Response letter

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 blue.

Editor comment

Manuscript ID JSR-22-468 entitled "Hierarchical Time Series Forecasting in Emergency Medical Services" which you submitted to the Journal of Service Research, has been reviewed. I read the manuscript and assigned it to an Associate Editor (AE). The AE selected the reviewers. Three expert reviewers responded with detailed and constructive comments. The AE wrote a summary letter with additional comments. I read all the reports and reread the manuscript. The reviewer and AE reports are located below.

With the exception of R3, the review team finds the research topic to be very important and sees the potential for your manuscript to have substantial impact. I also like the topic and see similar potential. The AE and I think the issues raised by R3 can be addressed in a revision.

R1, R2 and the AE raised many clarifying questions and had concrete suggestions for improve the work. I ask that you give serious consideration to these comments. The current MS is 33 pages, and we allow up to about 50 pages, and so you should have enough space to add the clarifications, possibly as an appendix.

I agree with many of R3's comments. JSR is not a methods journal, and the readers will generally be interested in substantive implications of the work. The evaluation currently focuses on several accuracy metrics such as MACE and MSSE, but what does improved forecasting accuracy mean to business metrics? Perhaps you could do some calculations to demonstrate the importance of your research in terms of business outcomes? Another way to address this is to give more detail about the natural of the decisions that are being made. How specifically can a planner "reconfigure units" and what sorts of options are available in "redeploying staff"? Help the reader understand the managerial decisions being made.

R3 also raises an issue about the generalizability. The AE has some suggestions for broadening the introduction slightly, and there could be a bit more discussion about other possible contexts in which your model might be used. R3 gives two cites that seem to address the problem; I am interested in whether these methods already solve your problem, or whether they would require substantial modification.

I liked your point about reproducibility very much. It is great that you will released you code on Github. Would it be possible to release your data as a benchmark data set for the research community? I understand that there may be NDAs involved, but publicly available data sets help the field advance.

This is minor, but there seem to be issues with your bibliography software. Sometimes it inserts initials in the text citations, other times not. Sometimes commas are missing, e.g., "Haugsbø Hermansen and Mengshoel (2021)". There are other writing errors, e.g., sometimes "neural network" is capitalized and other times not. It is not a proper noun.

Associate Editor

General comments to author: This manuscript looks at how to best forecast ambulance demand. These forecasts are crucial for planning and deploying staff and fleet. They are required at different levels (e.g., national, regional, and sub-regional levels) and are often made independently of each. So, the resulting forecasts at the different levels could be inconsistent. To avoid this problem, the authors exploit the hierarchical and grouped structure of the demand time series and apply several forecasting methods. In their empirical study with daily incident data from an ambulance service in Great Britain from October 2015 to July 2019, an ensemble of forecasting models yields the best forecasts that are also coherent at all levels.

We received three very excellent reviews that came up with mixed recommendations. Still, they share your excitement about the topic. They are mostly consistent in recognizing the manuscript's strengths and weaknesses. All three reviewers read the paper as a "forecasting exercise". Instead, you need to better highlight the substantive insights that your "forecasting exercise" provides. In addition to the recommendations of the reviewers, my recommendation below might help you achieve this aim.

All reviewers have questions and recommendations regarding the details of your time series. They seem to be justified and not in conflict with each other. So, consider implementing them.

All reviewers have questions regarding your look "beyond the point estimate". Consider them. It might help to also come up with an example that outlines the advantage of going beyond point estimates in greater detail.

R1 has a long list of precise and helpful suggestions for improving the manuscript. It looks like they are all justified, and it should be manageable to implement them. The same holds for R2 and R3. R3's remark regarding Figure 4 and the performance of the naïve model that seems to be not much worse is particularly crucial.

Response: Thank you for your feedback and for summarizing the main points from the reviewers. We appreciate your suggestion to highlight the insights provided by the forecasting exercise, and we have worked on the manuscript to make this more explicit in the revised manuscript. We agree that the reviewers have raised important questions and recommendations regarding the details of our study, and we have addressed these in the revised manuscript. Regarding the recommendation to look beyond the point estimate, we have taken the reviewers' suggestions into account and provide an example that illustrates the advantage of doing so. We have implemented R1's suggestions in the revised manuscript. We have also responded to R3's remark regarding the performance of the Naïve model and Figure 4.

AE comments

In addition to the reviewers, I recommend also addressing my following suggestions:

Comment 1. Consider broadening the scope of the manuscript by not focusing instantly on ambulance demand. Probably, other services also share the problem of having different levels. Outline them and explain why they all face the core problem that you address. R3 makes a comparable statement

("generalizability"). Such broadening might also enable you to better link your work to previous research in the Journal of Service Research.

Response: Thank you for the suggestion! We agree that broadening the scope of the study to include other services that face similar hierarchical and grouped time series forecasting issues would make it more relevant and interesting to the Journal of Service Research audience. We have added a new subsection to the manuscript in the conclusion section, where we discuss other services that share the same problem and how our approach could be applied to those services. We also linked our work to previous research in the Journal of Service Research to provide a more comprehensive understanding of our contribution.

Comment 2. 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: We agree that providing a detailed example can better illustrate the problems that arise when separate forecasts at different levels are inconsistent. We have added the following example to the manuscript, which outlines the issue of incoherent in hierarchical forecasts.

> For example, the annual budget for the whole organization is to be allocated to each area health board based on demand within the corresponding area. If the forecasts are incoherent, there is a mismatch between the total resources to be allocated, and the resources allocated to each area. Further, suppose the budget also needs to take account of the nature of incidents that occur within each area, with more money allocated for some types of incidents than others. Then we need forecasts of the demand disaggregated by nature of incidents and health board, but these data are often too noisy to forecast directly. Forecast reconciliation addresses these problems by ensuring the forecasts are coherent across all aggregated and disaggregated series (avoiding mismatches), and by using the signal in the aggregated data to allow forecasting of highly disaggregated data (allowing better targetting of the budget).

Comment 3: You sometimes mentioned that the data runs from 2015-2019 (e.g., in the abstract) and sometimes from 2015-2020 (page 3). Please harmonize.

Response: Thank you for bringing this to our attention. We have gone through the manuscript and made sure that we consistently and correctly report the duration of the dataset as 2015 to 2019.

Comment 4: Consider summarizing previous research in a table. Such a table should also enable you to better highlight your contribution (by also including your work in the table).

Response: Thank you. We agree that a table could provide a clearer overview of the related literature and our contribution. Therefore, we have now included a table in the literature review section summarizing previous research, including our study's contribution.

Comment 5: Make sure that all tables and figures are self-contained.

Response: We have reviewed each table and figure in the manuscript and ensure that they are complete and understandable without relying on the main text and all necessary information such as axis labels, legends, etc are provided.

Comment 6: Table 1: Make it easier for the reader to understand all values. For example, instead of "105", write "105 = 35 x3", etc.

Response: We have updated Table 1 to make it easier for readers to understand all values by providing a breakdown of the calculation.

Comment 7: If possible, display the number of observations in all tables and figures with statistical results.

Response: We agree that displaying the number of observations is important when reporting statistical results. We have added information about the size of the data set where relevant.

Reviewer 1

General comment: *In this study, the authors assess the efficacy of various hierarchical time series forecasting techniques within the domain of emergency medical services. Specifically, they consider the hierarchical and group structure of ambulance demand. Utilizing daily incident data from an ambulance service in Great Britain spanning from October 2015 to July 2019, the authors conduct forecasting and performance evaluations. They employ an ensemble of forecasting models, in which the predictions of multiple models are merged to produce superior results compared to any single forecasting model. The paper is commendably organized, clearly written, and easy to comprehend. However, it lacks sufficient detail on the methodology, as noted in my comments below. I encourage the authors to include more in-depth information to enhance the overall quality of the paper.*

Response: Thank you for taking the time to review our paper. We appreciate your positive comments and understand your concerns regarding the lack of detail on the methodology. We have addressed your comments in the following responses.

Major Comments: *The description of the approaches in section 3 and section 4 are missing important details that would enhance the readers' understanding of the paper.*

Response: Thank you for pointing out that our descriptions in sections 3 and 4 could benefit from more detail. We have revised those sections to include more specific information on the forecasting models and hierarchical and grouped time series techniques utilized in our study.

Comment 1*: The data points used in this time series forecasting pertain to "attended incidents", which refer to situations where a physical resource has been deployed to respond to an incident. However, this definition only accounts for actual responses, and does not consider potential failed responses. I suspect that the response rate will not always be 100% due to various reasons, such as lack of resources, and it is important to take these failed responses into account in forecasting future emergency medical services (EMS) needs. Is data on failed responses available? If not, it would be beneficial to at least acknowledge and discuss this aspect of the forecast in the analysis.*

Response: Thank you for your thoughtful comment regarding the potential impact of failed responses on our time series forecasting. We agree that failed responses could have an impact on forecasts, as the actual demand for ambulance services could be higher than what we have observed in the attended incidents data. Unfortunately, as you noted, the dataset we used only includes information on attended incidents, and we do not have access to information on failed responses. Therefore, we were not able to directly account for the potential impact of failed responses in our analysis. However, we have acknowledged this limitation in our manuscript and discussed its potential impact on our forecasting results. We agree that it would be valuable for future research to investigate the impact of failed responses on EMS forecasting, if data on these incidents becomes available. In the meantime, we hope that our study can provide some insights into the hierarchical time series forecasting techniques that can be used with attended incident data, and how these techniques can potentially improve the accuracy of EMS demand forecasting. It is also important to note that our methodology for hierarchical time series forecasting can be applied to any time series data in EMS, including those that may include failed responses. However, the forecasting results may differ depending on the specific characteristics of the data.

Comment 2. It is important to note the value of this research in the context of emergency medical services (EMS). Both high-level and low-level forecasts are necessary for effective decision-making in EMS. For example, control area EMS forecasts are high-level forecasts that inform strategic decisions about how to allocate limited resources to lower levels, such as health boards and stations/hospitals. These types of decisions are typically made periodically, not on a daily or weekly basis. In this context, the research presented in this paper has clear value. However, it is not clear to me what specific value the research brings at a lower level. It would be beneficial if this aspect could be further explained in the paper.

Response: Thank you for your insightful comment. We appreciate your point about the importance of both high-level and low-level forecasts for effective decision-making in EMS. In our study, we use a methodology that provides forecasts for all levels of a given hierarchy including high-level and low-level of the hierarchy. We believe that wherever decisions are made about the future in the hierarchy, forecasts would be used as an input. While you pointed out values on high-level forecasts, but forecasts are also valuable in at lower levels for shorter-term decision-making and planning. For example, hospitals or ambulance stations could use such forecasts to plan for staffing and resource allocation, ambulance dispatching, staff-to-shift assignment, staff rescheduling based on the anticipated volume and type of incidents. Additionally, generating forecasts at lower levels could potentially improve the accuracy of the high-level forecasts, by providing more detailed information on the nature and priority of incidents. This could help to identify patterns and trends in demand that may not be apparent at the higher level.

Comment 3. Second paragraph of page 9 mentions "the strength of trend" and "the strength of seasonality". Can you explain what those are or how they are calculated?

Response: Thank you for the comment. In this paper, the strength of trend and seasonality were calculated using the "STL" (Seasonal and Trend decomposition using Loess) decomposition method, which is a widely used and flexible method for decomposing time series data into trend, seasonal, and remainder components. The decomposition of a time series y_t is written as $y_t = T_t + S_t + R_t$, where T_t is the smoothed trend component, S_t is the seasonal component and R_t is a remainder component. The strength of trend is defined as:

$$F_T = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t + R_t)}\right)$$

For strongly trended data, the seasonally adjusted data should have much more variation than the remainder component. Therefore $\text{Var}(R_t)/\text{Var}(T_t + R_t)$ should be relatively small. But for data with little or no trend, the two variances should be approximately the same.

The strength of seasonality is defined similarly:

$$F_S = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t + R_t)}\right).$$

series with seasonal strength F_S, close to 0 exhibits almost no seasonality, while a series with strong seasonality will have F_S close to 1 because $\text{Var}(R_t)$ will be much smaller than $\text{Var}(S_t + R_t)$.

Comment 4. One of my main concerns is with regard to section 3.2. The current discussions of the forecasting approaches are inadequate in terms of important details. To improve the understanding of the paper, please provide a more detailed description of each of these approaches.

Response: Thank you for your feedback. We appreciate your concern regarding the level of detail provided in Section 3.2 on the forecasting approaches used in this study. While we agree that it is important to provide adequate information in the paper itself, we also understand that some readers may be interested in more technical details and code implementation. For this

reason, we have provided a Github repository with all technical details and R codes for the forecasting approaches used in this study. However, we acknowledge that more information can be provided in the paper itself to enhance readers' understanding. To address your concern, we have revised Section 3.2 to provide a more detailed description of each of the forecasting approaches used, including the mathematical framework, assumptions, and key parameters. We hope that this revised section, along with the Github repository, will provide readers with the necessary information to understand and replicate the methods used in this study.

Comment 4.1: In this research, the "naive forecasting approach" is used as a benchmark. Is this the forecasting approach currently used in practice? If not, why was it selected as the benchmark? I expect that more rigorous forecasting approaches will outperform the naive approach. It would be more compelling to demonstrate the real-world value of your approach if it were compared against the approach currently used in practice.

Response: Thank you for the suggestion. We appreciate your feedback. We would like to clarify that our intention in this paper is not to provide an exhaustive list of forecasting methods. Instead, we aim to provide a methodology that can incorporate forecasts produced by any method. Our focus is on developing a hierarchical time series forecasting approach that takes into account the hierarchical and group structure of ambulance demand. We have selected the naive method as a benchmark because it is widely used in forecasting literature and practice and is a simple and easily understandable method for users.

Forecasts are often used as input for other decision-making systems that often involve simulation, and it is quite common to use the empirical distribution of the demand as a forecast in simulations. Also, the naive method can be surprisingly accurate. Therefore, if any forecasting method should be able to provide more accurate results than naive, it would justify its use in practice, if not there is no need to use more complicated methods.

Comment 4.2: In your discussion of ETS, you mention that the trend, seasonality, and error terms can take various forms, such as additive, multiplicative, or mixed. Can you specify which form you used for each of these components in your research and explain the reasoning behind your choice?

Response: Thank you for the question. In our research, we used an automatic algorithm to determine the appropriate form for the trend, seasonality, and error terms in each time series. Specifically, we used the `ETS()` function in the fable package of R, which selects the most appropriate model for each time series based on statistical criteria such as Akaike's Information Criterion (AIC). Since we had a large number of time series (1530), it was not practical to manually select the appropriate form for each component for each time series. Instead, the automatic algorithm selects the most appropriate components based on the characteristics of each individual time series. As a result, a mix of different component forms (additive or multiplicative) were used, depending on the specific characteristics of each time series.

Comment 4.3: In your research, the Generalized Linear Model (GLM) approach is used for forecasting. Can you provide the equation used for this approach in a concise form? Additionally, in your description of the Fourier terms, it is mentioned that they are used to capture yearly seasonality. Does your research also take into account monthly seasonality and if so, how is it captured? Furthermore, for yearly seasonality, what is the specific form of the Fourier term used?

Response: Suppose the time series is denoted by $y_1, \ldots, y_T$, then the Poisson GLM can be written as

$$\begin{aligned} y_t &\sim \text{Poisson}(\lambda_t) \\ \text{where} \qquad \log(\lambda_t) &= \boldsymbol{x}_t'\boldsymbol{\beta}, \end{aligned}$$

and $\boldsymbol{x}_t$ is a vector of covariates. In our model, these include cubic splines for the time trend, day-of-week dummy variables, Fourier terms to capture the yearly seasonality, dummy variables

indicating public holidays, school holidays, Christmas Day and New Year's Day. The Fourier terms are as defined in Hyndman & Athanasopoulos (2021, Section 7.4).

This model takes account of weekly seasonality and annual seasonality. Monthly seasonality is exceedingly rare in time series data, and does not occur in ambulance demand. There is no reason, for example, for incidents to occur more at some times of the month than others.

Comment 4.4: Can you explain what "serial dependence" means in the TSGLM model in the context of EMS forecasting?

Response: In the context of EMS forecasting, serial dependence refers to the situations where ambulance demand at one point in time are correlated with the demand at previous points in time. In other words, the past values of the time series influence the current values, and this influence can be modeled using a variety of techniques. The TSGLM model, for example, explicitly accounts for serial dependence by including lagged values of the ambulance demand in the model. This is important in EMS forecasting because it allows the model to capture patterns and trends in the data that are dependent on the history of the time series, and which are not captured via the predictor variables.

Comment 4.5: It is clear that the naive method performed poorly in the results. Can you explain the reasoning behind its inclusion in the ensemble method? Have you also evaluated the performance of an ensemble method that combines forecasts from only ETS, GLM and TSGLM, without the naive method? I would be interested in seeing the results from this alternative ensemble method.

Response: We re-ran the results of our experience with the naive method excluded and …..

Comment 5: In the first paragraph of section 3.3, you stated that the model development and hyper-parameter tuning process is conducted using the training data alone. Is this a commonly employed approach within the literature? From my understanding, when hyper-parameter tuning is involved, a dataset is typically divided into three sets: training, validation, and test sets. The training set is used to develop the model, the validation set is employed for tuning the parameters, and the test set is utilized to report the error measurement metric. Can you provide further clarification on this methodology and how it compares to the standard approach in the literature?

Response: In many machine learning applications, datasets are divided into three sets: training, validation, and test sets, and the validation set is used to tune the hyperparameters of the model. But we use a more efficient approach than this, whereby we employ time series cross-validation (Hyndman & Athanasopoulos, 2021, Section 5.10) to assess the forecast accuracy, and thus the use of a separate "validation dataset" becomes unnecessary. Time series cross-validation involves dividing the data into multiple training and testing sets based on the time series structure. 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.

Comment 6: In Section 3.3, it is not clear what the forecasting horizon j represents. The unit of time, whether it is a day, a week, or multiple weeks, is not specified and this makes it difficult to understand the logic behind the point forecast error e_j . Additionally, it would be helpful to explain why the scale-independent measures such as MASE and MSSE were chosen and how it relates to other commonly used accuracy measures.

Response: To clarify, the forecasting horizon j represents the number of time periods ahead that we are forecasting, which is $j = 1, 2, \ldots, 84$ days in our study. We will make sure to include this information in the revised version of the paper.

In terms of the choice of accuracy measures, we chose MASE and MSSE because they are scale-independent. Using scale-independent measures like MASE and MSSE allows for more meaningful comparisons between time series at different levels and scales, as they are not affected by the magnitude of the data. This is particularly important in our study, as we have time series at various levels of the EMS hierarchy, which may have different scales and therefore different magnitudes of error. By using scale-independent measures, we can compare the accuracy of forecasts across the entire hierarchy in a more meaningful way.

MASE stands for Mean Absolute Scaled Error, and is a normalized measure of the mean absolute error (MAE) of the forecast relative to the naïve forecast, which is a simple forecasting method that uses the last observed value as the forecast for the next time period. Similarly, MSSE stands for Mean Squared Scaled Error and is a normalized measure of the mean squared error (MSE) of the forecast relative to the MSE of the naïve forecast.

Comment 7: In the last paragraph of page 13, the Continuous Ranked Probability Score (CRPS) is mentioned as a metric that rewards sharpness and penalizes miscalibration. Can you provide an explanation of what these terms mean in the context of your problem or dataset? Additionally, in the definition of p_j, the forecasted probability distribution and true probability distribution are referenced. Can you explain how these are derived from the dataset? If a specific distribution is assumed, it would be helpful to be specific about it.

Response: Thank you for your question. We have provided responses to each of the two separate questions below:

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 (EJOR, 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.

The comparison between the forecast distribution and the observations is then used to calculate the CRPS. In this context, the "true probability distribution" has all its weight on the observed value, and zero elsewhere.

We have now updated the explanation of CRPS along the lines described above.

Comment 8: Section 4.2 does not sufficiently relate to the current problem at hand. The definitions presented are given in a standard form without providing context or explaining how they relate to the problem or dataset being analyzed. For example, the term "bottom-level series" is used in both the text and Table 2, but it is not clear how this term applies to the specific dataset being studied. It would be beneficial

to provide a concrete example from the dataset to help readers understand the concept of "bottom-level series" in the context of this problem.

Response: We agree that further explanation and contextualization of the definitions presented in Section 4.2 would be beneficial for readers. In particular, we will provide a concrete example from our dataset to illustrate the concept of "bottom-level series" and how it applies to our problem. We will also revise the text to provide more context and explanation for each definition, to better relate them to the problem at hand.

Comment 9: I am experiencing confusion while reading section 4.2.2. The section discusses a "bottom-up approach" and it sounds like it is not a favorable approach. However, in the following paragraph, you mention "forecast reconciliation approaches." Are these two different approaches or is the "forecast reconciliation approach" a type of "bottom-up approach"? Additionally, you also mention the use of "linear reconciliation method" in this section, which leads me to assume it is a type of "forecast reconciliation approach." Can you provide further clarification on these different approaches and how they relate to each other in this context?

Response: Thank you for your comment on section 4.2.2. We apologize for any confusion that may have arisen from the text. Please allow us to provide some clarification.

While the bottom-up approach does lead to coherent forecasts, it is not a forecast reconcilation approach because no reconciling is done. In the bottom-up approach, forecasts are generated at the bottom (most disaggregate) level and then aggregated up to higher levels in the hierarchy. This approach has limitations because it only uses base forecasts from a single level of aggregation at the bottom level. On the other hand, the linear reconciliation method uses all base forecasts generated within a hierarchical structure to regenerate coherent forecasts at all level of the hierarchy. This means that it does not only use base forecasts from a single level of aggregation, but rather uses all available information at all levels to generate forecasts that minimises the total forecast variance of the set of coherent forecasts. We hope this clarifies the difference between the two approaches and how they relate to each other in the context of the paper.

Comment 10: In section 5, you present the results of your research. From my understanding, the "forecast reconciliation approach" aggregates lower-level forecasts. If this understanding is correct, I am confused as to why you are comparing base forecasts to the aggregation of those base-level forecasts as they appear to be fundamentally different things. Also, in the second paragraph of page 20, you mention that the reconciliation may not improve the forecast accuracy if the bottom series are "noisy" and lack systematic patterns. My question is: in such cases, what is the purpose or benefit of performing reconciliation at the bottom level?

Response:

The forecast reconciliation approach combines base forecasts (not only lower level forecasts). It is important to note that one of the main purposes of comparing base forecasts to the reconciliation of those forecasts is to assess the improvement in forecast accuracy that can be achieved by using the forecast reconciliation approach. 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 in the upper-level series. Therefore, even if the bottom-level forecasts are not 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.

Comment 11: In the second to last sentence of the paper, you state that "the reconciled forecast distributions are not counts." In your evaluation of forecast performance, did you round any of these non-integer values to whole numbers or were they kept as non-integers? For practical implementation, it is necessary to round these non-integers to whole numbers. Have you considered how your forecasting approaches perform if the non-integer forecast results are rounded up or rounded down?

Response: 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 forecast at bottom level of hierarchy, may be more susceptible to integer (rounding) effects. In our study, we kept the non-integer forecast values after reconciliation when evaluating the accuracy. The difference in forecast accuracy results is negiglible. [Perhaps we should check this assertion?] 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 (e.g., https://arxiv.org/abs/2207.09322), and it would be interesting to explore in future research how our approach could be adapted to generate count reconciled probabilistic forecasts.

*Comment 12: In Table 1, the number of parameters for "Priority * Nature of incident" is 104. Why is it not 105? Similar questions for the next two rows.*

Response: The reason for the discrepancy in the number of series is that for some combinations of priority and nature of incidents, there is no representation in the dataset. Therefore, these combinations are not included in the table For example, in the "Priority * Nature of incident" combination, one of the cells in the cross-tabulation have no incidents recorded for a combination of Red Green priority and Falls, so it is not included in the dataset. The same reasoning applies to the other two combinations in Table 1.

Comment 13: Please fix the headings in Table 2 for "Control areas" and "Health boards". Currently they are misaligned. I hope the authors will find those comments helpful in improving the paper.

Response: We have now fixed the headings in Table 2. We appreciate your feedback and we hope that our revised version will meet your expectations.

Reviewer 2

General comment: This paper proposes hierarchical algorithms to forecast the Emergency Medical Services (EMS) demands with both point estimates and distributional estimates. The EMS count data are naturally structured with hierarchies (with spatial levels corresponding to nation, sub-nation and region) and groups (including the priority and incident of the demand). The author starts with independent/base forecast with some classical benchmark time series algorithms such as moving average, GLM and Poisson regression. The hierarchical forecasts are then introduced with the bottom-up structure and optimal reconciliation

approaches. The outperformance of the algorithm is demonstrated by comparing the accuracy between combinations of methods and models under different spatial levels and groups. I appreciate the detailed background introduction and literature review for the application topic. The novel point of view that introduces the hierarchical and grouped forecasts to the EMS management community is promising. I provide some main suggestions, questions and minor comments listed below for reference.

Response: Thank you for your detailed review and positive feedback on our paper. We appreciate your valuable suggestions, questions, and comments, and we will address them accordingly.

Comment 1: Section 3.2 of Forecasting methods covers the independent/base forecasting methods, and Section 3.3 of Performance evaluation is designed for both base and grouped forecasts. I would suggest to look further into the logic structure of Sections 3 and 4. One advice may be to push back the overview of base forecast algorithms (Section 3.2) to a new section before the main contribution of grouped forecasts in Section 4.

Response: We appreciate your suggestion to improve the structure of the paper. We will consider moving the overview of base forecast algorithms to a separate section before Section 4 to provide a clearer delineation between the independent and hierarchical/grouped forecasts. This will help readers better understand the logic structure of the paper and facilitate easier navigation of the content.

Comment 2: In Section 3.2, the author builds the base forecasting models. Comment 2-1: There should be more detail added to ETS and Ensemble method. What exactly do they do? Ensemble method refers to Wang et al., but no specific reference. Adding explanatory details to the used methods and a short discussion on other methods should be sufficient.

Response: We agree that providing more detail on the ETS and Ensemble methods, as well as other methods, would enhance the understanding of our proposed approach. In the revised version, we have added more detailed explanations on these methods, including their mathematical formulations, assumptions, and limitations. We will also provide additional references to the Ensemble method and further explanations on how it is applied in our study.

Comment 2-2: In ETS, there are trend, seasonality and a property about the error terms. Have all combination options been considered? If not, how has it been evaluated?

Response: Since we had a large number of time series (1530), we did not manually evaluate all possible combinations of trend, seasonality, and error terms in ETS for each time series. Instead we used an automatic algorithm to select the most appropriate model for each time series based on statistical criteria such as Akaike's Information Criterion (AIC). The automatic algorithm considered a mix of different component forms, depending on the specific characteristics of each time series. We have added additional details on the automatic algorithm used in the revised version of the paper.

Comment 2-3: GLM and TSGLM are regression approaches, sometimes with exogenous variables like holiday indicators. I would suggest to include explicit model formulas (probably R-formatted like y ~ a + b + c) to list all the input predictors.

Response:

Comment 3: Section 3.3 stakes different measures for performance evaluation. I suggest to give more detailed references opposed to just the paper, i.e., equation numbers.

Response:

Comment 4: The author extracts the trend and weekly seasonality of the data, and demonstrates their strengths in Figure 2. The follow-up analysis and forecast depend on the Poisson assumption most of the time due to the nature of non-negative count time series data.

Response:

Comment 5: Is there any verification conducted to confirm the Poisson patterns in the data? Are over-dispersion and/or under-dispersion going to be potential issues? It may suffice to include statements from relative references.

Response:

Comment 6: ETS seems to have the disadvantage of not producing integer valued forecasts. There are certainly other ways to do forecasting for integer valued time series. What about INAR or INGARCH? On the other hand, tsglm() seems to allow for those options but it is not clear what has been used.

Response:

Comment 7: How to understand the stationarity of the time series? Is it going to impact the analysis if the errors are assumed additive and the heteroscedasticity is present?

Response:

Comment 8: And if the errors are assumed multiplicative in a Poisson sense, what are the possible adjustments to verify or handle the stationarity before forecasting?

Response:

Comment 9: The author introduces the algorithm for estimating the distributional forecasts of the data, and measures the accuracy using CRPS.

Response:

– With the distributional forecasts, how do the confidence intervals behave?

Response:

– As mentioned in P20 L14, the algorithm gives more accurate forecast results concerning the tails of the distribution. How does such outperformance visualized regarding this statement, and what to be read from the forecast distributions for risk management?

Response:

Comment 10: The author briefly explains the model setup and estimation routines. The conciseness of the paper is greatly appreciated. In the meantime, an appendix chapter or a separate file of supplementary material might be expected for brief elaboration on some ideas and technical derivations. For example:

Response:

-how are the strengths of trend and weekly seasonality estimated and scaled to [0, 1]?

Response:

– for the aggregated time series a_t = A b_t, how is the 'aggregation' matrix A determined?

Response:

– for the linear reconciliation method $\tilde{y}_h = S(S'W^{-1}S)^{-1}W^{-1}\hat{y}_h$*, how is it derived?*

Response:

– for the positive definite matrix W, what are the intuitions and strengths for the different solutions (Ordinary Least Squares (OLS), Weighted Least Squares (WLS) and Minimum Trace (MinT))? And how is W defined under these options respectively?

Response:

Comment 11: Minor comments Note that line count includes equations, and excludes (sub)section titles. Minus sign indicates counting from bottom.

Response:

- P13, L1 & L3, it might be good to not indent after the equation if the paragraph does not end. The same concern applies to the other equations in the paper.

Response:

- P14, L3, for the Continuous Rank Probability Score, how is the true probability distribution function F_j (x) determined?

Response:

- P17, L-6, "it is does" → "it does".

Response:

- P18, L8, "Mint" → "MinT".

Response:

In References, try to maintain a consistent format in terms of the inclusion (e.g., P22, L10 & L-4 & L-1) or exclusion of the url links.

Response:

P26, Figure 1(a) overlaps with Figure 1(b) and turns to be cropped at the right side.

Response:

In Table 2, the top row (of column names) is unclear with the line change ("areas" and "boards").

Response:

Reviewer 3

General comments: This paper looks at improving forecasts in emergency medical services by exploiting the hierarchical and grouped nature of the different timeseries that related to different levels of decision making. The paper is generally well written and the dataset is interesting. The abstract and introduction also have several elements that triggered my interest as a reader. Unfortunately, the rest of the paper did not live up to some of my expectations, which I will detail below.

Response:

Comment 1: Method Comment 1-a. The middle and backend of the paper feel more like a forecasting exercise. In many ways this paper reads more as a paper for a forecasting journal then a paper for JSR. Also, in this forecasting exercise many relevant choices, such as how the trend and seasonality enter the model happen in a black box for the reader, whereas the impact of these choices can be substantial, I think. For example, for a bottom level time series an additive seasonality component might be beneficial, but in its role for reconciling the higher level series it might be that a multiplicative term would have been better.

Response:

Comment 1-b. Related to point 1.a I was also wondering if an integrated estimation method would not be beneficial. This might go at the expense of reproducibility, as those models might rely less on built-in functions of R. I want to be careful in making concrete suggestions, but my first intuition would be to consider a Dynamic Hierarchical Linear Model (e.g., Gamerman and Migon 1993; Neelamegham and Chintagunta 2004). The transfer functions in these models could be based on the "summing" or "structural" matrices the authors mention in section 4.2.1. The authors could still keep the two-step approach as alternative to show how it differs from an integrated approach. If differences are small, the latter could still serve as a reproducibility/managerial tool.

Response:

Comment 2: Implications Comment 2-a: The implications of the paper are not clear to me. The authors mention that the performance is better for the ensemble method, but they do not show what the impact of this is. Looking at Figure 4 it does not seem like the ensemble method is performing a lot better than the other non-naïve methods. The paper would benefit from some economic interpretation of the better performances. For example, how much more efficient can allocation of medical services be? This comes back to my point 1.a that currently it feels more like a forecasting exercise, but misses the actual impact on the medical services.

Response:

Comment 2-b: Throughout the paper the authors mention that one of the strengths of their model is that they look beyond point estimates. What I understand is that they use the CRPS statistic for this. However, this is again very mechanic. It would be much more valuable to show when and how actual observations are within the prediction interval, as I can imagine that having both too wide or too narrow confidence intervals could lead to wrong allocation decisions.

Response:

Comment 3: Generalizability Comment 3-a: Another concern I have is the generalizability, especially for the service industry. The grouped and hierarchical nature seems very specific for the type of service of medical emergencies. And although the service is highly relevant and important, I am wondering whether the authors can't broaden the scope somewhat in the introduction and/or implications. If not I would specifically address this in the limitations.

Response:

I wish the authors all the best with their research further

Gamerman, D., & Migon, H. S. (1993). Dynamic hierarchical models. Journal of the Royal Statistical Society: Series B (Methodological), 55(3), 629-642.

Neelamegham, R., & Chintagunta, P. K. (2004). Modeling and forecasting the sales of technology products. Quantitative Marketing and Economics, 2(3), 195.