Authors' response to both Referee #1 and Referee #2

MS. Ref. No.: NHESS-2024-75 "The usefulness of Extended-Range Probabilistic Forecasts for Heat wave forecasts in Europe"
Natural Hazards and Earth System Sciences

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

Referee #1 (RC1):

General comments:
This study focuses on the sub-seasonal predictability of heat waves over Europe using the ECMWF model. The variable studied here is 'heat wave days,' which refers to the number of 5-day periods whose average temperature exceeds the 90th percentile of the climatological distribution. This approach highlights certain performances beyond week 2, particularly for the most intense and prolonged episodes, and mostly over the eastern half of the continent.
In general, the scientific question addressed and the method to answer it are entirely relevant and sound. Among other interesting results, I find particularly original and smart the evaluation of the capacity of the model to predict the life cycle of heat waves, taking into account the relative time of forecast issuance and heatwave initiation. However, in my view three aspects need to be revised or further elaborated before the manuscript can be accepted. These points, detailed below, concern first the structure of the manuscript which requires improvements, then the case of the summer of 2010 which needs to be further discussed, and also some missing specifications in the description of the method.

Response: We are very grateful for your comments, views and improvement suggestions, they really give a lot for the manuscript. It was especially nice to receive such encouraging comments about Figure 6! For point-by-point responses to parts that require improvement, please see below.

Specific comments
1- Paper organization:
I find the organization of the manuscript rather clumsy, in particular the discussion part that dwells into strategies of adaptation to heat, thereby repeating or elaborating some elements of the introduction. Additionally, it sometimes go quite far (too far!) into details when it comes to adaptation and preparedness measures, keeping in mind that the core of the paper is an evaluation of heat wave forecast skill.
The conclusion reads like a shorter repetition of the discussion.
Finally, the part of the discussion providing avenues for enhancing heat wave forecast skill (hence more aligned with the main work of this paper) refers very vaguely to "additional bias techniques" and cites a list of (sub-)seasonal forecast predictors without indicating in which manner they could contribute to refine the heat wave forecasts.

Response: Thank you for this comment. We have edited the Discussion strongly (lines 393-429). We have left out the details of adaptation and preparation measures, and do not refer any more to the additional bias techniques. We now focus more on discussing the skill of these forecasts in the revised manuscript. We also give more attention to the results of Figure 6, the heat wave life cycle figure, and the significance of our results considering the weight of the summer 2010 heat wave.
We would also like to mention here that we have changed the title to be: "The probabilistic **skill** of Extended-Range Heat wave forecasts over Europe" due to the comment by another referee:
"The word "usefulness" in the title made me as a reader expect something (more related to climate services) that is not shown in the study.-- "

2- Impact of the summer of 2010 in eastern Europe / western Russia:

That summer was characterized by a particularly long lasting heat wave over that region, and this is well reflected in your manuscript. Yet, that summer of 2010 seems to 'contaminate' your results and conclusions : it is particularly obvious when comparing your fig. 4g with 4i (and 4j with 4l), keeping fig. 2a in mind. Without that particular summer of 2010, most of the skill over Europe is gone in weeks 3 and 4. Could you elaborate on this, and discuss the significance of your results considering the huge weight of that event ?

Response: A good point, thank you. We now replotted figures 1, 2, 4, and 6 to the revised manuscript (see below) to compare our results for the period 2000-2019 with and without 2010. Figure 1 gives a spatial distribution, with 1 °C intervals, for the threshold of the heat wave days for the period 2000-2019 with (Figure 1, first column) and without 2010 (Figure 1, middle column). One can see that if summer 2010 is excluded, the north-west-southeast gradient is very close to that for the whole 2000-2019 period. The last column shows the impacts of including 2010: in most of the western and the southern Europe the difference is ±0.1°C, while in the eastern and north-eastern parts of Europe the impact is mostly between 0 and +0.55 °C, except for the very northern Fennoscandia where the impact is between -0.2 and 0 °C.

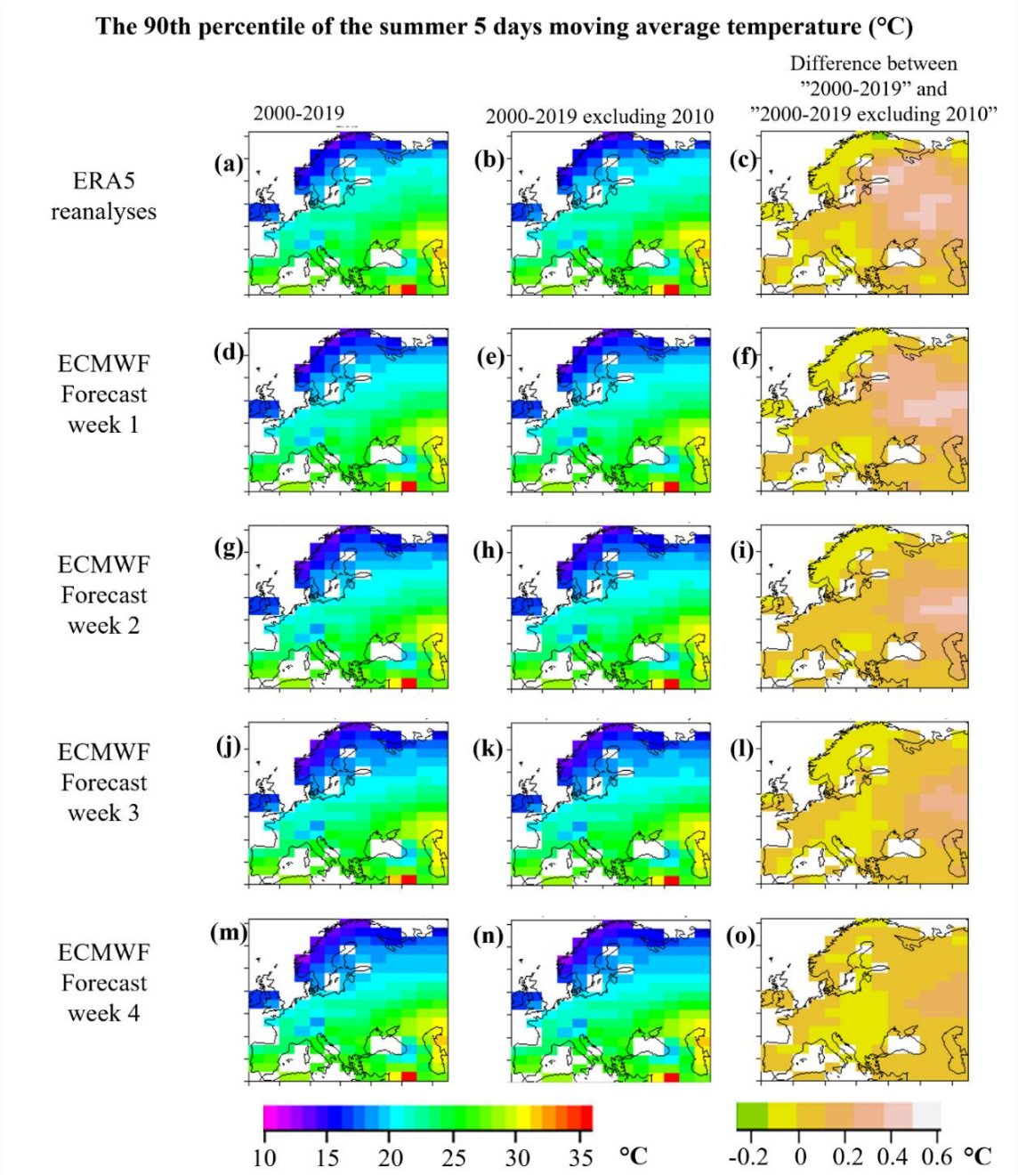


Figure 1: The lower thresholds of heat wave days: the 90th percentile of the 5-day moving average temperature in summers 2000-2019 (first column) and in summers 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010, middle column) of the ERA5 reanalyses (a and b), and (d,e,g,h,j,k,m, and n) of the ensembles of the ECMWF's hindcasts in different forecast weeks. The last column shows the difference between these two.

Figure 2b indicates that if the summer 2010 is excluded, other years (e.g., 2014) appear in eastern Europe / western Russia, compared to Fig. 2a, and the duration of the longest period of heat wave day get shorter there. In Fig 2d especially in those areas where 2010 had the longest period of heat wave days, excluding it means an increase in the number of periods with heat wave days, as the 10% of the hottest days are now distributed to a larger number of events.

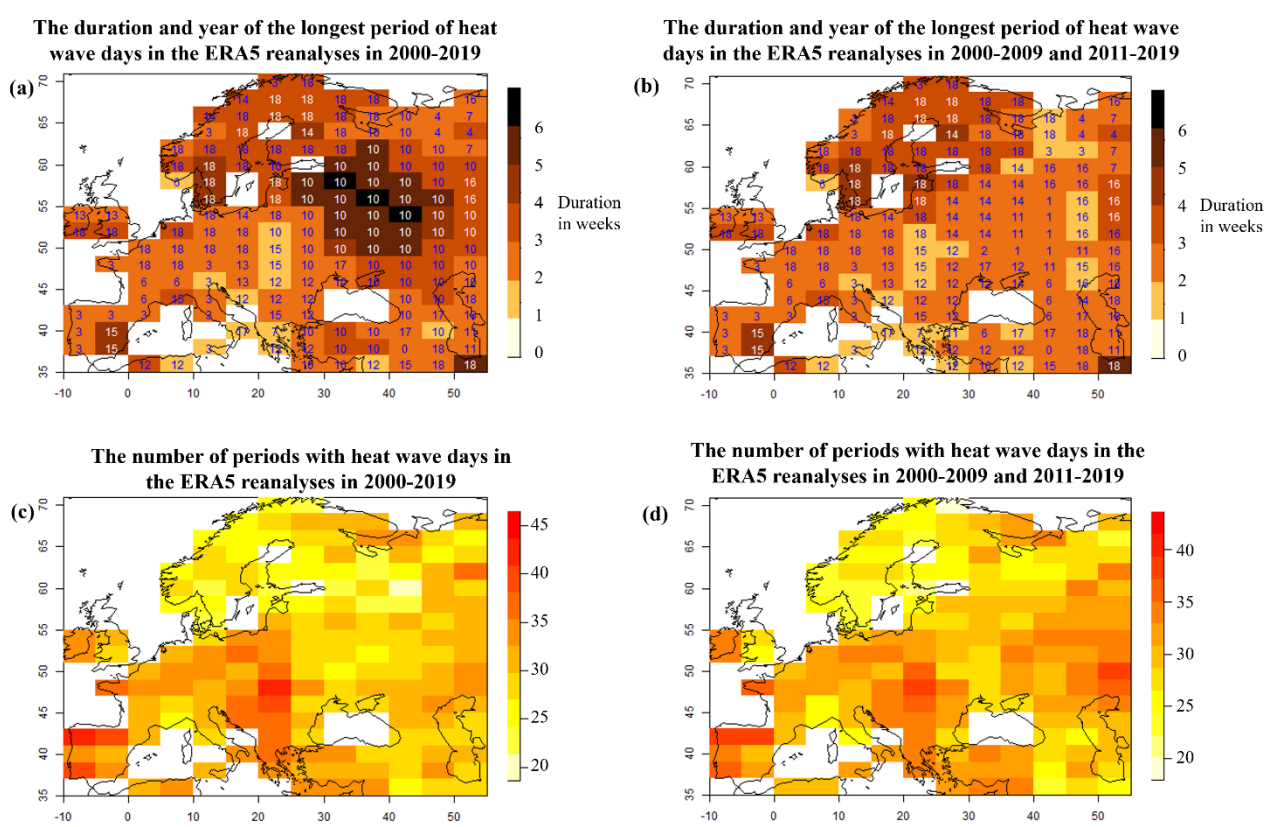


Figure 2: The duration and the year of the longest period of heat wave days defined from the ERA5 reanalysis data of (a) summers 2000-2019 and (b) summers 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010) (marked as 0-19), and the number of periods with heat wave days (b) in the ERA5 reanalyses during (c) 2000-2019 and (d) 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010).

In the revised Figure 4 (below), the last column now shows the BSS of the hindcasts excluding the summer 2010. Leaving out 2010 actually seems to have less impact than leaving out, in each grid point, the summer with the longest heat wave (the middle column). For example, in Finland the skill remains for the third week. Also the southeast parts of the study domain seem to remain with skill. Hence: the best of the skill seems to come from the longest period of heat wave days, whether it was the 2010 heat wave or a heat wave of some other year and also this has been written in the manuscript on lines 401-404.

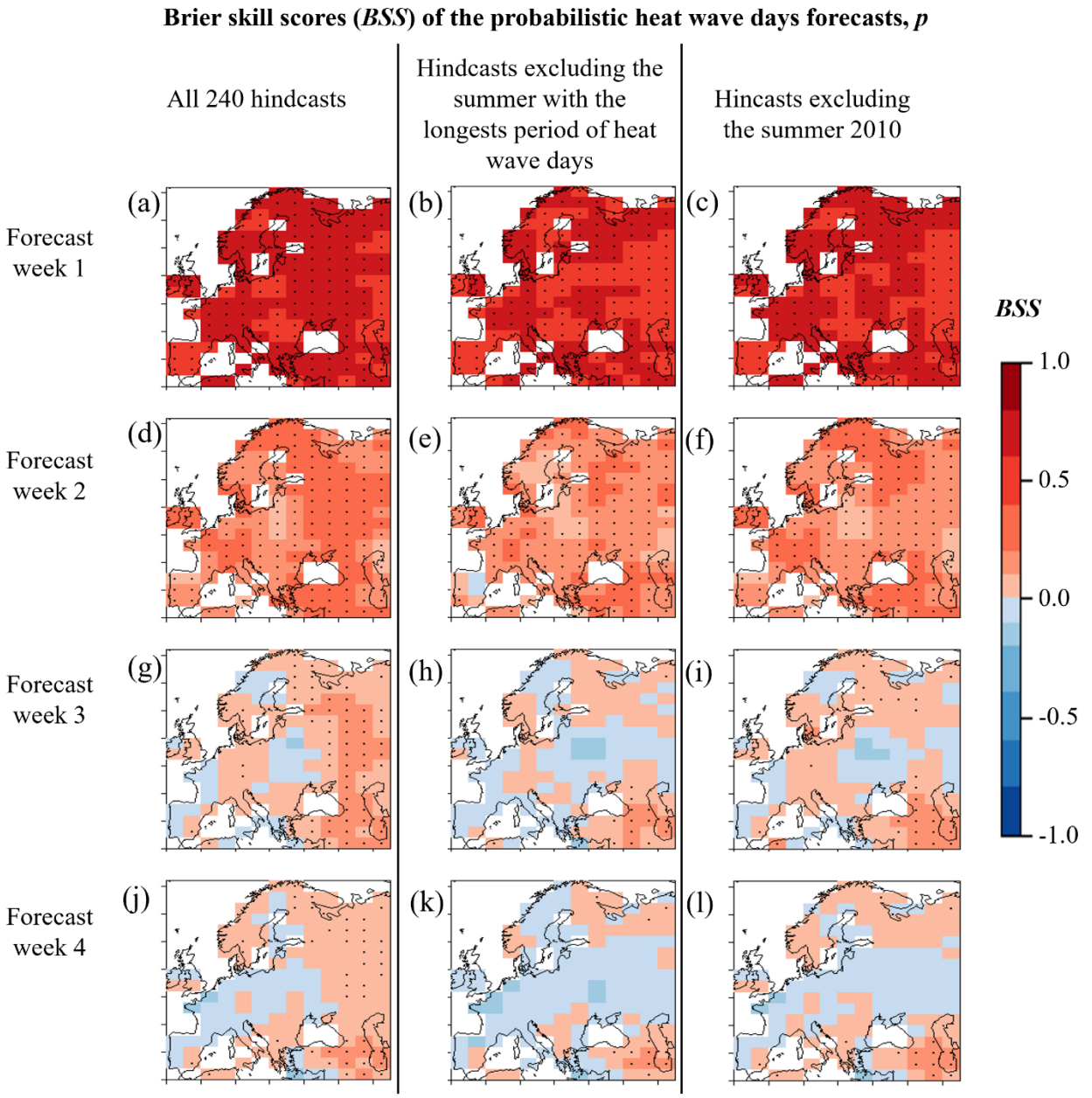


Figure 4: Brier Skill Scores (BSS) of the probabilistic heat wave days forecasts, *p*, during all summers 2000-2019 (first column), in hindcasts excluding the summer with the longest period of heat wave days (middle column), and in hindcasts excluding the summer 2010 (last column). The statistical occurrence p=0.1 for heat wave days were used as the reference forecasts. The dotted areas show where BSS is greater than zero with the false discovery rate no more than 10%.

In the revised Figure 4, we have left out the "summer with the longest heat wave" since another referee pointed out that:

Figure 4, ll. 288 – 295: While I think excluding the summers with the longest heat waves gives a good idea of how strongly the overall skill of the forecasts is influenced by these events, I don't think we can learn much from the skill for just the summer with the longest heat wave. While it seems to be in line with the conclusions from the right column in Fig. 4, I would argue that all the middle column might be telling us is that the reference forecast is particularly bad when you choose to basically look at one event alone (meaning ot in the BS is 1 most of the time and thus the BS of climatology, i.e. pt = 0.1, gets very high, because now your climatological forecast is not reliable anymore). Unless of course you recalculate the 90th percentile using only one summer, which is obviously problematic (representativeness), too.

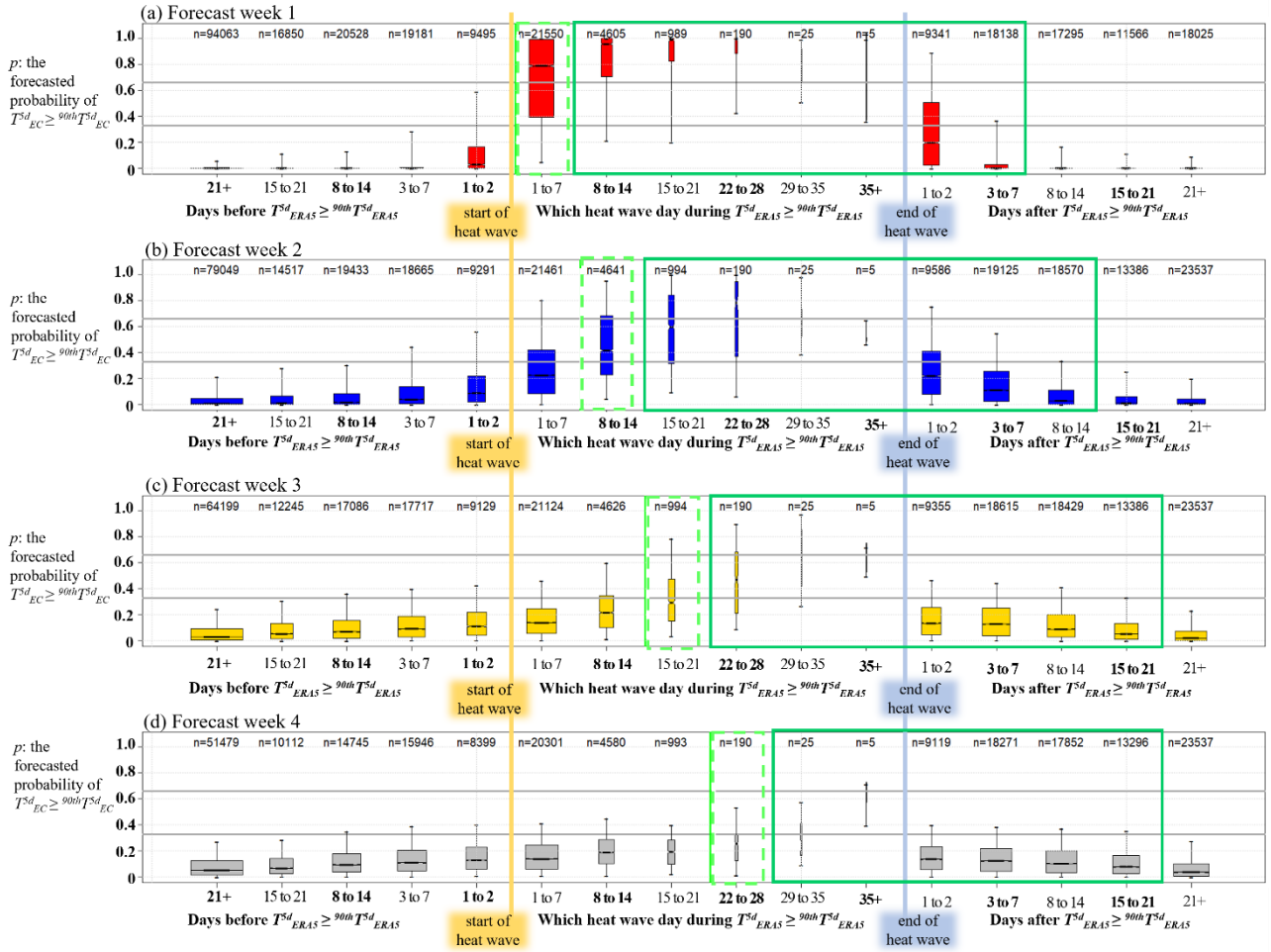


Figure 7 (DATA EXCLUDING YEAR 2010): The forecasted probabilities of heat wave days shown for days that (in ERA5) were 21 to 1 days before the heat wave(left); for the 1st to 35th heat wave day during the heat wave, and 1 to 21 days after the heat wave with lead times of a) one week, b) two weeks, c) three weeks, and d) four weeks. Dashed green boxes indicate forecasts where, at the time of issuance, a heat wave in that grid point was about to begin within a week. Solid green boxes indicate forecasts where, at the time of issuance, a heat wave was already ongoing in that grid point. The boxplots (as in Fig.5) include all forecast data across except summer 2010 the European region at each land-grid point.

We also plotted the heat wave life cycle figure (Figure 6) without year 2010, in the revised version as Figure 7.
Leaving out year 2010 removes most of the very longest heat waves, i.e, with lengths above 28 days. However, the same way as with including the year 2010, in the forecast weeks 1-3: there is still signal of enhanced accuracy in forecasting prolonged (here several weeks long) heat waves at the time when the heat wave had initiated prior to the forecast issuance. This is also written in the manuscript on lines 384-387.

These figures showing the effects of excluding year 2010 on the thresholds of heat wave days (Fig. 1), the duration and the year of the longest period of heat wave days (Fig. 2), probabilistic skill of the heat wave forecasts (Fig. 4) and on predicting the lifecycle of a heat wave (Fig. 7) are included in the revised manuscript.

3- Method:

L.115-117: It is not very clear if your take the lead time into account when computing the 90th$TEC5d$. For example, when computing this value for, say, July 1st : do you compute one single value by pooling together all the hindcast members that include the sequence July 1st- July 5th (regardless of the start date) ?
Response: No, we did not.
This is now explained and shown in Table 1 in the revised manuscript (lines 91-115) and below in Table 1A.

Or instead, do you compute different percentile values according to the lead time (i.e. one value if July 1st is part of week 1, another one if it is part of week 2 etc.) ?
Response: Yes, we did.
This is now explained and shown in Table 1 in the revised manuscript (lines 91-115) and below in Table 1A.

From my understanding, the former strategy allows a larger statistical sample to compute percentiles, but on the other hand, there is a potential impact of lead-time dependency. The latter one seems more accurate from this point of view but then of course the sample size is smaller. I guess the "lead time dependent climatology" indicated on l.125 refers to this strategy.
Response: Yes, the latter one is the one we used.
This is now explained and shown in Table 1 in the revised manuscript (lines 91-115) and below in Table 1A.

Additionally, you should specify the range of start dates used in the method. I believe this would help understand how you computed percentiles for the very first days of June in particular. In other words, did you include hindcasts initialized in early May, to ensure a homogeneous sample size throughout all summer days ?
Response: No, we did not.
This is now explained and shown in Table 1 in the revised manuscript (lines 91-115) and below in Table 1A.

Or did you only consider the first days of June as part of "week 1" lead time)?
Response: Yes we did.
This is now explained and shown in Table 1 in the revised manuscript (lines 91-115) and below in Table 1A.

Could you clarify (and potentially discuss) these method points in the manuscript ? Maybe include a schematic or a table if needed.

Response: Thank you for the comments. Sure, we have now added a Table (see below) to the manuscript together with explanations (lines 91-115).

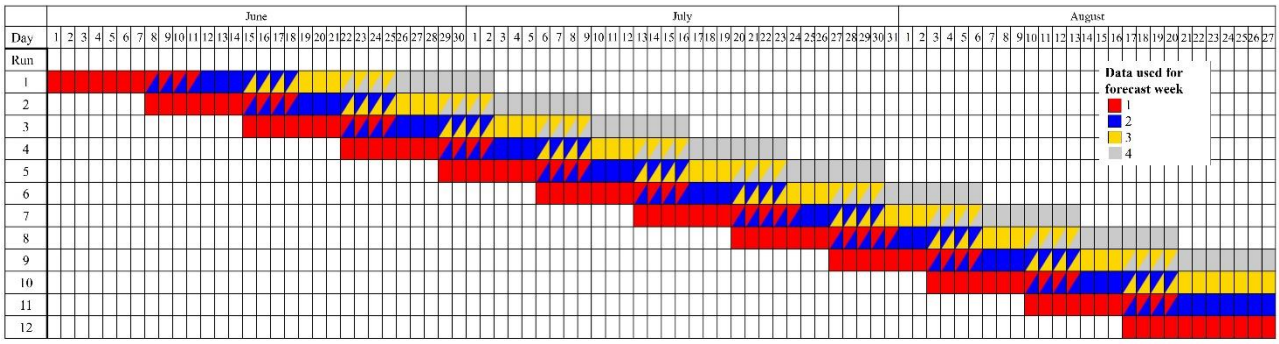


Table 1A. Table showing details of the investigated hindcasts. Each row contains one run, altogether 12 runs. The first red boxes on each row show the initiation date of the hindcasts, which are same for all years 2000-2019. The data of days marked with red are used for lead time 1 week, blue for 2 weeks,

yellow for 3 weeks, and grey for 4 weeks. The forecast data used for the forecast weeks were partially overlapping due to the use of 5-days moving averages with forward-looking window: the forecast week 1 used data of days 1 to 11, the forecast week 2 data of days 8 to 18, forecast week 3 data of days 15 to 25, and forecast week 4 data of days 22 to 32. The data used for two lead times are here marked with two colours. Note: for lead time 1 week we used data of 12 runs, for lead time 2 weeks we used data of 11 runs, for lead time 3 weeks we used data of 10 runs, and for lead time 4 weeks we used data of 9 runs (of years 2000-2019).

Technical corrections:

L.179 “by change” => “by chance” (?)

Response: Thank you, we corrected this as suggested.

L.196 Do you mean “how early a heat wave becomes…” or “how early heat wave days become” ?

Response: Thanks for noticing this, we mean “how early heat wave days become”, we corrected this.

L.248-250 : OK, but this sounds more like a rephrasing of what precedes than a new information.

Response: Thank you, we removed this sentence on lines 248-250.

L. 271-273: Nice result for week 3 but it would be fair to remind that the sample size of week 3 forecasts with p>0.5 is probably very small, considering Figure 3a. So I think this result should be considered with a pinch of salt.

Response: Yes, thank you, we agree and have removed these lines 271-273 of the first manuscript.

Figure 1: I would recommend to display the 4 ECMWF maps as “bias wrt. ERA5”, ie plotting the difference “ECMWF minus ERA5”. The associated comment L. 213 would be more convincing.

Response: Thank you for this comment. We replotted the Figure 1, however, not exactly as suggested here, but it should be more convincing now.

L. 268 and elsewhere : Choose between “Figure 3(b)” or Figure “3b” (even better: remove one of them) and choose also between “Fig.” and “Figure” .

Response: Thank you. We edited this.

Non-existing “Figure 3d” shows on L. 272. Better remove it, since it seems quite obvious that you keep commenting Figure 3b here.

Response: Thank you for mentioning this. We corrected 3d to 3b.

L. 294: Typo (?) . I was expecting : “Eastern Europe (summer 2010)“

Response: Thank you for mentioning this. We corrected 2018 to 2010.

L.319-326: Ok but this part is very unpleasant to read. Please rephrase by not repeating the exact same sentence 4 times !

Response: A good point, thank you. We have edited this in the manuscript (lines 325-332) to be:

In occasions the forecasted probability for heat wave days was low (p<0.33), heat wave days occurred in 2% (lead time one week), 7% (lead time two weeks), 10% (lead time four weeks), or 11% (lead time four weeks) of cases. Moreover, in occasions the forecasted probability for heat wave days was intermediate (0.33≤p≤0.66), heat wave days occurred in 45% (lead time one week), 39% (lead time two weeks), 30% (lead time four weeks), or 28% (lead time four weeks) of cases. In occasions the forecasted probability for heat wave days was high (p>0.66), heat wave days occurred in 86% (lead time one week), 68% (lead time two weeks), 67% (lead time four weeks), or 38% (lead time four weeks) of cases.

L. 406: heat-health action plans (2 typos)

Response: Thank you for mentioning this. We corrected this as suggested here.

MS. Ref. No.: NHESS-2024-75 "The usefulness of Extended-Range Probabilistic Forecasts for Heat wave forecasts in Europe"
Natural Hazards and Earth System Sciences

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

Referee #2 (RC2):

This study investigates the probabilistic skill of extended-range forecasts of mildly extreme land temperatures over Europe. It shows that these forecasts are overall reliable out to the third forecast week, but, except for Eastern/Southeastern Europe, do not significantly differ in skill from a much simpler climatological forecast. The skill of the forecasts appears to be strongly enhanced by the most long-lasting events. Excluding these events results in reduced skill over almost all of Europe. An analysis of the evolution of skill throughout the life cycle of the heat wave indicates that the models capture the persistence of anomalous temperatures well, whereas the onset and end of the events seem more difficult to predict. The study presents a relevant contribution to the field of evaluation of extended- range/subseasonal prediction for potentially impactful events. While previous studies have considered the prediction skill of the same extended-range forecasts for extreme temperatures before, this study adds a thorough assessment of the probabilistic skill of the forecasts by using some well-documented methods and scores (which facilitates comparability) and providing some more non-standard ways of looking at the prediction skill. Assessing the probabilistic instead of the deterministic skill of the forecasts is arguably much more important in the extended range, since their uncertainty is large, but the information in the spread of the ensemble could still make the forecasts reliable. The employed methods are sensible and the skill analysis for the heat wave life cycle is innovative and a highlight of the paper. I do, however, not entirely agree with the way the study is framed. The title implies that the study assesses the usefulness of the forecasts, which I don't think it does. The authors also stress the health impacts of heat waves a lot, which is of course a good motivation to investigate the skill at predicting heat extremes, but the study does not include any analysis that links the forecasts to health impacts/heat stress in any way. Furthermore, some of the methods should be explained and motivated more clearly. Finally, the writing could be made more concise in many places. I provide more detailed comments below.

Response: We are very grateful for your comments, views and improvement suggestions, which all give a lot for the manuscript. It was especially nice that you mentioned the skill analysis for the heat wave life cycle as the highlight of the paper! For point-by-point responses to parts that required improvement, please see below.

Major remarks:

1. The word "usefulness" in the title made me as a reader expect something (more related to climate services) that is not shown in the study. The usefulness of a forecast can only be determined by involving its user(s), which also means that the forecast, in most cases, will be useful only to some but not others. Furthermore, for a forecast to be useful, skill (which is what I think the article is actually focused on) is just one of many requirements. So, unless the analysis is extended significantly and involves this component, I suggest changing the word "usefulness" to something else here (maybe "skill" would be most accurate).

Response: Thank you for this comment, we agree, and we have changed the title to be:
"The probabilistic skill of Extended-Range Heat wave forecasts over Europe"

2. There is a lot of text concerning health impacts/risks of heat in the discussion (ll. 403 – 443). While I don't generally disagree with anything that is written about this, I don't think it deserves the amount of space it is given in the discussion, given there is no direct relationship with the presented results. The

study investigates the probabilistic skill of summer forecasts for mildly extreme (dry bulb) temperatures and the discussion should focus on this aspect. The authors offer an explanation for why they use the temperature measures that they use, and I think it is fair to focus on these, but there is evidence indicating that other measures of temperature are more strongly related to heat stress (involving radiation, humidity, wind) and thus more suitable for measuring health risk/impact of heat events, see e.g. Di Napoli et al. (2019), McGregor & Vanos (2018). Thus, I would suggest removing the too detailed discussion of health impacts of heat from Section 4.

Response: Thank you, we have removed the too detailed discussion of health impacts of heat from Section 4, as suggested. The revised Discussion is in lines 393-429.

Alternatively, if the focus on health impacts should be kept, I suggest considering the use of other, possibly more heat-stress-related, metrics.

Di Napoli, C., F. Pappenberger, and H. L. Cloke, 2019: Verification of Heat Stress Thresholds for a Health-Based Heat-Wave Definition. J. Appl. Meteor. Climatol., 58, 1177–1194, https://doi.org/10.1175/JAMC-D-18-0246.1.

McGregor, Glenn R., and Jennifer K. Vanos, 2018: Heat: a primer for public health researchers, Public Health, 161, 138-146, https://doi.org/10.1016/j.puhe.2017.11.0053.

If the current focus of the paper is kept, I think the discussion needs to be revised strongly. As mentioned above, the part ll. 403 – 443 seems very detached from the results of the study right now. The remaining text in Section 4 (ll. 374 – 401) is more of a summary and is to a large degree repeated in Section 5 (where it belongs, in my opinion). I think this part could be used better to discuss the implications, the potential and the limitations of your study (as you do in ll. 444 -452), see below for some suggestions:

• One question that I wonder about when seeing the results (although it is beyond the scope of the paper to answer this finally): Could it be that the forecasts are generally too persistent and thus lucky when a long-lasting heat wave happens, or do they actually "know" when to persist temperatures? In other words, are they right for the right reasons? The fact that the exclusion of the most long-lasting events basically removes all remaining skill from the week 3 & 4 forecasts makes me think that they might just have been lucky. Also, your Figure 6 could be interpreted further with this question in mind.

Response: Yes, thanks. Due to this we also have in Figure 6 shown how and how early the model forecasts the ending of heatwaves. Interpretation about the figure is on lines 353-371.

• You mention climate change in the discussion (l. 433). Against the backdrop of climate change, what do your results mean? Are we expecting better forecasts because we will see more (and potentially longer-lasting) heat extremes? Or might the predictability of these events also change?

Response: We mentioned climate change and the intensification of heat waves here as that means that heat wave forecasts are expected to be needed in the future as well. We actually removed this particular part (l.433) from the revised manuscript, however, in the beginning of the Introduction we still mention: "--due to the ongoing climate change, heat waves are expected to become even more common and intense (IPCC, 2021; Russo et al. 2014; Coumou and Rahmstorf, 2012; Kim et al. 2018, Vogel et al. 2020, Ruosteenoja and Jylhä, 2023)."

• Parts of your manuscript suggest that you would like to link this to the applicability of extended-range forecasts in early warnings of heat waves (e.g. l. 391). Could you elaborate on what your results mean, e.g. for an agency that would want to implement these forecasts for early warnings? Is the skill sufficient? Can the presented aggregation over large geographical areas (5 ˚ x 2 ˚) be useful in some way? Where can the forecasts contribute and where can't they, keeping in mind that they are

ok at predicting the persistence but not so good at predicting the onset far in advance?

Response: Thanks, this is a good question. We agree that at least this way used the forecasts do not predict the onset of a heat wave 3-4 weeks beforehand, but as it shows to capture the persistence of heat waves well. As the longest heat waves have high impact, even the smallest piece of information about them is reasonable to take into account. In the end of the Conclusions in the revised manuscript we have written: "These findings underscore the potential of these ECMWF's heat wave days forecasts to serve as early warnings for impending heat risks 1-2 weeks in advance. Notably, the higher-than-average predictability for intense and prolonged heat waves (at the time they have already started), offers a potential to early warnings even at a 3-week lead time. However, it is crucial to highlight the known uncertainty in the 3-week lead time forecast."

4. Since you try to address usefulness/applicability of the forecasts, it could be a good idea to assess reliability on a regional ("grid-point") level in addition to the BSS (Fig. 4). The reason is that reliability can be linked better to decision-making, see Weisheimer & Palmer (2014). Their paper shows a simple method of categorizing forecasts by the slope (and its uncertainty) of their reliability curve into 5 categories. This would address the usefulness aspect at least to some degree and could be a nice addition to the current results.

Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts, J. R. Soc. Interface, 11: 20131162. http://dx.doi.org/10.1098/rsif.2013.1162

Response: Thanks for the good idea. However, as we now changed the title of the manuscript (see Major remark 1.) to "The probabilistic skill of Extended-Range Heat wave forecasts over Europe" we would like to stay with the current analysis.

5. In general, Section 3 could use some additional explanations to make the results of the analysis easier to grasp for the reader. Generally, at the beginning of each subsection (3.X.), provide one sentence on why we're seeing this plot now and what it's supposed to tell us (like you do in ll. 276 – 277).

Response: Yes, thank you. This certainly is a good suggestion, and we have added explanations in the beginning of each subsection (3.X).

More specifically:
i. Section 3.5: Since this is not a very standard form of presenting forecast skill (at least not one I'm familiar with), I suggest explaining the reason for showing the skill in this form. I get the feeling it is relatively closely related to the reliability diagram. In what way does it differ/provide extra information? What can we learn from this way of looking at the forecasts? As a reference for the reader, give an example of what a good and a poor forecast would look like if displayed in this way (as you do in the part with the reliability diagram). A bit more information on this could also aid the interpretation of the next plot.

Response: In the reliability diagram (Fig 3b) the ERA5-based temperature data is used only as either no hot day (0) or hot day (1). However, this figure 5 shows also how near or far away the ERA5-based temperature was from the threshold of a hot day (the 90th percentile, the grey line). It is a good idea to add more information to explain and we have done that in the revised manuscript on lines 311-316.

ii. Figure 6, Section 3.6: I consider the life cycle plot a highlight of the manuscript, but it contains a lot of information, so I think it deserves a more thorough discussion (and to be picked up in Section 4!). One thing I find particularly noteworthy in this figure is that, while there seems to be an upward trend in the forecast probabilities leading up and into the heat waves, the highest probability class ($p > 0.66$) is only really predicted when the heat wave is already present in the initialization of the forecast.

Response: Thank you for this comment. We agree, Figure 6 is in the Results, Conclusions and Abstract "Nonetheless, persistence of prolonged heat waves seem to have higher-than-average level of predictability even at a 3-week lead time, offering early warning services an indication of the potential duration of an ongoing heat wave."
In addition, we have now mentioned Figure 6 also in the discussion on lines 404-410: "Figures 6 and 7 present a novel way for evaluating the ability of probabilistic heat wave day forecasts to capture the life cycle of heat waves, taking into account the timing of forecast issuance relative to heat wave onset. This approach could be developed further by adding information about the spread of the ensemble to the figure, and it could be applied to the verification of other extended-range models' heat wave forecasts in future studies."

Minor comments:

Title
"forecast" is used twice, could maybe reformulate?

Response: Yes, (see Major remark 1.) we changed the title to be:
"The probabilistic skill of Extended-Range Heat wave forecasts over Europe"

Intro
l. 26: 'intense and prolonged heat waves during the third forecast weeks' The study doesn't really address intensity, so the first part of this should be removed. I also think it would be more accurate to say that persistence of heat/extreme temperatures seem to have a higher level of predictability. The current sentence suggests that the forecasts are generally (onset, duration, intensity, ending) better for strong events.

Response: Thank you for this comment. We have removed the word "intense" here. And we have edited the sentence to be: "Nonetheless, persistence of prolonged heat waves seem to have higher-than-average level of predictability even at a 3-week lead time, offering early warning services an indication of the potential duration of an ongoing heat wave."

l. 28: one sentence linking back the results of the study to the motivation (early warning systems) would round off the introduction a bit more.

Response: Thank you, this is a good idea. We have added a sentence here: : "Nonetheless, persistence of prolonged heat waves seem to have higher-than-average level of predictability even at a 3-week lead time, **offering early warning services** an indication of the potential duration of an ongoing heat wave."

l. 32: 'in future' to 'in the future'

Response: Thank you, we have corrected this by adding word "the" here.

l. 37: 'particularly so in urban areas' can be removed since there is no relation of this to the question the study addresses.

Response: Thank you, we have removed 'particularly so in urban areas'.

ll. 46 – 54: I think it should be mentioned here that high (dry bulb) temperature is only one factor in heat stress, see references I provide above.

Response: Thank you for this comment, we agree. In the revised manuscript, this part has been moved to section 2.2.1 and there we have now added this text:

"Although high temperature (dry bulb) is a primary variable for assessing heat wave impacts, it is important to note that heat stress can also be influenced by other factors, such as humidity and wind speed. Nevertheless, in this study, we focus solely on the key driver of heat stress, the temperature."

ll. 55 – 63: I understand this paragraph as a motivation to consider the prediction of longer-term averages of temperature. If that's the case, be more explicit about it and say that due to the above reasons there could be value in considering the prediction of these averages. This could also be related to the fact that longer aggregations might be better predictable, see e.g.

Toth, Z. and R. Buizza (2019). "Weather Forecasting: What Sets the Forecast Skill Horizon?" In: Sub-Seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting. Ed. by A. Robertson and F. Vitart. 1st. Elsevier. Chap. Chapter 2, 17 –45.

Response: Yes, thank you, we agree, and we added this motivation in the last paragraph of the Introduction lines 65-66:
"As the uncertainty of extended-range forecasts is known to be large, we evaluated their probabilistic rather than deterministic skill."

ll. 59 – 62: I don't see the relevance of this with regards to the study. Can be removed.

Response: Ok, we removed suggested lines 59-62.

ll. 64 – 75: This fits more into the general motivation of the study at the beginning of the intro (potentially in a shortened form)

Response: Thank you for the comment. We have shortened this for some parts, but for the introduction we have kept the shortened ll. 64 – 75 where it was, as this paragraph ends with telling about the current length of heat wave forecasts in Europe, and from that it is good for us to continue in the next paragraph about the skill of extended range long heat wave forecasts.

l. 64: 'alleviate the tendency towards more frequent and intense heat waves' I don't understand what this means.

Response: Thank you for the comment. Here the idea of this sentence was that we mention the importance of mitigating the ongoing climate change to mitigate the intensifying of heat waves as they are projected to intensify the more the higher the atmospheric greenhouse gas concentration. However, it is not necessary to have this sentence here, so we have remove it (and made this chapter a bit shorter).

ll. 82 – 85: work out more clearly what your study is adding and providing beyond what has been done previously. Stress the probabilistic nature of the forecasts that you are evaluating and the analysis of the 'heat wave skill life cycle'

Response: Thanks, we have added this information here on lines 81-84: "The novelty of the study arises from the verification area encompassing the entirety of the European region allowing to highlight potential regional differences in the forecast skill, as well as from evaluating the model's ability to forecast the life cycle of heat waves, taking into account the relative time of forecast issuance and heat wave initiation."

ll. 86 – 89: This is already mentioned in ll. 55 – 62 and does not need to be repeated here

Response: Thank you, we removed lines 86-89.

l. 91: change 'forecasts' to 'hindcasts' or 're-forecasts'

Response: Thank you, we have used ‘hindcasts’.

l. 94/95: These two sentences seem a bit redundant as they are now. Can you be a bit more specific in guiding the reader through the paper here?

Response: Thanks, we have now specified here that we investigate the forecasts’ reliability, BSS and the model’s ability to forecast the life cycle of the heatwaves, taking into account the relative time of forecast issuance and heatwave initiation.

Methods

l. 96: The word ‘Materials’ seems a bit off in the context of the study. Maybe ‘Data’ is more appropriate?

Response: Yes, good point, thank you, we have changed the word ‘Materials’ to ‘Data’.

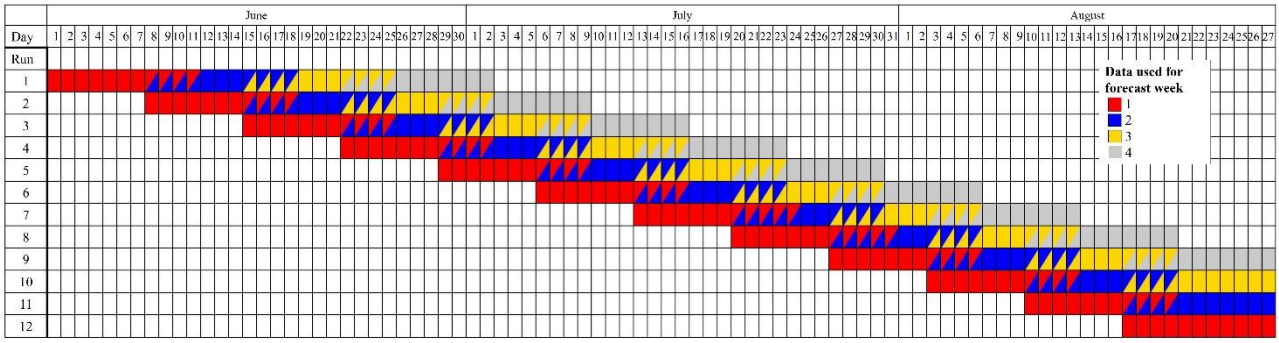
ll. 100 – 101: This could maybe be formulated more carefully. The skill of the hindcasts gives an indication of the skill of the forecasting system, but it is not necessarily the same (as you point out in ll. 126 – 134, so maybe merge these sentences).

Response: Yes, a good point, thank you. We merged these parts as:
”The hindcasts consisted of a control forecast and 10 perturbed ensemble members. It is important to distinguish between the hindcasts, consisting of 11 members, and the operational real-time forecasts, which initially had 51 members and now consist of 101 members (IFS Cycle 48r1). Therefore, the results obtained here from the 11-member hindcasts serve as a baseline measure of skill (see, e.g., Richardson 2001, Ferro et al. 2008) and the operational larger ensemble is expected to provide improved estimates of the normal distribution parameters, thereby enhancing skill to some extent.”

ll. 101: Meaning all forecasts initialized during JJA (which includes forecast and verification for September days) or all with verification dates in JJA?

Response: Meaning all forecasts initialized AND having verification dates in JJA. Hence, we did not include hindcasts initialized in early May (or those reaching September). We have now clarified this in the manuscript by a data Table 1A (below) and explanations:

Table 1A. Table showing details of the data of the investigated hindcasts. Each row contains one run, altogether 12 runs. The colouring of the boxes shows the coverage of the hindcasts’ data. The first red boxes on each row show the initiation date of the hindcasts, which are same for all years 2000-2019. The colours of the boxes indicate for which lead time (i.e., forecast week) the data were used: red for 1 week, blue for 2 weeks, yellow for 3 weeks, and grey for 4 weeks. The data used for the different lead times were partially overlapping due to the use of 5-days moving averages with forward-looking window: lead time 1 week used data of days 1 to 11, lead time 2 weeks data of days 8 to 18, lead time 3 weeks data of days 15 to 25, and lead time 4 weeks data of days 22 to 32. The data used for two lead times are here marked with two colours. Note: for lead time 1 week we used data of 12 runs, for lead time 2 weeks we used data of 11 runs, for lead time 3 weeks we used data of 10 runs, and for lead time 4 weeks we used data of 9 runs (of years 2000-2019).



ll. 102 – 104 & l. 106: What is the reason for only using Monday initializations instead of all available ones?

Response: Thank you for this comment. We have added this information to the manuscript, lines 95-98:
"As the 2m temperature has a large temporal autocorrelation, using both the Monday and Thursday initializations would not have added much information and would only have complicated the statistical analysis. We therefore decided to use only the Monday runs. The decision is arbitrary and we could have chosen to use only the Thursday runs as well."

l. 109: The ECMWF (re-)forecasts are run at higher horizontal resolution up to day 15 and then re-initialized at lower resolution from day 15 to 46.

Response: Thank you, we have added this information here, the text was edited on lines 99-103 to be: "The ECMWF reforecasts were initially run at a horizontal resolution of approximately 18 km for the first 15 days and then re-initialized at a coarser resolution of around 36 km for days 15 to 46. For our verifications, we used ECMWF's hindcasts at a horizontal resolution of 0.4° and ERA5 reanalysis data at 0.1°, both of which were bilinearly interpolated to a coarser 5° × 2° grid, considering only land grid points."

ll. 112 - : I suggest starting with defining heat wave days for the verification since the verification data is simpler (it only has one time dimension). Then you only have to explain how you handle the extra time dimension (lead time) in the hindcasts.

Response: Thank you, good point, we tried rearranging this, however, it was not easy, and therefore we still left the original order.

l. 117: Is this the 90th percentile of all (summer) days under consideration or for each calendar day individually?

Response: The 90th percentile is of all (summer) days under consideration.

l. 125: bias → frequency bias

Response: Thanks, we have corrected this.

ll.127 – 134: Maybe this could be re-structured a bit because it seems to be going back and forth between saying the hindcast ensemble is large enough to get an idea of the forecasting system's skill and saying it is not.

Response: Yes, thank you. We have re-structure this and merged it with ll 100-101, as suggested in a comment above, it is now on lines 116-120.
”The hindcasts consisted of a control forecast and 10 perturbed ensemble members. It is important to distinguish between the hindcasts, consisting of 11 members, and the operational real-time forecasts, which initially had 51 members and now consist of 101 members (IFS Cycle 48r1). Therefore, the results obtained here from the 11-member hindcasts serve as a baseline measure of skill (see, e.g., Richardson 2001, Ferro et al. 2008) and the operational larger ensemble is expected to provide improved estimates of the normal distribution parameters, thereby enhancing skill to some extent.”

l. 134: Another important difference between the skill shown in the study and the skill of the actual forecasting system is that in forecast mode, there is no information about the future, while you are using all years (including the evaluated one) when defining the percentiles. This is likely to lead to an overestimation of the skill. To simulate this setting, a leave-one-year-out cross validation could be employed. I’m not requesting the authors to do this, but I think it should be pointed out in addition.

Response: Thank you for the comment. As the 90th percentile of the 5 day mean temperature (of all the data) was used as a single threshold – and not, e.g., for bias-correction - in each grid point, we did not use leave-one-year-out cross-validation. However, in the revised Figure 1 (see below), the last column (Figures 1c,f,i,l,o) shows the difference between using data of "all years 2000-2019" (first column) and "years 2000-2019 excluding year 2010" (middle column) in the calculation of the 90th percentile.
Compared to the large northwest-southeast gradient of the absolute values of the 90th percentile in the first two columns, these differences are minor.

We have added following text on lines 177-186:
“As during the period 2000-2019, the summer 2010 was characterized by a particularly long-lasting heat wave over Europe, we investigated the weight of this event on our results by comparing our results for the period 2000-2019 with and without year 2010. In Figure 1 the middle column gives a spatial distribution, with 1 °C intervals, for the threshold of the heat wave days for the period 2000-2019 without summer 2010. The last column in Figure 1 (Figures 1c, 1f, 1i, 1l, and 1o) shows the impacts of including 2010: in most of the western and the southern Europe the difference is ±0.1°C, while in the eastern and north-eastern parts of Europe the impact is mostly between 0 and +0.55 °C, except for the very northern Fennoscandia where the impact is between -0.2 and 0 °C. Compared to the large northwest-southeast gradient of the absolute values of the 90th percentile in the first two columns, these differences are minor. When assessing the impact of the summer of 2010 (with the long heat wave in Europe) on the probabilistic skill of the heat wave forecasts, the threshold values of the middle column are utilized.”

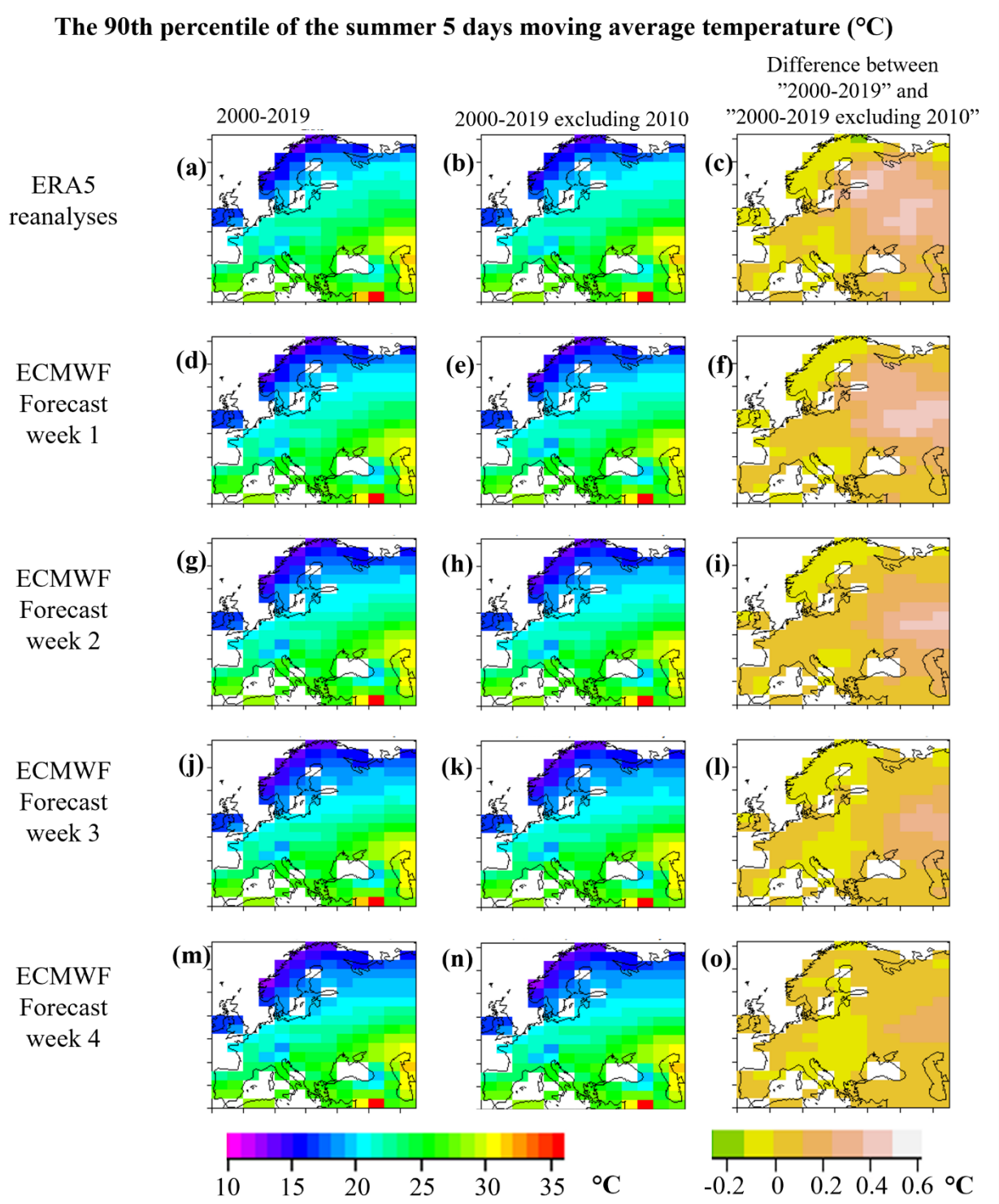


Figure 1: The lower thresholds of heat wave days: the 90th percentile of the 5-day moving average temperature in summers 2000-2019 (first column) and in summers 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010, middle column) of the ERA5 reanalyses (a and b), and (d,e,g,h,j,k,m, and n) of the ensembles of the ECMWF's hindcasts in different forecast weeks. The last column shows the difference between these two.

ll. 141 – 142: This sentence sounds like it is stating the obvious. Maybe better to say something like: "A single below-threshold day between two heat wave days was nevertheless classified as a heat wave day."

Response: Thank you, it is true that it is stating the obvious. However, we left is as it was as it clarifies what we mean.

ll. 144 – 149: see comment on Table 1 below.

Response: Yes, thank you, we edited the manuscript as suggested, i.e., we removed the old Table 1 and wrote the information about it here in section 2.2.1 as text only.

l. 168: do you mean “define this period as the summer containing the longest heat wave”? Is the entire summer taken out or just the period of the longest heat wave?

Response: We meant the entire summer, however, this will be removed (see comment for Figure 4 ll. 288-295)

ll. 175 - 1178: Could you provide a more detailed description of how the bootstrap resampling procedure works?

Response: Yes, thank you for the comment, here is a more detailed description that has been added to the revised manuscript on lines 235-240:
“For each grid point and lead time, we determined whether the hindcasts were considered more skilful than the reference forecasts by assessing the *BSS* using a bootstrap resampling procedure. First, we calculated the BSS *n* times (here *n*=5000), each time sampling the original data with replacement (i.e., the data points could be selected multiple times). The *BSS* was required to be statistically significantly above zero for the hindcasts to be considered more skilful than the reference forecasts. To assess this issue, we calculated the statistical significance level, i.e., the p-value under the null hypothesis that the *BSS* is zero. The p-value is then the proportion of the bootstrap samples greater the zero.”

l. 179: “change” → “chance”

Response: Thanks, the word ‘change’ was corrected to ‘chance’.

ll. 182 – 183: Explain in a few words how this procedure works.

Response: Yes, thank you for the comment, here is a more detailed description that has been added to the manuscript 240-284:
“However, because the statistical test on the map is repeated many times, small p-values are bound to occur by chance alone and the null hypothesis is rejected too often. Unadjusted p-values therefore overestimate the results (Wilks, 2016). We adjusted the p-values using the false discovery rate (FDR) method. Technically, the FDR controls for the expected proportion of false discoveries (hypotheses that should not have been rejected) among the rejected hypotheses. By setting this threshold *q* to 0.1 (twice the conventional 0.05, as suggested by Wilks 2016), and using the Benjamini-Hochberg (B-H) procedure (e.g., Benjamini and Hochberg, 1995), we ensured that on average no more than 10% of the rejected null hypotheses are false discoveries. In the B-H procedure, we first ordered the p-values from the smallest to the largest. Then we rejected the null hypothesis if $p_i < q * i/m$, where i was the position and m was the number of p-values. In practice, we can use readily available p-value adjustment functions (such as *p.adjust* in R) that change p-values to the smallest threshold *q* at which we would reject a particular null hypothesis.”

ll. 184 – 190: This seems to be better placed in the part where you explain how you generate a probabilistic forecast from the ensemble.

Response: Thank you, good idea, we moved it there as suggest here.

ll. 191 – 192: Why these categories? They seem rather arbitrary. Are they used somewhere, which would justify considering them here?

Response: Thank you for the comment, we clarified this in the manuscript by: "We conducted verification of heat wave day forecasts across all grid points in Europe based on forecasted probabilities falling within the ranges of here defined as low: $p<0.33$, intermediate: $0.33\leq p\leq 0.66$, and high: $p>0.66$."

l. 196: "a heat wave days become discernible" I don't understand this, please reformulate

Response: Yes, thank you. We mean: "(To investigate how early) heat wave days appear (in the forecasts)"

ll. 203 – 205: This part is a bit difficult to understand (especially before having seen Figure 6). Maybe reformulate this.

Response: Thanks, yes me can reformulated this and removed to Section 3.4.

Results
ll. 210 – 211: I think the information in this sentence is redundant here and already given where it is relevant.

Response: Thank you. We have removed this sentence on lines 210-211.

ll. 219 – 226, Table 1: What do you conclude from these numbers and how is this relevant for the forecasts or even their skill? Maybe this could rather become part of the method section (2.2.) if the point is to justify the definition of heat waves using the 5-day mean. To me, it wasn't clear why I'm seeing the table at this point in the paper. Since the information in the table is also entirely contained in the text, you could consider removing the table.

Response: Thank you, this is a good idea to remove the (old) Table 1 from here and include the text from here to the method section (2.2.1). We did this in the manuscript.

ll. 231 – 237: The same as the above comment applies to this subsection. This is just looking at ERA5, so it has nothing to do with the forecasts. I suggest moving this to Section 2 where the heat wave definition or the exclusion of the longest events is described. Alternatively, dedicate a short section at the beginning of Section 3 to the analysis that only deals with ERA5.

Response: Thank you for this comment. We moved this to Section 2.2.2.

l. 245: I think its noteworthy that this is not valid the other way around. You aren't claiming that, but I think it helps a reader who might be less familiar with the details of forecast verification to stress that sharpness is a property of the forecasts alone, i.e. 90% forecasts with p = 0 and 10% with p = 1 does not directly imply a perfect forecast (i.e. sharpness is a necessary but not a sufficient condition).

Response: A good point, thank you. We edited this as:
"If all the forecasts were perfect, then in Figure 3(a) 90% of the forecasts would have p=0 and 10% would have p=1, and in Figure 3(b) there would be only two points [0,0] and [1,1] for each forecast week."

l. 259: match → equal

Response: Thank you, we edited 'match' to 'equal'.

l. 267: by → with

Response: Thank you, we edited 'by' to 'with'.

l. 268: can drop the parentheses, it is mentioned in the sentence before.

Response: Thank you, we dropped the parentheses.

ll. 270 – 271: "reliability remained higher than that achieved by climatology alone" → this statement cannot be true since by the way you define climatology (i.e. without leaving the validation year out) it has perfect reliability by definition (but no resolution).

Response: Thank you for this comment. We removed this sentence.

ll. 271 – 273: I think there is a mix-up here between the "no skill-line" and the reliability of climatology. Climatology (as defined here) has perfect reliability, so no forecast can possibly have better reliability. It does, however, not have any resolution (it predicts p = 0.1 in all instances) and so its BS is higher than 0. If points lie above the "no skill-line" it means that they contribute positively to the BSS with climatology as reference. This is comparing the BS of the forecast to the BS of climatology, not just the reliability. For details see:

Mason, S. J., 2004: On Using "Climatology" as a Reference Strategy in the Brier and Ranked Probability Skill Scores. Mon. Wea. Rev., 132, 1891–1895, https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2.

Response: Thanks. We edited this to be:
"The points above the *no skill* line contribute positively to the BSS with climatology as reference." And we removed lines which were 271-273 in the first manuscript.

l. 280/281: "the predictions [...] demonstrates" → "the forecasts [...] demonstrate"

Response: Thank you, we have made the suggested editing.

l. 282: superior to the reference forecast → different from 0

Response: Thanks, we have edited this to "greater than 0".

l. 284: as before, here you basically say "BSS remains better than the reference forecast" while what you mean is that the BSS remains above zero, or alternatively, the forecasts remain better than the reference.

Response: Thank you, good point. We edited this to "forecast remaining better than the reference forecast".

Figure 4, ll. 288 – 295: While I think excluding the summers with the longest heat waves gives a good idea of how strongly the overall skill of the forecasts is influenced by these events, I don't think we can learn much from the skill for just the summer with the longest heat wave. While it seems to be in line with the conclusions from the right column in Fig. 4, I would argue that all the middle column might be telling us is that the reference forecast is particularly bad when you choose to basically look at one event alone (meaning ot in the BS is 1 most of the time and thus the BS of climatology, i.e. pt = 0.1, gets very high, because now your climatological forecast is not reliable anymore). Unless of course you recalculate the 90th percentile using only one summer, which is obviously problematic (representativeness), too.

Response: Ok, good point. We removed the column showing the skill for just the summer with the longest heat wave.

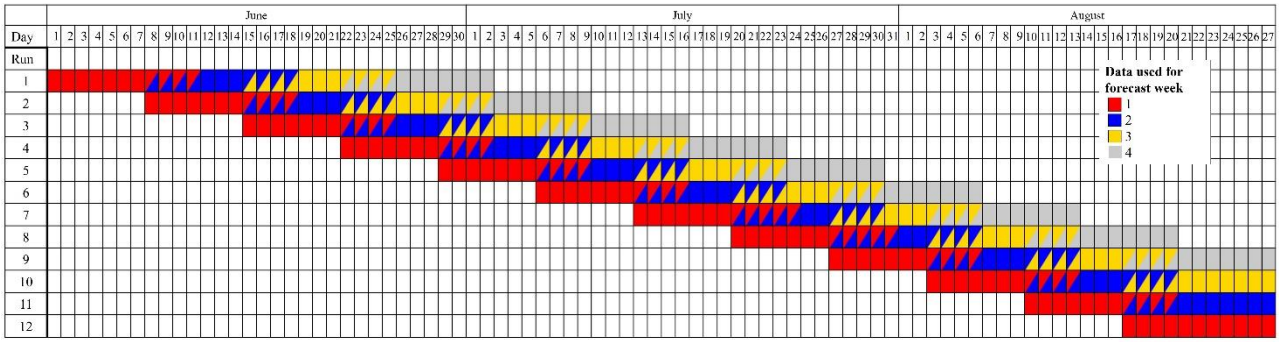
l. 317: refer back to Figure 3a?

Response: Ok, this is visible both in Fig 5 and Fig. 3a, so we added here “which was also visible in Fig. 3a”.

Figure 5: Why is the total n (sum of n for all 3 categories) for each subplot different? Shouldn’t this add up to the total number of forecast days within each forecast week times the number of considered grid points?

Response: As we did not include hindcasts initialized in early May (or those reaching September), there was actually the largest amount of data for forecast week 1 and smallest amount of data for forecast week 4. We have now clarified this in the manuscript by a data Table 1A (below) and explanations.

Table 1A. Table showing details of the data of the investigated hindcasts. Each row contains one run, altogether 12 runs. The colouring of the boxes shows the coverage of the hindcasts’ data. The first red boxes on each row show the initiation date of the hindcasts, which are same for all years 2000-2019. The colours of the boxes indicate for which lead time (i.e., forecast week) the data were used: red for 1 week, blue for 2 weeks, yellow for 3 weeks, and grey for 4 weeks. The data used for the different lead times were partially overlapping due to the use of 5-days moving averages with forward-looking window: lead time 1 week used data of days 1 to 11, lead time 2 weeks data of days 8 to 18, lead time 3 weeks data of days 15 to 25, and lead time 4 weeks data of days 22 to 32. The data used for two lead times are here marked with two colours. Note: for lead time 1 week we used data of 12 runs, for lead time 2 weeks we used data of 11 runs, for lead time 3 weeks we used data of 10 runs, and for lead time 4 weeks we used data of 9 runs (of years 2000-2019).



Ll. 334 – 335: I don’t quite understand what is meant by the notches here. The second sentence rather belongs into the results with a description of where we see this in the plot and what it implies.

Response: Thanks, we do not refer to the notches here anymore.

Section 3.6: I find it a bit confusing that the results are described from the longest to the shortest lead time here, when throughout the rest of the paper, the description starts with week 1. Maybe an option to invert the order?

Response: Thanks. Good point, we have now inverted the order in the manuscript.

l. 349: no need to put the “green box” in quotation marks.

Response: Thanks, we removed the quotation marks from the “green box”.

ll. 368 – 372 (caption Figure 6): what are the limits of the box plots? Same as in Figure 5, i.e. interquartile range and whiskers for 5th and 95th percentile?

Response: Yes, as in Fig 5, i.e., the horizontal line dividing each box into two parts shows the median of the data; the ends of the box show the lower and upper quartiles; and the whiskers indicate the 5th and 95th percentiles of the data in each group. We added this information to the caption of Figure 6.

l. 448: “as introduced to result from”; I don’t understand what this means.
l. 451: “the land-atmosphere interaction” → “land-atmosphere interactions”
ll. 444 – 452: Could you be more specific about how this could be used to refine the forecasts?

Response: Thank you for these remarks. This part of the discussion has been excluded as it is not about our methods and hence it will not be further edited.

ll. 458 – 462: This is almost an exact repetition of ll. 383 – 387. Keep it only in one place (I’d suggest Section 5).

Response: Thank you, we kept this only in Section 5.

ll. 473 – 478: Like the aforementioned part of the discussion (ll. 403 – 443), this paragraph seems very detached from the core results of the paper. Rather end the conclusions with some outlook for future work and how it could be continued to make it even more relevant in the context you bring up here.

Response: Thank you for the comment, we excluded this part (ll. 473-478) and end the conclusion with some outlook for future work:
“Building on these insights, 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 its occurrence, potentially enhancing early warning capabilities further.”