

TEXTE 00/2020

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

Project No. (FKZ) 3715 67 4200

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

Zusammenfassung

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.

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 Projekt-treffen. 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-Freilandstudien-datenbank - 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 Probenahme-flächen 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 Behandlungspartellen 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 Probenahmeterminale 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 Versuchspartellen wurden direkt nach der Applikation mit einem vom Traktor gezogenen Tankwagen mit mindestens 1000 l pro Parzelle (entsprechend 10 mm Niederschlag) bewässert. Die Versuchspartellen 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

Opinion“¹. 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-

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

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 EC_x-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 EC_x-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.