Response to Reviewers

**Reviewer #1:**

The manuscript applies a Bayesian forecast reconciliation method to forecast Swiss merchandise exports. There are several interesting aspects on the paper. The introduction provides a motivation of what follows and the methodology makes sense to me. The structure and the arguments are clear.

Section 1 provides, the introduction, motivation, the related literature and the contribution of the study. The motivation is strong and the section provides an overview on the related research. However the contribution of the paper could have been more clear. There is a related paragraph in the pg 3 and 4 that should be expanded. There should be a more direct comparison with the related literature and the advantages of the new approach.

We have expanded the paragraph on pages 3 and 4 to highlight the innovations of the proposed methods and compare it to the existing literature. In particular, our use of a state-space form to jointly reconcile at multiple forecast horizons, and use of a Bayesian approach with careful selection of informative priors are discussed in further detail.

Section 2 has the related methodology. The approach seems correct to me. The "more novel" part for me is section 2.3. I believe that the manuscript will benefit from extended simulations of the effect of the bias shrinkage (in addition to the simple one presented in figure 2)

We agree with the referee that this important aspect of the method has been neglected so far. In section 4.4, we have added simulations on bias shrinkage with a focus on our application and in particular draw your attention to Table 3. In addition, we have put further emphasis that the method is relevant for operational forecasting because it allows to reconcile forecasts not just with the objective of improved forecast accuracy.

Section 3 has the dataset. There is a description of the series under study, however I would like to see some summary statistics on the series under study. I wonder on the effect of time varying exchange rates on your figures and if they have any effect on the estimations that follow.

We have added summary statistics (Table 1) to provide a quick overview of the hierarchical structure of the dataset and the time series at important nodes.

The impact of exchange rate fluctuations on the data is now addressed in section 3: ﻿

*All values shown refer to the invoiced price of the goods in Swiss francs, including*

*transport and insurance costs as well as other expenditure up to the Swiss border. If the*

*invoice is in a foreign currency, the invoiced amounts are converted using the previous*

*day’s exchange rate. As a result, the figures are affected by exchange rate fluctuations.*

*However, prices respond in a way that mitigates the influence of exchange rate fluctuations,*

*due to a quick exchange rate pass-through, documented by Bonadio*

*et al. (2020) for imports as well as exports.*

Since we focus on the impact of reconciliation on forecasting accuracy rather than the methods used in the underlying base forecasts, it is not crucial to take exchange rate fluctuations into account. We have clarified this as well in section 4.

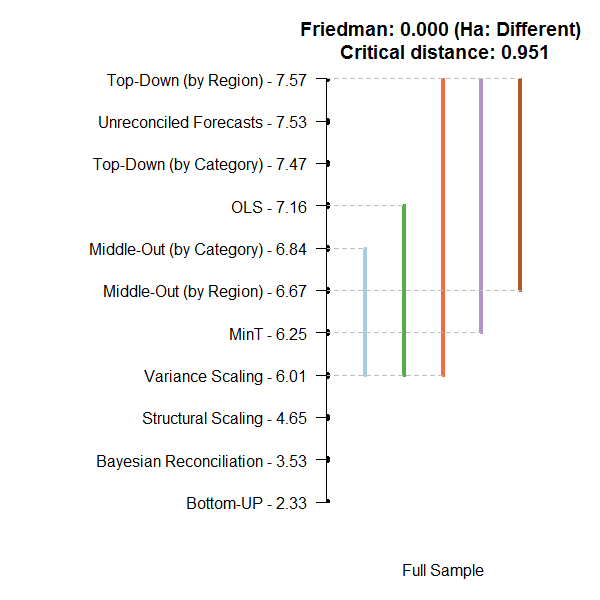
Section 4 presents the empirical analysis. You mention that there is a detail description of the methods in Appendix A.3 (footnote, 3). There is some related information but it is definetelly not detail. You need to expand this part. The forecast evaluation should been richer. The comparison is based on the RMSFE and several graphs. Are your results similar with other forecasting metrics?

Also in addition to the DB test, you can employ the Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. Econometrica, 74(6), 1545-1578. For your sub-periods: 2007 should be in the crisis. It is not a moderate year

We expanded the robustness checks and added detailed comments on the difference between our used metrics in the main part and alternative forecasting metrics in the appendix.

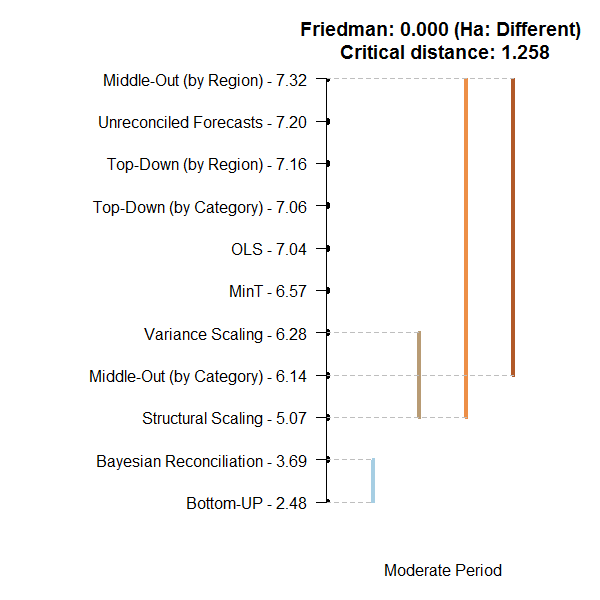
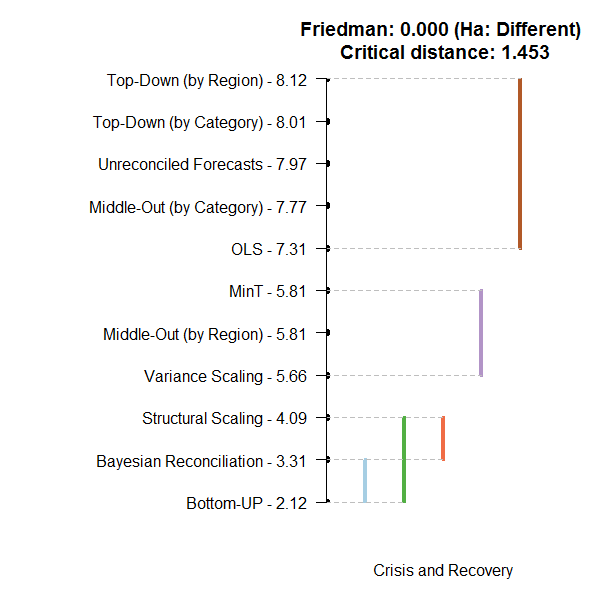
The use of appropriate test statistics is an important point. The Giacomini & White test might not be appropriate here because it requires that predictions be made with rolling windows and, more importantly, is used for nested models (as are other methods such as Clark & McCracken (2011) or Clark & West (2007)). Their use may lead to incorrect statements as reconciled and unreconciled predictions do not originate from nested models. We have discussed this issue in section 4.2.

We provide instead additional comparisons based on non-parametric Friedman and post-hoc Nemenyi tests (Hollander, Wolfe, & Chicken, 2013) for the top-level series at a forecasting horizon of 24 months. In a first step, the Friedman test establishes whether at least one forecast is significantly different from the others. If this is the case, the Nemenyi test is used to identify groups of forecasts which are not different with a 5% statistical significance. These tests have been used in the literature before, see for instance Kourentzes & Athanasopoulos (2019). The methods are sorted according to MSE rank and the bottom row shows, therefore, the best performing methods. For the full sample, they agree with the DM-tests that combination methods and middle-out reconciliation lead to more accurate results. In contrast to the DM-tests, the bottom-up approach also shows promising results:



Comparing the subperiods using Friedman and posthoc Nemenyi tests largely mirrors the results from the DM-tests. In particular, the combination methods work well during the crisis period.

We decided to include these results only in the response and leave them out of the paper in order to save space.

Regarding the definition of the crisis period, we have double checked our code and discovered that the original results in the table were in fact including 2007 in the crisis period rather than the moderate period. We actually made a mistake in mis-labelling the table (which we then carried through to the text). We have now corrected this. We note that our code is available publicly on Github (<https://github.com/florian-eckert/hierarch_bayes/blob/master/cls2/3_postprocessing.R>, with the DM-test starting at line 154).

Section 5 has the conclusions. I believe that this section is short. You should expand the comments related to the possible applications of your method to OR and forecasting. What are the implications of your study to practitioners and researchers?

The conclusion has been now been expanded. In particular, we now highlight the implications of shrinkage reconciliation for operational forecasting and their extension to other demand forecasting applications.

**Reviewer #2:**

In this paper authors have developed forecast reconciliation framework using Bayesian state space methods. They find that the use of state-space methods is promising for producing accurate forecasts.

The paper is an extension of the 'structural scaling' and 'MinT' approaches with the Bayesian flavor. I find the proposed method interesting, addresses a significant problem, and well structured. Several problems need to be addressed before the paper can be considered for publication. However, if the authors have done what I think they are trying to describe, then I see no problems with their methodology. For this reason, revisions are recommended.

General points:

\* When reading the manuscript, it is clear that the authors have not consulted the instructions for authors, which is provided by the publisher's website. These instructions need to be followed.

\* For example, there is no 'Highlights' section, which according to the instructions for authors "Highlights are three to five (three to four for Cell Press articles) bullet points that help increase the discoverability of your article via search engines. These bullet points should capture the novel results of your research as well as new methods that were used during the study (if any)."

We apologize for not meeting certain editorial standards at the submission stage. We have now included highlights.

\* The authors are often repeating the same information multiple times. The text should be shortened by omitting redundant information.

We have streamlined most sections of the text in order to avoid redundancies. In particular, the discussion of prior shrinkage in section Section 2.3 has been moved to section 4.4. in order to show it together with a simulation of certain shrinkage outcomes and to highlight the benefits for operational forecasting. While some redundancies in the introduction have been eliminated, the section is still slightly longer to accommodate additional comparisons to the related literature in the paragraph on pages 3 and 4. The conclusion in Section 5 has been rearranged to have a clearer structure.

\* The presented abstract and discussion are insufficient and need to be expanded.

The abstract as well as the conclusion have been expanded, with a focus on highlighting the relevance of the paper for operational forecasting.

\* The written English is substandard as detailed -but not limited to- the specific points below:

\* Specific points:

o Add a hyphen state-space, lower-level, single-level, top-level, state-dependent.

o Abstract, line 30: It is unclear what "This" refers to.

o Abstract, line 31: remove the redundant preposition "at"

o Page 1, line 5: "in nature" creates tautology.

o Page 1, line 5 and page 3 line 51 and Page 13: "comma before "therefore".

o Page 1, line 11: The sentence is unclear, rewrite the sentence start with "This".

o There are sentences that an "and" or "or" should be placed after the final comma to mark the end of the list. Examples include, but are not limited to, page 3 line, page 8, and page 12.

o Page 3, line 51: Add a comma before particularly

o Page 4, line 5: The word "policy makers" seems to be miswritten. Consider replacing it.

o Page 5, line 22: Add a comma before "for" and/or after "instance"

o Page 8, line 35 and page 10, line 32: It appears that you are missing a comma or two with the interrupter on the other hand. Consider adding the comma(s).

o Page 10, line 21: Add a comma before "with"

o Page 11, line 39: The noun phrase strength seems to be missing a determiner before it. Consider adding an article.

o Page 12, line 28: The noun phrase optimal hierarchical combination seems to be missing a determiner before it. Consider adding an article

o Page 16, line 51: The verb "do" does not seem to agree with the subject.

o Page 18, line 5: The word difficult is often overused. Consider using a more specific synonym to improve the sharpness of your writing.

o Page 19, line 21: Relying does not seem to work here

We are grateful to the referee for pointing out these issues. We have addressed each of these points and in addition, the text has undergone further proof-reading to meet the editorial standards.

\*\*\*