Response letter - Manuscript ID JSR-22-468.R1

Dear Professor Malthouse

Please find enclosed a revised version of our paper submitted for peer review for the *Journal of Service Research*. Your comments have been very much appreciated and we have tried to suitably modify/amend our paper for those comments to be reflected in the revised version.

We quote below the detailed referees' comments that are followed by our response. In all cases, we point out, where necessary, the corresponding amendments in the new version of the paper and highlighted them in teal (greenish-blue color).

Editor comment

My decision on this manuscript was not easy. On the one hand, the reviewers rate the importance of the work as "important" and "very important." They also see "major" potential impact. They acknowledge that you made major changes to the first version of the MS. I personally appreciate you highlighting the changes you made to the manuscript in a blue font and see your substantial improvements to R1. On the other hand the review team feels like many comments from the first round were not addressed. R1 and the AE are asking for model details, clarifications and elaborations. They continue to find typos (R2 notes issues with subscripts and both reviewers mention formatting issues). Simple suggestions like mentioning extended application areas in the introduction, which will help in the positioning of the article for JSR readers, were ignored. They appreciate the example you have added, but suggest adding details. I see these comments as constructive in that their goal is to improve the readability of your manuscript for the JSR and other readers, which will ultimately mean that readers will understand your work so that it realizes its full potential impact. While many of our readers have strong methodological training, JSR is also read by more managerial and policy-oriented scholars; it is important that articles are written in a way to reach both audiences. This expands the reach and impact of your work.

In order to gain final acceptance, I ask for you to go back and address the clarifications, details, etc. Please write a point-by-point response and highlight the changes in the manuscript by using a different colored font. I do not plan to send this out for review again. I reserve the right to reject the work if there is not sufficient progress made in addressing the issues identified by the review team in the next submission.

Response: Thank you for your detailed feedback and thorough review of the manuscript. We appreciate the acknowledgement of the significance of our work as well as the improvements made in response to the initial reviews. We have supplied a point-by-point response to the individual points identified, and the revised version includes the appropriate clarifications, details, and corrections. To indicate the changes made in the revised manuscript, we use a different colored font. We have addressed specific concerns mentioned about model specifics, explanations, and elaborations, as well as identified typos, formatting issues, and ideas for improvement. We also recognize the significance of catering to readers who are methodologically trained as well as managerial or policy-oriented. We presented the material in a way that effectively communicates the importance of our work to a wide range of audiences.

Associate Editor

We received two very good reviews that came up with justified concern and mixed recommendations. Let me add additional comments.

AE comments

Comment 1. The authors need to score higher on the economic insights. Their example did not impress me, and I was hoping to receive more intuition as to why their approach works better.

Response: Thank you for your feedback. In this paper, we introduce a comprehensive framework for probabilistic forecast reconciliation in Emergency Medical Services. Our work not only outlines the rationale behind the proposed approach's suitability for this specific context but also rigorously evaluates its effectiveness through statistical forecast accuracy metrics, as presented in Table 3 and Figure 4. In response to the comment on insights into the economic implications of our approach, we have added the following paragraphs to the paper to further discuss this point:

Our research establishes a strong basis for future investigations and practical implementation in EMS. Leveraging the hierarchical and grouped structure of demand time series, EMS can use this advanced forecasting framework to generate coherent point and probabilistic forecasts, making the most of all available data at every level of the hierarchy. We acknowledge that a forecast serves a greater purpose beyond its mere existence, ideally enabling the best utility in terms of efficient allocation of medical services, response time, and cost, all informed by the forecast. While we fully appreciate the importance of evaluating forecast quality based on its impact on decision-making processes, it is essential to address the data requirements and methodology involved in measuring this impact. For a comprehensive assessment of the forecasts' implications, access to additional data beyond ambulance demand, covering various decision types, capacity information, constraints in the decision system, and more, becomes necessary. This additional data would offer valuable insights into the specific decisions relying on the forecasts, resulting in a more accurate evaluation of their impact on medical services. Furthermore, measuring the actual impact of forecasts would necessitate an approach that goes beyond forecasting itself. This would involve developing and implementing simulation models capable of replicating decision-making processes based on the forecast inputs. These simulation models would then evaluate the quality of the final decisions, taking into consideration the utilities that are particularly significant in the context of EMS.

Comment 2. The authors miss opportunities. I appreciate that the authors summarize previous research in Table 1. Yet, there is nothing in Table 1 that helps the reader understand why the current research has the limitations that the authors state on page 8.

Response: Thank you for your comment. In Table 1, we present three dimensions that highlights the limitations identified in prior research. These dimensions include reconciliation, probabilistic forecasting, and reproducibility. It is noteworthy that there is a notable gap in the existing literature regarding the consideration of probabilistic forecast reconciliation while adhering to the principles of reproducibility. The first column in the table relates to our research, indicating 'YES' in all these dimensions, whereas previous studies have been characterized by a 'NO' in this regard. Table 1 in the previous version had only Reconciliation, and Probabilistic. We now added the reproducibility column to the table, aligning with limitations discussed in the literature section.

Comment 3. I highly appreciate that the authors intend to provide data and code via their GitHub repository. No doubt, the resulting transparency is a major strength of this manuscript. Yet, the authors could still do a much better job in outlining to the readers that the authors have done a good job.

Response: Thank you. We would like to emphasize that our commitment to transparency and reproducibility goes beyond merely sharing data and code. Our entire paper is authored in R using Quarto, setting it apart as a uniquely transparent and reproducible workflow and this could be used as an outstanding example of a transparent and reproducible research in academic communities. In this way, we provide not only the code and data but also the entire narrative of the research process. To further highlight that, we have expanded the the reproducibility section in the revised manuscript:

To enhance transparency and reproducibility, we not only provide data and the code, but also the entire paper that is written in R using [Quarto](https://quarto.org/). All materials to reproduce this paper is available at github.com/bahmanrostamitabar/forecasting-emergency-medicine. The repository contains the raw data, all R scripts used in experiments, the results used in the paper, as well as the quarto files for producing this paper. Full instructions are provided in the repository.

Comment 4.I felt that the authors did not address the following concern that I outlined in the last round of review: "It would also help to have a rather detailed example, maybe even a numerical example, to better describe why separate forecasts at different levels can cause problems. Right now, it sounds plausible that such problems occur, but you are still rather vague concerning the details of these problems."

Response: Thank you. We have now added a more detailed but simple example in the introduction based on the ambulance data and added the following text to the revised version of the paper:

To illustrate the problem, let's consider a very simple example where we have an EMS provider with a national level of governance, and two regions (A and B), each with a health board and a station. There is a total national budget to be split between the regions in proportion to the forecast number of incidents in each region. The two regions have very different incident patterns, and so must be forecast using different models. However, the data are noisy at regional level, so the national forecasts are best obtained by summing the demand from the two regions. The resulting national forecasts are not equal to the sum of the regional forecasts, and so are not coherent. In fact, the national forecasts show a decreasing trend in demand, and so the national governing body decides to cut the budget for the next year. But neither of the regional forecasts shows a trend, and so the regions argue that the budget cut is unfair. In addition, Region A has much more variable demand than Region B, and so to cope with periods of peak demand, Region A needs to hold more resources in reserve. So the budget distribution needs to be made in a way that ensures the probability of each region being unable to meet demand is equal. Our solution to this problem is to use a hierarchical forecasting approach that ensures the forecasts are probabilistically coherent. Then any trends or other forecast characteristics at national level will also be reflected in the regional forecasts, and the probabilistic forecasts allow for the different levels of uncertainty in the two regions. Budget can be allocated by controlling the probability of demand exceeding available resources, rather than being simply in proportion to the expected demand.

Comment 5. I also had the following comments in the previous round of review: "Make sure that all tables and figures are self-contained" and "If possible, display the number of observations in all tables and figures with statistical results". Unfortunately, Figure 1 is not self-contained (e.g., abbreviations are not explained). Figure 2 could also contain the correlation and the number of observations.

Response: Thank you for the comment. We have anonymized the name of the health board using two letters (e.g.) that hold no meaningful association, so these are not abbreviations. We have clarified this in the caption of Figure 1.

With regard to Figure 2, we appreciate the interest in examining correlations. However, we would like to clarify that the purpose of the scatter plot in Figure 2 is not to show the correlation. Each point on the plot corresponds to an individual time series, and the primary objective of this plot is to facilitate a clear understanding of time series features in our dataset such as the strength of trend and seasonality. Introducing correlation values in this specific plot may not add meaningful insights due to the nature of the data representation.

Comment 6. I was also hoping for an example that not only describes the problem that the authors solve, but that allows us to better understand the basic idea behind the authors' solution.

Response: Thank you. We've addressed this comment by including a detailed example in the introduction. This addition is designed to clarify the problem and elaborate on the idea behind the forecast reconciliation. We added the following paragraph to the revised version of the paper:

To illustrate the problem, let's consider a very simple example where we have an EMS provider with a national level of governance, and two regions (A and B), each with a health board and a station. There is a total national budget to be split between the regions in proportion to the forecast number of incidents in each region. The two regions have very different incident patterns, and so must be forecast using different models. However, the data are noisy at regional level, so the national forecasts are best obtained by summing the demand from the two regions. The resulting national forecasts are not equal to the sum of the regional forecasts, and so are not coherent. In fact, the national forecasts show a decreasing trend in demand, and so the national governing body decides to cut the budget for the next year. But neither of the regional forecasts shows a trend, and so the regions argue that the budget cut is unfair. In addition, Region A has much more variable demand than Region B, and so to cope with periods of peak demand, Region A needs to hold more resources in reserve. So the budget distribution needs to be made in a way that ensures the probability of each region being unable to meet demand is equal. Our solution to this problem is to use a hierarchical forecasting approach that ensures the forecasts are probabilistically coherent. Then any trends or other forecast characteristics at national level will also be reflected in the regional forecasts, and the probabilistic forecasts allow for the different levels of uncertainty in the two regions. Budget can be allocated by controlling the probability of demand exceeding available resources, rather than being simply in proportion to the expected demand.

Reviewer 1

Overall Review:

The authors have made revisions to address some of the concerns raised in the initial review, which has improved certain aspects of the manuscript. Additions have been made to explain the trend and seasonality strength measures, provide the GLM equation, and resolve minor formatting issues. These changes are beneficial.

However, critical concerns around model specifics, validation methodology, results analysis, and terminology remain unclear or insufficiently addressed. Major issues persist with detailing the ETS configurations, clarifying serial dependence in TSGLM, justifying the naive method's inclusion, explaining training/testing data splits for tuning, defining the forecast horizon, elaborating the CRPS metric,

providing dataset-specific examples for hierarchical forecasting concepts, distinguishing between hierarchical forecasting approaches, validating the comparison technique, and discussing practical implications of non-integer forecasts.

The lack of clarity and details in these important methodological areas continues to undermine the credibility and reproducibility of the study. Furthermore, the authors do not sufficiently leverage the results to provide managerial insights for EMS planning. More in-depth analysis of the forecast accuracy patterns across timescales, hierarchy levels, and model configurations would strengthen the practical implications.

In summary, while I acknowledge the efforts to improve the manuscript in certain aspects, there remains a critical need for enhancing model specifics, methodology descriptions, results analysis, and terminology usage.

Response: Thank you for the feedback. Wherever necessary, we responded to all comments raised in both the previous and current versions. We are confident that the current study is one of the first to publish a publication using a cutting-edge scientific report, such as Quarto, sticking to transparency and reproducibility principles. This study is completely reproducible and transparent, and the reviewer comments criticizing the research's credibility and reproducibility astound us.

Comment 1. The authors added brief explanations of the trend and seasonality strength measures in response to Comment #3. While this provides more context, a citation or reference for these specific definitions would further strengthen this section.

Response: Thank you. We have included Bandara, Hyndman, and Bergmeir (in press), which describes the approach to decomposing the time series used to compute the strength of trend and seasonality. Additionally, we have included Hyndman and Athanasopoulos (2021), which also describe the strength of trend and seasonality.

Comment 2. The revisions partially address the lack of model detail raised in Comment #4, but some components are still unclear:

Response: Thank you very much. Section 3.2 contains further information on all models. The section contains significantly more details than the first version of the manuscript. We have also made available the Github repository, which contains the whole manuscript written in Quarto, as well as the technical details and R codes for the forecasting methodologies used in this study. We think that the amended section, together with the Github repository, will provide readers with the information they need to understand and replicate the methodologies employed in this study.

Comment 2.1. For ETS, it is now stated that an automated algorithm selects the best configuration, but no information is given on what options are available for trend, seasonality, etc. Listing the choices would help the reader.

Response: Thank you. This has been included in the previosu version of the paper. The papers highlights that ETS models (Hyndman and Athanasopoulos 2021) can combine trend, seasonality, and error components in a time series through various forms that can be additive, multiplicative or mixed. The trend component can be none ("N"), Additive ("A") or damped ("Ad"); the seasonality can be none ("N"), Additive ("A"), or multiplicative ("M"); and the error term can be additive ("A") or multiplicative ("M"). Please refer to "Exponential Smoothing State Space model (ETS)" in section 3.2..

Comment 2.2. The equation for the GLM model is now provided, but the description still does not mention monthly seasonality. This should be explicitly addressed.

Response: Thank you. The GLM model employed in this study takes account of weekly seasonality and annual seasonality. Monthly seasonality in time series data is extremely rare, and it does not exist in the ambulance demand used in this study. There is no reason for occurrences to occur more frequently at certain times of the month than others.

We have highlighted this in the revised paper:

Monthly seasonality in time series data is extremely rare, and it does not exist in the ambulance demand used in this study. There is no reason for occurrences to occur more frequently at certain times of the month than others.

Comment 2.3. The concept of "serial dependence" in TSGLM needs more clarification and specificity to the EMS context.

Response: Thank you. We have clarified the concept of serial dependence and its specificity to the EMS context in the revised manuscript. Serial dependence in the context of EMS forecasting refers to scenarios in which ambulance demand at one point in time correlates with demand at previous points in time. In other words, the time series' previous values are associated with its current values, which can be described in a variety of ways. The TSGLM model, for example, clearly accounts for serial dependence by incorporating lagged ambulance demand numbers into the model. We have included the following paragraph in the revised manuscript to address this comment:

The term serial dependence refers to instances in which the number of incidents on a current day correlates with the number of incidents on previous days.

Comment 2.4. The rationale for including the naive method in the ensemble is still not provided. Its apparent poor performance suggests it should be excluded.

Response: Thank you very much. In response to your prior comment, we re-ran the experiment, removing the naive method from the ensemble. We then produced a graph that shows the outcomes with and without the naive method. We showed that removing the naive from the ensemble worsens overall performance, thus we kept it in the ensemble and included the original plot in the revised publication.

Comment 3. The explanation of the training/validation/testing data split in Comment #5 remains unclear. The authors should explicitly state if tuning is done on the training set only or if a separate validation set is used.

Response: Thank you. We explicitly mentioned that we employed time series cross validation (Hyndman and Athanasopoulos 2021, sec. 5.10) to assess the forecast accuracy. Time series cross-validation involves dividing the data into multiple training and testing sets based on the time series structure. Model development and hyper-parameter tuning is performed using the training data and the errors are assess using the corresponding test set. Each testing set consists of a number of observations equal to the forecast horizon, and the corresponding training set only includes observations that occurred before the test set. This ensures that future observations are not used in constructing the forecast, making it more robust to evaluate forecast accuracy on new data. Therefore, we believe that our methodology is appropriate for time series forecasting and provides a more robust approach for evaluating forecast accuracy on new data. This has been highlighted in the revised paper in section 3.3 :

Model development and hyper-parameter tuning is performed using the training data and the errors are assess using the corresponding test set.

Comment 4. The forecast horizon explanation in Comment #6 is still ambiguous. The time unit (days, weeks, etc.) should be clearly defined.

Response: Thank you very much. We explicitly stated in several sections of the paper, including the introduction and Section 3, that the forecasting horizon ranges from 1 to 84 days. Please see the Introduction, the beginning of section 3, and section 3.4.

Comment 5. Comment #7 regarding the CRPS metric lacks sufficient clarification. Explicitly defining concepts like sharpness and miscalibration would strengthen this section. Additionally, more specifics are needed on the forecast distribution derivation.

Response: We have provided detail explanation of CRPS and cleanly elaborated on the concepts of calibration and sharpness. We have also discussed how probabilsitic forecasts are computed. We have included the following text in the revised version to address this comment:

Calibration refers to the statistical consistency between the distributional forecasts and the observations. It measures how well the predicted probabilities match the actual probabilities. On the other hand, sharpness refers to the concentration of the forecast distributions — a sharp forecast distribution results in narrow prediction intervals, indicating high confidence in the forecast. A model is well-calibrated if the predicted probabilities match the actual probabilities, and it is sharp if it is confident in its predictions. The CRPS rewards sharpness and calibration by assigning lower scores to forecasts with sharper distributions, and to forecasts that are well-calibrated. Thus, it is a metric that combines both sharpness and miscalibration into a single score, making it a useful tool for evaluating the performance of probabilistic forecasts. We chose the CRPS as one of our evaluation metrics because it provides a comprehensive assessment of the forecast quality and allows us to compare the performance of different models.

To generate forecast probability distributions, we did not assume a specific distribution; instead we used a form of bootstrapping, described in Panagiotelis et al. (2023). This involves simulating 1000 future sample paths from each of the model, by bootstrapping the model residuals, taking into account the cross-sectional correlations between the different aggregated and disaggregated series. In this way, we can generate an empirical distribution of forecasts for each model. The ensemble forecast distribution is a simple mixture of these empirical distributions.

Comment 6. Section 4.2 still lacks concrete examples tying the hierarchical forecasting concepts to the EMS dataset, as noted in Comment #8. This example-based explanation needs enhancement.

Response: Thank you. We have now added a more detailed but simple example in the introduction based on the ambulance data and added the following text to the revised version of the paper:

To illustrate the problem, let's consider a very simple example where we have an EMS provider with a national level of governance, and two regions (A and B), each with a health board and a station. There is a total national budget to be split between the regions in proportion to the forecast number of incidents in each region. The two regions have very different incident patterns, and so must be forecast using different models. However, the data are noisy at regional level, so the national forecasts are best obtained by summing the demand from the two regions. The resulting national forecasts are not equal to the sum of the regional forecasts, and so are not coherent. In fact, the national forecasts show a decreasing trend in demand, and so the national governing body decides to cut the budget for the next year. But neither of the regional forecasts shows a trend, and so the regions argue that the budget cut is unfair. In addition, Region A has much more variable demand than Region B, and so to cope with periods of peak demand, Region A needs to hold more resources in reserve. So the budget distribution needs to be made in a way that ensures the probability of each region being unable to meet demand is equal. Our solution to this problem is to use a hierarchical forecasting approach that ensures the forecasts are probabilistically coherent. Then any trends or other forecast characteristics at national level will also be reflected in the regional forecasts, and the probabilistic forecasts allow for the different levels of uncertainty in the two regions. Budget can be allocated by controlling

the probability of demand exceeding available resources, rather than being simply in proportion to the expected demand.

Comment 7. Comment #9 about clarifying the differences between various hierarchical forecasting approaches does not seem to be addressed. This terminology is still confusing in Section 4.2.2 and requires revision.

Response: In the amended version, we have given a detailed overview of various hierarchical forecasting approaches, which are presented below:

The bottom-up (BU) approach is constrained by its reliance solely on base forecasts from a single level of aggregation at the bottom level. While it does result in consistent forecasts, the BU approach lacks forecast reconciliation since no reconciliation is performed.

Forecast reconciliation approaches bridge this gap by combining and reconciling all base forecasts to generate coherent forecasts. This technique utilizes all the base forecasts produced within a hierarchical structure to create consistent forecasts at every level of the hierarchy. As a result, it goes beyond relying solely on base forecasts from a single level of aggregation, and instead leverages all available information at each level to generate forecasts that minimize the total forecast variance of the set of coherent forecasts.

Ordinary Least Squares (OLS) is the simplest and most commonly used method for estimating the parameters in linear regression models. In this approach, the estimation of $\boldsymbol{W}$ is based on the assumption that all the errors have equal variance. Hence, $\boldsymbol{W}$ is simply defined as the identity matrix multiplied by a constant factor. The intuition behind OLS is that it minimizes the sum of squared residuals between the observed and predicted values of the dependent variable. The main weakness of this approach is that it does not take account of the different scales of the base time series; the aggregated series will usually have higher variance than the disaggregated series, simply because the values are larger, but OLS treats all series the same. A strength of the approach is that it is simple, and does not involve estimating a covariance matrix.

Weighted Least Squares (WLS) is an extension of OLS where the variance of the errors is assumed to be heteroscedastic, i.e., different for each series. But it assumes that the errors of each series are uncorrelated with each other. In this approach, $\boldsymbol{W}$ is defined as a diagonal matrix with the variance of the errors on the diagonal. The intuition behind WLS is that it assigns higher weight to series with smaller error variance, and thereby takes into account the different scales of the base time series. The main weakness of this approach is that it ignores the relationships between series. A strength of WLS is that it is relatively easy to compute $\boldsymbol{W}$ as it is based only on error variances which are readily estimated.

Minimum Trace (MinT) is a further generalization where $\boldsymbol{W}$ is defined as the covariance matrix of the base forecast errors. So it takes account of both the scale of each series, and the relationships between the series. Wickramasuriya, Athanasopoulos, and Hyndman (2019) showed that this approach gives the optimal reconciled forecasts in the sense that the sum of the forecast variances is minimized. The main weakness of this approach is that it is difficult to estimate the full covariance matrix. In practice, we usually need to use a shrinkage estimate where the off-diagonal elements are shrunk towards zero.

Comment 8. Comment #10 questioning the validation approach of comparing base and reconciled forecasts is not addressed. The authors should explain why this is an appropriate evaluation technique. (Rob)

Response: Thank you for the comment.

The forecast reconciliation approach does not aggregate lower level forecasts, instead it combines all base forecasts. The base forecasts are generated independently at each level of the hierarchy, ignoring the hierarchical/grouped structure of time series or any aggregation constraints. As

a result, the base forecasts are not guaranteed to be coherent, meaning that the sum of the lower-level forecasts will not necessarily add up to the higher-level forecast. On the other hand, the forecast reconciliation approach takes the base forecasts and regenerates them in a way that ensures coherence across all levels of the hierarchy, while also considering all available information. The comparison between the base forecasts and the reconciled forecasts allows us to see the improvement in forecast accuracy that can be achieved by using the forecast reconciliation approach.

While the approach is designed to improve forecast accuracy, it may not always be successful in doing so, particularly if the bottom-level series are too noisy and lack systematic patterns. In such cases, the benefit of performing reconciliation at the bottom level may still lie in creating coherent forecasts that can help align planning across different teams in an organization, improve coordination, and avoid conflicting decisions. Additionally, even if the bottom-level series are noisy and lack systematic patterns, reconciliation can still lead to more accurate forecasts at higher levels of the hierarchy by utilizing the information available across the hierarchy. Therefore, even if the bottom-level forecasts might not be very accurate on their own, reconciling them with higher-level forecasts can still provide a more consistent view of future demand and possibly more accurate forecasts at other levels.

We have added the following paragraph in the revised version of the manuscript to address this comment:

While the forecast reconciliation approach aims to enhance forecast accuracy, its effectiveness is not guaranteed, especially if the bottom-level series exhibit excessive noise and lack systematic patterns. Despite this, reconciling forecasts at the bottom level can offer advantages by generating coherent forecasts that facilitate alignment in planning across various teams within an organization, promote better coordination, and prevent conflicting decisions. Moreover, even when dealing with noisy and irregular bottom-level series, reconciliation can still improve forecast accuracy at higher levels of the hierarchy by leveraging the information available across the hierarchy. Therefore, although the bottom-level forecasts may not be highly accurate on their own, reconciling them with higher-level forecasts can still provide a more consistent view of future demand and potentially yield more accurate forecasts at other levels.

Comment 9. My concerns in Comment #11 about non-integer forecasts and rounding do not appear to be addressed. This practical issue needs to be discussed.

Response: Thank you for highlighting this point, we have now included a paragraph at the end of the result section to acknowledge this point.

Rounding forecasts up or down and its effect on the forecast accuracy depends on the level of hierarchy and scale of data. For some situations, with high volume demand, forecast accuracy calculations can ignore integer effects as rounding becomes negligible. However, low volume demand settings, such as forecasts at the bottom level of the hierarchy, may be more susceptible to integer (rounding) effects. In practice, we may need to use integer forecasts, especially when the forecasts are relatively small. Count forecast reconciliation is an active area of research, and it would be interesting to explore in future research how our approach could be adapted to generate count reconciled probabilistic forecasts.

Despite using Poisson regression models to create count distributions of attended incidents for the base forecasts, it is important to note that the reconciled forecast distributions do not maintain a count format. In practical scenarios, there might be a need to use integer forecasts. Count forecast reconciliation is an active area of research, and it would be interesting to explore how our approach could be adapted to generate count-reconciled probabilistic forecasts in future studies. Rounding the forecasts is one possible solution to this problem. However, the impact of rounding on forecast accuracy varies depending on the level of hierarchy and the scale of the

data. In situations with high volume demand, the effects of rounding may be negligible, and forecast accuracy calculations can overlook integer effects. On the other hand, in low-volume demand settings, such as forecasts at the bottom level of the hierarchy, integer (rounding) effects may have a more noticeable influence on forecast accuracy.

Reviewer 2

General comment: The author revised the manuscript upon the review comments and I appreciate the detailed introductions added to the body of the paper.I provide some follow-up comments below for consideration and reference.

Response: Thank you for your positive feedback and appreciation of the detailed added to the paper. We have carefully reviewed your follow-up comments and take them into consideration as we continue to refine the manuscript.

Comment 1. Might be good to at least briefly mention the extended applicability of the idea early in the introduction section.

Response: Thank you for the suggestion. We have included the following paragraph in the introduction to address this comment.

While our research focuses on emergency medical services, it is important to emphasize the suggested framework's adaptability, which expands its relevance to a variety of service sectors such as supply chains, tourism, finance, and call centers. Our approach can be generalized in cases with hierarchically structured and/or grouped time series data, which is common in many sectors.

Comment 2. It is confusing to have the abbreviations BME1 and BME2 listed with the same full name while no explanations or references are given. This is at Section 2, Page 7.

Response: Thank you for the feedback. We used the notation mentioned in the literature study. However, since it appears that this could be confusing, we have changed the wording to refer to these models as versions of bivariate mixed-effects models. We included a reference to the study that contains details regarding these variations.

Comment 3. The author added great details to the methods mentioned in Section 3.2, such as the well-explained model specifications of GLM and TSGLM. Make sure typos are corrected, especially notation-wise such as subscripts.

Response: Thank you for your suggestions. The document has been proofread thoroughly. We carefully corrected all typos, including punctuation and those relating to notation and subscripts, to ensure the manuscript's correctness and clarity.

_Comment 4. It may be good to consider an alternatively way of providing details to the linear reconciliation method. For instance, the main equation tilde{y}*h is still missing important reasoning about how it is derived. With some better descriptions of how it is derived, such as an objective function, the follow-up paragraphs for the three methods can be more concisely summarised with better readability.*

Response: Thank you for the comment. We have now reworded this section, and added some more details about the objective function which is optimized. We have included the following text in the section 4.2.2 of the revised manuscript:

When $\boldsymbol{W}_h$ is the covariance matrix of $\hat{\boldsymbol{y}}_h$, the resulting forecasts are optimal in the sense that the sum of the variances of the reconciled forecasts is minimized, provided the base forecasts $\hat{\boldsymbol{y}}_h$

are unbiased. However, $\boldsymbol{W}_h$ is difficult to estimate, and so there have been various suggested approximations to $\boldsymbol{W}_h$, leading to different types of reconciliation such as Ordinary Least Squares (OLS), Weighted Least Squares (WLS) and Minimum Trace (MinT).

Comment 5. The new Section 5.1 provided details about how to obtain and understand the distributional forecasts of the method. Indeed, this can be considered a major selling point of the work, and hence I would always appreciate further addressing and discussions about the probabilistic outperformance of the idea. For example, it might be good to include figures with fitting coverage of the training data, as well as some additional appendix figures for other example series probably at different hierarchy levels. It is not necessary to refer them in the main paper, but some self-explanatory captions should suffice. Besides, comparison with competing methods regarding the same aspect would be a plus.

Response: Thank you for your feedback. Although the distribution is generated for the whole 84-day prediction and all 1530 time series across the hierarchy, visualizing each of these distributions and forecast horizons to fit into the manuscript is not feasible. As a result, we've given a sample example at the total country level that illustrates the probability distribution over a 7-day period. This gives readers a clear understanding that can be accommodated within the constraints of the manuscript. The code is available on our GitHub repository for those interested in investigating additional time series or forecast horizons, allowing readers to reproduce the plot for any chosen time series and forecast length.

We have included the following plots in the revised version of the paper:

Figure 1 depicts the forecast distribution of total incidents in one health board over a 7-day period. It also gives the point forecast as well as the 80% and 90% prediction intervals. Figure 2 zooms in on the first day to show the histogram more clearly, illustrating the range of possible outcomes and their likelihood.

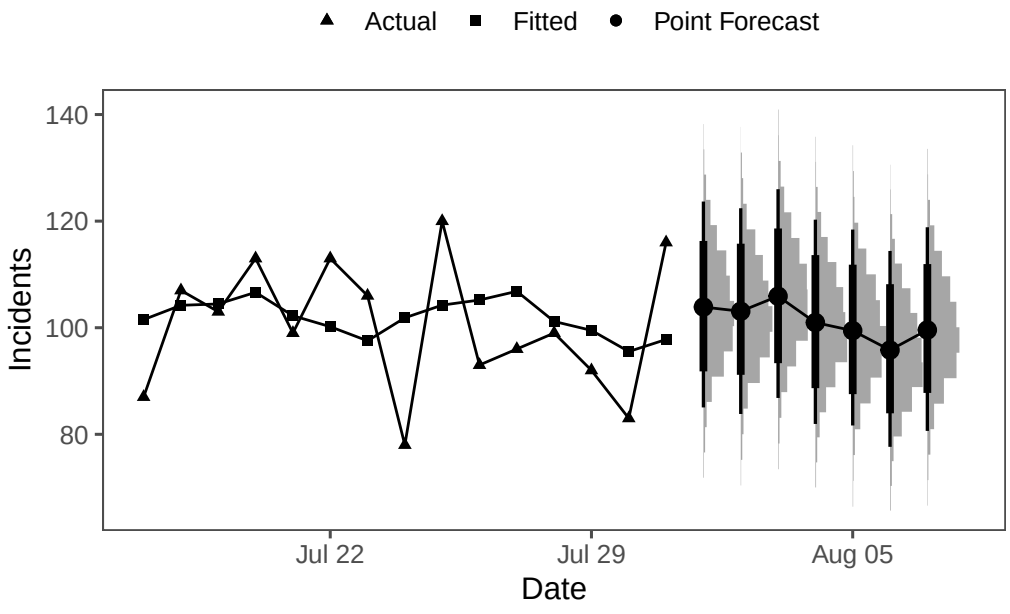

Figure 1: A graphical illustration of the forecast distribution of ambulance demand (i.e. total incidence attended) for the SB health board for a horizon of seven days. For each day, we display the point forecast (black point), the histogram, and 80% (thick line) and 90% (thin line) prediction intervals. It also shows a portion of a historical time series as well as its fitted values.

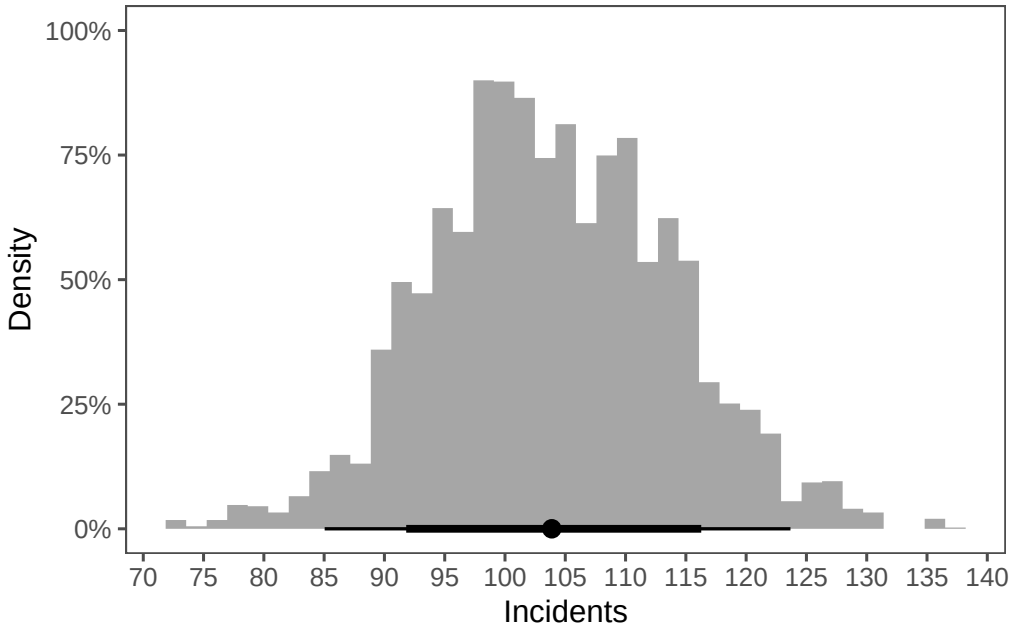

Figure 2: An illustrative example of the forecast distribution of ambulance demand (i.e. total incidence attended) for the SB health board for one day ahead. This corresponds to the first forecast distribution in Figure 5. The horizontal axis shows all possible outcomes that may occur, with their likelihood shown on the vertical axis. The point in the middle shows the point forecast. Two lines at the bottom of the distribution highlights 80% (thick line) and 90% (thin line) prediction intervals.

Comment 6. There are still formatting issues in the included tables.

Response: Thank you very much. We have now reviewed the format of all tables and resolved any issues that occurred. For more information, refer to Tables 1, 2, and 3.

References

Bandara, Kasun, Rob J Hyndman, and Christoph Bergmeir. in press. "MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns." *International J Operational Research*, in press.

Hyndman, Rob J, and George Athanasopoulos. 2021. *Forecasting: Principles and Practice*. 3rd ed. OTexts. https://OTexts.com/fpp3.

Panagiotelis, Anastasios, Puwasala Gamakumara, George Athanasopoulos, and Rob J Hyndman. 2023. "Probabilistic Forecast Reconciliation: Properties, Evaluation and Score Optimisation." *European Journal of Operational Research* 306 (2): 693–706.

Wickramasuriya, Shanika L., George Athanasopoulos, and Rob J. Hyndman. 2019. "Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization." *Journal of the American Statistical Association* 114 (526): 804–19. https://doi.org/10.1080/01621459.2018.1448825.