We thank the reviewers for their constructive feedback. Our responses follow. We use DF for 'distribution-free'.

[R1] ***'...if I understand the notation $\mathbf{\Omega(n)}$...'*** The interpretation of $\Omega(n)$ is correct; we would like to point out that n is #calibration points, which only depends on the desired calibration error ϵ (typically, $n \approx 1/\epsilon^2$), and does not have to be a constant fraction of #training points. As far as we know, Bayesian calibration does require distributional assumptions; examining whether they yield DF calibration is an interesting direction for future work.

'...examples that are in the appendix...move to the main paper...' We agree. If the paper is accepted, we will add explicit numbers and examples using the extra ninth page.

[R2] ***'...the paper does not contextualize and differentiate its contribution clearly...'*** The literature on calibration can be split in two parts: how to *measure* calibration (Brier score, ECE error, proper scoring rules, etc) and how to *achieve* calibration (Platt scaling, binning, etc). Our work falls in the second category. Within this, the only DF results we are aware of stem from "Venn prediction", which we discuss in Appendix E (however, their theoretical approach does not lead to an actionable algorithm with a DF guarantee). If the paper is accepted, we will contextualize our work as above (including its history in meteorology/statistics) in more detail using the extra ninth page.

'...more intuition for Theorem 1...' We agree. On line 127 we have added the following

$$\underbrace{\left|\mathbb{E}\left[Y \mid f(x)\right] - f(x)\right| \leqslant \epsilon(f(x))}_{\text{calibration}} \implies \underbrace{f(x) \in C(f(x))}_{\text{CI wrt } f} := [f(x) - \varepsilon(f(x)), f(x) + \varepsilon(f(x))],$$

and on line 128, we have called u, l, m as the left-endpoint, right-endpoint and midpoint functions respectively and ε as the constant function returning the largest interval radius. We also have intuition to share about Theorem 2; we have now added a paragraph just after describing how, and why, the proof works (not reproduced here due to limited space).

'...function C can only be (1-alpha)-CI if it is measurable...' We agree. Measurability is indeed required. We have now added a clarification/remark in the notations paragraph.

[R3] ***'...[21] previously showed histogram binning is needed for calibration... this has been partially shown before it other works (sic)...'*** We respectfully disagree, as we explain next. [21] does make claims about the difficulty of *evaluating/measuring* miscalibration of Platt/temperature scaling, but these are not formal impossibility results. Mathematically, [21, Thm 4.1] showed that binning *suffices* for calibration, and compares scaling+binning to the best within a fixed, regular, injective parametric class. In contrast, our results are new in the DF setting, showing not only that binning is necessary for calibration (Thm 3), but also sufficient (Cor 4). Thm 3 in particular implies that the best in the aforementioned parametric class in [21] itself cannot have a DF guarantee, and it also yields a formal impossibility result for Platt/temperature scaling. We will clarify this in our expanded discussion about related work.

'...not a single experiment...' Appendix D has 'proof-of-concept' simulation under covariate shift, but we agree that the focus of this paper is theoretical, specifically to shed light on what is or is not achievable without making (usually unverifiable) distributional assumptions. Luckily, our theory applies to a large swatch of popular existing methods, for which a lot of empirical work already exists in the literature.

'...a lot of (useful) theory, though it is hard to grasp without examples or visualizations...' We do have intuition/examples to share. If accepted, we will use the extra ninth page to add these.

'Line 241 "which would make the pi_b estimates less calibrated"... What exactly is meant here?...' We agree the current phrasing may be misleading. We intended to say something straightforward: with fixed-width binning, some bins may have much fewer samples relative to others, and calibration cannot be guaranteed/verified for those bins. This is the typical motivation for uniform-mass binning. We have updated the text to clarify this with a clearer word choice.