Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., author comment AC2
https://doi.org/10.5194/amt-2022-141-AC2, 2022

Response to Referee 2

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-AC2, 2022

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

Response to Referee #2

This manuscript evaluates the potential of two types of bootstrapping methods for quantifying confidence intervals of droplet freezing experiments. Therefore, this manuscript may contribute to a more consistent interpretation of results from droplet freezing studies. The authors have also provided a very thorough documentation of their coding efforts through a publicly available Github repository.

We thank the referee for their feedback on our manuscript. Each comment made is responded to below, along with any revisions as necessary.

However, I have major concerns regarding the discussion of limitations and requirements associated with the bootstrapping methods evaluated in this paper. These aspects (e.g., sample size requirements) need to be discussed in much more detail, e.g., as in figure S3 which should be moved to the main text. A more detailed discussion would also ensure that these methods are used „properly" as mentioned in the abstract. I would also recommend a stronger focus on the bootstrapping methods, instead of an in-depth discussion of binning and interpolation methods, to make the manuscript more concise and less verbose

In terms of limitations, besides the assumption of statistical independence stated in the manuscript there is only the issue of sample size/bootstrap sample size. For sample size requirements, we have brought what was Figure S3 into the main text and have also edited it – in the original figure, all curves included the freeze that was determined to be an outlier in Section 5.2, which caused the confidence bands on the N=98 and N=191 to be wider than they would normally be. Instead, we now randomly sample 50, 100, 150, and 200 droplets (along with the original 286) to better understand the effects of sample size on the width of confidence bands and the shape of the interpolations used.

Along with adding some additional text in Section 4.6, as you are right that it is important for any future implementation or interpretation of these methods. The last paragraph of Section 4.6 has been expanded significantly:

Although we cannot theoretically determine the sample sizes required for accurate confidence bands using empirical bootstrapping due to the same limitations discussed previously, the sample sizes required for accurate confidence bands can be empirically evaluated by testing how many assays, droplets, and simulated spectra are required for confidence bands to converge (therefore reducing the uncertainty of the confidence bands due to sample size). Fig. S5a displays interpolations and resulting confidence bands for the differential IN spectrum of aged volcanic ash when 50, 100, 150, 200, and 286 (where all droplets are included) droplets are randomly sampled from the six performed experiments. The width and shape of the confidence bands changes significantly but seem to be converging to a smooth curve exemplified when N=286. When N=50, the confidence bands span three or more orders of magnitude, indicating that 50 droplets may be too few to draw any conclusions from freezing spectra in our system. In this case, the minimum sample size is approximately 150 for useful conclusions at the 95% confidence level, and at least 200 is preferred (at least in our system) for more accurate confidence bands to ensure the entire probability space of droplet freezing is covered. More droplets or freezing assays improve the accuracy and reduce the width of the confidence bands, especially in differential IN spectra, further motivating the many recently developed microfluidic techniques (Brubaker et al., 2020; Reicher et al., 2018; Stan et al., 2009; Tarn et al., 2018; Weng et al., 2016; Roy et al., 2021). Additionally, the number of simulations (and resimulations if using studentized confidence bands) should be chosen carefully to ensure the full variability present in IN spectra is represented.

In Fig. 5b, the tskew confidence bands of the combined water aged volcanic ash IN spectra (all 286 droplets) are compared when the number of simulations (*nSim*) ranges from 50 to 10000. S4 shows the same bootstrap simulation number analysis when using 50 and 150 droplets randomly sampled from the initial 286 to test the effects of droplet number on the required bootstrap simulation sample size. Based on these plots, the number of bootstrapped spectra does not appear to impact the confidence intervals nearly as much as the number of droplets used. This is a contrast from requirements in other types of bootstrapping techniques, which can be easily calculated to require more than 10000 samples to be accurate within an acceptable margin of error (Hesterberg, 2015). While this insensitivity could be coincidence (errors in opposite directions cancelling out to result in confidence bands that are approximately correct), we speculate it is instead related to the fact that each bootstrap sample includes many droplets (286 in this case) which are also sampled, covering the probability space more completely than when single datapoints are sampled. Based on this, we recommend ensuring that the number of simulations multiplied by the number of droplets (the 'resample size') in each simulation exceeds 10000 – in the above spectrum this resample size ranges from around 14000 to 2.86 million. This is partially corroborated by Fig. S5, where the resample size reaches as low as 2500 and the confidence bands with low resample sizes often do not match those with high resample sizes. It is also important to consider the statistic being calculated – in the above case the standard deviation and skewness are being used to calculate confidence bands, but if quantiles were being used, the number of bootstrapped spectra would have to be large enough to calculate accurate 2.5^{th} and 97.5^{th} quantiles (likely about 200 spectra). Regardless, the effects of bootstrap sample size should be tested whenever empirical bootstrapping is applied to ensure that the accuracy

of the calculated confidence bands (or any other statistic) is never dependent on the number of simulations used. Similarly, each investigator must determine their own droplet sample size requirement to collect datasets that can answer their research questions.

Finally, Figs. 3-5 provide evidence that the interpolation technique used is not overfitting the data, as the quantiles and other confidence bands follow the general shape of the experimental spectra. Since these statistics are calculated from an aggregate of 1000 samples in most cases, they would be expected to smooth out random variation present in a single measured spectrum that could be causing the complex interpolated curve observed. Because the aggregated data maintains the same shape, it can be assumed that it is at least somewhat meaningful, and that the interpolation technique is using an appropriate smoothing factor, however, this should be tested regularly to minimize potential overfitting. Note that when droplet numbers are below 200 (as in some of the Fig, S4a spectra and in Fig. S5) the interpolated differential spectra have shapes that look unrealistic (e.g. many critical points within one or two degrees Celsius), but they also have error bars that span many orders of magnitude in those regions, meaning that neither the measured value nor the interpolation of the differential IN spectrum at that point is as useful because the uncertainty is so high. Regardless, the cumulative spectrum remains smooth and interpretable, and the portions of the differential spectrum with lower uncertainties are still meaningful."

We have also added a mention of these limitations/recommendations in both the last paragraph of the introduction and into the abstract to emphasize them.

However, we cannot generalize these requirements to every application of these techniques – the number of droplets and simulations required is dependent on the experimental setup, sample in question, and the questions the investigator is trying to answer. Each investigator will have to make their own choices about sample and resampling size depending on their findings. Figure S3 (now 5) is just an example of how this should be done and should not be taken as a general statement for the sample size requirements for empirical bootstrapping.

Regarding the in-depth discussion of interpolation/binning, interpolation is required for calculating continuous confidence bands and for comparing IN spectra quantitatively as discussed in Section 5, so we think it is integral to this manuscript. We are proposing to move away from the commonly-used binning techniques, so we offer a detailed explanation why and show the benefits of our preferred approach. However, we have moved Figure 2c and 2d to the SI, as most of the information in those panels is contained in panel a and b as well.

p.2, l. 58: HPLC is not defined - please make sure that abbreviations are defined consistently across the manuscript.

In this case, we are simply referring to a grade of chemical – we have changed it to the more conventional parenthetical "(HPLC grade, Sigma)".

p.3, Fig. 1: The naming convention is slightly confusing - maybe remove duplicate mentioning of „unnamed FUE" and „water aged FUE" in the legend? Also, the color scheme is not color-blind friendly and differences between samples are hard to see. How many runs were included per sample?

We have changed the legend to specify that the first two spectra are 'combined' spectra, while the rest are individual runs. We have also changed the color scheme and increased contrast (using the discrete rainbow scheme from Figure 19 of the following blog post: https://personal.sron.nl/~pault/), but our main point is not differentiating between individual runs, but rather showing that the combined spectra represent the individual runs relatively well. Unfortunately, it still is difficult to distinguish between all the spectra because by their very nature they overlap significantly – that's the point.

p.3, l.75ff: I don't see the advantage of representing the droplet freezing data as ns values within the context of this study. However, if applied, the concept of ice-active site densities needs to be introduced explicitly - it is unclear how ns is derived and which specific surface area values are used.

We chose to use ns values because that tends to be the most common normalization scheme in the ice nucleation community today. We have added an explicit equation describing the variables used to calculate ns and have added our source for the specific surface area we used. The concept of ice-active site densities is also introduced in the introduction in more detail.

We do not go into more detail on the subject in this paper because it is not the focus of this study, and many previous authors have discussed the concept of ice active site densities in detail.

p.4, l.88f: What does „many" droplets mean exactly? Throughout the manuscript, a stronger emphasis should be on a discussion of the impacts that sampling size has (i.e., in the original observed data, the re-sampled data etc.)

Changed to "hundreds of… in our estimation". In this case, we are basing this statement on previous work, but in other places we have emphasized the effects of sample size as mentioned in response to your major comment.

p.5, l. 36: I would debate whether the contact angle approach with its many underlying assumptions (e.g., using bulk properties for describing microscopic nucleation processes) is strictly speaking „physically-based" - maybe rephrase?

You are correct – we have simply removed "physically-based" – it is not necessary for the point we are making.

p.6, l. 63: „high/warm" instead of „low" temperature?

Changed, thank you for catching that.

p.7, Fig.2d: Why does the interpolated data (unaged FUE) stop at -12 degC?

Excellent question – in the version of the code used for this manuscript, a cutoff for the start of freezing events was used because without it, the interpolated spectra were unrealistic around the first few freezing events. We've reduced this cutoff so that more of the initial freezing events are captured, but we've also added the following line to explain this:

"Note that the interpolated spectra do not start until there is a sufficient density of freezing events (more than one per degree Celsius) to avoid overfitting and because the error on these initial points is much larger than that of the rest of the spectrum as will be seen later."

p.8, l.75ff: Many of these approximations are only valid for „larger" droplet ensembles - this limitation should be mentioned to emphasize in which situations (i.e., for which experimental setups) we need more flexible statistical approaches. Also, as many readers might be more familiar with t-intervals than Z-intervals, a short explanation of the differences would be great.

For the Z- vs t-interval, the line has been changed to: "a Z-interval (based on the normal distribution) or t-interval (based on Student's t-distribution) can be constructed."

As for the main point of your comment here – the number of droplets required for an accurate closed-form confidence limit to be calculated is not straightforward.

For example, according to the NIST engineering statistics handbook section 7.2.2.2[1], for a two-sided t-test (a procedure essentially similar to constructing confidence intervals for a single process e.g. 7.2.2.1 of the NIST handbook), the sample size required is:

Where N is the sample size, α is the confidence level for the test, β is the probability that we will fail to detect a shift of δ from the mean, and s is the standard deviation of the sample. We can set δ and β depending on how accurate we want our test to be. Say we are willing to accept a 1% chance of missing (failing to reject the null hypothesis that the mean has shifted) a shift of 1 standard deviation for a test with 95% confidence. Then, the required sample size is:

If we assume that the number of degrees of freedom for our sample is equal to the number of droplets, in our system approximately 50 (a debatable assumption that will be discussed momentarily), then we get:

The question becomes, what does a sample size of 22 mean? If we are testing (or constructing a confidence interval) for the temperature at which the average droplet freezes, then this means 22 droplets are required from a single sample. However, what if we want to test the difference in frozen fraction or IN active site density spectrum at a specific temperature or the difference in temperature at which a given frozen fraction of IN active site density occurs (a much more common and useful set of questions)? Then, the most common approach when using simple t- or Z-intervals is to re-run the experiment several times, interpolate in some way, and then use the mean and standard deviation at a specific temperature or IN activity measurement to calculate the confidence intervals or perform the test (e.g. Polen et al., 2018[2]). In this case, is the sample size the number of replicate runs (e.g. the three separate freezing assays performed on the same sample), or the total number of droplets in all three of those replicate assays? Strictly speaking, since we are only getting a single parameter out of each of the three runs, we must assume the former, and therefore for the Central Limit Theorem to drive the distribution of these parameters to a normal or t-distribution where the t- or Z-interval is valid, we need many more replicate runs of the same sample – potentially up to 22 if we assume the above analysis is correct – not a particularly realistic number for

more sophisticated scientific questions and limited samples or experimental time.

Additionally, the calculation above requires an estimation of the number of degrees of freedom for the t-statistic. Is that the number of droplets, because each droplet can freeze at a different time/temperature, or is that the number of replicates since those are the actual parameters we are comparing? Again, it is unclear.

Of course, the ensemble of droplets is much larger than the sample size of 22 and increasing the number of droplets frozen is usually observed to increase the accuracy of an experiment (and therefore statistical tests based on that experiment). In that case, how would those droplets contribute to the sample size requirement? To our knowledge, there is no quantitative estimate for this effect. So, when asking how many droplets and/or freezing assays are required for a given confidence interval to have a certain margin of error, we have essentially no way of giving a quantitative answer.

This problem has often been ignored in the past (e.g., there is no real mention of how the theory behind confidence intervals works with the sampling methods used in freezing experiments in previous papers we cited in this section). Since this manuscript is offering a new way of doing these statistics that does not have the same problem (although as you rightfully requested, sample size considerations are still important for our technique, albeit quite different), we do not feel that it is within the scope of our work to give quantitative estimates of how inaccurate these techniques are, as it would require building a significant amount of statistical theory for a technique that already has many other issues for this application.

Instead, we have added the following language after the sentence on Z- and t-intervals:

"While it is unclear how many droplets and freezing assays are required for these approximations to be valid under the Central Limit Theorem, in our experience it is unlikely that most existing freezing assay datasets achieve this sample size requirement, since confidence intervals calculated using these techniques often disagree with those calculated using other methods described below and those presented in this study. It is also unclear what exactly a required sample size would mean in this context: the number of droplets is not sufficient, because each droplet does not contribute to every point on the observed ice nucleation spectrum equally. However, the number of separate ice nucleation assays is also not sufficient, as techniques that measure hundreds of droplets in a single assay should require fewer overall assays to calculate accurate statistics because there are more droplets contributing to the accuracy of each point on the measured ice nucleation spectrum. Some combination of the two is required, but there is no existing method by which the accuracy of confidence intervals for an ice nucleation spectrum can be evaluated based on the relevant sample sizes."

This concept is also discussed in relation to our proposed techniques as follows:

"Although we cannot theoretically determine the sample sizes required for accurate confidence bands using empirical bootstrapping due to the same limitations discussed previously, the sample sizes required for accurate confidence bands can be empirically evaluated by testing how many assays, droplets, and simulated spectra are required for confidence bands to converge

(therefore reducing the uncertainty of the confidence bands due to sample size).”

p.13, l.40: The chosen sample sizes (e.g., n=91) seem to be arbitrary - please comment.

The parenthetical statement “(the number of droplets present in the two unaged FUE freezing experiments)” has been added to clarify this.

p.14, l.56: Replace „per degree Celsius measured“ with „per Kelvin“.

Both are valid SI unit choices, and we prefer to use degrees Celsius, in part because this is directly referenced to the melting point of water and thus very relevant for discussing ice nucleation.

[1] **NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, accessed August 16th, 2022**

[2] Polen, M., Brubaker, T., Somers, J., & Sullivan, R. C. (2018). Cleaning up our water: Reducing interferences from nonhomogeneous freezing of “pure” water in droplet freezing assays of ice-nucleating particles. *Atmospheric Measurement Techniques*, *11*(9), 5315–5334. https://doi.org/10.5194/amt-11-5315-2018