

Response to Reviewers

Thomas F. Whale, Trystan Surawy-Stepaney, and Sarah L. Barr

Colour key. Reviewer comments are shown in blue. Author responses are shown in black. Text quoted from, or explicitly added to, the revised manuscript is shown in green.

General response

We thank all three reviewers for their careful and constructive comments. We agree that the original manuscript presented the log-rank test too broadly and did not sufficiently assess the performance of the Kaplan–Meier confidence intervals or the effects of droplet number and curve shape.

We have therefore substantially revised the manuscript and changed its title to “Survival-analysis methods for droplet-freezing data: Kaplan–Meier confidence intervals and non-parametric curve comparison” to reflect its broader scope. The revised paper assesses the empirical coverage of Kaplan–Meier pointwise confidence intervals and evaluates the false-positive rates and power of several non-parametric comparison tests using realistic fitted freezing curves. We now treat the log-rank test as useful for particular proportional-hazards-like alternatives rather than as a default method. For more general curve-shape differences, we compare established empirical-distribution tests and introduce an Integrated Squared Difference permutation test operating directly on fraction-frozen curves. We have also added Supplementary Information describing the fitted distributions, statistical methods and Python implementation.

We believe these changes provide a more careful and practically useful framework for quantifying uncertainty and comparing droplet-freezing curves, and we are grateful to the reviewers for prompting this substantial improvement.

Detailed response to reviewer 1

1. Overall assessment and scope of the manuscript

Reviewer comment

In “Survival analysis for droplet-freezing data: Kaplan–Meier confidence intervals and log-rank tests,” Whale et al. present an approach to quantify uncertainty in droplet-freezing data and test differences in frozen fraction curves based on the non-parametric Kaplan–Meier survival function estimator. By applying this method, they are able to derive confidence intervals for frozen fraction

curves and cumulative ice nucleation activity spectra. They also adapt the log-rank test to droplet freezing data to hypothesis test whether two frozen fraction curves are identical or not. They then demonstrate both methods on selected literature and simulated datasets. The method of calculating confidence intervals they present is rigorous, easy to use, and addresses a persistent problem of a lack of standardization in ice nucleation statistics. However, I have much more significant concerns surrounding the use of the log-rank test for ice nucleation data. Specifically, the hypothesis being tested by the log-rank test is limited in the context of ice nucleation, where in many samples there may be multiple different populations of ice active species that dominate at different temperatures in the frozen fraction spectrum. The requirement of proportional hazards is also a concern as many frozen fraction curves in literature do not meet this assumption, further limiting the usefulness of this test. Given that it can also only be applied to frozen fraction curves (and so generally cannot be used to compare across instruments or across samples at different concentrations), I am not convinced of the broad applicability that the authors imply when recommending this statistical test be used. When combined with the fact that the method of calculating confidence intervals has been previously published (Kinney et al., 2024), I am not convinced that this manuscript in its current form constitutes a sufficiently novel and significant contribution to the field. I encourage the authors to consider whether refinements mentioned in the manuscript that address the shortcomings of the log-rank test or other approaches to statistical testing in the context of survival analysis might be more appropriate and useful to a wider range of applications. Specific comments and suggestions are below.

Response

We thank the reviewer for this careful and helpful summary of the manuscript, and for recognising the value of the Kaplan–Meier confidence interval approach. We also agree with the reviewer’s main concern. In the original manuscript we presented the ordinary log-rank test too strongly, and did not sufficiently delimit the question it answers or the assumptions under which it is likely to be useful. In particular, we agree that many ice-nucleation datasets may contain multiple ice-active populations that dominate in different parts of the freezing spectrum, and that the proportional-hazards assumption may often be doubtful. We also agree that comparison of fraction-frozen curves is not always the most physically meaningful approach, especially when comparing different concentrations, surface areas, masses, droplet volumes, or instruments.

We have therefore revised the manuscript substantially. The revised paper is no longer framed as a recommendation to use the log-rank test as a general-purpose comparison for droplet-freezing data. Instead, we now present the log-rank test as one useful test for a particular kind of alternative: a broadly persistent, proportional-hazards-like difference between two directly comparable droplet-level datasets. We have added a much fuller discussion of the proportional-hazards assumption, including the point that this assumption can fail even when curves do not visibly cross.

In response to the reviewer’s suggestion, we have also added and tested alternative approaches. Most importantly, we introduce an Integrated Squared Difference (ISD) permutation test. This test compares the accumulated squared separation between two empirical fraction-frozen curves over a specified temperature interval. It is designed to address the more general question of whether two directly comparable fraction-frozen curves differ somewhere over the analysed range, without requiring proportional hazards and without allowing differences in different parts of the curve to cancel. We

also compare this test with standard empirical-distribution tests, including Kolmogorov–Smirnov, Cramér–von Mises, and Anderson–Darling tests.

We have added simulation-based assessments of the false-positive rates and false-negative rates of these tests. These simulations use fitted representations of literature droplet-freezing curves, allowing us to test the methods against realistic curve shapes while knowing the underlying generating distributions. The results show that the ordinary log-rank test can perform well for simple shifted-curve alternatives, but can have much lower power for non-proportional or mixture-population cases. This directly addresses the concern that the original manuscript implied broader applicability than was justified.

We also agree that the novelty of the manuscript needed clearer framing, given that Kaplan–Meier confidence intervals were previously applied to droplet-freezing data in Kinney et al. (2024). However, that earlier application resulted from the work developed for the present study and did not provide a general explanation or validation of the method. The revised manuscript now assesses the approach systematically: it explains and implements Kaplan–Meier pointwise confidence intervals for fraction-frozen curves and derived cumulative spectra, tests their empirical coverage using realistic fitted freezing curves with known underlying distributions, and shows how their usefulness depends on droplet number, curve shape, and position within the freezing transition. To our knowledge, this evidence for the practical utility and limitations of Kaplan–Meier intervals for droplet-freezing data is not otherwise available in the literature.

Overall, we think the reviewer’s criticism was fair, and it has led us to reframe the manuscript substantially. The revised paper now aims to provide a practical non-parametric framework for uncertainty quantification and curve comparison, rather than promoting the ordinary log-rank test as a default method.

2. Droplet number and the asymptotic approximation

Reviewer comment

The calculation of confidence intervals with the Kaplan–Meier estimator uses the asymptotic normal distribution, which requires that the central limit theorem apply. It is not clear from the manuscript how many droplets are required for this assumption to be accurate. In Figure 2, the authors consider droplet freezing experiments with different numbers of droplets and conclude that the number of droplets used is “essentially a matter of taste” (excluding the simulation with 5 droplets). However, in using this statistical approach, the sample size matters to the accuracy of the confidence intervals – I suggest the authors at least discuss this in more detail and give guidelines on a population size required to meet the central limit theorem requirement, or to achieve a defined statistical power.

Response

We agree. The statement that droplet number was “essentially a matter of taste” was too casual and has been removed. The Kaplan–Meier intervals used here are pointwise intervals based on an asymptotic approximation, so their performance must be checked for the droplet numbers and curve shapes relevant to these experiments.

We have added an empirical coverage assessment using synthetic experiments generated from fitted literature freezing curves. Because the underlying generating curves are known, we can calculate directly how often nominal 95 % intervals contain the value they are intended to estimate. Figures 1–3 show that coverage is close to nominal over most of the curves for experimentally realistic droplet numbers, but deteriorates near the ends of the freezing transition and for very small experiments.

We have also removed any suggestion of a universal minimum droplet number. The revised text now states that useful precision depends on droplet number, curve steepness and the temperature region of interest. Experiments with tens of droplets can provide useful intervals over the central part of steep curves, while broad or multi-population curves generally require more droplets. Droplet number is therefore reported as essential contextual information rather than treated as a simple pass/fail threshold.

3. Differential ice-nucleation activity spectra

Reviewer comment

Is it possible to apply this estimator to differential ice nucleation activity spectra, in addition to the cumulative spectra presented here? This would extend its usefulness.

Response

We agree that this could be useful, but it is beyond the scope of the present study. Kaplan–Meier intervals propagate directly to cumulative quantities because these are monotonic transformations of $f(T)$. Differential spectra require an additional differencing, binning, smoothing or fitting step, and their uncertainty depends strongly on those choices. Developing a rigorously calibrated approach for differential spectra would be a useful subject for future work. We have added the following text after introducing the Kaplan–Meier intervals:

“Kaplan–Meier confidence intervals propagate naturally to cumulative spectra because these are direct transformations of $f(T)$. Differential spectra are less straightforward because they require an additional differencing, binning, smoothing, or fitting step, and their interpretation depends on how the freezing-temperature data are represented, as discussed by Vali (2023) in relation to Fahy et al. (2022). For this reason, we focus here on fraction-frozen curves and cumulative spectra, where the transformation from the Kaplan–Meier estimator is direct.”

4. Utility and interpretation of global tests

Reviewer comment

The foremost issue to me is that simply testing whether two frozen fraction curves are different has limited utility and could be misleading in some use cases or if not interpreted carefully. Consider the case of Figure 4c and d. The authors state that the log-rank test produces a correct result in distinguishing the two curves in Figure 4d, while in Figure 4c, the test produces a Type II error. This interpretation is correct under the hypothesis of the log-rank test. However, it is immediately clear from both panels that the curves crossing cause each of the two simulated samples to have

a higher ice nucleation activity at different points in the spectrum. In panel d, it would be far more physically meaningful to say that the curve in red has higher ice nucleation activity at high temperatures, and then meets the curve in grey, which then has a higher ice nucleation activity at low temperatures where its slope is large. These conclusions can be drawn from the confidence bands themselves, and could also be drawn from ice nucleation activity spectra. In contrast, applying the log-rank test tells the researcher that there is merely a difference between the samples, a far less useful result. The authors should further justify and discuss in what situations this specific test should be used and is more useful than simply comparing confidence intervals. They should also caution the reader about the limitations of the hypothesis being tested.

Response

We agree. The original manuscript placed too much emphasis on whether a test rejected the global null hypothesis and did not make sufficiently clear that this is only one part of the interpretation. A significant global test does not identify where the curves differ, which curve is more active over a particular temperature range, or what physical mechanism is responsible.

The revised manuscript now states this limitation explicitly in the abstract, statistical background and conclusions. We have also removed the former language describing individual test results simply as “correct” or “incorrect”. The revised analysis instead quantifies false-positive rates, power and false-negative rates over repeated simulations.

We nevertheless think there is value in a formal global comparison. Its role is to ask whether the observed separation between two finite-droplet curves is larger than expected from sampling variability alone. This is particularly useful in marginal cases where curves overlap, droplet numbers are modest, or the freezing transition is broad. Any significant result must still be interpreted using the curves, their confidence intervals and, where appropriate, cumulative spectra.

In response to this and related comments, we no longer recommend the ordinary log-rank test generally. We compare several empirical-distribution tests and introduce the ISD permutation test, which accumulates squared separation between fraction-frozen curves and does not rely on proportional hazards. The aim of the revised comparison framework is not to replace interpretation of the curves, but to provide a clearer and more reproducible statistical complement to it.

We now state in the Conclusions:

“The ISD test is a global test: a significant result indicates that two curves differ somewhere over the analysed temperature interval, but does not identify where they differ or why. Those questions still require inspection of the curves and, where appropriate, comparison of physically normalised quantities such as $K(T)$, $n_s(T)$, or $n_m(T)$.”

At the end of the Abstract, we now state:

“We note finally that a significant global test indicates that two curves differ somewhere over the analysed temperature interval, but does not identify where they differ or why; this still requires inspection of the curves.”

5. Proportional hazards and non-crossing curves

Reviewer comment

The second issue, which the authors acknowledge, is the requirement of proportional hazards in the log-rank test, i.e. that the ratio of freezing probabilities is constant across the experiment. Notably, the curves need not cross for this assumption to be invalid. The authors argue that violation of the proportional-hazards assumption, or specifically crossing of frozen fraction curves, is unlikely in literature. I disagree. Even in the specific use cases of ageing of laboratory samples that the authors advocate for using the log-rank test, many counterexamples exist in literature. Further discussion is needed in the manuscript to clarify how violating the proportional-hazards assumption influences the results of the log-rank test, and ideally the authors should implement a version of the log-rank test that is robust to non-proportional hazards.

Response

We agree. The original manuscript treated visibly crossing curves as the main example of non-proportional hazards and implied that such behaviour was relatively unusual. That was not justified. As the reviewer notes, proportional hazards can fail without literal curve crossing whenever the relative freezing probabilities change across the temperature range—for example, because of ageing effects, background freezing, or multiple ice-active populations.

We have removed the suggestion that non-proportional behaviour is uncommon and added a fuller explanation of how violation of proportional hazards affects the ordinary log-rank test. In particular, signed observed-minus-expected contributions from different parts of the temperature range can partly cancel, substantially reducing power.

We have also added simulations demonstrating this directly. The log-rank test performs well for some simple shifted-curve alternatives in Figs. 5–6, but has much lower power for the steep-versus-broad comparison in Figs. 7–8 and the mixture-population example in Figs. 9–10. The revised recommendation is therefore to use the log-rank test only when a proportional-hazards-like difference is a reasonable expectation. For more general curve-shape differences, we recommend the ISD test or an appropriate empirical-distribution test.

6. More difficult and multi-population simulations

Reviewer comment

The data used to demonstrate these approaches are generally well-behaved for this test, except for the data in Figure 4c and 4d, which break the assumptions of the test in very controlled, specific ways. I would encourage the authors to test the log-rank test with other, less well-behaved simulated datasets such as those generated from multiple separate Gaussians, and to demonstrate the limits of the log-rank test as a function of population size.

Response

We agree, and have replaced the limited illustrative examples in the original manuscript with a broader simulation study. The revised simulations use fitted literature-like curves spanning steep,

intermediate and broad freezing behaviour, together with deliberately constructed shifted, strongly non-proportional and two-component mixture scenarios.

The mixture example in Figs. 9–10 contains the same two Gaussian freezing populations in different proportions and was designed specifically to test the situation described by the reviewer. We now report false-negative rates as a function of droplet number, showing directly how power depends on sample size, curve shape and the magnitude and form of the difference. We have also added null simulations in Fig. 4 to establish that differences in power are not simply caused by large differences in false-positive behaviour.

7. Fraction-frozen curves versus physically normalised spectra

Reviewer comment

It is not clear to me what the benefit of testing differences between specifically two frozen fraction curves is, as opposed to the volume-, surface-area-, or mass-normalised cumulative ice-nucleation activity spectra $K(T)$, $n_s(T)$, or $n_m(T)$. A stronger argument should be made for the need for using this specific test. Other rigorous approaches to comparing differences between ice-nucleation spectra already exist, and confidence intervals or confidence bands can also be compared. Therefore, a stronger argument for using the log-rank test on frozen fraction curves over these other approaches should be presented in the text.

Response

We agree with much of this comment, and it has helped us clarify what the curve-comparison tests are for. In the original manuscript we did not explain this well enough, and we agree that we implied the log-rank test was more generally useful than it is. We also agree that simply asking whether two fraction-frozen curves are “different” is a limited question. A significant global test does not say where the curves differ, which sample is more active over a particular temperature interval, or what physical mechanism is responsible. Those questions still require inspection of the plotted curves and, where appropriate, comparison of normalised spectra.

However, we do think there is value in quantifying the evidence that two curves differ, provided the result is interpreted carefully. The purpose of a formal test is not to provide the physical interpretation of the difference, but to ask a narrower statistical question: is the observed separation between two finite-droplet curves larger than would be expected from sampling variability alone if the droplets were drawn from the same underlying freezing-temperature distribution? In that sense, a p value gives a reproducible measure of evidence against the no-difference null hypothesis, with a defined false-positive rate. This can be useful in marginal cases where curves overlap, droplet numbers are modest, or the freezing transition is broad.

In the Abstract we now state:

“These tests quantify whether the observed separation between two finite-droplet curves is larger than expected from sampling variability alone, while recognising that a significant global test does not identify where the curves differ or why.”

In the Introduction we now state:

“Although droplet-freezing datasets can be analysed in several defensible ways, the field lacks a widely adopted statistical framework for reporting uncertainty and quantitatively comparing fraction-frozen curves. Decisions about whether two curves represent meaningfully different freezing behaviour are often based on visual inspection or informal criteria. Visual comparison of fraction-frozen curves remains essential, because it shows where curves differ and suggests possible physical interpretations. However, visual comparison alone can be ambiguous when droplet numbers are small, curves are broad, or confidence intervals overlap over part of the temperature range. It also does not provide a pre-specified false-positive rate or an estimate of statistical power. Formal tests are therefore useful as a complement to graphical interpretation, indicating whether the observed separation between two finite-droplet curves is larger than expected from sampling variability alone.”

We have also changed the paper’s position on the log-rank test. We no longer argue that the ordinary log-rank test should be used as a general comparison for fraction-frozen curves, or that it is preferable to comparison of physically normalised spectra. Instead, we present it as useful only when a proportional-hazards-like difference is a reasonable expectation, such as a broadly shift-like change in freezing behaviour. For more general curve-shape differences, we now compare several non-parametric alternatives and introduce the ISD permutation test, which directly tests the accumulated separation between the plotted fraction-frozen curves.

Finally, we agree that normalised spectra are usually the more meaningful quantities when comparing across concentrations, surface areas, masses, droplet volumes, or instruments. We have added this caveat explicitly. The revised manuscript therefore does not imply that global tests on $f(T)$ replace physically normalised spectra.

Specific comments

8. Priority of Monte-Carlo confidence intervals

Reviewer comment

Line 187: These Monte-Carlo confidence intervals were first presented earlier than Whale et al. 2022, for example in the supplemental information of Jahl et al. (2021). I recommend removing this line, as the earliest use of this method is not important to this study.

Response

We agree. This was clumsy wording rather than an intended claim of priority. We were referring to the particular implementation used in the original comparison. That comparison is no longer central to the revised paper and the sentence has been removed. The revised manuscript instead assesses Kaplan–Meier interval performance directly through empirical coverage simulations.

9. Pointwise intervals versus simultaneous confidence bands

Reviewer comment

In the figures of this paper, the Kaplan–Meier confidence intervals are presented as continuous bands,

apparently through interpolations between the calculated confidence intervals. This should be stated somewhere, as continuous confidence bands are not precisely equivalent to interpolated discrete confidence intervals (which are less accurate than true confidence bands; see Sachs et al., 2022).

Response

We agree and thank the reviewer for identifying this. We have revised the manuscript throughout to describe the intervals as pointwise confidence intervals, not confidence bands. The shaded areas shown in the figures are visual connections between pointwise intervals and should not be interpreted as simultaneous confidence bands. This distinction is now stated explicitly, with reference to Sachs et al. (2022).

10. Software and example data

Reviewer comment

The provided script and example data work on my machine, and the methods used are appropriate. I appreciate the authors providing an easy-to-use script for this approach as a supplement to the manuscript.

Response

We thank the reviewer for checking the script and are pleased that it worked as intended. We have updated the accompanying notebooks and scripts to reflect the revised manuscript. The updated material now includes separate notebooks for calculation of Kaplan–Meier pointwise confidence intervals, the ordinary log-rank test, and the new ISD permutation test.

Detailed response to reviewer 2

We thank the reviewer for the careful and constructive review, and for recognising the need for clearer and more rigorous statistical treatment of droplet-freezing measurements. We agree with the two main concerns raised here. First, the original manuscript did not do enough to establish how the Kaplan–Meier pointwise confidence intervals behave as a function of droplet number and curve shape. Second, we presented the ordinary log-rank test too strongly without adequately addressing the proportional-hazards assumption or comparing it with suitable alternatives.

We have revised the manuscript substantially. The revised paper now includes an empirical coverage assessment of the Kaplan–Meier intervals, repeated simulations of false-positive rates and power, fitted multi-component freezing distributions based on literature data, and several non-parametric alternatives to the log-rank test. We have also introduced an Integrated Squared Difference (ISD) permutation test, which compares the accumulated separation between two fraction-frozen curves directly and does not require proportional hazards. The log-rank test is retained as a useful method for particular shift-like alternatives, but is no longer recommended as a general-purpose comparison for droplet-freezing data.

Major comments

1. Population size and the central limit theorem

Reviewer comment

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 calculating the confidence intervals?

Response

We agree that the original treatment of this issue was inadequate. In particular, the statement that the number of droplets was “essentially a matter of taste” was too casual and has been removed.

There is no sharp droplet-number threshold at which the asymptotic approximation suddenly becomes valid. Its usefulness depends on the number of droplets, the shape and steepness of the underlying freezing curve, and the part of the curve being interpreted. We have therefore addressed this empirically rather than trying to assign a universal central-limit-theorem threshold.

In the revised manuscript, we generate repeated synthetic experiments from fitted representations of three literature freezing curves. Because the underlying generating curves are known, we can calculate directly how often nominal 95 % Kaplan–Meier pointwise confidence intervals contain the values they are intended to estimate. These simulations show that the intervals behave reasonably well over much of the main freezing transition for droplet numbers typical of many experiments, but that coverage deteriorates near the tails, where finite experiments contain little information. Very small experiments produce coarse curves and broad intervals and should be treated cautiously.

We now state that experiments with tens of droplets can provide useful pointwise uncertainty estimates over the main freezing transition, particularly for steep curves, but that broad, long-tailed or multi-population curves generally require more droplets for comparable precision. We deliberately avoid presenting approximately 50 droplets, or any other number, as a universal threshold.

Reviewer comment

What are typical population sizes in the literature for frozen fraction curves?

Response

We agree that this context is useful and have added the following text to the Introduction:

“Many droplet-freezing datasets contain tens to hundreds of droplets per measurement, although the number varies widely with instrument type. Microlitre-scale assays typically freeze on the order of 40–100 droplets per experiment (Whale et al., 2015; Budke and Koop, 2015). High-throughput plate systems can use several hundred wells (Miller et al., 2021; Kunert et al., 2018), and microfluidic systems can generate hundreds to thousands of droplets (Stan et al., 2009; Tarn et al., 2025). On occasion, measurements with fewer than 10 freezing events in larger water volumes are reported, e.g. Kobayashi et al. (2016).”

This wide range reinforces the need to report droplet number alongside fraction-frozen curves and confidence intervals. The apparent smoothness of a plotted curve can otherwise obscure substantial

differences in the statistical resolution of different experiments.

2. Applicability and assumptions of the log-rank test

2a. Proportional hazards and crossing curves

Reviewer comment

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 analysing 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.

Response

We agree. The original manuscript placed much too much emphasis on literal curve crossing and incorrectly suggested that such behaviour would be unusual. We have removed that statement.

The revised manuscript now explains that crossing is only one obvious visual example of non-proportional hazards. The proportional-hazards assumption may also fail when curves do not cross, for example when different ice-active populations dominate different parts of the freezing spectrum or when a sample approaches the instrumental background at lower temperatures.

We have added simulations specifically addressing these cases. These include a comparison between steep and broad fitted freezing curves and a controlled two-component mixture example in which two samples contain the same freezing populations in different proportions. The results show that the ordinary log-rank test can lose substantial power in such cases.

We have also implemented the broader approach requested by the reviewer. The revised paper compares the log-rank test with Kolmogorov–Smirnov, Cramér–von Mises and Anderson–Darling tests, and introduces an integrated squared fraction-frozen distance (ISD) permutation test. The latter compares the accumulated squared separation between two empirical fraction-frozen curves and does not require proportional hazards. The revised manuscript therefore no longer presents the ordinary log-rank test as a general solution for crossing or multi-population curves.

2b. Non-proportional hazards without curve crossing

Reviewer comment

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 curves that have very similar means. One could make a case that the log-rank test could be a good way to check if two frozen fraction curves originate from the same distribution in the case of similar means. Consider adding this in your final statements

about the recommendation to use the log-rank test for general comparison.

Response

We agree that the original manuscript did not distinguish clearly enough between visible crossing and the more general proportional-hazards assumption. This has now been corrected throughout.

The revised text explains that proportional hazards can fail whenever the ratio of conditional freezing probabilities changes with temperature, whether or not the plotted curves cross. We therefore do not recommend the log-rank test simply because two curves have similar means, nor do we assume that non-crossing curves necessarily satisfy proportional hazards.

Instead, the ordinary log-rank test is now described as most useful when the expected difference is a broadly persistent, proportional-hazards-like change in freezing probability. A relatively simple temperature shift is one important example. Figures 5–6 show that the log-rank test can perform well for these shifted-curve alternatives. In contrast, Figs. 7–10 show that it can perform poorly for curve-shape and mixture-population differences.

We think this provides a clearer basis for deciding when the log-rank test is likely to be useful. It is retained as one test for a particular type of alternative, rather than recommended as a general comparison for all fraction-frozen curves.

2c. Usefulness of censoring

Reviewer comment

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, but is to my knowledge not useful for any experimental data of common instruments in ice-nucleation research. Can the authors think of an experimental setup where this property could be useful?

Response

We agree that censoring is unlikely to be an important advantage for most conventional droplet-freezing experiments, and we have reduced the emphasis placed on it.

A possible use would be an experiment that ends before all droplets have frozen, for example because the instrument reaches its minimum temperature. Droplets remaining unfrozen at the end of the experiment could then be treated as censored observations rather than assigned an artificial freezing temperature. Censoring could also arise if individual droplets became unobservable or were lost during an experiment, although such cases would need to be considered carefully because the censoring mechanism should not itself be related to freezing behaviour.

In practice, good experimental design should usually avoid these situations. We now present the ability to accommodate censoring as a technical feature of the Kaplan–Meier framework rather than as a major motivation for its use in ice-nucleation experiments.

3. Comparison with previously used approaches

Reviewer comment

It needs to be shown that the log-rank test actually is better than previously used methods. Of course this is a given for “visual inspection”, which is purely subjective and not scientific. For the ad hoc statistical comparisons, the authors 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 result, which—combined with additional parameters—offers 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.

Response

We agree that the original manuscript did not justify presenting the ordinary log-rank test as preferable to other quantitative approaches, and we no longer make that claim. We also do not think there would be much value in systematically comparing the revised methods with the ad hoc criteria previously used in the literature. These include, for example, comparing a shift in median freezing temperature with a multiple of the instrument temperature uncertainty. Such approaches may be useful for particular experiments, but they do not provide a general test of whether two complete fraction-frozen curves differ.

Instead, the revised paper compares the log-rank test with several established non-parametric tests and the new ISD permutation test. We assess their false-positive rates and power under several realistic curve shapes and alternatives. The results show that no single test is best in every situation, and the revised manuscript recommends choosing a test that is reasonably matched to the scientific question rather than treating the log-rank test as generally superior.

General comments

4. Upright sub- and superscripts

Reviewer comment

Sub- and superscripts that refer to an abbreviation should be upright. For example, as shown in Eq. (7): $N_{U,\text{tot}}(T)$.

Response

We agree and have revised the mathematical notation throughout.

5. Warmer and colder temperatures

Reviewer comment

Temperature can be high or low, but a temperature can never be warm or cold. An example is shown in lines 170–171.

Response

We appreciate the reviewer's preference, but we have retained "warmer" and "colder" where these terms improve clarity. These are established usages in scientific English, and in ice-nucleation work they help avoid ambiguity between numerical temperature and the magnitude of supercooling. For example, "warmer-freezing component" unambiguously describes a component active closer to 0 °C, whereas "higher" or "lower" can be interpreted in several ways for negative temperatures. We have nevertheless checked each use and replaced the terms with "higher temperature" or "lower temperature" where that wording is clearer.

6. Mathematical variables

Reviewer comment

Variables should be italic; this is also true for Greek symbols. An example is shown in line 212.

Response

We agree and have checked the equations and surrounding text for consistent mathematical formatting. Variables and Greek symbols are now italicised, while descriptive labels and abbreviations are set upright.

7. Table formatting

Reviewer comment

Tables should generally be set with the least amount of vertical and horizontal lines. While this is ultimately a stylistic choice, I would recommend removing vertical lines and all horizontal lines apart from the top rule, the middle rule and the bottom rule.

Response

We thank the reviewer for this suggestion. The tables referred to in the original manuscript have been removed as part of the substantially revised analysis. Any remaining tables use a simpler layout without unnecessary vertical or horizontal rules.

8. Number and unit spacing

Reviewer comment

There should be a space between the number and the unit. An example is shown in line 219.

Response

We agree and have checked the presentation of numerical values and units throughout. Following the convention used by Copernicus, a space is placed between the numerical value and the complete unit symbol, for example 2 °C. There is no space between the degree symbol and C.

Specific comments

9. Logarithm at zero

Reviewer comment

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.

Response

We agree and have changed the wording to state that the logarithm is not defined when its argument is zero.

10. Random-number generator website

Reviewer comment

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?

Response

We agree that there was no good reason to use an external website for this purpose. The website and associated discussion have therefore been removed. The revised simulations use pseudorandom-number generation within Python, rather than an external random-number service.

Comments on tables and figures

11. Table 1 precision

Reviewer comment

Table 1: The amount of significant digits should be equal across the table and especially for the mean and the standard deviation.

Response

We agree. The original table has been removed during the substantial restructuring of the manuscript. Numerical values reported in the revised tables, figures and captions have been checked for consistent and appropriate precision.

12. Figure 2 presentation and colour accessibility

Reviewer comment

Figure 2: The (c) is shifted slightly more to the left than to the right. The colours should also be chosen with consideration of colour 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.

Response

We agree. The original figure has been replaced as part of the revised simulation analysis. The new figures use consistent panel positioning, marker sizes and line styles. We have also used combinations of colour, marker shape and line style so that the figures do not rely on colour alone to distinguish datasets or tests.

13. Figure 3 labels, colours and ticks

Reviewer comment

Figure 3: Here the label is inconsistent with Figure 2, where “Fraction frozen” was used. The same axis labels should be chosen throughout the manuscript. The colours should also be chosen with consideration of colour deficiencies. The amount of tick labels should also be reduced, i.e. every 4 K.

Response

We agree. The original figure has been replaced, and the revised figures have been checked for consistent use of “fraction frozen” or “fraction unfrozen”, depending on the quantity actually plotted. Axis labels, tick spacing, colour choices and line styles have also been standardised throughout the revised manuscript.

14. Table 2 precision and p values

Reviewer comment

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.

Response

We agree that the original table reported more numerical precision than was useful. The table has been removed during the restructuring of the manuscript.

Where p values are reported in the analysis notebooks, they remain available for reproducibility, but the manuscript does not present long numerical p values that imply unjustified precision.

15. Figure 4 resolution

Reviewer comment

Figure 4: The graphic is very blurry and a version should be provided with more pixels.

Response

We agree. The original figure has been replaced by newly generated figures for the revised simulation study. All revised figures have been exported at higher resolution and checked for legibility.

Technical comments

16. Approximate symbol

Reviewer comment

Line 286: The approximate symbol should be used: "... , whereas $\approx$ 50 droplets...".

Response

We agree. This sentence has been removed as part of the revision because we no longer present approximately 50 droplets as a general threshold. Where approximate numerical values remain elsewhere in the manuscript, the appropriate symbol or wording is now used.

17. Missing full stop

Reviewer comment

Line 298: Full stop missing.

Response

Corrected.

Detailed response to reviewer 3

We thank the reviewer for this careful and constructive review, and for recognising the value of treating droplet-freezing measurements as survival data. We agree with the main methodological concerns raised. In particular, the first manuscript relied too heavily on a small number of illustrative Gaussian simulations, did not test the empirical coverage of the Kaplan–Meier confidence intervals, and treated the ordinary log-rank test more generally than was justified.

We have substantially reworked the manuscript in response. The revised paper now uses fitted representations of three literature droplet-freezing curves spanning steep, intermediate and broad freezing behaviour. We use repeated simulations to assess the empirical coverage of the Kaplan–Meier intervals, the false-positive rates of the comparison tests, and their power and false-negative rates under shifted, strongly non-proportional and multi-population alternatives. We have also added several non-parametric alternatives to the log-rank test and introduced an Integrated Squared Difference (ISD) permutation test that operates directly on fraction-frozen curves and does not require proportional hazards.

We think these changes address the principal concerns raised by the reviewer and have made the manuscript much stronger.

Major comments

1. Proportional hazards

Reviewer comment

The log-rank test really only delivers full power when hazards are proportional, i.e., when the ratio between two samples is constant in . The manuscript notes this but only flags the case where curves literally cross. In practice the hazard ratio can vary smoothly with even when the curves don't cross, for example, when two chemically different nucleators share the same but have different . PH is really only plausible when the two samples differ in the $n_s(T)$ prefactor and share the same hazard shape, which is rarely strictly true. It would be worth saying so.

Response

We agree. The original manuscript focused too much on literal curve crossing as the main example of non-proportional hazards. As the reviewer points out, proportional hazards may fail whenever the ratio of conditional freezing probabilities changes with temperature, regardless of whether the plotted curves visibly cross.

We have added a substantially expanded discussion of this issue. The revised manuscript now explains that non-proportional behaviour may arise when samples contain multiple ice-active populations, when different populations dominate different parts of the freezing spectrum, when a treatment affects only part of a curve, or when a sample approaches the instrumental background. We also explain why this reduces the power of the ordinary log-rank test.

We have also added simulations that demonstrate this directly. The log-rank test performs well for some relatively simple shifted-curve alternatives in Figs. 5–6, but has much lower power for the steep-versus-broad comparison in Figs. 7–8 and the two-population mixture example in Figs. 9–10.

The revised manuscript therefore no longer recommends the ordinary log-rank test as a general-purpose comparison for droplet-freezing data. We retain it as a useful test when a broadly proportional-hazards-like difference is a reasonable expectation, particularly for relatively simple shifts in freezing behaviour. For more general curve-shape differences, we compare several empirical-distribution tests and introduce the ISD permutation test, which does not require proportional hazards. These changes are reflected throughout the revised statistical background, results, abstract and conclusions.

2. Coverage check

Reviewer comment

The recommendation to report KM bands would be a lot more compelling with an empirical coverage study. If the answer is consistently near 95% the recommendation is on solid ground; if it isn't in some regime, that is itself important to know.

Response

We agree, and this is now a major part of the revised paper. We also now consistently call these pointwise confidence intervals, rather than confidence bands, in response to a comment from another reviewer

We fitted smooth two-component Gaussian mixture distributions to three published droplet-freezing curves: a steep *Betula pendula* pollen washing water curve, a broader background curve, and a still broader BCS375 feldspar curve. The fitting is briefly discussed in the SI. We then generated repeated synthetic experiments from these fitted distributions. Because the underlying generating curve is known, we can calculate directly how often nominal 95 % Kaplan–Meier intervals contain the value they are intended to estimate.

The results are shown in Figs. 1–3. Coverage is close to the nominal level over most of the fraction-frozen curve for experimentally realistic droplet numbers, but it is not uniform. Coverage deteriorates near the ends of the curve, where the true fraction frozen approaches 0 or 1 and finite experiments contain little information about the exact tail behaviour. This issue becomes more obvious when the intervals are transformed to logarithmic cumulative activity spectra.

We think this provides a much firmer basis for recommending the intervals, while also making their limitations clear. The revised manuscript now states explicitly that they are pointwise confidence intervals, not simultaneous confidence bands, and that the tails of finite droplet-freezing experiments should be treated cautiously.

3. Type I error and power for the log-rank test

Reviewer comment

Figure 4(b) shows one example of a false positive and labels it “Type I error from sampling variability.” A single example doesn’t tell anyone what the real false-positive rate is.

Response

We agree. The original individual examples were not sufficient to establish either the false-positive rate or the power of the test, and we have removed that approach.

The revised manuscript now estimates false-positive rates using repeated null simulations. For each of the three fitted freezing curves and each droplet number, we generate 1000 pairs of experiments from the same underlying distribution and apply each test at . Figure 4 shows the resulting empirical false-positive rates.

Most tests are close to the nominal 5 % false-positive rate for moderate and large droplet numbers. The clearest deviations occur at very small droplet numbers: the ordinary log-rank test is somewhat liberal at , while the Kolmogorov–Smirnov test is strongly conservative. This calibration exercise is important because it means that differences between tests in the subsequent alternative scenarios can mainly be interpreted as differences in power, rather than simply reflecting one test having a much higher false-positive rate.

We then estimate power and false-negative rates using repeated simulations under realistic shifted and non-proportional situations. This replaces the original reliance on isolated examples of Type I

and Type II errors.

4. Only Gaussian simulations

Reviewer comment

Table 1 shows that all the simulated datasets are Gaussian. Real fraction-frozen curves are routinely not and are often long-tailed at the warm end and sometimes bimodal. The “~50 droplets” heuristic in the conclusions should either be tested against some non-Gaussian shapes or qualified as distribution-dependent.

Response

We agree. The original single-Gaussian simulations were too limited to support a general recommendation about droplet number.

The revised simulations use two-component Gaussian mixtures fitted to three literature droplet-freezing curves. The fitted curves span a useful range of realistic behaviour, including a steep freezing transition, a broader background-like curve, and a broad BCS375 curve with freezing spread over a much wider temperature interval. The dual-Gaussian functions are used as flexible empirical representations of realistic curve shapes, not as mechanistic models of ice nucleation.

We have also added a deliberately constructed multi-population scenario in which two curves contain the same two Gaussian freezing components but in different proportions. This produces a shape difference rather than a simple temperature shift and provides a controlled example of non-proportional behaviour.

The revised results show that detectability depends strongly on curve shape. A modest shift in a steep curve can be detected with relatively few droplets, whereas the same shift in a broad curve may remain difficult to detect even with substantially larger droplet populations. We have therefore removed the approximately 50-droplet heuristic and now state that there is no universal droplet-number threshold. Useful precision and statistical power depend on droplet number, curve shape, effect size, temperature region and the test being applied.

5. Quantitative comparison with Fahy et al.

Reviewer comment

Figure 1 shows that the KM bands look similar to the studentized bootstrap. That’s good as far as it goes, but if the authors are recommending KM in preference to the existing method then “looks similar” isn’t quite enough. A quantitative comparison, say band width at matched temperatures or coverage against a known ground truth, would be straightforward to add and would fall out of the coverage study above.

Response

We agree. The qualitative comparison with the Fahy et al. intervals did not provide a sufficient

basis for assessing the statistical performance of the Kaplan–Meier intervals.

Rather than comparing interval widths for a single experimental dataset, we replaced this part of the analysis with the empirical coverage study suggested by the reviewer. Other methods for calculating confidence intervals necessarily involve choices concerning binning, smoothing, resampling, or parameterisation, making a direct comparison dependent on the particular implementation. We therefore considered it more useful to assess whether the Kaplan–Meier intervals themselves perform reasonably. To do this, we generated repeated synthetic experiments from fitted underlying freezing-temperature distributions and calculated directly how often nominal 95 % Kaplan–Meier intervals contained the known underlying value.

We think coverage is the more relevant quantitative test because two methods may produce intervals of similar width while having different coverage properties. Conversely, a wider interval is not necessarily worse if it gives correctly calibrated coverage. The revised analysis therefore focuses on whether the intervals contain the value they are intended to estimate, how that behaviour varies with droplet number and curve shape, and where along the fraction-frozen curve coverage deteriorates.

We have also softened the comparison with previous approaches. The revised manuscript does not claim that Kaplan–Meier intervals are universally preferable to the studentized bootstrap. Instead, we present them as a simple analytical approach that requires no resampling, binning or assumed parametric freezing distribution, and whose empirical pointwise coverage is shown to be sensible over most of the experimentally relevant curve.

6. The patch

Reviewer comment

Carrying the previous interval forward at the last point is pragmatic but it’s an arbitrary choice and warrants either some justification or at least a sensitivity check. Does the choice change any of the conclusions on the ash or birch-pollen data? If not, please say so.

Response

Response: We agree that carrying the penultimate interval forward is a pragmatic plotting convention rather than a formally derived confidence interval. We have now made this explicit in the manuscript by adding the following text:

“For plotting purposes, we assign the final point the same interval as the penultimate point. This is a pragmatic plotting convention rather than a formally derived confidence interval at $f(T) = 1$. It affects only the final plotted point and does not alter the curve-comparison tests or any of the conclusions, because these tests are calculated from the observed freezing temperatures and empirical fraction-frozen curves rather than from the confidence intervals. The final point should therefore not be interpreted as having a formally calibrated confidence interval

The choice therefore does not affect the simulation results or the conclusions of the revised manuscript. It also does not provide additional information for derived cumulative spectra, because quantities based on $-\ln[1 - f(T)]$ are themselves undefined at $f(T) = 1$.

We have removed the volcanic-ash and *Betula pendula* experimental examples from this part of the paper. On reflection, we realised that these individual datasets were not particularly helpful for explaining or validating the behaviour of the method. We have replaced them with the systematic empirical coverage assessment using synthetic experiments generated from fitted literature freezing curves.

Closing comment

Reviewer comment

I think this is a useful methodological contribution and well within scope for the journal. The framing is sound and I expect it to get picked up quickly once it's out. The calibration and assumption-testing points above would, in my view, change a good paper into a reference one. I look forward to the revision.

Response

We thank the reviewer for the encouraging assessment and for the specific suggestions. The comments on empirical coverage, false-positive rates, statistical power, realistic curve shapes and proportional hazards were particularly helpful. They led us to replace the limited illustrative simulations with a much broader calibration and power study, and to rethink the role of the ordinary log-rank test in the paper.

The revised manuscript is no longer primarily a recommendation to use Kaplan–Meier intervals and the log-rank test together. It instead presents a broader non-parametric framework for quantifying uncertainty and comparing droplet-freezing curves, including explicit discussion of the assumptions and limitations of each approach. We think the revision is substantially stronger as a result.