

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.

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?

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.

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.

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?

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”

l. 283: “the results” → would rather say the significance/importance 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 $p = 0.1$.

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) $p(f_0|o_0)$ 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) $p(f_1|o_1)$ drops from 86% in week 1 to 19% in week 4. In addition, the likelihood $p(f_0|o_1)$ 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.

l. 415: “(days 29..35)” → “(days 29 to 35)”?

l. 494: Does ERF stand for “extended-range forecast”? Don’t think this is defined anywhere.