Interactive comment on “Seasonal forecast verification and application in times of change” by Yoav Levi and Itzhak Carmona

Reply to Anonymous Referee #2

We appreciate the time and effort invested by the referee in reviewing this manuscript. We thank the reviewer for the helpful comments, from typos to fundamental correction, which we will address, as highlighted below in blue:

In a context of developing seasonal forecasts as climate services, this paper proposes an evaluation of seasonal temperature forecasts from ECMWF System 4 for the June-July-August season. Authors first propose the Fiasco score to evaluate the forecasts for risk assessment and cost-benefit analysis and relate it to the common RPSS and AUROC scores. They also present the trends in temperature from System 4 and ERAinterim.
They conclude that, with such trends, classifying forecasts between above normal, normal or below-normal conditions based on long-term thresholds will not be useful for end users as all future forecasts will be above-normal. If I understand correctly, the authors then analyze the System 4 temperature forecasts, along with two light post-processing: the detrended System 4 temperature forecasts and the series of differences in temperature forecasts between the current season and the previous one.

The performance of these three forecast scenarios is evaluated with the Fiasco score which counts the cases when an above average (respectively below average) temperature is forecast and a below average (respectively above average) temperature is observed.

We are grateful for these comment and we will improve the text by adopting the essence of the review.

General comment
Coming from a slightly different field, some methodologies were unclear to me, and would require some clarifications. This particularly targets the Fiasco next score and the time series it is based on. Detailing the variables and thresholds considered in each Fiasco score (standard, detrended, next) could help in that matter.

A case study will be added including a figure which illustrates the various fiasco scores. Some more explanation of the detrened and next score will be also added.

Results are discussed in the Conclusion section. I would have liked more in-depth discussions of the results, either in the Results section or in a Discussion section, and a more pragmatic link between the results and the general objective of the paper: “help end-users to understand better how to use seasonal forecasts”.

The discussion will include some risk assessment and reference to the cost benefit ratio which may benefit the end users. A case study which will be added will serve as a warning to end-users regarding using of a seasonal forecast without understand the origin of the model skill. The case study will emphasize the impact of climate trends on seasonal forecast skill.

To which extent does the Fiasco score actually measure/reflect end-user needs? This score is very specific to worst-case scenarios.

It is true that only the worst-case scenarios is addressed by the fiasco score. However given the low skill of seasonal forecast we believe that in the current situation it is the most important score for end user decision to trust or not to trust the seasonal forecast.

Other general questions and comments:
- Section 2.2: Is it correct to say that you used the forecasts issued in May, June and July to obtain the forecasts for June-July-August at one month lead? If so, don't you have 1350 hindcast runs? If you do limit yourself to the first month lead, then this study would be an evaluation of monthly forecasts rather than seasonal forecasts.
Lastly, which time step do you consider? From the rest of the article, it seems that the temperatures for June July and August are aggregated.

The forecast for all 3 JJA month is given in May, therefore it was an error to write a "one month lead". This is clarified now is section 2.1.

- The computation of the AUROC score needs to be better explained.

A case study will added to give an example so hopefully the AUROC will be better understood.

- L. 154-155 "However if both : : : 1981-2010 conditions." Just a comment: if these trends maintain in future years, couldn't the thresholds be adapted to allow for a fair evaluation of the models?

This is a simple solution that should be implemented for future forecasts. As this is a fundamental comment it will be added to the discussion:
"If using the 30 years climatology for future forecasts the average temperature trend of the forecast model should be added to the thresholds determining the tercile categories."

- Section 5 first §: Did I understand correctly: you change the variable of interest from simply being the temperature forecast to being the difference between the temperature forecast for one year and the previous one, thus resulting in 29 values instead of 30 values. Are the thresholds defining the three equal probability groups chosen within this sample of 29 values for each month? Within the sample of 29*3 values for all JJA months altogether? Additionally, if my understanding of this paragraph is correct and if all verification methods remain unchanged, the change in name "Fiasco next score" might be confusing as it would be the same score simply applied to a different variable.

You understood correctly the score. However the error using the term one month lead caused a problem in understanding the Fiasco next score. All the analysis was done for one season forecast from May aiming for the 3 month JJA. Therefore only 29 values remain and not 29*3. The term one month lead was removed.

- L.207 "The RPSS, which takes: : :is positive only in the tropics between 22S and 21N." Isn't the RPSS also positive for latitudes south of 47S and between 20N and 43N with some exceptions around 26N?

- L.208 ": : :the latitude average number of fiasco is 7%..." I could not find these values in Figure 6.

Both comments are correct, the text refers to a figure 6 which contained the latitude average including sea and land (see figure added). In the current version only the average over land was presented. To clarify both images can be included in the manuscript. Defiantly the text will be changed accordingly.

- Why not consider the time series of differences in forecast between one season and the previous, but calculated from the detrended forecasts?

The use of the differences between one season and the previous causes the de-trending to be needless assuming the trend from one year to the other are negligible. Although not checked using the "Fiasco next score" on the de-trend data seems possible but will introduce both disadvantages of using 1 year as a reference and assuming a constant trend.

- To which extent is the trend in temperatures responsible for the "good Fiasco scores" (Figure 5a) as compared to the Fiasco next and Fiasco detrended scores (Figure 5b and 5c)? To which extent do these results inform us on optimal strategies for end-users to detect "fiascos"?

A correlation of 0.5 (r=0.5) was found between the temperature forecast skill and the temperature trend (by a skill score which is not presented). The correlation between the 3 scores and the temperature trend will be calculated and added.

- L.243 "the end-user should consider using the coming season forecast relative to previous season or a shorter reference period than the traditional 30 years: : :": I would have liked to see this already in this paper to strengthen the analysis.

As full analysis of different reference periods may be too long however this analysis for the added case study can be performed.

Technical issues:
- Throughout the paper, spaces are missing between words or after punctuations and special characters. Extra parenthesis also appear, e.g. pages 7, 8 or 9. Several typos appear in the text. I listed some of them below.

Hopefully all typos are corrected (it seems they originated from an old word version ...)

- L.45 Change "observes" to "observed"

Corrected

- L.48 Many other methods exist to evaluate hindcast skill. This sentence should not be restrictive to the criteria enumerated here.
Corrected

- L.50 I would suggest to change for example to : ": : :the Area Under the Relative Operating Characteristic (AUROC) curve which considers jointly the hit rate (HR) and the false alarm rate (FAR), : : :"
corrected, thanks

- L.51-52 Maybe add: “Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. J. Appl. Meteor., 8, 985–987.” When introducing the RPS
Although it is an old paper it is still relevant and therefore added, thanks.

- L.69 could you detail the following: “their figures regarding their work”?
Some examples were added.
- L.80 “It became operational in November 2011”
Corrected
- L.83 “one month lead”
Corrected and clarified
- L. 131 “Spatial averaging of the model increases”
The latitude average was performed on the various skills not the forecast itself. Therfore the skill should stay.
- Figure 2: Rewrite the legend to make (a) appear
Corrected
- Figure 2: I could not see the dashed contour lines mentioned in the legend. Are they supposed to be seen in (a), (b) and (c)?
The significant contours where removed from the figure but forgotten in the caption. Will be corrected.
- L.146 “with the same scale as Figure 2a”
Corrected
- L.161 Replace “leg” with “lag”?
Corrected
- L.163 “hindcast”
Corrected
- L.198 “radon” change to “random”
Corrected
- L.210 change “significant” to “significantly”
Corrected