ChatGPT

Executive Summary

The manuscript proposes a novel distribution-free method for conditional density estimation and uncertainty quantification, based on partitioning the covariate space into bins that minimize a leave-one-out Continuous Ranked Probability Score (LOO-CRPS). The bins' empirical CDFs serve as predictive distributions, and a conformal prediction step (with CRPS as the nonconformity measure) yields finite-sample coverage guarantees. The authors derive a closed-form formula for the LOO-CRPS of a bin and use dynamic programming to find the optimal partition. Experiments on standard benchmarks (Old Faithful geyser and the Motorcycle dataset) show that the proposed method produces substantially *narrower* 90% prediction intervals than split-conformal baselines (Gaussian residual, CQR, CQR-QRF) while maintaining near-nominal coverage. Strengths of the paper include its clear motivation (aligning the partitioning criterion with the predictive score), the analytic tractability of the cost function, and competitive empirical performance. Weaknesses include limited scope (the method is essentially for a single continuous covariate), high computational complexity ($O(n^2K)$), and unclear treatment of multi-dimensional features and conditional validity. Overall, the idea is original and well-motivated, but the scope and evaluation are somewhat narrow. I recommend a **revision**: the paper should better situate itself in the literature on distributional conformal prediction, clarify assumptions, and strengthen empirical validation (e.g. on larger or multivariate tasks).

Originality and Relevance

Novelty: The paper introduces a novel use of CRPS as an objective for partitioning in conformal regression. While conformal methods for prediction intervals (e.g. Lei *et al.*, 2018 [1]; Romano *et al.*, 2019 [2]) and the idea of adaptive (heteroscedastic) intervals (e.g. Conformalized Quantile Regression (CQR) [2]) are well-known, the proposed *CRPS-optimal binning* approach appears to be new. No prior work optimizes bin boundaries directly with respect to CRPS. The approach is reminiscent of histogram or piecewise-constant regression models, but here the bins are chosen by an explicit proper scoring rule. Related literature includes "distributional conformal prediction" (Chernozhukov *et al.*, 2021 [3]) and the very recent bin-conditional conformal prediction (Randahl *et al.*, 2025 [4]) which adjusts coverage across outcome bins. However, those differ substantially: e.g. Chernozhukov *et al.* use the PIT and rank permutation to achieve (approximate) conditional validity [3], and Randahl *et al.* enforce coverage within pre-specified outcome subsets [4]. By contrast, the present method builds a full predictive distribution (via within-bin ECDFs) and directly optimizes interval quality via CRPS. To my knowledge this exact combination has not been studied in top-tier ML venues. Thus the paper makes an **original** contribution by linking CRPS-based scoring to conformal/regression partitioning.

Relevance to field: Estimating full conditional distributions (or valid prediction sets) is an important problem in ML and statistics. Conformal prediction methods provide *distribution-free* uncertainty quantification [1] [2], but many existing conformal methods produce intervals of roughly constant width (non-adaptive to heteroscedasticity) [5]. Adaptive methods like CQR [24] improve this, but still rely on a base quantile model. The proposed method directly tackles adaptivity via data-driven binning. This aligns with the current emphasis on valid uncertainty quantification (e.g. distributional forecasts, predictive inference [1] [3]). The focus on CRPS is also well-motivated, since CRPS is a **strictly proper scoring rule** for distribution forecasts [6] [7], bridging probabilistic forecasting with evaluation. The idea of optimizing the exact score that will be used for evaluation (CRPS) is elegant.

However, the scope is somewhat narrow: the method assumes a **single ordering covariate** (the bins are "covariate-sorted"), so it is essentially 1-dimensional in X. This limits its applicability compared to e.g. CQR or distributional learning methods. The authors should clarify if and how the method extends to multivariate X. In its current form it is primarily applicable to one-dimensional regression tasks. This is a significant limitation relative to the broader literature on regression uncertainty (which often addresses multi-dimensional inputs).

In summary, the paper tackles an important problem and introduces a **new approach**. It clearly fills a gap between non-parametric density estimation and conformal calibration. I did not find any existing work that directly does "CRPS-minimal binning," so the originality is high. It would help to cite and contrast more explicitly with related approaches (e.g. [13, 22, 24]) in the final version.

Technical Soundness

Methodology: The core idea is sound: for each candidate bin, compute the total leave-one-out CRPS and choose splits to minimize total CRPS. The authors derive a closed-form cost formula `cost(S) = m W / (m-1)^2` for a bin of size *m*, where *W* is the sum of pairwise absolute differences of responses within the bin. This formula (essentially an energy-distance expression of CRPS [7]) appears correct. The use of a Fenwick tree for efficient updates is a neat algorithmic detail (citing Fenwick [5]) and the dynamic programming for the global K-partition is classic (citing optimal binary search trees [6]).

However, some assumptions and details need clarification:

- **Exchangeability:** The validity guarantee of conformal prediction requires exchangeability of calibration and test points. Here, "full-data conformal" is used (all points serve in calibration via leave-one-out) [2]. The authors note this reuse improves efficiency but not in matched-sample regime (Sec. 10). It is not fully clear if the claimed *finite-sample marginal coverage* guarantee holds with this reuse. Standard split-conformal and Jackknife+-type methods do guarantee coverage [8]. It would be good to explicitly state any theoretical coverage guarantee (or lack thereof) of the proposed method. Does the conformal set guarantee at least (1–ε) coverage marginally? How does using the fitted ECDF as a predictive CDF affect independence assumptions?

- **Complexity:** The algorithm runs in **$O(n^2K)$** time and $O(n^2)$ memory. This severely limits scalability: even moderate datasets (n~10,000) would be intractable. The paper should discuss this explicitly. In practice the experiments use n up to ~300, which is small. If the method is to be used in ML practice, complexity is a major issue. The authors might mention possible approximations (e.g. greedy splits) or note that the method is best for small-to-medium data.

- **Choice of *K*:** The paper shows that minimizing training CRPS overfits (seemingly due to the usual "smaller bins always reduce in-sample loss"), so they instead select *K* via held-out CRPS on alternating splits (yielding a U-shaped validation curve). This is reasonable, but details matter: How many splits are used? Are splits random or deterministic alternating halves? Is the split fixed once, or cross-validated? The description ("alternating held-out split") is a bit opaque. More clarity on how *K* is chosen (e.g. k-fold CV on CRPS) would strengthen reproducibility.

- **Within-bin assumptions:** Implicitly, each bin is treated as having exchangeable residuals and equal variances. With many bins, some bins become small, risking unreliable ECDF estimates. The paper should discuss the trade-off: smaller bins can better capture heteroskedasticity but increase variance of ECDF (bias–variance tradeoff is mentioned but formal guarantees are not given). Are there safeguards (e.g. minimum bin size)?

- **Venn and conformal outputs:** The method produces two "predictive objects": a Venn prediction band and a conformal prediction set using CRPS. It's unclear how these differ. The Venn predictors (Venn-Abers style) guarantee coverage at each level (by construction) [9], but are typically more conservative. The paper should clarify the definitions and distinctions. It was not immediately clear whether the final reported intervals use Venn or conformal. (Table 2/3 captions imply the conformal intervals are shown.) More explanation of the role of each would help.

In summary, the technical ideas seem correct and well-founded, but the presentation should more explicitly articulate assumptions (iid, exchangeability) and limitations (scalability, 1D covariate). The derivations for CRPS and DP are solid. The authors should double-check that the conformal coverage claim is properly justified under their scheme.

Experiments and Reproducibility

Empirical evaluation: The experiments focus on two standard small-regression benchmarks with heteroskedastic noise:

- **Old Faithful geyser (n~272)**: This classic dataset exhibits relatively mild heteroscedasticity. Table 3 shows at nominal 90% level the proposed method achieves ~90.3% coverage (good) with mean interval width 1.20 minutes, versus 91–91.4% coverage and widths 1.33–1.68 for CQR, CQR-QRF, and Gaussian conformal. The new method is notably sharper.
- **Motorcycle accident (n=133)**: A heavily heteroskedastic dataset. Table 4 reports 91.0% coverage vs nominal 90% (slightly conservative) with width 78.9g, compared to 92.5–93.1% coverage and widths 87.9–172.4g for the baselines. Again, the proposed method gives much narrower intervals. Figure 11 and 12 (not shown here) illustrate how the binning adapts to high-variance regions.

These results suggest the method effectively captures local variability and yields tighter intervals. This aligns with [24]'s claim that CQR improves efficiency (shorter intervals) under heteroskedasticity. The authors' claim that they get "substantially narrower intervals while maintaining coverage" is supported by the numbers above (e.g. Motorcycle: width 78.9 vs 134–173).

Baselines and fairness: The comparison to standard split-conformal methods is appropriate. CQR (Romano *et al.*, 2019) [2] and CQR-QRF (using Quantile Regression Forest) are strong modern baselines. The Gaussian OLS+conformal baseline is weak (as expected). It's good that the authors used QRF with 500 trees (ref [16], Meinshausen 2006) and classical quantile regression for CQR. However, more details are needed: e.g. how were hyperparameters chosen? Did they use the same held-out split for all methods? The text suggests all competitors use 50/50 splits to match the proposed approach. That seems fair, but should be explicitly confirmed to ensure apples-to-apples.

Reproducibility: The paper mentions an R package for conformalInference [10] (for Lei's split and jackknife methods), but it is unclear if the authors release code for their own method. Given the complexity (Fenwick trees, DP, etc.), code release would greatly aid reproducibility. The manuscript should state whether code and datasets are available. Also, details like random seeds or variability (the tables show "±" over 200 splits) indicate robustness, which is good practice.

Scope of experiments: The experiments, while demonstrating the idea, are limited. Only two real datasets (n<300) are used, both one-dimensional X. It would strengthen the paper to test on more diverse data (e.g. UCI regression datasets with varying heteroskedasticity) or even synthetics with known distributions. Table 1 or synthetic examples are mentioned (e.g. a bimodal synthetic for illustration), but no quantitative results beyond the two benchmarks. I suggest adding at least one or two higher-dimensional examples (perhaps by discretizing or by projecting onto one feature) to show if the method can handle more complex inputs, even if only by using one covariate. Alternatively, clarify that the method is strictly univariate in X.

Statistical validity: The reported coverage results (Table 4) show that split-conformal methods often over-cover (e.g. 92–93% instead of 90%), which is typical (conservative). The new method's coverage is closer to nominal, which is good. It would be useful to include a simulation where true conditional distributions are known, to verify coverage more precisely and examine calibration (as in Chernozhukov *et al.*, 2021). Without ground truth, one relies on empirical averages. The paper does a decent job of averaging over random splits (200 runs) to reduce randomness in coverage.

Overall, the experiments support the claims, but are somewhat narrow. The paper would benefit from clearer mention of code availability and possibly additional datasets.

Clarity of Presentation

Organization: The paper is generally well-structured: introduction motivates the problem, Section 2 reviews conformal and CRPS, Section 3 presents the method (cost derivation, DP, K-selection), Section 4 gives experiments, etc. The outline is logical. The abstract and introduction clearly state the contributions.

Writing quality: The prose is mostly clear and technical. A few parts could be tightened for readability. For example, the introduction's second paragraph is a bit dense; splitting into shorter sentences may help. The term "conformal regression" in the title is somewhat vague (the method is really conditional distribution estimation). The keywords and early sentences mention "conformal regression" and "distribution-free", which aligns with the venue goals.

Notation and definitions: Most key terms are defined. CRPS is presumably defined (though I didn't see the formula, it is given by references [6]). The cost function `cost(S)` is clearly explained after Proposition 1. The algorithmic steps are described in text, but pseudocode or an explicit algorithm box would improve clarity. For instance, listing "Algorithm: Compute bins via DP" with inputs/outputs could help readers replicate.

Figures and tables: The figures (e.g. Fig. 10 "90% prediction intervals" and Fig. 11, 12 shown in PDF) illustrate the interval adaptation well. They are legible and appropriately captioned. Tables 2–4 are informative. A minor suggestion: in Table 3 and 4, the dataset names ("Old Faithful", "Motorcycle") could be in the caption or table heading for quick reference. Also, the ± confidence intervals (or std errors) for coverage/width are useful.

Clarity of claims: The paper should more explicitly state which guarantees hold. It mentions "finite-sample marginal coverage guarantee" in the abstract, but does not restate this in the main text or theorem. If there is no formal theorem, at least a remark should say "by the exchangeability arguments of conformal prediction [1], the conformal set has valid coverage". Also, the term "Venn prediction band" might be unfamiliar; a brief intuitive description would help.

Citations: The paper cites key algorithmic and statistical references (Fenwick [5], Knuth [6], Yao [7]). It cites conformal and CRPS literature such as Lei et al. (JASA) [1] and Barber *et al.* (Jackknife+) [11]. However, it should include more recent ML references on distributional prediction. For example, the Venn predictor framework (Vovk *et al.*) or neural approaches to predictive distributions. The newly published works we found (Randahl et al. 2026 [4] on bin-conditional CP, Zhang *et al.* 2023 [12] on conformal predictive systems) should be discussed. I recommend adding citations to [13], [22], and [18] where appropriate, as they are closely related in spirit.

In summary, the exposition is good but can be improved in places (definitions of terms, algorithmic clarity, explicit statements of guarantees, and some additional context citations).

References and Related Work

The manuscript's references cover many core topics (CRPS, Fenwick trees, dynamic programming, conformal basics, Jackknife+) but misses some important context:

- **Conformal Prediction & Coverage:** It cites Lei *et al.* (2018) [1] and Barber *et al.* (2021) [11], which are standard. It should explicitly cite Vovk *et al.* (2005, *and/or* 2020) on conformal prediction to acknowledge the framework's origin and coverage guarantee. The Chernozhukov *et al.* (2021) PNAS paper on distributional CP [3] is highly relevant (conditionally valid intervals via rank transform), and should be mentioned as a recent advance. The Bin-Conditional CP (Randahl 2026) [4] is directly related (enforcing coverage across bins). These works should be cited in the related work or introduction to situate the new method.

- **Adaptive Prediction Intervals:** The Conformalized Quantile Regression (CQR) paper [2] is cited (as [14] in PDF), which is good. Are there other methods like "EnbPI" or "Adaptive Confidence Bands" that could be mentioned? Possibly not needed, but the authors may want to ensure that any method claiming adaptive intervals is cited.

- **CRPS and Proper Scoring:** The classic reference Gneiting & Raftery (2007) on proper scoring rules (JRSS-B) should be cited when introducing CRPS [7]. Gneiting (2011) on quantiles as optimal forecasts may also be relevant. The Zamo & Naveau (2018) paper [6] on estimating CRPS is interesting but more about estimation error. It's good background but not essential to cite; however [20†L11-L15] gave enough justification for CRPS use.

- **Distributional Regression:** If the paper is about conditional distributions, some references on "distributional regression" from statistics could be cited (e.g. generalized additive models for location, scale, shape; or quantile regression forests [16]). The QRF reference is already [16] Meinshausen 2006. Also, mention of Gaussian Processes for quantiles (GP-VQR) or Bayesian quantile regression might be outside scope, but could be named.

- **Conformal and MCMC/Bootstrap:** The paper mentions "full-data conformal (which is related to Jackknife+)" but does not cite Liu & Lei 2018 (CV+) or other variants. The Jackknife+ [20†L31-L37] covers part of that.

Overall, I suggest:

- **Add**: Chernozhukov *et al.* (2021) PNAS or arXiv [3] , Randahl *et al.* (2026) [4] , Vovk *et al.* (2005) if space permits.
- **Ensure**: key conformal and CRPS references are properly cited (they have most).
- **Check** if any suitable references for adaptive binning or piecewise-constant density estimation exist; if not, note this is a novel use.

Strengths

- **Well-Motivated Idea:** Aligning the binning criterion with the CRPS objective is an elegant idea. It ensures that what is optimized during training (LOO-CRPS) directly matches the evaluation goal. This coherence is often lacking in conformal methods (as noted in Sec. 3).
- **Closed-form and Efficient Computation:** The derivation of the closed-form CRPS cost for a bin and the clever use of a Fenwick tree (for updating pairwise sums) are technical strengths. This makes an otherwise intractable leave-one-out computation feasible.
- **Dynamic Programming for Global Optimum:** Guaranteeing a globally optimal K-partition (for given K) is strong; many piecewise methods settle for greedy splits. This ensures the partition is truly CRPS-optimal.
- **Empirical Performance:** The method produces clearly sharper intervals than strong baselines (CQR, QRF) on the tested problems, validating its practical advantage in heteroskedastic settings. The results show both nominal coverage and reduced width, indicating improved efficiency.
- **Comprehensive Evaluation (Within Scope):** The experiments report coverage and width averaged over many trials, giving a robust comparison. The paper also provides illustrative figures (e.g. interval plots, U-shape selection curves) that convey the behavior well.
- **Clarity of Writing (Mostly):** The manuscript is written in a technically clear style, with well-labeled tables/figures and a logical flow.

Weaknesses

- **Limited Scope (Single Covariate):** The method as described only handles one-dimensional X (observations sorted by one covariate). There is no discussion of extensions to multivariate inputs. This is a major limitation: many practical problems have multivariate features. Without modification, the method cannot be applied when $d>1$. The paper should explicitly acknowledge this constraint and possibly discuss ideas (e.g. tree-based binning).
- **Computational Scalability:** The algorithm's $O(n^2K)$ time and $O(n^2)$ memory is prohibitive for large datasets. This needs emphasis. For modern ML tasks (n in the thousands or more), the method would be too slow. No discussion of this is present. Even moderate n (e.g. 5000) would make DP

infeasible. The authors should either propose approximations or state that this method is intended only for relatively small data.

- **Theoretical Guarantees:** There is no clear theorem or proof regarding coverage. The abstract claims "finite-sample marginal coverage guarantee", but the text does not formalize it. Given that the method uses all data in calibration (unlike split-conformal), it is unclear if the usual conformal coverage arguments fully apply. A careful statement of any assumptions (exchangeability within bins) and what guarantee holds is missing. Without this, it's hard to judge the reliability of the intervals in all cases.
- **Presentation of Method:** Some parts are dense. For instance, the discussion in Sec. 3.1 about different scoring criteria (CRPS vs Cramér distance) is interesting but somewhat buried in text and equations. The main algorithm is not given in pseudocode, which would aid understanding.
- **Empirical Evaluation Depth:** Only two small datasets are tested. No experiments on synthetic data where the true conditional distribution is known (to directly measure calibration of the predictive CDF) are shown. No large-scale or multi-dimensional experiments. This limits confidence in general usefulness. Also, runtime performance or sensitivity to K is not reported.
- **Reference Gaps:** As noted, some relevant recent works (Randahl's BCCP, Chernozhukov's DCP) are not cited or discussed. Also, the Venn prediction literature and related conformal distributional methods (e.g. Tibshirani's CV+) are not referenced. This makes the related work section incomplete.
- **Interpretability of Outputs:** The paper does not clearly explain the output format. Are we producing a full predictive CDF? A set of quantiles? The term "Venn prediction band" might confuse some readers. It should be clarified what the final user gets (e.g. "a 90% confidence band for the CDF" or "90% PI for Y at each x").

Suggestions for Improvement

1. **Clarify scope and assumptions:** Explicitly state that X is assumed 1-dimensional (or rankable). Discuss the limitation to 1D and potential extensions (e.g. use of decision trees or Voronoi partitions for multivariate X). Clarify i.i.d. assumptions and the role of within-bin exchangeability in enabling coverage.

2. **Strengthen theoretical claims:** If possible, include a theorem or proposition stating the coverage guarantee (even if marginal only). At minimum, a remark should explain why the conformal procedure here still yields valid coverage (citing Lei *et al.* [1]). Address how "full-data conformal" (using ECDF on all points) fits into existing theory (e.g. as in the Jackknife+ arguments [11]).

3. **Add pseudocode:** Provide an algorithm box summarizing the CRPS-based binning procedure, DP, and conformal step. This will help readers understand and implement it.

4. **Improve experiments:**

5. Include additional datasets. For example, choose one or two more from UCI or simulate data with known heteroskedastic patterns. Particularly, test on a dataset with 2D covariates by binning on one feature (or adapt the method to multiple features).

6. If feasible, measure calibration of the full predictive CDF (e.g. PIT histograms or coverage of quantiles) on synthetic data.

7. Report computation time (even for small datasets) to highlight the cost. Possibly compare runtime of the new method vs CQR.

8. Demonstrate the effect of different K values (e.g. show how interval width changes as bins increase).

9. **Discuss extensions or alternatives:** Suggest how one might approximate the DP for large n (e.g. greedy binning) or reduce complexity (e.g. coarser data summary). Mention whether kernel or nearest-neighbors approaches could achieve similar objectives.

10. **Clarify Venn vs Conformal outputs:** Provide a brief explanation of "Venn prediction band" for readers unfamiliar with it, and clarify what output is reported in experiments (Table 2/3 likely use the conformal CRPS set).

11. **Improve referencing:** Cite Chernozhukov *et al.* (2021) [3] when discussing related approaches to conditional prediction intervals; cite Randahl *et al.* (2026) [4] for bin-based calibration; cite Gneiting & Raftery (2007) [7] when introducing CRPS; cite the CQR paper explicitly if not already; and ensure Vovk's fundamental conformal prediction book or papers are mentioned.

12. **Minor edits:** Proofread for minor typos (e.g. "strategies is" → "strategy is" in Sec. 9.1) and ensure all notation is consistent (e.g. whether K or K_* is used). Improve figure/table labels as needed.

Questions for the Authors

1. **Multiple covariates:** How would your method handle multivariate inputs? Is there a straightforward extension (e.g. sorting on one key feature or recursive partitioning)? If not, this limitation should be clearly noted.

2. **Coverage guarantee:** Can you clarify the theoretical coverage guarantee of your final conformal interval? Does it hold for any data distribution (marginally)? What about conditional validity? How does reusing the data for calibration affect coverage?

3. **Choice of *K* and splits:** You use an "alternating held-out split" to select K. Could you elaborate on this procedure? How many splits were used, and how stable was the U-shaped CRPS curve? Would standard k-fold cross-validation give similar results?

4. **Small bin sizes:** In high-noise regions, the optimal binning may create very small bins (as seen for Motorcycle). Does this violate the implicit assumption that data within a bin are exchangeable? How do you guard against extremely small bins (e.g. one or two points)?

5. **Comparison with other distributional methods:** Did you consider comparing to non-conformal distributional models (e.g. Gaussian Processes or Bayesian Neural Nets) or other conformal variants (e.g. CV+ [13], vanilla Conformal Quantile Regression with different quantile models)?

6. **Code availability:** Will you release the implementation? In particular, code for the DP partition and CRPS calculations would be valuable for the community.

7. **Venn vs Conformal:** You mention both Venn prediction bands and a CRPS-based conformal set. Which of these is evaluated in Tables 3–4? Are these two approaches combined or is only one ultimately used for reporting?

Recommendation

Decision: *Revise (score ~7/10)*

Justification: This paper presents an **innovative and interesting method** that aligns binning with a proper scoring rule and integrates conformal calibration. The approach is well-motivated and shows promising results on benchmarks. However, there are notable gaps and limitations that need addressing before publication in a top-tier venue. In particular, the restriction to univariate inputs, the computational cost, and the unclear coverage guarantees are important concerns. The experimental evaluation should be broadened and the exposition tightened. The overall significance is high enough to merit publication, but **major revisions** are needed to strengthen the empirical and theoretical aspects and to properly position the work relative to recent literature.

Suggested Score: 7 (good paper, but incomplete in its current form; revision required).

---

1 8 10 13 stat.cmu.edu
https://www.stat.cmu.edu/~ryantibs/papers/conformal.pdf

2 5 Conformalized Quantile Regression
https://papers.neurips.cc/paper/8613-conformalized-quantile-regression.pdf

3 [1909.07889] Distributional conformal prediction
https://arxiv.org/abs/1909.07889

4 Bin-Conditional Conformal Prediction of Fatalities from Armed Conflict | Political Analysis | Cambridge Core
https://www.cambridge.org/core/journals/political-analysis/article/binconditional-conformal-prediction-of-fatalities-from-armed-conflict/4519907469FE9DF3B5088D88AA10D202

6 7 11 Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts | Mathematical Geosciences | Springer Nature Link
https://link.springer.com/article/10.1007/s11004-017-9709-7

9 12 A conformal predictive system for distribution regression with random features | Soft Computing | Springer Nature Link
https://link.springer.com/article/10.1007/s00500-023-07859-w