**We have implemented four major revisions based on feedback from the reviewers and editor:**

1. **Included Covariates:** asd
2. **Implemented Proper Scoring (and clarified evaluation methods):** asd
3. **Expanded Computational Complexity:** asd
4. **Included Simulation of Marginal Estimates from Random Effect Models:** asd

**A limited comparison at other sample sizes?**

**Reviewer 1**

Overall I enjoyed the manuscript, however there are a couple of suggestions that I would like to make to the authors in order to improve the article’s cohesiveness and presentation. Below are my comments and feedback for the authors.

1. Abstract: I think that the abstract would benefit from clearly stating that you are focused on data that features two measurements per observational unit, which allows you to use GJRM in your simulation and empirical studies. This aspect is stated too late in the manuscript.  
   **✅ Added a sentence to this effect to the abstract for clarity.**
2. Introduction: I enjoyed the presentation of the different methods considered throughout the manuscript. However, as mentioned in my previous point, I suggest that you state as early as possible that you only work with two measurements per observational unit.  
   **⚠️ I’ve made this more clear in the introduction but I think we need to figure out how to put this closer to the top. Temporarily I’ve bolded it.**
3. Page 3, lines 26-33: I think this paragraph could benefit from mentioning limitations of the listed copula packages / methods. For example, in your simulations and application you work with discrete responses. If I remember correctly, not all of the mentioned packages and methods for copula modelling support these types of responses directly. Interested readers would benefit from this information.  
   ⚠️
4. Page 3, lines 46-49: I suggest that you state clearly, in one paragraph, what are the selection criteria for the papers you selected for your review.  
   ⚠️
5. Page 4, lines 27-29: I find it puzzling that you limit the investigation to cases without covariates after you introduced the models in Page 2 with covariates. I suggest that you either re-introduce the models using a notation that does not include covariates early on or add a scenario with covariate(s) to your simulations. This will increase the scientific value of your investigation.  
   **✅ Added a whole section on covariates.**
6. Page 4, line 47: GJRM implements various distributions, however it does not feature all GAMLSS distributions.  
   ⚠️
7. Page 5, line 1: Again, please state as early as possible that you restrict your investigation to the case of T = 2 measurements per observational unit.  
   ⚠️
8. Page 5, line 14: I am very happy to see that you simulate data that does not originate from a copula, which could give GJRM an unfair advantage. I suggest that you highlight this in the Introduction where you list the key points that your manuscript demonstrates.  
   ⚠️
9. Page 8, line 28: Is there a reason why you omit for example gamCopula, VineCopula and / or Bayesian alternatives from your simulation study? Please state it clearly in the manuscript.  
   ⚠️
10. Figures 2, 3, 4 and 5: Please add solid colour lines to the plot mixed with the dashed and dotted lines. It was difficult to distinguish the different estimators.  
    ⚠️
11. Page 16, lines 53-57: Please include a scenario in your simulations where you compare the fit of a two-step or “inference for margins” estimation approach as you describe here to support your chosen method for the Application.  
    ⚠️

**Reviewer 2**

Review of the paper “A comparison between copula-based, mixed model, and estimating equation methods for analysis of bivariate correlated data”

Review Questionnaire

Question Response Justification

1. Does the manuscript contain new and significant information to justify publication? Yes It presents a novel empirical comparison of copula-based joint modeling (GJRM), GLMMs, and GEEs for bivariate non-normal data, a topic not thoroughly benchmarked before.  
   **✅**
2. Is the problem significant and concisely stated? See report The problem is important, but the paper frames itself as addressing regression, while it only considers joint modeling without covariates.  
   ✅ Covariates now added.
3. Are the methods appropriate and are they described comprehensively? See report The methods are well-chosen and competently implemented, but key aspects such as covariate omission and scoring criteria require clarification.  
   **✅** Covariate simulations added. Proper scoring now incorporated.
4. Are the interpretations and conclusions justified by the results? See report Conclusions are broadly supported, but limitations—especially computational cost and modeling assumptions—are not fully acknowledged.  
   ⚠️ Added expanded section on computational cost. Need to add more on computational cost to abstract + discussion / conclusion if relevant.
5. Is the summary (abstract) concise? Yes The abstract is concise and largely representative, although it omits mention of computational drawbacks.  
   ⚠️ Add computational drawbacks to abstract.
6. Is the language acceptable? Yes The manuscript is clearly written and well structured.  
   ✅

**Importance of the Problem and Overall Significance**

**Is the problem addressed by the manuscript important and relevant to the field?**

*Yes. The authors tackle the analysis of bivariate correlated outcomes with non-normal distributions, a problem that frequently arises in biostatistics and other applied fields. Choosing an appropriate modeling strategy for correlated data (especially when normality assumptions do not hold) is a significant challenge. This comparison of a copula-based method, a mixed-effects model, and GEE provides practical insight into how each approach performs. The problem is of clear interest: many researchers face decisions between using a joint modeling approach (like copulas), a GLMM (handling correlation through random effects), or a GEE (handling correlation through a working covariance) when analyzing paired or longitudinal outcomes. By focusing on these three classes of methods, the study addresses a relevant methodological question for applied statisticians.*✅

**Is the manuscript’s contribution significant?**

In principle, a systematic comparison of these methods could be quite valuable. There has been extensive development in models for correlated data, but direct comparisons across different paradigms (likelihood-based random effects, marginal quasi-likelihood, and copula dependence modeling) are not commonly found in one place. The manuscript’s contribution lies in bringing these approaches together under a unifying experiment and application, which could help readers understand the trade-offs. If executed well, the results could guide practitioners on when each method might be preferable (in terms of accuracy, computational cost, ease of interpretation, etc.). However, the significance of the contribution is currently limited by a mismatch between the paper’s stated goals and the actual analyses. The manuscript is framed as addressing “regression of non-normal correlated data,” suggesting a broad applicability, but in practice it only examines the case of two correlated response variables without any covariates. This narrower scope means some conclusions (especially regarding regression performance or covariate effects) cannot actually be drawn, potentially reducing the impact of the work. Once the scope is clarified and aligned with the content (or expanded to include covariates), the contribution will be more solid. Despite this issue, the comparative approach is novel enough to add value, since each method has different assumptions and it is instructive to see their performance side by side.

✅ Expanded to include covariates. ⚠️ Limitation of two variables is limitation of the copula approach today, this is added to intro.

**Novelty and Contribution to Knowledge**

**Does the manuscript offer new methodology or new insights?**

The paper does not introduce a brand-new statistical method; rather, it provides a comparative study of existing methods. The novelty here is in the application of three different modeling strategies to the same problem and the critical evaluation of their performance. This kind of comparative insight can be considered a contribution to knowledge, especially for practitioners who may not be aware of the relative advantages and disadvantages of each approach in practice. The manuscript appears to be one of the first to directly compare copula-based models with GLMMs and GEEs for the same dataset/simulation settings, which is an interesting exercise. In terms of insights, the authors presumably report on how the methods differ in terms of fit, criteria like AIC/BIC, and maybe parameter estimates or coverage. These results could be useful for guiding model choice (for example, understanding if the copula method provides better fit at the cost of more computation, or how GEE and GLMM differ in estimating marginal vs. conditional characteristics).

⚠️ No major suggestions here. I think this is covered. We will add more on computational complexity. We discuss marginal v conditional

**Is the work sufficiently innovative?**

As a comparison paper, the innovation is moderate. The value lies more in synthesis and evaluation than in methodological innovation. That said, given the importance of correctly analyzing correlated data, the paper addresses a practical gap. To ensure the work is seen as a meaningful contribution, the authors should clarify what new knowledge is gained. For instance, if the findings demonstrate scenarios where one approach outperforms the others (in accuracy or robustness), that is useful information. If all methods perform similarly for the given data, then the novelty might be in showing the interchangeability or subtle differences. At the moment, the novelty is undermined by the lack of any regression covariates in the analysis, this means the comparison is restricted to how each method captures the dependency between two outcome variables, rather than how they handle regression coefficients or adjust for predictors: • In the no covariate case, then there are many possibilities for modeling the joint distribution of two outcomes, either parametric or nonparametric, copula-based or not, etc. • Including covariates in a simulation scenario (or discussing results from a covariate-adjusted analysis) would greatly enhance the innovative aspect by extending insights to true regression settings (e.g., how do the methods compare in bias/efficiency for regression coefficients, which is a key question for GLMM vs GEE). In summary, the manuscript’s contribution is valuable as a comparative study, but it needs to better articulate and possibly broaden its scope to maximize its impact.

✅ Expanded to include covariates.

**Methodological Soundness and Rigor of Analysis**

**Are the methods chosen appropriate for the problem, and are they applied correctly?**

The three methods considered, copula-based modeling, GLMM, and GEE, are indeed appropriate choices for analyzing bivariate correlated data. Each represents a distinct philosophical approach: copulas model the joint distribution by specifying marginal distributions and a dependency structure; GLMMs introduce random effects to account for correlation in a conditional modeling framework; GEEs take a marginal modeling approach with a working correlation structure. It is appropriate and even insightful to compare these approaches. Generally, the manuscript demonstrates understanding of each method, and there is no fundamental objection to the methods themselves. The authors have selected relevant techniques for non-normal outcomes (for example, presumably using an appropriate copula family and marginal distributions matching the outcome types, and using a logistic or Poisson GLMM if outcomes are binary or count, etc., though this should be clearly described). One issue is that, when simply modeling the joint distribution of two outcomes, there are many non-regression related methods that could be used. For example, one could use a multivariate normal distribution with a covariance structure, or a nonparametric approach like kernel density estimation. Therefore, if the authors wish to stick to the current methods, then they should revise their simulation study and real data example to include covariates. Conversely, if the authors want to keep the current design, they should clarify that the goal and choose methods are focused on modeling the joint distribution of two outcomes, and not on regression modeling. This would help to avoid confusion about the purpose of the analysis and the relevance of the results.

⚠️Need to clearly describe models fit i.e. choosing appropriate response distribution, link etc. Added note to end of 2.2.

✅ Expanded to include covariates.

**Concerns about how the analysis is conducted**

While the choice of methods is sound, there are significant concerns with the design and execution of the analysis:

• Lack of covariates in analysis: As noted, neither the simulation study nor the real data example includes any predictor covariates (beyond perhaps an intercept). This is a surprising omission because it means the comparison is only testing how well each method captures the correlation between two response variables. In practice, one uses the chosen methods to assess covariate effects (e.g., treatment or risk factors) while accounting for correlation. By not including covariates, the authors have effectively removed a primary function of the chosen methods, making the comparison somewhat artificial. If the intent truly was to focus on the bivariate joint distribution modeling, the paper should explicitly state this and not frame it as a regression analysis. As it stands, this omission undermines the comparability of methods: e.g., a copula model in the same scenario is essentially just fitting a bivariate distribution, so it’s expected to perform better than the other two. The manuscript needs to address this either by adding covariate scenarios or by reframing the study’s aim.

✅ Expanded to include covariates.

• Simulation design and scope:: The manuscript should ensure that the simulation study (if included) is comprehensive. Important factors would include varying the sample size, number of covariates, outcome distributions (e.g., both continuous, one continuous and one binary, or both binary). It’s not fully clear if the simulations cover multiple scenarios or just one. A rigorous comparison would examine not only bias/SEs, but coverage and predictive performance under multiple settings. If only the currently limited scenario is presented, the authors should consider expanding it or at least discuss how generalizable they expect the results to be. Ensuring the simulation explores a range of conditions would make the conclusions more robust.

⚠️ We could potentially look at the difference at different sample sizes…?

• Correct application of model selection criteria:: The paper bases model selection on AIC and BIC. It is important that the authors clearly explain how these criteria are computed for each method and justify their use. For the GLMM, AIC/BIC can be used if the model is fitted by maximum likelihood (e.g., using Laplace approximation or adaptive quadrature). The authors should mention the method used to fit the GLMM (e.g., which software or package) to reassure that the AIC is reliable. In the context of GLMM, the use of AIC can be tricky (Greven & Kneib, 2010). From the manuscript, I could not understand which version have been used, i.e., based either on the marginal or on the conditional distribution. In the case of the copula-based model (likely fitted via the GJRM package), AIC/BIC are applicable since it’s a likelihood-based approach; again, detail on how the log-likelihood was obtained would be useful for transparency. The biggest concern is GEE: since GEE is not a likelihood-based method, one cannot compute a log-likelihood and AIC in the usual way. Couldn’t the authors use QIC (Quasi-likelihood Information Criterion) or some surrogate for AIC in the GEE context? This is not mentioned in the current manuscript. The manuscript should be revised to clarify this point. If no formal criterion was used for GEE, then the comparison across the three approaches is on unequal footing. This could be remedied by instead using a common metric for all methods (for example, evaluating predictive accuracy on held-out data, or using proper scoring rules as noted above).

✅ Added proper scoring. Also include QIC for comparison but it’s not very comparable.

• Use of proper scoring rules for comparison: The purpose of the chosen model selection method in this context is unclear to me. Why not use cross-validation with metrics related to probabilistic forecasting (Gneiting, & Katzfuss, 2014) like the energy score or variogram score, or log-density score? Or if probabilistic forecasting is not the goal, although in this context it seems to be, any metric that is considered of interest? To compare the methods fairly, especially in terms of predictive performance, the authors should consider proper scoring rules. For instance, they could simulate a test dataset (or use cross-validation) and evaluate how well each method predicts the bivariate outcomes. Scores like the Energy Score or Variogram Score are designed to compare the quality of multivariate probabilistic predictions (with the energy score being a generalization of the continuous ranked probability score to multi-dimensions, and the variogram score focusing on pairwise correlations). These scoring approaches would provide a principled way to compare copula vs GLMM vs GEE on equal terms. In the current manuscript, by relying solely on AIC/BIC for in-sample fit, the comparison might be biased or not reflective of achieving any specific goal. I strongly encourage the authors to incorporate a discussion of these scoring approaches, even if they do not carry out a full predictive comparison, to demonstrate awareness of modern model assessment tools. At minimum, the limitations of using AIC/BIC in this context (as discussed above) should be openly acknowledged.

✅ Added proper scoring.

• Computational considerations:: The GJRM approach is considerably more computationally intensive than the others, but this is only briefly mentioned. More transparent discussion, including in the abstract, is needed, as this may affect practical adoption. The analysis should comment on the computational feasibility of each method. Were any convergence issues encountered? A rigorous analysis would note if certain methods struggled to converge or required specific tuning. In a simulation study, computing time can be a significant practical factor, if one method is orders of magnitude slower, that is a relevant result for readers. Currently, only the only computation time table seems like it’s a bit swept under the rug, which downplays an important practical aspect. This should be addressed by explicitly discussing computation times and possibly summarizing them in a table or text.

⚠️ Need to add computational drawbacks to abstract. Need to also talk about convergence issues across methods.

In terms of correctness, there is likely no issue with the numerical results themselves. The key methodological shortcomings are in the scope (no covariates) and comparison metrics, as described. These need to be rectified to consider the analysis rigorous. Once the authors make these adjustments, the methods and analysis would be on much firmer footing.

✅ Added proper scoring. ✅ Expanded to include covariates.

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika, 97(4), 773-789. Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. Annual Review of Statistics and Its Application, 1(1), 125-151.

**Literature Review and Contextualization**

The manuscript provides an overview of copula models, GLMMs, and GEEs, citing classical references for each. However, there are some notable omissions and aspects that need improvement:

• Omission of recent GEE developments: The manuscript does not reference the recent R package glmtoolbox (Vanegas et al., 2023) which offers a state-of-the-art implementation of GEE​. This package includes advanced features for correlation structure selection and diagnostics in GEE analyses. Mentioning such a development is important for context, as it shows the authors are aware of current tools and performance improvements in the GEE arena. Additionally, I wonder why the authors did not use glmtoolbox for their GEE implementation or whether they considered it. If they did not, a brief discussion of the reasons would be helpful.

• Copula modeling references: The authors should cite relevant references for copula-based modeling beyond the regression context.

• Scoring rules and model assessment: If the authors decide to discuss proper scoring rules as suggested, they should cite key references in that area.

To summarize, the literature review is generally on the right track but needs updating. Including the omitted references and discussing the rationale for each method’s inclusion will improve the manuscript’s credibility. The lack of mention of glmtoolbox in particular stands out and should be remedied.

✅ Added glmtoolbox to software section and introduction. Updated all GEE model fits to use this package.

⚠️ Potentially still need to expand lit review for copulas, proper scoring rules.

**Clarity of Presentation and Structure**

The writing is mostly clear in terms of language (no major grammatical issues were noted), but there are some points of confusion in the exposition that need to be fixed:

• Introduction and framing: As mentioned, the introduction currently frames the paper as dealing with regression models for correlated data, which sets an expectation that covariates and regression coefficients will be part of the analysis. Because the actual content doesn’t include covariates, a reader can become confused or feel misled. To improve clarity, the authors should rewrite portions of the introduction to accurately reflect the scope. For example, if the focus is on “methods for joint modeling of two correlated responses,” say so explicitly. If the authors intend the results to generalize to regression settings, they need to explain the reasoning for studying the no-covariate case (perhaps as a simpler test-bed) and discuss this generalization in the Discussion. Clarity in motivation is crucial: the reader should understand why this comparison is being done and what precisely is being compared (joint distribution fit? accuracy of correlation estimates? Etc.).

✅ Expanded to include covariates.

• Discussion of limitations: A key part of clarity and transparency is acknowledging limitations. The manuscript should better discuss the limitations we have identified.

⚠️ Talk about limitations. Maybe sample size?

• Organization: The overall structure (Introduction, Methods, Simulation Study, Application, Discussion, Conclusion) is fine. Within that structure, ensure each section flows logically.

• Contribution of this paper: It seems like more than half of the section devoted to contribution of this paper is a literature review and should probably be moved to the introduction. The authors should better clarify what is novel about their work and how it builds on existing literature.

⚠️ Shorter, clearer section on contribution of this paper. I’ve currently moved everything to the appendix. We can take from some of the wording here on why it’s novel.

Overall, the manuscript’s readability will greatly improve once the mismatch in framing is corrected. After that, the authors should double-check that all terms are well-defined (for example, define AIC, BIC, any score if introduced, etc., for readers who might not be experts in all three areas). The tone of the paper is appropriately scholarly; the revisions needed are mainly about precise descriptions of what was done and why.

**Recommendations and Required Revisions**

The study addresses an interesting and important topic and, with a revision, has the potential to be published and to impact practice. However, the issues identified must be adequately addressed to ensure the manuscript’s conclusions are valid and clearly communicated. I list below the key revisions that the authors should undertake, based on the previous sections:

• Align scope with content ✅ • Incorporate missing literature and context ✅ • Discuss model selection criteria limitations ✅ • Consider proper scoring rules or predictive evaluation ✅ • Emphasize computational findings ✅

By addressing these, the authors will greatly improve the quality and clarity of their work. I believe the paper has a solid foundation in its comparative approach, and with these revisions, it can meet the high standards of the journal. I encourage the authors to take these critiques constructively and revise accordingly. I look forward to seeing a substantially improved version of this manuscript, as I do think it can make a meaningful contribution once these issues are resolved.