Review nhess-2024-75 “The usefulness of Extended-Range Probabilistic Forecasts for Heat wave forecasts in Europe”

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.

<u>Major remarks:</u>

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).
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. 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.2017.11.005

3. 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.
    - 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?
    - 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?

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

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). 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.
   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.

<u>Minor comments:</u>

Title

"forecast" is used twice, could maybe reformulate?

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.

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.

l. 32: ‘in future’ to ‘in the future’

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

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.

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.

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

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

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

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’

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

l. 91: change ‘forecasts’ to ‘hindcasts’ or ‘re-forecasts’

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?

Methods

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

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).

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

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

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.

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.

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

l. 125: bias → frequency bias

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.

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.

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.”

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

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?

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

l. 179: “change” → “chance”

ll. 182 – 183: 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.

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

l. 196: “a heat wave days become discernible” I don’t understand this, please reformulate

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

Results

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

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.

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.

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).

l. 259: match → equal

l. 267: by → with

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

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).

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.

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

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

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 o_t in the *BS* is 1 most of the time and thus the *BS* of climatology, i.e. $p_t = 0.1$, 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.

l. 317: refer back to Figure 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?

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.

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?

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

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 5^{th} and 95^{th} percentile?

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?

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