Reviewer Comment

April 14, 2026

The manuscript discusses a way to calculate confidence intervals for droplet freezing experiments and the importance of a rigorous approach that does not rely on prior assumption. In addition, they discuss a quantitative way to compare two frozen fraction curves and make a good case for why this is relevant and important. The authors thereby discuss a topic, which unfortunately is not as widely discussed as it should be. Robust statistical methods to compare frozen fraction curves and of course also provide a limit on the confidence one can have into its own data is definitely needed in ice nucleation research!

The authors quantify the confidence interval of frozen fraction curves using the non-parametric Kaplan-Meier survival function estimator and compare their results to a previous approach by Fahy et al. 2022. They provide a Github repository for their code, which makes it easy to compare their results and use their approach to other data sets. They use the log-rank test to provide a quantitative answer to the question, whether two frozen fraction curves are similar. They discuss that previous comparisons have used ad-hoc statistical comparisons of curves, which would be great to discuss in more detail. Which methods have been used and why are they less recommendable compare to the log-rank test? Specifically, the log-rank test does assume proportional hazards, which is an assumption that does not hold true to ice nucleation data in literature. Furthermore, the χ^2 statistic assumes that the central limit theorem applies, which again generally does not hold true either for a low number of droplets.

The manuscript can be a good contribution to ice nucleation research, but some major comments need to be addressed, before I can recommend its publication.

1 Major comments

1. At multiple parts of the manuscript, the authors discuss the central limit theorem and its implication to the robustness of their method of calculating the confidence intervals. Figure 2a) specifically highlights that 5 droplets are not enough to render the results meaningful. I realize that there is no perfect answer to the number of droplets needed for the central limit theorem to hold true, but this definitely should be something that needs to be discussed in more detail. Later on, this method is recommended to be used in ice nucleation research, therefore a short discussion on the typical number of droplets used in the literature would definitely add some weight to this statement. In summary, the two questions should be answered in the manuscript:
   a) How large does the population have to be for the central limit theorem to hold true, i.e. for which population sizes can you recommend your method of calculation the confidence intervals?
   b) What are typical population sizes in the literature for frozen fraction curves?

2. The log-rank test is presented as a quantitative way to compare different frozen fractions curves. For that to hold true, the test needs to be flexibly enough to implement it for results from different instruments. This is shown in the manuscript. Secondly, the underlying statistic needs to follow the assumptions made by the log-rank test:
   a) As discussed briefly, the log-rank test assumes proportional hazards, which is not given for ice nucleation. In line 270 the authors discuss that a crossing of two frozen fraction curves is unlikely, but that is certainly not true in general. Especially, when analyzing samples with multiple INP populations and comparing different populations, crossing will occur quite frequently. The authors also propose a way to handle those cases and I think this should be part of the manuscript. This would certainly strengthen the manuscript and broaden the usability of the log-rank test.
   b) Even in the case of two frozen fraction curves not crossing, the assumption of proportional hazards is not a given, for example when one population freezes at higher temperatures, while the other population freezes at lower temperatures. Figure 4 compares different results of the log-rank test, but generally focuses on the comparison of the curves, that have a very similar mean with 15 °C and 16 °C, respectively.

One could make a case, that the log-rank test could be a good way to check if two frozen fractions curves originate from the same distribution, in the case of similar means. Consider to add this in your final statements about the recommendation to use the log-rank test for general comparison.

c) One of the main features of the log-rank test, i.e. the ability to use it for censored data, is discussed in the manuscript (line 99), but is to my knowledge not useful for any experimental data of common instruments in ice nucleation research as the authors rightfully state. Can the authors think of an experimental setup, where this property could be useful?

and it needs to be shown that the log-rank test actually is better than previously used methods. Of course this is a given for the "visual inspection", which is purely subjective and not scientific. For the ad-hoc statistical comparisons, they do not provide examples. Even just a calculation of the absolute error between the two curves is a statistical approach that can certainly provide a quantitative results, which - combined with additional parameters - offer a more detailed comparison of the two curves. Providing examples and comparing them to their approach would certainly strengthen the manuscript and the case for using the log-rank test.

2 General comments

1. Sub- and superscripts that refer to an abbreviation should be upright. For example shown in equation (7): $N_{U,tot}(T)$ -> $N_{\mathrm{U,tot}}(T)$.
2. Temperature can be high or low, but a temperature can never be warm or cold. An example is shown in lines 170-171.
3. There should be a space between the number and the unit. An example is shown in line 219.
4. Variables should be italic, this is also true for Greek symbols. An example is shown in line 212.
5. Tables should generally be set with the least amount of vertical and horizontal lines. While this is ultimately a stylistic choice, I would recommend to remove vertical lines and all horizontal lines apart from the top rule, the middle rule and the bottom rule.

3 Specific comments

1. Line 120: The authors state that the logarithm cannot be calculated, when its argument is zero. While this is true, a more accurate statement would be that the logarithm is not defined in such a case.
2. Line 206: There are many methods available to generate random numbers, also for example within the Python framework. Can the authors provide an explanation for the use of the given website?

3.1 Comments on tables and figures

1. Table 1: The amount of significant digits should be equal across the table and especially for the mean and the standard deviation.
2. Figure 2: The (c) is shifted slightly more to the left than to the right. The colors should also be chosen with consideration of color deficiencies. While the use of different markers is good, the two black squares shown in panels (c) and (d) should be of the same size.
3. Figure 3: Here the label is inconsistent to Figure 2, where "Fraction frozen" was used. The same axis labels should be chosen throughout the manuscript. The colors should also be chosen with consideration of color deficiencies. The amount of tick labels should also be reduces, i.e. every 4 K.
4. Table 2: Same comment as for table 1. In addition, I do not see the value of reporting the numerical value of the p value.
5. Figure 4: The graphic is very blurry and a version should be provided with more pixels.

3.2 Technical comments

1. Line 286: The approximate symbol should be used: "..., whereas $\approx$ 50 droplets...".
2. Line 298: Full stop missing.

References

Fahy, W. D., C. R. Shalizi, and R. C. Sullivan (2022). "A universally applicable method of calculating confidence bands for ice nucleation spectra derived from droplet freezing experiments". In: *Atmospheric Measurement Techniques*. DOI: 10.5194/amt-2022-141.