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 are going to edit the Discussion strongly. We shall leave out the detailed adaptation and preparation measures, and the additional bias techniques. We shall focus more on discussing the skill of these forecasts shown in the manuscript. We shall 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 are going to change 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 plotted figures 1, 2, 4, and 6 (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.

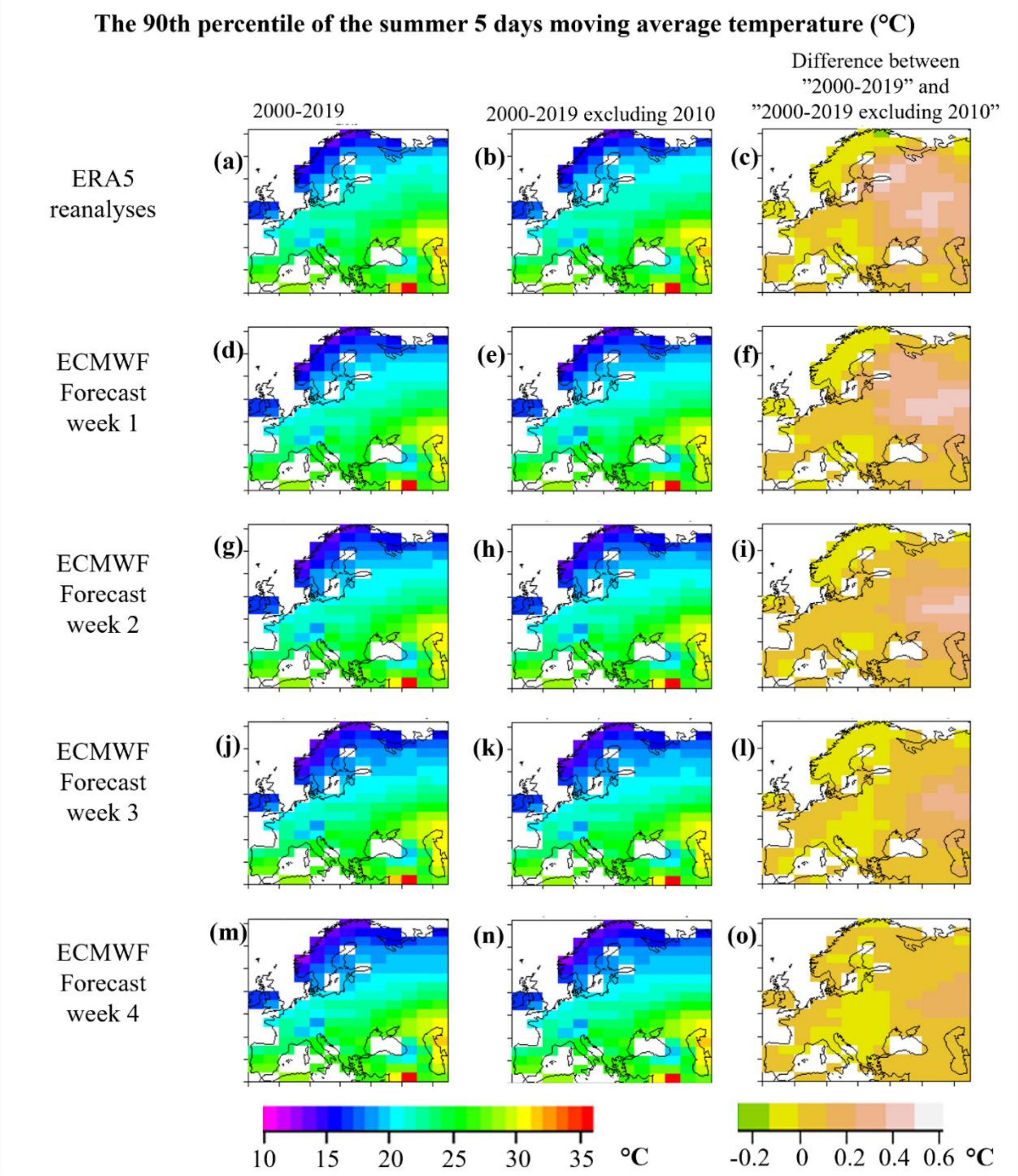


Figure 1: The lower thresholds of heat wave days: the 90th percentile of the 5-day moving average temperature in summers 2000-2019 (first column) and in summers 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010, middle column) of the ERA5 reanalyses (a and b), and (d,e,g,h,j,k,m, and n) of the 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.

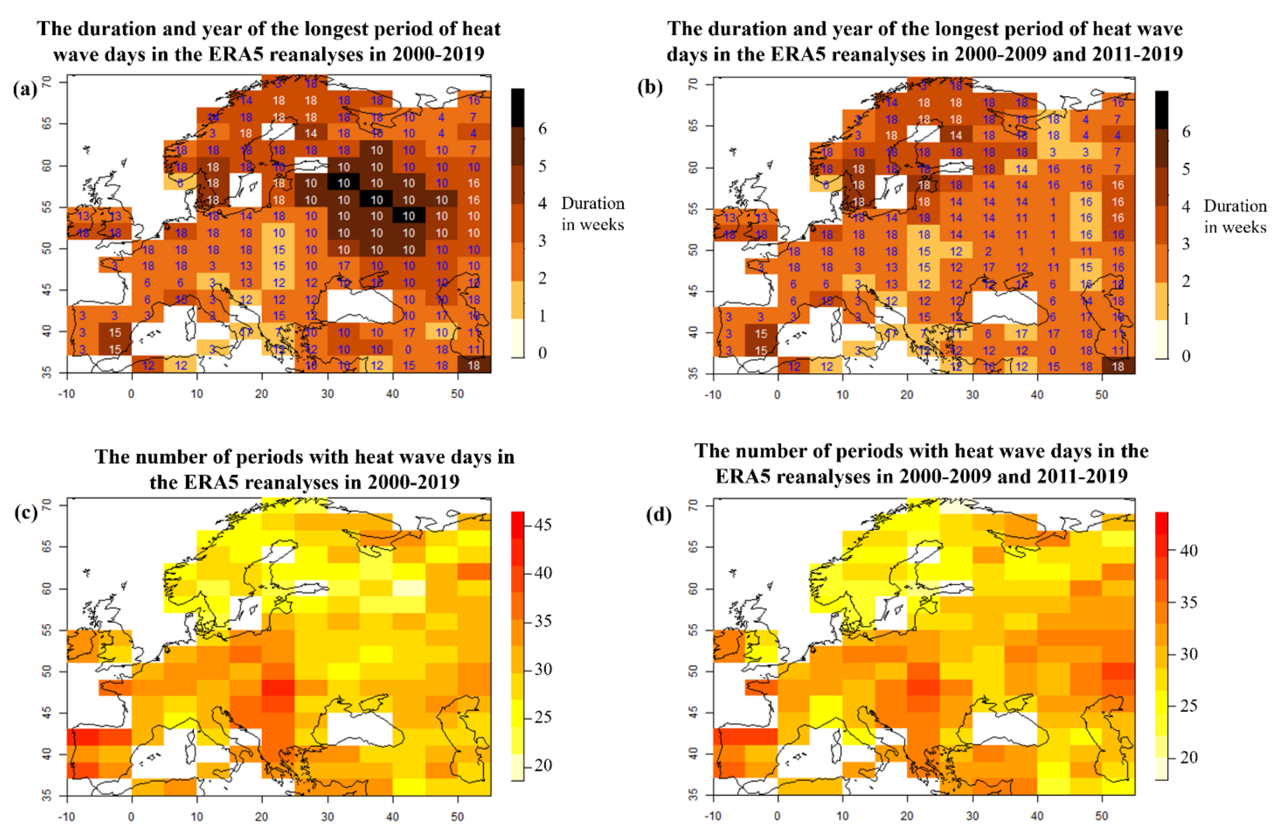

The duration and year of the longest period of heat wave days in the ERA5 reanalyses in 2000-2019
(a)
Duration in weeks
0
1
2
3
4
5
6
The duration and year of the longest period of heat wave days in the ERA5 reanalyses in 2000-2009 and 2011-2019
(b)
The number of periods with heat wave days in the ERA5 reanalyses in 2000-2019
(c)
20
25
30
35
40
45
The number of periods with heat wave days in the ERA5 reanalyses in 2000-2009 and 2011-2019
(d)
-10
0
10
20
30
40
50
35
40
45
50
55
60
65
70

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.**

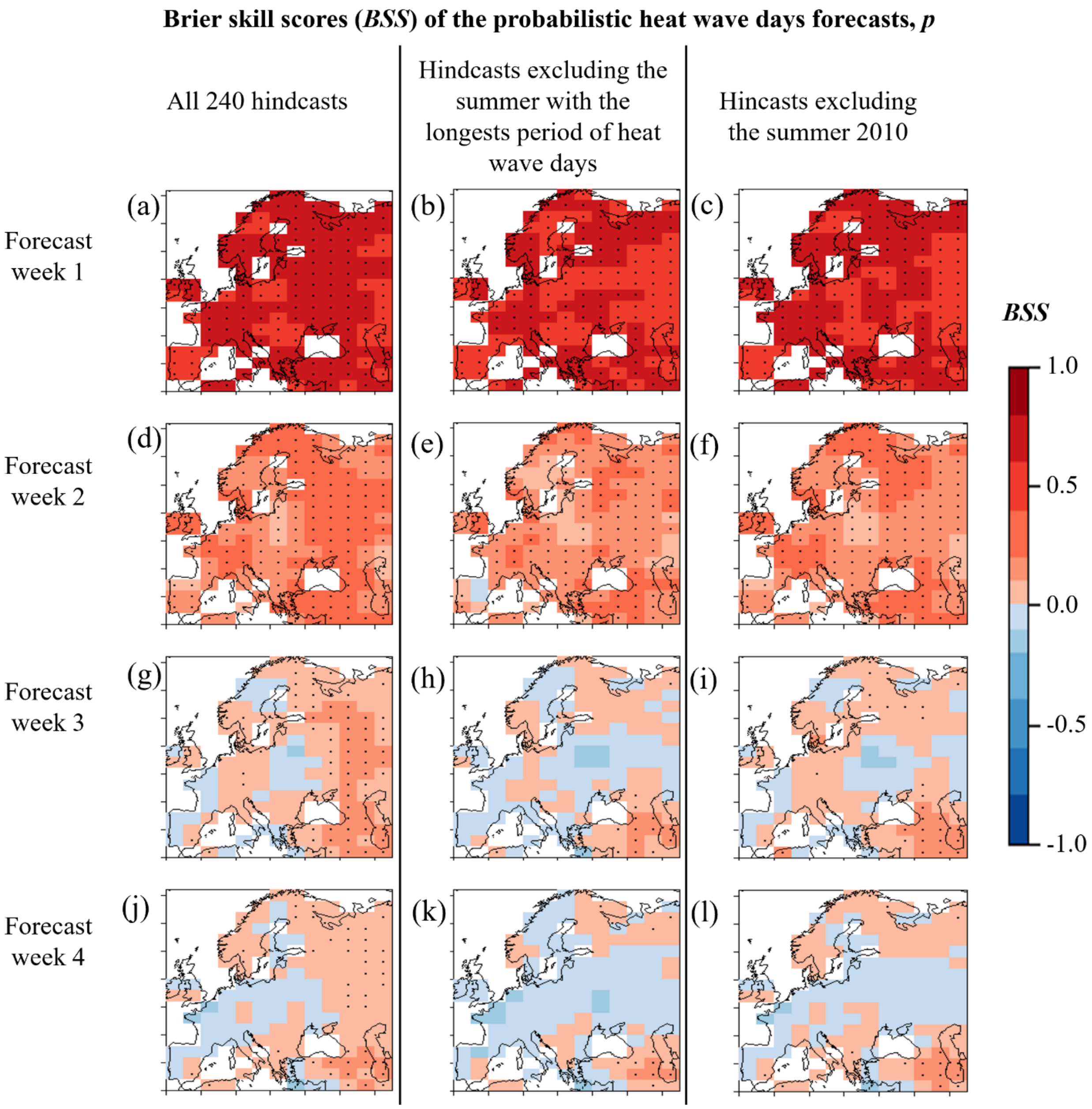

Brier skill scores (*BSS*) of the probabilistic heat wave days forecasts, *p*
All 240 hindcasts
Hindcasts excluding the summer with the longests period of heat wave days
Hincasts excluding the summer 2010
Forecast week 1
Forecast week 2
Forecast week 3
Forecast week 4
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
BSS
1.0
0.5
0.0
-0.5
-1.0

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%.

As you might have noticed in Figure 4, we have decided to leave out the “summer with the longest heat wave” as 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.

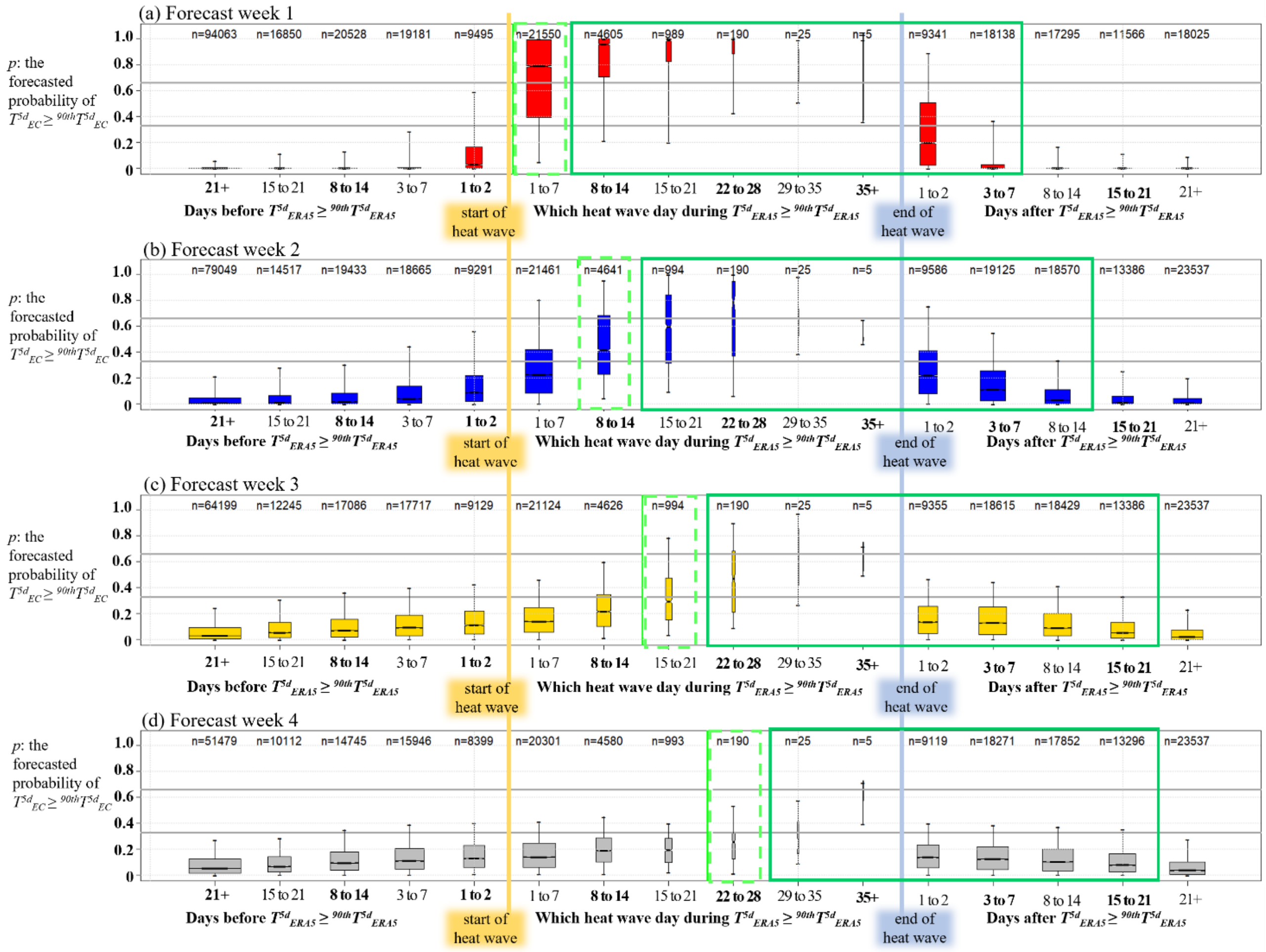


Figure 7 (DATA EXCLUDING YEAR 2010): The forecasted probabilities of heat wave days shown for days that (in ERA5) were 21 to 1 days before the heat wave, the 1st to 35th heat wave day during the heat wave, and 1 to 21 days after the heat wave with lead times of a) one week, b) two weeks, c) three weeks, 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, here 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 that the heat wave had initiated prior to the forecast issuance.

These figures showing the effect of excluding year 2010 will be 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.

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.

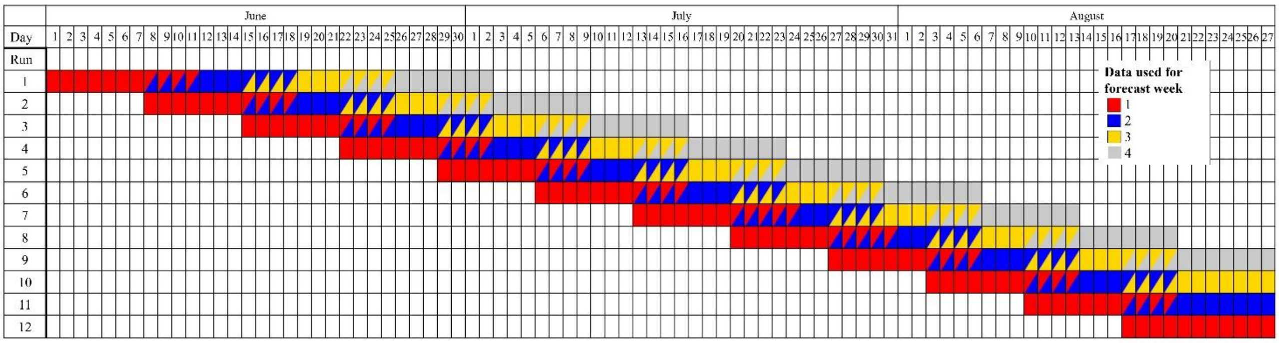
From my understanding, the former strategy allows a larger statistical sample to compute percentiles, but on the other hand, there is a potential impact of lead-time dependency. The latter one seems more accurate from this point of view but then of course the sample size is smaller. I guess the "lead time dependent climatology" indicated on l.125 refers to this strategy.
Response: Yes, the latter one is the one we used.

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.

Or did you only consider the first days of June as part of "week 1" lead time)?
Response: Yes we did.

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 will add a Table (see below) with explanations.



June July August

Day 1–30 (June), 1–31 (July), 1–27 (August)

Run 1 2 3 4 5 6 7 8 9 10 11 12

Data used for forecast week
🟥 1
🟦 2
🟨 3
⬜ 4

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 shall correct 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 shall correct this.

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

Response: Thank you, we shall remove 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. We shall point this out more clearly.

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 shall consider this, as technically this is easy to make.

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 shall edit 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 shall correct 3d to 3b.

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

Response: Thank you for mentioning this. We shall correct 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 shall edit this 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 shall correct this as you suggest.