# Review nhess-2024-75-2 “The probabilistic skill of Extended-Range Heat wave forecasts in Europe”

I appreciate the effort that you have put into addressing my comments and I think parts of the manuscript have improved. As stated in the previous round, I find the results interesting and relevant and the study worth publishing. In general, however, large parts of the paper are still difficult to comprehend and lack explanations and motivations. For clarity and readability, the manuscript requires some major revisions. I try to outline the major issues that I still see and what I think you could do about them below.

## Major general remarks:

*Introduction*

In my opinion, the paragraphs in the intro need to be linked more, both to each other and more explicitly to the specifics of the paper. It is currently difficult to see why certain things are mentioned where they are mentioned.

For instance, ll. 41 – 45 go into some details on the effects of heat waves on buildings in Northern Europe. When I read this for the first time, it seemed like an overly specific example, but I think that what you might mean to do here is to argue for why it makes sense to look at longer term averages of temperatures (l. 44: “heating of buildings has been observed to take 5 – 6 days”). If this is the case, please be more explicit about it. This applies to other parts of the intro, too.

Another example is the transition from l. 59 to l. 60. You nicely explain why early-warnings systems are needed and what general systems are in place. If you add one sentence on why “even earlier” warning systems (i.e. based on subseasonal forecasts) could be beneficial, you make a transition to introducing the extended range forecasts.

*Methods*

I like the new Table 1, which makes it very easy to grasp the structure of the data being used for the verification. I also appreciate that you took my suggestion of moving the part that relates solely to the verification data into the methods section. However, I think this section’s comprehensibility would benefit strongly from some major re-structuring. I think most parts of the text are there, and they could be re-arranged and slightly re-written. I recommend the following to improve the readability, but I acknowledge that this is somewhat subjective:
Try to go from the most basic to the details. To me, the definition of a heat wave (day) is the most essential and basic piece of information in the context of the paper*. It should be put first in the method section along with your explanation for why this is a meaningful threshold (ll. 161 – 166). Mention that with your definition, you transform a continuous variable (temperature) into a binary variable, which is your forecast target. Then you could introduce what you assume as reality/ground truth/verification (namely,

ERA5, maybe commenting quickly on shortcomings of reanalysis in representing reality) and go into some detail on what heat waves defined in this manner look(ed) like and how robust the definition is (ll. 167 – 175). Here, you could also use what is now Fig 2 (and potentially Fig. 1a or even 1a-c) and discuss the "outlier" 2010 and why it deserves some special attention.
I would only then move to the forecasts. Introduce the model and the ensemble system set-up. Then explain how the "extra" time dimension (lead time) is treated when defining heat waves days and that thresholds are defined with respect to the model climatology. Explain how you go from an ensemble forecast (essentially a "collection" of deterministic forecasts) to a probability forecast. Finally, you can talk about how to verify them (current section 2.3).

*I'd picture something like ll.73 – 74 but introducing variables such as $T^{5d}$, $T^{5d,90}$ and saying that you only include land areas.

Figure 1: From the current text, I'm not sure why this figure is shown (except maybe a-c which could be used as I indicated above). The main point of using a model-dependent threshold to define heat waves in the forecasts is to avoid any issues with differences in this threshold between models and re-analysis. So, what do you conclude from the fact that they are basically the same? What would be different if you saw that the $T^{5d,90}$ were very different in forecasts vs. re-analysis? If I'm just missing the point here, please give more concrete conclusions from this figure in the manuscript.

*Discussion*

I think the discussion should be extended. It is fair to say that the results are in line with other studies, but is this expected or unexpected given the employed methodologies/approaches? What are possible limitations of your study, where could it be extended (you mention this a bit in ll. 407 – 409) and why is it nevertheless important as it is? I like that you dedicate a section to the potential added value of probabilistic forecasts, but in its current form this section is for the most part a short literature review (ll. 411 – 424), which is better placed in the Intro. Ll. 423 – 429 go in the right direction in my opinion. Additionally, are there maybe examples of events where it is thought that even earlier warnings would have been beneficial in mitigating some of the effects of a heat wave? How important are things like the spatial resolution and temporal aggregation in this context? You mention the relevance of 5-6 day temperature averages for Northern Europe, but is this valid in other countries that might have a completely different building stock?

## Further comments:

Title: mix of capitalized and lower-case words

l. 19: in extended range → in the extended range

l. 27: "persistence [...] seem to have" → "persistence [...] seems to have a"

ll. 69 – 70: I think this last sentence might be better placed in the discussion.

l. 75: “have” → “has”

l. 106: “initiation” → “initialization”

l. 122 “capture” → “skillfully predict”

l. 141: “forecasting in the model’s climatology” I don’t quite understand how this is meant. Is it just to say that a heat wave in the forecast is defined relative to the forecast model’s climatology? Maybe you could reformulate.

l.144 (also see my earlier comment from the first round on l. 134): As you correctly say here, you are implicitly bias-correcting the hindcasts. Since you are not leaving out the year for which you forecast (this year would not be available in a real forecast, because it has not happened yet), this is a better correction than you could ever have access to in reality. This will lead to an overestimation of the skill (although the larger ensemble in forecast mode might counteract this). You do show that the 90$^{th}$ percentile does not change much in absolute terms when leaving out the most severe events, so it’s reasonable to assume that the effect on skill is not huge, but this does not mean that there is no effect. In conclusion, I think you should mention this point, as it is generally agreed upon that S2S hindcast verification should be done in a leave-one-out manner.

l. 175: It would be ideal to end this paragraph with a sentence about what you conclude from these statistics.

ll. 211 –220: would be good to add a subscript or something to distinguish the forecast probability from the base rate (currently, you call both $p$).

l. 220: “base rate of $p$” → “base rate $p_b$” (or whatever else you will call the base rate)

l. 236: “the BSS $n$ times (here $n$ = 5000)” → “the BSS 5000 times” (no need to define $n$ if you never use it again)

l. 243: “the FDR controls for the expected proportion of false discoveries” I don’t quite understand what this means. Isn’t the FDR the proportion of false discoveries? Could you maybe reformulate?

l. 245: Thanks for adding an explanation on the B-H procedure. It is still not entirely clear to me why this is necessary in addition to the p-value adjustment. Could you elaborate?

l. 267: “a heat wave days” → “heat wave days”

l. 273: “shorter forecast weeks” → “shorter lead times”

l. 275: It could be noted here again that there are a lot less samples in the higher probability bins (expect for maybe lead time 1 week, bin 0.9 – 1), so those points are a lot more uncertain.

l. 280: “of the all hindcasts” → “of all the hindcasts”

l. 295: add comma after “In the second week”

l. 299: what do you take from this analysis? Are the results strongly influenced by the longest heat wave (or 2010)? It looks to me like the skill in weeks 2 – 4 is systematically lower when the longest heat waves are excluded with only few exceptions. In most areas differences are small, so maybe it only really matters in Eastern Europe/Russia (skill goes from being significant to being not significant). Also, I think the left and middle rows might be enough to show. Or what extra information do you gain from excluding 2010 everywhere? If you decide to show it, you should discuss it more.

l. 300: “In the Figure 4” → “In Figure 4”

l. 305: two commas at the end of the line

Section 3.3: I appreciate that you included some more background on why to look at the forecasts in this way. However, I am still a bit confused about what we learn from this plot, so I think it would help to add what you are concluding from this analysis at the end of the section. You say the point is “to assess the severity of the over- or under-forecasts”. So, based on these plots, how severe is it for different lead times? Is it possible to relate this type of evaluation to any of the fundamental properties of a forecast, e.g. is it related to discrimination or resolution (in the forecast verification sense)?
I’m also thinking about the bins/categories you use. Now, they are basically: extremely elevated likelihood ($p > 0.66$), strongly elevated likelihood ($0.33 < p < 0.66$) and everything from moderately elevated likelihood to lower likelihood ($p < 0.33$) for having a heat wave. I think it would make this plot a lot more interesting if the forecasts were split at $p = 0.1$ and $p$ was expressed relative to the base rate. Below that threshold, forecasts indicate a lower-than-normal likelihood of a heat wave and above, they indicate a higher likelihood. You could have one “basically normal/climatological likelihood” category with something like $0.05 < p < 0.2$ and one below (reduced likelihood) and one above (increased likelihood). Also, the statement “we are 5 times more likely than normal to have a heat wave in week X” has a very different psychological effect than saying “the chances of having a heat wave in week X are 50%”, which makes me think it might be more interesting to see $p$ relative to the base rate.

L. 321 – 324: I think this is of little help in understanding the plot, because your perfect forecast would only ever issue $p=0$ or $p=1$, which is a very hypothetical situation. Could you rather say what a good (but not infinitely sharp) vs. a poor or no-skill forecast would look like?

l. 322: “p<0.33 be” → “$p < 0.33$ would be”

l. 339: “amount” → “fraction”

l. 343: “the relative time of forecast issuance and heat wave initiation” do you mean the forecast initialization (date) relative to the onset of the heat wave?

l. 344: “corresponsive” → “corresponding”?

l. 357 – 362: Some statements from ll. 353 – 356 are repeated verbatim here. Is it an option to wrap these into one, as in: “In both forecast week 1 and 2, there is …”

l. 371: “indicating ongoing heat” → “indicating an ongoing heat wave”

caption Figure 6: add what *n* and the width of the boxes mean. Also see my comment on Section 3.3/Figure 5: maybe an option to express p relative to the base rate?

Figure 7 and ll. 384 – 387: I’m not sure this figure is needed. What extra information does it provide? The main difference is that there remains hardly any data in the “29+ day inside the heat wave” categories, which just shows that there is basically no event inside the sample that is as long-lasting as 2010. But this does not really tell us any more about the forecasts. Since the differences are small in the categories that are well populated, you could just mention that the differences are negligible.

ll. 393 – 397: This is a short recap of the method. Why is it relevant in the context of the discussion here? Is there some relation (similarities/differences) to the methods used in the papers you cite in l. 398?

l. 403: “the best of the forecast skill seems to come from the longest period of heat wave days” Do you mean that the skill in forecasting heat waves decreases when excluding the event?

l. 439: would add here that this significant skill largely vanishes when 2010 is excluded (Fig. 4)

l. 448: “its” → “heat wave occurrence”

l. 448: “further” → remove