Author's comments to referees

Again, we express our gratitude to the referees and the handling editor for the time they invested in helping to improve our manuscript. As before, we agree with all comments and believe that the implemented changes further enhance the quality of the paper. Below you find our detailed response to the referee comments:

AC Comment to RC1 – Implemented changes

> Code: I reviewed the code on the Zenodo repository and was able to follow the README to check that both scripts and notebooks worked correctly. The scripts (mainly the tuner) I didn't run until termination as it was not my intention to recreate any models, only checking if they were working correctly. My only suggestion would be to add a ´requirements.txt´ file to the Github repo containing the same information as the 'environment.yml' as not everyone uses Anaconda as their package manager (I don't). I think it's worth the trouble because the Github repo can be a very good starting point for someone just getting started in a similar application of ARIMA models and/or LSTMs.

A requirements.txt was added to the Github repo as well as to Zenodo

> Section 4.1: I would be careful in using words such as "exceptional" or "outstanding" to describe the evaluation results of a particular model since performance assessment should rely on quantitative metrics rather than qualitative judgment. Also, there are cases in which there is no "winner" depending on the metric and/or year analysed, therefore the use of these words feels overly positive and overstate the authors' findings.

We agree with the referee on this point. We rephrased all relevant sections that sounded overly positive and excluded wordings like "exceptional" or "outstanding".

> ARIMAX: In the current manuscript, a direct comparison between the ARIMA and eLSTM models can be made, but I feel there's still a gap due to the absence of a comparable benchmark for the PBHM-LSTM model. An ARIMAX model could fill this gap, though I understand its exclusion aligns with the original research focus on improving upon the commonly used ARIMA model. That said, the authors' response shows their test of an ARIMAX model in which they finding no significant differences compared to the ARIMA model included in the paper. I believe that including this detail in the manuscript would be beneficial as it adds context.

We now mentioned that an ARIMAX model was tested but did not yield significant differences to the presented ARIMA model in the discussion section. We also emphasized on the fact that this was true for this particular study, but should not be seen as a general statement.

AC Comment to RC2 – Implemented changes

> timing error and peak error should already introduced in the "Methods Section 3.2". The symbol for the timing error is somewhat misleading: maybe eΔt would be more appropriate.

The peak magnitude and timing errors are now introduced in the methods section 3.2. We also agree that the symbol of the timing error could be misleading. We changed this accordingly in the revised manuscript.