Authors' response to both Referee #1 and Referee #2 (Report 2)

Referee #1 (RC1):

Review for "The probabilistic skill of Extended-Range Heat wave forecasts over Europe", by Korhonen et al.

General comments:

I acknowledge the efforts of the authors to tackle all the issues raised by the reviewers, and improve the manuscript accordingly.
I still find the lengthy discussion part in lines 411-424 quite odd. In my view, it would better fit the introduction section, since it does not discuss the results found but rather explains the motivation to carry out this study. However, I would not make a strong case for another round of revision, and I suggest the paper be accepted for publication.
I am listing below a few typos found in the revised manuscript.

Typos:
L. 280: typo "all the hindcasts"
Fig.4 : typo in the title of the second column: "longest"
L.349: typo "if the forecasts were perfectly.."

Response: Thank you for these comments. We have corrected the mentioned typos and moved lines 411-424 in a shortened form to the Introduction section.

Referee #2 (RC2):

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.

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:

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.

Thank you for these comments. We have now made several modifications to the introduction to link the paragraphs to each other, point out why the things are mentioned where they are, and motivate the use of 5 days mean temperature.

First, we added to the end of the first paragraph of the Introduction (now lines 35-37): "This growing occurrence of heat waves underscores the urgent need to understand their dynamics and improve forecasting methods, especially for prolonged events with severe impacts."

l.41-45 was edited as (now lines 46-52): "In Northern Europe, where apartments are typically not equipped with mechanical cooling systems, the thermal inertia of buildings plays a critical role. For instance, a Finnish study observed that buildings required 5-6 days to reach overheating conditions, highlighting the importance of the 5-day mean temperature as a predictor for indoor heat stress (Velashjerdi Farahani 2024a). 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. These findings emphasize the relevance of forecasting tools capable of predicting not only the occurrence but also the persistence of heat waves.”

l. 51 was connected to the heat-health action plans by (now lines 57-59): "Recognizing this, many countries in Europe and other parts of the world have developed heat-health action plans over the past 20 years to mitigate heat-related health risks (Kotharkar et al. 2022; Martinez et al. 2022)."

l. 59-60 we added (now lines 68-70): "Extending these lead times could significantly enhance preparedness by allowing for earlier adaptive measures and better resource allocation, particularly for prolonged heat waves."

and we start the next paragraph by (now lines 72-73): "Sub-seasonal forecasts, which cover the extended range of 2 weeks to 1 month, offer a promising avenue for improving early warning systems."

Parts of the discussion ll. 411-424 have been placed into the third paragraph of the introduction, now lines 59-65 (and these lines have been completely removed from the discussion):
“As health effects of heat exposure occur quickly, at the same day or a few days lag (Baccini et al. 2008), it is imperative that the protection measures are implemented rapidly when a potentially dangerous heat wave is forecasted. However, organization of the response measures requires coordination of actions between many stakeholders and distribution of workforce, equipment, and other resources, which take time.”

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 $T5d$, $T5d$,90 and saying that you only include land areas.

Thank you for these recommendations to improve readability of the Methods. We have now strongly re-arranged and where needed re-written the text so that:
The Definition of the heat wave days is now first, in Section 2.1.
Thereafter, in Section 2.2. we introduce the ERA5 data (together with its Fig 1a and Fig 2).
Next, in Section 2.3 we move on to the hindcasts and probabilistic forecasts, and Section 2.4 introduced the used skill scores.

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 $T5d$,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.

Thank you for this comment. We have now included contents of the Figure 1 in the following text added in Section 2.3.2 (this text was related to the comment on l.144 also) now lines 245-250:
"In the verification, the forecast model-based probability of a heat wave day, p, was compared to the observed heat wave days (Section 2.2.1) derived from the ERA5 dataset. Since we used the data from the entire period (years 2000–2019) to define the heat wave day thresholds, we may achieve an overestimation of the forecast skill in the verification compared to using a leave-one-out method (in which one year is excluded at a time from the dataset when defining the threshold). However, as shown in the last column of Figure 1, excluding even the most extreme year has only a minimal impact on the threshold definition. Therefore, it is reasonable to assume that the effect on the skill is not substantial."

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?

Thank you for these comments. We edited the Discussion for many parts.

The first paragraph of the Discussion (now lines 438-443) we edited to be:
"We examined the skill of hindcasts of the ECMWF in forecasting the probability of heat wave days over Europe 1 to 4 weeks ahead. The assessed hindcasts demonstrated varying levels of accuracy across different regions, and decreasing levels with increasing forecasting lead times, which is in line with many earlier studies, e.g., Wulff and Domeisen (2019), and Pyrina and Domeisen (2023). This outcome could be seen as expected, as we employed the same forecasting model and verification region as in these previous works. However, our method for determining the probability of a heat wave day was novel, providing a fresh perspective that sets our study apart from earlier research using the same model and verification region."

Additionally, parts of the discussion ll. 411-424 have been placed into the introduction (and these lines have been completely removed from the discussion).

Further, considering question: *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?*, we have now added the following information to the Introduction, now lines 45-50:
"The warm-up time related to outdoor temperature depends on building properties (U-value, ventilation airflow rate, and thermal mass of buildings). In typical Nordic well-insulated apartment buildings, e.g., brick and concrete apartment buildings, the warm-up time is 5- 6 days. In not well-insulated buildings and/or light structures, e.g., wooden structures, the warm-up time is much shorter. In those cases, it could be only 1- 2 days."

Considering question: *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?*
We have now added to the discussion, now lines 463-473:
"To the knowledge of the authors, there has been no published research on how warning lead time contributes to the effectiveness of heat-health warning systems. However, considering the short lag between heat exposure and worsening of health conditions, extending warning lead times from the current level of few days is acknowledged to be valuable to public health, as prevention and emergency measures need to be in place and operational at the onset of a hazardous heat event (WHO 2021). Organization of the measures, such as communication campaigns, establishing cooling centers, arrangements to protect vulnerable population groups, and ensuring adequate supply and distribution of workforce, equipment, and other resources, require time and would benefit from receiving early warnings 1–2 weeks ahead, particularly because heat waves often occur at times when organizations and services are already short-staffed due to summer holiday season. Longer lead time is especially important regarding exceptionally severe and prolonged hot periods, which challenge the functioning of society on a wider scale and may require large-scale interagency and even

transboundary response. The likelihood for these types of events can be expected to increase in Europe as climate change progresses."

Regarding to possible limitations of our study:
We acknowledge in the methods section (now lines 209-213) that the ensemble size of the hindcast dataset (11 members) differs from that of the current operational forecasts (101 members). Using operational forecasts would involve significantly more effort and extend beyond the scope of this study. Moreover, operational forecasts might cover a shorter historical period, limiting their utility for our analysis.
Another limitation is that the ensemble spread could have been examined in greater detail, as briefly mentioned in the discussion (now lines 451-453). However, since the probabilistic forecast utilizes the entire ensemble, the spread is inherently accounted for in the analysis.
Further possible limitation of our study is that we used only one model (ECMWF), as we write in the conclusion (now lines 490-492): "-- future research could investigate at which stage of the heat wave development extended-range weather forecast models in general, not only the specific model system considered here, begin to predict heat wave occurrence, --"
The selected definition "heat wave day" also might be a limitation, however as we mention in section 2.1. (now lines 114-115) "Our definition of heat wave days is meaningful as it aligns with thresholds commonly used in epidemiological studies on heat-related health effects, --"
Another limitation is that this study focuses on forecasting heat wave days; however, it is important to note that what ultimately concerns people are the impacts, such as health effects. Predicting these impacts is beyond the scope of this work.

Further comments:

Title: mix of capitalized and lower-case words
Thank you for pointing this out. We now corrected this to title case, i.e., The Probabilistic Skill of Extended-Range Heat Wave Forecasts Over Europe.

l. 19: in extended range → in the extended range
l. 27: "persistence [...] seem to have" → "persistence [...] seems to have a"
Thank you for these corrections, we have edited the text as suggested.

ll. 69 – 70: I think this last sentence might be better placed in the discussion.
Thank you for the comment. We reformulated this and left it in the introduction, now lines 86-88: "In theory and practice, probabilistic forecasts have been shown to contain more information and should be more valuable to users than categorical, deterministic forecasts (Murphy 1977, Richardson 2001), though their practical utility depends on users' ability to incorporate such information into decisions (e.g., Lopez & Haines 2017; Ramos et al. 2013)."

l. 75: "have" → "has"
l. 106: "initiation" → "initialization"
l. 122 "capture" → "skillfully predict"
Thank you for these three corrections, we have edited the text as suggested.

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.
Thank you for this remark. We have reformulated this as: Hence, a heat wave in the forecast is defined relative to the forecast model's climatology.

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 90th 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.
Thank you for this remark. We have added this text here in Section 2.3.2, now lines 245-250:
"In the verification, the forecast model-based probability of a heat wave day, p, was compared to the observed heat wave days (Section 2.2.1) derived from the ERA5 dataset. Since we used the data from the entire period (years 2000–2019) to define the heat wave day thresholds, we may achieve an overestimation of the forecast skill in the verification compared to using a leave-one-out method (in which one year is excluded at a time from the dataset when defining the threshold). However, as shown in the last column of Figure 1, excluding even the most extreme year has only a minimal impact on the threshold definition. Therefore, it is reasonable to assume that the effect on the skill is not substantial."

l. 175: It would be ideal to end this paragraph with a sentence about what you conclude from these statistics.
Thank you for this remark. We added here (now lines 146-148): "These statistics show that the 5-day moving average definition covers nearly all longer heat wave events (such as 3- to 4-day heat waves), but only a portion of shorter ones (1- to 2-day heat waves). This indicates that the 5-day moving average is particularly useful for identifying sustained heat wave events.".

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 *pb*" (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)
Thank you for these improvement suggestions, we have edited the text as suggested. Hopefully it is now clearer.

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?
Thank you for these remarks. We reformulated these and now it is hopefully more clear that the FDR is more like a concept and the B-H is one of the possible procedures to implement it.

l. 267: "a heat wave days" → "heat wave days"
l. 273: "shorter forecast weeks" → "shorter lead times"
Thank you for these corrections, we have edited the text as suggested.

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.
Thank you for pointing this out. We continued the sentence by (now lines 316-317)"; however, it should be noted that for lead times of 2 weeks (and longer) there are far fewer samples in the higher probability bins, making these points considerably more uncertain."

l. 280: "of the all hindcasts" → "of all the hindcasts"

l. 295: add comma after “In the second week”
Thank you for these corrections, we have edited the text as suggested.

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.
Thank you for this comment. We added (now lines 344-345) “These results suggest that the skill in forecasting heat waves decreased when excluding the longest period of heat wave days, whether it was the 2010 heat wave or a heat wave from another year.”

l. 300: “In the Figure 4” → “In Figure 4”
l. 305: two commas at the end of the line
Thank you for these corrections, we have edited the text as suggested.

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.
Thank you for this comment. To clarify how the severity of the over – or underforecasts can be assessed, we have now added about Fig. 5 (now lines 382-385):
“It should be noted that Figure 5 also shows how often forecasts were followed by a heat wave or *near-heat wave* conditions (e.g., temperatures exceeding the 85th percentile) in the ERA5 dataset. For instance, in situations where $p > 0.66$, temperatures surpassing the 85th percentile (rather than the 90th percentile) occurred even in 95% (lead time one week), 78% (lead time two weeks), 74% (lead time four weeks), or 44% (lead time four weeks) of cases.”
For the threshold p=0.1 we added to the text (now lines 378-380):
“Additionally, $p < 0.33$ provides a good indication that a heat wave is unlikely. Based on the data, the lower the p (below 0.33), the less likely a heat wave is to occur, as, e.g., in occasions the $p < 0.1$ (no figure), heat wave days occurred only in 1% (lead time one week), 4% (lead time two weeks), 6% (lead time three weeks), or 8% (lead time four weeks) of cases.”

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?
Thank you, good point. We have now added here (as lines 368-370):
“All in all, the forecast skill improves when more of the data points in $p < 0.33$ fall below the grey line,

and those in p > 0.66 are above the grey line. At a glance, forecast week 1 (Fig. 5a) appears to have good skill, while forecast week 4 (Fig. 5d) shows relatively poor skill."

l. 322: "p<0.33 be" → "$p < 0.33$ would be"
l. 339: "amount" → "fraction"
Thank you for these corrections, we have edited the text as suggested.

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?
Thank you for pointing this out. We mean indeed the forecast initialization (date) relative to the onset of the heat wave, and we have now edited the text accordingly.

l. 344: "corresponsive" → "corresponding"?
Thank you for this remark, we have edited the text as suggested.

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 …"
Thank you for this comment. We have edited this as (now lines 406-410):
"In heat wave day forecasts both one week in advance (Figure 6a) and two weeks in advance (Figure 6b), the forecasts show clearly higher p for days within the heat wave than outside, especially for the forecasts which are in the green boxes indicating that the heat wave was just starting or already underway when these forecasts were issued. Additionally, there is some overestimation, particularly 1-2 days before or after the heat waves indicating slight inaccuracy in forecasting the exact day of the start and ending of the heat wave."

l. 371: "indicating ongoing heat" → "indicating an ongoing heat wave"
Thank you for this remark, we have edited the text as suggested.

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?
Thank you for these comments. We have added the explanation for *n* and the width of the boxes. Regarding the suggestion to express p relative to the base rate, we believe that the current presentation of Figures 5 and 6 works best for our purposes, and therefore, we have decided not to make changes in this regard.

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.
Thank you for this comment. We agree and we have now put this Figure 7 as Supplementary Material 1 and we have also included here in the text: "Thus, the differences remain 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?
Thank you for this remark. We have removed from here the short recap of the method and further edited this paragraph to be (now lines 438-443):
"We examined the skill of hindcasts of the ECMWF in forecasting the probability of heat wave days over Europe 1 to 4 weeks ahead. The assessed hindcasts demonstrated varying levels of accuracy across different regions, and decreasing levels with increasing forecasting lead times, which is in line with many earlier studies, e.g., Wulff and Domeisen (2019), and Pyrina and Domeisen (2023). This

outcome could be seen as expected, as we employed the same forecasting model and verification region as in these previous works. However, our method for determining the probability of a heat wave day was novel, providing a fresh perspective that sets our study apart from earlier research using the same model and verification region."

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?
Thank you for this comment. Yes, we edited this to be (now lines 447-448): "We found that the skill in forecasting heat waves decreased when excluding the longest period of heat wave days, whether it was the 2010 heat wave or a heat wave of some other year."

l. 439: would add here that this significant skill largely vanishes when 2010 is excluded (Fig. 4)
Thank you for the comment. We here remove the word "Eastern" and leave the text as (now lines482-483) "--in the forecast weeks 3-4: statistically significantly better skill than the reference forecast only in some grid points across South-Eastern Europe" it mentions that significant skill is only in *some* grid points across South-Eastern Europe, and those are the areas in which statistically significantly better skill than the reference forecast remained even when 2010 was excluded.

l. 448: "its" → "heat wave occurrence"
l. 448: "further" → remove
Thank you for these remarks, we have edited the text as suggested.