Review of *Ecotoxicology is not normal - How the use of proper statistical models can increase statistical power in ecotoxicological experiments*

General comments

The manuscript compares a number of methods that are used in ecotoxicology for analyzing non-normally distributed data.

The purpose of the manuscript is highly commendable. Unfortunately, the present manuscript has some major shortcomings: The simulation results, which constitute the key contribution, are far from being as unequivocal as the title may lead one to think. Also, it is disturbing to see a fairly technical (and for ESPR non-standard) formulation of statistical models and yet the authors seem to lack some understanding of the models in some cases.

Specific comments

- About the title: Please consider a more informative title such as "Comparison of statistical approaches for analysis of non-normally distributed response in ecotoxicology". Also in view of the fact that the results are not very clearly favouring the generalized linear models (GLMs).

  p. 2, left column, lines 2–12: The issue of LOEC/NOEC versus regression is interesting but not really relevant to the aim of the manuscript. Consider skipping this part. However, you're right that for small sample sizes it is extremely important to choose the most efficient statistical analysis and the better the model describes the data the more efficient.

p. 2, left column, lines 56–59: You aren't mentioning the issue with back-transformation So, in particular, βx is not additive changes from control *on the original scale*. To many ecotoxicologists such estimates on a transformed scale aren't *directly interpretable* as claimed by the authors. A more detailed explanation or at least discussion on how to back-transform is warranted (this is an issue regardless of whether GLMs or models for transformed responses are used!).

p. 2, right column, line 1: The real advantage of using a generalized linear model for count data shows up in case of many small counts and in the presence of ties. For large counts (e.g., corresponding to a large μ_C) there is practically no difference. However, this difference may not show up in simulations when the sample size is kept between 3 and 9.

p. 2, right column, lines 19–21: It is not the response variable that is linked to the linear predictors, it is the mean of the response variable (McCullagh & Nelder, 1989). This is a basic fact of generalized linear models as you also show in Eqn. (3). Please explain Eqn. (3) in more detail (explain the parameters).

- McCullagh, P. & Nelder, J. A. (1989). Generalized Linear Models, Chapman & Hall, 2nd edition.

p. 2, right column, lines 25–27: There exists no *quasi-Poisson* distribution!!! You use **R** terminology, but this is not in this case statistical terminology. Please consult McCullagh & Nelder (1989) to understand that over-dispersion is dealt with by means of an ad hoc scaling of the standard errors that takes place after having fitted the ordinary Poisson GLM. So Eqn. (4) doesn't make sense.

p. 2, right column, Eqn. (5): Have you explained the parameters in the text?

p. 3, left column, lines 26–42: What about over-dispersion for binomial data? Shouldn't that also be mentioned now that you consider models for over-dispersion for count data.

p. 3, left column, line 33: Why not specify the mean next to the variance in all equations? Or skip the variance where not needed (as in this case).

p. 3, right column, Fig. 1: Please be precise in the figure text: By using the Poisson model the *width* of confidence intervals is underestimated in the presence of overdispersion.

p. 3, right column, line 28: How is multiple testing taken care of? This may also affect the simulation results.

p. 4, left column, line 2: The simulations will not be very informative for such small sample sizes (3–9) as all methods will have low power. Suggestion: Do also consider larger sample sizes: 12, 25, 50, 100. Not completely unrealistic any more as experiments tend to get larger. And then instead keep the μ_C small: 2, 4, 8, 16 to get more ties.

p. 4, left column, lines 10 and 28: In simulation studies it is quite common to use 1000 simulated datasets for each scenario (it would also reduce the sampling variability in the results).

p. 4, left column, line 36: Please do define the type I error properly.

p. 4, right column, lines 26–29: The lack of convergence is not surprising as the simulated datasets are simply too small to fit complex models. Please re-consider the entire concept of the simulation (in particular which scenarios to consider).

p. 4, right column, lines 31–33: One explanation for the reduced power is certainly the multiple testing issue. Perhaps this point could be addressed?

p. 4, right column, lines 57–58: Did LM really maintain the nominal significance level in all cases? This seems to be a subjective assessment (as the level is above in some cases for $N = 3$).

p. 6, right column, lines 43–46: Why didn't you also use the transformation suggested by Brock *et al.* (2015)? Would be useful to see the comparison.

p. 6, right column, line 49: Dunnett test is **R** terminology (yes, it may be found in other recent publications, but it doesn't make it more correct). Please use a term like "comparisons to the control".

p. 7, right column, lines 7–10: A nice result that is not found in many publication: non-parametric approaches lack power (due to less assumptions being made). However, your results for the parametric models are all quite similar and as it stands the paper does not present a strong case for the use of GLMs; and a more balanced abstract would be appropriate.

p. 7, Simulations: A suggestion: remove all diverging discussion of LOEC/NOEC, GLMs for multivariate data, GLMMs, sample size calculation and concentrate on discussing the simulation results. Or, alternatively, discuss in much more detail that GLMMs may actually offer a difference approach for handling over-dispersion in binomial and count data (instead of considering completely different designs).