

Response to Referee 1

William D. Fahy et al.

Author comment on "A universally applicable method of calculating confidence bands for ice nucleation spectra derived from droplet freezing experiments" by William D. Fahy et al., Atmos. Meas. Tech. Discuss., <https://doi.org/10.5194/amt-2022-141-AC1>, 2022

*The original referee comments are pasted below with our response and how we revised the manuscript following in **bold text**.*

Response to Referee #1

The manuscript titled "A universally applicable method of calculating confidence bands for ice nucleation spectra derived from droplet freezing experiments" by Fahy et al., demonstrates a method to derive ice nucleation (IN) spectrum and confidence bands for ice nucleation studies. They use this to determine an IN spectrum normalized to particles surface area, a quantity referred to as $ns(T)$, the number of ice nucleation events normalized to particle surface area as a function of temperature. Bootstrapping is used in this method to simulate freezing temperatures multiple times by random sampling and calculating $ns(T)$ from these simulated data. Variability in simulated freezing and $ns(T)$ is used to derive confidence bands. Finally, the authors are then able to deduce statistical similarity between datasets. Overall, this paper stresses in the importance of deriving uncertainties in $ns(T)$ and should be of great benefit to ice nucleation community.

Unfortunately, there are many instances when clarity and precision are lacking in describing the methods, as well as when the authors mix what can be stated as options or as claims with evidence. I cannot recommend publication until the comments below are addressed. The first major comment relates to poor writing, due to poorly explaining (or not explaining at all) complex topics/methods that are the main subjects of this manuscript. Another major comment is about assuming variability on the droplet by droplet basis in Eq 1 and 2. A final major comment is about the method for assessing the similarity of IN spectra is not explained at all. Although, the paper is quite novel, addresses relevant scientific questions, and has important conclusions. They demonstrate sufficient evidence through their simulations to support their conclusion.

We thank the referee for their detailed review of our manuscript. Each comment made is responded to below, along with any revisions as appropriate.

- The article and especially the introduction has many vague terms, descriptions and opinions. I have outlined most in the minor comments below, however, a few major comments are described. Most notably the introduction is poorly written, with insufficient details about what the author will present and how it relates to previous work. The authors have not reviewed previous literature and what advances were made

in ice nucleation simulation to lead them to this work. They are not the first to model or simulate ice nucleation. Instead, their intro is a narrative about the failure in the community. This is unacceptable and must be rewritten.

Our original manuscript was formatted such that the review of previous literature and previous advances is contained mostly in later sections, where each specific technique is introduced and explained. To improve the introduction, some language summarizing previous work has been added, along with a sentence directing the reader to the later sections where this work is discussed in more detail.

After p.2 l.46: "In our experience, there is no widely implemented approach to reporting uncertainty in IN temperature spectra derived from freezing experiments. Instead, methods vary between groups, often relying on different assumptions about the nature of ice nucleation measurements, the forms of distributions that the random variables involved take, and the quantification of the derived uncertainties. In the simplest case, standard deviations, errors, and/or confidence intervals have been calculated from repeated experiments either by assuming that variability follows a normal distribution (Polen et al., 2018; Jahn et al., 2019; Worthy et al., 2021; Chong et al., 2021; Roy et al., 2021; Losey et al., 2018), a Poisson distribution, Koop et al. (1997) (Yun et al., 2021; Alpert and Knopf, 2016; Knopf et al., 2020; Kaufmann et al., 2017), or that droplet freezing follows a binomial distribution (Gong et al., 2020, 2019; McCluskey et al., 2018; Suski et al., 2018; Wex et al., 2019). In other cases, authors have used a model of ice nucleation to simulate their experiments and use that simulated distribution to estimate the uncertainty present in their experiment. In the simplest case, droplet freezing is modelled as a Poisson point process (Vali, 2019; Fahy et al., 2022b; Jahl et al., 2021). In more sophisticated models, variables such as the number of sites, mass of material, and temperature variations are parameterized to run completely new simulated experiments (Wright and Petters, 2013; Harrison et al., 2016). Even in these sophisticated models, either additional measurements are required, or assumptions must be made about the distribution of each variable. Until the inherent variability behind ice nucleation can be measured to prove or disprove the assumptions being made, all the above methods are only as reliable as the assumptions themselves. In Section 4, each method, their required assumptions, and the validity of those assumptions are discussed in detail.

Empirical bootstrapping is an alternative approach to estimating statistics for a dataset that to our knowledge has not been applied in the context of ice nucleation. In this technique, a series of random samples of the measured dataset is taken to generate estimated statistics that converge on the actual values as the number of samples increases (Efron, 1979; Shalizi, forthcoming). No assumptions are required about the distributions of random variables underlying ice nucleation and it can be applied to any system where the freezing temperatures or times of droplets are measured. Here we present a set of generalized and statistically rigorous methods based on empirical bootstrapping for quantifying uncertainty in IN spectra. When accompanied by interpolation methods presented in Section 3, this approach can be used to calculate continuous confidence bands and statistically test differences between IN spectra (Section 5). We also address the effects of interpolation techniques, droplet sample size, and bootstrap sample size to direct the field towards more rigorous and repeatable methods of experimentation and data analysis. An implementation of all presented statistical methods along with documentation and instructions for its use is provided freely for use or reference to assist in future research and improve the statistical treatment of ice nucleation data in

the field.”

- “ice nucleation processes” The authors must be more specific about what this process is. As a reader, the phrase “it is simply a fact of the ice nucleation process” is meaningless because these facts are not outlined or given in the manuscript up to this point. I read only that ice nucleation is a process, and it does not help understanding. Please be specific, what is the process.

Changed this sentence to:

“This variability is inherent to ice nucleation. Using the singular-stochastic model most recently discussed in Vali (2014) and terminology proposed in Vali and DeMott et al., (2015), ice nucleation activity (or rate) is an accumulation of many ice nucleation sites with variable critical active temperatures dispersed randomly throughout a material. In turn, the material is distributed randomly throughout droplets which can have varying sizes, shapes, and environments. Therefore, a measured IN activity can be affected by heterogeneity in the distribution of ice active sites across a material, heterogeneity in the mass or surface area of material suspended in each droplet, differences between droplet sizes and environments, and variations in temperature between droplets.”

- “IN spectra”, “ $n_s(T)$ ”, “ k ”, “ K ”, These are defined in many places throughout the first pages of the manuscript. Would the authors please consolidate a common and nonredundant definitions of these. i. p. 3, l. 75: k and K are ice active site density spectra ii. p. 3, l. 76: k is a differential spectra and K is a cumulative spectra iii. p. 4, l. 81: K is the number of ice active sites at temperature T per unit of suspension volume. K can also be nm . K can also be ns . iv. p. 4, l. 6-7: K are the cumulative ice nucleation active site density curves

Two of these definitions have been removed (“ k and K ”, p. 3, l. 75 and “For simplicity and generality... cumulative or differential IN spectrum”, p. 4 l. 85-86). The remaining instances either define the terms themselves or the connection between k/K and other terms.

- p. 2, l. 44-46, This sentence insults the ice nucleation community. Yes, ice nucleation analysis is complex, however, the statistical tests, approximations, and methods of previous work are not inaccurate. They are peer-reviewed and explained. Please rewrite this sentence, emphasize previous work that has been done on which your manuscript is based, and how the authors fill a missing gap.

It was never our intention to insult the ice nucleation community, and we apologize that it was written to imply that. Indeed, the previous statistical methods used are peer-reviewed and explained. However, as we discuss in detail later in the manuscript, their accuracy varies depending on the validity of the assumptions and approximations made to derive the statistics used. This is not a failing of the ice nucleation community – every statistic has an error associated with it, and for reasons discussed elsewhere in the manuscript, ice nucleation presents a difficult problem for many standard statistical methods. We merely wished to point out that the approaches being used for some studies (including our own previous works) may not have been as accurate as many people reading them may have assumed.

To remove any offensive language, this sentence has been changed to:

“In our experience, there is no widely implemented approach to reporting uncertainty in IN temperature spectra derived from freezing experiments.

Instead, methods vary between groups, often relying on very different assumptions about the nature of ice nucleation, the form of distribution which the random variables involved take, and the visualization of the derived uncertainties.”

- p. 4, l. 81: “...number of ice active sites...” This is a vague term. What is a site and how does Eq 1 and 2 give sites? In fact, eq1 and eq2 are using counted freezing events and there are no ice active sites in them. There are many assumptions made between Eq 1 and 2, and then making a claim about quantitative measures of ice active sites.

Much of this material is thoroughly discussed in literature that we cite in this paper. We have added some additional discussion of what ice nucleation sites are and the model we are using to conceptualize ice nucleation earlier in the text to point the reader in the direction of that literature.

- The word “complex” is overused. On p. 5, l. 23, what make a curve complex and what makes a curve not complex? On p. 2, l. 55, what makes a dataset complex and not complex? On p. 2, l. 45, what makes ice nucleation complex and what makes it not complex? On p. 11, l. 80, math is complex. It appears that when the authors do not want to take the time to explain ice nucleation, IN spectra/curves on graphs, and bootstrapping, they call it “complex” and move on. This shows lack of care and work put in to write a comprehensive manuscript, as these complex things are the main subjects of the manuscript. Please search for the word complex in the manuscript and try to replace it with specific details and explanation. These should be briefly explained.

Most examples of ‘complex’ have been removed, we agree it was overused. However, in some cases the word complex is used rightfully – in the case of what makes ice nucleation complex, there is an extensive body of previous literature explaining the complexity. We have incorporated a brief discussion the model that we are referencing, but most of that material is the subject of a different manuscript. Similarly, for derivations of IN spectra, k and K are widely used and accepted in the community, and have already been explained and discussed in detail elsewhere – redoing those calculations and explanations is outside of the scope of this manuscript.

We have added some citations about the mathematics behind bootstrapping, but once again, this manuscript is not intended to prove the efficacy of bootstrapping in general – it is an application of a well-defined and previously studied tool in statistics. The new parenthetical statement is:

“(see Efron and Tibshirani, (1994) or Davison and Hinkley, (1997) for a thorough treatment of the mathematics behind bootstrapping and Canty et al., (2006) for a thorough discussion of inconsistencies and errors that can be encountered when using bootstrapping),”

- The words/phrases “superior”, “best”, “most powerful” are used through the manuscript. In all context, these are opinions of the authors. I suggest removing all.

These words have been removed, except for in the following case:

- **14, l. 344, ‘best-performing’ remains because we define a metric for what we are interpreting that to mean, and studentized bands are known to be the most accurate confidence bands that we used.**
- Equation 1 and 2 has an important and unclaimed assumption. Why are the authors so sure that the normalization constant, X , is identical for each droplet? On l. 35-37, it states many differences that could cause variability from droplet to droplet. If so, these

equations are not valid. If X were to be different for each droplet, then it would need to be accounted for. However, this equation inherently assumes some identical normalization. In other words, one constant for all liquid droplets. In other derivations besides Vali, such as in the Pruppacher and Klett textbook, eqs 7-64 and 9-56, the natural log appears due to the assumption that all drop volumes are the same, or surface area in drops is the same. The authors must claim this assumption in the manuscript, and that their analysis only holds if this assumption is valid.

We have added the following language to explicitly claim this assumption in deriving IN active site density spectra. You are correct that the calculation of these spectra only holds if this assumption is valid, as we now make explicit:

“The derivation of these equations requires that X be identical for every droplet being analyzed – an important assumption and source of error.”

We have also made explicit the connection between this normalization constant and our statistical methods – specifically, if empirical bootstrapping is used, then variations in the normalization constant X are automatically incorporated into the error analysis:

“However, as will be discussed later, the empirical bootstrapping approach quantifies this source of error, meaning these parameters can be used and interpreted even when the assumption does not strictly apply as long as the uncertainty is also incorporated into the interpretation.”

And later on in Section 4:

“Since variations in droplet size, sample mass suspended, or distributions of surface area among droplets (the parameters behind the normalization constant X) also contribute to the variability observed in experiments, the error caused by assuming X is constant between droplets is also included into the model.”

- p. 15, l. 92 - p. 18, l. 97: This method descriptions is far too short. The authors claim taking a ratio or subtracting two different $n_s(T)$ curves is supposed to be between 0 and 1? That is the test of similarity. When taking a ratio or difference, the authors are doing this on a \log_{10} scale? When they take the ratio, does the larger spectra need to be in the denominator? When taking the difference, the smaller values need to be subtracted from the larger values to get only positive numbers? The step-by-step procedure here is not so clear and so it is difficult to review Fig. 5 and any discussion around it. It would help to guide the reader a bit more slowly here. Finally, can the authors give a name or equations for "these metrics" stated p. 15 l. 91?

We have added some additional detail. To clarify, all we are doing is subtracting or dividing two spectra and calculating resulting confidence bands – there is not much more information to provide beyond how confidence intervals are calculated, which is also explicitly stated. To answer your questions, we are not working on a \log_{10} scale except to plot the data clearly. No, the larger spectrum does not need to be in the denominator, nor does the smaller value have to be subtracted from the larger value – note that in Figure 5, neither of these conditions are true. We do not claim that the ratio or (mathematical) difference between two n_s spectra is ‘supposed’ to be between 0 and 1 – we are discussing the confidence bands of each metric and making conclusions about whether two IN spectra are different based on whether they contain 0 or 1.

We have clarified the language. We use the term ‘difference spectrum’ to represent both metrics we are discussing to make this section less verbose. The

paragraph now reads:

“For a more quantitative (and interpretable) method to compare IN spectra can simply be divided or subtracted. We will use the term ‘difference spectrum’ to refer to a continuous plot of either this ratio or difference, as both are calculated using the same procedures and provide similar information. When interpolated IN activity spectra are used, a continuous difference spectrum can easily be generated by calculating the ratio (or difference) between two interpolations at each point in a dense grid of temperatures, then interpolating between those points. A difference spectrum can be plotted as a function of temperature with its own confidence bands and can be used to test whether two IN spectra are statistically significantly different at any temperature where the two spectra overlap at any confidence level. Stated precisely, in the case of a ratio-based difference plot with confidence bands, the hypothesis that the two IN spectra are different can be tested. If the confidence bands do not contain one at a given temperature, then the null hypothesis (that the two IN activities compared are insignificantly different at that temperature) is rejected. If they do contain one, then that claim cannot be made. If a difference between IN spectra is used instead of a ratio, then zero is used for this hypothesis test instead of one. Therefore, if confidence bands can be accurately calculated for a difference spectrum, then continuous statistically rigorous claims about differences between IN spectra can be tested.”

Some additional language in later paragraphs has also been changed to match the terminology used here.

- Abstract, p. 1, l. 15: “...and if used properly...” Would you please rephrase this? It reads as if there is a way to also improperly use your methods.

This phrase has been changed to:

“...and when large sample sizes ($\sim > 150$ droplets and ≥ 1000 bootstrap samples in our system) can capture...”

- p. 1, l. 19: “By improving the statistical tools available, this work will improve...” Can you rephrase this? It reads like the authors still need to improve their methods. I do not think that is their intension.

This has been rephased to:

“By providing additional statistical tools to the community, this work will improve...”

- p. 1, l. 19: “...improve the quality and accuracy...” Accuracy is improved with instrumentation that is more accuracy. Statistical analysis does not make instruments more or less accurate. Would you please state exactly what is more accurate here?

This sentence now reads:

“improve the quality and accuracy of statistical tests and uncertainties in...”

- p. 1, l. 39: The authors write the word "this", but do they refer to all of the uncertainties and variabilities in experimental investigation mentioned above? Or are they still talking about a perfect experimental setup?

Changed to "This variability"

- Intro, p. 2, l. 33: Would the authors please precisely define "IN activity"? There should be a distinction between what is measured and what is derived.

Inserted the parenthetical statement:

"(here we use the term 'IN activity' as a general term to refer to any measured or derived variable which quantifies ice nucleation rate with respect to temperature.)"

In this case we do not make a distinction between measured and derived values because they are different quantifiers of the same underlying process, and as such have the same sources of error. Later in the manuscript we specify which variables are measured and which are derived in the framework we are using.

- p. 2, l. 42: "If we cannot eliminate experimental error, it must..." The word "we" is usually referred to the authors. I suggest to replace with "Experimental error is always present and must be ..."

This change has been made.

- p. 2, l. 43-44: It is the authors opinion that there is no widely accepted approach. Please remove this sentence, or write this is an opinion.

This has been changed to "In our experience there is no..."

- p. 2, l. 50: "...remaining questions..." What are the questions the author is referring? What are those questions that remain? Please be precise in formulating your scientific questions.
- p. 2, l. 50: "...experimental parameters..." What are the specific experimental parameters the authors are referring to. Please be specific.

To replace both of these statements this line has been changed to:

"We also address the effects of interpolation and sample size on measured IN activity spectra."

- p. 2, l. 51: "...these approaches ..." What are the approaches the authors are referring to. Please be specific.

Changed to "all presented statistical methods" and added "based on empirical bootstrapping" to a previous sentence for more specificity.

- p. 2, l. 55: What is the difference between a toy IN dataset and an IN dataset. Maybe the authors would like to change this to read, "example of an IN dataset"?

This now reads "we selected an example IN dataset".

- p. 2, l. 61: please state the diameter

Added in text, "(1.5 mm)."

- p. 3, l. 68-73: These sentences are redundant. Please rewrite.

The second to last sentence has been removed ("The validity of... suspensions.").

- p. 4, l. 99: Should "...temperature density of freezing events..." be "...number of freezing events depending on temperature..."?

Phrasing has been changed to "number density of freezing events with respect to temperature".

- p. 5, l. 49-40: This is a strange way to end. This is a strange way to end. Please tell the reader exactly what you are looking for, instead of telling them you are looking "elsewhere".

Changed to "we must look for an interpolation method that can capture an ice nucleation spectrum with any shape"

- p. 6, l. 44 and l. 49: What is the difference between the terms "accurate interpolation", "faithful interpolation", and simply "interpolation". Can an interpolation be "unfaithful" or "inaccurate"? I can find no quantitative meaning of these in the paper. I suggest to remove "accurate" and "faithful". Please search through the manuscript for these.

When referring to interpolations, these terms have been removed in most cases. However, the term 'accurate' is used in many places in the manuscript to indicate that a particular approach either qualitatively matches the shape and value of the 'actual' values (either measured or theoretical). When referring to confidence limits, for example, there does exist an actual exact confidence limit at every point, even though we cannot measure it without knowledge of the entire population. Thus, it is still valid (and informative) to use the term 'accurate' in relative terms between methods.

- p. 8, l. 83: "...are unreliable." should be changed to "...may be unreliable."

Changed the wording to "distribution, making this assumption unreliable."

- p. 8, l. 88-89: I tried looking up the reference Kaufmann et al. 2017, but it is not in the references list at the end of the manuscript. Would you please check that all references are actually included?

We have double-checked the reference list to make sure nothing is missing. The requested citation is as follows:

Kaufmann, L., Marcolli, C., Luo, B., & Peter, T. (2017). Refreeze experiments with water droplets containing different types of ice nuclei interpreted by classical nucleation theory. *Atmospheric Chemistry and Physics*, 17(5), 3525–3552.
<https://doi.org/10.5194/acp-17-3525-2017>

- p. 10, l. 45: "...originally proposed..." What exactly was originally proposed? There are many things proposed up until this point reading the manuscript.

Removed this parenthetical statement – empirical bootstrapping was proposed before any type of parametric bootstrapping, but it is not important for the manuscript, so we removed it to avoid confusion.

- p. 10, l. 47: ..."with replacement..." Would the authors please be specific about what this means? I thought it was a typo at first, but appears like it means something important to them.

Added the parenthetical statement "i.e. the same datapoint can be sampled more than once"

- p. 10, l. 61: "bye" is a typo.

Thank you – changed to 'by'

- p. 10, l. 62-65: This sentence is unnecessary, negative and offensive. Please remove it. It is affronting members of the ice nucleation community as unfamiliar with statistics and as unable to access information. This is not true.

We have removed this statement – we intended no offence, we simply wished to supply some rationalization for why this very simple technique that is well-suited for the problem of ice nucleation (empirical bootstrapping) has not been used previously.

- Through the manuscript there is use of single quoted words and phrases. It is not clear to the reader why these have to have single quotes. Would the authors please elucidate the reason for this, or just remove the quotes and explain things clearly.

Single quotes are used to denote a term that we are defining. Much as you would quote someone else's term with double quotes, we are using single quotes to separate the specific term from the surrounding text as an important definition. For example, when comparing 'observed' vs 'sampled' lists, these words have single quotes to denote that we are defining these terms in this context. There are a few places where the single quotes were redundant, and those have been removed.

- p. 11, l. 78: What is a "re-interpolation", and how can this be accurate or inaccurate?

Changed to "interpolated exactly using a simple spline fit" for clarity. We used re-interpolation to indicate that the curve had previously been interpolated, then discretized, and that we are now interpolating it again.

- p. 12, l. 4: Please write out what alpha is.

We have added "respectively, where alpha is the threshold value chosen for statistical significance"

- p. 15, l. 82: Please add commas so that the words "...method such as empirical bootstrapping rooted..." is changed to "...method, such as empirical bootstrapping, rooted...".

Changed.

- p. 17, l. 22-24: This is plenty of evidence that droplets on a substrate surrounded by oil or air can be used to measure homogeneous freezing. Yes, other studies have a background of heterogeneous ice nucleation occurring. Generalizing this to all substrate based approaches is not correct. Please remove this sentence.

Changed the first word to "Most" instead of "All" and removed "measured as the ice active site density normalized to the volume of water in each droplet" for concision. However, background subtraction is important for many systems - even droplet-on-substrate methods that can measure homogeneous freezing usually have some heterogeneous freezing signal earlier in the spectrum and many systems that are intended to measure rare heterogeneous ice nucleation

events have higher background freezing activities that should be accounted for. As such, we have elected to keep the sentence, but have removed the generalization.