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 require 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 will change 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 will remove the too detailed discussion of health impacts of heat from Section 4, as suggested.

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. It is a good idea to add interpretation about that here.

• 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 mention 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 will clarify this.

• 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.

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, and maybe rather mention this in the discussion.

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 shall add 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 will do that.

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, and we shall add a more thorough discussion of this figure!

Minor comments:

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

Response: Yes, (see Major remark 1.) we will change the title to be:

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 shall remove word “intense” here. And we shall edit the sentence to be: “Nonetheless, persistence of heat waves seems to have a higher level of predictability--”

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 shall add a sentence here.

l. 32: ‘in future’ to ‘in the future’

Response: Thank you, we shall correct 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 shall remove ‘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: Thanks, we shall mention it here.

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 will add this motivation and reference.

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

Response: Ok, we shall remove 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 shall shorten this for some parts, but for the introduction we would keep 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 can remove it (and make 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 shall add this information here.

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

Response: Thank you, we shall remove lines 55-62.

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

Response: Thank you, we hall use '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 shall specify 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 shall change 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, good point, thank you. We shall merge these parts as you suggest.

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 shall clarify this by a data Table 1A (below) and explanations.

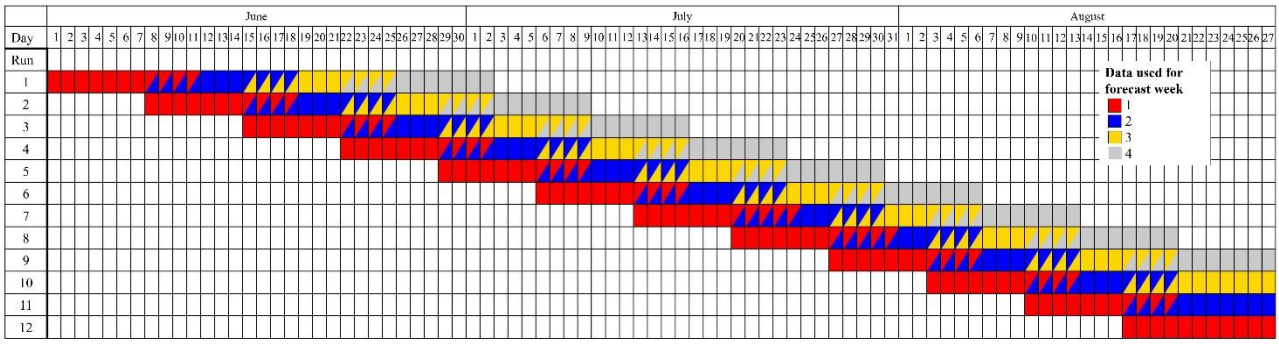

Table 1A. Table showing details of the investigated hindcasts. Each row contains one run, altogether 12 runs. The first red boxes on each row show the initiation date of the hindcasts, which are same for all years 2000-2019. The data of days marked with red are used for lead time 1 week, blue for 2 weeks,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

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. The reason for using only the Monday initialization, was to have each date in each lead time only once to avoid autocorrelation. We could have chosen to use the Thursday runs only as well.
We shall add this information to the manuscript 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 shall add this information here.

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 shall rearrange this as you suggest.

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, sure we can correct 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 shall re-structure this and merge it with ll 100-101, as suggested in a comment above.

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. We shall mention this.

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, we shall edit this.

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

Response: Yes, thank you, we shall do as suggested, i.e., remove the old Table 1 and write the information about it here in section 2.2 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, sure we can add here a more detailed description of how the bootstrap resampling procedure works.

l. 179: "change" → "chance"

Response: Thanks, the word 'change' is corrected to 'chance'.

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

Response: Yes, sure we can here explain in a few words how this procedure works.

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 shall move it there as you suggest.

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 shall clarify this 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 shall edit this to: “(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 reformulate this to make it more clear.

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 shall remove 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). We shall do that.

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 haven’t decided yet which alternative to adopt but shall either move this to Section 2 or have it in a short section at the beginning of Section 3.

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 shall mention this here.

l. 259: match → equal

Response: Thank you, we shall edit ‘match’ to ‘equal’.

l. 267: by → with

Response: Thank you, we shall edit ‘by’ to ‘with’.

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

Response: Thank you, we shall drop 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 shall remove 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 shall check this and correct the mix-up.

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

Response: Thank you, we shall make the suggested editing.

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

Response: Thanks, we shall edit 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 shall edit 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 will remove 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 shall add 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 shall clarify this by a data Table 1A (below) and explanations.

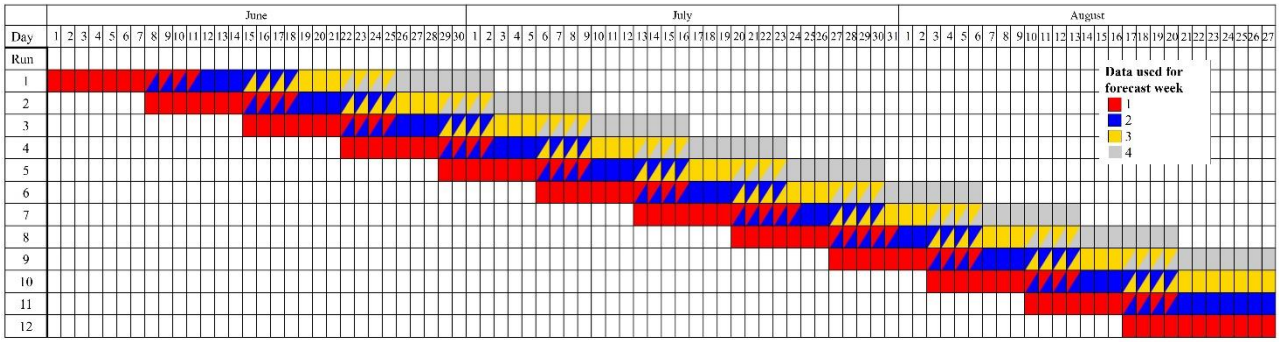

Table 1A. Table showing details of the investigated hindcasts. Each row contains one run, altogether 12 runs. The first red boxes on each row show the initiation date of the hindcasts, which are same for all years 2000-2019. The data of days marked with red are used for lead time 1 week, blue for 2 weeks, yellow for 3 weeks, and grey for 4 weeks. 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 shall clarify and edit this.

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 shall invert the order.

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

Response: Thanks, we shall remove 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.

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 might be 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 keep this only in one place.

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 shall exclude this part (ll. 203-443) and end the conclusion with some outlook for future work.