECx-design in the pilot earthworm field study. C = control; T1-T6 = treatments	57
Table 5:	Application rates of the earthworm pilot field study. Concentrations are given in kg active substance (a.s. carbendazim)/hectare (ha).....	57
Table 6:	Physical and chemical characterization of the field soil (0 – 10 cm depth)	60
Table 7:	History of the field site with regard to crop and application of fertiliser and plant protection products.....	62
Table 8:	Characterization of the test chemical	64
Table 9:	Test Design, application rates, test chemical concentration in the spray solution and application rate per plot.....	65
Table 10:	Actual applied volumes of spray solutions of the test chemical.....	67
Table 11:	On-site air and soil temperature and wind velocity during spray application	68
Table 12:	Mean, minimum and maximum monthly air temperature, mean monthly soil temperature and monthly cumulated precipitation [mm] during the field trial period (April 2017 – April 2018)	72
Table 13:	Air and soil temperature, soil moisture and general weather conditions at the four earthworm sampling dates of the study	75
Table 14:	Chronology of the study	76
Table 15:	Earthworm species found during the pilot field study across both treated and untreated plots	77
Table 16:	Mean abundance [ind/m ²] and biomass fresh weight [g/m ²] of total earthworms (adults and juveniles) during the pilot field study (± standard deviation). T1 – T6: treatment rates with carbendazim (kg a.s./ha)	78
Table 17:	Abundance and biomass [% of control] of total earthworms (adults and juveniles) during the pilot field study. T1 – T6: treatment rates with carbendazim (kg a.s./ha).....	78

Table 18:	Mean coefficients of variation for different endpoints from control treatments in earthworm field studies (pilot study and database, mean of all sampling time points) on plot level (1.0 m ² for database studies and 1.5 m ² for pilot study) and sample level (0.25 m ²).....	87
Table 19:	Scaling of magnitude of effects (= "Effect classes") according to the EFSA Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms (EFSA PPR 2017)	91
Table 20:	Overview of the percentage of test procedures with significant effects (p < 0.05) Calculations according to Williams, Dunnett and CPCAT (all sampling times). Database= available field studies with standard design. Pilot field study = extended design	99
Table 21:	Number of plots and treatments for the ECx- and the mixed-design in earthworm field tests. More information on the design type in the text above. C control; T 1-x treatments; R reference substance	110

List of abbreviations

AITC	Allyl isothiocyanate
a.s.	Active substance
BBA	Federal Biological Research Center for Agriculture and Forestry, Brunswick
C	Control
CA	Correspondence analysis
CCC	Chlormequat chloride
CEC	Cation exchange capacity
CIP	Chemisches Institut Pforzheim
CP	Closure principle
CPCAT	Closure principle computational approach test
CPFISH	Closure principle and Fisher-Freeman-Halton test
C_{org}	Organic carbon
CRO	Contract research organization
CV	Coefficient of variation
DAA	Days after application
DBA	Days before application
dm	Dry matter
DNA	Deoxyribonucleic acid
DT90	90% dissipation time
DWD	Deutscher Wetterdienst, Offenbach
EC	European Community
ECT	ECT Oekotoxikologie GmbH, Flörsheim
EC_x	X % effective concentration
EFSA	European Food Safety Authority
ERA	Environmental risk assessment
EU	European Union
GD	Guidance document
GLM	Generalized linear model
GPS	Global Positioning System
GSIG	Global Soil Interest Group
ISIS	Information System Chemical Safety
ISO	International Organization for Standardization
K_{oc}	Octanol carbon partition coefficient

K_{ow}	Octanol water partition coefficient
LOEC	Lowest observed effect concentration
LUFA	Landwirtschaftliche Untersuchungs- und Forschungsanstalt, Speyer
MDD	Minimum detectable difference
MSD	Minimum significant difference
NEC	No effect concentration
NOEC	No observed effect concentration
N_{tot}	Total nitrogen
OECD	Organization for Economic Cooperation and Development
OM	Organic matter
PPR	Panel on Plant Protection Products and their Residues
PRC	Principal response curve
QSAR	Quantitative structure–activity relationship
RDA	Redundancy analysis
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen
SC	Suspensible concentrate
SETAC	Society of Environmental Toxicology and Chemistry
SSD	Species sensitivity distribution
T	Treatment
TG	Test guideline
TKTD	Toxicokinetic-toxicodynamic models
TME	Terrestrial model ecosystem
UBA	German Environment Agency
US EPA	United States Environmental Protection Agency
VR	Validation report
WHC_{max}	Maximum water holding capacity
WNT	Working Group of National Co-ordinators of the Test Guidelines Programme
WP	Work package

Summary

Introduction

Since 1994, the risk of chemicals for earthworms in the field is assessed by a test that was originally standardised by the German Federal Biological Institute (BBA). Since 1999, an international guideline standardised by International Organisation for Standardisation (ISO) is available that has been updated several times up to now (last in 2014) without changing the basic approach (ISO 11268-3). However, ISO guidelines focus on the assessment of (potentially) contaminated compartments (water bodies, sediments, waste materials as well as soils), i.e. they are used in a retrospective approach for an environmental risk assessment. In contrast, OECD test guidelines serve in general the purpose of a prospective assessment of individual chemicals and defined chemical mixtures such as pesticide formulations. As a consequence, several ISO guidelines used in the testing of chemicals were transcribed to the OECD format during the past 10 years. In the course of this conversion, which in the case of the earthworm field test is performed under German lead since April 2013 as OECD project no. 2.47 ('New Test Guideline on Determination of Effects on Earthworms in Field Studies'), it was also checked whether -apart from formal adjustments- further amendments were necessary. This assessment was performed by an ad hoc sub-group of the Global Soil Interest Group (GSIG) of the Society for Environmental Toxicology and Chemistry (SETAC) gathering representatives of academia, industry and authorities. Based on the experiences made during the past 20 years it was decided that several aspects of the guideline need adjustment to reflect the scientific progress. Specifically, besides technical details, the study design and the statistical evaluation of the test results had to be optimised. Regarding the study design, the ISO Guideline already mentions the possibility of performing studies according to a dose-response design, an option that is deemed to "clearly facilitate environmental risk assessment compared to single dose studies" (ISO 2014). In particular, due to the variability of the endpoints assessed in the field, the test design and evaluation needed improvement, so to increase the statistical significance of the results of the field test and the level of safely detectable effects of the tested chemicals. In addition, some assessment criteria proposed by the European Food Safety Authority (EFSA PPR 2017) needed to be translated in measurable endpoints. To address these issues, scientifically robust and practical information was missing. The generation of this information was the objective of this project. In close cooperation with the ad hoc SETAC GSIG sub-group, the following aims were reached by performing three work packages (WP):

- ▶ WP1: Evaluation of existing data and development of proposals for an optimized design of the earthworm field test: Compilation and critical evaluation of information available in the literature and the database of the German Environment Agency (UBA) regarding the standardised performance of earthworm field studies to develop an improved test design;
- ▶ WP2: Experimental investigations and statistical analyses: (1) Performance of a pilot field study according to the new test design. (2) In-depth statistical analysis of the pilot field study in combination with the existing database regarding natural variability in earthworm communities. (3) Calculation of effect thresholds, effect concentrations and community analysis. (4) Formulation of design requirements for earthworm field studies and identification of limitations and open questions;
- ▶ WP3: Participation in the OECD process: Formulation of a new draft OECD test guideline (TG) based on the existing ISO guideline 11268-3 but following the formal requirements of

the OECD, using the experiences made in the pilot study as well as the evaluation of the UBA database. Discussion of this draft guideline within the ad hoc SETAC GSIG sub-group in a final project meeting. The combined results of the development and discussion process will be submitted to OECD.

Evaluation of existing data and development of proposals for an optimized design of the earthworm field test (WP 1)

In the course of the preliminary analyses and investigations, the ISIS database (“Information System Chemical Safety”) of the UBA was identified as a useful source for data analysis of earthworm field tests. The database held 150 entries for field studies on earthworms. Quality criteria for data were initially defined with regard to further statistical investigations. Raw data “abundance” and “biomass” on sample level (0.25 m²) were extracted from original study reports. A unified database was developed for further statistical analysis. The subsequent systematic procedures of descriptive metadata analysis and advanced statistical calculations were performed.

Earthworm field study database – compilation and quality check

Only earthworm field studies possessing the following characteristics were used for statistical analyses: Earthworms should have been sampled by a combination of formalin/allyl isothiocyanate (AITC) extraction and hand-sorting. A bias of the sampled species composition due to the use of the octet sampling was therefore prevented. Moreover, the technical reports should include raw data collected on the sample (= subplot) level. This prerequisite enabled an analysis of test data at sample level in comparison to the conventional evaluation at plot level. The 21 field studies that fulfilled these characteristics were divided into two classes: Tests with only one treatment and one reference compared to the control (limit test) were assigned to class 1, while tests with several treatment levels were classified as class 2. Eleven field studies were classified into class 1 (limit-tests), two field studies assessed two different substance concentrations next to the control, and another eight field studies were designed with three treatments (class 2). In addition, further 5 studies with digitalized raw data at sample or plot level were integrated into the database, each with a slightly different sampling method. In total, data of 26 field tests of the ISIS database (+test data of the pilot study performed in this project) were used for statistical calculations. The processed field studies were carried out according to the ISO guideline 11268-3 or in consideration of the BBA (Biologische Bundesanstalt) guideline part VI, 2-3. Therefore, the analyzed test procedures followed a common approach. All reports contained information on earthworm species, numbers, and biomass collected for sampling plots treated with a test substance in a randomized arrangement (four replicates per treatment) and compared with those collected from control and reference plots. Every replicate (=sampling plot) consisted of four aggregated samples (= subplots) of 0.25 m² per sample (1 m² sampling plot in total). The sampling dates were usually set about 1-3 months, 4-6 months and 12 months after application. Tests usually started in April/May. The calculations of effects within the test procedures were mainly limited to the evaluation of abundance and biomass on species level and for total earthworms. Juvenile earthworms were summarized and evaluated on genus level (morphological groups: *tanylobous* and *epilobous*). In addition, the ecological groups of endogeic, epigeic and anecic earthworms were differentiated.

Data collection: environmental and agricultural variables

Descriptive metadata of the field studies revealed that the composition of species among all field studies consisted of 6 to 14 species per study. The respective Shannon Diversity Index was between 0.3 and 1.6 (mean: 1.2). The diversity index was slightly higher on grassland sites (mean: 1.44) than on other land use types (bare soil: 1.27; crop sites: 1.05). Accordingly, the minimum

number of species in grassland was at least 10. The mean number of individuals sampled was about 372 per m² on grassland, 356 on bare soil and about 196 on crop sites. The dataset available did not allow for an in-depth analysis of the potential systematic impact of environmental conditions or land use type on the earthworm community.

Field study data: Species composition, variability and MDDs

Based on the ISIS-database pre-processing, data of field studies for earthworm communities were subsequently analysed. The sampled individuals of the 21 field studies belonged to 17 different species. As a statistical measure, the minimum detectable difference (% MDD) between control and treatment of all field studies was calculated. Although the most likely value of the MDD for abundance data of total earthworms in the database was 45%, the probability of obtaining an MDD smaller than 50% of the control was 42%. An MDD between 10% and 35% (proposed in the EFSA soil opinion (EFSA PPR 2017) as small effects on the protection goals) was calculated with a probability of 8%. The same calculations for total biomass gave even lower power values than for total abundance: an MDD smaller than 50% was only detected for 32% of all sampling time points. For the aggregated group of total earthworms, the most powerful MDDs were calculated. For the most dominant species in the database, *Aporrectodea caliginosa*, the possibility to detect statistically significant effects in the field studies was even worse. Individuals of *A. caliginosa* had a very low probability to show MDDs less than 50% (12% of all sampling time points within the database). The most likely value of the calculated probability distribution for MDDs of *A. caliginosa* was 66%. Again, even lower MDDs were calculated for the endpoint biomass. In an overall picture, best-practice studies (using a combination of hand-sorting and formalin/AITC extraction for earthworm sampling) revealed low power to detect differences between control and treatment plots for aggregated taxa. Thus, based on statistical considerations, the testing and adaption of a new field study test design in the course of this project was justified. The limitations of the old design, covering limit-tests as well as NOEC-approaches, became evident. Therefore, an adapted test design should contain an option to perform regression approaches as an alternative to the NOEC approach.

Development of a pilot study test design

In a joint discussion between the UBA and the project consortium, the results of the evaluation described above led to a first proposal of the earthworm pilot field study design to be performed in 2017. This design of the experimental pilot study was characterized by combining a so-called NOEC- with an ECx-design and was called “mixed omni-design”:

- ▶ Four sampling dates, covering a total test duration of one year (as in ISO guideline 11268-3);
- ▶ One control (C) and six test chemical treatments (T) (only limit test in the ISO guideline);
- ▶ Number of plots per treatment six (C, T2, T5) or three (T1, T3, T4, T6) (four in the ISO guideline);
- ▶ Five samples per plot (four in the ISO guideline).

Running such a study meant that in total 30 plots with 150 samples per sampling date had to be covered. This original proposal was considered by the project team as large but still practical in terms of handling (e.g. number of days needed for sampling, field size etc.).

This proposal of the test design for the pilot study was discussed during the meeting of the ad hoc SETAC GSIG sub-group in February 2017. Further recent contributions addressing different aspects of the planning, performance or evaluation of earthworm field studies were presented to

the group. In the following discussion during the meeting various changes to the “mixed omnidesign” were proposed, all of them with the intention to improve the quality of the study output but without strongly increasing the efforts at the same time. The resulting final test design for the pilot study was called “balanced design”. It was decided to take six samples per plot in the NOEC- as well as in the ECx-plots and the number of replicate NOEC- and ECx-plots were six and three per treatment, respectively.

The selected test chemical was carbendazim, since it is by far the best-studied pesticide in soil ecotoxicology due to its use as reference substance in earthworm laboratory and field tests. Using the available information, various carbendazim concentration ranges were discussed. The following six application rates (plus a negative, i.e. water-only, control) were finally selected to cover a range spanning from concentrations where no effects are expected to concentrations where strong effects are likely: 0.6, 1.8, 3.2, 5.8, 10.5, and 31.5 kg carbendazim/ha. In the currently used ISO guideline 11268-3, the reference substance carbendazim should yield a statistically significant difference of at least 50 % on overall abundance and/or biomass compared to the control at least at one sampling date, when applied at rates of 6 to 10 kg a.s. carbendazim/ha. Thus, such effects should be detectable at the three highest application rates. Accordingly, and referring to the experiences made in an EU project focusing on the development of a standard semi-field method where Terrestrial Model Ecosystems (TME) have been employed, no detectable effects should appear at the two lower rates. A priori analyses have shown that an EC₅₀ could be expected at rates around 2.5 kg carbendazim/ha.

Experimental investigations and statistical analyses (WP 2)

Performance of the pilot field study

Arable land was chosen for the trial. It was surrounded by agricultural fields and pathways. The experimental plots were installed within an area of approximately 55 m by 107 m. Winter wheat was grown on the field before the study took place. To free the experimental site from vegetation without soil tillage that would have impacted the earthworm community, glyphosate was applied at a rate of 1.8 kg a.s./ha. For each treatment, i.e. control (C) and six different test chemical (carbendazim) treatments (T1 to T6), six (C, T2, T5) or three (T1, T3, T4, T6) plots (= replicates), each 10 m by 10 m, were installed at the field site and assigned randomly. The distance between two neighbouring plots was 3 m and the distance to the surrounding fields or cart tracks was at least 5 m. The test chemical was applied as the suspensible concentrate (SC) formulation Carbomax 500 SC once on 11 April 2017. The water (control) and the test chemical were applied onto the bare soil surface at a wind velocity below 3 m/sec to avoid any risk of cross contamination due to possible drift during application. All experimental plots were irrigated directly after application by means of a tractor-pulled tank wagon with at least 1000 l/plot (equivalent to 10 mm precipitation). The experimental plots were left to natural development of vegetation. No agricultural practices such as tillage, application of plant protection products or fertilizers, were undertaken. On 25 August 2017 all plots were mowed with a string trimmer and all cuttings were left on the plots.

Eight to six days prior to the first application of the test chemical, earthworms were sampled on all plots. The mean total number and the mean biomass of earthworms were determined for each of the thirty plots, designated either for test chemical treatment or to serve as untreated controls. The mean number of earthworms collected (hand-sorting and AITC-extraction) before application ranged from 413 to 512 ind./m² - hence fulfilling the requirements of the ISO guideline 11268-3. Earthworms were sampled at each sampling time point by a combined hand-sorting and AITC extraction method. Six random samples of an area of 0.25 m² (50 cm x 50 cm) to a depth of approximately 20 cm were taken per plot. Hence, there were 18 (3 plot replicates)

or 36 (6 plot replicates) individual samples per treatment and sampling time point. The distance between two samples taken on the same date and plot was at least 2 m. The sampled area was marked and not used again at subsequent sampling dates. Samples were taken at least 2 m apart from the plot border. Five to ten litres of an AITC solution (0.1 g/l) were poured uniformly into the remaining cavity to catch earthworms from deeper soil layers. The soil was carefully searched for earthworms by hand-sorting. These worms and those extracted by AITC were preserved in a 70% ethanol solution in watertight containers.

The worms were identified by means of a binocular microscope, using morphological characters. Adult worms were determined to the species level. Juveniles were classified according to the genus level, but in some cases a distinction of small worms belonging to closely related genera was not possible (e.g. *Allolobophora* and *Aporrectodea* were combined). All adult worms of one sample belonging to a particular species and all juvenile worms belonging to a particular genus were weighed together. The field site was inhabited by an earthworm population which can be considered typical for central European arable land (ISO 11268-3) including the ecological most important groups of anecic and endogeic earthworms. In total, nine different species of earthworms were found during the study. The lumbricid biocoenosis was dominated by juveniles of the endogeic genera *Aporrectodea/Allolobophora* and *Allolobophora chlorotica* was the most abundant species.

The test chemical Carbomax 500 SC (a.s. carbendazim) caused a clear reduction in total abundance and biomass at all three post-application sampling time points. Compared to the control, mean abundance and mean biomass in the test chemical treated plots were 15-59% and 11-55%, respectively at 34-36 days after application (DAA), 45-90% and 69-111%, respectively at 188-190 DAA, and 38-74% and 80-113% respectively at 377-379 DAA.

Statistical analysis: field study and database

A set of different statistical data analysis procedures were conducted for both data of the pilot study and existing test data from the UBA database. The main focus was to improve the conventional statistical methods to evaluate earthworm field studies (ISO 11268-3) and to acquire insights for statistical considerations regarding an adapted test design for earthworm field studies. Raw data for biomass and abundance at all sampling dates and for all taxa and morphological or functional earthworm groups were integrated at sample- and plot level into the existing database of the project. Single species calculations and analyses did not show any significant effects at the last sampling time point (377-379 DAA). With “*Aporrectodea/Allolobophora* spp. juvenile”, a statistically significant effect could be observed in a taxonomic group after one year. Due to the high dominance of this group in the overall data set, a reduction of abundance and biomass after 12 months was also indicated in other aggregated groups. However, this was exclusively caused by juveniles of *Aporrectodea/Allolobophora* spp. This example illustrates the need for assessments of different types of endpoints and earthworm groups (e.g. species level and group level), to avoid only general conclusions for effects of test substances based on an aggregated endpoint such as total abundance of all earthworms.

The natural, heterogeneous scattering of earthworm species within a field is a decisive factor for the statistical visibility of possible effects caused by applied environmental chemicals. The variability of tested endpoints in database and pilot field studies was assessed using the coefficient of variation (CV) of field study control treatments to derive conclusions and suggestions for improvement regarding the test power. The natural variability of the species groups in field studies was illustrated descriptively as the variance of the control treatments and used as a basis for multiple sample planning. Aggregated earthworm groups had the lowest CVs while rare species showed a comparatively high relative scattering between plots. Results indicated that particularly

aggregated species groups with high abundances and biomass values provide powerful end-points (especially low variation in controls and treatments). On average, the scattering at the single species level seemed for many species too high to prove statistically significant effects. A high variation in control treatments thus leads to a lower detectability of possible effects of the test substance (= high MDD).

The impact of variance on the number of required replicates to achieve a certain test power was determined for the standardised Dunnett test. Calculations were based on CVs for control treatments in earthworm field tests and applied for a dynamical sample size planning for an MDD that should be achieved. For the development of an adapted test design we investigated how many samples (=replicates) should be used given a desired target-test power and a given natural variability of data. By default, the desired test power is usually set to 80% for statistical hypothesis testing. The MDD that can be achieved with the respective sample sizes was classified into four different classes in the simulation, adapted to the scaling of magnitude of effects on the protection goals as proposed in the EFSA soil opinion¹. Up to 10% difference between control and treatments was defined as negligible deviation, up to 35% as small effects, between 35 and 65% as medium effects and higher than 65% as large effects. Even if it is not required to measure these effect ranges in field tests, a comparison with the available data was performed. Results of the sample size simulation for mean total earthworm variability indicated that standardized earthworm field tests might have an insufficient number of replicates to detect small effects with a test power of 80% for “total earthworms”, which was the group with the lowest CVs in earthworm field test data. Accordingly, the ability to detect effects for other earthworm groups is even more limited.

For this reason, it was investigated if a NOEC calculation using samples (=subplots) as statistical replicates would result in an improvement with regard to MDDs. We calculated the sample size planning with the measured CVs on plot level (current standard method), and on sample level to assess shifts in test power. Results of the pilot study showed that the increase in plot numbers (n=6) and the slightly lower CVs (1.5 m² sampled instead of 1.0 m²) increased the test power. The number of required replicates to achieve a certain threshold of MDDs decreased compared to database field studies. Nevertheless, using this test design, medium effects (35% - 65% effect) will very often not be detectable with a power of 80%. The comprehensive detection of small effects (10% - 35%) with a test power of 80% appears not to be achievable in this simulation considering realistic numbers of replicates. Nevertheless, a 35% difference would be detectable at sample (= subplot) level. In this case, the mean CV at subplot level would be slightly higher than at plot level (34.56%). For this reason, at least 14 replicates would be necessary to detect medium effects. By using the single samples as replicates there would be 36 replicates available in this statistical design. This switch in the assessment level allows the identification of more significant differences, especially at single species level. In general, the statistical detectability of effects always improves if the evaluation is carried out at the subplot level.

The calculation of the effect thresholds was carried out for all studies in the database and at all sampling times and a comparison of the Dunnett method with results from the so-called CPCAT approach (Closure Principle Computational Approach test) was performed. The theoretical distribution assumption of earthworm abundance field test data follows a Poisson model. Therefore, the application of the CPCAT approach is highly recommended for abundance data due to more powerful test statistics. This is the first time in which the performance of CPCAT was assessed within a comprehensive meta-analysis of field study data. It was shown that the use of the CPCAT

¹ EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues) (2017): Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. EFSA Journal 15(2):4690, 225 pp. doi: 10.2903/j.efsa.2017.4690

procedure in comparison to the Dunnett test increased the probability that significant effects were identified, even at small effects (10% - 35%). CPCAT is therefore generally "more selective", i.e. significant effects of the test substance are already indicated for smaller differences to the control. The differences in test power between the two procedures also became visible in the separate examination of the NOEC calculations for different species groups. The Poisson distribution used in CPCAT procedures describes the earthworm community data in outdoor tests mathematically and statistically more accurately than the normal distribution used in conventional t-tests (e.g. Dunnett or Williams). Thus, the use of the CPCAT approach increases the test power for earthworm field data. However, for the relatively new CPCAT approach there is currently no procedure for the generic calculation of a quantitative measure of test power and sample planning using the CPCAT procedure is not yet possible.

In contrast to the database studies, the pilot study was carried out in a test design with several concentration levels. Probit curve regressions were conducted and at all three samplings after application, a significant dose-response relationship was identified in the group "total earthworms". Unlike the NOEC approach (Dunnett test), the choice of an EC_x design allowed revealing significant relationships between the carbendazim concentration applied and the measured effect on the earthworm population (total abundance in the pilot study) across the whole range of concentrations. In addition, the comparison of EC₅₀ values across sampling times indicated recovery effects that could be assumed in case of the total earthworm community between 1 and 6 months after application. The calculated EC₅₀ increased by a factor of 10 during this period, whereas it did not change in any further comparison at the sampling time after 12 months. By contrast, the EC₅₀ for (*Aporrectodea/Allolobophora* spp.) juveniles increased only by a factor of approximately 3 until the end of the study. The results of the study showed that the use of an EC_x design to derive effect concentrations on the earthworm population in the field is generally feasible. However, the choice of a suitable concentration range for adequate testing of all species and aggregated groups poses a challenge.

A Principal Response Curve (PRC) was used to answer the question whether there was a significant relation between community structure and treatments. The PRCs revealed a highly significant effect of the treatment on the earthworm community (p-value < 0.05). A clear dose-response relationship was visible and with increasing concentrations the deviation from the control increased. According to the PRC, a recovery of the community (abundance of adults for single species regaining initial state) could be assumed at all test concentrations after approximately one year.

Based on these investigations, generic derivations of recommendations are limited due to the high variability in the various earthworm data sets of different field tests and due to the expectable impact of local site conditions. However, the following basic recommendations and requirements regarding the implementation and evaluation of earthworm field tests were identified:

1. There is still a need to determine and evaluate biomass and abundance at species level, as the aggregated morphological or functional groups used may disguise effects on single species.
2. The EC_x design is a meaningful alternative to the NOEC design in the earthworm field test. At least a mix design would be advisable. In fact, the EC_x design leads to stronger/more protective statements for environmental risk assessment (ERA) especially at lower effect ranges, a masking of possible effects as in the NOEC evaluation is avoided.
3. The calculation of effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure for given prerequisites. If possible, the CPCAT approach is preferred. If data are metric (e.g. biomass), multiple t-test procedures such as Dunnett's or Wil-

liams' test ($\alpha = 0.05$, two-sided for unclear direction of response) should be performed for multiple comparisons in a randomized plot design. The prerequisite of normally distributed data and variance homogeneity has to be tested using e.g. Shapiro-Wilks and Levene's test procedures, respectively. If data do not fulfil the criterion of normality, generalized linear models or non-parametric tests e. g. the Bonferroni U-test or the Jonckheere-Terpstra Step-down-test (homogeneity of variance required) can be applied. The theoretical distribution assumption of earthworm abundance field test data follows a Poisson model. Therefore, the application of the CPCAT approach is highly recommended for abundance count data due to more powerful test statistics. Nevertheless, if abundance data show homogeneity of variances, the null-hypothesis of normal distribution is not rejected and absolute abundances per replicate are > 5 , the application of parametric test procedures (Williams, Dunnett) is also feasible. For multiple t-test procedures and with unequal replication, the table t-values must be corrected as suggested by Dunnett and Williams. In addition, an inappropriate log-transformation of data during the calculation procedure should be avoided.

4. After data revision, it should be decided whether a simple two-parameter Probit (Logit, Weibull) regression, a nonlinear regression or the integration of a so-called hormesis model for the calculation of effect concentrations (EC_x) is necessary. In case of a monotonous increase of the measured endpoint (biomass, abundance), the derivation of significant effect concentrations should also be taken into account.
5. If there are no ecological reasons for not using the data at sample level (i.e., no proven interdependency between samples from the same plot), the evaluation and interpretation of the data at plot (pooled samples of 1 m^2 in total used as replicates) and sub-plot level (single samples as replicates of 0.25 m^2) should be requested.
6. Principal response curves are generally applicable within the EC_x -design and a powerful tool for community analyses. They should be carried out in addition to uni-variate methods when appropriate data are available, for tests with multiple treatments (e.g. EC_x design).

Some limitations and open questions regarding the proposed changes need to be kept in mind. The recommendations towards adjustments of the field study test design reveal two opposing trends whose benefits and downsides for the significance of the test have to be balanced. On the one hand, as many test-concentration levels as possible should be considered for a meaningful EC_x design. From a strictly statistical point of view, replication of the concentration levels is not needed for the subsequent regression analysis. A strong design for calculating robust NOEC values requires, as shown, a substantial increase in the number of replicates per control and treatments. These two demands need to be weighed and integrated into a new design depending on the underlying test concept and desired endpoints. However, this decision is not a strictly statistical one, but primarily a question of feasibility in the field (plot numbers and field sizes to be handled) and a question of regulatory prioritization of various endpoints.

In addition, the analyses and underlying data presented above have a few limitations that should not be forgotten: The results for the implementation of an EC_x design in field studies are based on a proof-of concept pilot field study at one site and with the well-known reference substance carbendazim. In this case, a sound prior knowledge and experience from earlier field studies on possible effect widths and dynamics was available. This is not the case, in particular, for new substances in regulatory practice. In such cases, the choice of concentration ranges in earthworm field tests might be considerably more difficult. Furthermore, the pilot field study demonstrates that an applied concentration range provides different dose-response curves for earthworm species and groups due to their different sensitivities, as also in the previously used NOEC design. If some species do not react to the test chemical, then no dose-responses and NOEC can be derived. However, the statistical endpoint of the NOEC disguises this to a large extent.

For the derivation of NOEC values with abundance data, CPCAT represents a meaningful alternative to the standardized test procedures of t-test statistics. However, it should be mentioned that there is still no established methodology for the calculation of test power and corresponding sample planning for CPCAT. Also, CPCAT should achieve higher acceptance as an appropriate tool for assessing the results of ecotoxicological tests, for example by being applied as a standard analysis method in a wider range of standard ecotoxicological test methods.

The CPCAT procedure is not suitable for metric data because the Poisson distribution does not adequately describe this type of data. To improve the statistical test procedures for metric data, it might be considered to integrate the closure principle into multiple t-test procedures to prevent alpha inflation.

The use of the samples as replicates for the calculation of NOEC values leads to an improvement of the test power. A general investigation of the effects in earthworm field tests at both plot and sample (= subplot) level is therefore recommended based on these results (provided that ecological conditions exist for the use of subplots as replicates). Whether this is a useful option in consideration of the debate on pseudoreplicates in field studies remains to be discussed. Within a regulatory framework, the following steps could be considered: A respective endpoint is evaluated at both subplot and plot level. If the same NOEC values are obtained as results, these are considered; if other (smaller) NOEC values are calculated at subplot level, the following procedure is suggested: If it is not possible to reliably demonstrate a relic of the plot effect at this level, the smaller NOEC should be used for the regulatory process. This is not necessarily a decision based on purely scientific considerations, but a regulatory, protective decision based on the precautionary principle.

Derivation of a new test design

The experience gained during the performance of the pilot study as well as the statistical evaluation of this study and the UBA database were applied to derive a proposal for a new test design (Table 1).

Table 1: Number of plots and treatments for the ECx- and the mixed-design in earthworm field tests. More information on the design type in the text above. C control; T 1-x treatments; R reference substance

Test design	Plots per treatment (No.)									Plots (sum)	Samples (total No.)
	C	T1	T2	T3	T4	T5	T6	(T7)	R		
ECx Design	3	3	3	3	3	3	3	(3)	3	24 (27)	96 (108)
Mixed Design	6	2	6	2	2	6			3	27	108

Participation in the OECD process (WP 3)

The experience gained in the more than 20 past years of performing earthworm field studies based on the existing BBA and ISO guidelines and during the project was used to formulate a new draft OECD test guideline including a proposal for a new test design. The draft OECD test guideline was distributed to the ad hoc SETAC GSIG sub-group in March 2019 as a basis for discussion during the final project meeting at UBA in Dessau. A multitude of comments were provided during and after the meeting which were compiled in a commenting table according to the

OECD process. This table is currently under review to create an updated version of the draft test guideline that will then be subject to the further OECD process.

Conclusions and outlook

The purpose of this project was to provide scientifically robust and practical information on (1) the variability of the endpoints assessed in earthworm field studies, (2) the statistical evaluation of the results and (3) the level of the statistically detectable effects of the chemicals tested. The final aim was to provide suggestions for an improved test design. Critical evaluation of information available in the literature and the database of the UBA revealed the following shortcomings of the currently used earthworm field test design according to ISO standard 11268-3:

- ▶ The evaluated best-practice studies (i.e. using a combination of hand-sorting and formalin/AITC extraction) reveal low statistical power to detect differences between control and treatment plots for aggregated taxa. For single species, this statistical potential for a reliable identification of effects is even lower. The overall MDD is not low enough for a comprehensive detection of small or medium effects.
- ▶ NOEC and related concepts have long been criticized in the ecotoxicological literature. Furthermore, the actual MDD calculations of field studies revealed that potentially relevant effects are not detectable in many field situations by the current standardized statistical procedures.
- ▶ An adapted test design should contain an option to perform regression analyses, which have been suggested as an addition to the NOEC approach. The resulting estimated concentrations (ECx values) from fitting a curve to the data have been proposed as a more meaningful alternative to the NOEC-value. Thus, the number of concentration levels in the pilot field study has to be increased to investigate the suitability of an ECx-design for earthworm field studies.
- ▶ To still include the possibility of deriving NOEC values as well as to improve the statistical power of this procedure compared to the old design, the number of replicates on the plot level for the control and test concentration treatments need to be increased.
- ▶ The number of samples per replicate should be increased to examine the changes in variance and to estimate if these samples can be used as individual replicates to improve statistical test power.
- ▶ As the field conditions and practical feasibility of the pilot field study limited the total number of plots, the enlargement of the concentration levels and the increase of plots and samples (=subplots) per treatment had to be adjusted in such a way that both research questions (feasibility of ECx design and improvement of NOEC design) could be addressed.
- ▶ Based on these evaluations, a pilot field study was performed according to a newly developed combined NOEC- and ECx-test design with the test chemical carbendazim. One control (C) and six treatments (T) were used. The number of plots per treatment were six (C, T2, T5) or three (T1, T3, T4, T6). The number of samples per plot was six. The results of the pilot field study and the in-depth statistical evaluation of additional earthworm field studies yielded the following design requirements for earthworm field studies:

- ▶ Abundance and biomass should be determined and evaluated at species level as aggregated morphological or functional groups may disguise effects on single species.
- ▶ The ECx design is a meaningful alternative to the NOEC design but at least a mixed design would be advisable. The ECx design leads to more robust conclusions for ERA, a masking of possible effects as in the NOEC evaluation is avoided.
- ▶ The calculation of additional effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure for given prerequisites. If possible, the CPCAT approach is preferred.
- ▶ If there are no ecological reasons for not using the data at sample level, the evaluation and interpretation of the data at plot level (pooled samples of 1 m² in total used as replicates) and sub-plot level (single samples as replicates of 0.25 m²) should be requested.
- ▶ Principal response curves are generally applicable within the ECx-design and a powerful tool for community analyses. They should be carried out in addition to uni-variate methods when appropriate data are available, i.e. for tests with multiple treatments (e.g. ECx design).

Some limitations and open questions regarding the proposed changes need to be kept in mind:

- ▶ There are two opposing trends whose benefits and downsides for the significance of the test have to be balanced: On the one hand, as many concentration levels as possible should be considered for a meaningful ECx design (with no replication of concentration levels required) while on the other hand a strong design for calculating robust NOEC values requires a substantial increase in the number of replicates per control and each treatment. This question is not a strictly statistical one, but it is also related to the feasibility in the field (plot number and field size) and of the regulatory prioritization of statistical endpoints;
- ▶ The results for the implementation of an ECx design in field studies are based on a proof-of-concept pilot field study at one site and with the well-known reference substance carbendazim. For new chemicals, the choice of concentration ranges might be considerably more difficult;
- ▶ There is still no established methodology for the calculation of test power and corresponding sample planning for CPCAT;
- ▶ The CPCAT procedure is not suitable for metric data because the Poisson distribution does not adequately describe this type of data. To improve the statistical test procedures for metric data, it might be considered to integrate the closure principle into multiple t-test procedures to prevent alpha inflation;
- ▶ The use of samples as replicates for the calculation of NOEC values leads to an improvement of the test power. A general investigation of the effects in earthworm field tests at both plot and sample (= subplot) level could therefore be recommended (provided that ecological conditions exist for the use of subplots as replicates). This is not necessarily a decision based on scientific principles, but a regulatory, protective decision based on the precautionary principle.

According to the experiences made in the more than 20 past years of performing earthworm field studies based on the existing BBA and ISO guidelines and during the project, a draft OECD test guideline (TG) was formulated and provided to the ad hoc SETAC GSIG sub-group for discussion. As of now, the discussion of the draft TG is ongoing.

Zusammenfassung

Seit 1994 wird das Risiko von Chemikalien für Regenwürmer im Freiland durch einen Test bewertet, der ursprünglich von der Biologischen Bundesanstalt für Land- und Forstwirtschaft (BBA) standardisiert wurde. Seit 1999 steht eine von der ISO standardisierte internationale Richtlinie zur Verfügung (ISO 11268-3), die seitdem (zuletzt 2014) mehrmals aktualisiert wurde, ohne den grundlegenden Ansatz zu ändern. ISO-Richtlinien konzentrieren sich jedoch auf die Bewertung (potenziell) kontaminierter Umweltkompartimente (Gewässer, Sedimente, Abfallstoffe sowie Böden), d. h. sie werden in einem retrospektiven Ansatz zur Bewertung des Umwelttrisikos verwendet. Im Gegensatz dazu dienen OECD-Prüfrichtlinien der prospektiven Bewertung einzelner Chemikalien und definierter chemischer Gemische wie Pestizidformulierungen. Daher wurden die ISO-Richtlinien für die Prüfung von Chemikalien in den letzten 10 Jahren in das OECD-Format übertragen. Im Zuge dieser Umstellung, die im Falle des Regenwurmfreilandtests seit April 2013 unter deutscher Leitung als OECD-Projekt Nr. 2.47 („Neue Testrichtlinie zur Bestimmung der Auswirkungen auf Regenwürmer in Freilandstudien“) durchgeführt wird, wurde auch geprüft, ob neben formalen Anpassungen weitere Änderungen erforderlich sind. Diese Bewertung wurde von einer Ad-hoc-Untergruppe der „Global Soil Interest Group“ (GSIG) der „Society for Environmental Toxicology and Chemistry“ (SETAC) durchgeführt. Aufgrund der Erfahrungen der letzten 20 Jahre wurde entschieden, dass einige Aspekte der Richtlinie der wissenschaftlichen Entwicklung angepasst werden müssen. In Bezug auf das Studiendesign wird in der ISO-Richtlinie bereits die Möglichkeit erwähnt, Studien gemäß einer Dosis-Wirkungs-Anordnung durchzuführen, eine Option, die „im Vergleich mit Einzeldosis-Studien die umweltbezogene Risikobeurteilung deutlich unterstützt“ (ISO 2014). Insbesondere mussten neben technischen Details primär das Studiendesign und die statistische Auswertung der Testergebnisse optimiert werden. Vor allem die Variabilität der im Freiland erfassten Endpunkte, die statistische Signifikanz der Ergebnisse des Freilandtests und die Höhe der sicher nachweisbaren Wirkungen der getesteten Chemikalien sollten verbessert werden, da sonst von der EFSA (2017) vorgeschlagene Beurteilungskriterien nicht verwendbar wären. Um diese Probleme zu adressieren, fehlten wissenschaftlich belastbare und praktische Informationen. Die Generierung dieser Informationen war das Ziel dieses Projekts. In enger Zusammenarbeit mit der Ad-hoc SETAC GSIG Untergruppe wurden folgende Ziele im Rahmen von drei Arbeitspaketen (AP) erreicht:

- ▶ AP1: Auswertung vorhandener Daten und Entwicklung von Vorschlägen für ein optimiertes Design des Regenwurmfreilandtests: Zusammenstellung und kritische Auswertung von Informationen aus der Literatur und der Datenbank des Umweltbundesamtes (UBA) zur standardisierten Durchführung von Regenwurmfreilandstudien, um ein verbessertes Testdesign zu entwickeln;
- ▶ AP2: Experimentelle Untersuchungen und statistische Analysen: (1) Durchführung einer Pilotfreilandstudie mit einem verbesserten Testdesign. (2) Eingehende statistische Analyse der Pilotfreilandstudie in Kombination mit der vorhandenen Datenbank zur natürlichen Variabilität in Regenwurmgemeinschaften. (3) Berechnung von Wirkschwellen, Wirkkonzentrationen und Gemeinschaftsanalyse. (4) Formulierung von Designanforderungen für Regenwurmfreilandstudien und Identifizierung von Einschränkungen und offenen Fragen;
- ▶ AP3: Teilnahme am OECD-Prozess: Formulierung eines neuen Entwurfs einer OECD-Prüfrichtlinie auf der Grundlage der bestehenden ISO-Richtlinie 11268-3, jedoch gemäß den formalen Anforderungen der OECD, unter Verwendung der in der Pilotstudie gemachten Er-

fahrungen sowie der Auswertung der UBA-Datenbank. Diskussion dieses Prüfrichtlinienentwurfs innerhalb der Ad-hoc SETAC GSIG Untergruppe in einem abschließenden Projekttreffen. Die kombinierten Ergebnisse des Entwicklungs- und Diskussionsprozesses werden der OECD vorgelegt.

Auswertung vorhandener Daten und Entwicklung von Vorschlägen für ein optimiertes Design des Regenwurmfreilandtests (AP 1)

Im Rahmen der Voranalysen wurde die ISIS-Datenbank („Information System Chemical Safety“) des UBA als nützliche Quelle für die Datenanalyse von Regenwurmfreilandtests identifiziert. Die Datenbank enthielt 150 Einträge für Freilandstudien an Regenwürmern. Für statistische Untersuchungen wurden zunächst Qualitätskriterien für diese Daten definiert. Die Rohdaten zur Abundanz und Biomasse auf Probenebene (0,25 m²) wurden aus den ursprünglichen Studienberichten extrahiert. Für die weitere statistische Analyse wurde eine vereinheitlichte Datenbank entwickelt und befüllt. Anschließend wurden systematische Verfahren der deskriptiven Metadatenanalyse und anschließende statistische Berechnungen damit durchgeführt.

Regenwurm-Freilandstudienbank - Zusammenstellung und Qualitätsprüfung

Für statistische Analysen wurden nur Freilandstudien zu Regenwürmern mit den folgenden Merkmalen verwendet: Regenwürmer sollten durch eine Kombination aus chemischer Austreibung und Handauslese untersucht worden sein. Eine Verzerrung der Zusammensetzung der untersuchten Artengemeinschaft aufgrund der Verwendung der Oktettmethode wurde daher verhindert. Darüber hinaus sollten die technischen Berichte Rohdaten enthalten, die auf der Ebene der Einzelprobe (= Teilparzelle) gesammelt wurden. Diese Voraussetzung ermöglichte eine Analyse der Testdaten auf Probenebene im Vergleich zur konventionellen Auswertung auf Parzellenebene. Die 21 Freilandstudien, die diese Bedingungen erfüllten, wurden in zwei Klassen unterteilt: Tests mit nur einer Behandlung und einer Referenz im Vergleich zur Kontrolle (Limittest) wurden der Klasse 1 zugeordnet, während Tests mit mehreren Behandlungsstufen als Klasse 2 kategorisiert wurden. Elf Freilandstudien wurden in Klasse 1 (Limittests) eingeteilt, zwei Freilandstudien bestanden aus zwei verschiedenen Substanzkonzentrationen und weitere acht Freilandstudien wurden mit drei Behandlungen durchgeführt (Klasse 2). Darüber hinaus wurden zusätzliche 5 Studien mit digitalisierten Rohdaten auf Einzelproben- oder Parzellenebene mit jeweils leicht unterschiedlicher Beprobungsmethodik in die Datenbank integriert. Insgesamt wurden Daten von 26 Freilandtests der ISIS-Datenbank (+ Testdaten der in diesem Projekt durchgeführten Pilotstudie) für statistische Berechnungen verwendet. Die verarbeiteten Freilandstudien wurden gemäß der ISO-Richtlinie 11268-3 oder der BBA-Richtlinie Teil VI, 2-3 durchgeführt. Daher folgten die analysierten Testverfahren einem gemeinsamen Ansatz. Alle Berichte enthielten Informationen zu Regenwurmart, -zahlen und -biomasse, die auf Probenahmeflächen gesammelt wurden, die mit einer Testsubstanz in einer zufälligen Anordnung behandelt wurden (vier Replikate pro Behandlung) und mit den Daten verglichen wurden, die aus Kontroll- und Referenzflächen stammten. Jedes Replikat (= Parzelle) bestand aus vier aggregierten Proben (= Teilparzellen) von 0,25 m² pro Probe (insgesamt 1 m² Probenfläche). Die Probenahmedaten lagen normalerweise bei etwa 1-3 Monaten, 4-6 Monaten und 12 Monaten nach der Applikation. Die Tests begannen üblicherweise im April oder Mai. Die Berechnung von Effekten innerhalb der Testverfahren beschränkte sich hauptsächlich auf die Auswertung der Gesamthäufigkeit und der Biomasse auf Artenebene und für alle Regenwürmer. Juvenile Regenwürmer wurden zusammengefasst und auf Gattungsniveau ausgewertet (morphologische Gruppen: tanylob und epilob). Zusätzlich wurden die ökologischen Gruppen der endogäischen, epigäischen und anözischen Regenwürmer unterschieden.

Datenerfassung: Umwelt- und Agrarvariablen

Beschreibende Metadaten der Freilandstudien zeigten, dass die Zusammensetzung der Arten in allen Freilandstudien aus 6 bis 14 Arten pro Studie bestand. Der jeweilige Shannon Diversitätsindex lag zwischen 0,3 und 1,6 (Mittelwert: 1,2). Der Diversitätsindex war an Grünlandstandorten etwas höher (Mittelwert: 1,44) als in anderen Landnutzungstypen (unbedeckter Boden: 1,27; Ackerstandorte: 1,05). Die Artenzahl im Grünland betrug mindestens 10. Die durchschnittliche Anzahl der untersuchten Individuen betrug etwa 372 pro m² auf Grünland, 356 auf unbedecktem Boden und etwa 196 auf Ackerflächen. Die Artenzusammensetzung der Regenwürmer innerhalb der Freilandtests wurde analysiert und unter Verwendung einer Korrespondenzanalyse für die Abundanzdaten der Arten aller Datensätze miteinander verglichen. Leider war die Datenbasis nicht ausreichend, um die mögliche systematische Auswirkung der Umweltbedingungen und der jeweiligen Landnutzungsformen auf die Gemeinschaft zu untersuchen.

Freilandstudien: Artenzusammensetzung, Variabilität und MDDs

Basierend auf der Vorverarbeitung der ISIS-Datenbank wurden anschließend Daten von Freilandstudien für Regenwurmgemeinschaften analysiert. Die in die Stichprobe einbezogenen Individuen der 21 Freilandstudien gehörten 17 verschiedenen Arten an. Als statistische Maßzahl wurde der minimale nachweisbare Unterschied (% MDD, *minimum detectable difference*) zwischen Kontrolle und Behandlung aller Freilandstudien berechnet. Obwohl der wahrscheinlichste Wert des MDDs für Abundanzdaten (Modus der Wahrscheinlichkeitsdichtefunktion) der Regenwürmer in der Datenbank 45% betrug, lag die Wahrscheinlichkeit, einen MDD zu erhalten, der kleiner als 50% der Kontrolle war, bei 42%. Ein MDD zwischen 10% und 35% (in der EFSA „Soil Opinion“ (2017) als geringer Effekt definiert) wurde mit einer Wahrscheinlichkeit von 8% beobachtet. Dieselben Berechnungen für die Gesamtbiomasse ergaben noch niedrigere Wahrscheinlichkeitswerte als für die Gesamthäufigkeit: Ein MDD von weniger als 50% wurde nur für 32% aller Probenahmezeitpunkte festgestellt. Für die aggregierte Gruppe der Gesamtabundanz der Regenwürmer wurden die niedrigsten MDDs berechnet. Bei der dominantesten Art in der Datenbank, *Aporrectodea caliginosa*, war die Möglichkeit, statistisch signifikante Effekte in den Freilandstudien festzustellen, noch geringer. Die Wahrscheinlichkeit, MDDs von weniger als 50% für Endpunkte der Art *A. caliginosa* zu erhalten, war sehr gering (12% aller Probenahmezeitpunkte in der Datenbank). Der wahrscheinlichste Wert der berechneten Wahrscheinlichkeitsverteilung für MDDs der Abundanz von *A. caliginosa* betrug 66%. Auch hier wurden noch höhere MDDs für den Endpunkt Biomasse berechnet. Insgesamt zeigten Best-Practice-Studien (unter Verwendung einer Kombination aus Handauslese und chemischer Austreibung für die Regenwurmprobenahme) eine geringe Trennschärfe, um Unterschiede zwischen Kontroll- und Behandlungspartikeln für aggregierte Taxa festzustellen. Aus statistischen Gründen waren daher die Erprobung und Anpassung eines neuen Freilandstudientestdesigns im Rahmen dieses Projekts gerechtfertigt. Die Einschränkungen des alten Designs, das sowohl Limittests als auch NOEC-Ansätze abdeckte, wurden deutlich. Daher sollte ein angepasstes Testdesign eine Option zur Durchführung von Regressionsansätzen als Alternative zum NOEC-Ansatz enthalten.

Entwicklung eines Testdesigns für die Pilotstudie

In einer gemeinsamen Diskussion zwischen dem UBA und dem Projektkonsortium führten die Ergebnisse der oben beschriebenen Auswertungen zu einem ersten Vorschlag für das Design der Regenwurm-Pilotfreilandstudie, die 2017 durchgeführt werden sollte. Dieses Design war durch die Kombination eines sogenannten NOEC- mit einem ECx-Design gekennzeichnet und wurde "Mixed Omni-Design" genannt:

- ▶ Vier Probenahmeterminen bei einer Gesamttestdauer von einem Jahr (wie in der ISO-Richtlinie 11268-3);
- ▶ Eine Kontrolle (C) und sechs Testchemikalienbehandlungen (T) (nur Limittest in der ISO-Richtlinie);
- ▶ Anzahl der Parzellen pro Behandlung sechs (C, T2, T5) oder drei (T1, T3, T4, T6) (vier in der ISO-Richtlinie);
- ▶ Fünf Proben pro Parzelle (vier in der ISO-Richtlinie).

Die Durchführung einer solchen Studie bedeutete, dass insgesamt 30 Parzellen mit 150 Proben pro Probenahmedatum abgedeckt werden mussten. Dieser ursprüngliche Vorschlag wurde vom Projektteam als groß, aber hinsichtlich der Handhabung als immer noch praktikabel angesehen (z. B. hinsichtlich der Anzahl der für die Probenahme benötigten Tage, Feldgröße usw.).

Der Vorschlag des Testdesigns für die Pilotstudie wurde auf dem Treffen der Ad-hoc SETAC GSIG Untergruppe im Februar 2017 erörtert. Weitere aktuelle Beiträge zu verschiedenen Aspekten der Planung, Durchführung oder Auswertung von Regenwurmfreilandstudien wurden der Gruppe vorgestellt. In der folgenden Diskussion während des Treffens wurden verschiedene Änderungen am „Mixed Omni-Design“ vorgeschlagen, alle mit der Absicht, die Qualität der Studienergebnisse zu verbessern, ohne jedoch gleichzeitig den Aufwand stark zu erhöhen. Das resultierende endgültige Testdesign wurde als „Balanced Design“ bezeichnet. Es wurde beschlossen, sechs Proben pro Parzelle sowohl in den NOEC- als auch den ECx-Parzellen zu entnehmen. Die Anzahl der Replikate der NOEC- und ECx-Parzellen betrug sechs bzw. drei pro Behandlung.

Die ausgewählte Testchemikalie war Carbendazim, da es aufgrund seiner Verwendung als Referenzsubstanz in Regenwurmlabor- und -freilandtests bei weitem das am besten untersuchte Pflanzenschutzmittelwirkstoff in der Bodenökotoxikologie ist. Unter Verwendung der verfügbaren Informationen wurden verschiedene Carbendazim-Konzentrationsbereiche diskutiert. Die folgenden sechs Aufwandmengen (plus eine Negativkontrolle, d. h. nur Wasser) wurden schließlich ausgewählt, um einen Bereich abzudecken, der von Konzentrationen, bei denen keine Auswirkungen zu erwarten sind, bis zu Konzentrationen reicht, bei denen starke Auswirkungen wahrscheinlich sind: 0,6, 1,8, 3,2, 5,8, 10,5, und 31,5 kg Carbendazim/ha. In der derzeit verwendeten ISO-Richtlinie 11268-3 sollte die Referenzsubstanz Carbendazim einen statistisch signifikanten Unterschied von mindestens 50% in Bezug auf die Gesamtabundanz und/oder -biomasse im Vergleich zur Kontrolle an mindestens einem Probenahmezeitpunkt hervorrufen, wenn sie in Raten von 6 bis 10 kg Carbendazim/ha angewendet wird. Daher sollten solche Effekte bei den drei höchsten Aufwandmengen nachweisbar sein. Dementsprechend und unter Bezugnahme auf die Erfahrungen, die in einem EU-Projekt gemacht wurden (das sich auf die Entwicklung einer Standard-Halbfreilandmethode konzentrierte, bei der terrestrische Modellökosysteme (TME) eingesetzt wurden), sollten bei den beiden niedrigeren Raten keine nachweisbaren Effekte auftreten. A-priori-Analysen haben gezeigt, dass eine EC₅₀ bei Raten um 2,5 kg Carbendazim/ha zu erwarten ist.

Experimentelle Untersuchungen und statistische Analysen (AP 2)

Durchführung der Pilotfreilandstudie

Für die Testdurchführung wurde ein Ackerlandstandort ausgewählt. Es war von landwirtschaftlichen Feldern und Wegen umgeben. Die Versuchsparzellen wurden auf einer Fläche von ca. 55 m x 107 m installiert. Vor Beginn der Studie wurde auf dem Feld Winterweizen angebaut. Um die Versuchsfläche ohne Bodenbearbeitung, die sich auf die Regenwurmgemeinschaft ausge-

wirkt hätte, von Vegetation zu befreien, wurde Glyphosat in einer Rate von 1,8 kg a.s./ha angewendet. Für jede Behandlung, d. h. Kontrolle (C) und sechs verschiedene Testchemikalienbehandlungen (T1 bis T6), wurden sechs (C, T2, T5) oder drei (T1, T3, T4, T6) Parzellen (= Replikate), jede 10 m x 10 m, am Versuchsstandort installiert und zufällig auf die Behandlungen verteilt. Der Abstand zwischen zwei benachbarten Parzellen betrug 3 m und zu den umliegenden Feldern oder Feldwegen mindestens 5 m. Die Testchemikalie wurde am 11. April 2017 einmal als suspensierbares Konzentrat (SC; Formulierung Carbomax 500 SC) appliziert. Das Wasser (Kontrolle) und die Testchemikalie wurden bei einer Windgeschwindigkeit unter 3 m/s auf die Bodenoberfläche aufgetragen, um jegliches Risiko einer Kreuzkontamination aufgrund möglicher Drift während der Applikation zu vermeiden. Alle Versuchspartzellen wurden direkt nach der Applikation mit einem vom Traktor gezogenen Tankwagen mit mindestens 1000 l pro Parzelle (entsprechend 10 mm Niederschlag) bewässert. Die Versuchspartzellen wurden der natürlichen Entwicklung der Vegetation überlassen. Es wurden keine landwirtschaftlichen Praktiken wie Bodenbearbeitung oder Applikation von Pflanzenschutzmitteln oder Düngemitteln durchgeführt. Am 25. August 2017 wurden alle Parzellen mit einem Fadenschneider gemäht und aller Verschnitt auf den Parzellen belassen.

Acht bis sechs Tage vor der Applikation der Testchemikalie wurden auf allen Parzellen Regenwürmer beprobt. Die mittlere Gesamtabundanz und die mittlere Biomasse von Regenwürmern wurden für jede der dreißig Parzellen bestimmt, die entweder zur Behandlung mit der Testchemikalie oder als unbehandelte Kontrolle vorgesehen waren. Die mittlere Anzahl der vor der Applikation gesammelten Regenwürmer (Handauslese und AITC-Austreibung) lag zwischen 413 und 512 Ind./m² und erfüllte damit die Anforderungen der ISO-Richtlinie 11268-3. Regenwürmer wurden zu jedem Probenahmezeitpunkt durch eine kombinierte Handauslese und Allylthiocyanat (AITC)-Austreibungsmethode beprobt. Pro Parzelle wurden sechs zufällig verteilte Einzelproben mit einer Fläche von 0,25 m² (50 cm x 50 cm) bis zu einer Tiefe von ca. 20 cm entnommen. Daher gab es 18 (3 Parzellenreplikate) oder 36 (6 Parzellenreplikate) Einzelproben pro Behandlung und Probenahmezeitpunkt. Der Abstand zwischen zwei am selben Datum und in derselben Parzelle entnommenen Proben betrug mindestens 2 m. Die Probenahmestelle wurde markiert und an späteren Probenahmeterminen nicht mehr verwendet. Die Proben wurden mindestens 2 m vom Parzellenrand entfernt entnommen. Fünf bis zehn Liter einer AITC-Lösung (0,1 g/l) wurden gleichmäßig in den verbleibenden Hohlraum gegossen, um Regenwürmer aus tieferen Bodenschichten auszutreiben. Der Boden wurde sorgfältig durch Handauslese nach Regenwürmern durchsucht. Diese und die durch AITC extrahierten Würmer wurden in einer 70%igen Ethanollösung in wasserdichten Behältern aufbewahrt.

Die Würmer wurden unter einem Binokularmikroskop unter Verwendung externer Merkmale identifiziert. Adulte Würmer wurden auf Artenebene bestimmt. Juvenile wurden auf Gattungsebene klassifiziert, aber in einigen Fällen war eine Unterscheidung von kleinen Würmern, die zu eng verwandten Gattungen gehörten, nicht möglich (z. B. wurden *Allolobophora* und *Aporrectodea* zusammengefasst). Alle adulten Würmer einer Probe einer bestimmten Art und alle juvenilen Würmer einer bestimmten Gattung wurden zusammen gewogen. Das Feld war von einer Regenwurmgemeinschaft besiedelt, die als typisch für mitteleuropäisches Ackerland angesehen werden kann (ISO 11268-3), einschließlich der ökologisch wichtigsten Gruppen anektischer und endogäischer Regenwürmer. Insgesamt wurden während der Studie neun verschiedene Arten von Regenwürmern gefunden. Die Lumbricidenbiozönose wurde von Juvenilen der endogäischen Gattungen *Aporrectodea*/*Allolobophora* dominiert, *Allolobophora chlorotica* war die am häufigsten vorkommende Art.

Die Testchemikalie Carbomax 500 SC (a.s. Carbendazim) verursachte zu allen drei Zeitpunkten der Probenahme nach der Applikation eine deutliche Verringerung der Gesamtabundanz und -

biomasse. Im Vergleich zur Kontrolle betragen die mittlere Abundanz bzw. Biomasse in den mit Testchemikalien behandelten Parzellen 34-56% bzw. 11-55% (34-36 Tage nach der Applikation; DAA), 45-90% bzw. 69-111% (188-190 DAA) und 38-74% bzw. 80-113% (377-379 DAA).

Statistische Analyse: Freilandstudie und Datenbank

Eine Reihe verschiedener statistischer Datenanalyseverfahren wurde sowohl für Daten der Pilotstudie als auch für vorhandene Testdaten aus der UBA-Datenbank angewandt. Das Hauptaugenmerk lag auf der Verbesserung der konventionellen statistischen Methoden zur Auswertung von Regenwurmfreilandstudien (ISO 11268-3) und auf der Gewinnung von Erkenntnissen für statistische Überlegungen hinsichtlich eines angepassten Testdesigns für Regenwurmfreilandstudien. Rohdaten für Biomasse und Abundanz zu allen Probenahmeterminen sowie für alle Taxa und morphologischen oder funktionellen Regenwurmgruppen wurden auf Einzelproben- und Parzellenebene in die bestehende Datenbank des Projekts integriert. Berechnungen und Analysen einzelner Arten zeigten zum Zeitpunkt der letzten Probenahme (377-379 DAA) keine statistisch signifikanten Effekte. Für „*Aporrectodea/Allolobophora* spp. juvenile“ wurde ein statistisch signifikanter Effekt nach einem Jahr beobachtet. Aufgrund der hohen Dominanz dieser Gruppe im Gesamtdatensatz wurde auch in anderen aggregierten Gruppen eine Verringerung der Abundanz und Biomasse nach 12 Monaten angezeigt. Dies wurde durch Jungtiere von *Aporrectodea/Allolobophora* spp. verursacht. Dieses Beispiel zeigt die Notwendigkeit, verschiedene Arten von Endpunkten und Regenwurmgruppen (z. B. Arten- und Gruppenebene) zu bewerten, um allgemeine Schlussfolgerungen für die Auswirkungen von Testsubstanzen auf der Grundlage eines aggregierten Endpunkts wie der Gesamtabundanz aller Regenwürmer zu vermeiden.

Die natürliche, heterogene Streuung von Regenwurmartarten innerhalb eines Feldes ist ein entscheidender Faktor für die statistische Sichtbarkeit möglicher Auswirkungen von angewandten Chemikalien. Die Variabilität der getesteten Endpunkte in Datenbank- und Pilotfreilandstudien wurde unter Verwendung des Variationskoeffizienten (CV) von Freilandstudien-Kontrollbehandlungen ausgewertet, um Schlussfolgerungen und Verbesserungsvorschläge hinsichtlich der Testtrennschärfe abzuleiten. Die natürliche Variabilität der Artengruppen in Freilandstudien wurde deskriptiv als Varianz in den Kontrollbehandlungen dargestellt und als Grundlage für die multiple Stichprobenplanung verwendet. Aggregierte Regenwurmgruppen hatten die niedrigsten CVs, während seltene Arten eine vergleichsweise hohe relative Streuung zwischen den Parzellen zeigten. Die Ergebnisse zeigten, dass besonders aggregierte Artengruppen mit hohen Abundanz- und Biomassewerten statistisch starke Endpunkte bieten (insbesondere geringe Variation bei Kontrollen und Behandlungen). Im Durchschnitt schien die Streuung auf der Ebene der einzelnen Arten oft zu hoch zu sein, um statistisch signifikante Effekte nachzuweisen. Eine starke Variation der Kontrollbehandlungen führt somit zu einer geringeren Nachweisbarkeit möglicher Wirkungen der Testsubstanz (= hoher MDD).

Der Einfluss der Varianz auf die Anzahl der erforderlichen Replikate, um eine bestimmte Testtrennschärfe zu erreichen, wurde für den standardisierten Dunnett-Test bestimmt. Die Berechnungen basierten auf den CVs für Kontrollbehandlungen in Regenwurmfreilandtests und wurden für eine dynamische Probengrößenplanung für einen MDD angewendet, der jeweils erreicht werden sollte. Für die Entwicklung eines angepassten Testdesigns wurde untersucht, wie viele Proben (= Replikate) bei einer gewünschten Zieltesttrennschärfe und einer bestimmten natürlichen Variabilität der Daten verwendet werden mussten. Standardmäßig ist die gewünschte Testtrennschärfe für statistische Hypothesentests auf 80% eingestellt. Der MDD, der mit den jeweiligen Probengrößen erreicht werden kann, wurde in der Simulation in vier verschiedene Klassen eingeteilt, angepasst an die Skalierung der Größenordnung der Effekte in der EFSA „Soil

Opinion². Bis zu 10% Unterschied zwischen Kontrolle und Behandlung wurden als vernachlässigbare Effekte definiert, bis zu 35% als kleine Effekte, bis zu 65% als mittlere Effekte und ab 65% als große Effekte. Obwohl die Angaben der EFSA (EFSA PPR 2017) sich auf die Effekte auf Schutzgüter beziehen und nicht zwangsweise in Feldstudien detektiert werden müssen, wurden die Größenordnungen der Effekte analysiert. Die Ergebnisse der Stichprobengrößensimulation für die mittlere Gesamtvariabilität von Regenwürmern zeigten, dass standardisierte Regenwurmfreilandtests möglicherweise nicht genügend Replikate aufweisen, um kleine Effekte mit einer Testtrennschärfe von 80% für die Gesamtsumme aller Regenwürmer zu erkennen. Dies war die Gruppe mit den niedrigsten CVs in Regenwurmfreilandtestdaten. Dementsprechend ist die Möglichkeit, Effekte für andere Regenwurmgruppen zu erkennen, noch geringer.

Aus diesem Grund wurde untersucht, ob eine NOEC-Berechnung unter Verwendung von Einzelproben (= Teilparzellen) als statistische Replikate zu einer Verbesserung hinsichtlich der MDDs führen würde. Die Planung der Stichprobengröße wurde mit den gemessenen CVs auf Parzellenebene (aktuelle Standardmethode) und auf Stichprobenebene berechnet, um Verschiebungen der Testtrennschärfe zu bewerten. Die Ergebnisse der Pilotstudie zeigten, dass die Erhöhung der Parzellenzahlen ($n = 6$) und die geringfügig niedrigeren CVs ($1,5 \text{ m}^2$ anstelle von $1,0 \text{ m}^2$ Fläche beprobt) die Testtrennschärfe erhöhten. Die Anzahl der erforderlichen Replikate, um einen bestimmten Schwellenwert für MDDs zu erreichen, verringerte sich im Vergleich zu den Freilandstudien in der Datenbank. Trotzdem konnten mit diesem Testdesign mittlere Effekte (35% - 65% Effekt) mit einer Trennschärfe von 80% nicht erkannt werden. Die umfassende Detektion kleiner Effekte (10% - 35%) mit einer Testtrennschärfe von 80% erschien in dieser Simulation unter Berücksichtigung einer realistischen Anzahl von Replikaten unmöglich. Trotzdem wäre in einem dargestellten Fallbeispiel für die Gesamtabundanzzahlen der beprobten Regenwürmer in der Pilotstudie ein Unterschied von 35% auf Einzelprobenebene (= Teilparzellen) erkennbar gewesen. In diesem Fall war der mittlere CV auf Teilparzellenebene geringfügig höher als auf Parzellenebene (34,56%). Aus diesem Grund wären mindestens 14 Replikate erforderlich gewesen, um mittlere Effekte nachzuweisen. Bei Verwendung der Einzelproben als Replikate waren in diesem statistischen Design jedoch 36 Replikate verfügbar. Dieser Wechsel in der Auswertungsebene ermöglichte die Identifizierung von mehr statistisch signifikanten Unterschieden, insbesondere auf der Ebene einzelner Arten. Im Allgemeinen verbesserte sich die statistische Erkennbarkeit von Effekten, wenn die Auswertung auf der Ebene der Teilparzellen erfolgte.

Die Berechnung der Effektschwellen wurde für alle Studien in der Datenbank und zu allen Probenahmezeitpunkten durchgeführt und ein Vergleich der Dunnett-Methode mit den Ergebnissen des sogenannten CPCAT-Ansatzes (Closure Principle Computational Approach test) durchgeführt. Die theoretische Verteilungsannahme der Freilandtestdaten für die Abundanz von Regenwürmern folgt einem Poisson-Modell. Daher wird die Anwendung des CPCAT-Ansatzes für Daten zur Abundanz aufgrund trennschärferer Teststatistiken dringend empfohlen. Dies war das erste Mal, dass die Teststärke von CPCAT im Rahmen einer umfassenden Metaanalyse von Freilandstudien bewertet wurde. Es wurde gezeigt, dass die Verwendung des CPCAT-Verfahrens im Vergleich zum Dunnett-Test die Wahrscheinlichkeit erhöhte, dass signifikante Effekte identifiziert wurden, selbst bei kleinen Effekten (10% - 35%). CPCAT ist daher im Allgemeinen statistisch weniger konservativ, d. h. signifikante Wirkungen der Testsubstanz werden bereits für kleinere Unterschiede zur Kontrolle angezeigt. Die Unterschiede in der Testtrennschärfe zwischen den beiden Verfahren wurden auch bei der getrennten Untersuchung der NOEC-Berechnungen für verschiedene Artengruppen sichtbar. Die im CPCAT-Verfahren ver-

² EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues) (2017): Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. EFSA Journal 15(2):4690, 225 pp. doi: 10.2903/j.efsa.2017.4690

wendete Poisson-Verteilung beschreibt im Bereich oft vorkommender, geringer Abundanzwerte die Daten der Regenwurmgemeinschaft in Freilandtests mathematisch und statistisch genauer als die in herkömmlichen t-Tests (z. B. Dunnett oder Williams) verwendete Normalverteilung. Somit erhöht die Verwendung des CPCAT-Ansatzes die Testtrennschärfe für Regenwurmfreilanddaten. Für den relativ neuen CPCAT-Ansatz gibt es derzeit jedoch kein Verfahren zur generischen Berechnung eines quantitativen Maßes für die Testtrennschärfe, eine Stichprobenplanung mit dem CPCAT-Verfahren ist noch nicht möglich.

Anders als die Datenbankstudien wurde die Pilotstudie in einem Testdesign mit mehreren Konzentrationsstufen durchgeführt. Probitkurven-Regressionen wurden durchgeführt, und in allen drei Probenahmen nach der Applikation wurde eine signifikante Dosis-Wirkungs-Beziehung in der Gruppe "Gesamtregenwürmer" identifiziert. Im Gegensatz zum NOEC-Ansatz (Dunnett-Test) ermöglichte die Wahl eines ECx-Designs die Erkennung signifikanter Beziehungen zwischen der angewendeten Carbendazim-Konzentration und dem gemessenen Effekt auf die Regenwurmpopulation (Gesamtabundanz in der Pilotstudie) über den gesamten Konzentrationsbereich. Darüber hinaus zeigte der Vergleich der EC₅₀-Werte über die Probenahmezeitpunkte hinweg Erholungseffekte, die im Fall der gesamten Regenwurmgemeinschaft zwischen 1 und 6 Monaten nach der Applikation angenommen werden konnten. Die berechnete EC₅₀ stieg in diesem Zeitraum um den Faktor 10 an, während sie sich im weiteren Zeitraum bis 12 Monate nach der Applikation nicht änderte. Im Gegensatz dazu stieg die EC₅₀ für Jungtiere (*Aporrectodea/Allolobophora* spp.) bis zum Ende der Studie nur um den Faktor 3 an. Die Ergebnisse der Studie zeigten, dass die Verwendung eines ECx-Designs zur Ableitung von Effektkonzentrationen auf die Regenwurmpopulation im Freiland im Allgemeinen möglich ist. Die Wahl eines geeigneten Konzentrationsbereichs für eine angemessene Prüfung aller Arten und aggregierten Gruppen ist jedoch eine Herausforderung.

Eine Principal Response Curve (PRC) wurde verwendet, um die Frage zu beantworten, ob ein signifikanter Zusammenhang zwischen der Struktur der Regenwurmlebensgemeinschaft und der Behandlung bestand. Die PRCs zeigten einen hoch signifikanten Effekt der Behandlung auf die Regenwurmgemeinschaft (p-Wert <0,05). Eine klare Dosis-Wirkungs-Beziehung war sichtbar, und mit zunehmender Konzentration nahm die Abweichung von der Kontrolle zu. Gemäß der PRC könnte nach etwa einem Jahr von einer Erholung der Gemeinschaft (Abundanzen von Adulten einzelner Arten, die den Ausgangszustand wiedererlangen) ausgegangen werden.

Basierend auf diesen Untersuchungen sind generische Ableitungen von Empfehlungen aufgrund der hohen Variabilität in den verschiedenen Regenwurmdatensätzen verschiedener Freilandtests und aufgrund der zu erwartenden Auswirkungen der örtlichen Standortbedingungen begrenzt. Die folgenden grundlegenden Empfehlungen und Anforderungen in Bezug auf die Durchführung und Auswertung von Regenwurm-Freilandtests konnten jedoch identifiziert werden:

1. Es besteht weiterhin die Notwendigkeit, Biomasse und Abundanz auf Artenebene zu bestimmen und zu auswerten, da die verwendeten aggregierten morphologischen oder funktionellen Gruppen Auswirkungen auf einzelne Arten verschleiern können.
2. Das ECx-Design ist eine sinnvolle Alternative zum NOEC-Design im Regenwurm-Freilandtest. Zumindest ein gemischtes Design wäre ratsam. Tatsächlich führt das ECx-Design zu stärkeren/protektiveren Aussagen für die Umweltrisikobewertung gerade in niedrigen Wirkungsbereichen, eine Verschleierung möglicher Effekte wie bei der NOEC-Auswertung wird vermieden.
3. Die Berechnung der Effektschwellen (NOEC/LOEC) sollte mit dem trennschärfsten Mehrfachtestverfahren für die gegebenen Voraussetzungen durchgeführt werden. Wenn möglich, wird der CPCAT-Ansatz bevorzugt. Wenn die Daten metrisch sind (z. B. Biomasse), sollten multiple t-Testverfahren wie der Dunnett- oder Williams-Test ($\alpha = 0,05$, zweiseitig für eine

unklare Richtung des Effekts) für mehrere Vergleiche in einem randomisierten Parzellendesign durchgeführt werden. Die Voraussetzungen für normalverteilte Daten und Varianzhomogenität müssen unter Verwendung von z.B. Shapiro-Wilk- bzw. Levene-Test geprüft werden. Wenn Daten das Kriterium der Normalverteilung nicht erfüllen, können verallgemeinerte lineare Modelle oder nichtparametrische Tests, z. B. der Bonferroni U-Test oder der Jonckheere-Terpstra Step-Down-Test (Varianzhomogenität erforderlich) angewendet werden. Die theoretische Verteilungsannahme der Freilandtestdaten für die Abundanz von Regenwürmern folgt einem Poisson-Modell. Daher wird die Anwendung des CPCAT-Ansatzes aufgrund trennschärferer Teststatistiken für Zählraten zur Abundanz dringend empfohlen. Wenn jedoch Abundanzdaten eine Varianzhomogenität zeigen, die Nullhypothese der Normalverteilung nicht verworfen wird und die absoluten Abundanzen pro Replikate >5 sind, ist auch die Anwendung parametrischer Testverfahren (Williams, Dunnett) möglich. Bei multiplen t-Testverfahren und bei ungleicher Replikation müssen die Tabellen-t-Werte wie von Dunnett und Williams vorgeschlagen korrigiert werden. Darüber hinaus sollte eine unangemessene log-Transformation von Daten während des Berechnungsvorgangs vermieden werden.

4. Nach der Datenrevision sollte entschieden werden, ob eine einfache Probit-Regression (Logit, Weibull) mit zwei Parametern, eine nichtlineare Regression oder die Integration eines sogenannten Hormesmodells zur Berechnung der Effektkonzentrationen (EC_x) erforderlich ist. Auch bei einer monotonen Erhöhung des gemessenen Endpunktes (Biomasse, Abundanz) sollte die Ableitung signifikanter Effektkonzentrationen berücksichtigt werden.
5. Wenn es keine ökologischen Gründe gibt, die Daten nicht auf Einzelprobenebene zu verwenden (d. h. keine nachgewiesene Interdependenz zwischen Proben aus derselben Parzelle), sollte die Auswertung und Interpretation der Daten auf Parzellenebene (gepoolte Proben von insgesamt 1 m² als Replikate) und Teilparzellenebene (Einzelproben als Replikate von 0,25 m²) gefordert werden.
6. PRC sind im Allgemeinen innerhalb des EC_x-Designs anwendbar und ein leistungsfähiges Werkzeug für Gemeinschaftsanalysen. Sie sollten zusätzlich zu univariaten Methoden für Tests mit mehreren Behandlungen (z. B. EC_x-Design) durchgeführt werden, wenn geeignete Daten verfügbar sind.

Einige Einschränkungen und offene Fragen zu den vorgeschlagenen Änderungen müssen berücksichtigt werden. Die Empfehlungen zur Anpassung des Testdesigns der Freilandstudie zeigen zwei gegensätzliche Trends auf, deren Vor- und Nachteile für die Aussagekraft des Tests abgewogen werden müssen. Zum einen sollten für ein aussagekräftiges EC_x-Design möglichst viele Testkonzentrationsstufen berücksichtigt werden. Aus rein statistischer Sicht ist für die nachfolgende Regressionsanalyse keine Replikation der Konzentrationsstufen erforderlich. Ein starkes Design zur Berechnung robuster NOEC-Werte erfordert zum anderen, wie gezeigt, eine erhebliche Erhöhung der Anzahl der Replikate pro Kontrolle und Behandlung. Diese beiden Anforderungen müssen abhängig vom zugrundeliegenden Testkonzept und den gewünschten Endpunkten abgewogen und in ein neues Design integriert werden. Diese Entscheidung ist jedoch nicht streng statistischer Natur, sondern in erster Linie eine Frage der Machbarkeit vor Ort (zu behandelnde Parzellenzahlen und Feldgrößen) und der regulatorischen Priorisierung verschiedener Endpunkte.

Darüber hinaus weisen die oben dargestellten Analysen und zugrunde liegenden Daten einige Einschränkungen auf, die nicht vergessen werden sollten. Die Ergebnisse für die Implementierung eines EC_x-Designs in Freilandstudien basieren auf einer Machbarkeitsstudie (Pilotfreilandstudie) an einem Standort mit der gut untersuchten Referenzsubstanz Carbendazim. In diesem Fall standen fundierte Vorkenntnisse und Erfahrungen aus früheren Freilandstudien zu möglichen Effektbreiten und -dynamiken zur Verfügung. Dies ist insbesondere bei neuen Stoffen in

der Regulierungspraxis nicht der Fall. In solchen Fällen kann die Wahl der Konzentrationsbereiche in Regenwurmfreilandtests erheblich schwieriger sein. Darüber hinaus zeigt die Pilotfreilandstudie, dass ein angewandter Konzentrationsbereich normalerweise aufgrund ihrer unterschiedlichen Empfindlichkeit nicht für alle Regenwurmart und -gruppen ableitbare Dosis-Wirkungs-Kurven liefert. Dieses Problem trat auch bereits in den zuvor verwendeten NOEC-Designs aufgrund der unterschiedlichen Empfindlichkeiten der Spezies auf. Der statistische Endpunkt der NOEC verschleierte dies bisher jedoch weitgehend.

Für die Ableitung von NOEC-Werten mit Abundanzdaten stellt CPCAT eine sinnvolle Alternative zu den standardisierten Verfahren der t-Test-Statistik dar. Es sollte jedoch erwähnt werden, dass es noch keine etablierte Methodik für die Berechnung der Testtrennschärfe und die entsprechende Stichprobenplanung für CPCAT gibt. Außerdem sollte CPCAT zunächst eine höhere Akzeptanz als geeignetes Instrument zur Auswertung der Ergebnisse ökotoxikologischer Tests erreichen, beispielsweise durch Anwendung als Standardanalysemethode in einem breiteren Spektrum von ökotoxikologischen Standardtestmethoden.

Das CPCAT-Verfahren ist nicht für metrische Daten geeignet, da die Poisson-Verteilung diese Art von Daten nicht angemessen beschreibt. Um die statistischen Testverfahren für metrische Daten zu verbessern, könnte erwogen werden, das im CPCAT-Verfahren verwendete Closure Principle auch in multiple t-Testverfahren zu integrieren, um die Korrektur des Signifikanzniveaus α für diese Tests so zu verbessern.

Die Verwendung der Einzelproben als Replikate zur Berechnung der NOEC-Werte führt zu einer Verbesserung der Testtrennschärfe. Auf der Grundlage dieser Ergebnisse wird daher eine allgemeine Untersuchung der Effekte in Regenwurmfreilandtests sowohl auf Parzellen- als auch auf Einzelprobenebene (= Teilparzelle) empfohlen, vorausgesetzt, es bestehen die ökologischen Bedingungen für die Verwendung von Teilparzellen als Replikate. Ob dies angesichts der Debatte über Pseudoreplikation in Freilandstudien eine nützliche Option ist, sollte unbedingt stärker diskutiert werden. Innerhalb eines regulatorischen Rahmens könnten die folgenden Schritte in Betracht gezogen werden: Ein entsprechender Endpunkt wird sowohl auf Teilparzellen- als auch auf Parzellenebene ausgewertet. Wenn die gleichen NOEC-Werte als Ergebnisse erhalten werden, werden diese berücksichtigt. Werden andere (niedrigere) NOEC-Werte auf Teilparzellenebene berechnet, wird das folgende Verfahren vorgeschlagen: Wenn es nicht möglich ist, zuverlässig ein Artefakt des Parzelleneffekts auf dieser Ebene nachzuweisen, sollte die niedrigere NOEC für den Regulierungsprozess verwendet werden. Dies ist nicht unbedingt eine Entscheidung, die auf rein wissenschaftlichen Erwägungen beruht, sondern eine regulatorische Entscheidung, die auf dem Vorsorgeprinzip beruht.

Ableitung eines neuen Testdesigns

Die während der Durchführung der Pilotstudie sowie in der statistischen Auswertung dieser Studie und der UBA-Datenbank gesammelten Erfahrungen wurden verwendet, um einen Vorschlag für ein neues Testdesign abzuleiten (Tabelle 1).

Tabelle 1: Anzahl der Parzellen und Behandlungen für das ECx- und das Mixed Design in Regenwurmfreilandtests. Weitere Informationen zum Designtyp im obigen Text. C = Kontrolle; T1-x = Behandlungen; R = Referenzsubstanz

Testdesign	Parzellen pro Behandlung (Anzahl)									Parzellen (Summe)	Proben (Gesamtanzahl)
	C	T1	T2	T3	T4	T5	T6	(T7)	R		
ECx Design	3	3	3	3	3	3	3	(3)	3	24 (27)	96 (108)
Mixed Design	6	2	6	2	2	6			3	27	108

Teilnahme am OECD-Prozess (AP 3)

Die in den letzten mehr als 20 Jahren gesammelten Erfahrungen mit der Durchführung von Regenwurm-Feldstudien auf der Grundlage der bestehenden BBA- und ISO-Richtlinien und während des Projekts wurden genutzt, um einen neuen Entwurf einer OECD-Prüfrichtlinie mit einem Vorschlag für ein neues Testdesign zu formulieren. Der Entwurf der OECD-Prüfrichtlinie wurde im März 2019 als Diskussionsgrundlage während des abschließenden Projekttreffens am UBA in Dessau an die Ad-hoc SETAC GSIG Untergruppe verteilt. Während und nach dem Treffen wurden mehrere Kommentare abgegeben, die gemäß dem OECD-Prozess in einer Kommentartabelle zusammengestellt wurden. Diese Tabelle wird derzeit überprüft, um eine aktualisierte Version des Prüfrichtlinienentwurfs zu erstellen, die dann dem weiteren OECD-Prozess unterzogen wird.

Schlussfolgerungen und Ausblick

Ziel dieses Projekts war es, wissenschaftlich belastbare und praktische Informationen über (1) die Variabilität der in Regenwurmfreilandstudien ausgewerteten Endpunkte, (2) die statistische Auswertung der Ergebnisse und (3) das Ausmaß der statistisch nachweisbaren Auswirkungen der getesteten Chemikalien zu generieren. Das endgültige Ziel war es, Vorschläge für ein verbessertes Testdesign zu machen. Die kritische Auswertung der in der Literatur und in der Datenbank des UBA verfügbaren Informationen ergab die folgenden Mängel des derzeit verwendeten Freilandtestdesigns für Regenwürmer gemäß ISO-Standard 11268-3:

- ▶ Die überprüften Best-Practice-Studien (d. h. unter Verwendung einer Kombination aus Handauslese und chemischer Austreibung) zeigen eine geringe statistische Aussagekraft, um Unterschiede zwischen Kontroll- und Behandlungspartzellen für aggregierte Taxa festzustellen. Für einzelne Arten ist die statistische Power für eine zuverlässige Identifizierung von Effekten noch geringer. Der Gesamt-MDD ist oft nicht niedrig genug, um kleine oder mittlere Effekte umfassend zu detektieren.
- ▶ NOEC und verwandte Konzepte werden in der ökotoxikologischen Literatur seit langem kritisiert. Darüber hinaus haben die tatsächlichen MDD-Berechnungen von Freilandstudien gezeigt, dass potenziell relevante Effekte in vielen Freilandsituationen mit den derzeit standardisierten statistischen Verfahren nicht nachweisbar sind.
- ▶ Ein angepasstes Testdesign sollte eine Option zur Durchführung von Regressionsanalysen enthalten, die als Ergänzung zum NOEC-Ansatz vorgeschlagen wurden. Die resultierenden abgeschätzten Effektkonzentrationen (ECx-Werte) aus der Anpassung einer Kurve an die Daten wurden als sinnvolle Alternative zum NOEC-Wert vorgeschlagen. Daher musste die An-

zahl der Konzentrationsstufen in der Pilotfreilandstudie erhöht werden, um die Eignung eines ECx-Designs für Regenwurmfreilandstudien zu untersuchen.

- ▶ Um weiterhin die Möglichkeit zu haben, NOEC-Werte abzuleiten und die statistische Aussagekraft dieses Verfahrens im Vergleich zum alten Design zu verbessern, muss die Anzahl der Replikate auf der Parzellenebene für die Kontroll- und Testkonzentrationsbehandlungen erhöht werden.
- ▶ Die Anzahl der Proben pro Replikat sollte erhöht werden, um die Änderungen der Varianz zu untersuchen und um abzuschätzen, ob diese Proben als einzelne Replikate zur Verbesserung der statistischen Testtrennschärfe verwendet werden können.
- ▶ Da die Freilandbedingungen und die praktische Durchführbarkeit der Pilotfreilandstudie die Gesamtzahl der Parzellen begrenzten, mussten die Erhöhung der Konzentrationsstufen und die Zunahme der Parzellen- und Probenzahl (= Teilparzellen) pro Behandlung so angepasst werden, dass beide Forschungsfragestellungen (Machbarkeit des ECx- und Verbesserung des NOEC-Designs) adressiert werden konnten.

Basierend auf diesen Auswertungen wurde eine Pilotfreilandstudie gemäß einem neu entwickelten kombinierten NOEC- und ECx-Testdesign mit der Testchemikalie Carbendazim durchgeführt. Eine Kontrolle (C) und sechs Behandlungen (T) wurden verwendet. Die Anzahl der Parzellen pro Behandlung betrug sechs (C, T2, T5) oder drei (T1, T3, T4, T6). Die Anzahl der Proben pro Parzelle betrug sechs. Die Ergebnisse der Pilotfreilandstudie und der eingehenden statistischen Auswertung zusätzlicher Regenwurmfreilandstudien ergaben die folgenden Designanforderungen für Regenwurmfreilandstudien:

- ▶ Abundanz und Biomasse sollten auch auf Artenebene bestimmt und ausgewertet werden, da aggregierte morphologische oder funktionelle Gruppen die Auswirkungen auf einzelne Arten verschleiern können.
- ▶ Das ECx-Design ist eine sinnvolle Alternative zum NOEC-Design. Zumindest ein gemischtes Design wäre ratsam. Das ECx-Design führt zu stärkeren Aussagen für die Umweltrisikobewertung, eine Verschleierung möglicher Effekte wie bei der NOEC-Auswertung wird vermieden.
- ▶ Die Berechnung zusätzlicher Effektschwellen (NOEC/LOEC) sollte mit dem trennschärfsten Mehrfachtestverfahren für die gegebenen Voraussetzungen durchgeführt werden. Wenn möglich, wird der CPCAT-Ansatz bevorzugt.
- ▶ Wenn es keine ökologischen Gründe gibt, die Daten nicht auf Einzelprobenebene zu verwenden, sollte die Auswertung und Interpretation der Daten auf Parzellenebene (gepoolte Proben von insgesamt 1 m² als Replikate) und Teilparzellenebene (Einzelproben als Replikate von 0,25 m²) gefordert werden.
- ▶ PRC sind im ECx-Design allgemein anwendbar und ein leistungsstarkes Werkzeug für Gemeinschaftsanalysen. Sie sollten zusätzlich zu univariaten Methoden durchgeführt werden, wenn geeignete Daten verfügbar sind, d. h. für Tests mit mehreren Behandlungen (z. B. ECx-Design).

Einige Einschränkungen und offene Fragen zu den vorgeschlagenen Änderungen müssen berücksichtigt werden:

- ▶ Es gibt zwei gegensätzliche Trends, deren Vor- und Nachteile für die Aussagekraft des Tests abgewogen werden müssen: Einerseits sollten so viele Konzentrationsstufen wie möglich für ein aussagekräftiges ECx-Design (in dem eine Replikation der Konzentrationsstufen nicht erforderlich ist) berücksichtigt werden, während andererseits ein starkes Design zur Berechnung robuster NOEC-Werte eine erhebliche Erhöhung der Anzahl der Replikate pro Kontrolle und jeder Behandlung erfordert. Diese Frage ist nicht streng statistischer Natur, sondern hängt auch mit der Machbarkeit im Freiland (Parzellenanzahl und Feldgröße) und der regulatorischen Priorisierung statistischer Endpunkte zusammen.
- ▶ Die Ergebnisse für die Implementierung eines ECx-Designs in Freilandstudien basieren ausschließlich auf einer Pilotfreilandstudie an einem Standort und mit der häufig verwendeten Referenzsubstanz Carbendazim. Bei neuen Chemikalien kann die Auswahl der Konzentrationsbereiche erheblich schwieriger sein.
- ▶ Es gibt noch keine etablierte Methode zur Berechnung der Testtrennschärfe (statistische Power) und der entsprechenden Stichprobenplanung für CPCAT.
- ▶ Das CPCAT-Verfahren ist nicht für metrische Daten geeignet, da die Poisson-Verteilung diese Art von Daten nicht angemessen beschreibt. Um die statistischen Testverfahren für metrische Daten zu verbessern, könnte erwogen werden, das Closure Principle in mehrere t-Testverfahren zu integrieren, um eine Alpha-Inflation zu verhindern.
- ▶ Die Verwendung von Einzelproben als Replikate zur Berechnung von NOEC-Werten führt zu einer Verbesserung der Testtrennschärfe. Eine allgemeine Untersuchung der Auswirkungen von Regenwurmfreilandtests sowohl auf Parzellen- als auch auf Probenebene (= Teilparzelle) könnte daher empfohlen werden (vorausgesetzt, es bestehen die ökologischen Bedingungen für die Verwendung von Teilparzellen als Replikate). Dies ist nicht unbedingt eine Entscheidung, die auf wissenschaftlichen Grundsätzen beruht, sondern eine regulatorische Schutzentscheidung, die auf dem Vorsorgeprinzip beruht.
- ▶ Nach den Erfahrungen, die in den letzten mehr als 20 Jahren mit der Durchführung von Regenwurm-Freilandstudien auf der Grundlage der bestehenden BBA- und ISO-Richtlinien und während des Projekts gemacht wurden, wurde ein Entwurf einer OECD-Prüfrichtlinie formuliert und der Ad-hoc SETAC GSIG Untergruppe zur Diskussion zur Verfügung gestellt. Die Diskussion über den Entwurf der Prüfrichtlinie ist derzeit noch nicht abgeschlossen.

1 Introduction

Since 1994, the risk of chemicals for earthworms in the field has been assessed by a test that had originally been standardised by the German Federal Biological Institute (BBA 1994). Since 1999, an international guideline standardised by the International Organization for Standardization (ISO) is available that has been updated several times up to now (last in 2014) without changing the basic approach (ISO 11268-3; 2014). However, according to an agreement between ISO and OECD, ISO guidelines should focus on the assessment of (potentially) contaminated compartments (water bodies, sediments, waste materials as well as soils), i.e. they are used in a retrospective approach to environmental risk assessment. In contrast, OECD guidelines serve the purpose of a prospective assessment of individual chemicals and defined chemical mixtures such as pesticide formulations. As a consequence of this agreement, several ISO guidelines used in the testing of chemicals were transcribed to the OECD format during the past 10 years. In the course of this conversion, which in the case of the earthworm field test is performed under German lead management since April 2013 as OECD project no. 2.47 ('New Test Guideline on Determination of Effects on Earthworms in Field Studies'), it was also checked whether apart from formal adjustments further changes were considered necessary. This assessment was performed within the framework of the "OECD Test Guidelines Programme" by the national experts that were supported by an ad hoc sub-group (led by Dr. J. Römbke) of the Global Soil Interest Group (GSIG) of the Society for Environmental Toxicology and Chemistry (SETAC). In this working group, experts from contract research organisations (CROs), industry, universities and public authorities have been working together for about seven years (i.e. all relevant stakeholders were involved in this process from the very first beginning). Based on the experiences made during the past 20 years, it was decided that several aspects of the guideline need to be adjusted. In this context, also technical recommendations compiled by the ad hoc SETAC GSIG sub-group could be taken up (Kula et al. 2006), in particular because they have already been followed in regulatorily required tests in close consultation with the national and international competent authorities and they have been included in the newest ISO-Guideline (2014). Specifically, besides technical details (e.g., exchange of the extraction fluid used in earthworm sampling and the review of the reference substance to be used in this test), primarily the study design and the statistical evaluation of the test results had to be optimised. Regarding the study design, the ISO Guideline already mentions the possibility of performing studies according to a dose-response design, an option that is deemed to "clearly facilitate environmental risk assessment compared to single dose studies" (ISO 2014). In particular, the variability of the community assessed in the field, the (lacking) statistical significance of the results of the field test and the level of safely statistically detectable effects of the tested chemicals needed improvement. To address these issues, scientifically robust and practical information was missing. The generation of this information was the objective of this project. In close cooperation with the ad hoc SETAC GSIG sub-group, the following aims were reached by performing three work packages (WP):

- ▶ WP1: Evaluation of existing data and development of proposals for an optimized design of the earthworm field test: Compilation and critical evaluation of information available in the literature and the database of the UBA regarding the standardised performance of earthworm field studies with the aim of developing suggestions for improving the test design. Compilation of field study data in a database, quality-check and characterization regarding their environmental and agricultural variables. Analysing the data concerning species composition, natural variability of earthworm biomass and abundances, and calculating the minimum detectable difference (% MDD) between control and treatment of all field studies. Dis-

discussion of suggestions to improve the existing test design and transforming them into detailed proposals for an improved test design for the earthworm field test;

- ▶ WP2: Experimental investigations: Performance of a pilot field study according to the new test design proposals. In order to use the available resources as optimal as possible and answer the different research questions, the design of this pilot field test allowed to analyse possible variations of the test design:
 - Combination of a so-called NOEC- and ECx-design, i.e. variation of the number of plot replicates (three or six);
 - Use of six instead of four individual samples per plot;
 - Testing of six concentrations of the test chemical carbendazim.
- ▶ In-depth statistical analysis of the pilot field study in combination with the existing database regarding natural variability in earthworm communities. Calculation of effects thresholds, effect concentrations and community analysis. Formulation of design requirements for earthworm field studies and identification of limitations and open questions.
- ▶ WP3: Participation in the OECD process: Formulation of a new draft OECD TG based on the existing ISO guideline 11268-3 (2014) but following the formal requirements of the OECD, using the experiences made in the pilot study as well as the evaluation of the UBA database. Discussion of this draft TG within the ad hoc SETAC GSIG sub-group in a final project meeting. The combined results of the development and discussion process will be submitted to OECD.

2 Evaluation of existing data and development of proposals for an optimized design of the earthworm field test (WP1)

In the course of the preliminary analyses and investigations, the ISIS database (“Information System Chemical Safety”) of the UBA was identified as a useful source for data analysis of earthworm field tests. The database held 150 entries for field studies on earthworms (date of query: October 10th, 2016).

Raw data on earthworm abundances and biomass per sample (0.25 m²), which are essential for further statistical examination of field test data, were not included in the ISIS database. Nevertheless, technical reports for most of the field studies were available as pdf-files from the UBA. For the available studies, quality criteria for data were initially defined with regard to further statistical investigations (chapter 2.1). Subsequently, the quality of data for each field test was checked and the eligible datasets were manually digitalised. Raw data “abundance” and “biomass” of earthworm species and groups on sample level (0.25 m²) were extracted from original study reports. A unified database was developed for the collected earthworm raw data for further statistical analysis.

The subsequent systematic procedure of descriptive metadata analysis and advanced statistical calculations were performed using the software R (3.1.1) with R Studio (1.0.136). Scripts were compiled for the various evaluation procedures, which access the common database of the digitized raw data.

2.1 Earthworm field study database – compilation and quality check

The project consortium decided that the earthworm field studies for subsequent statistical analyses should possess following characteristics:

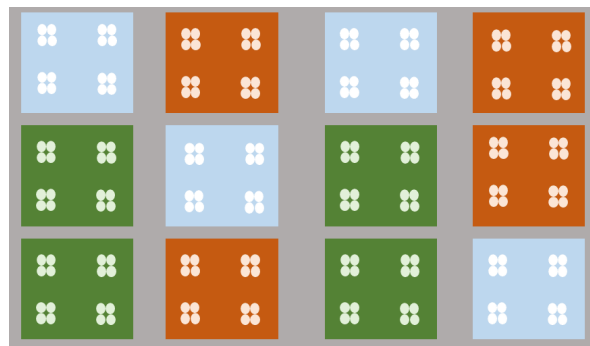
The extraction method of earthworm sampling should include formalin/allyl isothiocyanate (AITC) extraction and hand-sorting. A bias of the sampled species composition due to the use of the octet sampling is therefore prevented. Moreover, the technical reports should include raw data collected on sample (= subplot) level. This prerequisite enables an analysis of test data at sample level in comparison to the conventional evaluation at plot level.

The field studies that fulfilled these characteristics were divided into two classes: Tests were assigned to class 1 with only one test chemical treatment and one toxic reference treatment compared to the control (limit test), while in class 2, several test chemical treatment levels were considered in the test. However, it has to be stated that the limit test set-up was used far more frequently than test with several treatments. This can be explained by the fact that in the BBA test guideline only limit tests were listed. The newest ISO guideline from 2014 includes already the dose-response test design.

We considered in total 21 different field studies, including 1-3 treatments (+ control and reference treatment). Eleven field studies were classified into class 1 (limit-tests), two field studies assessed two different substance concentrations next to the control, and another eight field studies were designed with three treatments (class 2). In addition, further five studies with digitalized raw data at sample or plot level have been integrated into the database, each with a slightly different sampling method. This data, however, was not used for meta-analyses of the species composition, but only for feasible test-internal statistical evaluations. In total, data of 26 field tests of the ISIS database (+test data of the following pilot study were used) for statistical calculations were used.

The processed field studies were carried out according to the ISO guideline 11268-3 (2014) “Soil quality - Effects of pollutants on earthworms - Part 3: Guidance on the determination of effects in field situations” or in consideration of the BBA (Biologische Bundesanstalt) guideline part VI, 2-3. (1994) “Richtlinien für die amtliche Prüfung von Pflanzenschutzmitteln, Nr. VI, 2-3, Auswirkungen von Pflanzenschutzmitteln auf Regenwürmer im Freiland”. Therefore, the analysed test procedures follow a common approach (Figure 1).

Figure 1: Exemplary illustration of an earthworm field study test design (random design). The different colours of the boxes represent a control treatment and different concentrations of the tested substance (=treatments). The white dots correspond to the samples (= subplots) collected at each time point of sampling. Four samples (=0.25 m²) per time of testing are aggregated to one replicate according to the current guideline



Source: RWTH Aachen University

All reports contain data on earthworm species, numbers, and biomass collected for sampling plots treated with a test substance in a randomized arrangement (four replicates per treatment) and compared with those collected from control and reference plots (e.g. those treated with carbendazim or benomyl). Every replicate (=sampling plot) consists of four aggregated samples (=subplots) of 0.25 m² per sample (1 m² sampling plot in total). The sampling dates are usually set shortly before application and about 1-3 months, 4-6 months and 12 months after application of the test chemical. Tests usually start in April/May. The calculations of effects within the test procedures included the evaluation of total abundance and biomass on species level and for earthworm groups. Juvenile earthworms were summarized and evaluated on genus level (morphological groups: *Tanylobous* and *Epilobous*). In addition, the ecological groups of endogeic, epigeic and anecic earthworms were differentiated. As already mentioned above, the field studies usually include a control treatment, a reference and 1-3 concentration levels. Univariate statistical analyses for multiple (tests with more than one treatment) or pairwise comparisons (control vs. treatment) were applied to the recorded data. For multiple comparisons, Dunnett/Williams tests for normally distributed and homogeneous data were used, otherwise a Bonferroni U-test or Jonckheere-Terpstra Step-down-test was performed. For pairwise comparisons, Student's t-test or Mann-Whitney U-test were used.

The existing raw data of endpoint measures earthworm biomass and abundance on sample level for all sampling time points, treatments, species and aggregated earthworm groups, as well as available data on land use and covering vegetation were collected and integrated into a field study database.

2.2 Data collection: environmental and agricultural variables

Table 2 shows an overview of the earthworm studies that were collected and used within the statistical analyses, the number of identified species and agricultural and environmental parameters.

Table 2: Environmental and agricultural variables, study class and number of sampling time points of selected earthworm field tests from the ISIS Database (UBA)

Study ID	Land use application date	Study class	No. of sampling dates	Country	No. of species	Diversity Shannon	Mean individuals /m ² control	Mean biomass g/m ² control	Vegetation	Texture (USDA)	Soil pH
test2674	bare soil	2	4	GER	10	1.5	279	152	grass-clover mixture	sandy loam (SaLo)	7.7
test2818	bare soil	2	4	GER	9	1.1	337	155	winter wheat	no information	7.3
test2863	bare soil	2	4	GER	9	1.2	451	189	grass-clover mixture	sand (Sa)	6.6
test1777	crop	1	3	GER	7	1.1	105	59	grain maize, spring wheat	no information	6.5
test1941	crop	2	3	GER	6	0.3	123	59	winter oilseed rape	sand (Sa)	5.4
test2225A	crop	2	6	GER	11	1.0	275	92	maize	sandy loam (SaLo)	6.7
test2225B	crop	2	6	GER	12	1.0	275	92	maize	sandy loam (SaLo)	6.7
test2225C	crop	2	6	GER	11	1.0	275	92	maize	sandy loam (SaLo)	6.7
test2237	crop	1	4	UK	12	1.5	194	104	maize	sand (Sa)	6.3
test2268	crop	1	4	UK	12	1.5	194	104	maize	sand (Sa)	6.3
test2594	crop	2	3	GER	14	1.0	191	68	barley	sand (Sa)	5.1
test2678	crop	1	12	GER	13	1.3	253	208	grass-clover mixture	no information	7.1
test2740	crop	2	4	GER	14	1.1	201	79	spring barley	no information	5.1
test2764	crop	1	5	GER	10	1.0	216	146	sugar beet	no information	6.9
test3014	crop	2	4	GER	10	1.0	170	50	maize	sand (Sa)	5.71
test3064	crop	1	3	GER	11	0.9	70	41	barley	loamy sand (LoSa)	7

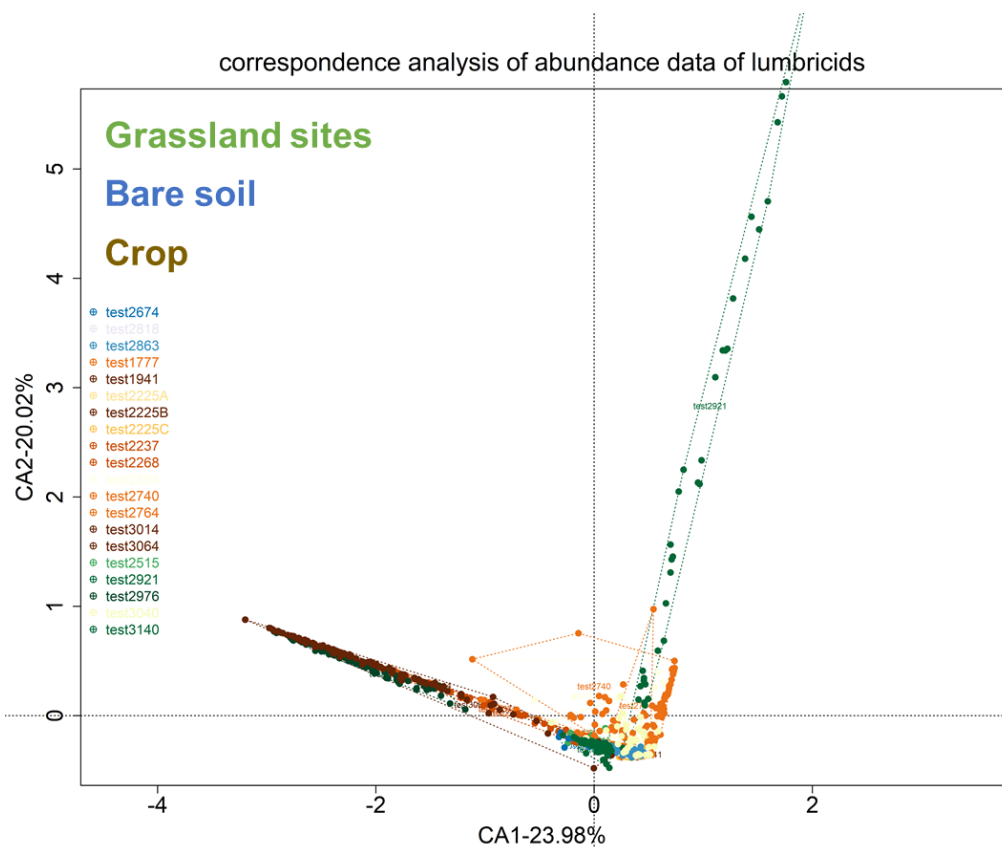
Study ID	Land use application date	Study class	No. of sampling dates	Country	No. of species	Diversity Shannon	Mean individuals /m ² control	Mean biomass g/m ² control	Vegetation	Texture (USDA)	Soil pH
test2515	grassland	1	4	GER	13	1.6	421	134	grass	loamy sand (LoSa)	6.9
test2921	grassland	1	3	GER	14	1.5	475	114	grass	loamy sand (LoSa)	7.43
test2976	grassland	2	4	GER	11	1.3	124	71	grasses	loamy sand (LoSa)	6.24
test3040	grassland	1	3	GER	11	1.4	119	75	grass	sand (Sa)	5.83
test3140	grassland	1	11	GER	13	1.4	724	254	grass	no information	6.9

For all data analyses in the project, the group of undetermined earthworm individuals in field studies were excluded. Descriptive metadata of the field studies reveal that the composition of species among all field studies report 6 - 14 species per study. The respective Shannon Diversity Index was between 0.3 and 1.6 (mean: 1.2). It can be noted that the diversity index is slightly higher on grassland sites (mean: 1.44) than on other land use types (bare soil: 1.27; crop sites: 1.05). Accordingly, the minimum number of species in grassland is at least 10. This trend can also be observed for the mean individuals sampled, which is about 372 individuals per m² on grassland, 356 individuals/m² on bare soil and about 196 individuals/m² on crop sites.

These metadata show a slight tendency towards better interpretability and evaluation of grassland field studies in terms of different endpoints at species and community level. It should be noted, however, that data may vary widely between field studies and within studies. In addition, the history of land use for the investigation sites in the past is often unknown. For studies of the database, only data at the date of application was taken into account. The history of database field studies was not included in the database and was therefore not further evaluated in this project.

The composition of earthworm communities within the field tests were analysed and compared to each other using a correspondence analysis for abundance data of all data sets (adults only, Figure 2). Thus, a potential systematic impact of environmental conditions on the community was tried to be investigated.

Figure 2: Correspondence analysis of earthworm species abundance data for field studies of the database (adults only, all sampling time points and treatments)



Source: RWTH Aachen University

The field studies were color-coded according to their different land use types (land use at the date of application). The first, horizontal axis of the correspondence analysis (CA) accounts for 23.98% and the second axis for 20.02% of the variance in the examined data set. In this coarse analysis, the communities of the various field studies do not have any significantly definable patterns according to their land use, if all sampling time points and treatments are included in the analysis. However, a gradient within the test data, which leads to a partitioning of the data by different land use type, can be interpreted for this illustration as grassland sites (green dots) being separated from crop sites (brown dots). A more detailed evaluation separating e.g. sampling time points in pre-sampling and first until fourth post application sampling could show more separated cluster. To justify a distinct classification, e.g. based on environmental influences, the seasonality or the impact of treatment, an in-depth analysis of species presence and abundance development during the course of the tests would be necessary. The current set of collected test data and the number of differentiated species per test approach is not sufficient to take these influencing and interacting factors adequately into account.

2.3 Field study data: Species composition, variability and MDDs

Based on the ISIS-database pre-processing, data of field studies for earthworm communities were subsequently analysed. The sampled earthworm individuals of the 21 field studies belong to a total of 17 different species (Table 3).

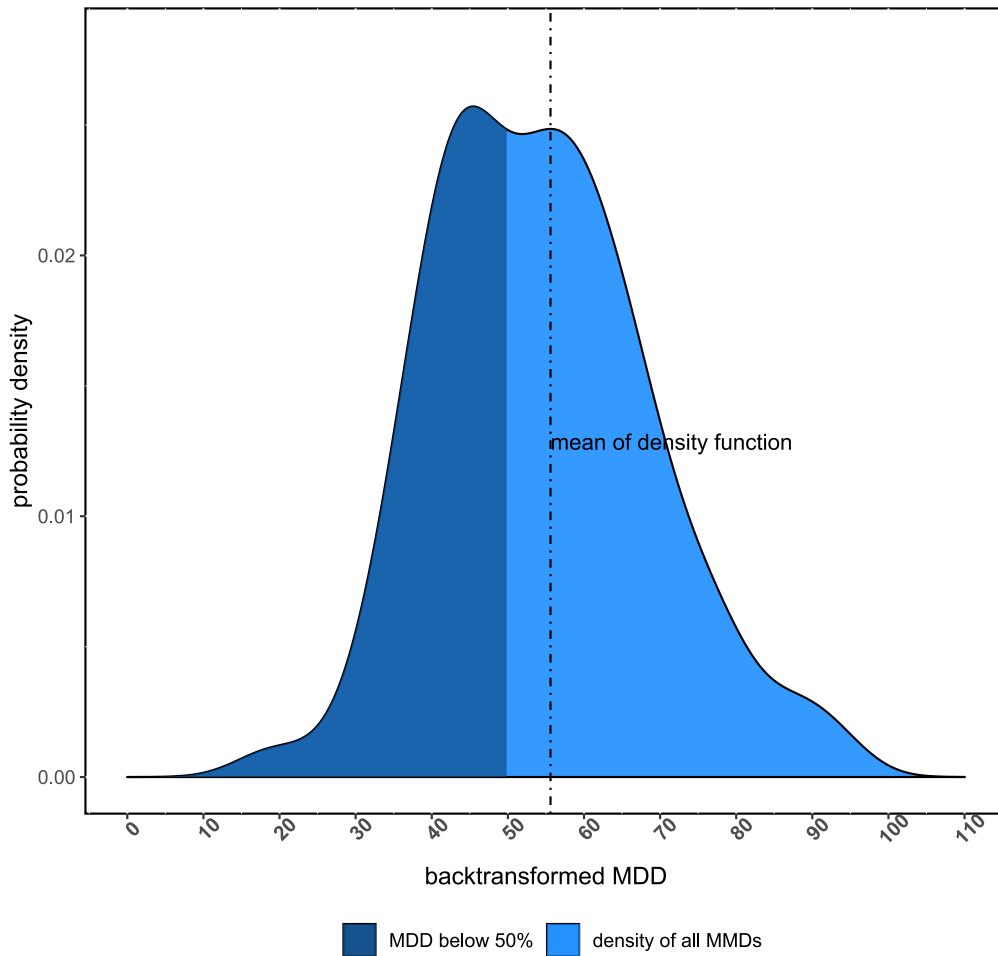
In addition to these analyses, aggregated taxa groups and genus level juveniles were subdivided. Aggregated groups were developed for “Total earthworms”, “Total adults”, “Total juveniles”, “Total anecics”, “Total endogeics”, “Total epigeics”, “Total epilobous adults”, “Total epilobous juveniles”, “Total tanylobous adults” and “Total tanylobous juveniles”. Undetermined individuals were excluded from further analyses. A detailed calculation of the natural variability of earthworm biomass and abundances, illustrated by the coefficient of variation for control treatments, is calculated and presented in comparison to results of the pilot field study in chapter 3.2.3.1 (‘Analysis of natural variability in earthworm communities’). Their implications towards the statistical test power of earthworm field studies are outlined in chapter 3.2.3.1.3 (‘Sample size modelling approach’).

In a preliminary analysis, however, the statistical characteristics of the database field studies were already reviewed. As a statistical measure, the minimum detectable difference (% MDD) for the endpoint total earthworm abundance between control and treatment of all field studies was calculated (Figure 3). The analysis of the MDD was based on the approach of Brock et al. (2015). However, the equation used was adjusted to some extent: Brock et al. sets a statistical test power of 50% by default for the testing procedure. However, since this does not correspond to the general conventions of a desired test power of ecotoxicological testing (e.g. OECD 2012), we have included a term to consider and adapt test power. This was fixed at 80% (type-II error of 20%; see Duquesne et al. 2020).

Table 3: Percentages of sampled earthworm species and assigned ecological and morphological groups

Species	dominance [%]	ecol. group	morph. group
<i>Aporrectodea caliginosa</i> (Savigny, 1826)	46.3	endogeic	epilobous
<i>Aporrectodea rosea</i> (Savigny, 1826)	17.1	endogeic	epilobous
<i>Lumbricus terrestris</i> Linnaeus, 1758	11.9	anecic	tanylobous
<i>Allolobophora chlorotica</i> (Savigny, 1826)	9.6	endogeic	epilobous
<i>Lumbricus castaneus</i> (Savigny, 1826)	4.4	epigeic	tanylobous
<i>Aporrectodea longa</i> (Ude, 1885)	2.9	anecic	epilobous
<i>Lumbricus rubellus</i> Hoffmeister, 1843	2.5	epigeic	tanylobous
<i>Aporrectodea limicola</i> (Michaelsen, 1890)	2.1	endogeic	epilobous
<i>Octolasion tyrtaeum</i> (Savigny, 1826)	1.7	endogeic	epilobous
<i>Murchieona minuscula</i> (Rosa, 1906)	0.883	endogeic	epilobous
<i>Octolasion cyaneum</i> (Savigny, 1826)	0.5	endogeic	epilobous
<i>Proctodrilus antipae</i> (Michaelsen, 1891)	0.052	endogeic	epilobous
<i>Eisenia fetida</i> (Savigny, 1826)	0.027	epigeic	epilobous
<i>Satchellius mammalis</i> (Savigny, 1826)	0.019	epigeic	epilobous
<i>Aporrectodea cupulifera</i> (Tetry 1937)	0.018	endogeic	epilobous
<i>Lumbricus festivus</i> (Savigny, 1826)	0.005	endogeic	tanylobous
<i>Dendrobaena illyrica</i> (Cognetti, 1906)	0.003	epigeic	epilobous

Figure 3: Distribution of the probability density for the minimum detectable difference (MDD in %) of total earthworm abundance data in the earthworm field study database extracted from ISIS (all sampling dates, empirical MDD between 11% and 100.2%)

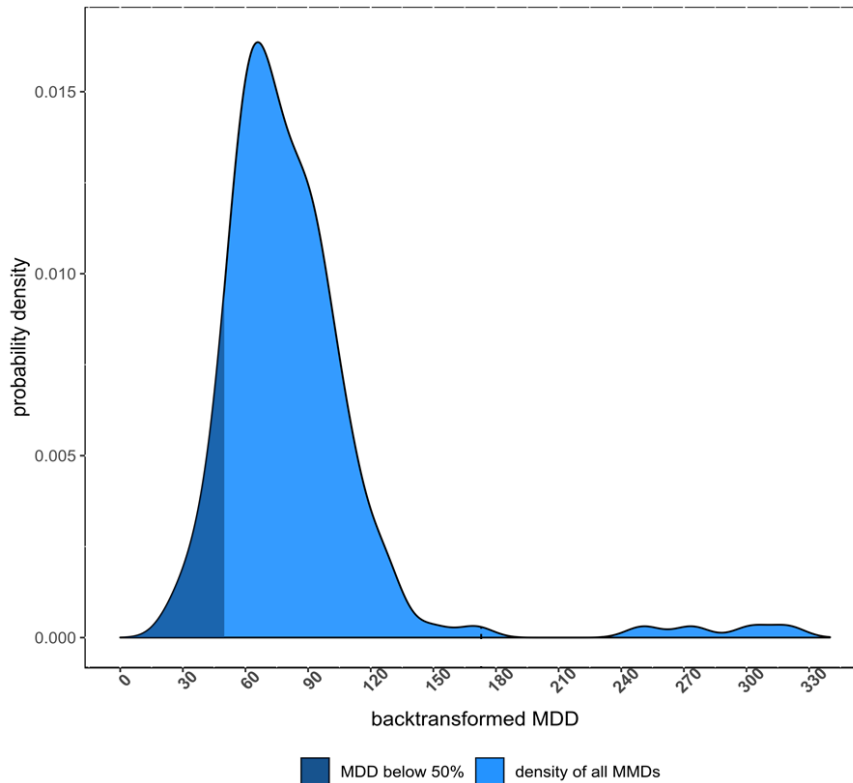


Source: RWTH Aachen University

Although the most likely value of the MDD for abundance data of total earthworms in the database is 45%, the probability of obtaining a minimum detectable difference smaller than 50% of the control is 42%. The probability of obtaining an MDD between 10% and 35% was 8% (small effects according to EFSA PPR 2017). However, this range relates to effects on the protection goals and not to the measurement endpoints as the analysed earthworm field data. The same calculations for total biomass give even lower power values than for total abundance: a minimum detectable difference smaller than 50% was only detected in 32% of the cases (mode at 67%, which is the most likely value of the density function).

For the aggregated group of total earthworms, the most powerful MDDs have been calculated. Even for the most dominant species in the database, *Aporrectodea caliginosa* (46% of all sampled individuals), considerably lower probability was found to detect statistically significant effects in field studies (Figure 4).

Figure 4: Distribution of the probability density for the minimum detectable difference (MDD in %) of *Aporrectodea caliginosa* abundance data in the earthworm field study database (all sampling dates, lowest empirical MDD at 15.4%)



Source: RWTH Aachen University

The probability to calculate a minimum detectable difference less than 50% between treatment and control for the species *Aporrectodea caliginosa* is low, only 12% of the probability density distribution can be assigned to this range. The most likely value of the calculated probability distribution for minimum detectable differences for abundance data of *Aporrectodea caliginosa* is 66% (mode density function). This corresponds to a reduction of earthworm abundances of two third in treatments compared to a control. A less substantial reduction would not provide a statistically significant effect, even if biologically relevant. Again, even lower minimum detectable differences were calculated for the endpoint biomass. An MDD smaller than 50% was only calculated for 8% of the data. The respective mode of the density function is about 70%. Statistically significant identification of medium-sized effects is rare for this endpoint due to the variability of data.

In an overall picture, best-practice studies (hand-sorting and formalin/AITC extraction) reveal low power to detect differences between control and treatment plots for aggregated taxa. For single species, this statistical potential for a reliable statistical identification of effects is even lower, as the example of *Aporrectodea caliginosa* shows. The even lower abundances of other species and the resulting increasing variability in the datasets (see detailed description of these correlations in chapter 3.2.3.1.1) lead to an even weaker distribution of the MDD for other single species. Boxplots and tables of MDD calculations for single species and earthworm groups are shown in the Appendix (A.3, Table A3-1 and Figures A3-1). There is a chance of obtaining sufficient MDD for single taxa to identify small effects, as shown for single instances within the database (Figures A3-1). However, the overall MDD for species and earthworm groups show, that a comprehensive detection of small effects (10 to 35% difference to control, see chapter 3.2.3.1.2 for corresponding effect classification) is not given in earthworm field tests. For all sampled spe-

cies, an MDD of 35% (threshold value for small effects) lies below the interquartile range in the distribution of the MDD considering all tests in the database (A.3, Figures A3-1). Across all sampling points, endpoints, species and earthworm groups of all database studies, only 8.4% of the measured MDD are below 35%.

With regard to statistical considerations, there are clear indications justifying the testing and adaption of a new field study test design in the course of this project. The limitations on the old design, covering limit-tests as well as NOEC-approaches, became evident: NOEC and related concepts have long been criticized in ecotoxicological literature (see chapter 3.2.1). Furthermore, these actual MDD calculations of the earthworm endpoints from the field studies of the ISIS database have revealed that potentially biologically relevant effects are not detectable in many field situations by standardized statistical procedures (regarding effect classes, see chapter 3.2.3.1.2).

2.4 Development of a pilot study test design

In order to meet the shortcomings of the current earthworm field test, possible designs for the planned pilot study were discussed.

An adapted test design should contain an option to perform regression approaches, which have been suggested as an alternative to the NOEC approach (chapter 3.2.1). The resulting estimated concentrations (EC_x) from fitting a curve to the data have been suggested as a more powerful alternative to the NOEC-value (e.g. Fox 2009). The number of concentration levels in the pilot field study has to be increased to investigate the suitability of an EC_x-design for earthworm field studies. In order to still include the possibility of deriving sound NOEC values in the field and improve statistical power of this procedure compared to the old design, we also increased the number of replicates on plot level for control treatment and two concentrations in the pilot field study. Number of samples per replicate should be increased in order to examine the changes in variance and, finally, to estimate if these samples can be used as individual replicates instead of aggregating data on plot level to improve statistical test power. This is done by a sample size modelling approach (chapter 3.2.3.1.3).

As the field conditions and practical feasibility of the earthworm pilot field study limited the total number of plots, the enlargement of the number of different treatments and the increase of plot and sample (= subplots) number per treatment must be adjusted in such a way that both research questions (feasibility of EC_x design and improvement of NOEC design) can be addressed. Also, the final proposal for an adapted design for earthworm field studies will need to be fit for the purpose to detect chemical effects but also take practicability into account.

As an output of the pilot study, an amended design for earthworm field studies is to be proposed, which will have a smaller replicate design than the pilot study – but will better address the statistical power of the study results in an amended set-up compared to the current design.

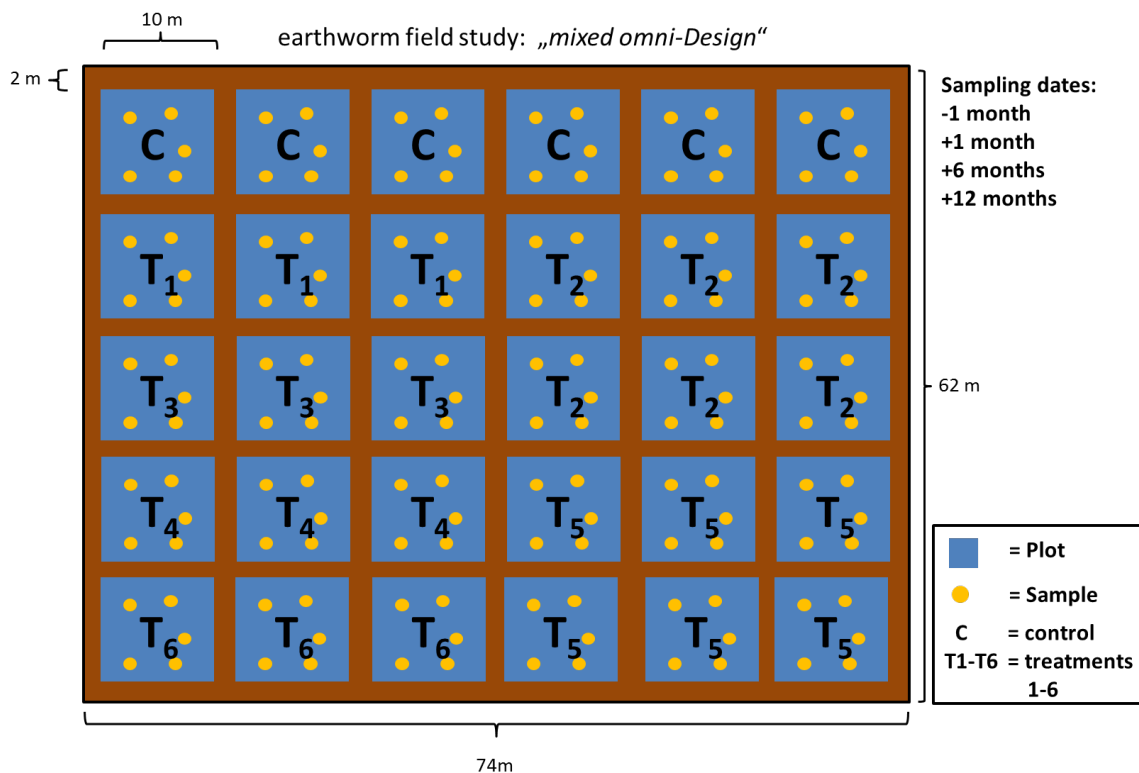
2.4.1 First proposal

In a joint discussion between UBA and the project consortium, the results of the evaluation described above led to a first proposal of the earthworm pilot field study design to be performed in 2017 (Figure 5). This design was characterized by combining a so-called NOEC- with an ECx-design and was called “mixed omni-design”:

- ▶ Four sampling dates, covering a total test duration of one year (as in ISO guideline 11268-3);
- ▶ One control (C) and six test chemical treatments (T) (only limit test in the ISO guideline);
- ▶ Number of plots per treatment six (C, T2, T5) or three (T1, T3, T4, T6) (four in the ISO guideline);
- ▶ Five samples per plot (four in the ISO guideline).

Running such a pilot study meant that in total 30 plots with 150 samples per sampling date had to be covered. This original proposal was considered by the project team as large but still practical in terms of handling (e.g. number of days needed for sampling, field size etc.).

Figure 5: Original proposal for the design of the pilot earthworm field study provided to the involved stakeholders prior to the meeting in Flörsheim



Source: RWTH Aachen University

The proposal of the test design for the pilot study was discussed during the meeting of the SETAC-GSIG earthworm field group in Flörsheim (February 20th – 21th, 2017). In the following, the most important parts of this discussion are summarized. The complete minutes of this meeting are given in the Appendix A.5.1 of this report.

2.4.2 Discussion of the pilot study in the ad hoc SETAC GSIG sub-group

2.4.2.1 Final test design

Before starting the discussion on the design of the earthworm pilot study itself, further recent contributions addressing different aspects of the planning, performance or evaluation of earthworm field studies were presented to the ad hoc SETAC GSIG sub-group. Vollmer et al. (2016) performed an assessment of the results of 26 standard earthworm field studies performed in Germany, France and Spain according to the ISO guideline 11268-3 (2014). In the context of this project, these results of their work were important:

- ▶ The statistical power of the current earthworm field test is suitable to detect medium effects (from 35 – 65%; EFSA PPR Panel in soil opinion, 2017) for total abundance or the most dominant species in most of the tests, but it is not sufficient for small effects (10 – 35%) with a desired test power of 80%, especially for individual species;
- ▶ An increase in the number of plot replicates from four to six will theoretically reduce the MDD by 5 to 10% of the original value;
- ▶ Increasing the number of plot replicates beyond 6 is not a reasonable option as this would increase the overall variability of earthworm populations which will probably diminish benefits on statistical power and the study design will become practically unfeasible.

The outcome of six standard earthworm field studies was summarized by Andrade et al. (2017). These authors conclude “that the standard test design of current earthworm field studies provide a suitable degree of statistical power when earthworm density is sufficiently high (i.e. >50 ind./m²), considering the magnitude of effects that are relevant at the earthworm community level (minimum significant difference (MSD) values up to 70% of the control mean value).” In addition, they state that statistical robustness could be improved by increasing the number of samples per plot: a decrease in the minimum significant difference (MSD) values was observed when increasing the number of samples per plot from four to eight.

In the discussion during the meeting, various changes to the “mixed omni-design” were proposed, all of them with the intention to improve the quality of the pilot study output but without strongly increasing the efforts at the same time. While the number of replicate plots dedicated to the NOEC- and ECx-components of the study was kept constant, the number of samples per plot was increased. In the original proposal it was five, but this idea was not supported by the majority of the group. In particular, the difference between four and five seemed to be too small to have an impact on the final result of such a test. In addition, the results published by Andrade et al. (2017) were also taken into account.

The resulting final test design was called “balanced design”. It was decided to take the same number of samples per plot in the NOEC- as well as in the ECx-plots (six), whereby the number of replicate NOEC- and ECx-plots will be six and three per treatment, respectively (Table 4).

Table 4: Number of plots and treatments for the combined NOEC- and ECx-design in the pilot earthworm field study. C = control; T1-T6 = treatments

Test design	Number of plots per treatment							Total number of plots	Number of samples per plot	Total number of samples
	C	T1	T2	T3	T4	T5	T6			
Balanced design (combined NOEC- and ECx-design)	6	3	6	3	3	6	3	30	6	180

2.4.2.2 Identification of the test chemical concentrations

The selected test chemical was carbendazim, since it (or its parent compound benomyl) has been used in earthworm field tests as a reference substance following the publication of the first earthworm field test guideline (BBA 1994). Actually, regarding the soil ecosystem, it is probably one of the best investigated chemicals and it is by far the best-studied pesticide in soil ecotoxicology:

- ▶ It has been used as reference substance in ISO earthworm field studies for more than 20 years (partly in parallel with the parent active substance benomyl). Part of these studies were submitted to UBA during pesticide authorization processes;
- ▶ Carbendazim was used in an EU project focusing on the development of a standard semi-field method where Terrestrial Model Ecosystems (TME) have been employed (e.g. Knacker et al. 2004; Römbke et al. 2004).

Using the available information from these different sources, various carbendazim concentration ranges were discussed. In detail, information from regulatory field studies (in total 16 studies) from the ISIS database were analyzed by the consortium, together with data from the literature (especially the EU TME ring test, see above). For this exercise, it was assumed that the effect of the test substance carbendazim on earthworm endpoints assessed at 4 to 6 months after application would be the most suitable in order to decide which treatment rates to select for the pilot field study.

The following six application rates (plus a negative, i.e. water-only, control) were finally selected in order to cover a range spanning from concentrations where no effects are expected to concentrations where strong effects are likely (Table 5).

Table 5: Application rates of the earthworm pilot field study. Concentrations are given in kg active substance (a.s. carbendazim)/hectare (ha)

Treatments	T1	T2	T3	T4	T5	T6
	0.6	1.8	3.2	5.8	10.5	31.5

It should be noted that the spacing factor is not fixed between the different treatments. This approach is already used in laboratory tests following the ECx design (e.g. in earthworm reproduction tests according to OECD 222, 2004) where it is stated that “The spacing factor may vary, i.e. less than or equal to 1.8 in the expected effect range and above 1.8 at the higher and lower concentrations”. While in the proposed test design for the pilot field test this factor is as high as 3 between the lowest test rates (and between the highest rates), it is about 1.8 between the rates at the centre of the treatment range. In the currently used ISO guideline 11268-3 (2014), the reference substance carbendazim should yield a statistically significant difference of at least

50 % on overall earthworm abundance and/or biomass compared to the control at least at one sampling date, when applied at rates of 6 to 10 kg a.s. carbendazim/ha. Thus, such effects should be detectable at the three highest application rates. Accordingly, and referring to the experiences made in the EU project mentioned above, it was postulated that no detectable effects should appear at the two lower rates. A priori analyses have shown that an EC₅₀ could be expected at rates around 2.5 kg carbendazim/ha.

3 Experimental investigations and statistical analyses (WP2)

3.1 Performance of the pilot field study

3.1.1 Experimental site

3.1.1.1 Characterisation of the experimental site

Arable land was chosen for the trial in a distance of less than 10 km from the laboratory of the ECT GmbH. It was owned by a local farmer and had been leased by ECT GmbH for the duration of the field trial. The field belonged to the cadastral area no. 35 of Flörsheim, Wicker (Germany), land parcels no. 4 to 6, named “im Strengen”. It was surrounded by agricultural fields and pathways (Figure 6). The experimental plots were installed within an area of approximately 55 m by 107 m and were at least 5 m away from neighbouring fields or pathways.



Figure 6: Aerial view of the experimental site in Flörsheim Wicker

Source: Google Maps, modified by ECT Oekotoxikologie GmbH

The soil of the field site was characterised by ECT GmbH (pH, water holding capacity) and Landwirtschaftliche Untersuchungs- und Forschungsanstalt Speyer (LUFA; all other parameters) using standardised methods. The data are given below (Table 6).

Table 6: Physical and chemical characterization of the field soil (0 – 10 cm depth)

Parameter	Measured value
pH (CaCl ₂)	7.2
C _{org} [% dm]	1.46
Organic matter [% dm] ^a	2.51
N _{tot} [% dm]	0.17
CaCO ₃ [% dm]	0.4
Soil type (USDA)	Silt loam
Clay (<0.002 mm) [%]	21.2
Silt (0.002 – 0.050 mm) [% dm]	55.9
Sand (0.050 – 2.000 mm) [% dm]	22.9
Water holding capacity [% dm]	55.3
Cation exchange capacity [cmol/kg dm]	14.6

^a Approximated from C_{org} by applying a factor of 1.72 (AG Boden 2005). dm = dry mass.

3.1.1.2 Field site history

The history of the field site with regard to crop and application of fertiliser and plant protection products is summarised in Table 7. Winter wheat was grown on the field before the study took place. To free the experimental site from vegetation without soil tillage that would have impacted the earthworm community, glyphosate was applied once on 16 March 2017 at a rate of 1.8 kg a.s./ha (Figure 7).



Figure 7: Experimental site on 28 March 2017, i.e. 12 days after glyphosate application

Tyre tracks lie offside the experimental area.

Source: ECT Oekotoxikologie GmbH

Table 7: History of the field site with regard to crop and application of fertiliser and plant protection products

Year	Crop	Fertiliser	Plant protection products
2014	Winter wheat	KAS 27 (80+60+70 kg/ha N)	Herbicides: Atlantis (295 g/ha); Starane XL (1.2 l/ha) Fungicides: Aviator X-pro (0.5 l/ha); Skyway X-pro (1.0 l/ha); Tebucur (0.5 l/ha) Insecticides: Biscaya (300 ml/ha) Growth regulators: CCC 720 (0.7 l/ha); Medax Top (0.5 l/ha)
2015	Winter barley	EPSO Combitop (10 kg/ha); KAS 27 (35 kg/ha N); N/P/K 15/15/15 (60 kg/ha N, 60 kg/ha P ₂ O ₅ , 60 kg/ha K ₂ O)	Herbicides: IPU (2.5 l/ha); Stomp (2.0 l/ha) Fungicides: Amistar Opti (1.5 l/ha); Imput Classik (0.8 l/ha); Tebucur (0.5 l/ha) Insecticides: Fastacse (100 ml/ha); Karate Zeon (75 ml/ha) Growth regulators: Medax Top (0.6 l/ha)
2016	Winter oilseed rape	Boron (1.6 l/ha); Digester liquor (1.57 t TM/ha: 4.6 kg N, 1.08 kg P ₂ O ₅ , 5.1 kg K ₂ O, 2.0 kg CaO); EPSO Microtop (6+6 kg/ha); Sulphan 24 + 6S (88 kg/ha N, 22 kg/ha S); Sulphan 24 + 6S (77 kg/ha N, 19 kg/ha S)	Herbicides: Butisan Gold (2.25 l/ha); Panarex (1.0 l/ha) Fungicides: Cantus Gold (0.5 l/ha); Tebucur (0.5+0.4 l/ha); Tilmor (0.5 l/ha) Insecticides: Biscaya (300 ml/ha); Hunter (150 g/ha); Karate Zeon (75 ml/ha); Teban (200 ml/ha)
2017	Winter wheat	none	Herbicides: Glyphosate (1.8 kg/ha)

3.1.1.3 Installation of experimental plots

The experimental plots were installed on 28 March 2017 (Figure 8). For each treatment, i.e. control (C) and six different test chemical treatments (T1 to T6), six (C, T2, T5) or three (T1, T3, T4, T6) plots (= replicates), each 10 m by 10 m, were installed at the field site (Figure 9) and assigned randomly. The distance between two neighbouring plots was 3 m. The distance to the surrounding fields or cart tracks was at least 5 m.

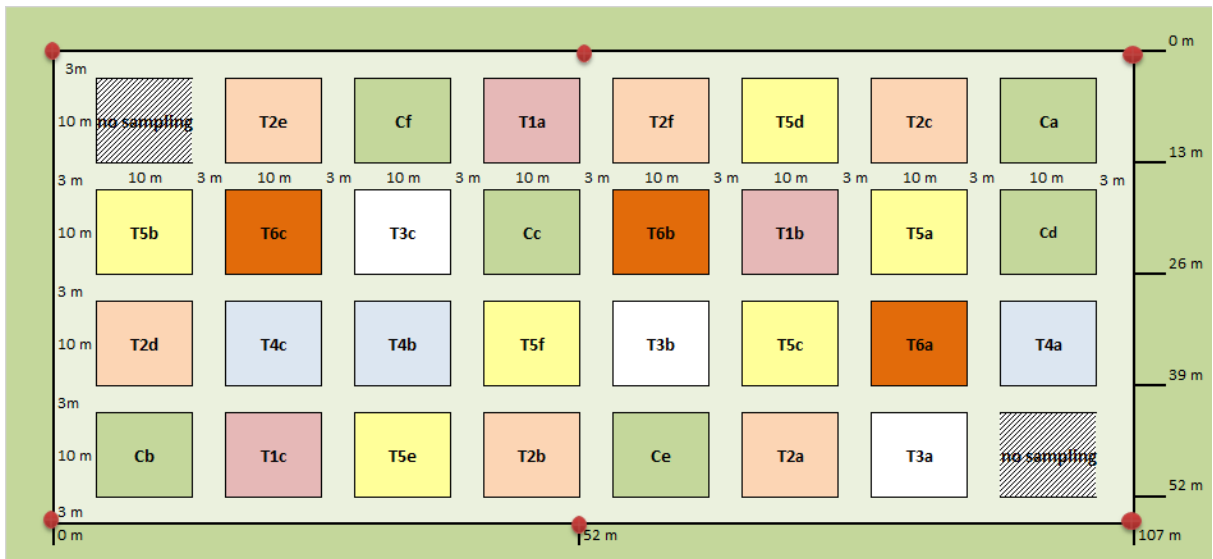
Figure 8: Experimental site on 30 March 2017 after installation of the plots



Tyre tracks lie offside the experimental area.

Source: ECT Oekotoxikologie GmbH

Figure 9: Scheme of the trial area with randomized allocation of the treatment to the plots (squares)



C (a-f; control), T1, T3, T4, T6 (a-c) and T2, T5 (a-f; test chemical treated). The size of each plot was 10 m by 10 m and the distance between plots was 3 m

Source: ECT Oekotoxikologie GmbH

3.1.2 Test chemical, test performance and application

3.1.2.1 Test chemical (a.s. carbendazim)

The test chemical was applied as the suspensible concentrate (SC) formulation Carbomax 500 SC (Table 8). The determination of the content of the active substance (a.s.) carbendazim was performed by Chemisches Institut Pforzheim GmbH (CIP) on 16 June 2016. No pre-treatment such as solution in an organic solvent was necessary. Aqueous suspensions were prepared in the field immediately before spray application.

Table 8: Characterization of the test chemical

Name:	Carbomax 500 SC
Active substance (a.s.):	Carbendazim
Classification:	Broad-spectrum benzimidazole fungicide
Content (a.s.):	490 g/l
IUPAC name (a.s.):	Methyl 1 <i>H</i> -benzimidazol-2-ylcarbamate
Chemical formula (a.s.)	C ₉ H ₉ N ₃ O ₂
CAS No. (a.s.):	10605-21-7
Batch No.:	0002-16-14400T/B
Density [g/cm ³]:	1.161 at 20°C
Physical appearance:	Odourless, greyish beige viscous liquid
Water solubility:	Suspensible
Re-analysis date:	June 2018

3.1.2.2 Test design and application rates

The day before application the appropriate amounts of test chemical for each of the six test chemical treated plots were weighed into separate vessels in the laboratory. These appropriately labelled vessels were brought to the field on the day of application. The spray solutions for the application of the test chemical (based on 400 l/ha, equivalent to 4.0 l per plot) were freshly prepared in the field. For each plot a volume of 4.5 l of spray solution was prepared (Table 9).

Table 9: Test Design, application rates, test chemical concentration in the spray solution and application rate per plot

Carrier (water) application rate: 400 l/ha; single plot area: 100 m²; application rate per plot (nominal): 4.0 l; volume of spray solution prepared per plot: 4.5 l

Code	No. of plots	Test chemical application rate [kg a.s./ha]	Test chemical per plot (nominal) [g a.s.]	Amount of test chemical prepared per plot [g]	Spray solution concentration [g a.s./l]
C	6	0.0	0.0	0.00	0.00
T1	3	0.6	6.0	6.75	1.50
T2	6	1.8	18	20.25	4.50
T3	3	3.2	32	36.00	8.00
T4	3	5.8	58	65.25	14.50
T5	6	10.5	105	118.125 ^a	26.25 ^a
T6	3	31.5	315	354.375	78.75

^a Due to an insufficient supply of test chemical, for three plots (T5a, T5c, T5f) the amount of test chemical was lower (112.941 g a.s.) and thus the spray solution concentration was lower (25.10 g a.s./l). Accordingly, a higher volume (4.18 l) was applied per plot.

3.1.2.3 Calibration of spray equipment

Water (control) and test chemical were applied using a mobile parcel sprayer (plot sprayer PL 1, Baumann Saatzuchtbedarf, Waldenburg, Germany) which was equipped with a spray boom carrying 5 nozzles (type Lechler AD 120 04; Lechler GmbH + Co KG, Metzingen, Germany), used commonly in agricultural practice. The spray nozzles were arranged in line at 0.5 m above the ground with a distance of 0.5 m between the nozzles. The total working width was 2.5 m. The lateral distribution of the spray was regularly checked at an official test bench for agricultural spray booms. The parcel sprayer PL 1 was calibrated prior to use in order to assure the uniform flow rate of its 5 individual spray nozzles. This was done by repeated determination of the output of spray solution per spray nozzle over time. Output per spray nozzle was assessed three times and the velocity of the PL 1 was calculated according to the formula:

$$Velocity [km/h] = \frac{\text{total output [l/min]} * 600}{\text{field rate [l/ha]} * \text{working width [m]}}$$

with: total output = sum of all 5 nozzles [l/min]
 field rate = volume of spray solution [l/ha]
 working width = effective size of spray area [m]

3.1.2.4 Performance of application

The test chemical was applied once on 11 April 2017. The water (control) and the test chemical were applied onto the bare soil surface of arable land. The pneumatic spray device consisted of a hand driven one-wheel frame equipped with a spray tank, compressed air cylinder and a boom with five spray nozzles (Figure 10). The test chemical was applied at a wind velocity below 3 m/sec to avoid any risk of cross contamination due to possible drift during application. The method of treatment was done as close to field application procedures as possible.

Figure 10: Application of the test chemical on 11 April 2017



Source: ECT Oekotoxikologie GmbH

The application was performed plot by plot, beginning with the application of water onto the control plots, followed by the test chemical application in ascending order. A volume of 4.5 l of the respective spray solution was filled into the tank of the parcel sprayer ("initial volume") prior to the application of each plot. Before starting to move the parcel sprayer, the release button was pressed until all nozzles released spray solution uniformly. The spray solution released during this procedure was collected ("pre-release"). Thereafter the plot was sprayed by crossing the plot in four parallel rows of 2.5 m width each. The speed necessary had been determined during the calibration of the parcel sprayer immediately prior to application in the field. After terminating the application of one plot, the release button was pressed to completely empty in the tank of the parcel sprayer. The amount of spray solution released during this procedure ("post-release") was collected too and comprised together with the "pre-release" the "rest-volume".

To calculate the amount of spray solution actually applied to the plot, the "rest volume" was subtracted from the "initial volume". Nominally, 4.0 l should have been applied per plot. Therefore, a "rest-volume" of 0.5 l should remain after application of the plot. In general, the application volume was within the expected range on all plots of the test chemical treatments. Maximum deviations from nominal treatment volume were -11.2% (T5f) and +4.3% (T5b). The actually applied volumes are given in Table 10.

Table 10: Actual applied volumes of spray solutions of the test chemical

Plot code	nominal [l]	applied [l]	% of nominal
T1a	4.00	3.93	98.3
T1b	4.00	3.70	92.5
T1c	4.00	3.94	98.5
T2a	4.00	4.00	100.0
T2b	4.00	4.11	102.8
T2c	4.00	4.10	102.5
T2d	4.00	4.14	103.5
T2e	4.00	3.89	97.1
T2f	4.00	4.01	100.3
T3a	4.00	3.87	96.8
T3b	4.00	3.91	97.8
T3c	4.00	4.13	103.3
T4a	4.00	4.06	101.5
T4b	4.00	4.07	101.8
T4c	4.00	4.09	102.3
T5a	4.18 ^a	4.20	100.5
T5b	4.00	4.17	104.3
T5c	4.18 ^a	3.90	93.3
T5d	4.00	4.16	104.0
T5e	4.00	3.80	95.0
T5f	4.18 ^a	3.71	88.8
T6a	4.00	4.06	101.5
T6b	4.00	3.95	98.8
T6c	4.00	4.15	103.8

^a Due to an insufficient supply of test chemical, for three plots (T5a, T5c, T5f) the spray solution concentration was lower (25.10 g a.s./l). Accordingly, a higher volume (4.18 l) was applied per plot.; T1a-T1c = plots treated with test chemical (0.6 kg a.s./ha); T2a-T2f = plots treated with test chemical (1.8 kg a.s./ha); T3a-T3c = plots treated with test chemical (3.2 kg a.s./ha); T4a-T4c = plots treated with test chemical (5.8 kg a.s./ha); T5a-T5f = plots treated with test chemical (10.5 kg a.s./ha); T6a-T6c = plots treated with test chemical (31.5 kg a.s./ha).

3.1.2.5 Weather conditions during application

Air and soil temperature (approximately 5 cm depth) on the day of application and wind speed during application were measured on site. Precipitation was recorded at the nearest weather station (Raunheim), located approximately 3 km from the trial location, run by the German Weather Service (Deutscher Wetterdienst, DWD). Environmental conditions as recommended

by the ISO guideline 11268-3 (2014) for test chemical application (wind velocity during application <3 m/sec; no rain for at least one hour after finishing the application) were fulfilled. Details are summarised in Table 11.

Table 11: On-site air and soil temperature and wind velocity during spray application

Plot [code]	Time window [hh:mm]	General weather conditions	Mean air temperature [°C]	Mean soil temperature [°C]	Wind velocity [m/sec]	Precipitation [mm]
T1a-c	08:05 – 08:31	Sunny	10.5	8.4	0.2 – 1.6	0.0
T2a-f	08:42 – 09:51				0.7 – 1.6	
T3a-c	10:00 – 10:27				1.8 – 2.3	
T4a-c	10:38 – 11:06				2.2 – 2.8	
T5a-f	11:22 – 12:34				1.4 – 2.8	
T6a-c	12:45 – 13:10				2.1 – 2.6	

3.1.2.6 Irrigation of experimental plots

All experimental plots were irrigated directly after application on 11 April 2017 by means of a tractor-pulled tank wagon. The plots were irrigated with at least 1000 l/plot (equivalent to 10 mm precipitation).

3.1.3 Conditions of the experimental site during study duration

3.1.3.1 Maintenance of experimental plots

The experimental plots were left to natural development of vegetation (Figure 11 to Figure 13). No agricultural practices such as tillage, application of plant protection products or fertilizers, were undertaken. On 25 August 2017 all plots were mowed with a string trimmer (Figure 14) and all cuttings were left on the plots (Figure 15).

Figure 11: Experimental site on 24 May 2017



Source: ECT Oekotoxikologie GmbH

Figure 12: Experimental site on 12 June 2017



Source: ECT Oekotoxikologie GmbH

Figure 13: Experimental site on 25 August 2017 prior to mowing



Source: ECT Oekotoxikologie GmbH

Figure 14: Mowing of the experimental site on 25 August 2017 with a string trimmer



Source: ECT Oekotoxikologie GmbH

Figure 15: Experimental site on 28 August 2017 after mowing



Source: ECT Oekotoxikologie GmbH

Figure 16: Experimental site on 23 April 2018 during the last earthworm sampling



Source: ECT Oekotoxikologie GmbH

3.1.3.2 Weather conditions during the study period

During the trial, air and soil temperature and precipitation were recorded at the nearest weather stations, run by the German Weather Service (Deutscher Wetterdienst, DWD): Wiesbaden-Auringen (temperature) and Raunheim (precipitation), located approximately 13 km and 3 km from the trial location, respectively. Monthly precipitation, monthly mean air temperature with minimum (min) and maximum (max) temperature and monthly mean soil temperature during the experimental phase of the study are summarised in Table 12.

Table 12: Mean, minimum and maximum monthly air temperature, mean monthly soil temperature and monthly cumulated precipitation [mm] during the field trial period (April 2017 – April 2018)

Month year	Air temperature [°C]			Soil temperature [°C]	Precipitation [mm]
	Mean	Min	Max	Mean	Cumulative
April 2017	8.5	-1.9	20.7	11.5	10.6
May 2017	14.6	0.2	30.8	17.7	69.4
June 2017	18.4	6.0	33.1	22.1	28.4
July 2017	18.9	7.0	33.3	21.8	106.1
August 2017	18.0	8.6	28.3	21.3	87.7
September 2017	12.8	3.3	23.8	15.5	64.0
October 2017	10.8	0.5	20.2	11.7	30.9
November 2017	4.9	-2.2	13.6	5.2	75.8
December 2017	2.4	-3.1	11.2	1.7	86.6
January 2018	4.5	-2.8	10.5	3.9	69.4
February 2018	-1.2	-10.7	7.2	0.2	12.5
March 2018	3.6	-8.5	13.5	4.2	45.5
April 2018	13.4	-1.4	26.9	15.1	46.6

Measured at the nearest meteorological station (temperature: Wiesbaden-Auringen; precipitation: Raunheim) of the German weather service (DWD).

3.1.4 Assessment of the earthworm community

3.1.4.1 Sampling of earthworms

Earthworms were sampled at each sampling time point by a combined hand-sorting and AITC extraction method according to ISO 23611-1 Version 11/2005 and Zaborski (2003). Six random samples were taken per plot. Hence, there were 18 (3 plot replicates) or 36 (6 plot replicates) individual samples per treatment and sampling time point. The soil of an area of 0.25 m² (50 cm x 50 cm) was excavated by means of a spade to a depth of approximately 20 cm and placed in a large bucket (Figure 17). The distance between two samples taken on the same date and plot was at least 2 m. The sampled area was clearly marked with a blue stick and was not used again at subsequent sampling dates. Samples were taken at least 2 m apart from the plot border. Five to ten litres of an AITC solution (0.1 g/l) were poured uniformly into the remaining cavity in

order to catch earthworms from deeper soil layers. The soil in the bucket was carefully hand-sorted and searched for earthworms (Figure 18). All earthworms sampled by hand-sorting and AITC extraction were preserved in a 70% ethanol solution in watertight containers (Figure 19).



Figure 17: Buckets containing soil for hand-sorting and watering cans containing AITC solution

Source: ECT Oekotoxikologie GmbH

Figure 18: Hand-sorting and AITC-extraction of earthworms



Source: ECT Oekotoxikologie GmbH

Figure 19: Sampling vessel containing 70% ethanol and earthworms



Source: ECT Oekotoxikologie GmbH

Air and soil temperature, soil moisture and general weather conditions at the four earthworm sampling dates of the study were recorded and are summarized in Table 13.

Table 13: Air and soil temperature, soil moisture and general weather conditions at the four earthworm sampling dates of the study

Sampling	Dates	Time point	Mean air temperature [°C]	Mean soil temperature [°C]	Mean soil moisture [% dw]	General weather conditions
1 st sampling	03-05 Apr 2017	8-6 DBA	10.2 – 14.4	9.7 – 10.7	0-12 cm: 17.4 12-24 cm: 18.1	sunny
2 nd sampling	15-17 May 2017	34-36 DAA	17.1 – 19.7	13.7 – 15.8	0-5 cm: 10.9 5-10 cm: 17.2	sunny
3 rd sampling	16-18 Oct 2017	188-190 DAA	14.9 – 16.0	12.5 – 13.0	0-5 cm: 20.0 5-10 cm: 18.8	sunny
4 th sampling	23-25 Apr 2018	377-379 DAA	16.0 – 20.2	15.9 – 16.8	n.d.	sunny/cloudy

DBA = days before first test chemical application; DAA = days after test chemical application. n.d. = not determined.

3.1.4.2 Identification of earthworm species

The worms were identified by means of a binocular microscope, using external characters (mainly the form of the prostomium, the distribution of the setae, the form and place of the clitellum and the tubercula pubertatis). The determination was performed according to Graff (1953), Sims & Gerard (1985) and Bouché (1972). The nomenclature follows Sims & Gerard (1985). Adult worms were determined to the species level. Juveniles were classified according to the genus level, but in some cases a distinction of small worms belonging to closely related genera was not possible (e.g. *Allolobophora* and *Aporrectodea* were combined).

3.1.4.3 Weighing of earthworms

All adult worms of one sample belonging to a particular species and all juvenile worms belonging to a particular genus were weighed together. Before weighing, the specimens were briefly dried on a piece of tissue. Afterwards, the worms were transferred back to the vessels containing ethanol.

3.1.5 Chronology of the study

The chronology of the study is summarized in Table 14.

Table 14: Chronology of the study

Date	Time point	Action
16 Mar 2017	26 DBA	Application of glyphosate.
28 Mar 2017	14 DBA	Preparation of experimental field (plots marked)
03-05 Apr 2017	8-6 DBA	Earthworm sampling (pre-application)
11 Apr 2017	0 DAA	Application of test chemical and control (water); irrigation of the experimental field
15-17 May 2017	34-36 DAA	Earthworm sampling (first post application)
25 Aug 2017	136 DAA	Mowing the vegetation on all plots
16-18 Oct 2017	188-190 DAA	Earthworm sampling (second post application)
23-25 Apr 2018	377-379 DAA	Earthworm sampling (third post application)

DBA = days before first test chemical application; DAA = days after test chemical application.

3.1.6 Results of the study

3.1.6.1 Species diversity of earthworms

The field site was inhabited by an earthworm community which can be considered typical for central European arable land (ISO 11268-3; 2014) including the groups of anecic and endogeic earthworms as required by the ISO guideline. In total, nine different species of earthworms were found during the study (Table 15).

During the trial, the lumbricid biocoenosis was dominated by juveniles of the endogeic genera *Aporrectodea/Allolobophora* (by number and biomass). *Allolobophora chlorotica* was the most abundant species.

Table 15: Earthworm species found during the pilot field study across both treated and untreated plots

Genus	Species	Ecological group	Morphological group	8-6 DBA	34-36 DAA	188-190 DAA	377-379 DAA
<i>Allolobophora</i>	<i>chlorotica</i>	endogeic	epilobous	X	X	X	X
<i>Aporrectodea</i>	<i>caliginosa</i>	endogeic	epilobous	X	X	X	X
<i>Aporrectodea</i>	<i>longa</i>	anecic	epilobous	X	X	X	X
<i>Aporrectodea</i>	<i>rosea</i>	endogeic	epilobous	X	X	X	X
<i>Lumbricus</i>	<i>castaneus</i>	epigeic	tanylobous	X	X	X	---
<i>Lumbricus</i>	<i>terrestris</i>	anecic	tanylobous	X	X	X	X
<i>Octolasion</i>	<i>cyaneum</i>	endogeic	epilobous	X	X	X	X
<i>Octolasion</i>	<i>tyrtaeum</i>	endogeic	epilobous	---	---	X	---
<i>Proctodrilus</i>	<i>antipae</i>	endogeic	epilobous	X	X	---	---

X = present; --- = absent; DBA = days before the first application; DAA = days after application.

3.1.6.2 Abundance and biomass of earthworms before application

Eight to six days prior to the first application of the test chemical, earthworms were sampled on all plots. The mean total number and the mean biomass of earthworms was determined for each of the thirty plots, designated either for test chemical treatment (“test chemical plots”) or to serve as untreated controls (“control plots”).

The mean number of earthworms collected (hand-sorting and AITC-extraction) before application ranged from 413 to 512 ind./m², hence fulfilling the requirements of the ISO guideline 11268-3 (2014). Please see also result tables in the appendix A.1 to this report.

3.1.6.3 Effects of the test chemical

The test chemical Carbomax 500 SC (a.s. carbendazim) caused a clear reduction in total abundance and biomass at all three post application sampling time points (Table 16, Figure 20, Figure 21). Compared to the control, mean abundance and mean biomass in the test chemical treated plots were 15-59% and 11-55%, respectively at 34-36 DAA, 45-90% and 69-111%, respectively at 188-190 DAA, and 38-74% and 80-113% respectively at 377-379 DAA (Table 17).

Table 16: Mean abundance [ind/m²] and biomass fresh weight [g/m²] of total earthworms (adults and juveniles) during the pilot field study (± standard deviation). T1 – T6: treatment rates with carbendazim (kg a.s./ha)

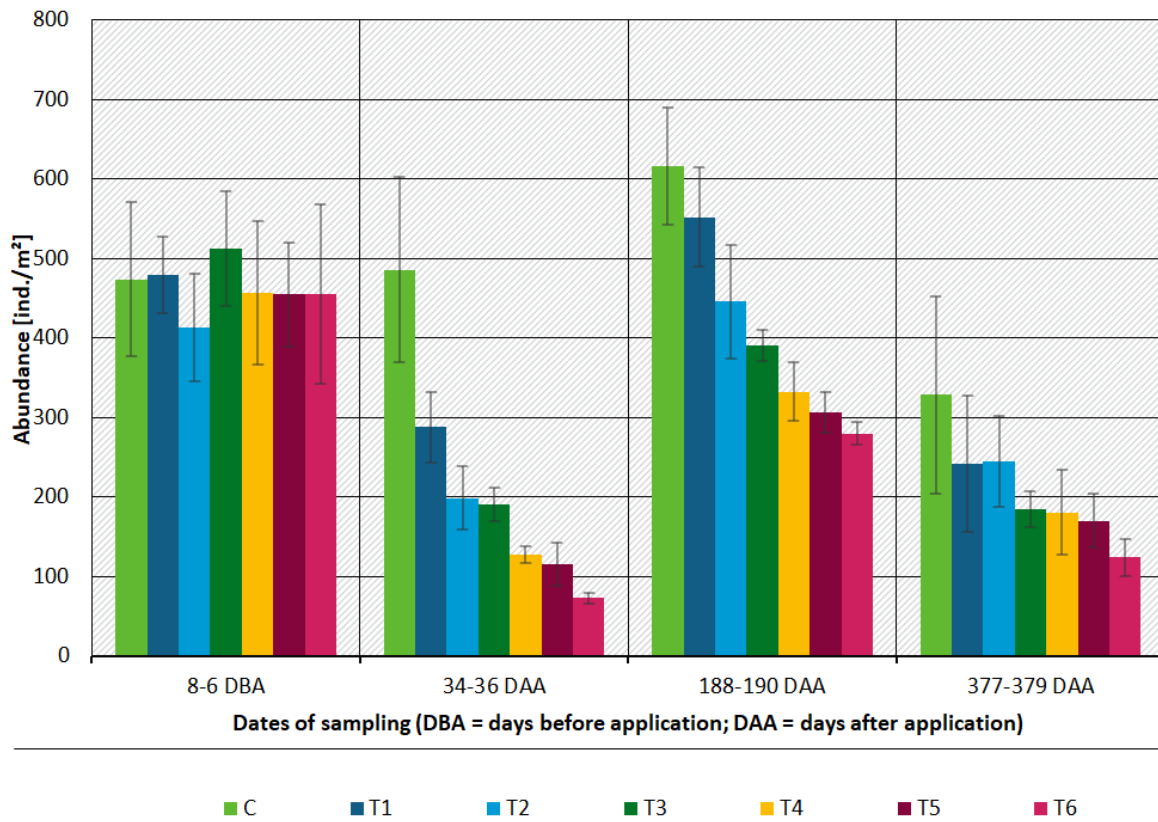
Time point	Control	T1 (0.6)	T2 (1.8)	T3 (3.2)	T4 (5.8)	T5 (10.5)	T6 (31.5)
Abundance [ind/m²]							
8-6 DBA	474 ± 97.5	479 ± 48.0	413 ± 67.5	512 ± 72.2	457 ± 90.6	454 ± 65.7	456 ± 112.9
34-36 DAA	486 ± 117	288 ± 44.7	198 ± 39.6	191 ± 21.4	127 ± 10.2	115 ± 27.4	72.7 ± 6.7
188-190 DAA	616 ± 73.9	552 ± 62.4	445 ± 71.9	390 ± 19.7	332 ± 37.1	306 ± 24.9	280 ± 13.9
377-379 DAA	328 ± 124	241 ± 85.2	244 ± 56.8	185 ± 22.7	180 ± 53.1	170 ± 34.1	124 ± 23.4
Biomass [g/m²]							
8-6 DBA	127 ± 21.6	117 ± 11.0	102.1 ± 18.4	126 ± 26.4	99.3 ± 13.5	98.9 ± 10.5	106 ± 1.1
34-36 DAA	93.3 ± 12.6	51.0 ± 16.8	36.4 ± 4.4	37.6 ± 7.0	23.9 ± 3.0	17.8 ± 7.8	10.3 ± 2.3
188-190 DAA	177 ± 30.8	196 ± 27.5	154 ± 31.4	143 ± 7.8	123 ± 15.7	122 ± 8.9	125 ± 11.1
377-379 DAA	73.4 ± 20.9	82.8 ± 21.3	80.1 ± 19.3	60.2 ± 8.2	68.7 ± 37.8	71.6 ± 14.7	58.4 ± 1.3

Table 17: Abundance and biomass [% of control] of total earthworms (adults and juveniles) during the pilot field study. T1 – T6: treatment rates with carbendazim (kg a.s./ha)

Time point	Control	T1 (0.6)	T2 (1.8)	T3 (3.2)	T4 (5.8)	T5 (10.5)	T6 (31.5)
Abundance [% of control]							
8-6 DBA	100	101	87	108	96	96	96
34-36 DAA	100	59	41	39	26	24	15
188-190 DAA	100	90	72	63	54	50	45
377-379 DAA	100	74	74	56	55	52	38
Biomass [% of control]							
8-6 DBA	100	92	80	99	78	78	84
34-36 DAA	100	55	39	40	26	19	11
188-190 DAA	100	111	87	81	70	69	70
377-379 DAA	100	113	109	82	94	98	80

Figure 20: Total earthworms abundance [ind./m²] during the pilot field test . C = control; T1 - T6: treatment rates with carbendazim (T1 = 0.6, T2 = 1.8, T3 = 3.2, T4 = 5.8, T5= 10.5, T6 = 31.5 kg a.s./ha)

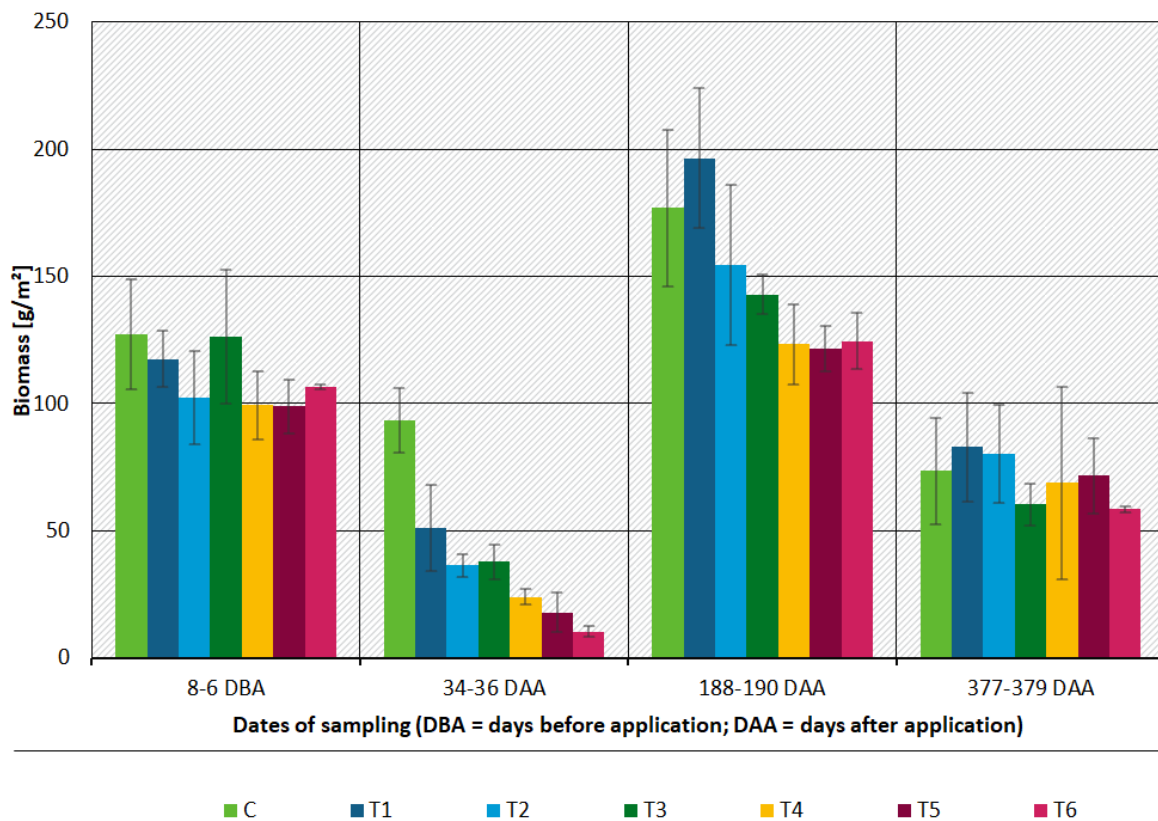
Total earthworms abundance [ind./m²] during the pilot field study



Source: ECT Oekotoxikologie GmbH

Figure 21: Total earthworms biomass [g/m²] during the pilot field test. C = control; T1 - T6: treatment rates with carbendazim (T1 = 0.6, T2 = 1.8, T3 = 3.2, T4 = 5.8, T5 = 10.5, T6 = 31.5 kg a.s./ha)

Total earthworms biomass [g/m²] during the pilot field study



Source: ECT Oekotoxikologie GmbH

3.2 Statistical analysis: field study and database

A set of different statistical data analysis procedures were conducted for both data of the pilot study and test data of an existing UBA database. We primarily focus on an improvement of the conventional statistical methods to evaluate earthworm field studies (ISO 11268-3, 2014) and to acquire insights for statistical considerations regarding an adapted test design for earthworm field studies. Therefore, the state-of-the-art of recently used statistical tests for comparable ecotoxicity data is reviewed (chapter 3.2.1), data characteristics, storage and processing is presented (chapter 3.2.2) and the results of statistical analyses are shown in chapter 3.2.3. Afterwards the design requirements for earthworm field tests, derived from statistical analyses, are summarized (3.2.4) and limitations are discussed (3.2.5).

3.2.1 State of the art of statistical procedures to analyse ecotoxicological field tests

The NOEC (No Observed Effect Concentration, aka NOEL or NOER) represents the concentration (aka level or rate) of a chemical at which no statistically significant effects were observed in the specifically assessed experimental set-up on the endpoint evaluated (e.g. species abundance). NOECs have intensively been used as risk assessment endpoints due to their ease of computation and the possibility to provide an easily justifiable risk threshold. In higher-tier experiments,

a NOEC design has the advantage that the choice of test concentrations is simplified in the way that at least only the intended application rate may be tested (limit test).

According to the latest guidance document on determination of effects in field situations for earthworms (ISO 2014), NOECs are to be deduced from an analysis of variance to determine differences between treatments (Landis & Chapman 2011). The test is followed by a multiple comparison post-hoc test against a control for randomized complete block design including a correction for alpha-inflation (Dunnett-test or Williams-test, $\alpha=0.05$, one-sided; Dunnett 1955, 1964; Williams 1971, 1972) which is in fact a t-distribution based analysis of variance. Also, other guidelines (e.g. OECD 2006b) recommend pairwise comparison of fractional responses between each treatment (concentration or rate) and the control with Dunnett's multiple comparison test for NOEC calculation. The lowest exposure that is not statistically different from the control is reported as NOEC (or NOEL or NOER). To test the necessary assumptions of normality and variance homogeneity, Shapiro-Wilks (Shapiro 1965) and Levene's test (Levene 1960, Brown & Forsythe 1974) are recommended, respectively. If data do not fulfil the criteria, they are allowed to be transformed (logarithmic or square-root, respectively, ISO 2014). For quantal responses, the application of arc-sine square root transformation is recommended in order to stabilize variabilities and to make the distribution closer to the normal (OECD 2006a). Alternatively, generalized linear models (Nelder & Wedderburn 1972) or non-parametric tests, e.g. Bonferroni U-test (Holm 1979) or Jonckheere-Terpstra Step-down-test (Jonckheere 1954) can be used. Software to perform NOEC calculation is widely available (e.g. GraphPad, SigmaPlot, R, ToxRat). Detailed recommendations about how to transform data and which test to use can be found in the OECD test guidelines (e.g. OECD 2006a). In addition to uni-variate methods, multi-variate statistical tools, such as PRC (Principal response curves; van den Brink & ter Braak 1998), can be helpful in the interpretation of study results.

NOEC and related concepts have long been criticized in ecotoxicological literature (Laskowski 1995; Koijman 1996; Walter et al. 2002; Warne & van Dam 2008; Jager 2011, 2012; Landis & Chapman 2011; Fox & Landis 2016; Tanaka et al. 2018). Landis & Chapman (2011) pointed out that the NOEC is flawed due to several reasons. (1) It "ignores critical data". Only a small subset of the data is used and evidence of an effect from lower or higher concentrations is not considered. (2) It "uses a lack of evidence as no-effect" as they perform null hypothesis testing. Not rejecting the null hypothesis can e.g. simply result from a badly replicated experiment. (3) It is "inconsistent between studies". NOECs depend on the experimental design which is often specific for a certain study. (4) It is "not associated with any measure of uncertainty" (like standard deviation).

In a thorough simulation study, Tanaka et al. (2018) recently proved that the NOEC performs significantly worse compared to other ecotoxicological statistics (especially ECx-values) when the coefficient of variation in responses between replicates in treatments or in the control was larger than 10%. Laskowski (1995) pointed out that NOECs strongly depend on the experimental design and the number of replicates used in the statistical test and that a large type-II-error might be hidden within the resulting NOEC value. NOECs also depend on the power of the statistical test which decreases with smaller sample size and a reduced number of test concentrations. Also, larger variation in the experimental data decreases the power of the null hypothesis test. Thus, the NOEC may be underestimated by weak testing power due to inappropriate design of experiments (Tanaka et al. 2018). It is said to reward bad experiments (Fox 2009). There have been some approaches to circumvent these shortcomings (Green et al. 2012). Recent testing guidelines development demands at least 75% power (less than 25% type-II-error) (OECD 2012). For regulatory sciences, it is generally recommended that type-II-errors (β -errors) "should be of greater concern than type-I-error in regulatory sciences ... because the decision for

protection of the environment must be biased ... towards safety rather than certainty of positive results” (Tanaka et al. 2018).

In a mixture toxicity study, Walter et al. (2002) showed that NOEC values of single substances are uninformative especially in the case of simultaneous presence of multiple toxic substances in low concentrations with different modes of action. Observed mixture toxicity was clearly higher than can be expected from single substance toxicities of substances at NOEC concentration level. In this mixture toxicity context, NOECs of single substances are judged as “no safe guard against unwanted toxicity from mixtures”. Landis & Chapman (2011) pointed out that the NOEC “does not meet the criterion of adequately describing the exposure-response curve”. They “advocate adoption of curve-fitting as the standard interpretation of laboratory test data and urge rejection of the NOEC approach”. Despite the criticism about NOEC, it is still common practice in regulatory contexts and scientific publications (Jager 2011; Landis & Chapman 2011).

Reproduction data (raw data as well as quantal data calculated as fractions of integer numbers $[0, \infty]$) are generally not assumed to be normally distributed but Poisson distributed (Szoecs & Schafer 2015; Delignette-Muller et al. 2014; Chapman et al. 1996). There are several reasons why theory limits the usage of statistics based on normal distribution and variance homogeneity (like Dunnett-test or Williams-test) in the case of reproduction data: (1) Reproduction data are discrete, whereas the normal distribution is continuous. Approximating a discrete distribution with a continuous distribution can lead to conclusions which are not in accordance with the information content of the data (sometimes a continuity correction can help). (2) The lower limit of reproduction is 0, whereas the normal distribution is an asymptotic distribution in both directions. This can for example lead to wrong estimations of confidence intervals. (3) Variance under the Poisson model is always equal to the mean. Thus, decreasing mean values with increasing concentrations (the normal case for toxicological effect, especially in acute toxicity) inevitably result in decreasing variances and variance homogeneity must be rejected. Besides Dunnett- and Williams-test, it should be noted that also the alternatives of Jonckheere-Terpstra (Jonckheere 1954) and Kruskal-Wallis test (Kruskal & Wallis 1952) suffer from such inhomogeneous variances (Lehmann et al. 2016).

Although one can test experimental count data on normal distribution and variance homogeneity, subsequent application of Dunnett-test or Williams-test is only an approximation of the true Poisson distribution (see the central limit theorem; Dudley 2014) and gives only approximative statistics valid under the according assumptions ($E(X) \geq 5$; Gupta & Guttman 2014). The possibility for generalization is limited, because reproduction data can appear to be normally distributed by pure chance in a special experiment. Transformation of raw data towards normal distribution or variance homogeneity also results only in an approximation of the theoretical Poisson distribution. Above these limitations, it must be remembered that tests on normal distribution and variance homogeneity always favour the null hypothesis which is rewarded by small sample sizes, that strongly enhance the type-II-error for these tests.

Besides normal distribution and variance homogeneity, the Williams-test assumes a monotonic trend in the effects as a theoretical a priori assumption. A general fault is to test the data for monotonicity using a statistical test. It has to be remembered that an empirically found monotonic trend in the test data is not a valid argument for the application of the Williams test.

Lehmann et al. (2016) pointed out that NOECs obtained from t-test statistics must always be questioned and recommend CPCAT (Closure Principle Computational Approach test) as a non-approximative Poisson-based test on differences against a control. This test does not require any assumptions about normality or variance homogeneity as the t-based Dunnett- or Williams-test

and avoids the problem of alpha-inflation and inappropriate data transformations. It is strongly recommended to replace the t-test especially in OECD guidelines.

Up to now, CPCAT is not demanded in any testing guidelines, however it is planned to include CPCAT in the future. Although the R-scripts to calculate CPCAT are freely available, it is not yet available as an easy to use function in any of the statistic software packages mentioned. Up to now, there is no CPCAT available that appropriately takes into account overdispersion in count data (a case of so-called generalized Poisson distribution). If the variance within the data is significantly higher than the mean (so-called overdispersed count data, can be tested using the Hampel identifier; Hampel et al. 2005) Lehmann et al. (2018a) showed that the statistical power of CPCAT might be reduced and approaches that of other statistical tests.

In chronic toxicity tests, regression approaches have been suggested as an alternative to the NOEC approach since decades (Stephan & Rogers 1985). These methods are much more robust than analysis of variance in terms of violation of assumptions and experimental variability and allow to estimate the effect level for any concentration as well as confidence intervals. The resulting estimated concentrations (EC_x) from fitting a curve to the data have been suggested as an alternative to the NOEC-value (e.g. Fox 2009). Landis & Chapman (2011) point out that “curve-fitting approaches can use all of the data, can express the uncertainty of the data and the model and provide information on the slope of the response”. Additionally, “because replication at each exposure is not required, a broader range of exposure-response interactions can be observed at the same level of effort”. It was shown that NOECs from a set of laboratory data (lacking any information about uncertainty) typically respond to an EC₁₀ to EC₃₀ on a dose-response curve (Moore & Caux 1997), thereby hiding a 10 to 30% effect. Landis & Chapman (2011) proposed to follow three principles in dose-response modelling: (1) To establish dose-response relationships curve-fitting should preferably be used, (2) the calculation of confidence or credibility intervals should be included, and (3) findings based on NOECs (including SSDs) should be scrutinized. They also requested regulatory agencies to remove statistical hypothesis tests like the NOEC approach for the reporting of exposure-response from their guidance documents and refer to the corresponding actions from the US EPA (Crump 1995). This has been taken up e.g. in the data requirements in place for active substances and plant protection products since 2013, but only regarding laboratory data (EC 2013a+b).

Modelling dose-response data is e.g. described in ISO (2014) and in great detail in OECD (2006a). Depending on the variable scale, different models can be chosen (mainly Logit, Probit, Weibull, linear, Hill etc.) to produce a sigmoidal response curve that can be used to derive effective concentrations (like EC₁₀, EC₂₀, EC₅₀) and their confidence limits. For complex experimental designs and toxicological effects, more flexible models can be used (e.g. Hormesis, additive, nested or GLM). Software to perform EC_x curve fitting is widely available (e.g. GraphPad, SigmaPlot, R, ToxRat) and even Bayesian approaches are available to incorporate a priori information about the uncertainty of the estimated effect concentrations (Fox 2010).

Although EC_x approaches avoid some of the shortcomings of NOEC approaches, they are not without critics. Some authors present justified criticism against the use of EC_x as a measure of toxicity (e.g. Koijman 1996; Jager 2011; Green et al. 2012).

Jager (2011) presents several reasons why also EC_x are limited as a measure of toxicity, especially with respect to extrapolation from the actual experimental conditions to more general conditions and the comparability from different studies. Some of these aspects are also relevant for NOEC calculations. (1) EC_x/LC_x depend on exposure time and constancy. For a longer exposure, the values can decrease (e.g. acute toxicity) or even increase (e.g. chronic toxicity). (2) EC_x/LC_x depend on the choice of endpoint (e.g. body volume vs. reproduction rate) and the

measure chosen to quantify it (e.g. body volume vs. body length). (3) EC_x depends strongly on the standardized experimental conditions. As a statistical approach, EC_x cannot be used to extrapolate to other conditions than the ones from the experiment or to learn anything about the mechanism of toxicity. (4) Also, the usefulness of compiling EC_x values from different species in comparative approaches like SSDs or QSARs and the value of databases is questioned. As an alternative, a mechanistic modelling approach incorporating the dynamic aspects of toxicity is asked for. According to the methods used in fate modelling, the TKTD approach (toxicokinetic-toxicodynamic, e.g. Ashauer & Escher 2010) is suggested. Green et al. (2012) point out that effect concentrations from EC_x experiments are only useful if scientists could agree on the value of *x* for each study type, species, or endpoint.

Koijman (1996) criticizes the often-used log-logistic or log-probit curve, characterized by 50% point estimate (LC₅₀) and a maximum slope (gradient parameter). They point out that choosing log-logistic or log-probit regression is arbitrary and state three major problems: (1) The smaller the effect level, the larger the confidence interval. (2) Data points with full or no mortality have to be excluded due to mathematics. In this case, the estimation of small effect concentrations has to be an extrapolation. One can use maximum likelihood estimation of regression parameters on untransformed data to avoid this. (3) Log-logistic regression predicts smaller effects in case of small concentrations when exposure time is elongated. Koijman (1996) suggests to replace NOEC with the model-based NEC (No effect concentration). In this concept, null hypothesis and alternative hypothesis are exchanged. Not rejecting the null hypothesis (=NEC is zero) “leads to the safe conclusion that each molecule of the compound might have an effect”. Unfortunately, “the NEC cannot be built into standard response models ... but has to be replaced by mechanistically underpinned models” (e.g. Dynamic Energy Budget model; Koijman 1993).

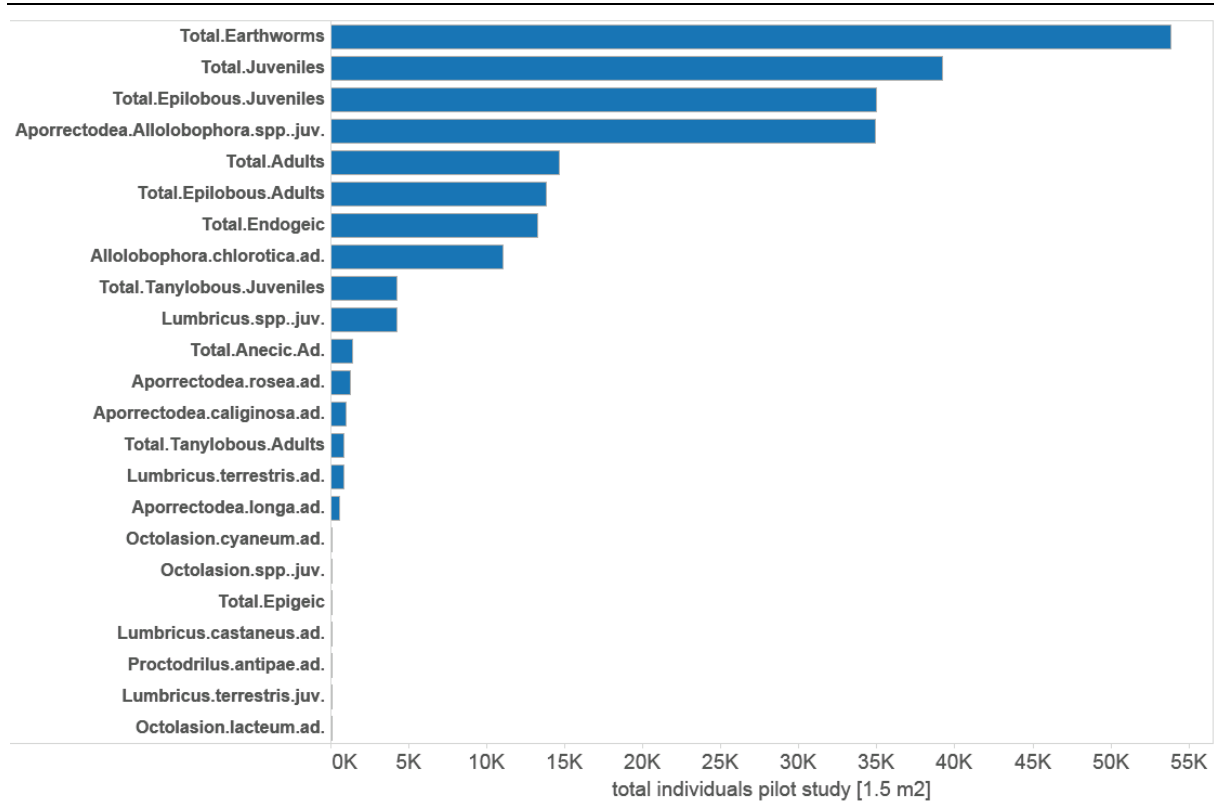
Principal response curves (PRC; van den Brink & ter Braak 1998, 1999) have been developed as a multivariate technique to evaluate community treatment effects resulting from complex higher tier experiments under field or semi-field conditions (Maund et al. 1999). They have been used in many ecological and ecotoxicological studies to test whether there is a significant relationship between community composition and the treatment applied. Although PRCs have proven to be a powerful tool to show effects on interacting communities that can often not be elucidated by univariate methods (e.g. van den Brink et al. 2009; Moser et al. 2007; Heegaard & Vandvik 2004; van den Brink et al. 2003; Frampton et al. 2000; Kedwards et al. 1999) and have been judged as eventually helpful (ISO 2014), they are not included as standard tools in guidelines yet.

3.2.2 Data description of the pilot study

The raw data generated in the pilot field study with earthworms regarding biomass and abundance at all times of sampling and for all taxa and morphological or functional earthworm groups were integrated at sample- and plot level into the existing database of the project (chapter 2).

An overview of the total sampled individuals per classified earthworm group during the pilot field study is given in Figure 22. This is the underlying pilot field study dataset for the following statistical analyses. Please note that these are not distinct classifications for the single individuals. Earthworms are categorized into the respective single species group (e.g. *Allolobophora chlorotica*, *Aporrectodea longa*, *Lumbricus terrestris*) as well as into aggregated cluster (“total earthworms” etc.) and into morphological or functional groups (e.g. “total anecic adult”, “total tanylobous juveniles” etc.). The few undetermined individuals were not included into the calculations.

Figure 22: Overview of sampled total individuals per taxonomic/morphological group or earthworm species in the performed pilot field study



Note: *Octolasion lacteum* is a synonym of the valid species *O. tyrtaeum*.

Source: RWTH Aachen University

Mainly juvenile representatives of the group *Aporrectodea/Allolobophora* (34,927 individuals) were determined in the pilot study, adult individuals are dominated by the endogeic species *Allolobophora chlorotica* (11,000 ind.). *Aporrectodea caliginosa* (959 ind.) was found in comparable abundances as *Aporrectodea rosea* (1,230 ind.). The anecic species *Aporrectodea longa*, another representative of this genus, was identified only less frequently (573 ind.).

Additional information on abundance and biomass data, as well as statistical calculations for all identified morphological and taxonomic groups can be found in the respective fact sheets, Appendix A.2. The fact sheets include descriptive boxplots over time, dose-response curves (linear Probit regression) for the tested chemical carbendazim, visual illustrations of derived EC₁₀ and EC₅₀-values for earthworms exposed to carbendazim, as well as the calculated NOEC values using the Dunnett and CPCAT approach.

Analyses of endpoints for single species do not show any statistically significant effects at the last sampling time point (377-379 DAA). For the endpoint "*Aporrectodea/Allolobophora* spp. juvenile", a statistically significant effect of the test chemical carbendazim can be observed in a taxonomic group after one year. Due to the high dominance of this group in the overall data set, a reduction of abundance and biomass after 12 months is also indicated in other aggregated groups such as "total earthworms" or "total epilobous juveniles". However, this is exclusively caused by the effects on juveniles of *Aporrectodea/Allolobophora* spp. This example illustrates the need for assessments of different types of endpoints and earthworm groups (e.g. species level and group level), in order to avoid general conclusions for effects of test substances based on single endpoints.

3.2.3 Advanced statistical procedures - database and pilot field study

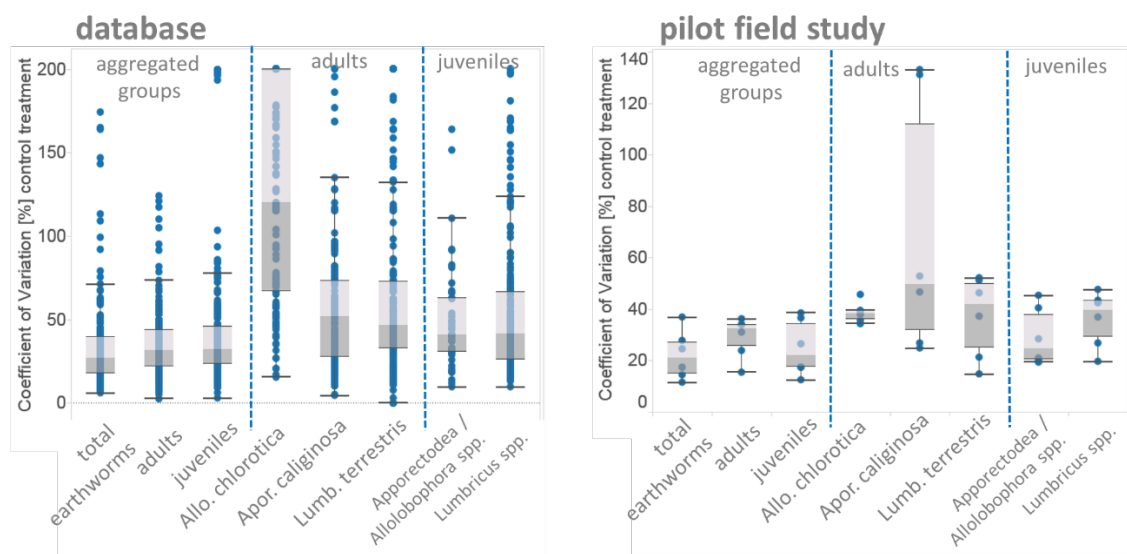
3.2.3.1 Analysis of natural variability in earthworm communities

The natural, heterogeneous scattering of earthworm species within a field is a decisive factor for the statistical visibility of possible effects caused by applied chemicals in the environment. Within the setup of a field study, this variability can be expressed by dispersion measures for untreated control treatments, such as the coefficient of variation. In the following chapter, the variability of tested endpoints in database and pilot field studies is assessed using the coefficient of variation of field study control treatments. Results were used to derive conclusions and suggestions for improvement regarding the test power of the database and pilot study setup. This is done using a sample size modelling approach (chapter 3.2.3.1.3)

3.2.3.1.1 Coefficient of variation within control treatments

In a subsequent assessment step, the natural variability of the species groups in field studies was illustrated descriptively as the variance of the control treatments and used as a basis for multiple sample planning. The coefficient of variation results from the quotient of standard deviation and arithmetic mean (in percent). An overview for the respective distributions of the most dominant earthworm groups in field studies is given in Figure 23.

Figure 23: Distribution of coefficients of variation for control treatments (pilot study and database) on plot level (1.0 m² for database studies and 1.5 m² for pilot study) for earthworm biomass and abundance data at all tested times of sampling



Source: RWTH Aachen University

The presented boxplots combine data for both endpoint measures, biomass and abundance data. Aggregated earthworm groups “total earthworms”, “adults” and “juveniles” have the lowest coefficients of variation for database studies and for the pilot field study compared to the most dominant species for adults (*A. chlorotica*, *A. caliginosa*, *L. terrestris*) and for juveniles (*A. caliginosa* / *A. spp.* and *L. terrestris*). The lowest coefficient of variation was observed for the aggregated group “total earthworms”.

A comparison between the two types of data (available database studies and pilot field study) shows, that the calculated coefficients of variation are at a comparable level for both types of field studies. The mean dispersion measures in the pilot study generally appear to be slightly

lower, especially for aggregated groups ("total earthworms", "total adults", etc.). This might be a consequence of the larger sampling area (1.5 m² instead of 1.0 m²) and therefore slightly higher mean abundance and biomasses (see Figure 24).

The lowest coefficient of variation within the pilot study was measured for 'total earthworm abundance' data at the 2nd sampling after application (188 DAA). The coefficient of variation was calculated to be 11.5% at this sampling point. The mean for all sampling time points was 32.9% for both endpoint measures abundance and biomass of total earthworms in control plots of the database field studies. Large deviations from the calculated values in the database with exceptionally high coefficients of variation typically only occurred if the underlying expected value (mean value) was particularly low, usually for single findings of individuals or similar. Mean coefficients of variation of control treatments for all recorded species groups are shown on plot and sample (= subplot) levels for the pilot study and the additional database studies in Table 18.

Table 18: Mean coefficients of variation for different endpoints from control treatments in earthworm field studies (pilot study and database, mean of all sampling time points) on plot level (1.0 m² for database studies and 1.5 m² for pilot study) and sample level (0.25 m²)

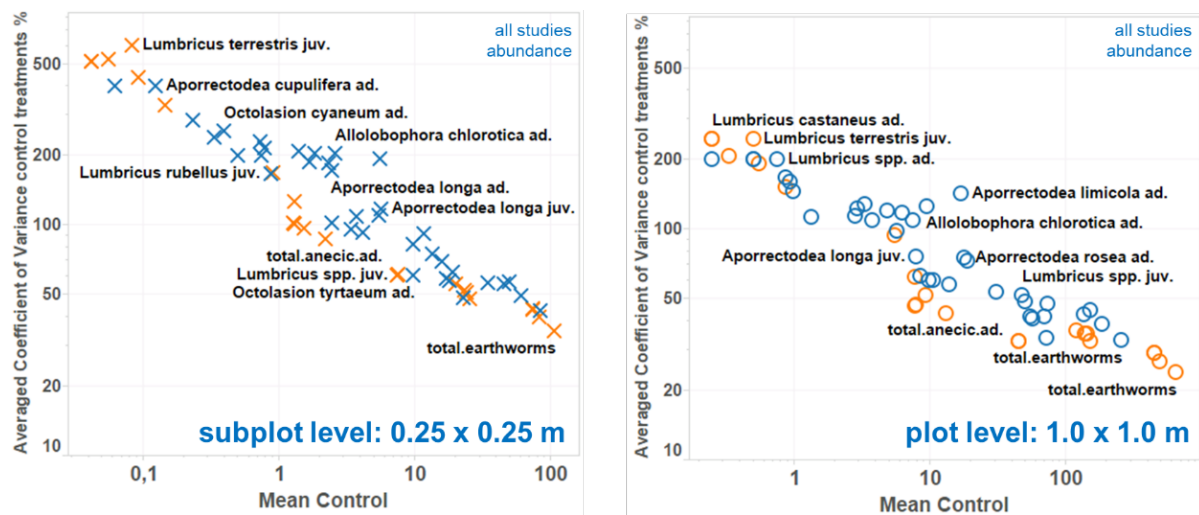
Level	Mean coefficients of variation for control treatments [%]							
	Plot				Sample (= subplot)			
	Database [1.0 m ²]		Pilot study [1.5 m ²]		Database [0.25 m ²]		Pilot study [0.25 m ²]	
Statistical measure	Abundance	Bio-mass	Abundance	Bio-mass	Abundance	Bio-mass	Abundance	Bio-mass
<i>Allolobophora chlorotica</i> adult	125.8	123.4	36.4	38.5	203.8	198.2	55.0	58.6
<i>Allolobophora chlorotica</i> juvenile	127.6	122.0	NA	NA	207.8	207.6	NA	NA
<i>Aporrectodea Allolobophora</i> spp. adult	109.1	117.0	NA	NA	109.1	117.0	NA	NA
<i>Aporrectodea Allolobophora</i> spp. juvenile	47.4	52.6	29.1	26.9	62.2	70.3	43.1	37.3
<i>Aporrectodea caliginosa</i> adult	53.2	56.6	61.9	59.0	82.1	85.5	126.2	130.1
<i>Aporrectodea caliginosa</i> juvenile	41.9	42.5	NA	NA	56.0	58.8	NA	NA
<i>Aporrectodea cupulifera</i> adult	200.0	200.0	NA	NA	400.0	400.0	NA	NA
<i>Aporrectodea limicola</i> adult	142.2	154.5	NA	NA	192.8	232.2	NA	NA
<i>Aporrectodea longa</i> adult	98.2	101.9	93.5	98.9	171.8	180.8	167.6	176.5
<i>Aporrectodea longa</i> juvenile	75.6	74.7	NA	NA	117.4	118.0	NA	NA

	Mean coefficients of variation for control treatments [%]							
<i>Aporrectodea rosea</i> adult	74.8	71.0	51.6	54.6	109.7	110.5	96.4	101.4
<i>Eisenia fetida</i> adult	200.0	200.0	NA	NA	400.0	400.0	NA	NA
epilobous juveniles	44.4	50.9	NA	NA	56.6	66.3	NA	NA
<i>Lumbricus castaneus</i> adult	117.2	119.2	244.9	244.9	204.2	207.9	509.1	519.3
<i>Lumbricus rubellus</i> adult	119.6	116.9	NA	NA	186.9	183.3	NA	NA
<i>Lumbricus rubellus</i> juvenile	166.7	178.0	NA	NA	166.7	178.0	NA	NA
<i>Lumbricus</i> spp. adult	200.0	200.0	NA	NA	200.0	200.0	NA	NA
<i>Lumbricus</i> spp. juvenile	51.6	57.7	32.7	41.0	74.7	85.3	60.4	75.4
<i>Lumbricus terrestris</i> adult	62.8	64.5	46.3	36.4	101.7	102.8	102.2	95.7
<i>Lumbricus terrestris</i> juvenile	72.6	75.9	244.9	244.9	91.6	94.6	600.0	600.0
<i>Murchieona minuscula</i> adult	114.3	119.7	NA	NA	215.1	220.5	NA	NA
<i>Octolasion cyaneum</i> adult	145.6	146.0	151.5	153.5	256.0	258.7	330.2	342.9
<i>Octolasion cyaneum</i> juvenile	200.0	200.0	NA	NA	200.0	200.0	NA	NA
<i>Octolasion</i> spp. juvenile	112.2	119.5	193.2	195.8	240.1	257.4	434.0	436.3
<i>Octolasion tyrtaeum</i> adult	60.1	64.7	NA	NA	60.1	64.7	NA	NA
<i>Proctodrilus antipae</i> adult	161.1	149.5	206.1	200.3	284.6	283.1	521.0	509.8
total epigeic adults	109.2	112.6	244.9	244.9	186.0	189.6	509.1	519.3
total endogeic adults	41.8	43.8	35.3	36.9	58.5	63.2	51.7	57.4
total anecic adults	60.1	63.4	43.2	35.8	95.5	98.2	86.3	83.5
total epilobous juveniles	42.7	48.4	29.1	27.3	55.1	63.8	43.0	37.2
total epilobous adults	40.6	43.0	35.0	37.2	57.3	63.1	50.9	56.9
total tanylobous juveniles	48.6	56.3	32.5	43.2	69.5	80.1	60.6	87.4
total tanylobous adults	57.2	57.7	46.9	36.4	92.6	93.7	100.9	95.7
total juveniles	38.5	44.3	26.7	22.2	49.1	59.1	39.7	37.1
total adults	33.5	36.1	32.6	24.0	48.3	57.8	47.6	52.1
total earthworms	32.9	32.9	23.9	18.7	42.4	46.9	34.6	34.3

The mean values again show a tendency for the aggregated groups to have comparatively low coefficients of variation. Earthworm species that do not usually occur in dominant abundances and are rather rarely sampled, however, show a comparatively high relative scattering between plots. This pattern inherently has implications for the test power of the different endpoints in terms of identifying significant effects of substances (see chapter 3.2.3.1.3). During the temporal course of the field tests, no systematic changes in the coefficients of variation for control treatments at specific sampling time points were detected. Boxplots (database) and tables (pilot study and database) of the single sampling time points for dominant species and earthworm groups are presented in Appendix A.3 (Tables A3-8 – A3-10, Figure A3-3).

Figure 24 and Figure 25 illustrate the relationship between mean control values for the endpoints earthworm abundance and biomass and the respective coefficients of variation of the control on plot and sample (= subplot) level.

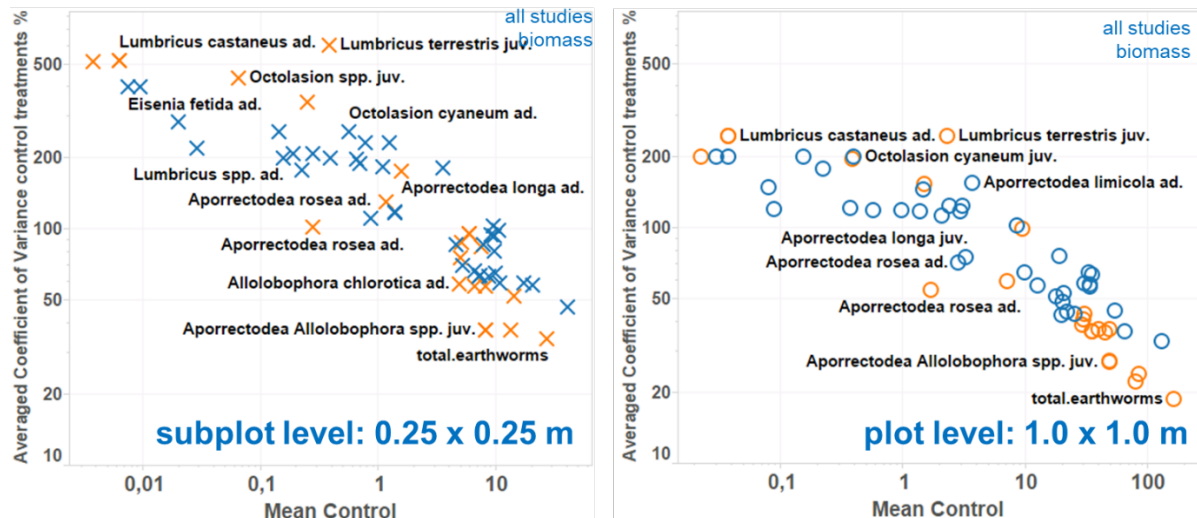
Figure 24: Relationship between mean control earthworm abundances and mean coefficients of variation of the control treatments [%] for the identified species and groups of all studies (orange: pilot study, blue: database studies; only selected species are labelled). Illustration at sample (= subplot) level (left, 0.25 x 0.25 m) and at plot level (right, 1.0 x 1.0 m)



Source: RWTH Aachen University

There is a clear correlation between the coefficient of variation and the mean value of the control at both plot and subplot levels. This relationship is also present for biomass data at subplot and plot level, as seen in Figure 25.

Figure 25: Relationship between mean earthworm control biomass and mean coefficients of variation of the control treatments [%] for the identified species and groups of all studies (orange: pilot study, blue: database studies; only selected species are labelled). Illustration at sample (= subplot) level (left, 0.25 x 0.25 m) and at plot level (right, 1.0 x 1.0 m).



Source: RWTH Aachen University

This correlation has already been stated in earlier studies (cf. Ekschmitt et al. 1998). With regard to the research question of the project that focusses on an improvement of the test design for earthworm field studies, this natural variability of the earthworm community has implications for the statistical detectability of effects: Results indicate that particularly aggregated species groups with high abundances and biomass values will provide powerful endpoints, considering the required data characteristics for a powerful derivation of effect thresholds in statistical test procedures (especially low variation in controls and treatments). On average, the scattering at the single species level seems too high to prove statistical effects. This has also been shown in the MDD calculation of the database studies (chapter 2.3). A high variation in control treatments (especially with low abundances and biomass values) thus leads to a lower visibility of the possible effect of the test substance (= high MDD), i.e. possible effects cannot be statistically detected.

3.2.3.1.2 Assessing the statistical strength of the test setup

The impact of variance on the number of required replicates to achieve a certain test power was determined for the standardised Dunnett test in a separate calculation step. Calculations were based on coefficients of variation for control treatments in earthworm field tests, applied for a dynamical sample size planning with regard to a detectable difference (MDD %) that should be achieved.

At this point, the project consortium decided to use the sample size calculation approach as a measure for the respective test power. A comparative analysis of the percentage test power as the reciprocal of the type-II error is not useful here (post-hoc test): Since the “observed” test power is a direct function of the p-values of the multiple t-test (Hoenig & Heisey 2001), it is not more informative as a post-hoc value than the respective calculated p-values. It is known that the database studies hardly show any statistically significant effects (see also limit test design and Dunnett test procedure, chapter 3.2.3.2.1), for this reason the p-values and thus the calculated post-hoc test power-values are lower than for the pilot field study. Observed (or post-hoc)

power and p-values are directly related, therefore a direct comparison of the impact of the new test design cannot be derived.

For the development of an adapted test design for earthworm field tests, it is more informative to raise the research question of how many samples (=replicates) should be used theoretically given a desired target-test power and a given natural variability of data (chapter 3.2.3.3). This question enables us to draw conclusions regarding an adapted, upcoming test design. The desired test power should be determined beforehand. By default, this is usually set to 80% for statistical hypothesis testing, which is also applied in the following analyses.

The detectable difference that can be achieved with the respective sample sizes was classified into four different classes in this simulation. The class sizes were adapted to the scaling of magnitude of effects of the EFSA soil opinion (EFSA PPR 2017), although there they refer to the possible effects on the protection goals (assessment endpoints) and not specifically to the measurement endpoints in the field.

Table 19: Scaling of magnitude of effects (= "Effect classes") according to the EFSA Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms (EFSA PPR 2017)

Scaling of magnitude of effects	Per cent reduction (%)
Negligible effects	0-10%
Small effects	10-35%
Medium effects	35-65%
Large effects	65-100%

Following effect ranges in a general dose response curve, up to 10% difference between control and treatments was defined as negligible deviation from control ranges. Small effects were set up to a limit of 35% difference from controls, 35 to 65% difference were defined as medium effects and from 65% onwards as large effects. The acceptability of effects is not considered in these ranges, but needs to be defined in further steps taking other parameters into account (e.g. potential for dispersal). For instance, only small effects might be acceptable for some organisms, while for others also medium effects could be acceptable for a given time frame.

3.2.3.1.3 Sample size modelling approach

The theoretical sample size calculation for earthworm field studies was conducted according to Dunnett's multiple test procedure (Horn & Vollandt 1995) using the following equations to calculate the number of replicates for substance treatments (1.) and the control (2.):

$$\left[(t_{\alpha, \infty, r, 1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{\delta^2} \left(1 + \frac{1}{\sqrt{a}} \right) \right] = n \quad (1.)$$

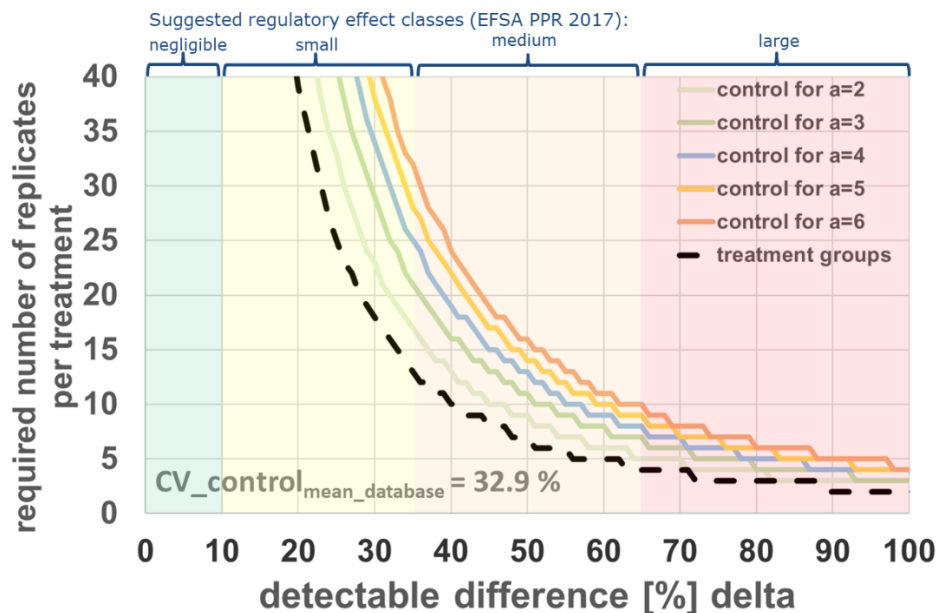
and

$$\left[(t_{\alpha, \infty, r, 1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{\delta^2} (1 + \sqrt{a}) \right] = n_0 \quad (2.)$$

where a is the number of treatments per field study, $t_{\alpha, \infty, r, 1-\alpha}$ is the one-sided threshold for an a -dimensional t-distribution (significance level α set to 0.05, tabulated according to Horn & Vollandt 1995), $z_{1-\beta}$ is the quantile of the standard normal distribution (type-II-error= $\beta=0.2$), σ is the coefficient of variation for control treatments (in percent, chapter 3.2.3.1.1) and δ the difference to be detected (%).

As an example of application, Figure 26 illustrates the simulation of required replicate numbers that correspond to the mean coefficients of variation of the group “total earthworms” (=32.9%, field tests available in the database). Test power was set to 80%. The required number of replicates is plotted against the desired detectable difference between control and treatments with varying number of treatment levels.

Figure 26: Number of required replicates (plots) per treatment in earthworm field tests plotted against the detectable difference (in percent) between treatment and control. Variation of control was set to 32.9%, which is the mean variability of total earthworms in available field studies. Coloured lines: Required replicates for controls using different numbers of test treatments (a); dotted line: Required replicates of test treatments



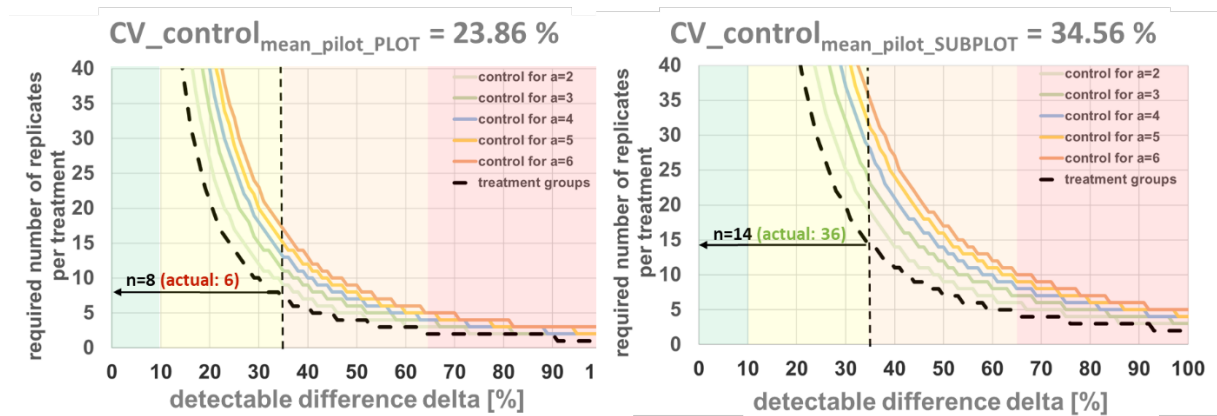
Source: RWTH Aachen University

Results of the sample size simulation for mean total earthworm variability indicate that standardized earthworm field tests might have an insufficient number of replicates to detect small effects with a test power of 80% for “total earthworms”, which is the group with the lowest coefficients of variation in earthworm field test data. Accordingly, the ability to detect effects for other earthworm groups is even more limited. With a coefficient of variation of 32.9%, only large effects are detectable in a standardized field test if four plots as replicates per treatment are used in combination with a test power of 80%.

These results indicate that earthworm field tests might have a general shortcoming of inadequate number of replicates, that hampers a solid identification of small to medium effects among different taxa and groups and with a high statistical power. Even for the lowest coefficient of variation for total earthworms observed in the pilot study (11.5%), the identification of all small effects with a test power of 80% would only be possible in the pilot field study test design with 19 replicates for the concentration level and about 47 control replicates.

For this reason, it was investigated subsequently if a NOEC calculation using the samples (=subplots) as statistical replicates would result in an improvement with regard to detectable differences. We calculated the sample size planning with the measured coefficients of variation on plot level (standardized method), and on sample level to assess shifts in test power.

Figure 27: Detectable difference (in percent) of a treatment in earthworm field tests compared to the control depending on the number of required replicates for a given variability of the community (coefficient of variation of the control) at plot level (left) and at sample (= subplot) level (right). A type-II error of 0.2 respectively a test power of 80% was fixed for the sample size simulations. Coloured lines: Required replicates for controls using different numbers of test treatments (a); dotted line: Required replicates of test treatments



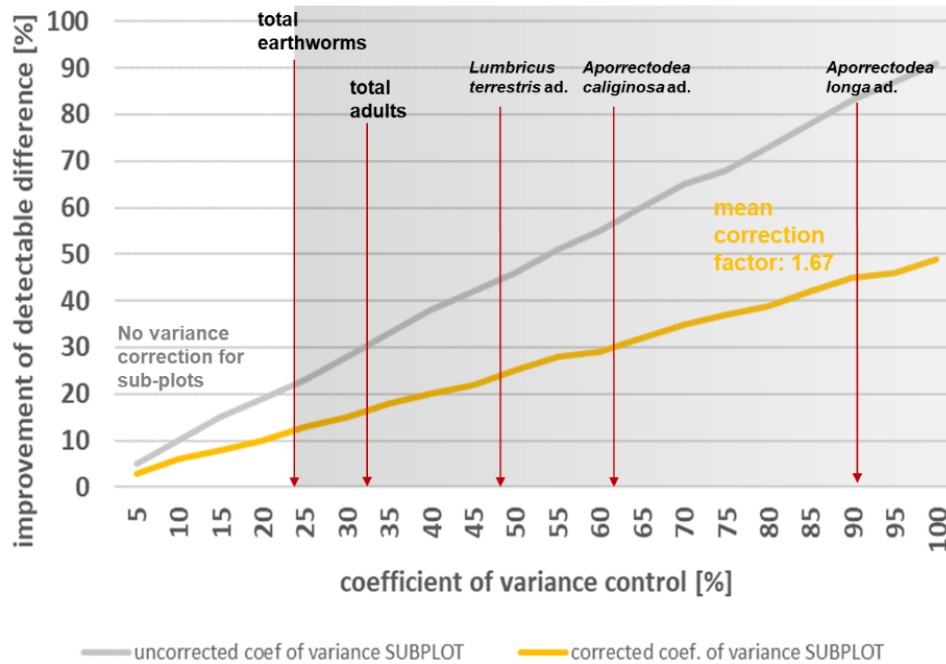
Source: RWTH Aachen University

Results of the pilot study shows that the increase in plot numbers ($n=6$) and the slightly lower coefficients of variation in combination with 1.5 m^2 sampled plot area instead of 1.0 m^2 will increase test power (left graph). The number of required replicates to achieve a certain threshold of detectable difference decreases compared to database field studies (see also Figure 26). Nevertheless, using this test design, medium effects (35% - 65% effect) will not be detectable with a power of 80%. The type-II-error would still be higher than 20%. In this case, the required number of replicates to reach this threshold of 35% would be eight plots, instead of the used six plots. If the aim would be to detect small effects (10% - 35%) with a test power of 80%, it seems unrealistic in this simulation considering realistic number of replicates (8 - >40 plots).

Nevertheless, in the test setup of the pilot study, a 35% difference between treatment and control would be detectable at sample (= subplot) level (right graph). In this case, the mean coefficient of variation at subplot level is slightly higher than at plot level (34.56% compared to 23,86%); for this reason, at least 14 replicates (for treatment groups and controls in a test design with $a=2$ treatments) would be necessary to detect medium effects of 35%. However, by using the single samples as replicates, there would be 36 replicates available in this statistical design. This calculation illustrates, that the assessment of the pilot study on subplot-level enables to detect medium effects with a test power of 80% (regarding the mean coefficients of variation). For the detection of a large effects with 65% difference from the control, 4 replicates would be needed.

Based on these findings, the potential improvement of detectable differences at subplot level compared to the plot level for the data and design of the pilot study was analysed generally in the following section.

Figure 28: Extrapolation analysis - improvement of the detectable difference in earthworm field studies [%] depending on theoretically assumed coefficients of variation. The calculation shown is based on the following fixed parameters, similar to the earthworm pilot field study (type-I error: 0.05; test power: 0.8, number of treatments: 6, number of plots: 6; total number of samples (= subplots): 36. More details in the text



Source: RWTH Aachen University

The grey graph in Figure 28 illustrates the theoretical improvement of the detectable difference (in percent) in relation to the respective coefficient of variation of control treatments by evaluating the study results at sample (= subplot) level instead of plot level. Red arrows in the figure indicate the mean observed coefficients of variation for single earthworm groups. It can be shown that a proportional improvement always happens, but especially in cases where comparatively low abundance or biomass values have been measured. Thus, this switch in the assessment level allows the identification of more significant differences, especially at single species level, which would mean a substantial benefit for the identification of sensitive individual species.

However, the grey graph does not take into account the fact that at sub-plot level, due to the lower scores, a higher variability of the endpoints has to be expected compared to the plot level (see Table 18). The mean increase of the coefficient of variation at subplot level is a factor of 1.67. Taking this into account as a correction factor for the coefficients of variation in the theoretical simulation at subplot level (yellow graph), there is still an improvement in statistical detectability across all species and earthworm groups. For the group of total earthworms this is about 13 %, for single species - for example *Aporrectodea caliginosa* (29%), *Lumbricus terrestris* (22%), *Allolobophora chlorotica* (18%) or *Aporrectodea longa* (46%) - the relative improvement is substantially higher.

Moreover, it was calculated in the simulations that by changing the assessment level from plot to subplot level, a statistical improvement in detectability can always be achieved as long as the coefficient of variation of the data is at most 2.45 times (or less) greater at subplot level than at plot level (calculated for the study design of the pilot study). However, this threshold factor of

2.45 is not reached on average for any of the differentiated species or earthworm groups in the database. We can therefore state that, in general, the statistical detectability of effects in earthworm field studies generally improves if the evaluation is carried out at the subplot level.

The findings lead to the legitimate question whether it is acceptable to statistically evaluate the field tests at subplot level. This can indeed be affirmed if the communities of the single samples (=subplots) of a plot do not interfere with each other (Schank & Koehnle, 2009). However, due to the expected migratory behaviour of earthworms, an interaction between the individuals of the single samples cannot be excluded. The authors of this report at least raise concerns that there may also exist interdependencies between samples of adjacent plots that may be more pervasive than between samples within a plot due to the heterogeneity of the soil matrix. This, in turn, varies from field to field and cannot be conclusively clarified. The existing dogma that the sampled subplots of one plot describes a dependent entity and therefore should not be regarded as single replicates ("pseudoreplication") is just as difficult to clarify on the available data basis. A statistical verification of these declarations, for example by considering a factor for the impact of the plot arrangement (e.g. in a GLM) is not meaningful due to the low number of replicates in an earthworm study and would preferably also require data for the location of the individual plots or samples on the entire test field. This cannot be realized with the current data situation.

Even if we cannot give a final and generally valid judgement on the choice of the assessment level, since we acknowledge that this and related topics has been controversially and not conclusively discussed under the term "pseudoreplication" for decades (Hurlbert 1984; Ruxton & Colegrave 2017; Schank & Koehnle 2009; Davies & Gray 2015), we would like to emphasize the advantages of statistical evaluations at the subplot level.

The general improvement of test power in the statistical evaluation at subplot level compared to plot level is also evident in the reduction of the percentage MDD. This is shown as tables for the database studies (table A3-4) and the earthworm pilot study (table A3-5) in Appendix A.3. This trend can also be seen within the pilot study between the treatments with six replicates (36 subplots, treatment T2 and T5) and three replicates (18 subplots, treatment T1, T3, T4, T6). Both at the plot level (Table A3-6) and at the subplot level (Table A3-7), an increase in the number of replicates tends to improve the MDD (%). The only exceptions to this trend include species and sampling time points in which statistically unrepresentative numbers are sampled (single findings etc.). Small effects (10-35% effect), however, cannot be detected comprehensively with a test power of 80% with this design using the Dunnett test. The MDDs of tested species and earthworm groups in the pilot study (plot level) are shown in table A3-2 and figures A3-2.

3.2.3.2 Calculation of effect thresholds (NOEC)

3.2.3.2.1 Standardized procedures - multiple t-tests

The calculation of the effect thresholds was carried out for all available studies of the database and at all sampling times using the statistical software R (version 3.5.0, package: multcomp, 1.4-8). According to the ISO guideline 11268-3 for the determination of effects on earthworms in field situations (ISO 2014), pilot study data and tests from the database were initially evaluated using a Dunnett's t-test ($\alpha=0.05\%$, two-sided for unclear direction of response). The calculated NOEC values of the species groups at all sampling times are shown in the fact sheets, Appendix A.2. A comparison of the Dunnett method with results from the CPCAT approach will be presented in the following chapter.

In all cases, a prior data transformation is not recommended, the significance of statistical test results using transformed data cannot be interpreted straightforwardly for the non-transformed data. It is also noted that from an ecotoxicological point of view relevant increases in abundance

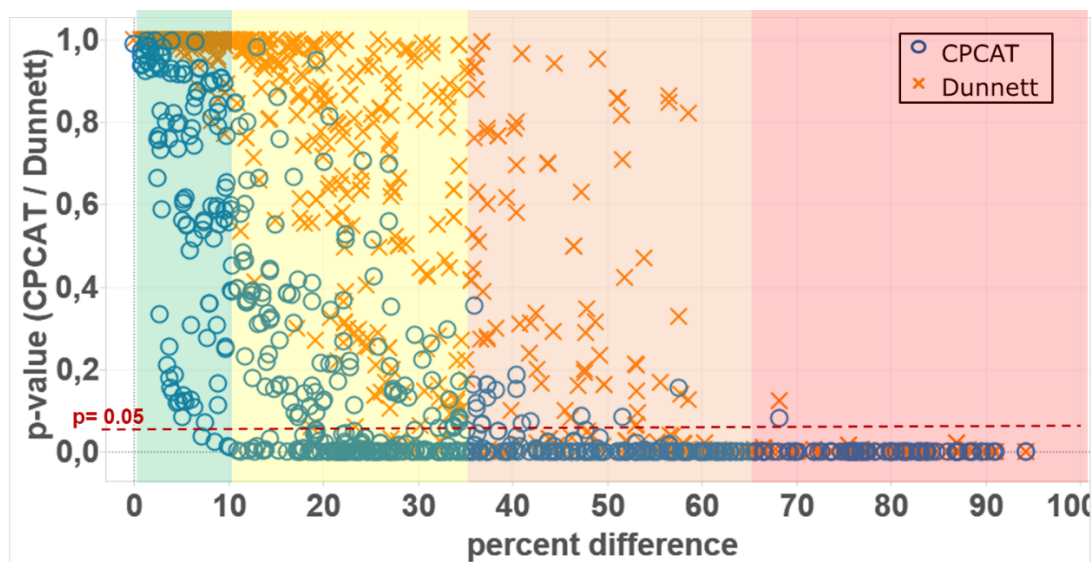
and/or biomass are in principle considered as abundance and/or biomass decreases: they are deviations from control situation. Increases, as well as decreases in measured endpoints, need to be considered for their biological relevance and statistical significance (two-sided test procedures).

3.2.3.2.2 CPCAT

The theoretical distribution assumption of earthworm abundance field test data follows a Poisson model (see chapter 3.2.1). Therefore, the application of the CPCAT approach (Lehmann et al. 2016) is highly recommended for abundance count data due to more powerful test statistics (Lehmann et al. 2018a). This is the first time that the performance of CPCAT is assessed within a comprehensive meta-analysis of field study data.

For the implementation of the CPCAT procedure, an R-script was generated which is based on the original script for CPCAT analyses (see Lehmann et al. 2016). A comparison of the calculated p-values from the CPCAT procedure with corresponding values of the Dunnett calculation, plotted against the respective percentage differences of the treatments for control plots for the pilot study, is shown in Figure 29.

Figure 29: Percentage difference between control and treatments for all single species and aggregated earthworm groups abundances in the pilot field study plotted against respective calculated p-values calculated with the CPCAT (blue dots) and Dunnett (orange crosses) method for all sampling time points. Background colours: Scaling of magnitude of effects as suggested in the Scientific Opinion on Soil Organisms (EFSA PPR 2017)



Background colour: effect classes according to the Scientific Opinion on Soil Organisms (EFSA PPR 2017), green: negligible effects, yellow: small effects, orange: medium effects, red: strong effects. The red dotted line corresponds to the conventional threshold value ($p=0.05$), a statistically significant effect is suggested if the value is below it. For reasons of clarity, only the data points with mean control values ≥ 10 are shown here (explanation in chapter 3.2.3.1, all results shown in Appendix A.4).

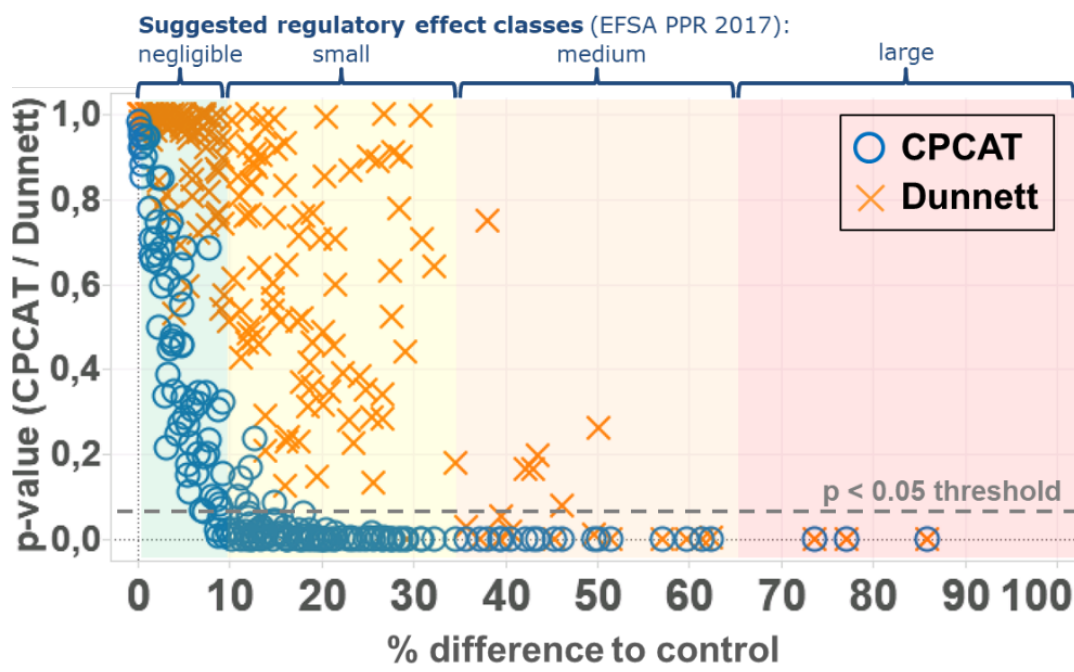
Source: RWTH Aachen University

It is shown that the use of the CPCAT procedure in comparison to the Dunnett test increases the probability that significant effects are identified, even in effect classes with smaller effect sizes (especially for small and medium effects, between 10% and 65% difference to control). CPCAT is therefore generally "more sensitive", i.e. significant effects of the test substance are already indicated for smaller differences in control. This is evident in the existing earthworm field test of the

database as well as in the pilot study and especially in species groups with relatively high abundance and biomass numbers and comparatively low coefficients of variation of the control. As already described in chapter 3.2.2, these "powerful" species groups are mainly aggregated groups such as "total earthworms", "total juveniles" or "total adults" (see Annex A.2). These are thus statistically well assessable, but do not replace the ecological significance of the individual assessment of all detected earthworm species (chapter 3.2.3.1).

This very crucial shift in calculated p-values for the same data depending on the test procedure employed to assess the data (here Dunnett vs. CPCAT) can also be found in existing field studies of the database, as shown for the statistical measure "total earthworms – abundance" in Figure 30.

Figure 30: Percentage difference between control and treatments for total earthworm abundances in database and pilot field study plotted against p-values calculated with the CPCAT (blue dots) and Dunnett (orange crosses) method for all sampling time points. Background colours: Scaling of magnitude of effects as suggested in the Scientific Opinion on Soil Organisms (EFSA PPR 2017)



Background colour: effect classes according to the Scientific Opinion on Soil Organisms (EFSA PPR 2017), green: negligible effects, yellow: small effects, orange: medium effects, red: strong effects. The red dotted line corresponds to the conventional threshold value ($p=0.05$), a statistically significant effect is suggested if the value is below it. For reasons of clarity, only the data points with mean control values ≥ 10 are shown here (explanation in chapter 3.2.3.1, all results shown in Appendix A.4).

Source: RWTH Aachen University

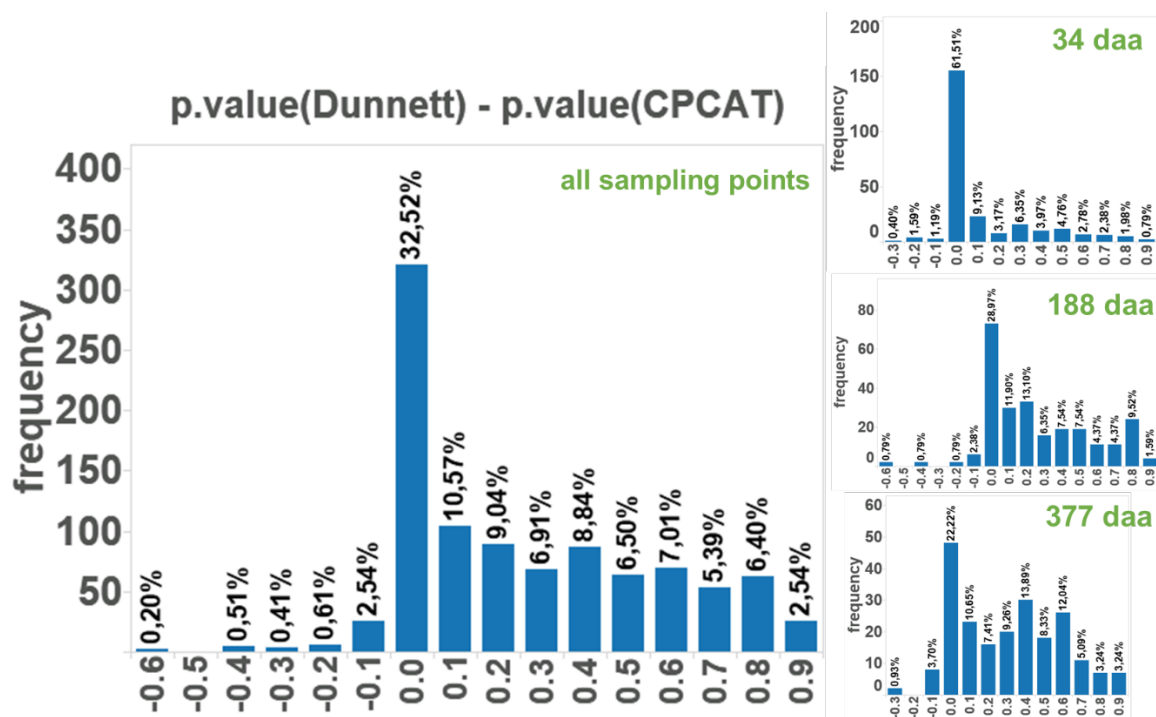
Results for single species and other functional /morphological earthworm groups are shown in this kind of illustration in the Appendix A.4. The differences in test power between the two procedures also become visible in the separate examination of the NOEC calculations for different species groups (fact sheets, Appendix A.2): As an example, at the sampling time point 12 months after application, a NOEC of 5,800 g/ha was calculated according to testing with Dunnett for abundance data of total earthworms. From the fifth treatment level (10,500 g/ha) onwards, significant differences between the treatment and control abundance were detected with Dunnett. This NOEC disguised a mean decrease of earthworms of 46.1% at 5,800 g/ha, due to a relatively

high scattering of data in the control treatment. In contrast, the NOEC according to CPCAT is < 600 g/ha. At this lowest test concentration there is already a mean decrease of 27.5%.

The fact sheets in the Appendix (A.2) contain assessment results of CPCAT for the endpoints abundance and biomass. These were carried out in this project to test the general applicability of CPCAT for field study data. Nevertheless, biomass data are metric responses which do not correspond to the testing requirements of a Poisson distribution. However, in the course of Poisson estimation, the CPCAT algorithm calculates the mean values from the replicates in any case (averages of Poisson distributions again result into a Poisson distribution), so that numbers with decimals can also be included initially in the calculation. In CPCAT, no pre-test for integers is implemented, even if this would be reasonable for a Poisson estimation. Therefore, we are able to use biomass data as a proof of concept of the CPCAT approach -they are treated in a similar way to abundance data- even though we are aware that the formal requirements for the statistical evaluation of biomass data are not fulfilled.

The absolute differences between the calculated p-values of the two test procedures of all concentration levels, sampling times and species groups in the pilot study are shown in Figure 31.

Figure 31: Histogram of the classified frequencies (class width: $p=0.1$) for the difference of p-values between Dunnett test and CPCAT method for all tested earthworm species groups, treatments and sampling time points of the pilot field study. Daa = days after application

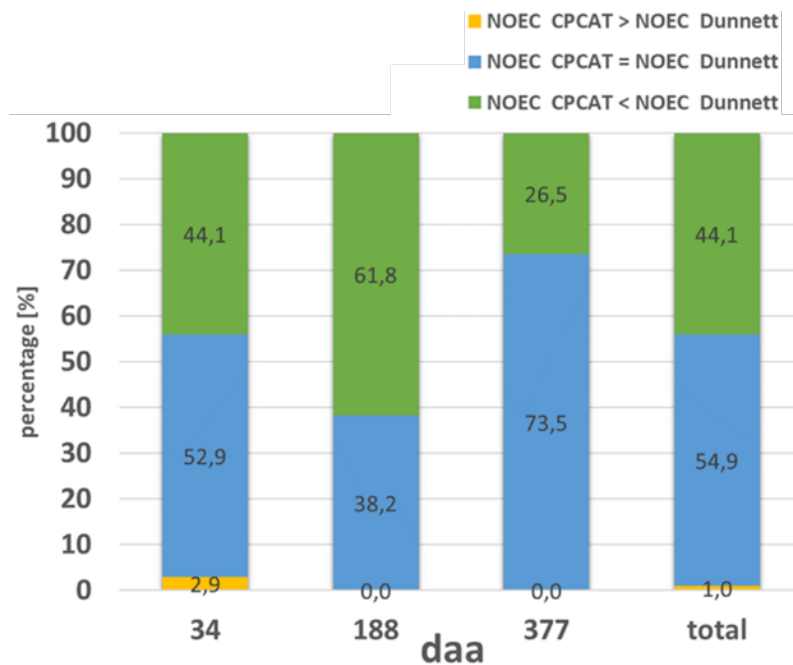


Source: RWTH Aachen University

This illustration shows that CPCAT usually generates similar or lower p-values than the Dunnett test. The differences to the Dunnett-test procedure are often not recognizable in case of clear effects of the test substance, e.g. 34 days after application. In this case, 61.5% of the p-values of all procedures are in the same range. For minor effects (377 days after application), on the other hand, the differences become more apparent, with more than 70% of the tests showing a lower p-value due to the use of CPCAT.

The effect of the different test procedures on the actually derived NOEC values is shown in Figure 32.

Figure 32: Percentage of cases where the calculated NOEC of endpoints in the earthworm pilot field study according to CPCAT is higher (yellow), lower (green) or equal (blue) to the NOEC of the Dunnett procedure. Daa = days after application



Source: RWTH Aachen University

These plots again reveal that CPCAT has an overall higher test power than the Dunnett procedure. This can be detected for the pilot study as well as for the database studies with classical study design. Thus, the percentages of significant treatments and results in both study types are summarized in Table 20.

Table 20: Overview of the percentage of test procedures with significant effects ($p < 0.05$) Calculations according to Williams, Dunnett and CPCAT (all sampling times). Database= available field studies with standard design. Pilot field study = extended design

Statistical toxicity measure	Database – NOEC determination [%]	Pilot field study – NOEC determination [%]
Williams NOEC	3.02 %	40.85 %
Dunnett NOEC	2.54 %	34.75 %
CPCAT NOEC	32.14 %	63.41 %

number of test procedures (=possible numbers of NOEC)= 4206 (database) & 164 (pilot field study)

The Poisson distribution used in CPCAT procedures describes the earthworm community data in outdoor tests mathematically and statistically more accurately than the normal distribution used in conventional t-tests (e.g. Dunnett or Williams). Thus, the use of the CPCAT approach increases the test power for earthworm field data (Lehmann et al. 2016). Although CPCAT does not perform perfectly when data are not following an exact Poisson distribution due to the approximation of a generalized Poisson distribution (overdispersed data), at least CPCAT is able to take into account the binomial distributed characteristic of count data, which is a major advantage in contrast to approaches of the t-test family, especially in case of small count numbers (as seen in the earthworm field tests for many species). However, it should be mentioned that for the relatively new CPCAT approach there are on-going debates on the best approach to calculate test power. Based on this, sample planning using of CPCAT procedure is not yet implemented. There is still a need for further research to increase the acceptance and usability of CPCAT.

In addition, the test design immanent shortcomings of the NOEC design are also retained if using CPCAT, even if the inefficiencies in the effect threshold calculation can be reduced. This increase in test power can be explained by the more accurate underlying Poisson model compared to the normal distribution, but also due to the use of the closure principle (CP), a powerful tool in multiple testing (Bretz et al., 2011) which avoids duplicate testing of hypotheses and alpha-inflation (Lehmann et al. 2018b). However, the calculated threshold values are still a priori defined concentration levels, so possible effects could still be disguised by the use of CPCAT. For this reason, the evaluation of the applicability of the EC_x design tested in the pilot study is presented in the following chapter.

The study reveals that the application of standard multiple testing procedures leads to a disguising of possible effects due to relatively high differences to be achieved between control and treatments. This consequently results in uncertainties regarding the actual level of effects at the NOEC. The CPCAT approach offers a more powerful and statistically proper evaluation for these earthworm field studies because data distribution and variance are adequately considered and smaller differences between control and treatments can be detected.

3.2.3.3 Calculation of effect concentrations - pilot study

In contrast to the available earthworm field studies in the database, the pilot study was carried out in a test design with several application rates. The chosen substance application rates were selected so that a dose-response relationship could be demonstrated along the different endpoint measures. The calculation of potential dose-response curves and the derivation of EC_x-values, the application rates that causes x% of an effect on test organisms within a given exposure period when compared with a control, is described in the following chapter. In this case, to be more precise, the applied rates that results in x% effects were used for calculations. Therefore, the resulting statistical measures are technically ER_x-values. In this chapter, the terms EC_x/ED_x/ER_x are used interchangeably in order to emphasize the focus on the statistical approach, which is independent of whether the related parameter is a concentration, dose or rate.

The Probit curve regressions were conducted using the statistical software ToxRat (ToxRat Solutions, version 3.2) in order to ensure standardized procedures.

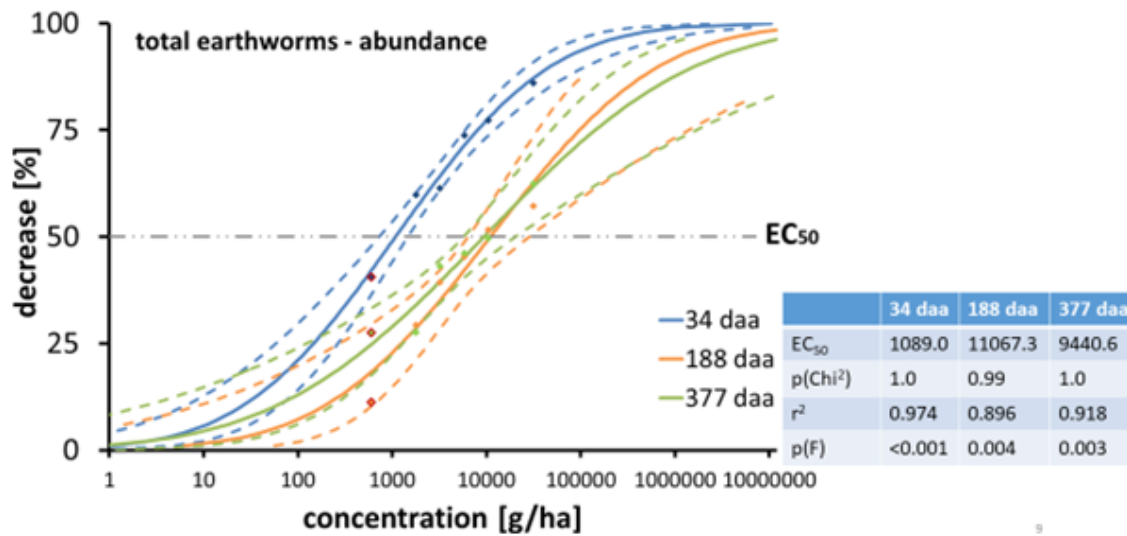
3.2.3.3.1 Probit regression

For Probit regression procedures, the effect is modelled as a per cent or proportion of the control mean response. In these cases, the normal sigmoid curve is fitted to the results using the probit regression procedure (Finney 1971). The regression uses a maximum likelihood approach. EC_x/ER_x values are computed by inserting a value corresponding to x% of the control mean into the equation found by regression analysis. 95%-confidence limits are calculated ac-

ording to Fieller (Finney 1971). For the metric responses, a weighting function has to be adjusted for the computation of variances and confidence limits as given by Christensen (1984).

The dose-response relationship for the total abundances of the earthworm community in the pilot study is shown in Figure 33 based on a standardized probit regression. All dose-response curves and statistical calculations of curve fits for single species and earthworm groups can be found in the statistical fact sheets (Appendix A.2).

Figure 33: Dose-response curves of the group "total earthworms" using the endpoint measure total abundance (adults & juveniles) for pilot study data (regression method: Probit). Daa = days after application



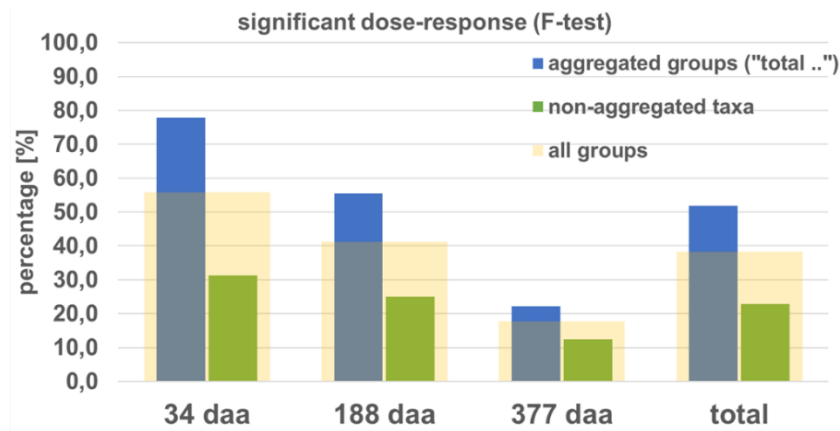
Source: RWTH Aachen University

At all three times of sampling after application, a significant dose-response relationship has been identified in the group "total earthworms" ($p(F) \leq 0.05$, i.e. the slope of the curve is significantly different from zero). Thus, for the total abundance in the pilot study, the choice of an EC_x design allows revealing significant relationships between the carbendazim concentration applied and the measured effect on the earthworm population. In addition, the comparison of EC₅₀ values across sampling times indicates recovery effects that can be assumed in case of the total earthworm community between 1 and 6 months after application (during the summer period). The calculated EC₅₀ increases by a factor of 10 during this period, whereas it does not change in any further comparison at the sampling time after 12 months. By contrast, the EC₅₀ for (*Aporrectodea/Allolobophora* spp.) juveniles were increasing only by a factor of approximately 3 until the end of the study.

The results of the study show that the use of an EC_x design to derive effect concentrations on the earthworm population in the field is generally feasible. However, when comparing the dose-response relationships and the quality of the resulting curve fit for different species (Appendix A.2), it becomes apparent that the choice of a suitable concentration range for adequate testing of all species and aggregated groups poses a challenge. This problem is in principle also present in the NOEC design, but is concealed due to the test statistics applied (chapter 3.2.1): A calculated NOEC does not reveal this as does a non-derivative EC_x value.

The proportion of significant dose-response curves in the pilot study all sampling time points is shown in Figure 34.

Figure 34: Percentage of significant dose-response relationships for all calculated regression procedures with the endpoints of the of the earthworm pilot field study (F-test, threshold $p < 0.05$). Single findings of species were not considered. Daa = days after application



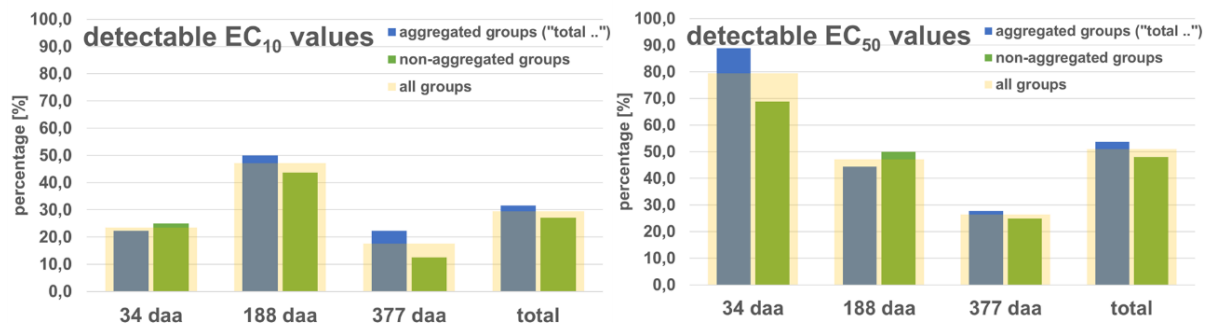
Source: RWTH Aachen University

The significance of the concentration-response relationship was evaluated with a F-test for the slopes of the calculated curves. $P(F)$ is the probability that the data points of the curve randomly simulate a dose-response relationship, even if there is none. The smaller $p(F)$ (the larger the F-value), the better the explanatory contribution of the substance on the slope.

Regarding the pilot study, it was found that a significant dose-response relationship is most frequently detectable for initial, strong effects (34 days after application). There is a decline over time regarding the significances of the slopes: after one year, a significant dose-response relationship is only detected in <20% of the groups. Aggregated earthworm groups show in general more significant dose-response relationships (51.9%) than non-aggregated species groups (22.9%). This is true for every time of sampling after application during the pilot field study.

Figure 35 illustrates the percentage of detectable EC_{10} and EC_{50} values during the test. According to general convention, the EC values that are located between tested concentration levels and can thus be interpolated on a predictable regression line are assigned as "detectable" in this diagram. Extrapolation factors to derive EC_x values were not considered.

Figure 35: Percentage of detectable EC₁₀ (left) and EC₅₀ values (right) of all calculated dose-response curves for endpoints of the earthworm pilot field study data. EC values are detectable in this representation if they lie between concentration levels and can therefore be interpolated on a calculable regression line. The significance of the curve as a pre-test (see above) as well as single findings of species were not considered here. Daa = days after application.



Source: RWTH Aachen University

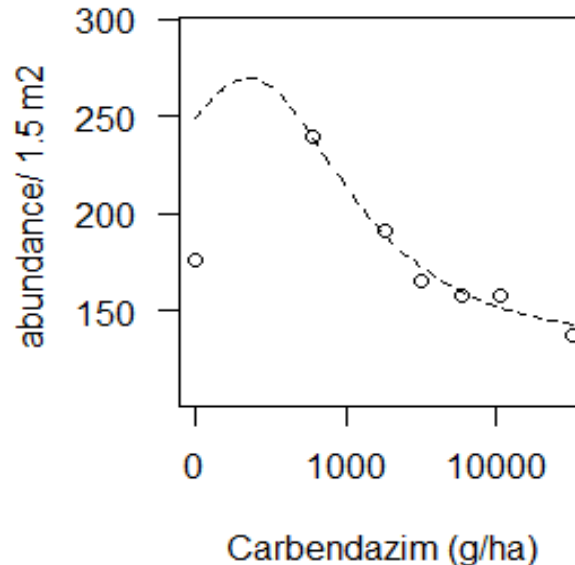
This chart illustrates to what extent it was possible to derive adequate effect concentrations across all tested groups due to the chosen concentration range in the study. It can be seen that at the first sampling time, the range of EC₅₀ values was initially covered by most of the groups (~80%). Overall, the selected concentrations were too high after one month, as only about 20% of the EC₁₀ could be reliably estimated. After 188 days after application, the proportion of EC₁₀ and EC₅₀ values is similarly high (~50%). Since several groups showed no effect after one year, reliable EC_x values are rarely reached (~20%). The major challenge of choosing suitable concentration ranges for earthworm field studies in the course of the year becomes evident from these analyses.

3.2.3.3.2 Alternative regression methods

The use of two-parameter probit regression is the well-established method of choice for describing dose-response curves in ecotoxicology. Nevertheless, in the pilot study, effects of the substance on species groups were also observed where the dynamics cannot be satisfactorily described by the point-symmetric shape of the probit curves. These include left-skew concentration interactions, which can usually be adequately represented by a two-parameter Weibull regression. However, slight increases of the considered endpoint may occur in low concentration ranges (so-called hormesis-like responses) as well as general increases due to exposure. In these cases, the data situation should always be examined specifically and, if reasonable, the regression procedure can deviate from a two-parameter regression. Guidelines for these scenarios of modelling quantal dose-response data, e.g. for the evaluation of earthworm reproduction tests, have been provided by the OECD (e.g. OECD 2006a; OECD 2016).

In the case of a monotonous increase of the effect measure, two-parameter regression procedures could also be used; for the occurrence of so-called hormesis, the use of an adequate, alternative curve adaptation procedure was shown as an example in Figure 36. The abundance data of the group *total epilobous adults* 188 days after application is illustrated.

Figure 36: Exemplary illustration of an elaborated curve adaptation method by integrating a so-called hormesis function for the earthworm data set *total epilobous adults* of the earthworm pilot field study (sampling time: 188 days after application, end-point: total abundance). A modified four-parametric log-logistic model according to Brain Cousens was used to model an hormesis-like response

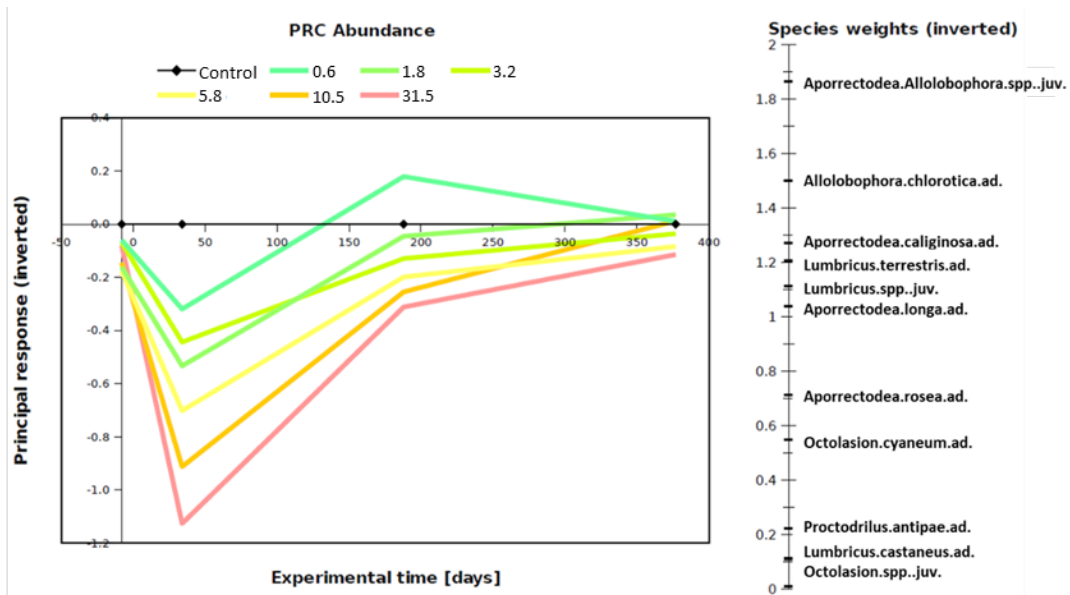


Source: RWTH Aachen University

3.2.3.4 Community analyses (PRC) – pilot study

A multivariate community analysis was conducted in order to summarize the response of the community to a disturbance, in this case the response of the earthworm community in the field exposed to carbendazim in several treatments. Principal Response Curve (PRC; van den Brink & ter Braak 1998, 1999) is often used to summarize the response of a community to some disturbance. It is a special type of redundancy analysis (RDA) applied to answer the question whether there a significant relation between community structure and environment (here: treatment). A permutation test is used to test for significance to ascertain the effect represented by the PRC. The resulting response curve shows the extent and direction of development of samples (communities) under different experimental treatments, compared with the control. Additionally, the directions of such composition changes can be interpreted using the response of individual species.

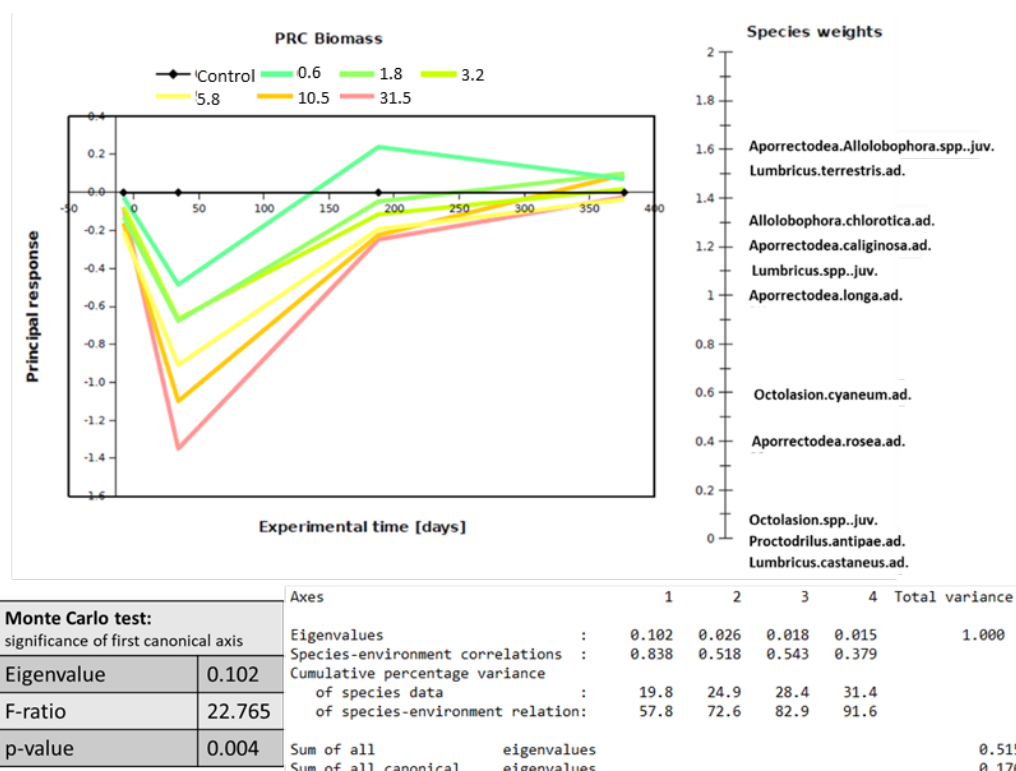
Figure 37: Principal-Response-Curve (PRC) for species abundance data of the earthworm pilot study. Different treatments (application rates from 0.6 to 31.5 kg carbendazim/ha) have different colours



Monte Carlo test:		Axes					
significance of first canonical axis		1	2	3	4	Total variance	
Eigenvalue	0.106	Eigenvalues	0.103	0.040	0.020	0.014	1.000
F-ratio	19.394	Species-environment correlations	0.837	0.740	0.411	0.410	
p-value	0.002	Cumulative percentage variance of species data	17.4	24.1	27.6	29.9	
		of species-environment relation:	51.1	70.9	81.0	87.8	
		Sum of all eigenvalues					0.593
		Sum of all canonical eigenvalues					0.202

Source: RWTH Aachen University

Figure 38: Principal-Response-Curve (PRC) for species biomass data of the earthworm pilot study. Different treatments (application rates from 0.6 to 31.5 kg carbendazim/ha) have different colours



Source: RWTH Aachen University

The PRCs reveal a highly significant effect of the treatment on the earthworm community (p -value < 0.05). A clear dose-response relationship is visible and with increasing concentration the deviation from the control increases. According to the PRC, the recovery of the community (regaining initial state) is indicated after approximately one year. This is in line with the assessment of effects on single species (adult individuals) and for total adults (Appendix A2). However, when considering earthworm groups with juvenile individuals, the NOEC (CPCAT) and percentage decrease compared to control treatments of the measured endpoints is still detectable after one year.

3.2.4 Design requirements for earthworm field studies -conclusions from statistical procedures

Based on the analyses of the existing data on earthworm field tests, generic derivations of recommendations are limited due to the high variability in the various earthworm data of different field tests and due to the expectable impact of local site conditions. However, the following basic recommendations and requirements regarding the implementation and evaluation of earthworm field tests could be defined:

1. There is still a need to determine and evaluate the endpoints of biomass and abundance at species level, as the aggregated morphological or functional groups used may disguise effects on single species (chapter 3.2.2 and Appendix A.2, statistical fact sheets).
2. The ECx design is a meaningful alternative to the NOEC design in the earthworm field test, at least one mix design would be advisable; the ECx design leads to stronger/more protective

statements for Environmental Risk Assessment of chemicals especially at lower effect ranges, a masking of possible effects as in the NOEC evaluation is avoided.

3. The calculation of effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure available for given prerequisites. If possible, the CPCAT approach is preferable. A possible scheme to conduct multiple test procedure based on data characteristics is shown in Figure 39.

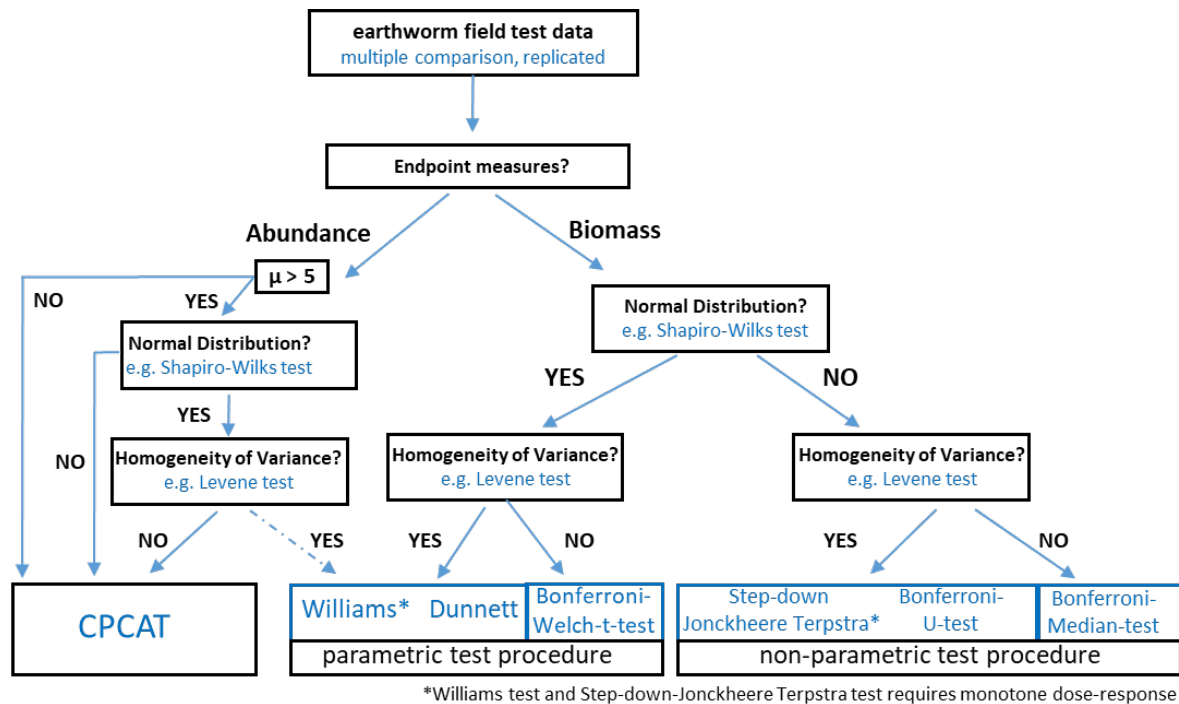
If data are metric (e.g. biomass), multiple t-test procedures such as Dunnett's or Williams' test ($\alpha = 0.05$, two-sided for unclear direction of response) should be performed (Dunnett 1955; 1964) for multiple comparisons in randomized plot design. The prerequisite of normally distributed data and variance homogeneity has to be tested using e.g. Shapiro-Wilks and Levene's test procedure, respectively. If data do not fulfil the criterion of normality, generalized linear models or non-parametric tests e. g. the Bonferroni U-test in accordance with Holm (1979) or the Jonckheere-Terpstra Step-down-test (homogeneity of variance required) can be applied (Figure 39).

The theoretical distribution assumption of earthworm abundance field test data follows a Poisson model. Therefore, the application of the CPCAT approach (Lehmann et al. 2016) is highly recommended for abundance count data due to more powerful test statistics (Lehmann et al 2018a). Nevertheless, if abundance data show homogeneity of variances, the null-hypothesis of normal distribution is not rejected (which is seldom reached for small sample numbers) and absolute abundances per replicate are > 5 (Gupta & Guttman 2014), the application of parametric test procedures (Williams, Dunnett) is also feasible. For multiple t-test procedures and with unequal replication, the table t-values must be corrected as suggested by Dunnett and Williams.

In addition, an inappropriate log-transformation of data during the calculation procedure should be avoided: In comparison to the evaluation of e.g. aquatic mesocosms, in which single species regularly appear with highly dominant abundances, this is not given for the expected earthworm abundances on one square meter plot sizes. In addition, the significance of statistical test results using transformed data cannot be interpreted straightforwardly for the non-transformed data as well. For this reason, the statistical assessment of the earthworm field data should be conducted without data transformation.

4. After data revision, it should be decided whether a simple two-parameter Probit (Logit, Weibull) regression, a nonlinear regression or the integration of a so-called hormesis model for the calculation of effect concentrations (EC_x) is necessary. In case of a monotonous increase of the measured endpoint (biomass, abundance), the derivation of significant effect concentrations should also be taken into account.
5. If there are no ecological reasons for not using the data at sample level, the evaluation and interpretation of the data at plot (pooled samples of 1 m^2 in total used as replicates) and subplot level (single samples as replicates of 0.25 m^2) should be requested. A discussion regarding possible limitations of this approach is given in chapter 3.2.5.
6. Principal response curves (PRC) are generally applicable within the ECX-design and a powerful tool for community analyses (van den Brink et al. 2003). They should be carried out in addition to univariate methods when appropriate data are available, for tests with multiple treatments (e.g. ECX design).

Figure 39: Scheme of the statistical testing procedure for earthworm field study data when assessing differences between treatments and controls (e.g. for No Observed Effect Concentrations, NOEC; calculation in Mixed Design)



Source: RWTH Aachen University

3.2.5 Limitations and open questions

The recommendations towards adjustments of the earthworm field study test design reveal two opposing trends whose benefits and downsides for the significance of the test have to be balanced: On the one hand, it is evident that as many concentration levels as possible should be considered for a meaningful EC_x design. From a strictly statistical point of view, replication of the concentration levels is not needed for the subsequent regression analysis.

A strong design for calculating robust NOEC values requires, as shown, a substantial increase in the number of replicates per control and treatments. These two demands need to be weighed and integrated into a new design depending on underlying test concept and desired endpoints. However, this decision is not a strictly statistical one, but primarily a question of feasibility in the field (plot numbers and field sizes to be handled) and a question of regulatory prioritization of various endpoints and protection needs.

In addition, the analyses and underlying data presented above have a few limitations that should not go unmentioned: The results for the implementation of an EC_x design in field studies are based on a proof-of-concept pilot field study at one site and with the well-known reference substance carbendazim. Thus, a sound prior knowledge and experience from earlier field studies on possible effect ranges and dynamics was available. This is not the case, in particular, for new substances in regulatory practice. In such cases, the choice of concentration ranges in earthworm field tests might be considerably more difficult. Furthermore, the pilot field study demonstrates that an applied concentration range usually does not provide derivable dose-response curves for all earthworm species and groups due to varying sensitivities. This problem has also occurred in the previously used NOEC designs due to the different sensitivities of the species. However, the statistical endpoint of the NOEC disguises this to a large extent.

For the derivation of NOEC values with abundance data, CPCAT represents a meaningful alternative to the standardized test procedures of t-test statistics. However, it should be mentioned that there is still no established methodology for the calculation of test power values and corresponding sample planning for CPCAT. In addition, CPCAT should achieve higher acceptance as an appropriate tool for assessing the results of ecotoxicological tests – for example by being applied as a standard analysis method in a wider range of standard ecotoxicological test methods.

The CPCAT procedure is not suitable for metric data because the Poisson distribution does not adequately describe this type of data. In future, however, in order to improve the statistical test procedures for metric data, it might be considered to integrate the closure principle into multiple t-test procedures to prevent alpha inflation. This has already been implemented for binary data (closure principle and Fisher-Freeman-Halton test (CPFISH); Lehman 2018b) and counting data (CPCAT; Lehman et al. 2016). Initial studies on a CP-Williams procedure were also published (Bretz 2011), but its applicability for ecotoxicological test procedures has not yet been investigated.

The use of the samples as replicates for the calculation of NOEC values leads, as shown, to an improvement of the test power. A general investigation of the effects in earthworm field tests at both plot and sample (= subplot) level could therefore be recommended based on these results (provided that ecological conditions exist for the use of subplots as replicates). Whether this is a useful option in consideration of the debate on pseudoreplicates in field studies remains to be discussed. Arguments and counterarguments as well as additional literature on this topic were presented in detail in the study (chapter 3.2.3.1). Within a regulatory framework, the following steps could be considered: A respective endpoint is evaluated at both subplot and plot level. If the same NOEC values are obtained as results, these are considered; if other (smaller) NOEC values are calculated at subplot level, these should be discussed accordingly. If it is not possible to reliably demonstrate a relic of the plot effect at this level, the smaller NOEC should be used in the regulation process. This is not necessarily a decision based on the scientific principle, but a regulatory, protective decision based on precautionary principles.

3.3 Derivation of a new test design

The experience gained during the performance of the pilot study as well as the statistical evaluation of this study and the UBA database were applied to derive a proposal for a new test design. In particular, the following considerations were taken into account:

- (1) The results of best-practice studies (earthworm sampling by combined hand-sorting and formalin/AITC extraction performed according to the current ISO guideline 11268-3) revealed low statistical power to detect differences between control and treatment plots using standardized multiple t-test procedures. The number of replicates (4 plots per treatment) was insufficient to detect small effects (10% to 35%), even for total earthworms which was the group with the lowest variation in the earthworm field data set (see chapter 3.2.3.1.3).
- (2) Field conditions, e.g. the availability of sufficiently large and homogeneous experimental sites, as well as the practical feasibility of field studies limit the total number of treatments, plots and samples (= subplots) per treatment that can be implemented in such a study.
- (3) The results of the pilot study showed that the increase in plot replicates from 4 to 6 increased the test power. However, medium effects (35% to 65%) were often still not detectable with a power of 80%. Thus, the comprehensive detection of small effects with this test power appears to be very unlikely considering the practical limitations regarding the number of plot replicates. Therefore, the use of the individual samples (= subplots) as replicates is suggested (see chapter 3.2.3.1.3). Considering the mean coefficients of variation for control treatments in

the pilot study at subplot level, at least 14 replicates would be necessary to detect medium effects with a statistical test power of 80%. As a best compromise between statistical power and practicability (with regard to suitable test field sizes and a feasible, not too long-lasting sampling), the use of 6 plot replicates with 4 samples per plot is proposed, resulting in 24 replicates at subplot level. The calculation of effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure for given prerequisites. If possible, the CPCAT approach should be used (e.g. for abundance data, see chapter 3.2.3.2.2).

However, as shown in chapter 3.2.1 ('State of the art of statistical procedures to analyse ecotoxicological field tests'), NOEC and related concepts have long been criticized in ecotoxicological literature for good reasons. Therefore, as the first option, the performance of a dose-response design (Table 21 'ECx Design'; effect concentration for x% effect, e.g. EC₅₀ or EC₂₀) with at least 6 test chemical treatments (in order to have at least 3 treatments in the range of the slope of the dose-response curve) plus a control and reference treatment and 3 plots per treatment is recommended. Further concentrations may be added to improve fit of the resulting regression curve. The application rates for the dose-response testing have to be estimated with sufficient confidence before the definitive tests, based on existing information (e.g. laboratory test results, range-finding tests).

Otherwise, a mixed NOEC/ECx design with 2 treatments of 6 plots (for sufficient statistical power in the determination of NOEC/LOEC values) and at least 3 more treatments with 2 plots (for ECx determination) should be carried out. In this case, at least 6 untreated control plots are required (Table 21, 'Mixed Design'). The treatments for the NOEC/LOEC calculation should have the second lowest (T2) and one of the two highest application rates among all tested treatments. The range and spacing of the treatments should be chosen in such a way that it is most likely to obtain both a NOEC and a LOEC at these application rates which will also increase the possibility to derive robust ECx values.

The test follows a randomized design with 4 samples per plot and sampling time point. Depending on the chosen test design, ECx and/or NOEC, LOEC values can be determined. The proposed test design thus mitigates the identified shortcomings of the currently used test design according to ISO guideline 11268-3 while still being practically feasible. To fully exploit the potential of the test design, up to date and selective statistical methods need to be applied to derive robust and meaningful effect values.

Table 21: Number of plots and treatments for the ECx- and the mixed-design in earthworm field tests. More information on the design type in the text above. C control; T 1-x treatments; R reference substance

Test design	Plots per treatment (No.)									Plots (sum)	Samples (total no.)
	C	T1	T2	T3	T4	T5	T6	(T7)	R		
ECx Design	3	3	3	3	3	3	3	(3)	3	24 (27)	96 (108)
Mixed Design	6	2	6	2	2	6			3	27	108

4 Participation in the OECD process (WP3)

The experience gained in the more than 20 past years of performing earthworm field studies based on the existing BBA and ISO guidelines, during the performance of the pilot study and following the statistical evaluation of the results both of the pilot study and of the studies in the UBA database were applied to formulate a new draft OECD TG including a proposal for a new test design. The version of the draft OECD TG that was distributed to the ad hoc SETAC GSIG subgroup in March 2019 can be found in Appendix A.6. This draft was a basis for discussions during the final project meeting at the UBA in Dessau on 28./29. March 2019. The minutes of this meeting are contained in Appendix A.5.2 of this report. Several comments were provided during and after the meeting, which were compiled in a commenting table according to the OECD process. This table is currently under review to create an updated version of the draft guideline that will then be subject to the further OECD process.

5 Conclusions and outlook

The purpose of this project was to provide scientifically robust and practical information on the variability of the endpoints assessed in earthworm field studies, the statistical evaluation of the results and the level of the statistically detectable effects of the chemicals tested with the aim of developing suggestions for improving the test design. Critical evaluation of information available in the literature and the database of the UBA revealed the following shortcomings of the currently used earthworm field test design according to ISO standard 11268-3 (ISO 2014):

- ▶ The evaluated best-practice studies (i.e. using a combination of hand-sorting and formalin/AITC extraction) reveal low statistical power to detect differences between control and treatment plots for aggregated taxa. For single species, this statistical potential for a reliable identification of effects is even lower. The overall MDD is not low enough for a comprehensive detection of small or medium effects.
- ▶ NOEC and related concepts have long been criticized in ecotoxicological literature. Furthermore, the actual MDD calculations of field studies revealed that potentially relevant effects are not detectable in many field situations by the current standardized statistical procedures.
- ▶ An adapted test design should contain an option to perform regression approaches, which have been suggested as an addition to the NOEC approach. The resulting estimated concentrations (EC_x values) from fitting a curve to the data have been proposed as a more meaningful alternative to the NOEC-value. Thus, the number of concentration levels in the pilot field study has to be increased to investigate the suitability of an EC_x-design for earthworm field studies.
- ▶ In order to still include the possibility of deriving NOEC values as well as to improve the statistical power of this procedure compared to the old design, the number of replicates on the plot level for the control and test concentration treatments needs to be increased.
- ▶ The number of samples per replicate should also be increased in order to examine the changes in variance and to estimate if these samples can be used as individual replicates to improve statistical test power.
- ▶ As the field conditions and practical feasibility of the pilot field study limited the total number of plots, the enlargement of the concentration levels and the increase of plots and samples (=subplots) per treatment had to be adjusted in such a way that both research questions (feasibility of EC_x design and improvement of NOEC design) could be addressed.
- ▶ Based on these evaluations, a pilot field study was performed according to a newly developed combined NOEC- and EC_x-test design with the test chemical carbendazim. One control (C) and six treatments (T) were used. The number of plots per treatment differed between six (C, T2, T5) and three (T1, T3, T4, T6). The number of samples per plot was higher than in the currently used ISO guideline 11268-3 (six instead of four). The results of the pilot field study and the in-depth statistical evaluation of additional earthworm field studies yielded the following design requirements for earthworm field studies:

- ▶ Abundance and biomass should be determined and evaluated at species level as aggregated morphological or functional groups may disguise effects on single species.
- ▶ The ECx design is a meaningful alternative to the NOEC design but at least one mixed design would be advisable. The ECx design leads to more robust conclusions for ERA, a masking of possible effects as in the NOEC evaluation is avoided.
- ▶ The calculation of effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure for given prerequisites. If possible, the CPCAT approach is the preferred option.
- ▶ If there are no ecological reasons for not using the data at sample level, the evaluation and interpretation of the data at plot level (pooled samples of 1 m² in total used as replicates) and sub-plot level (single samples as replicates of 0.25 m²) should be requested.
- ▶ Principal response curves (PRC) are generally applicable within the ECx-design and a powerful tool for community analyses. They should be carried out in addition to uni-variate methods when appropriate data are available, i.e. for tests with multiple treatments (e.g. ECx design).

Some limitations and open questions regarding the proposed changes need to be kept in mind:

- ▶ There are two opposing trends whose benefits and downsides for the significance of the test have to be balanced: On the one hand, as many concentration levels as possible should be considered for a meaningful ECx design (with no replication of concentration levels required) while on the other hand a strong design for calculating robust NOEC values requires a substantial increase in the number of replicates per control and each treatment. This question is not a strictly statistical one, but it is also related to the feasibility in the field (plot number and field size) and of the regulatory prioritization of statistical endpoints;
- ▶ The results for the implementation of an ECx design in field studies are based on a proof-of-concept pilot field study at one site and with the well-known reference substance car-bendazim. For new chemicals, the choice of concentration ranges in earthworm field tests might be considerably more difficult;
- ▶ There is still no established methodology for the calculation of test power and corresponding sample planning for CPCAT;
- ▶ The CPCAT procedure is not suitable for metric data because the Poisson distribution does not adequately describe this type of data. In order to improve the statistical test procedures for metric data, it might be considered to integrate the closure principle into multiple t-test procedures in order to prevent alpha inflation;
- ▶ The use of samples as replicates for the calculation of NOEC values leads to an improvement of the test power. A general investigation of the effects in earthworm field tests at both plot and sample (= subplot) level could therefore be recommended (provided that ecological conditions exist for the use of subplots as replicates). This is not necessarily a decision based

on scientific principles, but a regulatory, protective decision based on the precautionary principle.

According to the experiences made in the more than 20 past years of performing earthworm field studies based on the existing BBA and ISO guidelines, during the performance of the pilot study and following the results of the statistical analyses, a draft OECD TG was formulated and provided to the ad hoc SETAC GSIG sub-group for discussion. As of now, the discussion of the draft TG is ongoing.

6 List of references

- Andrade, T.O.; Bergtold, M.; Kabouw, P. (2017): Minimum significant differences (MSD) in earthworm field studies evaluating potential effects of plant protection products. *J. Soils Sediments* 17, p. 1706 – 1714
- Ashauer, R.; Escher, B.I. (2010): Advantages of toxicokinetic and toxicodynamic modelling in aquatic ecotoxicology and risk assessment. *J. Environ. Monit.* 12, p. 2056 – 2061
- BBA (Biologische Bundesanstalt) (1994): Richtlinien für die amtliche Prüfung von Pflanzenschutzmitteln, Nr. VI, 2-3, Auswirkungen von Pflanzenschutzmitteln auf Regenwürmer im Freiland, Braunschweig.
- Bretz, F.; Hothorn, T.; Westfall, P. et al. (2011): Multiple comparisons using R. CRC Press, Boca Raton
- Brock, T.C.; Hammers-Wirtz, M.; Hommen, U.; Preuss, T.G.; Ratte, H.T.; Roessink, I.; Strauss, T.; Van den Brink, P.J. (2015): The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. *Environ. Sci. Pollut. Res. Int.* 22(2), p. 1160 – 1174
- Brown, M.B.; Forsythe, A.B. (1974): Robust tests for the equality of variances. *Journal of the American Statistical Association* 69, p. 364 – 367
- Chapman, P.F.; Crane, M.; Wiles, J.; Noppert, F.; McIndoe, E. (1996): Improving the quality of statistics in regulatory ecotoxicity tests. *Ecotoxicology* 5, p. 169 – 186
- Christensen, E.R. (1984): Dose-response functions in aquatic toxicity testing and the Weibull model. *Water Research* 18, p. 213 – 221
- Crump, K.S. (1995): Calculation of benchmark doses from continuous data. *Risk Anal.* 15, p. 79 – 89
- Davies, G.M.; Gray, A. (2015): Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). *Ecol. Evol.* 5(22), p. 5295 – 5304
- Delignette-Muller, M.L.; Lopes, C.; Veber, P.; Charles, S. (2014): Statistical handling of reproduction data for exposure-response modeling. *Environ. Sci. Technol.* 48, p. 7544 – 7551
- Dudley, R.M. (2014): Central limit theorems. Cambridge University Press, Cambridge
- Dunnett, C.W. (1955): A multiple comparison procedure for comparing several treatments with a control. *Amer. Statist. Ass. J.* 1955 50, p. 1096 – 1121
- Dunnett, C.W. (1964): New tables for multiple comparisons with a control. *Biometrics* 20, p. 482 – 491
- Duquesne, S.; Alalouni, U.; Gräff, T.; Frische, T.; Pieper, S.; Egerer, S.; Gergs, R.; Wogram, J. (2020): Better define beta-optimizing MDD (minimum detectable difference) when interpreting treatment-related effects of pesticides in semi-field and field studies. *Environ. Sci..Pollut. Res.* 27, p. 8814 – 8821
- EC (European Commission) (2013a): Commission Regulation (EU) No 283/2013 of 1 March 2013 setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market (Text with EEA relevance). *Official Journal of the European Union L 93*, p. 1 – 84
- EC (European Commission) (2013b): Commission Regulation (EU) No 284/2013 of 1 March 2013 setting out the data requirements for plant protection products, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market (Text with EEA relevance). *Official Journal of the European Union L 93*, p. 85 – 152
- EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues) (2013): Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA Journal* 11(7):3290

- EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues) (2017): Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. *EFSA Journal* 15(2):4690, 225 pp. doi: 10.2903/j.efsa.2017.4690
- Ekschmitt, K. (1998): Population assessments of soil fauna: General criteria for the planning of sampling schemes. *Applied Soil Ecology* 9, p. 439 – 445
- Finney, D.J. (1971). *Probit Analysis* (3rd ed.). Cambridge Univ. Press, p. 19 – 76
- Fox, D.R. (2009): Is the ECx a legitimate surrogate for a NOEC? *Integrated Environmental Assessment and Management* 5, p. 351 – 353
- Fox, D.R. (2010): A Bayesian approach for determining the no effect concentration and hazardous concentration in ecotoxicology. *Ecotoxicol. Environ. Saf.* 73, p. 123 – 131
- Fox, D.R.; Landis, W.G. (2016): Don't be fooled-A no-observed-effect concentration is no substitute for a poor concentration-response experiment. *Environmental Toxicology and Chemistry* 35(9), p. 2141 – 2148
- Frampton, G.K.; van den Brink, P.J.; Gould, P.J.L. (2000): Effects of spring precipitation on a temperate arable collembolan community analysed using Principal Response Curves. *Applied Soil Ecology* 14, p. 231 – 248
- Green, J.W.; Springer, T.A.; Staveley, J.P. (2012): The drive to ban the NOEC/LOEC in favor of ECx is misguided and misinformed. *Integr. Environ. Assess. Manag.* 9, p. 12 – 16
- Gupta, B.C.; Guttman I. (2014): *Statistics and probability with applications for engineers and scientists*. Wiley, Hoboken
- Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J.; Stahel, W.A. (2005): *Robust statistics: the approach based on influence functions*. Wiley, New York
- Heegaard, E.; Vandvik, V. (2004): Climate change affects the outcome of competitive interactions - an application of principal response curves. *Oecologia* 139, p. 459 – 466
- Hoening, J.M.; Heisey, D.M. (2001) The Abuse of Power. *The American Statistician* 55(1), p. 19 – 24
- Holm, S. (1979): A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, p. 65 – 70
- Horn, M.; Vollandt, R. (1995): *Multiple Tests und Auswahlverfahren*. Biometrie, Gustav Fischer, Stuttgart
- Hurlbert, S.H. (1984): Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs* 54(2), p. 187 – 211
- ISO (International Organization for Standardization) (2014): *Soil quality - Effects of pollutants on earthworms, Part 3: Guidance on the determination of effects in field situations*. ISO 11268-3:2014(E)
- Jager, T. (2011): Some good reasons to ban ECx and related concepts in ecotoxicology. *Environ. Sci. Technology* 45, p. 8180 – 8181
- Jager, T. (2012): Bad habits die hard: The NOEC's persistence reflects poorly on ecotoxicology. *Environmental Toxicology and Chemistry* 31(8), p. 228 – 229
- Jonckheere, A.R. (1954): A distribution-free k-sample test against ordered alternatives. *Biometrika* 41, p. 133 – 145
- Kedwards, T.J.; Maund, S.J.; Chapman, P.F. (1999): Community level analysis of ecotoxicological field studies: II. Replicated-design studies. *Environmental Toxicology and Chemistry* 18(2), p. 158 – 166
- Knacker, T.; Van Gestel, C.A.M.; Jones, S.E.; Soares, A.M.V.M.; Schallnaß, H.-J.; Förster, B.; Edwards, C.A. (2004): Ring-testing and field-validation of a Terrestrial Model Ecosystem (TME) - An instrument for testing potentially harmful substances: Conceptual approach and study design. *Ecotoxicology* 13, p. 9 – 27

- Koijman, S.A.L.M. (1993): Dynamic energy budgets in biological systems. Theory and applications in ecotoxicology. Cambridge University Press, Cambridge
- Koijman, S.A.L.M. (1996): An Alternative for NOEC exists, but the standard model has to be abandoned first. *Oikos* 75(2), p. 310 – 316
- Kruskal, W.H.; Wallis, W.A. (1952): Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 47, p. 583 – 621
- Kula, C.; Heimbach, F.; Riepert, F.; Römbke, J. (2006): Technical Recommendations for the update of the ISO Earthworm Field Test Guideline (ISO 11268-3). *J Soils & Sed.* 6, p. 182 – 186
- Landis, W.G.; Chapman, P.M. (2011): Well past time to stop using NOELs and LOELs. *Integrated Environmental Assessment and Management* 7(4), p. vi – viii
- Laskowski, R. (1995): Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology. *Oikos* 73(1), p. 140 – 144
- Lehmann, R.; Bachmann, J.; Maletzki, D.; Polleichtner, C.; Ratte, H.T.; Ratte, M. (2016): A new approach to overcome shortcomings with multiple testing of reproduction data in ecotoxicology. *Stoch. Environ. Res. Risk Assess.* 30, p. 871 – 882
- Lehmann, R.; Bachmann, J.; Karaoglan, B.; Lacker, J.; Lurman, G.; Polleichtner, C.; Ratte, H.T.; Ratte, M. (2018a): The CPCAT as a novel tool to overcome the shortcomings of NOEC/LOEC statistics in ecotoxicology: a simulation study to evaluate the statistical power. *Environmental Science Europe* 30, p. 50
- Lehmann, R.; Bachmann, J.; Karaoglan, B.; Lacker, J.; Polleichtner, C.; Ratte, H.T.; Ratte, M. (2018b). An alternative approach to overcome shortcomings with multiple testing of binary data in ecotoxicology. *Stoch. Environ. Res. Risk Assess.* 32(1), p. 213 – 222
- Levene, H. (1960): Robust tests for equality of variances. In: Ingram Olkin, Harold Hotelling et al. (eds.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press., p. 278 – 292
- Maud, S.; Chapman, P.; Kedwards, T.; Tattersfield, L.; Matthiessen, P.; Warwick, R.; Smith, E. (1999): Application of multivariate statistics to ecotoxicological field studies. *Environmental Toxicology and Chemistry* 18(2), p. 111 – 112
- Moore, D.R.J.; Caux, P. (1997): Estimating low toxic effects. *Environ. Toxicol. Chem.* 16, p. 794 – 801
- Moser, T.; Römbke, J.; Schallnaß, H.-J.; van Gestel, C.A.M. (2007): The use of the multivariate Principal Response Curve (PRC) for community level analysis: a case study on the effects of carbendazim on enchytraeids in Terrestrial Model Ecosystems (TME). *Ecotoxicology* 16, p. 573 – 583
- Nelder, J.A.; Wedderburn, R.W.M. (1972): Generalized Linear Models. *Journal of the Royal Statistical Society A* 135(3), p. 370 – 384
- OECD (Organisation for Economic Co-operation and Development) (2004): Guidelines for the testing of chemicals No. 222. Earthworm Reproduction Test (*Eisenia fetida/Eisenia andreii*). Paris, France
- OECD (Organisation for Economic Co-operation and Development) (2006a): Current approaches in the statistical analysis of ecotoxicity data: A guidance to application. OECD Series on testing and assessment, 54. ENV/JM/MONO(2006)18.
- OECD (Organisation for Economic Co-operation and Development) (2006b): Guidelines for the Testing of Chemicals 201-Freshwater Alga and Cyanobacteria, Growth Inhibition Test.

- OECD (Organisation for Economic Co-operation and Development) (2012): Guidelines for the Testing of Chemicals 210-Fish Early-Life Stage Toxicity Test: draft revised version.
http://www.oecd.org/env/ehs/testing/Draft%20revised%20TG210-Clean%20version_04-Sept.pdf
- OECD (Organisation for Economic Co-operation and Development) (2016): OECD Guideline for the Testing of Chemicals 222 - Earthworm reproduction test (*Eisenia fetida*/*Eisenia andrei*).
- Römbke, J.; van Gestel, C.A.M.; Jones, S.E.; Koolhaas, J.E.; Rodrigues, J.M.L.; Moser, T. (2004): Ring-Testing and Field-Validation of a Terrestrial Model Ecosystem (TME) – An Instrument for Testing Potentially Harmful Substances: Effects of Carbendazim on Earthworms. *Ecotoxicology* 13, p. 105 – 118
- Ruxton, G.D.; Colegrave, N. (2017): *Experimental Design for the Life Sciences*. Oxford University Press
- Schank, J.C.; Koehnle, T.J. (2009): Pseudoreplication is a pseudoproblem. *J. Comp. Psychol.* 123(4), p. 421 – 433
- Shapiro, S.S.; Wilk, M.B. (1965): An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, p. 591 – 611
- Stephan, C.E.; Rogers, J.W. (1985): Advantages of using regression analysis to calculate results of chronic toxicity tests. In: Bahner R.C. and Hansen D.J (eds.): *Aquatic toxicology and hazard assessment: Eighth symposium*, ASTM STP 891. American Society for Testing and Materials, Philadelphia, p. 328 – 338
- Szoecs, E.; Schafer, R.B. (2015): Ecotoxicology is not normal: A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology. *Environ. Sci. Poll. Res.* <https://doi.org/10.1007/s1135-015-4579-3>
- Tanaka, Y.; Nakamura, K.; Yokomizo, H. (2018): Relative robustness of NOEC and ECx against large uncertainties in data. *PLOS ONE* 13(11), e0206901.
- van den Brink, P.J.; ter Braak, C.J.F. (1998): Multivariate analysis of stress in experimental ecosystems by Principal Response Curves and similarity analysis. *Aquatic Ecology* 32, p. 163 – 178
- van den Brink, P.J.; ter Braak, C.J.F. (1999): Principal response curves: analysis of time-dependent multivariate responses of biological community to stress. *Environmental Toxicology and Chemistry* 18(2), p. 138 – 148
- van den Brink, P.J.; van den Brink, N.W.; ter Braak, C.J.F. (2003): Multivariate analysis of ecotoxicological data using ordination: demonstrations of utility on the basis of various examples. *Australasian Journal of Ecotoxicology* 9, p. 141 – 156
- van den Brink, P.J.; den Besten, P.J.; bij de Vaate, A.; ter Braak, C.J.F. (2009): Principal response curves technique for the analysis of multivariate biomonitoring time series. *Environmental Monitoring and Assessment* 152, p. 271 – 281
- Vollmer, T.; Klein, O.; Frank, S.; Knaebe, S. (2016): Statistical power and MDDs in Earthworm Field Testing. Poster presented at the SETAC Europe Conference in Nantes.
- Walter, H.; Consolaro, F.; Gramatica, P.; Scholze, M.; Altenburger, R. (2002): Mixture toxicity of priority pollutants at no observed effect concentrations (NOECs). *Ecotoxicology* 11, p. 299 – 310
- Warne, M.S.J.; van Dam, R. (2008): NOEC and LOEC data should no longer be generated or used. *Australasian Journal of Ecotoxicology* 14, p. 1 – 5
- Williams, D.A. (1971): A test for differences between treatment means when several dose levels are compared with a zero-dose control. *Biometrics* 27, p. 103 – 117
- Williams, D.A. (1972): The comparison of several dose levels with a zero-dose control. *Biometrics* 28, p. 519 – 531
- Zaborski, E.R. (2003): Allyl isothiocyanate: an alternative chemical expellant for sampling earthworms. *Applied Soil Ecology* 22, 87 – 95