Reviewer 2:

The paper addresses an important current issue regarding the applicability of AI-generated weather forecasts for impact-oriented predictions of mortality.

The analysis is presented in a clear and structured manner throughout, making it easy to follow the line of reasoning.

Overall, the authors achieve the paper's objectives with their presentation.

Nevertheless, I recommend making a few minor adjustments to the technical and visual presentation, primarily to facilitate the interpretation of the results.

Thank you for the time you have invested in reviewing our manuscript and your constructive feedback. Please find our point-by-point proposed responses below.

1. Both the title and the abstract give the impression of a systematic comparison of data-driven and physics-based forecasts. However, the paper compares only two deterministic models for one summer, so the generalisability of the findings is not obvious. In the interest of transparency, it would therefore be advisable to address this limitation already in the abstract.

   Thank you for highlighting this. We propose clarifying the sentence on line 6 by adding the word 'case' as follows:

   This case study compares European…

   We agree that these results should not be generalised, nonetheless, we performed this based on the data that was available at the time, and would like to advocate for the continued release of data-driven forecasts. In particular, the release of a set of hindcasts would be particularly useful for allowing for this study to be extended to a systematic evaluation. It is also worth noting that the two forecasting systems being compared are the leading deterministic systems by ECWMF, see also answer below.

2. The authors may add their reasons for the choice of the two particular models.

   These two models were chosen because they are the flagship, state-of-the art models developed by ECWMF, and are widely considered some of the best weather forecasting models in the world. The rationale behind this choice was that since the two models were developed by the same institution, they would largely have the same priorities for aspects of forecast performance, thus making an illustrative case study as a proof of concept for applying an epidemiological framework to data-driven forecasts. We will also clarify this in the text.

3. It might be helpful for the reader if the ordering in section 2.1 would be changed such that the used models are mentioned before the specifics are explained.

Thank you for highlighting this. In answer also to comment 2 of Reviewer 1 we would clarify the use of ERA5 and the forecast models at the beginning of section 2.1.

4. There is no mentioning of the population data that is used for the weighting. This should be added to increase transparency. Additionally, a map or a list of the used cities would help the reader to assess the geographical distribution of the data.

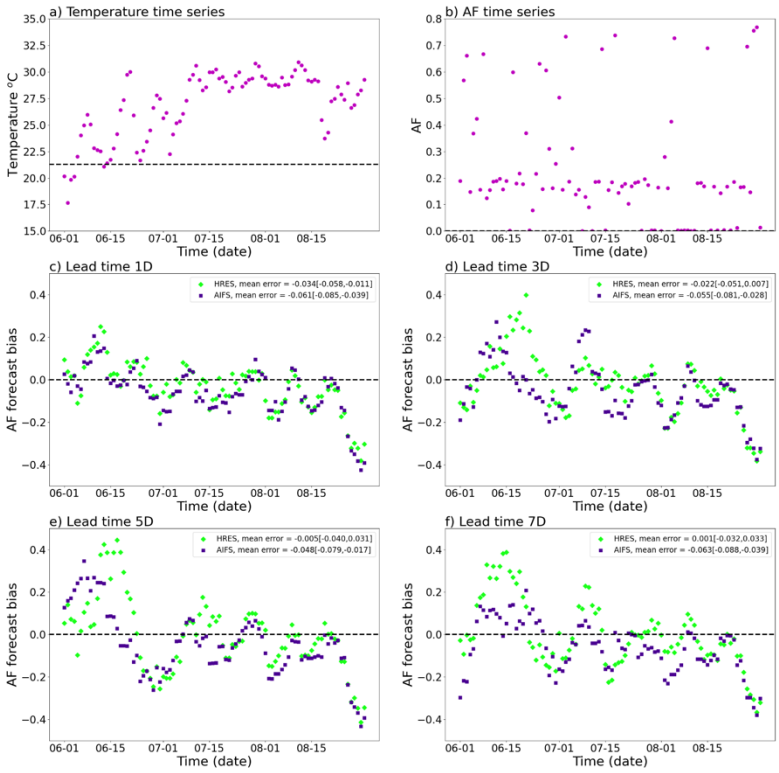
Thank you for highlighting this, we apologise for missing this point. The population data used was that published by Masselot et al. (2023) together with the temperature-mortality associations, which we will clarify in the manuscript.

5.Since several results are visually on the edge of significance, a more thorough use of statistical methods could be helpful for the interpretation of the stated results. In particular this holds for the following points:

a. The means of figures 1c-f are compared (lines 117-118) without stating the uncertainty of the calculated means, which should be accessible from the mentioned bootstrapping procedure. The results of a statistical test on the significance of the difference would then help to interpret the result.

Thank you for this suggestion. We show below the confidence interval estimated through bootstrapping, although we would like to highlight that we have performed this bootstrapping on a very small sample size meaning that these findings should not be generalized. We are thus hesitant to include these numbers in the manuscript so as not to mislead the reader on overstating the robustness of our results. Since the bootstrapping is also based only on the data from the study period, any multi-year extremes not observed during the study period would also not be represented in the confidence intervals, which is why we believe it to be best to focus on

descriptive rather than inferential claims in this respect.



Revised Fig. 1: Time series of temperature (a) and AF (b) for the summer of 2024 in Rome. The minimum mortality temperature for Rome is shown by the black dashed line in (a). Time series of the AF forecast bias for Rome during the summer of 2024, where green denotes the physics-based forecast (HRES) and purple denotes the data-driven forecast (AIFS), smoothed with a 7-day rolling mean (c--f). Lead times 1 day (c), 3 days (d) 5 days (e) and 7 days (f) are shown. The dashed black represents the ground truth obtained when using ERA 5 as meteorological input data.

b. The same is true for the interpretation of figure 7c-d where the significance of the difference would improve the reported observations in line 167-170.

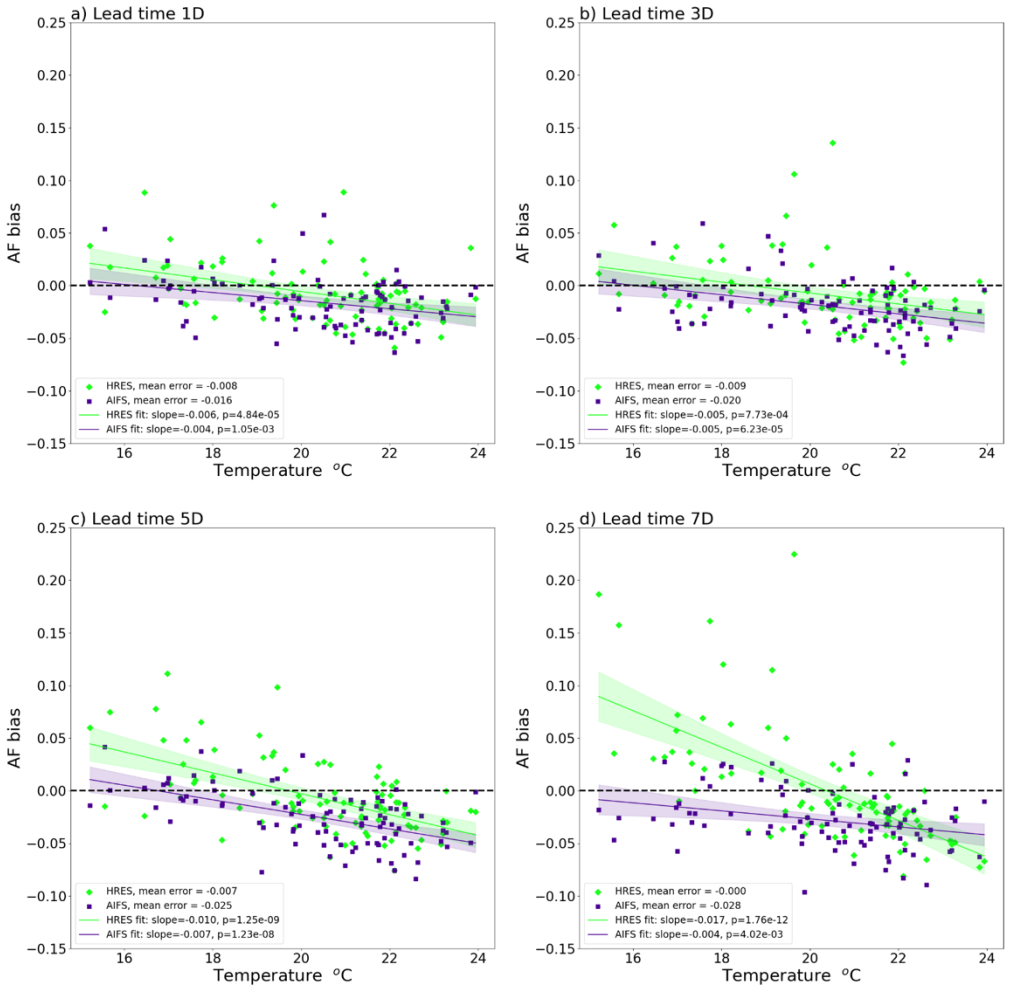
Here we have made our interpretations considering when the centre lines fall outside the confidence intervals as a way of showing some statistical robustness. As for the previous point, we do not want to mislead the reader by over-implying the statistical robustness of our results given that we consider only one summer as a case study.

c. The comparison in line 156 mentions a Kolmogorov-Smirnov test whose results should be included for transparency.

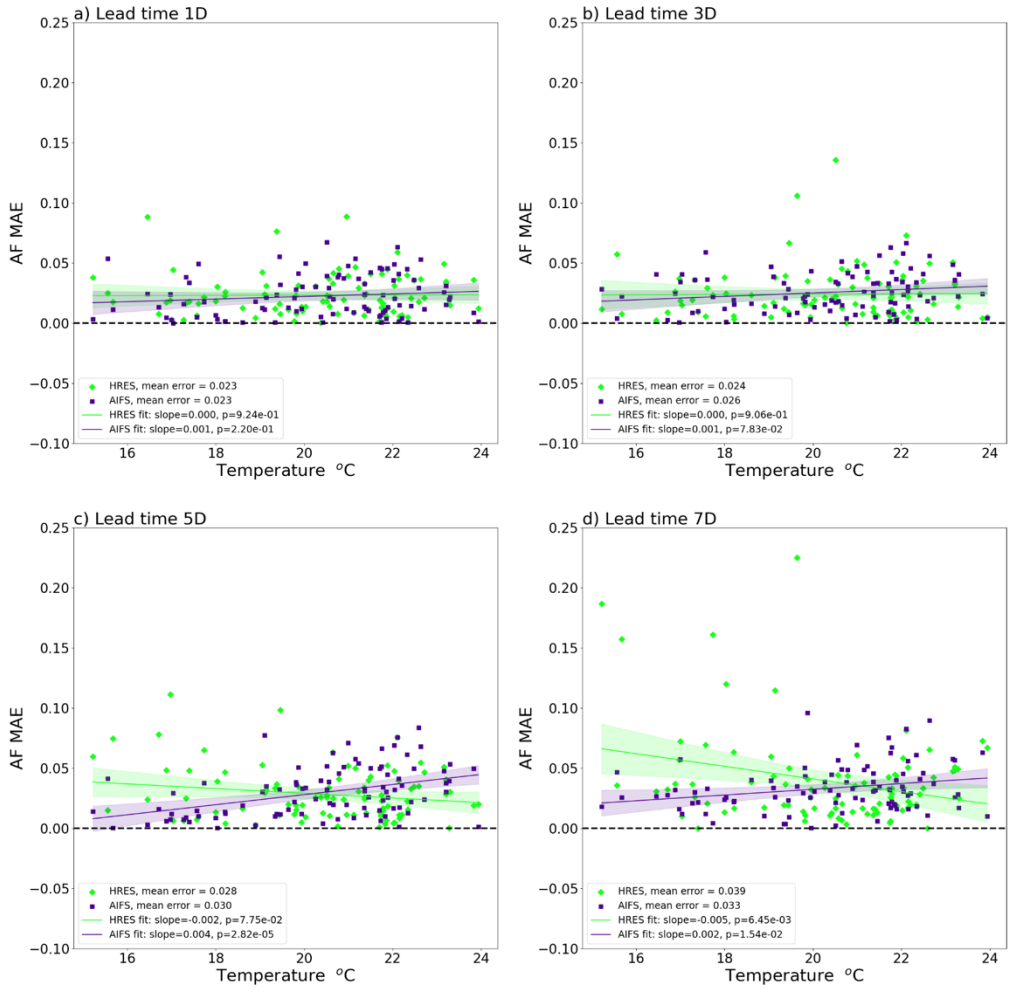
We propose including the results for the tests in the sentence on line 156 as follows:
This was further verified by performing a Kolmogorov-Smirnov test, the results of which indicated that for lead times 3, 5 and 7 days, the physics-based and data-driven AF forecasts correspond to different distributions (p values = 0.026, 0.026 and < 0.001 respectively)

6. The performed fits in figure A4 and A5 are subject to uncertainties that could visually be shown by plotting the confidence band. Since the results of figure A4 are a central result discussed in section 4 (lines 175-179), the significance of the difference in the fits, especially at high temperatures, is important for the interpretation of the stated conclusion. Also testing whether the fitted slope differs significantly from zero would be an interesting information.

We have revised the figures to now include confidence intervals and the p value associated with the fit. We note here that we now use the package Scipy to perform the statistical analyses.



Revised Fig. A4: Scatter plot of the population-weighted mean AF forecast bias vs temperature for the summer of 2024; green denotes the physics-based forecast (HRES) and purple denotes the data-driven forecast (AIFS) (a--d). Lead times 1 day (a), 3 days (b) 5 days (c) and 7 days (d) are shown. The lines show a linear fit performed using the 'linregress' routine from the Python package Scipy.



Revised Fig. A5: Scatter plot of the population-weighted mean AF forecast MAE vs temperature for the summer of 2024; green denotes the physics-based forecast (HRES) and purple denotes the data-driven forecast (AIFS) (a--d). Lead times 1 day (a), 3 days (b) 5 days (c) and 7 days (d) are shown. The lines show a linear fit performed using the 'linregress' routine from the Python package Scipy.

6. Lines 151-159 describe the results of figure 6, which compares the distribution of the daily averaged AF forecast biases with the distribution of daily averaged temperatures in form of a QQ plot. However I do not see what information can be obtained from such a comparison, which is also not used in any further discussion.

The authors should clarify why these two distributions are expected to be related and what the intended interpretation is, or replace the figure with the ones in figure A4 (see next point).

With this plot our aim was to show that the two distributions are not the same, thus highlighting the complex relation between errors in AF forecasts and temperature, particularly for hot temperatures where AF forecasts would be most relevant for application in heat warning systems. We will add this in the text.

7. The central result that the forecasts underestimate AF at hot temperatures (lines 175-179) is visible only in figure A4, which is relegated to the appendix. This should at least be mentioned via a reference and/or by moving the figure to the main part of the paper.

Thank you for this suggestion, we propose moving this figure to the main text.

8. Several plots contain inconsistencies which should be corrected.

a. The values of the mean bias in figures 1c-f and figures A2a-d are different, however they should be equal. Further, the y axes in figures A2a-d do not show the whole data range which is from -0.4 to 0.4 as shown in figures 1c-f.

Thank you for highlighting this, upon inspection of the code we discovered an error where the y limits had been hard coded for Fig. A2, which has now been fixed, as shown below. We apologise for the mistake. Nonetheless, the values are still not the same between the two figures because for figure 1c-f we performed a 7-day rolling mean. In particular, we did not wrap this rolling mean since it was for a 3-month time series rather than a climatology. This means that the contribution of the first and last 3 days of the time series would lead to slightly different mean values between the two figures.

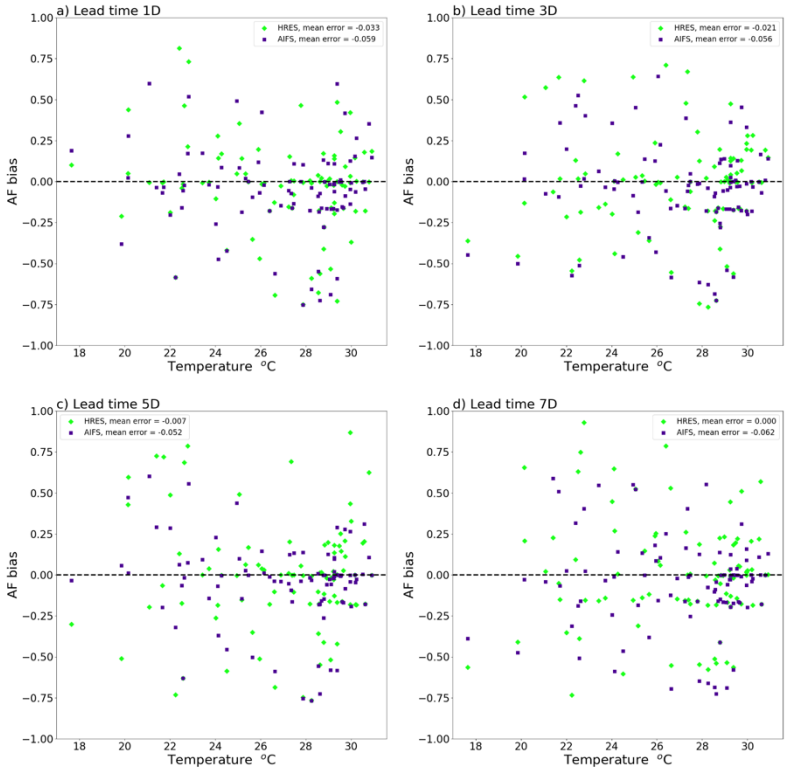


Fig. A2: Time series of the AF forecast bias for Rome during the summer of 2024, where green denotes the physics-based forecast (HRES) and purple denotes the data-driven forecast (AIFS) (a--d). Lead times 1 day (a), 3 days (b) 5 days (c) and 7 days (d) are shown.

b. The text explanation to figure 6 (line 152) mentions a deviation from the diagonal, however, the plot shows a horizontal line. If this really is a QQ-plot, a diagonal line should be added.

Thank you for pointing out this. We agree with the reviewer that the figure is not technically a qq-plot since it is centered around 0 and not a diagonal axis representing the ground truth (which is p.d. always 0 here, since we are looking at bias from the ground definition). The figure shows the AF forecast bias vs temperature (for each quantile), so we have proposed to remove the QQ-plot terminology, and rephrase the label to " population-weighted mean AF forecast bias vs temperature for lead time 1 day (a), 3 days (b), 5 days (c)

and 7 days (d). Green denotes the physics-based forecasts, purple denotes the data-driven forecasts (a–d). The black horizontal line denotes a bias of 0."

c. Figure A3 gives mean MAE values, which I assume are the mean values over all temperatures, however, from the shown data it is obvious that these values do not match the mean of the shown data. Is there data missing in the plot? If so, the axes should be changed such that all data is included.

We apologise for the error here (this was the same bug as mentioned in 5.1), and have now fixed this, as shown below.

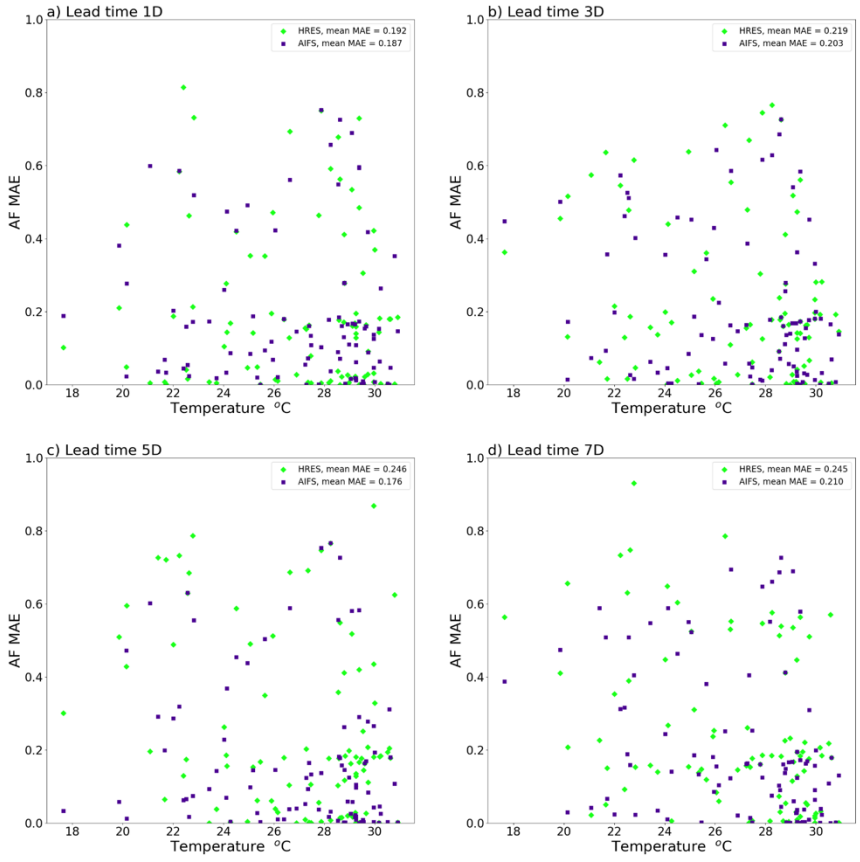


Fig. A3: scatter plot of the AF forecast MAE vs temperature for Rome during the summer of 2024; green denotes the physics-based forecast (HRES) and purple denotes the data-driven forecast (AIFS) (a--d). Lead times 1 day (a), 3 days (b) 5 days (c) and 7 days (d) are shown.

d. The data in figures A4a-d should be the same as in figures 3c-f. Therefore, the magnitude of the values should match, which is not the case. For example figure A4d has values between -0.1 and 0.25 while figure 3f has values between -0.05 and 0.12.

   Here Figure 4 refers to the population-weighted average, whilst figure 3 refers to the Rome case study.

Finally, the following list contains several purely technical issues.

1. In line 93-94 the order should be reversed to "difference between forecast and reference" to match the given formula 4.

   Thank you for highlighting this, we propose amending lines 93-94 accordingly.

2. Formula 4 uses j as index while the text and the following formula use i. This should be aligned.

   Thank you for highlighting this, we propose changing Formula 4 to use the index i for consistency.

3. Figures 1, 3, A2 and A4 describes the mean of the time series as "mean error", a name which is already used as alternative to "bias". It should thus be "mean mean error" or better "mean bias".

   Thank you for this suggestion, we agree that the term ‘mean bias’ would be clearer, and propose updating the manuscript accordingly (or to mean MAE where appropriate).

4. The caption of figure 1 misses a "line" in the last sentence.

   Thank you, we propose implementing this.

5. The caption of figure 3 has a wrong sub-figure labelling a-d -> c-f

   We apologise for the error, and propose amending the labelling accordingly.

6. Figure 5 has different y-scales for the four plots. For a better comparison these should be aligned.

   We propose implementing the hamonisation of y-axis limits across all panels.

7. For a better comparison the histograms in figure A1 should use the same width of the bins not the same number of bins for both forecasts.

   Thank you for this suggestion, we show the revised figure below with a fixed bin width:

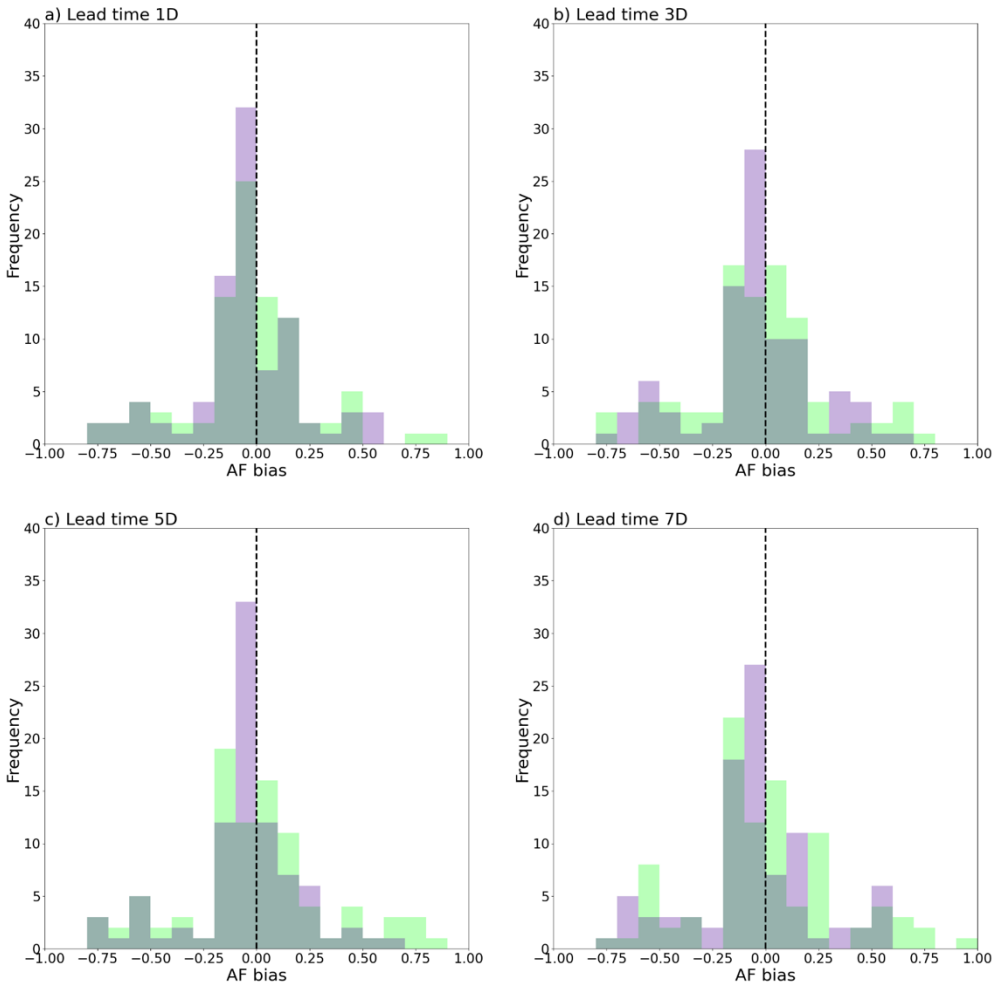


Fig. A1: Histogram of the distribution of AF forecast biases for lead times 1D (a), 3D (b), 5D (c) and 7D (d) for Rome. Green denotes the physics-based forecast whilst purple denotes the data-driven forecast.

8. The caption to figure A2 mentions a time series, but it is a bias-temperature scatter-plot.

The Reviewer is correct and we propose correcting the caption as follows:

Figure A2. Scatter plot of the AF forecast bias vs temperature for Rome during….

References:
Masselot P, Mistry M, Vanoli J et al. Excess mortality attributed to heat and cold: a health impact assessment study in 854 cities in Europe**.** *The Lancet Planetary Health* **7**, e271-e281 (2023). DOI: 10.1016/S2542-5196(23)00023-2