Response letter - Manuscript ID JSR-22-468

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

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.

Response: Thank you very much for considering our paper for publication. We are sincerely grateful to the reviewers for providing such valuable feedback, and we highly value the comments raised by all reviewers and the Associate Editor. We have devoted significant effort to carefully reflect on the concerns raised and have made every effort to address them in the revised version of the manuscript.

We agree that the reviewers have raised important questions and provided valuable recommendations regarding the finer details of our study, and we have taken all of these into account in the revised manuscript. In response to the suggestion to look beyond the point estimate, we have incorporated an example that highlights the advantages of doing so. Furthermore, we have thoroughly addressed the specific suggestions provided by Reviewers 1 and 2 in the revised manuscript. Additionally, we have considered and responded to Reviewer 3's remarks concerning the performance of the Naïve model and Figure 4. To emphasize the significance of our study, we have included additional paragraphs in the paper that highlight its importance and potential extensions. Specifically, in response to R3's remark, we have discussed how the framework can be extended to incorporate the decision-making process and assess the implications of forecasts on crucial aspects like cost allocation and resource management, which are vital to planners. Finally, we elaborate on how our current study has broader implications and can be generalized to other areas within the service industries. This highlights the wide applicability and potential impact of our research in various contexts.

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).

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 responded and implemented suggestion provided by R1 and R2 in the revised manuscript. We have also responded to R3's remark regarding the performance of the Naïve model and Figure 4. We have also included additional paragraphs in the paper to highlight the importance of the current study and how it can be extended to incorporate the decision making process and evaluate the implication of forecasts on decision impacts and utilities important to planners such as cost, allocating resources, etc. Finally, we added a paragraph to highlight how the current study can be generalized to other areas in service industries.

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 paragraph 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.

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 introduction section in the manuscript, which outlines the issue of incoherent in hierarchical forecasts.

For example, suppose 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: 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: 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. We also included two additional references in the table and added the relevant texts in the literature section.

Table 1: Summary of some studies in forecasting FOR AMBULANCE SERVICES

Reference	Year	Variable	Horizon	Method	Metric	Probabilistic	Reconciliation
Current study	2023	Ambulance demand	84 days	Naïve; Exponential Smoothing State Space (ETS), Poisson regression using Generalized Linear Model (GLM) and tscount (TSGLM), a simple empirical distribution and an ensemble method	MASE, MSSE, CRPS	YES	YES
Al-Azzani et al.	2021	Call volume	7 , 30 , 90 days	ARIMA, Holt Winters, Multiple Regression, and Singular Spectrum Analysis	RMSE, MAPE	NO	NO
Haugsbø et al.	2021	Ambulance demand in Spatio-temporal	1hour	MLP, LSTM	MAE, MSE, Cross-Entropy	Yes	NO
Grekousis et al.	2019	Locations of incidents	1 hour	MLP and Hungarian algorithm	RMSE	NO	NO
Ibrahim et al.	2016	Call volume	1, 7 , 14 days	multiplicative univariate forecasting, univariate mixed-effects, bivariate mixed-effects model, and bivariate mixed-effects	RMSE, prediction interval coverage	Partial	NO
Vile et al.	2012	Ambulance demand	7, 14, 21, 28 day	Singular Spectrum Analysis, ARIMA, Holt-Winters	RMSE	No	NO
Sasaki et al.	2010	Ambulance demand	5 years	OLS regression	R^2	NO	NO

Al-Azzani, Davari, and England (2021) utilises real data from the Welsh Ambulance Service to explore the forecast accuracy of four forecasting approaches: ARIMA, Holt Winters, Multiple Regression, and Singular Spectrum Analysis (SSA) in predicting call volume demand accurately. The aim is to compare these approaches with the current method across various planning horizons (7 days, 30 days, and 90 days) for both total call volume and category-specific demand. Forecast accuracy performance is evaluated using root mean square error (RMSE) and mean absolute percentage error (MAPE) are calculated. The findings indicate that ARIMA performs the best in predicting weekly and monthly demand. However, when it comes to long-term demand, the SSA method proves to be the most effective. Ibrahim et al. (2016) conducted a case study to assess the effectiveness of multiple forecasting methods: the multiplicative univariate forecasting model (MU), univariate mixed-effects model (ME), bivariate mixed-effects model (BME1), and bivariate mixed-effects model (BME2). Real-life call centre data was utilised to forecast for periods of 1, 7, and 14 days ahead, using only a limited dataset of 42 days. The performance of these forecasting methods was evaluated using two metrics: RMSE for point forecasts and coverage probability for the 95% prediction interval. The findings indicate that the ME consistently produces the most accurate point forecasts. On the other hand, BME1 and BME2 demonstrate superior coverage probabilities when forecasting for one day or one week ahead. For a two-week leading period, MU shows better coverage probability.

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. Please refer to notes or captions in each figure and table in the revised manuscript.

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, where feasible. Please see the footnote added to Table 1:

Due to certain combinations of the nature of incident with other variables, there is a lack of representation in the dataset. As a result, for example, instead of the calculation 3 * 35 = 105, it would be modified to 3 * 35-1 = 104.

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 could be important when reporting results. We have added information about the size of the data set where relevant. Please see captions od figures and tables for reference.

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: We have now revised sections 3 and 4 to include more specific information on the forecasting models and hierarchical and grouped time series techniques utilized in our study. Please refer to section 3 and 4 in the revised manuscript where we highlighted changes in teal.

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: 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. We have added the following text in the conclusion section of the revised manuscript:

The dataset used in this study only includes information on attended incidents. It would be valuable for future research to investigate the impact of failed responses on EMS forecasting, if data on these incidents becomes available. 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.

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: 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 for a given daily temporal period. The methodology could be used for any temporal interval such as sub-daily, weekly, monthly or quarterly. 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.

We have included the following text in the introduction section of the revised manuscript:

Forecasts produced at both higher and lower levels of hierarchy are necessary for effective decision-making in EMS. For example, control area EMS forecasts can inform strategic decisions about how to allocate limited resources to lower levels, such as health boards and stations/hospitals. At the lower levels, 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 priority and nature 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 in demand that may not be apparent at the higher level. Therefore, employing forecasting techniques that consider the hierarchical and/or grouped patterns of time series in EMS aligns naturally, offering the possibility to enhance forecast accuracy and facilitate coordination.

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. We have added the following text to the revised manuscript to address this comment:

In this paper, the strength of trend and seasonality were calculated using the "STL" (Seasonal and Trend decomposition using Loess) decomposition method, as described by Kang, Hyndman, and Smith-Miles (2017). STL 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: To address your concern, we have revised Section 3.2 to provide a more detailed description of each of the forecasting approaches used. 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 containing the entire paper written in Quarto with all technical details and R codes for the forecasting approaches used in this study. 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.

Please see our response to comments 4.1, 4.2 and 4.3 below, where we included further description of the methods and added the related texts to the revised manuscript.

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: 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 data. We have selected the naive method as a benchmark because it is widely used in the 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.

We have chosen the naive method as a benchmark due to its widespread usage and simplicity, making it easily understandable for users. Forecasts serve as inputs for various decision-making systems that frequently employ simulations, wherein it is common to utilize the empirical distribution of demand as a forecast. Additionally, the naive method has shown surprisingly high accuracy. Hence, any forecasting approach that can offer superior results compared to the naive method would validate its practical use, otherwise there is no necessity for employing more complex 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: In response to your comment, we have included the following text in the revised version of the manuscript:

In our study, we use an automated algorithm to determine the suitable form for the trend, seasonality, and error terms in each time series. Specifically, we utilize the 'ETS()' function in the fable package of R, which employs statistical criteria like Akaike's Information Criterion (AIC) to identify the optimal model for each time series. Given the large number of time series we work with (1530), it is impractical to manually select the appropriate form for each component in every time series. Consequently, the automated algorithm selects the most fitting components based on the unique characteristics of each individual time series. As a result, a combination of additive or multiplicative forms for the components are employed, depending on the specific attributes 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: Thank you for your comment. We have included the following text in revised manuscript to address your comment:

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, β is a vector of coefficients, and λ_t is the parameter of Poisson distribution. In our model, these include cubic splines for the time trend, day-of-week dummy variables (from Monday to Sunday), Fourier terms to capture the yearly seasonality, dummy variables indicating public holidays (1 when is public holiday, 0 otherwise), school holidays (1 when is school holiday, 0 otherwise) and Christmas Day (1 when is Christmas Day, 0 otherwise) and New Year's Day (1 when is New Year's Day, 0 otherwise). The Fourier terms are as defined in Hyndman and 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.

We have now included the following text in revised manuscript:

The Poisson TSGLM is similar to GLM with an aditional component accounting for serial dependence (i.e. autoregressive terms). The term serial dependence refers to situations where the ambulance demand at a given time exhibits correlation with the demand at previous time intervals. Put simply, past values of the time series influence the current values. TSGLM can be written as

$$\begin{aligned} y_t &\sim \text{Poisson}(\lambda_t) \\ \text{where} \qquad \log(\lambda_t) &= (\boldsymbol{y}_{t-k}' + \boldsymbol{x}_t') \boldsymbol{\beta}, \end{aligned}$$

y_{t-k} is a vector of k lagged values. The TSGLM model explicitly accounts for serial dependence by including lagged values (i.e. past values) of the ambulance demand in the model. This is important in EMS forecasting because it allows the model to capture patterns in the data that are dependent on the past values of the time series, which might not be captured via the predictor variables.

We use the 'tsglm()' function in the tscount package in R (Liboschik, Fokianos, and Fried 2017) to model the attended incidents. Again, the logarithmic link function is used to ensure that the parameter of Poisson distribution is always positive.

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 experiment with the naive method excluded and name it "Ensemble 2". We illustrated the results in the following plot. As the plot shows, excluding the naive from the ensemble deteriorates the performance overall, so we keep it in the ensemble and show the original plot in the revised manuscript.

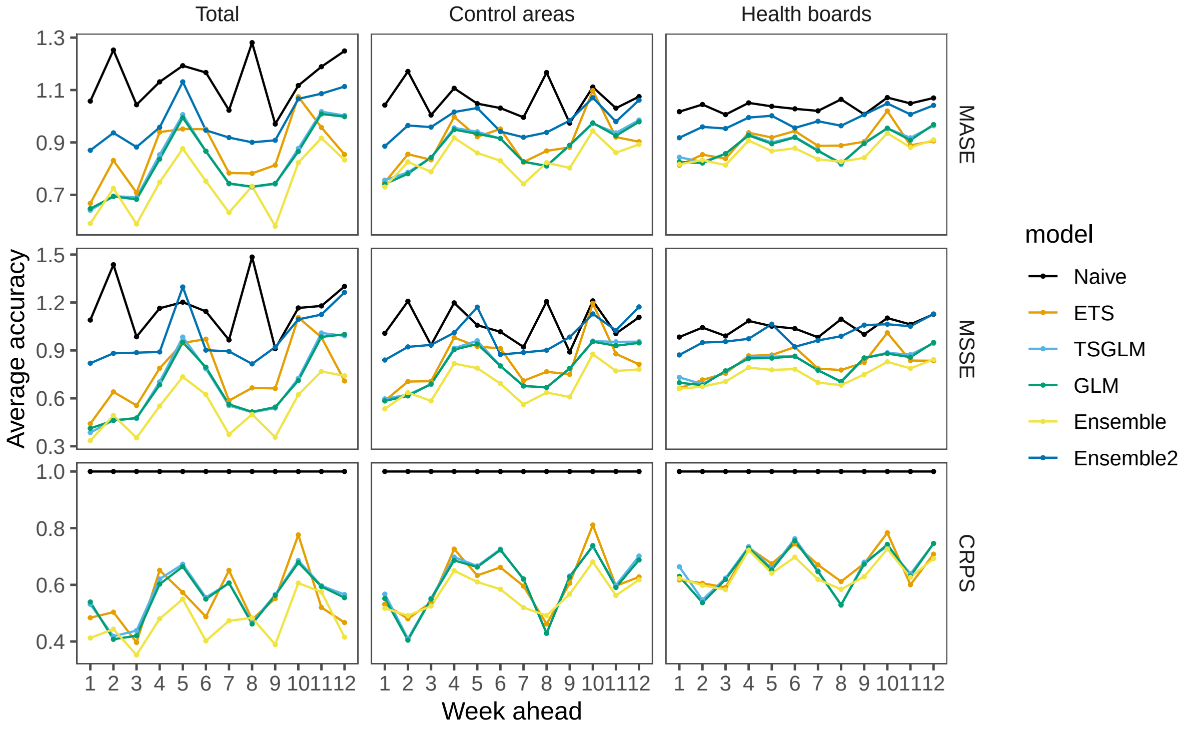


Figure 1: **?(caption)**

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 machine learning applications, datasets might be 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 and Athanasopoulos 2021, sec. 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.

The error metrics provided below consider a forecasting horizon denoted by j, representing the number of time periods ahead we are predicting. In our study, this forecasting horizon ranges from 1 to 84 days, $j = 1, 2, \ldots, 84$.

Using of scale-independent measures, such as MASE and MSSE, enables more significant comparisons between time series at different levels and scales, as these measures are not influenced by the data's magnitude. This aspect holds particular importance in our study, as we work with time series at various levels of hierarchy. These time series show varying scales, resulting in different magnitudes of error. By employing scale-independent measures, we can meaningfully assess the forecast accuracy across the entire hierarchy, ensuring a more robust comparison.

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: 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. (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.

We have added the following text to section 3.2 in the revised version of the paper:

To generate forecast probability distributions using above methods, we use 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.

Calibration refers to the statistical consistency between the distributional forecasts and the observations. It measures how well the predicted probabilities match the actual. 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, 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.

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 of the definitions presented in Section 4.2 would be beneficial for readers. In particular, we will provide an example from our dataset to illustrate the concept of "bottom-level series" and how it applies to our problem. We have ioncluded the following text in the revised version of the manuscript to address this comment:.

The term "level bottom series" relates to the most disaggregated series within the hierarchical and grouped time series structure. For instance, in Table 2, each distinct combination of values in Control area (e.g. South & East), Health board (e.g. CV), Priority (e.g. Green), and Nature of incident (e.g. Chest pain) corresponds to one individual time series. In the dataset at hand, there are 691 unique combinations, resulting in 691 bottom level time series. The "aggregate time series" describes how these bottom-level series are combined to create higher-level series. For instance, to obtain the incidents at the national level (i.e. all country level), the time series are aggregated across all Control areas, Health boards, Priorities, and Natures of incidents. Any desired aggregation level can be achieved based on the data structure, utilizing the bottom level series available.

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: 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. There are also different approaches in the linear reconciliation method, we have now updated section 4.4 and included the following paragraphs to address your cocnerns:

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 W is based on the assumption that all the errors have equal variance. Hence, 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, 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 W as it is based only on error variances which are readily estimated.

Minimum Trace (MinT) is a further generalization where 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 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: 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 may not always be 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 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: 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. One possible approach to address this is by rounding the forecasts. 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 become 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.

*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 has no incidents recorded for the 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. we have added a note to Table 1 to explain this further. Please refer to Table 1 in the revised paper.

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: Thank you for your valuable suggestion to enhance the paper's structure. We have thoroughly considered your proposal of moving the base forecast to a new section. However, after careful consideration, we believe that the current structure maintains a smooth and easily understandable flow. As it stands, Section 3 provides a detailed discussion of the forecasting models, while Section 4 focuses on using the models described in Section 3 to generate forecasts for the hierarchical structure, including both base and reconciled forecasts. We hope this arrangement aligns with your expectations and requirements.

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: In the revised version, we have added more detailed explanations on ETS, Ensemble as well as Naive, GLM and TSGLM. Please refer to section 3.2 to see all details:

We added the following text in the revised version of the manuscript:

In our study, we use an automated algorithm to determine the suitable form for the trend, seasonality, and error terms in each time series. Specifically, we utilize the 'ETS()' function in the fable package of R, which employs statistical criteria like Akaike's Information Criterion (AIC) to identify the optimal model for each time series. Given the large number of time series we work with (1530), it is impractical to manually select the appropriate form for each component in every time series. Consequently, the automated algorithm selects the most fitting components based on the unique characteristics of each individual time series. As a result, a combination of additive or multiplicative forms for the components are employed, depending on the specific attributes of each time series.

Finally, one effective strategy for improving forecast accuracy includes the simultaneous application of multiple forecasting methods on a given time series, followed by averaging the forecasts, rather than relying on separate forecasts generated by each individual method (Clemen 1989). In this paper, we use an ensemble method that combines the forecasts generated from the Naive, ETS, GLM and TSGLM models using a simple average to form a mixture distribution (Wang et al. 2022).

For the rest of the methods, we added:

We have chosen the naive method as a benchmark due to its widespread usage and simplicity, making it easily understandable for users. Forecasts serve as inputs for various decision-making systems that frequently employ simulations, wherein it is common to utilize the empirical distribution of demand as a forecast. Additionally, the naive method has shown surprisingly high accuracy. Hence, any forecasting approach that can offer superior results compared to the naive method would validate its practical use, otherwise there is no necessity for employing more complex methods.

Suppose the time series is denoted by $y_1, \dots, y_T$, 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, β is a vector of coefficients, and λ_t is the parameter of Poisson distribution. In our model, these include cubic splines for the time trend, day-of-week dummy variables (from Monday to Sunday), Fourier terms to capture the yearly seasonality, dummy variables indicating public holidays (1 when is public holiday, 0 otherwise), school holidays (1 when is school holiday, 0 otherwise) and Christmas Day (1 when is Christmas Day, 0 otherwise) and New Year's Day (1 when is New Year's Day, 0 otherwise). The Fourier terms are as defined in Hyndman and 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.

The Poisson TSGLM is similar to GLM with an additional component accounting for serial correlation (i.e. autoregressive). The term serial dependence refers to situations where the ambulance demand at a given time point exhibits correlation with the demand at previous time points. Put simply, past values of the time series influence the current values, and various techniques can be employed to model this influence. TSGLM can be written as

$$\begin{aligned} & y_t \sim \text{Poisson}(\lambda_t) \\ \text{where} \qquad & \log(\lambda_t) = (\boldsymbol{y}_{t-k}' + \boldsymbol{x}_t') \boldsymbol{\beta}, \end{aligned}$$

y_{t-k} is a vector of k lagged values. The TSGLM model explicitly accounts for serial dependence by including lagged values (i.e. past values) of the ambulance demand in the model. This is important in EMS forecasting because it allows the model to capture patterns in the data that are dependent on the past values of the time series, which might not be captured via the predictor variables.

We use the ‘tsglm()‘ function in the ‘tscount‘ package in R (Liboschik, Fokianos, and Fried 2017) to model the attended incidents. Again, the logarithmic link function is used to ensure that the parameter of Poisson distribution is always positive.

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.

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

In our study, we use an automated algorithm to determine the suitable form for the trend, seasonality, and error terms in each time series. Specifically, we utilize the ‘ETS()‘ function in the fable package of R, which employs statistical criteria like Akaike’s Information Criterion (AIC) to identify the optimal model for each time series. Given the large number of time series we work with (1530), it is impractical to manually select the appropriate form for each component in every time series. Consequently, the automated algorithm selects the most fitting components based on the unique characteristics of each individual time series. As a result, a combination of additive or multiplicative forms for the components are employed, depending on the specific attributes of each time series.

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: We have revised the manuscript to include the model formulas, as you suggested for both GLM and TSGLM. We have also included the entire publication in Quarto, including R codes, in a GitHub repository for further details:

Suppose the time series is denoted by $y_1, \dots, y_T$, 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, β is a vector of coefficients, and λ_t is the parameter of Poisson distribution. In our model, these include cubic splines for the time trend, day-of-week dummy variables (from Monday to Sunday), Fourier terms to capture the yearly seasonality, dummy variables indicating public holidays (1 when is public holiday, 0 otherwise), school holidays (1 when is school holiday, 0 otherwise) and Christmas Day (1 when is Christmas Day, 0 otherwise) and New Year’s Day (1 when is New Year’s Day, 0 otherwise). The Fourier terms are as defined in Hyndman and 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.

The Poisson TSGLM is similar to GLM with an additional component accounting for serial correlation (i.e. autoregressive). The term serial dependence refers to situations where the ambulance demand at a given time point exhibits correlation with the demand at previous time points. Put simply, past values of the time series influence the current values, and various techniques can be employed to model this influence. TSGLM can be written as

$$\begin{aligned} y_t &\sim \text{Poisson}(\lambda_t) \\ \text{where} \qquad \log(\lambda_t) &= (\boldsymbol{y}'_{t-k} + \boldsymbol{x}'_t)\boldsymbol{\beta}, \end{aligned}$$

y_{t-k} is a vector of k lagged values. The TSGLM model explicitly accounts for serial dependence by including lagged values (i.e. past values) of the ambulance demand in the model. This is important in EMS forecasting because it allows the model to capture patterns in the data that are dependent on the past values of the time series, which might not be captured via the predictor variables.

We use the 'tsglm()' function in the 'tscount' package in R (Liboschik, Fokianos, and Fried 2017) to model the attended incidents. Again, the logarithmic link function is used to ensure that the parameter of Poisson distribution is always positive.

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: We have revised the manuscript to include a more detailed reference for each equation used in the performance evaluation. Specifically, we have included the paper citation along with the equation number, so that readers can easily locate the relevant equation and read more about its derivation and interpretation. Please refer to section 3.3 to see changes, also highlighted below:

The error metrics provided below consider a forecasting horizon denoted by j, representing the number of time periods ahead we are predicting. In our study, this forecasting horizon ranges from 1 to 84 days, $j = 1, 2, \ldots, 84$.

Using of scale-independent measures, such as MASE and MSSE, enables more significant comparisons between time series at different levels and scales, as these measures are not influenced by the data's magnitude. This aspect holds particular importance in our study, as we work with time series at various levels of hierarchy. These time series show varying scales, resulting in different magnitudes of error. By employing scale-independent measures, we can meaningfully assess the forecast accuracy across the entire hierarchy, ensuring a more robust comparison.

Calibration refers to the statistical consistency between the distributional forecasts and the observations. It measures how well the predicted probabilities match the actual. 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, 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.

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. 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: Thank you for your comment. Provided accidents occur independently, they will follow a Poisson distribution by definition (Feller (1991), p156-158). So it is reasonable to assume a Poisson distribution here. We allow for the mean of the Poisson distribution to change over time by including trend and seasonality covariates, and public holiday effects. If there are other factors that affect the mean of the Poisson distribution, but which aren't in our model, we may observe over-dispersion or under-dispersion. We have developed a flexible framework that can incorporate any forecasting model, and it is important to note that the intention of this paper is not to provide a comprehensive list of models, or to argue for a particular model class. Instead, our focus is on demonstrating the usefulness of our framework for incorporating base forecasts generated by any model and producing hierarchical coherent forecasts.

To address this comment, we have added the following to section 3 before ensemble method:

Provided accidents occur independently, they will inherently follow a Poisson distribution (Feller 1991, 81:p156–158). Hence, it is reasonable to assume a Poisson distribution in this context. To account for changes over time, we incorporate trend and seasonality covariates, as well as public holiday effects, allowing the mean of the Poisson distribution to vary. However, it is important to note that if there are additional factors influencing the mean of the Poisson distribution, which are not accounted for in our model, we might observe over-dispersion or under-dispersion in the data.

We have also added the following paragraph to the end of section 3:

It is important to emphasize that the aim of this study is not to provide an exhaustive compilation of forecasting models or promote a particular model class. Instead, we have developed a flexible framework that can accommodate any forecasting models. Our primary objective is to demonstrate its practicality and effectiveness in integrating base forecasts from any model and generating coherent forecasts within a hierarchical structure.

Comment 5: 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: Thank you for the comment. We used ETS models because they are relatively simple, easy to understand, widely used, and well-established in the time series forecasting literature. They are particularly useful for modeling time series with multiple components, such as trend, seasonality, and/or cycles, and they are flexible and can be adapted to time series data with different patterns. This makes them appropriate for this study given the various patterns in the 1530 time series we analyzed.

We acknowledge that ETS models do not necessarily produce integer-valued forecasts, which may be a disadvantage for certain applications. However, we note that rounding the forecasts can be a practical solution to address this issue. The impact of rounding on forecast accuracy depends on the level of hierarchy and the scale of the data. For high volume demand situations, rounding effects may be negligible, and forecast accuracy calculations can ignore integer effects. However, low volume demand settings, such as forecasting at the bottom level of the hierarchy, may be more susceptible to rounding effects.

Furthermore, we did consider other models specifically designed for integer-valued time series, such as TSGLM and GLM. TSGLM is also capable of accounting for autoregressive lags (i.e. previous observations) and the number of lags to beconsidered can be determined by user.

We acknowledge that there are other methods available, such as INAR or INGARCH models. However, neither INAR or INGARCH are able to handle the non-stationary trends and seasonality, or public holiday effects, that are evident in the data. Further, we note that the intention of this study was not to provide a comprehensive list of forecasting models, but rather to develop a framework that can incorporate any forecasting model.

Comment 6: 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: Thank you for your comment. All of the time series in our data set are non-stationary, due to the changing rate of accidents with the time-of-year, day-of-week, and other factors. All our models, other than Naive, allow for these non-stationary features of the data. None of our models assume additive errors or homoscedasticity. The ETS models allow for multiplicative errors and heteroscedasticity where appropriate. The GLM and TSGLM models also do not have additive errors and have implicity heteroscedasticity due to the Poisson distribution.

Comment 7: 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: Thank you for your comment. First of all, we apologize if our understanding of the question is not entirely clear. Despite thoroughly discussing it, we acknowledge that we may not have grasped it completely. Building on the response to the previous comment, we want to emphasize that the data used in this study exhibit non-stationarity. Nevertheless, the models employed possess the capability to handle non-stationarity through various methods, as explained in our response to the previous concern.

Comment 8: The author introduces the algorithm for estimating the distributional forecasts of the data, and measures the accuracy using CRPS. With the distributional forecasts, how do the confidence intervals behave?

Response: Thank you for the comment.

The performance of the distributional forecasts can be evaluated using the CRPS, which is a scoring rule that accounts for the shape and width of the predictive distribution. 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. A specific prediction interval could be evaluated using a Winkler score (Winkler 1972). CRPS can be considered an average of all possible Winkler scores (Hyndman and Athanasopoulos 2021, sec. 5.9), and thus provides an evaluation of all possible prediction intervals.

We have clarified this further in the revised manuscript:

CRPS can be considered an average of all possible Winkler scores (Winkler 1972; Hyndman and Athanasopoulos 2021, sec. 5.9) or percentile scores (Hyndman and Athanasopoulos 2021, sec. 5.9), and thus provides an evaluation of all possible prediction intervals or quantiles. A specific prediction interval could be evaluated using a Winkler score. Certain situations may also require assessing accuracy for a particular quantile, such as lower (e.g 5%) or higher (e.g. 95%) quantiles. In such cases, a percentile score becomes useful in meeting this specific requirement.

Comment 9: 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: Thank you for this valuable question. While the central part of this distribution may indicate the most probable scenario, it might not include the entire spectrum of potential outcomes that could significantly impact EMS decision-making. Thus, gaining a deeper understanding of the distribution's tails can offer valuable insights to managers, enabling them to better prepare for and manage risks effectively. Ensuring the accuracy of these tail forecasts becomes crucial in emergency planning and response, as it plays a pivotal role in anticipating and addressing critical situations. In our research, we use CRPS (Continuous Ranked Probability Score) as a metric to assess the accuracy of probabilistic forecasts. CRPS provides an average measure across all quantiles, including both low and high quantiles. However, in some cases, there might be a need to calculate the accuracy for a specific quantile only like 5% or 95%. In such situations, a percentile score can be used to fulfill this requirement.

As a response to your comment, we have incorporated an extra subsection (section 5.1) that includes a visualization representing the forecast distribution for a specific time series of incidents. We have added the following new section to the revised manuscript:

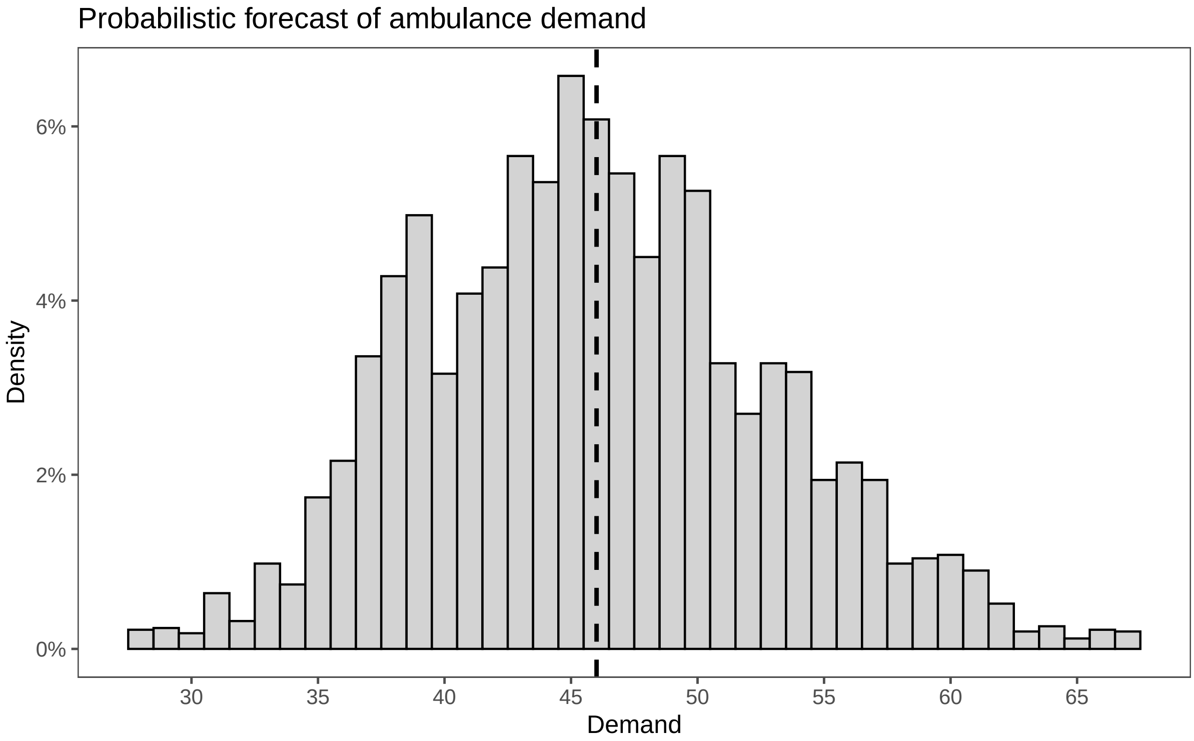


Figure 2: An example of forecast distribution of ambulance demand (i.e. total incidence attended) in PO health board for 1 day ahead. x-axis shows all possible outcomes that may occur, with thier likelihhod in y-aixs

Figure 2 provides an illustrative example of a probabilistic forecast for future demand, one day ahead, regarding the total attended incidents in the PO health board. The black dashed line in the graph represents the point forecast, which presents a single numerical estimate for the future number of attended incidents. Due to the complexity of including such plots for the entire hierarchy and 84 days ahead, only one example is presented here for 1 day ahead. However, it is feasible to generate these plots for the entire hierarchy and for any forecast horizon if necessary.

In practice, point forecasts are commonly used, but they have limitations as they ignore the uncertainty associated with the forecast, leading to wrong decisions. In contrast, probabilistic forecasts offer an alternative approach to anticipate future demand. Rather than providing a single value, they assign likelihoods to all possible demand outcomes, acknowledging that different numbers of attended incidents are possible, but with varying likelihoods.

The purpose of probabilistic forecasting, as demonstrated in Figure 2, is to quantify uncertainty. Decisions based on these forecasts could focus on the tails of the distribution: unexpected high demand leading to crowding and inefficiency, or unexpectedly low demand resulting in wasted resources. Such forecasts are valuable tools for decision-makers and planners, especially when dealing with low-probability, high-cost situations. Different EMS managements may have varying risk attitudes depending on resource availability, making it crucial to consider the entire distribution when making decisions. For instance, these forecasts enable management to calculate the probability of demand exceeding a certain threshold of available resources (e.g., 90%), which can serve as an informative early warning measure for crowdedness.

It is important to note that while point forecasts and prediction intervals can be obtained from the probabilistic forecasts, the reverse is not possible. A single number cannot be used to directly derive a probabilistic forecast. Prediction intervals, although helpful in indicating possible ranges, do not provide information on the probabilities of low or high demand.

In EMS planning, future demand is just one aspect to consider. Other inputs, such as capacity, should also be treated as probability distributions to adopt a probabilistic approach to planning. To extract valuable insights and make informed decisions from probabilistic forecasts, specialized numerical tools are required, as the forecasts themselves are typically represented as explicit probability density functions or Monte Carlo generators.

We have also included the following paragraph in section 3.3:

CRPS can be considered an average of all possible Winkler scores (Winkler 1972; Hyndman and Athanasopoulos 2021, sec. 5.9) or percentile scores (Hyndman and Athanasopoulos 2021, sec. 5.9), and thus provides an evaluation of all possible prediction intervals or quantiles. A specific prediction interval could be evaluated using a Winkler score. Certain situations may also require assessing accuracy for a particular quantile, such as lower (e.g 5%) or higher (e.g. 95%) quantiles. In such cases, a percentile score becomes useful in meeting this specific requirement.

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: We appreciate your suggestion of adding supplementary material to provide more technical details and elaboration on some ideas. To respond to your queries, we have included the GitHub repository which includes all files use to produce the paper including Quarto files, data and R codes used to generate forecasts, evaluate accuracy and produce plots and tables. Additionally, we have included further explanation in the revisions of the paper highlighted them in teal.

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

Response: We have added the following text to the revised manuscript to discuss how the strengths of trend and weekly seasonality are estimated :

In this paper, the strength of trend and seasonality were calculated using the "STL" (Seasonal and Trend decomposition using Loess) decomposition method, as described by Kang et al (2017). STL 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:

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

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 Var(R_t) will be much smaller than Var($S_t + R_t$).

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

Response: We have included the following paragraph in section 4.3 to address this comment:

The aggregation matrix A is determined by the structure of the hierarchy. It maps the bottom-level time series to the corresponding higher-level time series. The matrix A is typically predefined based on the structure of the hierarchy, and it remains fixed throughout the forecasting process. For example, if there are two bottom-level series, and one aggregated series (equal to the sum of the two bottom-level series), then $A = [11]$.

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

Response: We have added the following text to section 4.4 to further clarify this:

Linear reconciliation involves projecting the base forecasts onto the coherent space. It is derived by minimizing the sum of the variances of the reconciled forecasts subject to the resulting forecasts being coherent and unbiased (Wickramasuriya, Athanasopoulos, and Hyndman 2019).

– 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: Thank you for asking. We have clarified different solutions (Ordinary Least Squares (OLS), Weighted Least Squares (WLS) and Minimum Trace (MinT)) and how W is defined. We have added the following paragraphs to section 4.4 in the revised manuscript:

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 W is based on the assumption that all the errors have equal variance. Hence, 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, 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 W as it is based only on error variances which are readily estimated.

Minimum Trace (MinT) is a further generalization where 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 11: Minor comments Note that line count includes equations, and excludes (sub)section titles. Minus sign indicates counting from bottom.

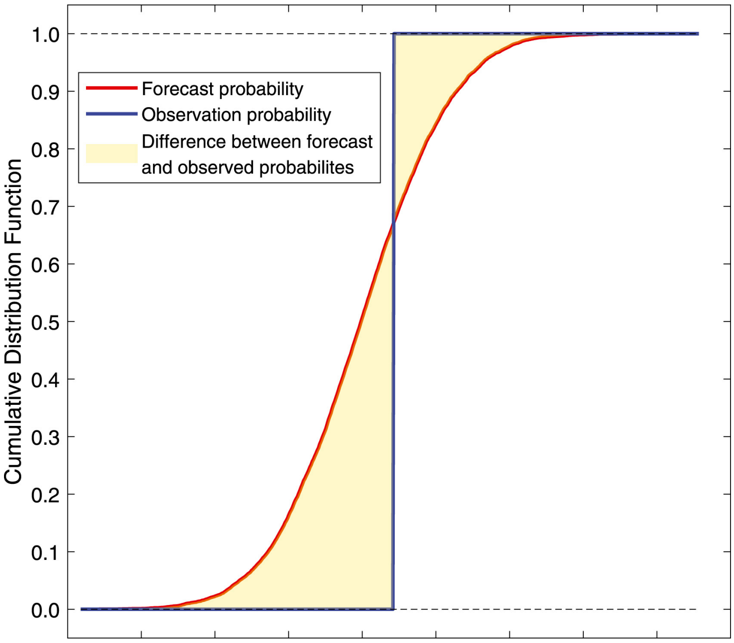
Response: Thank you. We believe that line counts are automatically inserted by the system following the paper's submission.

- 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: We have removed the indentation after equations in paragraphs that do not end.

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

Response: CRPS measures the integrated squared difference between the forecasted cumulative distribution function (CDF), $G_j(x)$$, and the true CDF of the forecast variable, $F_j(x)$. We estimate the latter using the empirical distribution function. Thus, once observed, $\hat{F}_j(x)$ is the empirical distribution function taking value 0 before $x = y_t$ and value 1 after $x = y_t$, as shown in the figure below. The CRPS measures the difference between the forecasted probability distribution of the demand and the actual probability distribution of the demand. The yellow area in the plot of the forecasted CDF and the actual CDF highlights this difference.



- P17, L-6, "it is does" → "it does".

Response: This is now corrected.

- P18, L8, "Mint" → "MinT".

Response: The typo has been fixed now.

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: Thank you. We have removed URL links from the references. Throughout, we ensured that a consistent format was maintained in the References section. Also, we note that the entire citation and reference management is automatic and handled by Quarto and the format of the journal.

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

Response: The issue has been fixed now.

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

Response: We have reviewed Table 2, revised the table to make the column names more clear and consistent with Table 1.

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: We are glad to hear that the abstract and introduction piqued your interest as a reader and that you found the dataset engaging. However, we are sorry to learn that other aspects of the paper did not meet your expectations. Your feedback is valuable to us, and we have carefully considered it in our revision to address your concerns. Thank you for bringing these points to our attention.

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: We completely understand your concerns and acknowledge that transparency in our modeling process is crucial, and we would like to provide further information on how different forms of the trend and seasonality enter the model. For the exponential smoothing model (ETS), using an automatic algorithm, specifically the `ETS()` function in the fable package of R, to select the most appropriate model for each time series based on statistical criteria such as Akaike's Information Criterion (AIC). This approach allowed us to select the most appropriate components based on the individual characteristics of each time series, which was necessary given the large number of time series (1530) in our dataset. As a result, a mix of different component forms (additive or multiplicative) were used, depending on the specific characteristics of each time series at different level of the hierarchy. We hope that this information provides a better understanding of our approach to modeling.

We have included the following paragraph to further clarify this in the revised manuscript:

In our study, we use an automated algorithm to determine the suitable form for the trend, seasonality, and error terms in each time series. Specifically, we utilize the 'ETS()' function in the fable package of R, which employs statistical criteria like Akaike's Information Criterion (AIC) to identify the optimal model for each time series. Given the large number of time series we work with (1530), it is impractical to manually select the appropriate form for each component in every time series. Consequently, the automated algorithm selects the most fitting components based on the unique characteristics of each individual time series. As a result, a combination of additive or multiplicative forms for the components are employed, depending on the specific attributes of each time series.

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: Thank you for your comment. Indeed, it is possible to define a state space model that ensures coherent forecasts. This was proposed by Pennings and Van Dalen (2017), and Villegas and Pedregal (2018). However, in these models, the covariance matrices are extremely difficult to estimate with anything other than very small hierarchies. The hierarchy used in our study is extremely big with 1530 time series. The strength of our approach is that it allows for any type of forecasting method to be used for the base forecasts, including different methods for different series. It decouples the time series models from the reconciliation step, thus allowing much more flexibility in the types of models that can be considered.

We have added the following paragraph in the revised version of the paper for further clarification:

Certainly, other approaches can be applied to hierarchical forecasting problems. Pennings and Van Dalen (2017) and Villegas and Pedregal (2018) proposed the idea of using a state space model to ensure consistent forecasts. However, when dealing with larger hierarchies, these models encounter difficulties in estimating covariance matrices. In contrast, our approach provides a clear advantage by allowing the incorporation of different forecasting methods for the base forecasts, and even accommodating distinct methods for individual series. The decoupling of time series models from the reconciliation step adds significant flexibility in exploring a wide range of models.

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: We appreciate your comment and agree that providing an economic interpretation of the improved performances would greatly enhance the significance of our findings. We acknowledge that a forecast serves a purpose beyond its own existence and ideally it should enable the best utility such as the efficiency of allocating medical services, response time, and cost, informed by the forecast.

While we recognize the importance of evaluating the quality of a forecast by considering its impact on decision-making processes, it is crucial to address the data requirements and methodology involved in measuring this impact. To comprehensively assess the implications of the forecasts, it would be necessary to have access to further data beyond ambulance demand across the hierarchy, such as different decision types, capacity data, constraint in the decision system, etc. This additional data would provide insights into the specific decisions that rely on the forecasts, allowing for a more accurate evaluation of the impact on medical services. Unfortunately, obtaining such data is not feasible and accessible within the scope of our current study. Moreover, measuring the impact of forecasts would require a different approach that extends beyond the forecasting itself. It would involve developing and implementing simulation models that can simulate the decision-making processes based on the forecast inputs. These models would evaluate the quality of the final decision, considering the utilities that are of particular importance to the EMS context.

While we were unable to measure the specific economic impacts and utilities of forecasts largely due to data limitations in the current study, we believe that our research lays a strong foundation for future research investigations, as well as practical implementation in EMS. The proposed forecasting framework and forecast reconciliation methods offer a robust framework that can be extended and integrated with operational information to evaluate decision quality and forecast utility in EMS. EMS can use this advanced forecasting framework that exploit the hierarchical and grouped structure of demand time series, and apply forecast reconciliation methods to generate coherent point and probabilistic forecasts that use all available data at all levels of disaggregation.

Moreover, future research can build upon this foundation by incorporating additional data and implementing simulation modeling techniques. This would enable a more comprehensive evaluation of the economic impact, efficiency gains, and utility of the forecasted results in the context of EMS. In future investigations, we aim to explore this avenue by incorporating operational information, simulating decision processes, and assessing the impact on utilities that are significant to the EMS. To that end, we have started presenting this study to different ambulance services and we hope to establish a collaboration that allow us to access data and decisions made across the hierarchy to evaluate its impact.

We have added the following paragraphs to the conclusion section to further discuss this crucial point and address the comment:

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.

Future research can build upon this study by incorporating additional data and implementing simulation modeling techniques. In future investigations, we aim to explore this avenue by incorporating operational information, simulating decision processes, and assessing the impact on utilities that are significant to the EMS. Linking forecasts with its utilities (response time, resource utilization, etc) can offer an opportunity to maximize benefits through a more holistic planning approach.

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: In our study, we chose the Continuous Ranked Probability Score (CRPS) as one of our evaluation metrics because it provides a comprehensive assessment of forecast quality and allows for the comparison of different models. The CRPS strikes a balance between calibration and sharpness of forecast distributions. Calibration refers to the statistical consistency between the forecasted distributions and the actual observations. It measures how well the predicted probabilities align with the observed probabilities. On the other hand, sharpness refers to the concentration of the forecast distributions — a sharp forecast distribution results in narrow prediction intervals, indicating a high level of confidence in the forecast. A well-calibrated model matches the predicted probabilities to the actual probabilities, while a sharp model 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.

In the context of emergency medical services, it is crucial to consider the implications of forecasts that are not well-calibrated or not sharp. If a forecast is not well-calibrated, it means that the predicted probabilities do not accurately reflect the actual probabilities of events, leading to potential biases in decision-making processes. For example, if the forecast consistently underestimates the demand for ambulance services, it could result in inadequate resource allocation and delayed responses to emergencies. Conversely, if the forecast consistently overestimates the demand, it may lead to unnecessary allocation of resources and increased costs. On the other hand, the sharpness of a forecast is an indicator of the level of confidence in the predictions. When a forecast is not sharp, it suggests a lack of confidence in the estimated outcomes. This can have implications for decision-making processes within emergency medical services. If the intervals are excessively wide, decision-makers may adopt overly conservative strategies, leading to underutilization of resources and potentially delayed responses. Conversely, if the intervals are excessively narrow, decision-makers may take overly aggressive actions, risking resource shortages and potential errors in resource allocation. Therefore, it is crucial to assess both calibration and sharpness of forecast to ensure effective decision-making within emergency medical services.

We have added the following paragraph to section 3.3 to furthe clarify this:

CRPS can be considered an average of all possible Winkler scores (Winkler 1972; Hyndman and Athanasopoulos 2021, sec. 5.9) or percentile scores (Hyndman and Athanasopoulos 2021, sec. 5.9), and thus provides an evaluation of all possible prediction intervals or quantiles. A specific prediction interval could be evaluated using a Winkler score. Certain situations may also require assessing accuracy for a particular quantile, such as lower (e.g 5%) or higher (e.g. 95%) quantiles. In such cases, a percentile score becomes useful in meeting this specific requirement.

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: Thank you for raising the concern about the generalizability of our proposed framework in the context of the service industry. We appreciate your feedback and have addressed it in the revised paper.

While our study primarily focuses on the domain of emergency medical services, we would like to emphasize that the framework we propose is applicable to a wide range of industries, including the service industry. Wherever there are collections of related time series structured hierarchically and/or in a grouped manner, our framework can be utilized to provide accurate forecasts. The grouped and hierarchical structure of time series data is a common characteristic across many sectors, and it arises when data can be naturally organized into various levels of aggregation or when there are dependencies and relationships among different entities within the system.

For example, in the supply chain industry, accurately forecasting demand at different levels of the distribution network, such as regional warehouses or retail stores, is essential for efficient inventory management and minimizing stockouts. Our framework enables the reconciliation of forecasts to ensure consistency and alignment across the supply chain, leading to improved decision-making and operational efficiency.

In the financial industry, where investments span multiple asset classes, geographical regions, or customer segments, our framework can be applied to forecast portfolio performance, asset allocation, or customer demand. By considering the hierarchical structure of the data, financial institutions can make informed investment decisions, manage risk effectively, and meet the specific demands of different customer segments.

In the transportation sector, our framework can support forecasting transportation demand at various levels, such as across different states or regions, to optimize route planning and resource allocation. Similarly, in the hospitality and tourism industry, our framework can be utilized to forecast demand rates at state, regional, and department levels, facilitating strategic pricing, capacity planning, and revenue management for hotels and other travel-related businesses.

Furthermore, in call centers, accurately forecasting call volumes at different levels of the call center hierarchy is crucial for workforce management and resource allocation. By applying our framework, call centers can generate accurate forecasts for different skill groups, shifts, and locations, enabling efficient staffing and optimal service levels to meet customer demands.

In summary, the strength of our proposed framework lies in its ability to handle and reconcile time series data structured hierarchically and/or in a grouped manner. This characteristic can be found across a multitude of service industries, making our framework applicable in various contexts. By considering the relationships and dependencies within the data, our approach can contribute to more accurate forecasts, more coordinated activities due to consistent forecasts and potentially improved decision-making.

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.

We have added the following paragraph to the end of the conclusion section to discuss how our approach is generalisable across different service industries:

Although our study primarily focuses on Emergency Medical Services, it is essential to emphasize that the framework we propose has broad applicability across various service industries (Ostrom et al. 2010). Our approach is particularly valuable in situations where time series data is structured hierarchically and/or grouped, a common characteristic found in many sectors. This occurs when data can be naturally organized into different levels of hierarchies or when dependencies and relationships exist among entities within the system. For instance, in the supply chains (Shugan and Xie 2000), demand forecasting at different levels of the distribution network, such as regional warehouses or retail stores, is vital for efficient inventory management and minimizing stockouts. Our framework allows the reconciliation of forecasts, ensuring consistency and alignment throughout the supply chain, leading to improved decision-making and operational efficiency. In the financial industry (Kimes and Chase 1998) , where investments span multiple asset classes, geographical regions, or customer segments, our framework can be applied to forecast portfolio performance, asset allocation, or customer demand. Similarly, in transportation, the framework supports forecasting transportation demand at various levels, optimizing route planning and resource allocation. Likewise, in the hospitality and tourism industry (Dekimpe, Peers, and Heerde 2016), it facilitates forecasting demand rates at state, regional, and department levels, enabling strategic pricing, capacity planning, and revenue management for hotels and other travel-related businesses. Additionally, in call centers, accurate call volume forecasting at different levels of the call center hierarchy or grouped structure is crucial for workforce management and resource allocation. Implementing our framework, call centers can generate accurate forecasts for different skill groups, shifts, and locations, ensuring efficient staffing and optimal service levels to meet customer demands.

References

Al-Azzani, Mohamed AK, Soheil Davari, and Tracey Jane England. 2021. “An Empirical Investigation of Forecasting Methods for Ambulance Calls-a Case Study.” *Health Systems* 10 (4): 268–85.

Clemen, Robert T. 1989. “Combining Forecasts: A Review and Annotated Bibliography.” *International Journal of Forecasting* 5 (4): 559–83.

Dekimpe, Marnik G, Yuri Peers, and Harald J van Heerde. 2016. “The Impact of the Business Cycle on Service Providers: Insights from International Tourism.” *Journal of Service Research* 19 (1): 22–38.

Feller, William. 1991. *An Introduction to Probability Theory and Its Applications, Volume 2*. Vol. 81. John Wiley & Sons.

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

Ibrahim, Rouba, Han Ye, Pierre L’Ecuyer, and Haipeng Shen. 2016. “Modeling and Forecasting Call Center Arrivals: A Literature Survey and a Case Study.” *International Journal of Forecasting* 32 (3): 865–74.

Kang, Yanfei, Rob J Hyndman, and Kate Smith-Miles. 2017. “Visualising Forecasting Algorithm Performance Using Time Series Instance Spaces.” *International Journal of Forecasting* 33 (2): 345–58.

Kimes, Sheryl E, and Richard B Chase. 1998. “The Strategic Levers of Yield Management.” *Journal of Service Research* 1 (2): 156–66.

Liboschik, Tobias, Konstantinos Fokianos, and Roland Fried. 2017. “Tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models.” *Journal of Statistical Software* 82 (5): 1–51. https://doi.org/10.18637/jss.v082.i05.

Ostrom, Amy L, Mary Jo Bitner, Stephen W Brown, Kevin A Burkhard, Michael Goul, Vicki Smith-Daniels, Haluk Demirkan, and Elliot Rabinovich. 2010. “Moving Forward and Making a Difference: Research Priorities for the Science of Service.” *Journal of Service Research* 13 (1): 4–36.

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.

Pennings, Clint LP, and Jan Van Dalen. 2017. “Integrated Hierarchical Forecasting.” *European Journal of Operational Research* 263 (2): 412–18.

Shugan, Steven M, and Jinhong Xie. 2000. “Advance Pricing of Services and Other Implications of Separating Purchase and Consumption.” *Journal of Service Research* 2 (3): 227–39.

Villegas, Marco A, and Diego J Pedregal. 2018. “Supply Chain Decision Support Systems Based on a Novel Hierarchical Forecasting Approach.” *Decision Support Systems* 114: 29–36.

Wang, Xiaoqian, Rob J Hyndman, Feng Li, and Yanfei Kang. 2022. “Forecast Combinations: An over 50-Year Review.” *International Journal of Forecasting*. robjhyndman.com/publications/combinations/.

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.

Winkler, Robert L. 1972. “A Decision-Theoretic Approach to Interval Estimation.” *Journal of the American Statistical Association* 67 (337): 187–91.