Editor

It has been decided that your paper should be reconsidered for publication in the EUROPEAN JOURNAL OF OPERATIONAL RESEARCH after a major revision taking into account the enclosed referees' comments. (You may wish to argue that some comments are invalid). Please also ensure that you have cited recent and relevant publications in EJOR and other OR journals.

We thank all reviewers for their invaluable/positive comments. All responses are in Blue font, any additions/changes to the paper are marked in Red. To the best of our knowledge, we have now added all recent and relevant publications in EJOR and other OR journals. This is particularly aided by the requests of the Reviewers.

Reviewer #1:

The manuscript provides a novel approach to probabilistic reconciliation using score optimization. It is concise, clearly structured and has a strong motivation, pointing out the necessary gaps in the literature. Following a detailed derivation of the SGD algorithm, the reconciled projections are evaluated using simulated and empirical data against a number of benchmark models. As the paper is very comprehensive, provides code as well as documentation, and appears quite mature, I have only a few minor comments.

- In the simulations, what are the assumptions on the covariance matrix in the joint Gaussian base forecasts?

  In Section 6.2 we previously stated that we use *"the variance covariance matrix of the residuals"*. We now modify the statement to make sure that this is clear to now read: *"the variance covariance matrix of the residuals of the fitted models"*. In principle different choices could be used depending on assumptions made about how the covariance process should be modelled (e.g. time varying correlation), however such an issue lies beyond the scope of the simulation.

  Why is it that the evaluated methods produce better scores for jointly distributed prediction errors?

  In general, all methods score better for jointly distributed errors. This is due to the fact that assuming independence is likely to represent a (more severe) misspecification. The difference is substantial in the empirical application, especially with assuming Gaussian independent errors, as shown in Figure 8. We highlight this with the correlation heatmap we present in Figure 7. We also show some departures from normality in Figure 6. We comment on this stating *"Therefore, independent Gaussian probabilistic forecasts are likely to represent severe misspecification"* just before Figure 6.

In the simulations, all DGPs assume a joint error process and hence we expect that there is less misspecification error when probabilistic forecasts are generated using a joint distribution rather than assuming independence. As these differences are less pronounced we opt not to comment on these here and distract the reader from the main feature which is the performance evaluation across the reconciliation approaches.

- How well does the model scale to large, complex (grouped) hierarchies? How large is the computational demand compared to benchmark methods?

  Relative to OLS and MinT, Score Optimisation is slower taking 2-3 minutes, mainly since it requires finding base forecasts over a rolling window (see line 2 of Algorithm 1). We note that implementing methods such as OLS and MinT for probabilistic rather than point forecasting using Theorem 3.5, is in itself novel and a significant contribution of the paper.

  In an operational setting, where forecasts are made every period and kept, then these forecasts will already be available. The computational cost can also be mitigated by exploiting parallelisation (in our case we parallelised over different simulation/empirical scenarios rather than in the SGD itself). The stochastic gradient descent itself converges very quickly in most cases in less than 20 iterations.

- The authors might expand the discussion of the findings, in particular in section 7.2. How come the OLS approach performs so well given that it has been shown to lead to mediocre results in point forecasts?

  A crucial point to make here is that OLS/MinT applied in the point forecasting setting are different and will thus have different properties to OLS/MinT applied in the probabilistic forecasting setting. In particular, for point forecasting, MinT will minimise the expected total mean squared error, although it should be noted that OLS also has some desirable properties for a loss function based on the L2 norm (see Panagiotelis et al 2021 for details). In the probabilistic setting we use scoring rules to evaluate forecast quality and these previous results do not necessarily apply. We now add the following statement to the paper (see Page 26, Lines 43-50):

  “We suggest two possible reasons for the good performance of OLS in the probabilistic case. First, the energy score depends on the L2 norm of the difference between realizations and draws from the probabilistic forecast, which is similar to the setting for which OLS has optimal properties for point forecasts (see Panagiotelis et al, 2021). Second, for OLS there is less estimation uncertainty as fewer parameters need to be estimated.”

- In the electricity example, how is the joint distribution of the base forecasts estimated if they are generated using independent neural networks?

For both generating joint Gaussian and joint bootstrap probabilistic forecasts we use the residual covariance matrix of the fitted neural networks (fitted using the NNETAR function in the fable package, references are provided in the paper we do not repeat them here). This is identical to how the joint base forecasts are generated from the models in the simulations in Section 6. We have now added this information on Page 24, lines 40-42, by stating:

"Four situations were considered where base forecasts are assumed to be either Gaussian or bootstrapped from residuals, and either independent or dependent (we use the residual covariance matrix of the fitted neural networks, in a similar fashion as in Section 7)."

- A matter of preference, but the figures might look better when using a lighter theme than the out-of-the-box ggplot theme.

  We acknowledge that this is a matter of taste. Having tried a few lighter themes our preference is to use the default ggplot theme.

Reviewer #2:

We thank the reviewer for the positive and constructive feedback on our paper. We respond below in detail to the comments that require a response.

Summary and General points

1. The paper is relevant to EJOR and continues a lineage of hierarchical forecasting papers recently published in the journal

   No response required. Thank you for your positive view/comment.

2. The objective of the paper is to extend hierarchical forecasting methodology to the probabilistic setting.

   No response required.

3. The paper is a difficult read due to the level of mathematical knowledge assumed. Whilst it is undoubtedly mathematically sophisticated, a good deal more effort could be made to explain the reasoning behind the linear algebra and borel sets. This sets a high bar for the empirical results - which in terms of the score optimisation algorithm in particular provide nothing by way of a reward for struggling through to the end. I'm sure the authors mathematical credentials are impeccable, but this is of little benefit to the journal if the reader is strongly inclined to give up reading on page 2.

   We have now added a new section, after the introduction but before the main discussion of theoretical results (See Section 2 Outline of Main Results). The purpose of this section is to give a more intuitive and non-technical description of the main theoretical results. We believe this new section provides the reader less interested in technical details sufficient information to skip the relevant sections.

Something that we now also emphasise in this new section (and in the introduction) is that the OLS and MinT methods are used in a novel way in this paper that relies on the way the theoretical results extend point reconciliation to probabilistic reconciliation. As a consequence, the assertion that the score optimisation algorithm offers little reward on the basis that it is outperformed by OLS and MinT in some settings, does not fairly represent the contributions of the paper. “OLS” and “MinT” as used here are not the existing point forecasting techniques, but probabilistic extensions of these techniques. These are introduced in this paper and are made possible by our definition of probabilistic coherence.

4. The paper makes significant and novel contributions in two main areas:
   a. Firstly, in formally establishing how scoring rules can and should be applied in multivariate hierarchical contexts.

      No response required.

   b. Secondly the paper posits a valuable and far-reaching result that sampling from underlying base forecast distributions, and reconciling the samples delivers valid probabilistic forecasts.

      No response required.

5. These two contributions are important and worthwhile, and the paper should focus on explaining and exploring them in more detail, and making them accessible to the forecasting and OR community.

   Both of these results are highlighted in the new section summarising the key theoretical results of the paper.

6. Additionally, the authors present a HF algorithm which seeks to produce probabilistic forecasts by optimising a scoring rule. This part of the paper is much less successful:
   a. The empirical results are weak, with the somewhat complex methodology presented only outperforming simpler methods when the base forecast models are clearly and obviously very badly miss-specified. The reader is left somewhat short-changed in that the level off effort required to understand and deploy the algorithm is in no way justified by improved empirical results compared to simply sampling from the base forecast distributions.

   We respectfully disagree with the referee here. Please allow us to make a few points which we now include in the paper in order to highlight the contributions of the paper.

1. The novel contribution of the paper is not only using score optimisation in forecast reconciliation. What is labelled OLS and MinT Shrink in the empirical application (and of course in the simulation setting) is also novel and goes beyond the point forecast setting (which has been previously explored by us or other authors). (Please see added text in the Introduction Page 4, lines 43-50)
2. From our analysis we find that there are two possible sources of misspecification: (i) assuming independence and (ii) assuming Gaussian errors. Such misspecifications are not unusual in practice. For example, organisational sections or silos generating their forecasts independently, is common practice. We argue that one of the novel contributions of the hierarchical literature, including this paper, is to overcome such situations using forecast reconciliation. (Please see added text in the Introduction Page 3, lines 19-36).
3. Despite an increasing recognition of the importance of probabilistic forecasts, it is not always the case in practice that organisations produce probabilistic forecasts. On the other hand it is more common for organisations to at least provide a predicted mean and a predicted variance. Having only a mean and variance available is also common with judgementally adjusted forecasts. Where only a mean and variance are available the logical parametric assumption is a normal distribution and bootstrapping (due to the lack of residuals) is not possible. Therefore the assumption of normality, while a misspecification for our empirical example, is not unrealistic in practice. We have now added discussion regarding the importance of the Gaussian setting in Section 4.2, Page 13, lines 8-15.
4. Our results clearly show that score optimisation in forecast reconciliation is worth considering in such commonly observed misspecification settings. We believe this is a very strong, and also useful, result in practice.

b. The empirical approach has a major conceptual weakness in that it fails to account for parameter uncertainty in the reconciliation process. HF processes are not parsimonious - most reconciliation models require the estimation of a vast dimensional covariance matrix. It is well known that such estimation exercises are fraught with difficulty, for example in the similar context of VAR estimation, it is well established that accounting for parameter uncertainty is critical, and the vast majority of published research in the area adopts a Bayesian approach. The fact that the author's own (full covariance) MinT approach only works well when shrinkage based estimation is adopted supports this viewpoint. The authors make this point in their conclusion, but it is of critical importance, and in my view undermines their empirical results here.

We agree with the referee to some extent here, but believe that much of what is being proposed goes beyond the scope of our paper. We agree with the importance of shrinkage whether it be used for estimating the parameters of a large variance covariance matrix (as required for MinT) or in regularising the reconciliation coefficients in the score optimisation algorithm. Regarding the latter, we certainly believe that this is a worthwhile avenue for research, and for this reason discuss it in the conclusion. Given that this paper has a number of theoretical results, uses these results to extend OLS and MinT into the probabilistic setting (in a novel way) and proposes the score optimisation algorithm, we feel that considering shrinkage in the score optimisation setting is best handled in another paper. We note that we comment on this as *“A promising future research avenue”* in the conclusion of the paper (see Page 27, lines 58-61 and Page 28 lines 1-2).

c. In my view the authors should focus on applying their own simple and robust idea of sampling from probabilistic forecasts and reconciling. The additional complexity of score optimisation without accounting for parameter uncertainty is certainly not justified by the results presented here.

We would contend that we do apply *“a simple and robust idea of sampling from probabilistic forecasts and reconciling”*. This method is used to produce joint probabilistic base forecasts either by assuming Gaussianity or via joint bootstrapping and then reconciling via popular point forecasting methods such as OLS and MinT. As argued elsewhere, these are novel extensions of existing point forecasting methods motivated by the definitions of probabilistic reconciliation proposed in this paper.

The novel score optimisation algorithm may not work in all cases. However, we have shown an important practical situation (as we argue above in our response to 6a) where it does work and works significantly better than the other reconciliation approaches. This is a novel contribution that one may choose to implement in a different setting and may possibly get better results. As discussed in our response to 6b we agree that shrinkage/regularisation is a potentially fruitful area of future research.

7. For publication in a Journal where real world application and decision making is important, the authors should make the effort to provide much more in the way of explanation and general reasoning behind their results.

We believe that this has been addressed in our response to the other points raised by the referee. In particular the new Section 2 as well as further discussion in the section on the empirical application provide further explanation.

Detailed Comments

1. Abstract… 'This method improves on base forecasts…' this is hardly a contribution… I think the same authors prove elsewhere that reconciliation always improves on base forecasts…

The results referred to here from our other work apply to the point forecast setting. The proposed manuscript explores the probabilistic setting, therefore the results here, even for MinT and OLS are all new. Even in the point forecasting setting, the statement made here by the referee is only partially true. In Panagiotelis et al. (2021) we do present general results but for two objectives: (i) to guarantee that reconciled forecasts improve upon base forecasts and (ii) to find the reconciliation method that is best on average. To guarantee objective (i) the loss function has to be matched with the weights matrix used in the projection reconciliation approach.

Panagiotelis, A., G. Athanasopoulos, P. Gamakumara, and R. J. Hyndman (2021). Forecast reconciliation: A geometric view with new insights on bias correction. International Journal of Forecasting 37(1), 343–359.

2. Section 2.2 I think this definition will be completely meaningless to many readers. Surely there should be a way of setting this more clearly without recourse to borel sets etc? Of course the definition is important (and therefore merits more explanation) but In a journal focused on applications this would be better off in an appendix?

   As well as the discussion added in the new Section 2, we have added further explanation to this definition (see Page 9 lines 5-15, following Definition 3.1). In particular we talk about assigning probabilities to “intervals”, “rectangles” or “regions”, since these are all cases of Borel sets when giving intuitive explanations. We continue to use Borel sets in the definitions only to ensure full rigour.

3. Section 2.3 - Again much more effort could be made to explain what is going on here

   We now summarise this result more succinctly in the new Section 2. We believe that the additional discussion around Section 3.2 (what was previously 2.2) as well as the use of Figure 2, makes the rest of this section clear.

4. Theorem 3.5 - I think this is a really important result (and one of the key contributions of the paper). A proof is of course provided, and while the theorem makes intuitive sense, no attempt is made to clarify and explain the logic of what is going on, and as the proof builds on the earlier results, and many readers will have to take it on trust…

   We have added additional explanation here (see the new text following Theorem 4.5 - previously Theorem 3.5). This proof also leans heavily on the concepts from Definitions 2.1 and 2.2 (now 3.1 and 3.2) which we believe are now more clearly explained.

5. Page 16, line 37. 'bottom level series have lower signal to noise than higher level'? It is a little surprising that this does not occur automatically, in practice the noise level in the higher level series would normally reduce via a diversification effect?

   Yes we agree. The bottom-level series did automatically have a lower signal-to-noise ratio.The additional noise was added to the bottom-level to make this difference more pronounced and replicate a realistic/practical setting. We have revised the statement now to read:

"After simulating from the ARIMA models, additional noise is added to ensure bottom-level series have a considerably lower signal-to-noise ratio than upper-level series with details provided in Appendix D of the online supplement."

Note we have also corrected a typo in the paper changing the second instance of the word *"bottom"* to *"upper"*.

6. Section 7.2 - As noted by the authors, the assumptions of independence and gaussian errors are clearly both badly violated in the data. The score optimisation algorithm generates 'statistically significant' improvements to base forecasts generated using naïve and inaccurate base forecasts! Is it really worth going to the trouble of reconciling such poor forecasts? When more sensible base forecast assumptions are made, simpler and more robust methods based on Theorem 3.5 perform more strongly.

   As we argue in our response to comment 6a above, and we argue again here, it is absolutely worth reconciling such forecasts, or at least it is worth having the knowledge and a tool, such as score optimisation, that can successfully reconcile these:

   1. We never know the quality of our forecasts a priori. Any evaluation of these will always be in sample and many times such evaluation is not possible, e.g., when these are judgmentally generated or in general not model based or the model is not made available.
   2. In commonly observed organisational silos, independence is the only assumption possible.
   3. A parametric assumption of Gaussianity for the errors is the usual alternative when only the mean and variance are available and/or when bootstrapping is not a possibility.

Reviewer #3:

This paper is concerned with the highly relevant problem of forecast reconciliation in a probabilistic setting. The authors (i) present a theoretical framework and theoretical results for this problem, (ii) propose an algorithm for probabilistic forecast reconciliation based on score optimization, and (iii) present results on both simulated and real data. In general, the paper is well written and of high quality. Considering all these elements, in my opinion, this paper is suitable for publication in EJOR after a minor revision.

We thank the reviewer for the positive and constructive feedback on our paper. We respond below in detail to the comments that require a response.

Please find below my comments.

General comments:

I very much like that the paper is concise and to the point. This is true for both the theoretical and empirical parts. The empirical sections are very good: nice presentation + insightful discussion for both the simulation and case study results. However, the introduction and the theoretical sections (2 to 5) are hard to follow at times, which I think is a consequence of these sections being somewhat too concise. Extra context is needed in some cases to make the text more comprehensible for a more general audience.

We have now added a new section (Section 2) giving a non-technical explanation of the main theoretical results. We have provided more context to some of the theorems at the request of Reviewer 2. We believe the changes have made the more challenging sections of the paper much more comprehensible to a general audience.

Some examples:

- P2L53: "Prior to the development of forecast reconciliation, the focus was on finding a subset of variables that could be subsequently aggregated or disaggregated to find forecasts for all series." Please clarify.

  We have now revised the statement as follows:

  “Prior to the development of forecast reconciliation, the focus was on forecasting a subset of variables at some selected level of aggregation, and subsequently aggregating or disaggregating these to generate forecasts for all series.”

- P3L1: "These papers formulated reconciliation as a regression model, however subsequent work has formulated reconciliation as an optimisation problem where weights are chosen to minimise a loss." Please clarify.

  Let us please note that this statement does not end where the reviewer indicates it ends. We reproduce below the full statement:

  “These papers formulated reconciliation as a regression model, however subsequent work has formulated reconciliation as an optimisation problem where weights are chosen to minimise a loss, such as a weighted squared error (Van Erven and Cugliari, 2015; Nystrup et al., 2020), a penalised version thereof (Ben Taieb and Koo, 2019), or the trace of the forecast error covariance (Wickramasuriya et al., 2019).”

  In order to further clarify we break the statement into two sentences and also add to it. The revised text now reads as follows:

  “These papers formulated reconciliation as a regression model, reconciling the base forecasts by projecting them onto a subspace for which aggregation constraints hold. Subsequent work has formulated reconciliation as an optimisation problem where weights are chosen to minimise a loss, such as a weighted squared error (Van Erven and Cugliari,

2015; Nystrup et al., 2020), a penalised version thereof (Ben Taieb and Koo, 2019), or the trace of the forecast error covariance (Wickramasuriya et al., 2019).”

- P3L9: "The accuracy and popularity of forecast reconciliation methods can be attributed to a number of factors". However, only one factor is discussed, i.e., breaking down organizational silos. Please also discuss/mention other factors.

  We have now revised the paragraph to read as follows:

  “The popularity of forecast reconciliation methods can be attributed to a number of factors. Forecasts across different aggregation levels may be generated by different departments or ‘silos’ within an organisation, using different sets of predictors, modelling approaches, or expert judgement. Potentially, these are viewed as optimal within these divisions. Reconciliation represents a way to combine information via the sharing of forecasts, thus breaking down these silos. Although it may be difficult to share forecasting processes and associated information across different parts of a large organisation, the forecasts themselves are much easier to share and reconcile. In contrast to bottom-up and top-down approaches, which effectively discard the forecasts of all but one level, the combination of forecasts across all levels also leads to improved forecast accuracy.”

- Ben Taieb et al. (2020) explanation: Either discuss it in more detail or in less detail, but now it is rather unclear what this method is exactly about. What is meant by 'reordering' base forecasts? To some extent this comment is also true for the discussion of Jeon et al. (2019): "ranking draws from independent base probabilistic forecasts before reconciliation is effective"?

  To elaborate on “reordering”, both JPP and BTTH obtain draws from the probabilistic forecast from each variable independently (let's think of these as vectors of length L). To reconcile these methods we take a single draw from each variable and combine these in an n-vector. In this case, the ordering of the draws within each vector of length L becomes critical and both BTTH and JPP attempt to do this in a way that captures dependence.

  JPP do this in a way that is equivalent to reconciling quantiles (which is how we have described on Page 3 lines 56-57, Page 4 lines 21-23, and Page 19 lines 45-55). We have rewritten the sentence *“ranking draws…is effective”* since the connection between this and reconciling quantiles may not be clear. Regarding BTTH, we have expanded on our discussion of their method adding the following on Page 4 lines 1-6

  “In particular, Ben Taieb et al.(2020) draw a sample of size L from the probabilistic forecasts of univariate models for the m bottom-level series and stack these in an L x m matrix. To induce dependence, the columns of this matrix are reordered so that the copula of the data matrix created, matches the empirical copula of the residuals. Samples of the aggregate series are obtained in a bottom-up fashion.”

- P14L44: "These are typically high dimensional problems, deep neural networks handle millions of parameters, so this tool is well suited to our problem." What do you mean exactly by this and 'these'? Gamma is not necessarily high-dimensional, right?

  The above sentence follows on from *"There is also a recent but growing literature on using SGD to optimise scoring rules (see Gasthaus et al., 2019; Janke and Steinke, 2020; Hofert et al., 2020, and references therein for examples)."* which highlights the recent literature using SGD to optimise scoring rules. *"These…problems"* refers to the problems tackled within these papers.

  We have now rewritten the sentence to make it clear. The sentence now reads (see Page 16 lines 35-36):

  "These papers typically deal with high dimensional problems, deep neural networks handle millions of parameters, so this tool is well suited to our problem."

  Also although Gamma does not necessarily contain millions of parameters (although in some applications it may do so), we simply make this point to assuage any concerns a reader may have about the applicability of our methods to larger hierarchies.

- P14L61: Discuss in more detail how the proposed score optimization algorithm differs from the method proposed by Rangapuram et al. (2021). At first glance, an obvious one is that your approach allows using general techniques (including ets() for example) to generate probabilistic base forecasts, which is not the case for Rangapuram et al. (2021). However, how is your method different from the one presented in Rangapuram et al. (2021) as for the reconciliation procedure (inclusion of translation, orthogonality of projection(?)…), and what are the implications? Just to be clear: I do not expect you to add this method to the empirical results in Sections 6 and 7.

  We have now extended the discussion of Rangapuram et al. (2021) as follows (see Page 16 lines 60 and Page 61 lines 38-43):

  "Rangapuram et al. (2021) use a similar approach in their end-to-end forecasting process. Their method is more restrictive than what we propose here in that the projection must be orthogonal, base forecasts are not translated, and base forecasts must be generated by a DeepVAR."

The conclusion could be improved a bit by adding some nuance to the discussion:

- "Since the scores are approximated by Monte Carlo simulation, stochastic gradient descent is used for optimisation." One could also use batch gradient descent using all Monte Carlo simulations in one go, isn't it?

  There is some subtlety to this point. In many contexts, (e.g. training neural networks) what makes stochastic gradient descent, *"stochastic"*, is that the training data are subsampled for computational reasons. In our setting, the Energy (and Variogram) Scores cannot be computed in closed form and must be estimated using a Monte Carlo estimate. This is what requires us to use SGD even when all data are used. We have now added the following (see Page 16, lines 39-46):

“An important distinction is that the use of SGD, rather than gradient descent in these contexts, arises due to computational considerations as it is not efficient to use all data. In contrast we use all data and the ‘stochastic’ nature of our gradient descent arises since the score functions contain integrals that must be estimated by Monte Carlo.”

- "This method is shown to lead to significant improvements over base forecasts, bottom-up methods and existing probabilistic reconciliation approaches across a wide variety of simulated and empirical examples." Refer also here to the degree of misspecification of the base forecasts.
-

  We have now revised the statement and have added the suggestion. The statement now reads as follows (see Page 27 line 49):

  This method is shown to lead to significant improvements over base forecasts, bottom-up methods and existing probabilistic reconciliation approaches across a wide variety of simulated and empirical examples, particularly when the base forecasting models are severely misspecified.

Minor comments:

- Add references to the first sentence of the Intro.

  Thank you for the suggestion. We have now added references to these.

- P10L60: Add dot at the end of footnote 1.

  Done.

- P12L18: Replace 'and' by 'are'?

  Done.

- P14L60: Is 'projection' the correct term here (also see P14L2)?

  Yes ‘projection’ is the correct term here.

- P17L35: Replace period by colon.

  Done.

- P18L10: Variogram score is only introduced by referring to Scheuerer and Hamill (2015), but no definition or explanation is provided.

  We have now added the definition and appropriate additional references and discussion on the discrimination ability of scoring rules (see Page 15 lines 1-13).

- P24L56: Remove dot after footnote 3 in text + add dot at the end of footnote 3.

  Done.

- How many errors do you use for the construction of bootstrapped probabilistic forecasts? I think that this info is missing.

  We have added the following on Page 19 lines 22-25.

  “The number of bootstrap samples is set equal to the sample size both here and in Section 8.”

- Can you provide an explanation for the deviant performance of BTTH for the Gaussian DGP in terms of variogram scores? The fact that a definition of the variogram score (and a discussion of how this score is different from the energy score) is missing makes it impossible for the reader to reason about this observation.

  We are unable to explain the deviant performance of BTTH in this setting. However, we have now added the definition of the variogram score. We also provide references that discuss scenarios in which the variogram score has good discriminatory power (see Page 15 lines 1-13).