Environmetrics Australia
Environment ● Statistics ● Science

Norwegian Institute for Water Research

The Benefits of Bayesian Statistics

Examples from the Revised OECD Document 54 Annexes

Raoul Wolf
raoul.wolf@ngi.no
Norwegian Geotechnical Institute (NGI)

David R. Fox
david.fox@environmetrics.net.au
Environmetrics Australia

S. Jannicke Moe
jannicke.moe@niva.no
Norwegian Institute for Water Research (NIVA)

OECD Document No. 54 and Its Annexes

- **OECD No. 54 (2006)** is a guide for analysing ecotoxicity study data.
- It stresses **good study design** and **clear, transparent reporting**.
- It compares three main analysis options: **concentration-response models**, **hypothesis testing**, and **mechanistic effect models**.
- The annexes to OECD No. 54 include **real case studies** (OECD TGs 201, 202, 211, 215, 236) for **algae, invertebrates, and fish**.
- OECD No. 54 is **currently being revised** to reflect updated practice since 2006 (Daniels *et al.*, 2025, Wolf *et al.*, 2026).
- **Bayesian methods** will be better represented in the revised OECD No. 54.

Bayesian Concentration-Response Models and Hypothesis Testing

- **Bayesian methods** combine the initial assumptions (the **prior** probability) with the data (the **likelihood**) to give an updated result (the **posterior** probability). **All assumptions are explicit** and the full **uncertainty is quantified**.
- **Models are fully versatile**: they can handle unusual data (non-normal distributions, measurement error, zero inflation, etc.) and combine several assumptions and endpoints.
- **Concentration-response endpoints** (*e.g.*, **EC_{50}**) can be derived directly from the posterior, with uncertainty included.
- Using the region of practical equivalence (**ROPE**; Kruschke *et al.*, 2005), both **difference and equivalence of effects** relative to the control can be tested, supporting statements about **effect presence** *and* **effect absence**.

Concentration-Response Analysis

Model formulation

Situation

In an OECD TG 201 experiment, *R. subcapitata* has been exposed to eight concentrations (x) of atrazine. Growth rates have been reported for individual replicates following the procedure suggested in OECD TG 201; growth rates are reported as estimate (y_n) and (observational) standard error (s_n).

Assumptions

Growth rates likely decrease with increasing atrazine concentrations, following a non-linear concentration-response relationship (log-logistic model, LL.4, *cf.* Ritz *et al.* 2019); x is a continuous variable. Overall, growth rates are assumed to arise from a normal distribution. No hard upper or lower asymptote limits are assumed (growth rates < 0 are biologically possible). The reported standard errors should be used in the modelling approach.

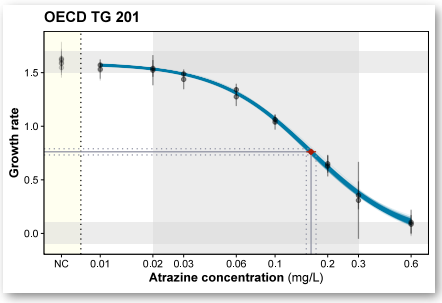
Model

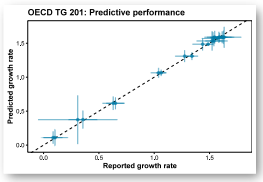
$$y_n \sim \text{Normal}(\mu_n, \sigma) \qquad \mu_n = \begin{cases} d, & x_n = 0 \\ c + \dfrac{d-c}{1+(x_n/e)^b}, & x_n > 0 \end{cases} \qquad \sigma = \sqrt{\sigma_{\text{res}}^2 + s_n^2}$$

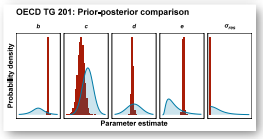
Prior formulation

	b	c	d	e	σ_{res}
Question	What is the expected non-linear slope?	What is the assumed maximum effect on the growth rate?	What is a typical control group growth rate?	What is an expected value for the EC_{50}?	What is the expected residual standard deviation?
Answer	Difficult to answer. The non-linear slope is not easily comparable to a linear slope.	High treatment concentration result in a full reduction of the growth rate.	OECD TG 201 states that typical growth rates for *R. sub-capitata* are between 1.5 and 1.7.	Based on the experimental design, the EC_{50} should be between the second lowest and second highest concentration.	Difficult to answer because of the observation standard error; likely the residual standard deviation is low.
Inter-pretation	A standard Student t distribution allows for a wide range of values due to its tails.	The growth rate should be 0, with the same standard deviation as for the d.	1.5 and 1.7 serve as 95% interval on the normal scale.	0.02 and 0.3 mg/L serve as a 95% interval on the log-normal scale.	A half-Student t distribution places most probability near 0, while also allowing for larger values due to its tail.
Prior	$b \sim \text{StudentT}(3,0,1)$	$c \sim \text{Normal}(0,0.1)$	$d \sim \text{Normal}(1.6,0.1)$	$e \sim \text{LogNormal}(-2.6,0.7)$	$\sigma_{\text{res}} \sim \text{HalfStudentT}(3,0,1)$

Framework	Estimate	Dispersion	Lower	Upper
Bayesian EC_{50}	0.161	0.005	0.151	0.171
*Frequentist EC_{50}**	*0.149*	—	*0.128*	*0.173*







Unbiased EC_{50} estimate (incl. observational error) and credible intervals

Hypothesis Testing

Model formulation

Situation

In an OECD TG 211 experiment, *D. magna* were exposed to six concentrations (x) of an unknown substance. Cumulative offspring (y_n) has been reported per individual, as well as mortality.

Assumptions

Cumulative offspring decreases under treatment conditions b_1–b_6. Cumulative offspring are assumed to be over-dispersed counts, thus following a negative binomial distribution on log-scale. The model will be partially-pooled to minimize type I and type II errors; x is a categorical variable. The region of practical equivalence (ROPE; Kruschke 2010) for cumulative offspring is set to the interval [60,120]. *Not shown: mortality was also modelled simultaneously.*

Model

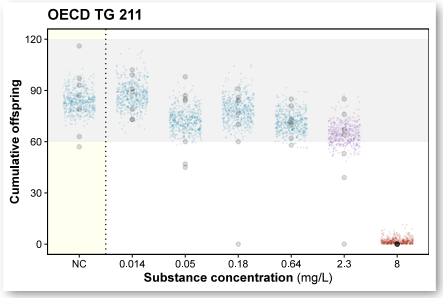
$$y_n \sim \text{NegBinomial}(\eta_{x_n}, \phi) \qquad \eta_{x_n} = a + x_n \cdot b_{c_n} \qquad b_{c_n} = \tau \cdot z_{c_n} \qquad \phi = (1/u)^2$$

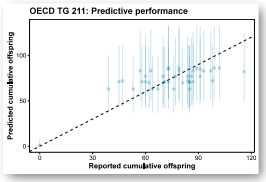
Prior formulation

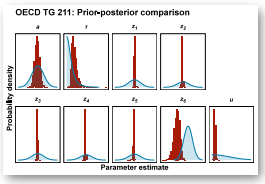
	a	τ	z	u
Question	What is the expected range of cumulative offspring?	What is the assumed average treatment effect on cumulative offspring?	What is a difference to the control cumulative offspring?	What is a realistic value for over-dispersion?
Answer	OECD TG 211 gives a lower limit of 60 for validity. 120 is a commonly observed higher value.	Difficult to estimate without prior knowledge.	Difficult to estimate without prior knowledge.	Over-dispersion itself can be difficult to estimate. Choosing a numerically stable proxy is better.
Inter-pretation	60 and 120 define a 95% interval on log-scale.	The effect size should be positive, and conservatively also allow for larger values.	A standard normal prior will be permissive enough, especially given τ.	Using a half-normal prior on the reciprocal square root is considered safe.
Prior	$a \sim \text{Normal}(4.4,0.2)$	$\tau \sim \text{HalfStudentT}(3,0,1)$	$z \sim \text{Normal}(0,1)$	$u \sim \text{HalfNormal}(0,1)$

Test	Pr(→)	Pr(∈ ROPE)	Pr(∉ ROPE)	Result
$b_1 \Leftrightarrow$ ROPE	0.624	0.996	0.004	Equivalent
$b_2 \Leftrightarrow$ ROPE	0.901	0.963	0.037	Equivalent
$b_3 \Leftrightarrow$ ROPE	0.725	0.990	0.010	Equivalent
$b_4 \Leftrightarrow$ ROPE	**0.921**	**0.956**	**0.044**	**Equivalent**
$b_5 \Leftrightarrow$ ROPE	0.985	0.755	0.245	Inconclusive
$b_6 \Leftrightarrow$ ROPE	**>0.999**	**<0.001**	**>0.999**	**Different**

Framework	"NOEC"	"LOEC"
Bayesian	0.64 mg/L	8.00 mg/L
*Frequentist**	*2.30 mg/L*	*8.00 mg/L*







Partial-pooling for robustness, equivalence and difference testing of the posterior

This poster demonstrates two Bayesian workflows examples for OECD TGs 201 and 211, producing familiar endpoints (EC_{50} and NOEC/LOEC equivalents) while making uncertainty explicit and decision-ready.

Bayesian methods are "good to go": They are fully compatible with routine data from experimental OECD TGs, without additional data requirements.

Clearer uncertainty for decisions: Bayesian outputs support credible intervals and decision-relevant probability statements (*e.g.*, the probability of exceeding a relevant effect level), improving transparency in regulatory interpretation.

Less reliance on "significant": Bayesian decision rules move away from binary *P*-value thresholds that can be sensitive to study power and design choices, instead focusing on interpretations of the posterior probabilities.

More robust EC_x/LC_x reporting: Bayesian model averaging combines predictive distributions across candidate models and can reduce sensitivity to choosing a single curve parametrisation.

ROPE-based hypothesis testing answers both questions on control equivalence (absence of effect) and difference (presence of effect), *de facto* enabling "NOEC/LOEC"-like interpretations with more robustness.

Fits "real world" ecotoxicology data: Bayesian models naturally accommodate endpoint-specific data structures (*e.g.*, quantal outcomes and zero-inflated reproduction counts), enabling consistent reporting across different TGs.

Regulatory advantage: robustness can be demonstrated using predictive diagnostics and prior sensitivity analysis, supporting traceable, reviewable conclusions. Bayesian workflows for "standard" OECD TGs are fully compatible with higher-tier analysis strategies, such weight-of-evidence approaches.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101036756.

References

Kruschke, 2010. Rejecting or accepting parameter values in Bayesian estimation. *Adv. Methods Pract. Psychol. Sci.* **1**:2.
Moe *et al.*, 2025. Updating statistical practice in ecotoxicology: reflections and recommendations. *Integr. Environ. Assess. Manag.* **21**:3.
Daniels *et al.*, 2025. High time to update statistical guidance in ecotoxicology—a workshop synthesis on the revision of OECD document no. 54. *Integr. Environ. Assess. Manag.* **vjaf113** (in print)
Wolf *et al.*, 2026. Progressing statistical analysis for regulatory ecotoxicology: developments, processes, and opinions. *Integr. Environ. Assess. Manag.* **vjag029** (in print)