Detailed Response to Referee 1: Summary

The authors provide a geometric interpretation for reconciliation of hierarchical forecasts. They show why and how reconciliation via projection is guaranteed to improve squared forecast errors. They explore a couple of different ways for dealing with biased base forecasts in an application to Australian tourism flows. Overall the paper is well written and the geometric interpretations are an important contribution to the growing literature on forecast reconciliation. The authors do a very good job explaining the geometric aspects, which lead to new insights. That being said, the contribution of the paper in its current form is mainly theoretical, as the empirical evaluation is not much of a contribution. I do not think the paper meets the high standard set by the IJF in terms of empirical evaluation. I recommend that the authors revise their paper with a particular focus on strengthening its empirical contribution to more clearly show the practical value of the geometric insights they derive. I hope that the authors will find my comments useful for improving their paper.

We thank the referee for their comments acknowledging the theoretical contributions of the paper. While we believe that more theory focused papers lie within the scope of the International Journal of Forecasting, we also agree with the referee that the empirical section could be greatly improved. We have endeavoured to do this in ways outlined in our responses below.

Detailed Response to Referee 1: Major comments

1. *Lack of empirical contribution: The authors apply a couple of different transformations to the tourism data before constructing base forecasts. Failing to correctly transform the base forecasts back before they are reconciled introduces a bias. That the highest accuracy is obtained when applying the correct back-transformation is a very small contribution.*

   We agree that if our empirical results only showed that applying the correct back-transformation improves accuracy, that this would be a very marginal contribution. The referee's suggestions have allowed us to make additional insights that we believe have greatly improved the paper. We discuss these in detail below.

   When the authors introduce the log-transformation and in the graphs in Figures 6 and 7, they do not clarify the interplay between the transformation and the reconciliation constraints. If $A + B = C$ *then clearly* $\log A + \log B \neq \log C$*. In other words, a log-transformation affects the reconciliation constraints. This needs to be explained more clearly; e.g., is the reconciliation constraint imposed on the transformed or the raw data?*

The observation regarding the constraints not holding on the log scale is correct. When series are transformed to the log scale, forecasts are first produced on the log scale and these are back transformed to the original scale (possibly with bias correction) prior to carrying out reconciliation. Therefore while all modelling is done on the log scale, all reconciliation is carried out on the original scale. The fact that the constraints do not hold on the log scale is therefore not relevant.

On re-reading the paper we do agree with the referee that there is scope for confusion here. Therefore we have added some additional exposition to the beginning of section 4.1. The aim here is to make it clear that after producing forecasting on some tranformed scale, the correct sequence of steps is to 1) back-transform to the original scale, 2) bias correct and 3) reconcile.

When the empirical evaluation is focused on bias, then it would make sense to include an error that measures bias in addition to the squared error.

We now also report the total error for all methods to highlight bias. Here we see that bias correction method 1 removes bias, while bias correction method 2 in fact worsens bias. A new insight that we are able to make is to recognise that reconciliation without bias correction can still mitigate bias. For example, even if bias correction is not used at all, OLS reconciliation significantly reduces bias for both the log transformation and Box Cox transformation. This is likely to occur since bias lies in a direction that is almost orthogonal to the coherent subspace.

Moreover, simply showing the MSE without any confidence intervals or measures of significance is not sufficient for the reader to assess the results. It would also be useful to show the MSE relative to the base forecasts or the percentage improvement that is obtained.

We now report all results relative to the base forecasts. We also have added stars to the tables that report our results to indicate when a method significantly upon base forecasts.

The best performing reconciliation method is MinT with shrinkage, but the authors never state the value of the shrinkage parameter or how it was chosen. Similarly, they compare with variance scaling without explaining what they mean by variance scaling.

The shrinkage estimator is that of Schäfer and Strimmer (2005). Details on this method (including the choice of shrinkage parameter) are now provided. Similarly, we now provide more detail on variance scaling, which simply refers to using the

variance of one step ahead in-sample forecast errors.

In addition to the above mentioned shortcomings, I think the authors should reconsider their empirical evaluation. Maybe a second case study or a simulation study is needed to show the value of the geometric insights provided. We already know that MinT is better than OLS and WLS. What is the new and better reconciliation approach that has come from the geometric insights?

The objective of this paper is not to propose a new and better reconciliation approach. Rather it is to establish new results that lead to clearer understanding of the properties of existing methods that in turn motivate research into new methods. Therefore, while the paper does not include a new method, it does provide insights into how new methods should (or should not) be developed. For examples of this, see the discussion on page 19 just before Section 4.1 and comments in the conclusion.

In this paper, the purpose of the empirical evaluation is to provide a demonstration of the theoretical results that we have established. We agree that this could be improved. We have greatly expanded Section 5.2 to more clearly investigate three different loss functions and demonstrate the different ways in which reconciliation methods can be considered optimal.

Improving the existing empirical application in the ways suggested by the referee has lengthened the paper considerably. Therefore, we prefer to refrain from including a second empirical demonstration.

2. *Improvement guarantees: The boxplot in Figure 8 shows that OLS always improves MSE, while this is not the case for the other reconciliation approaches. To gain a better understanding of the implications of Theorem 3.2, it would be useful to show that the other approaches always improve accuracy in their transformed spaces.*

   We now expand section 5.2 to report results for three different error metrics. Furthermore we have added additional insights in Section 3.4 that are now also demonstrated in this section. In particular we show that while the distance reducing property of Theorem 3.2 depends entirely on the choice of loss function, minimising the expected loss function can always be achieved by a MinT approach and does not depend on the weights used in a loss function.

   What is the interpretation of the transformed spaces and can the authors make the connection between these spaces and the choice of reconciliation approach and

error measure more clear?

Reflecting on this issue has led us towards making a number of changes. To reiterate, we may define a loss function based on the notion of Euclidean distance. In this case an orthogonal projection is guaranteed to lead to reconciled forecasts that are better than base forecasts. Alternatively we may define a more general loss function, for instance by taking weights into account. We now consider two such loss functions, one based on the structure of the hierarchy (so-called 'structural scaling') and another based on average spend in each region. To highlight how the transformed space can be interpreted in some contexts we have added the following sentence to the manuscript: "By using average spend per region as weights, the error metric (and transformed space associated with this metric) can be interpreted in terms of revenue measured in dollars rather than raw tourist numbers."

In cases where the transformed space cannot be easily be interpreted in terms of the empirical example, it is nevertheless it is an important construction in our proofs. This includes a new proof that MinT minimises expected loss even for loss functions based on generalised Euclidean distance.

For example, Hyndman et al. (2011); van Erven and Cugliari (2015) argued for selecting OLS to increase the importance of forecasting the aggregate. What is the argument for WLS or MinT and what is the corresponding consistent error measure?

We would like to point out that any argument made by Hyndman et al. (2011) or van Erven and Cugliari (2015) that OLS reconciliation somehow targets and improves the aggregate series is based on empirical rather than theoretical evidence. Addressing this misconception, which appears to be common, has in fact been a major motivation behind us writing the paper. We hope that with the revision to the paper these distinctions are clearer.

Detailed Response to Referee 1:Minor Comments

1. *In the first half of the paper it feels like every other sentence includes a however. I suggest reducing the use of however.*

   We have reduced the usage of the word "however" substantially, from twelve instances to just three.

2. *P. 2, l. 11: In several places the authors talk about adjusting forecasts ex-post. Although I understand what is meant, it gives the impression that forecasts are*

adjusted after observing the realized values.

We have either removed all use of the term 'ex post' or stated 'ex post of base forecasting' to avoid the potential for confusion.

3. *P. 2, l. 12: The authors discuss the regression formulation of forecast reconciliation. It would be useful to also make the connection to the optimization formulation considered by, e.g., van Erven and Cugliari (2015); Nystrup et al. (2020). This could also be useful for clarifying the connection between reconciliation approaches and error measures.*

   We now discuss the connection to the optimisation formulation considered by these two papers on page 2 line 12 as requested. We believe that some of the other revisions we have made, particularly the new discussion in 3.4 further clarify the connection between reconciliation approaches and error measures.

4. *P. 4, l. 22: forf*

   This has been corrected.

5. *P. 10, l. 12: the comma should not be there.*

   The comma has been removed.

6. *P. 11, Figure 3: usually a small square is drawn in the corner of the triangle to show orthogonality.*

   We have made this change to all Figures. TO DO (GEORGE)

7. *P. 17, l. 18: i,e.*

   We have made this correction.

8. *P. 26, l. 14: the authors mention that the full results are available upon request. I suggest including them in an online supplementary appendix.*

   We now include these an an online supplement. TO DO

9. *P. 27, Conclusions: The authors should comment on the implications of the non-uniqueness of the S matrix for future work on cross-temporal reconciliation (Kourentzes and Athanasopoulos, 2019).*

   TO DO (GEORGE).

Detailed Response to Referee 2

This type of paper makes me regret not investing more time into geometric interpretation because as shown here, it offers an elegant and intuitive way to showcase results related to data integration and reconciliation. The paper is extremely well written, with a great flow and thoughtful considerations. Figure 4 and its description are exemplary successful in their simplicity and effectiveness. The discussion of theorem 3.1 on page 11 is another example of thoroughness and clever insight.

We thank the referee for these kind comments.

I found only one statement in the paper that could be better supported by evidence on page 8 lines 14 when the author(s) refer to multivariate modeling. State-space approaches have also been shown to be theoretically successful is solving these problems although maybe not to the large scale needed for very detailed and complex hierarchical systems. A comment or comparative discussion to the multivariate modeling may be useful.

We have rewritten this sentence and now explicitly include the case of state space models in our discussion.

In the context of real-life application and either for small discussion here or future work, I am wondering if and how the author(s) coherent subspace that would be defined with hard boundaries. For example, the set of Australian tourism flow data and any forecasts that would be considered useful should be non-negative and likely upper bounded (if only by the size of the global population or other more realistic subject matter expert opinion). In many other reconciliation problems, these boundary constraints affect the feasible space. In the context here, could a convoluted case lead to an orthogonal projection be coherent but outside the desired constrained subspace?

It is indeed possible for an orthogonal projection to reconcile a set of positive base forecasts into a vector that includes some negative values. This is a fairly pathological case and does not arise in the application that we consider (and many other applications that we have considered in other work). We do agree that this problem may arise in other contexts so we now include some discussion of this issue in the conclusion and cite some recent work that addresses this problem.

Please correct the minor typo just before section 2 (forf).

We have made this correction.