Authors' response to Referee #2 (Report 3)

Referee #2 (RC2):

Review nhess-2024-75-3 "The probabilistic skill of Extended-Range Heat wave forecasts in Europe"
Thank you for the responses and clarifications to my comments and questions. As before, I think the analysis you carried out is interesting and relevant. I also think that after your latest revisions, the method and all technical aspects of your analysis have become very clear. I still think the discussion part and to some degree the introduction could be improved in terms of readability, but I see these as minor points and leave it up to the authors to decide how much they would like to address these.

We appreciate your constructive comments as they further helped to enhance the quality of our manuscript. The following are point-by-point answers in blue colour:

Some more detailed comments:

ll. 43 – 52: This paragraph is overly specific. It is fair to cite your studies here since they are somewhat relevant to an aspect of this paper, but the level of detail is unnecessary in my opinion, as it distracts the reader from the main message. E.g.,
l. 43: Add something like "As an example, during heat waves, apartments… ". You go from listing heat wave impacts very generally to something extremely specific (buildings in the Nordics) here.
l. 45 – 46: I think this level of detail is not necessary here and this sentence could be dropped.
ll. 49 – 50: This sentence could be dropped, too. I don't see it adding any relevant reason for why to consider longer term temperature averages. If anything, doesn't it make the case weaker?
Thank you for these comments, we have removed these suggested parts:
l. 45-46 "The warm-up time of buildings related to outdoor temperature depends on building properties (U-value, ventilation airflow rate, and thermal mass of buildings)."
and ll.49-50 "In not well-insulated buildings and/or light structures, such as wooden ones, the warm-up time can be significantly shorter, often only 1–2 days."

ll. 84 – 86: To me, this sentence just says, that a lot of people (in which case there should be more than one example study) have written about this topic. I think it can be dropped.
Thank you for this comment, we have dropped this sentence: "There is a large literature in statistics and decision analysis on the use of probabilistic information in so-called decision making under uncertainty (e.g. Clemen 1996)."

ll. 90 – 102: I think this paragraph gets slightly too specific on some details of the analysis and could rather give a more general overview over what you're doing in the paper (for which the motivation should be covered at this point of the intro).
ll .95 – 97: You have motivated this already earlier (ll. 43 – 52), I don't think it needs to be motivated again.
Thank you for these comments, we have dropped this suggested sentence from ll 95-97: "Moreover, in an empirical study conducted in Finland, indoor temperatures were found to be more strongly correlated with outdoor 5-day moving average temperature than with average temperatures of a few days only, suggesting impacts of building's thermal inertia (Velashjerdi Farahani et al., 2024b)."

Figure 2, ll. 176 – 180: In (d) you have one year less, so doesn't it make more sense to show events per year instead of total number of events?
Thank you for this remark, however, here we are leaving this comparison as it is.

l .186: initiated → initialized

l. 192: “which are *the* same”
l.195 – 196: “runs *per summer/year*”
l. 203 – 205: Could replace these two sentences by: “We therefore *arbitrarily* decided to use only the Monday runs.”
l. 217: “*daily* mean”
ll. 226 – 227: I think this could be dropped, since it has been clearly defined already.
l. 230: Could mention here that this difference does not influence your verification because you consider model-specific thresholds.
l. 242: Remove “Moreover,”
l. 255: remove “, *p,*”
ll. 258 – 259: “probability (of the heat wave day)” → “probability of a heat wave day”
l. 259: good one :)
l. 281: “greater *than* zero”
Thank you for these remarks, we have corrected these as suggested.

l. 283: “the results” → would rather say the significance/importance of the results
Thank you for this remark, we agree, and we have edited this to be “the significance of the results”.

ll. 316 – 317: “for lead times of 2 weeks (and longer) there are far fewer samples” → I would say this is true for all lead times. Except maybe week 1 p = 0.9, although also those are far fewer than p = 0.1.
Thank you for this remark, however the text has been left as it was, as the forecast week 1 stands out with a larger n also in Figure 5a in the category p>0.66 (in comparison to other weeks’ n in Figures 5b-d.

ll. 356 – 357: could maybe explicitly say that here, you transform the probabilistic forecast to a categorical one. I find this quite appealing by the way, because I think this is a much more likely scenario of how such a forecast would be used in practice. I think it could also be fair to mention that.
Figure 5: If you want to, there is a lot more that you could unpack from this figure and the way you look at the forecasts here (i.e., as categorical forecasts). While in this figure, it is easy to see what the outcome was, given a certain forecast (category) was issued, it could be quite insightful to see what the forecast was, given a certain outcome (heat wave/no heat wave). You can actually do this (approximately) with the numbers in the plot. For instance, the forecasts for all lead times have a high likelihood (decreasing with lead time but > 90% for all) p(f0|o0) of forecasting no heat wave when there turned out to be no heat wave, i.e. they have high specificity. However, when you wrap up the (0.33 < p < 0.66) and (p > .66) categories into one “forecast of heat wave” category, the true positive rate (sensitivity) p(f1|o1) drops from 86% in week 1 to 19% in week 4. In addition, the likelihood p(f0|o1) of forecasting “no heat wave” when there was actually a heat wave, goes up from 18% in week 1 to almost 99% in week 4. This means that if a forecast at longer lead times indicates “heat wave”, there are good chances there will be one, but if it shows “no heat wave”, you shouldn’t rely on there being none. In my eyes, this could be quite important information to someone designing an early-warning system. I leave it up to you, but in my opinion something like this would make a very nice addition to the paper and it could add some points that could be addressed in the discussion.
Thank you for these remarks, you are making good points. We have edited this to be “Next, we conducted verification of heat wave day forecasts based on forecasted probabilities falling within the ranges of here defined as low: p <0.33, intermediate: 0.33 ≤ p ≤ 0.66, and high: p > 0.66, i.e., we transformed the probabilistic forecast to a categorial one.”

l. 415: “(days 29..35)” → “(days 29 to 35)”?
Thank you for this remark, we have edited this to be “(days 29 to 35)”.

l. 494: Does ERF stand for “extended-range forecast”? Don’t think this is defined anywhere.
Thank you for this remark, we have added the definition for ERF here.