

Dear Prof. Rust,

Thank you for the opportunity to revise our manuscript, and the time that you have invested in handling our manuscript thus far. Please find enclosed our revised manuscript, file with tracked changes, and point-by-point responses to the Reviewers.

The most notable changes are, in accordance with the Reviewers' feedback, a more detailed description of the data used, and the replacing of one figure from the main with two figures from the Appendix. Additionally, we have made a number of minor text clarifications, and updated the formatting of some figures.

Yours sincerely,

Emma Holmberg and Leonardo Olivetti

Reviewer 1:

The manuscript addresses an important and timely topic, namely the development of impact-based forecasts for heat-related mortality. Overall, the manuscript clearly achieves its stated objectives and is well written, with a logical structure that makes the analysis easy to follow. I suggest some minor revisions, mainly related to clarification of certain aspects, as outlined below.

Thank you for taking the time to review our manuscript and your constructive feedback. Please find our point-by-point proposed revisions below. We further note that we have revised lines 80-81 of the revised manuscript to correctly state the epidemiological study design we build upon in this study.

1. The introduction clearly provides a clear overview of the motivation and research gap, it may benefit from a brief explanation of what is meant by “data-driven” versus “physics-based” weather forecasts. Since the journal attracts an interdisciplinary audience, including researchers working on climate impacts and hazards, a short clarification of these approaches could improve accessibility for readers who may be less familiar with weather forecasting.

Thank you for this comment. We have revised the introduction to include the following from lines 24-31 of the revised manuscript.

Currently, two main frameworks exist for generating weather forecasts. Physics-based forecasts refer to weather forecasts issued from numerical weather prediction models. These models develop predictions by solving numerical equations representing the physics of the atmosphere, as well as other components of our earth system such as oceans, and have been the core framework for producing weather predictions for the past decades. Data-driven forecasts refer to weather forecasts where techniques from the field of statistics and artificial intelligence have been used to train weather models based on large amounts of climate data, typically in the form of reanalysis datasets such as ERA5. The rapid development of these data-driven weather models contrasts the so called ‘quiet revolution’ of traditional, physics-based numerical weather prediction (NWP) models (Bauer et al., 2015).

2. Section 2.1: the source of the reanalysis is not specified when first introduced (line 55), and only later in the paragraph it becomes clear that ERA5 is used. For clarity, the specific dataset (ERA5) could be mentioned at the beginning of the paragraph.

Thank you for highlighting this, we clarified the use of ERA5 on line 60 of the revised manuscript.

3. Section 2.1: The authors state that the meteorological data were aggregated to daily resolution. It would be useful to clarify the choice of daily mean temperature rather than

daily maximum temperature (Tmax), which is commonly used in studies examining heat-related health impacts.

Daily mean temperature was used in the study of Masselot et al (2023) on which we base our results, and has also been used extensively for heat-related mortality studies (Vicedo-Cabrera et al. 2021; Ballester et al. 2023; Janos et al 2025, e.g.). We now clarify that we follow Masselot et al. (2021) in using daily mean temperature on line 63 of the revised manuscript.

4. Section 2.1: The authors also state:

"We also use 2m temperature from archived forecasts from two different types of weather prediction models, one physical model (IFS HRES cycle 48r1), and one data-driven model (AIFS single v1)." It would be helpful if the authors briefly explained why these specific models were selected.

We have added the following subclause in the highlighted passage (lines 70-71 of the revised manuscript):

'..., which were the model versions available at the time this analysis was performed'.

5. Section 2.1: The manuscript states that the analysis focuses on a selection of European cities. However, the specific cities included are not clearly listed or illustrated. Providing a table or a map showing their spatial distribution across Europe would improve transparency and help readers understand the geographical area of the analysis.

We clarified the geographical focus adding a map of the cities in the appendix (New Fig. A1, shown below) and as follows for the sentence on line 66 of the revised manuscript:

Specifically, this study focuses on a selection of 854 cities across 30 European countries further detailed in Section 2.2.

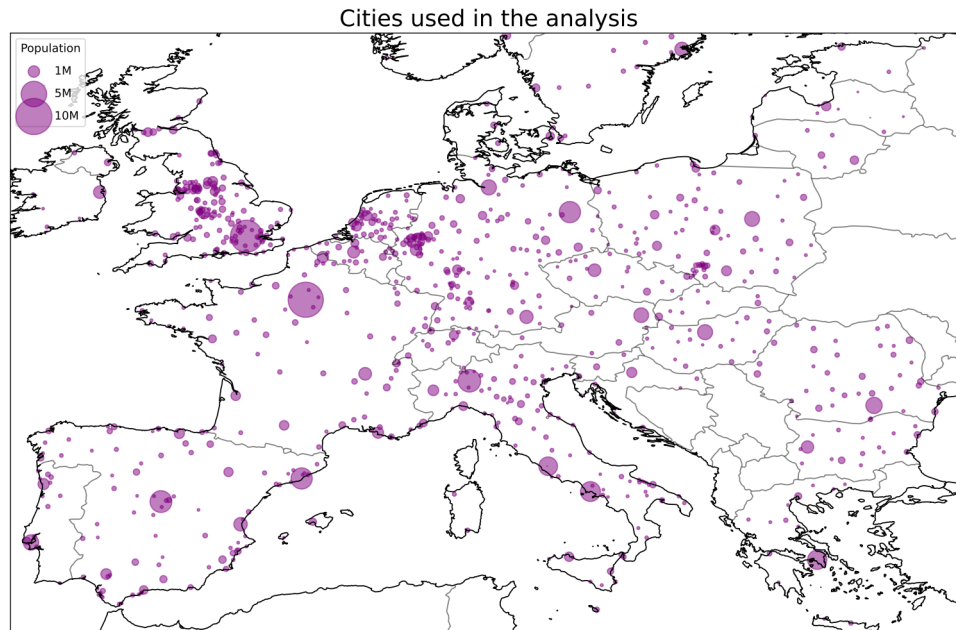


Figure A1. Map of the cities included in this study, with marker size indicating the population.

6. Section 2.2:

The epidemiological framework relies on exposure–response functions from the multi-city analysis of Masselot et al. (2023), which is based on **all-cause mortality data**. It would be helpful to state this explicitly, as the estimated impacts therefore correspond to temperature-related mortality across all causes rather than specific cause-of-death categories (e.g., cardiovascular or respiratory mortality which are the most common causes that are used and studied in health impact studies). A brief clarification regarding this would help.

We amended line 77 of the revised manuscript to state explicitly that we focus on all-cause mortality.

7. Section 3: In lines 111–112 the authors state that the continental-scale assessment uses population-weighted averages. However, the population dataset used to compute these weights is not specified. For transparency and reproducibility, the authors could clarify the population data source (dataset and spatial aggregation).

Thank you for highlighting this, we apologise for the missing information. We clarified the city-level population data source on lines 123–124 of the revised manuscript as that provided together with the temperature-mortality associations by Masselot et al. (2023).

8. Section 3.2: Figure 6 (“QQ plot of population-weighted mean AF forecast bias vs temperature...”) appears to be either mislabeled or incorrectly described. The figure is referred to as a QQ plot in both the caption and the text (lines 151–159), yet the axes

represent temperature and AF forecast bias rather than quantiles of two distributions. Consequently, the reference to deviations from a diagonal line is unclear. The authors may wish to clarify or revise the description of the figure.

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). In relation to comments also from Reviewer 2 we have opted to remove the figure and replace it with a Kolmogorov-Smirnov test to compare the two distributions.

9. Section 3.2: It could also be informative to distinguish heatwave days from typical summer days and assess forecast performance separately, as this may provide additional insight into how well physics-based and data-driven forecasts capture the attributable fraction of heat-related mortality during extreme heat conditions. I understand that this may not be possible in this type which focuses only on the summer of 2024 with limited data, so this is just a suggestion for future work including multi-year analysis.

Thank you for the suggestion, we added this as a point to the discussion on potential further extensions on lines 238 of the revised manuscript.

... or extended to focus specifically on heatwaves.

10. Section 4: The continental-scale population-weighted analysis provides a useful summary of forecast performance across Europe. However, this approach may also hide some regional differences in forecast skill. Given that temperature-mortality relationships vary substantially across Europe (Masselot et al. (2023)), future work could explore city-level or regional patterns of forecast performance. In addition, examining cause-specific mortality rather than all-cause mortality may help identify where data-driven and physics-based forecasts perform better or worse under different climatic and vulnerability conditions. While this point is briefly mentioned in the discussion, it could benefit from stronger emphasis.

Thank you for highlighting these points, we expanded the sentence on lines 238-240 of the revised manuscript as follows:

A further extension of the work presented here would be to focus on specific subregions within Europe, as well as regions beyond Europe, particularly those that have been highlighted as being especially vulnerable to climate change.

Regarding cause-specific mortality, we would like to highlight that this study was based on temperature-mortality associations published by Masselot et al. (2023). We are not aware of a similar database for cause-specific mortality, but would warmly welcome openly accessible data of such nature. We expanded lines 241-243 of the revised manuscript as follows:

The association between heat and health outcomes beyond all-cause mortality is an active area of research within the health community, and we warmly welcome efforts to increase accessibility to this data.

References:

Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015). <https://doi.org/10.1038/nature14956>

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

Vicedo-Cabrera, A.M., Scovronick, N., Sera, F. *et al.* The burden of heat-related mortality attributable to recent human-induced climate change. *Nat. Clim. Chang.* **11**, 492–500 (2021). <https://doi.org/10.1038/s41558-021-01058-x>

Ballester, J., Quijal-Zamorano, M., Méndez Turrubiates, R.F. *et al.* Heat-related mortality in Europe during the summer of 2022. *Nat Med* **29**, 1857–1866 (2023). <https://doi.org/10.1038/s41591-023-02419-z>

Janoš, T., Quijal-Zamorano, M., Shartova, N. *et al.* Heat-related mortality in Europe during 2024 and health emergency forecasting to reduce preventable deaths. *Nat Med* **31**, 4065–4074 (2025). <https://doi.org/10.1038/s41591-025-03954-7>

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. We further note that we have revised lines 80-81 of the revised manuscript to correctly state the epidemiological study design we build upon in this study.

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 clarified the sentence on line 6 of the revised manuscript 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 (lines 71-72 of the revised manuscript). 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.

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 clarified the use of ERA5 at the beginning of section 2.1. We clarified here only the use of ECMWF forecast data because there is a larger quantity of information conveyed regarding this choice later in the paragraph, and we did not want to break the flow of the passage by moving all of this information to be earlier in the paragraph.

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.

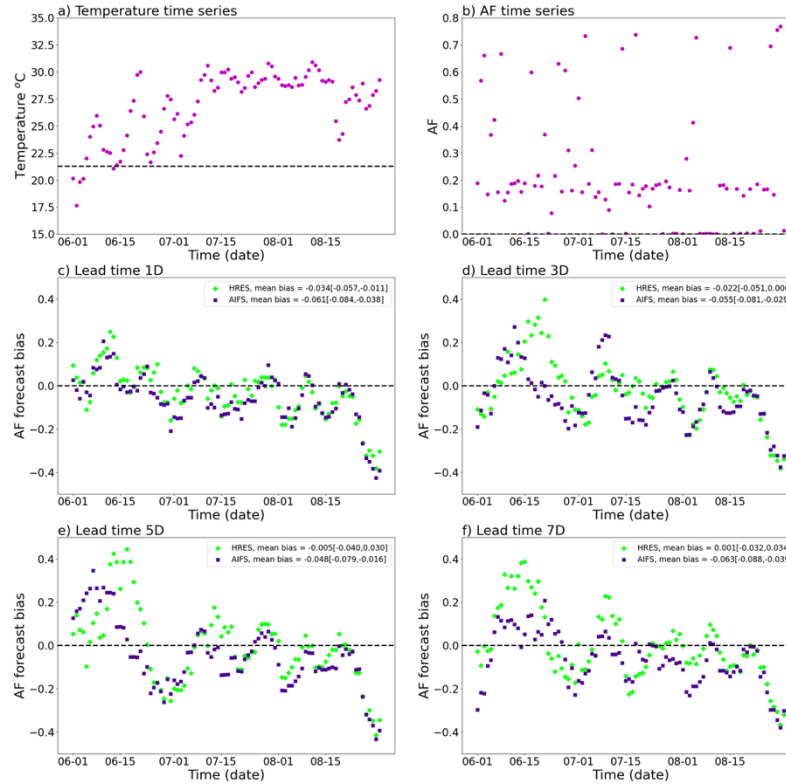
Thank you for highlighting this, we apologise for missing this point. The population data used was that published by Masselot et al. (2023) together with the temperature-mortality associations, which we clarified on lines 123-124 of the revised manuscript. Additionally, we have added a map showing the cities included in the study in Appendix A (Figure A1).

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 cautious about including 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.

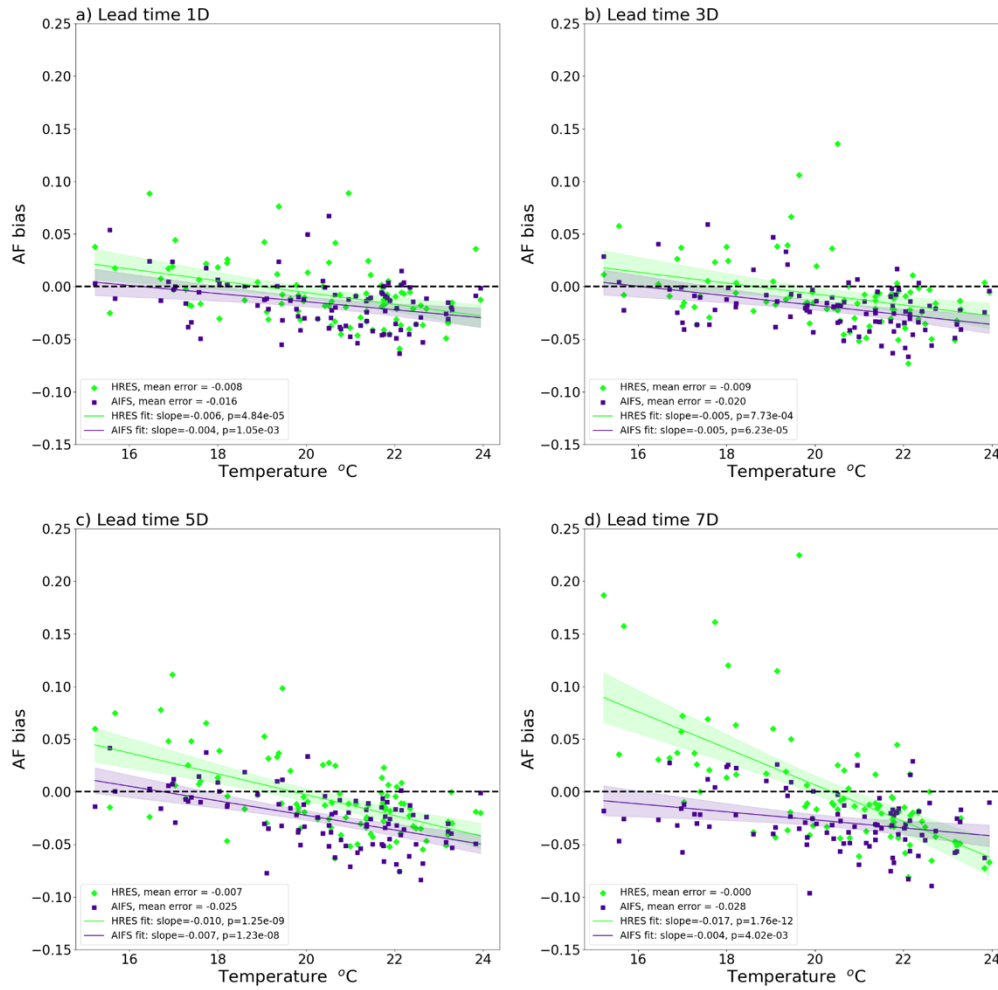
- c. The comparison in line 156 mentions a Kolmogorov-Smirnov test whose results should be included for transparency.

Following this comment, we have now included the results for the tests in the sentence on line 162-165 of the revised manuscript as follows:

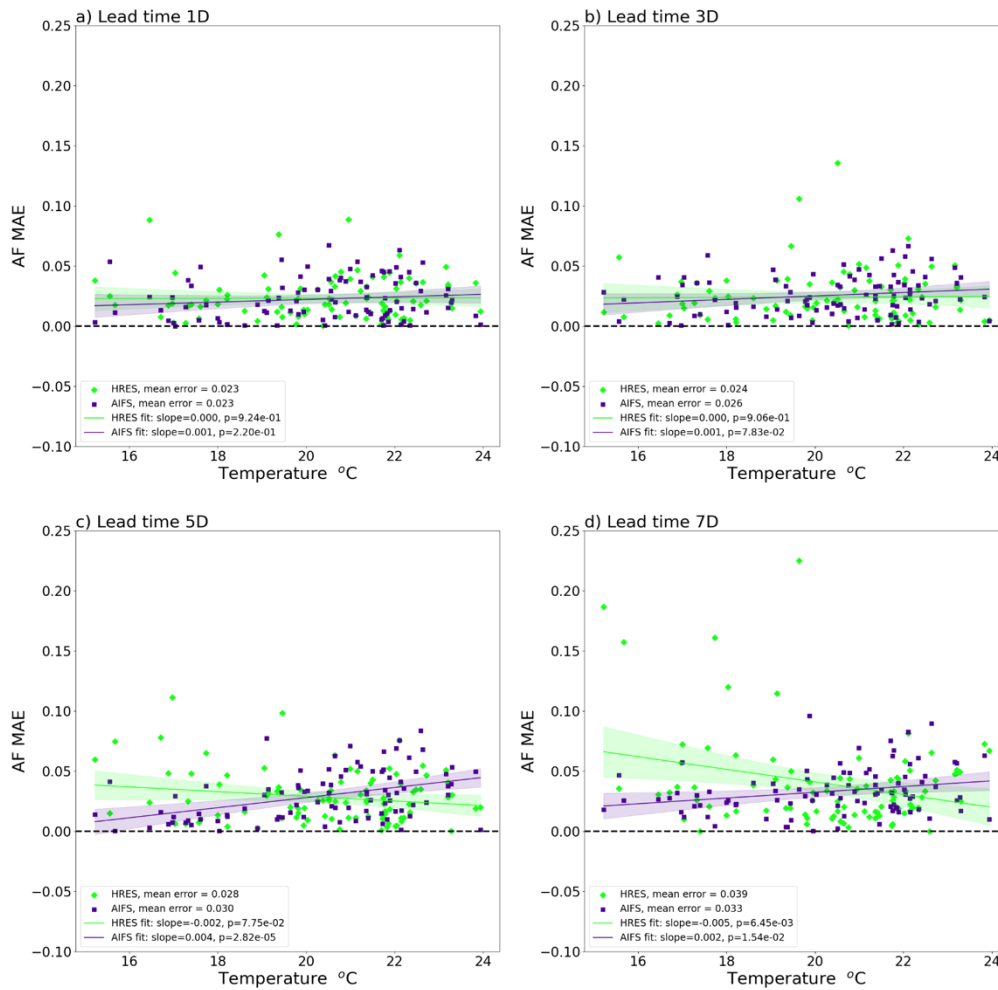
To investigate the relationship between AF forecast bias and temperature further, we compare the distributions of AF forecast bias of the two models by performing a Kolmogorov-Smirnov test. The results indicated that for lead times 3, 5 and 7 days, the physics-based and data-driven AF forecasts correspond 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 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. Please note that we have moved these figures to the main in light of another comment from the reviewer.



Revised Fig. 6: 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 and the shading denotes the 95% confidence interval.



Revised Fig. 7: 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 and the shading denotes the 95% confidence interval.

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).

In light of the addition of the Kolmogorov-Smirnov test and moving two figures to the main text, we have opted to remove Fig. 6 of the original manuscript and replace it with a Kolmogorov-Smirnov test to compare the two distributions. We believe that the information that we wanted to convey is now conveyed a clearer manner by other figures, and thank the reviewer for these helpful comments.

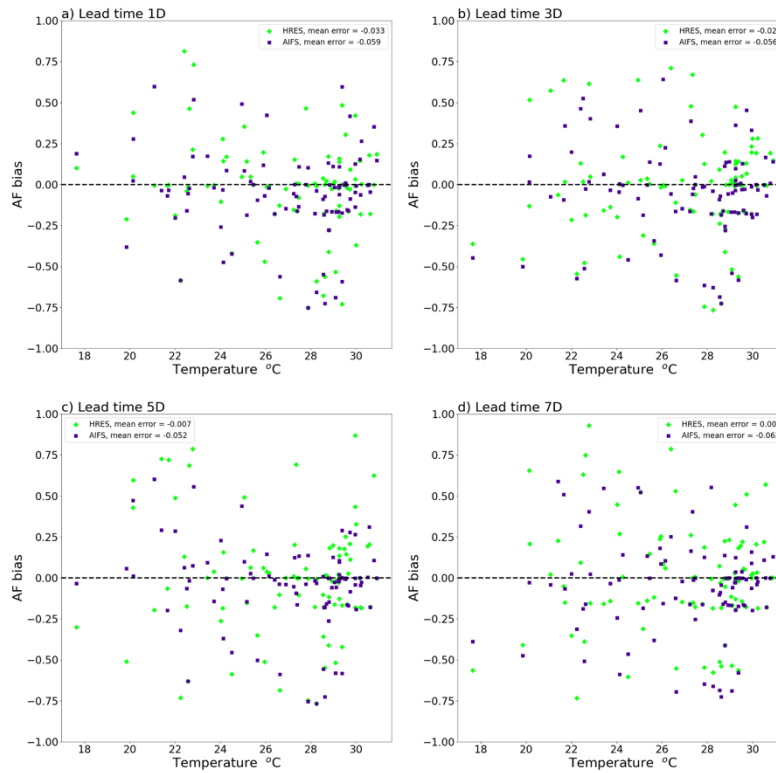
7. The central result that the forecasts underestimate AF at hot temperatures (lines 175-179) is visible only in figure A4, which is relegated to the appendix. This should at least be mentioned via a reference and/or by moving the figure to the main part of the paper.

Thank you for this suggestion, this figure and its partner figure, Fig. A5 have been moved to the main text (Figs. 6,7 of the revised manuscript).

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. A3 of the revised manuscript, 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.



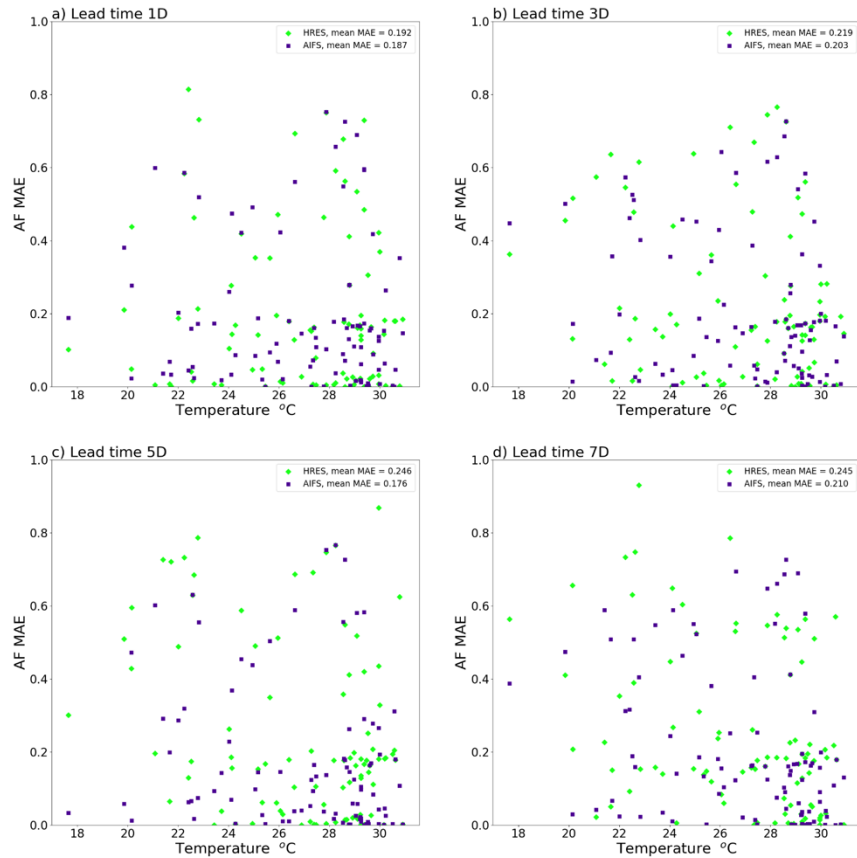
Revised Fig. A3: Time series of the AF forecast bias for Rome during the summer of 2024, where green denotes the physics-based forecast (HRES) and purple denotes the data-driven forecast (AIFS) (a-d). Lead times 1 day (a), 3 days (b) 5 days (c) and 7 days (d) are shown.

- 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.

In response to a previous comment of the reviewer, and a comment from reviewer 1, we have removed Fig. 6 from the manuscript.

- 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.



Revised Fig. A4: 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.

As for comment 8a, Fig. 3 of the revised manuscript has been smoothed with a 7-day rolling mean whilst Fig. 6 of the revised (A4 of the original) manuscript has not, leading to a difference in the plotted values.

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 amended lines 103-104 of the revised manuscript 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 changed 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 updated 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 implemented this.

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

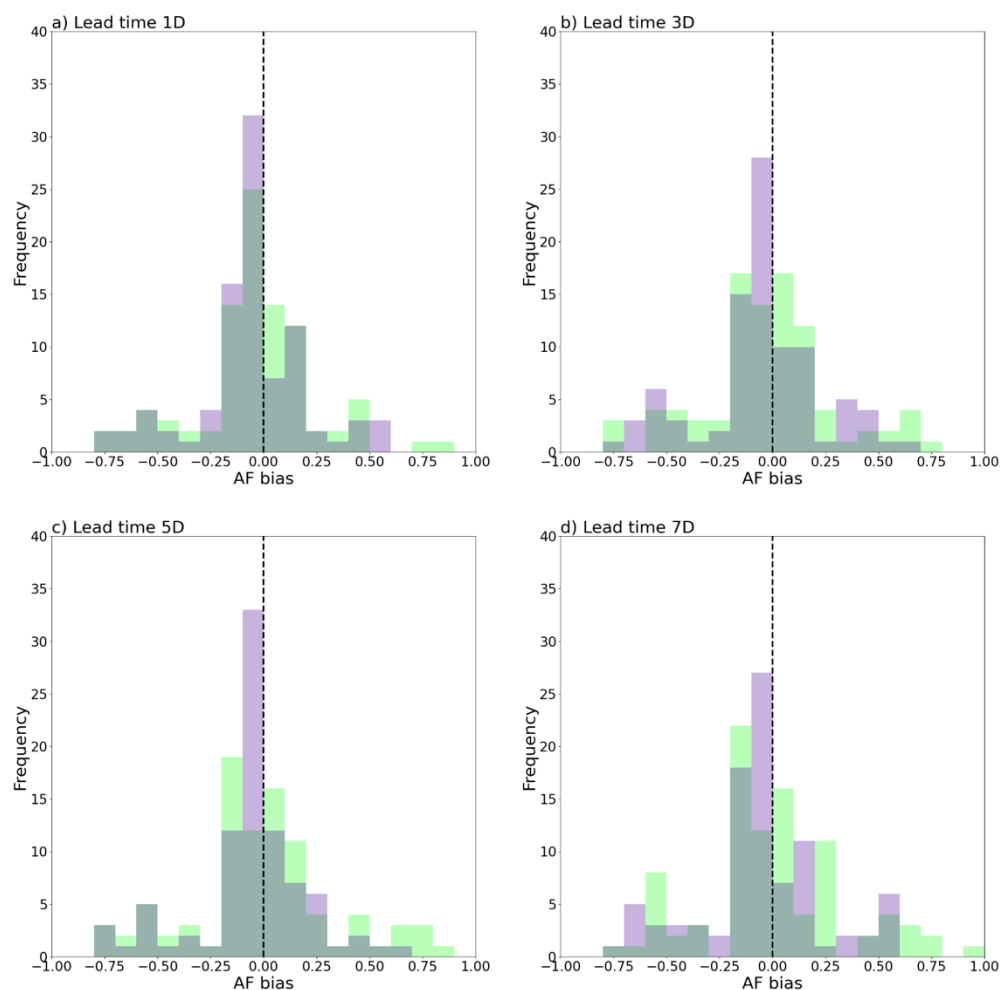
We apologise for the error, and amended the labelling of Fig. 3 accordingly.

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

We hamonised the y-axes limits across all panels of Fig. 5.

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:



Revised Fig. A2: 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 corrected the caption of Fig. A3 of the revised manuscript as follows:

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

