Dear Reviewers,

The reviewers of our paper have obviously spent considerable effort in reading our work and pointing out many directions in which the paper can be improved. We are really grateful for this and want to thank all reviewers upfront for their diligence. Following the review comments, we have strived to revise the manuscript based on the comments accordingly. Our revisions are marked in blue in the manuscript to ease reading. In the following, we describe the changes that we did to the paper in response to these comments and suggestions. To make reading easier, we mark the reviewers’ comments in *blue italic*. Our answers follow in plain text. Also, for brevity, we omitted in the text below the positive appreciations in the reviews and also comments which did not point to issues to be improved or changes to be effectuated.

Thanks again for reviewing our paper. We hope everyone stays healthy in this time!

Sincerely,

The Authors

**Summary of main changes**

* We have revised the visual designs, including
  + a
* We have added more discussions, including
  + a
* We have proofread the paper and made several corrections, including
  + a
* We have updated the video to the latest design.

**Summary Review:**

*issues with the analysis tasks (R1, R3).*

* elaborate on how the analysis tasks were derived, possibly discussing existing task taxonomies from literature or, if none exist, outlining this gap in the literature (R3)
* relate the design criteria to the analysis tasks (R3)
* add task "detection of overall patterns of prediction" (R1)

aaaaaaa

*reformulate the contribution (R1).*

* reformulate the contribution as a general framework which is instantiated on example of the two given transformations (R1)

aaaaa

*missing explanations for visual designs (R1) as well as details about model (R3).*

* add explanations of the "inspector view" (R1)
* provide some more details about the model train/test split and its validation (R3)
* As the page limit does not allow explaining all visual representations in detail, please extend the video by explaining the functionality of visual representations (R1)

aaaaaa

*issues with case studies (R3, R4).*

* In the case studies, please include an evaluation methodology and elaborate on the missing details, e.g., which of the user groups’ feedback is presented, were there any drawbacks/dislikes/issues? (R3)
* clarify why a VA system is needed to achieve the goals of "Mike" in the cases studies (R4)

aaaa

*required justification of some metrics (R2, R3).*

* justify choice of RSME metric (R3)
* The example in Figure 9 allows users to visually confirm the efficient smoothing method and factor for the prediction by Deep learning. However, SHAP was designed for a similar purpose. So, further clarifications and comparisons are required in the revision. (R2)

aaaaa

*too limited discussion scalability of the system (R2).*

* discussion of scalability of TimeTuner wrt. number of variables in time-series data must be added (R2)

aaaa

*revised references and typos (R1, R3)*

* *fix the inconsistencies with the acronyms (R3)*
* *include DOIs in the references, references to the tools and libraries used, and citations to the tests that were used in the time-series forecasting sections (R3)*
* *add reference to "Visualization of time-oriented data" by W Aigner, S Miksch, H Schumann, C Tominski (R1)*

aaaaaa

-------

Below we list responses to questions raised by individual reviewer. Detailed answers for the summary reviews can be found accordingly.

**Reviewer 1:**

*The related literature is more or less complete. However, it is surprising to see a submission dealing with time series analysis and NOT citing the THE TIME book "Visualization of time-oriented data" by W Aigner, S Miksch, H Schumann, C Tominski - please fix!*

aaaaaa

*The system of tasks seems to be incomplete. T3.1 addresses detection of prediction outliers. However, a necessary per-requisite for this task is detection of overall patterns of prediction - how prediction quality is distributed in the overall space of available transformations Vs. time. This task is not formulated and not supported by the system.*

aaaaaa

*Implemented transformations are quite limited and rather trivial. As expert interviews suggest (last paragraph of sect.6.3), there exist many other very important transformations. I suggest reformulating the contribution as a general framework which is instantiated on example of the two given transformations. This requires, however, consideration of scalability of the framework and implementation in respect to the number of transformations and their cardinality (count of different options).*

aaaaaa

*The same idea is applicable to the considered pre-computed counter-factual explanations. Generally, I would suggest reducing sections 4.2 and 4.3 for giving more space for explaining visual representations. Thus, the so-called "inspector view" requires further explanations. What is the meaning of differently shaped rectangles in the plot? Why some of them are full squares, but other are aligned to top or bottom?*

aaaaaa

*I am not convinced that the VSUP palette works really well. I see on the plots mostly variants of color intensity, thus discarding the 2nd variable. The alternative design (Fig.6) has its disadvantages. Therefore I recommend considering either paired time lines or paired heatmaps for representing pairs of quality indicators over time. More over, this approach would enable inspecting more than 2 quality measures simultaneously.*

aaaaaa

*Fig.7(A) is not referred in the text - pls fix.*

aaaaaa

*Supplementary material describes well the use case. As the page limit does not allow explaining all visual representations in detail, please extend the video by explaining the functionality of visual representations.*

aaaaaa

**Reviewer 2:**

*There are no significant contributions in terms of visualization/ visual analytics*

aaaaaa

*Besides the limitation on the time-step parameters (as mentioned in section 7, Discussion), I also think that TimeTuner is not scalable with the number of variables in the time series data.*

aaaaaa

*Second, the paper employs different metrics to measure traffic flow prediction but it does not provide enough insights to explain the opposite conclusions on different measurements. In case studies, the author does not describe the connection among the trends of* *PRMSE, uncertainty coefficient, and CORR, either. The author needs to add more evaluation on the metrics instead of simply listing them in multiple views.*

aaaaaa

*I am unsure if the color usage in Figure 8 is correct. In particular, in the first column of the matrix where x=Pm25, the color is supposed to be darker toward the right side of the cells as Pm25 is the target variable.*

aaaaaa

*There is a minor typo in the video script at 2:35.*

aaaaaa

**Reviewer 3:**

*The sections about the analytical tasks may benefit from some parallels to existing literature on task taxonomies, grounding them better, and if there are no suitable taxonomies in this case the limitations of these taxonomies for this specific problem domain can be outlined. It's a bit unclear to me how these analytical tasks were derived some further elaboration would significantly improve this section and make it more transparent to the reader.*

aaaaaa

*In the section about transformation, forecasting, and explanations, please add references about the tests that are used (ACF, ADF, etc.).*

aaaaaa

*Regarding the validation of the model, is there a specific reason that an 80/20 train/test data split was selected and not another form of validation? I assume this because it is the usual approach in ML/DL. Please provide a short justification as to why.*

aaaaaa

*In order to validate the time-series forecasting models you use the RMSE metric, which you argue, initially, overlooks the uniqueness of sliding windows and fluctuations (see Introduction):*

*"Firstly, the evaluation of time-series forecasting is currently limited to single numeric metrics, such as root mean square error (RMSE), which overlooks the uniqueness of individual sliding windows and does not explicitly reveal which parts of the data representations are responsible for model behavior and resulting predictions [46]." Please justify this choice.*

aaaaaa

*In my opinion, the case studies (evaluation) can be improved by adding more details. This section lacks a formal evaluation methodology and more details are necessary about the expert interviews and amount of independent visualization researchers that participated. Furthermore, the outcomes of the case study with visualization researchers should be discussed in the paper, or is the entire feedback from both groups combined? I recommend the authors re-write the case study section, including an evaluation methodology and elaborating on the abovementioned details.*

aaaaaa

*Finally, I'm curious if the proposed approach is available online - It would be greatly appreciated to provide a link to a repository or online version of the prototype.*

aaaaaa

Revisions Required:

In my opinion, the paper presents a valuable contribution, however, in order to accept this for presentation at the conference, I believe the following revisions are required:

- Elaborate on how the analysis tasks were derived (possibly discussing existing task taxonomies from literature or, if none exist, outlining this gap in the literature).

- Provide some more details about the model train/test split and its validation.

- In the case studies, please include an evaluation methodology and elaborate on the missing details (which of the user groups’ feedback is presented, were there any drawbacks/dislikes/issues?).

- Fix the inconsistencies with the acronyms - some are used before they are defined, some are defined multiple times, and the sunspot dataset is referred to as both SILCO and SILSO.

- Relate the design criteria to the analysis tasks.

- Include DOIs in the references.

- Include references to the tools and libraries used (@online)

- Include citations to the tests that were used in the time-series forecasting sections.

**Reviewer 4:**

*I am a bit puzzled by the case studies. Both seem to aim at demonstrating how TimeTuner enables Mike to identify the optimal setting of smoothing and sampling for achieving the highest prediction accuracy. Wouldn't it be sufficient for this purpose to automatically analyze the content of the representation view and "search" for the brightest representation? Is it necessary at all for this purpose to have a visual analytics system? Mike could also be supported by automatically sorting the representations according to overall brightness.*

aaaaaa

End of response letter