Answer to Reviewer n.2

We would like to thank the reviewer for their time and valuable feedback to improve our manuscript. In the following, we answer (in blue) to the points raised by the reviewer (in black), and we indicate how we adapted the manuscript (in green).

The authors aim to investigate the predictive ability of ECMWF S2S reforecast, focusing on intense precipitation events. S2S hindcast cover a range between medium-range and seasonal prediction. Currently this is one of the first work covering the skill of predicting extreme events at these long time ranges, is therefore interesting. They propose a relatively new index in measuring forecast accuracy (Binary Loss Index) although probably the same consideration would have been emerged using the Critical Success Index, the latter widely used in the meteorological community.

While the index computations and results are well presented, it is also true that discussion is weak in terms of meteorological implications. A part the obvious dependence of seasonality, with convective precipitation leading to less skill during summer, and spatial and temporal aggregation, would have been nice to investigate with greater detail the spatial variability of the lead time. I found the lead time definition useful and interesting, not as a number per se, as you correctly comment it depends on the level of event detection you want to achieve and it is user dependent, but as an index to investigate predictability and its dependency on other factors. Taking two regions of example, would have been interesting to composite days with very long time day, and events with a shorter long time day and show the differences in some upper level variable to infer the role of the precursors dynamical evolution leading to the precipitation extreme. In a way this goal was also mentioned in the introduction "*Skill information is also useful to identify potential sources of predictability and windows of opportunity (i.e. intermittent time periods with higher skill Mariotti et al. (2020))*". The long time day could be used to detect those windows of opportunity; in which conditions they occur and for which regions are stronger. In that respect I think some elements in the paper could be inserted.

Thank you for this remark, suggesting clarifying the goals of our manuscript. Our aim is to quantify the skill of weather forecast in predicting precipitation extremes on a S2S timescale. We mentioned the concept of windows of opportunities in the introduction as a motivation, as it is opening a whole new research question beyond the scope of our manuscript. Our analysis paves the way for deeper studies of the skill limitation and potential opportunities. Hypotheses regarding the physical reasons (such as moisture origin and orography) behind the spatial and seasonal heterogeneity of the skill are formulated in the discussion section, in the second paragraph. To improve the manuscript thank to the reviewer's remark, we modified the manuscript by:

a) adding the word "skill" in the title: "Assessment of S2S ensemble extreme precipitation forecast skill over Europe". The title is now more precise regarding the content of the manuscript, which is centered on the skill quantification of extreme precipitation forecast (with the binary loss index and the last skilful day), rather than the analysis of the processes responsible for skill or absence of skill in forecasting precipitation extremes.

b) extending the discussion section: we added a paragraph about the (absence of significant) results regarding the dependance of the skill on the NAO configuration.

“We additionally investigated the effect of European weather regimes on the forecast skill (as defined in Grams et al., 2017), as the forecast skill of the weather regimes themselves can largely differ (Büeler et al., 2021). We computed the forecast skill independently for positive phases and negative phases of the NAO. The forecast skill does not exhibit a strong dependence on the NAO phase, although the data was also spatially aggregated to increase robustness (not shown). This absence of signal should be confirmed with a deeper analysis, by considering some time lag or seasonality for the influence of the teleconnection patterns (Tabari and Willems, 2018) or by aggregating over larger spatio-temporal neighborhoods, to increase the robustness. Other teleconnection patterns could be investigated, such as Scandinavian and East Atlantic patterns, El Niño southern oscillation, the Atlantic multidecadal oscillation (Casanueva et al., 2014) or the state of the stratosphere (Domeisen et al., 2019).”

Figures 1_r2 and 2_r2 illustrate the limited signal obtained when computing the BLI on the different NAO regimes.The results for the Brier skill score are similar.

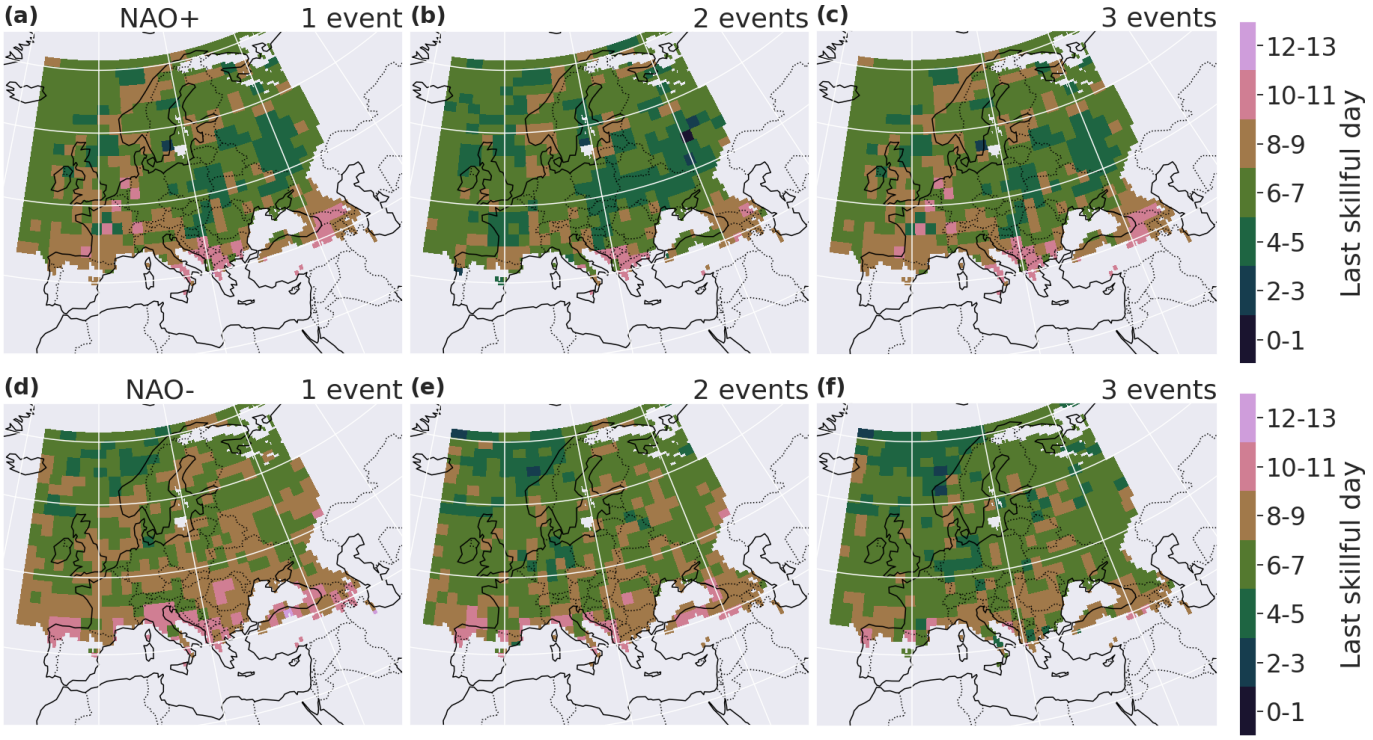

Figure 1_r2: Last day of skill for the BLI in positive NAO phase (a-c) and negative NAO phase (d-f) for a minimum of 1 (first column), 2 (second column) and 3 (last column) events in neighborhoods of 150x150km.

Forecasts for northern Norway have a higher skill during positive NAO phases than during negative phases, and forecasts for southern Europe have higher skill in negative phases of the NAO (Figure 1.). Apart from these regions, no clear patterns appear and the results are noisier than for the seasonal analysis. In general, in the negative phase of the NAO there is higher skill than in the positive phase and in the positive phase there is a larger latitudinal gradient of skill. However these results are not robust (Fig 2_r2.). The zonal mean of the skill length does not have a stronger gradient in one of the NAO phases when compared to the skill in the extended seasons. 20 years of hindcast data seem to not include a sufficient number of extreme precipitation days to distinguish the skill between the different NAO phases with our method.

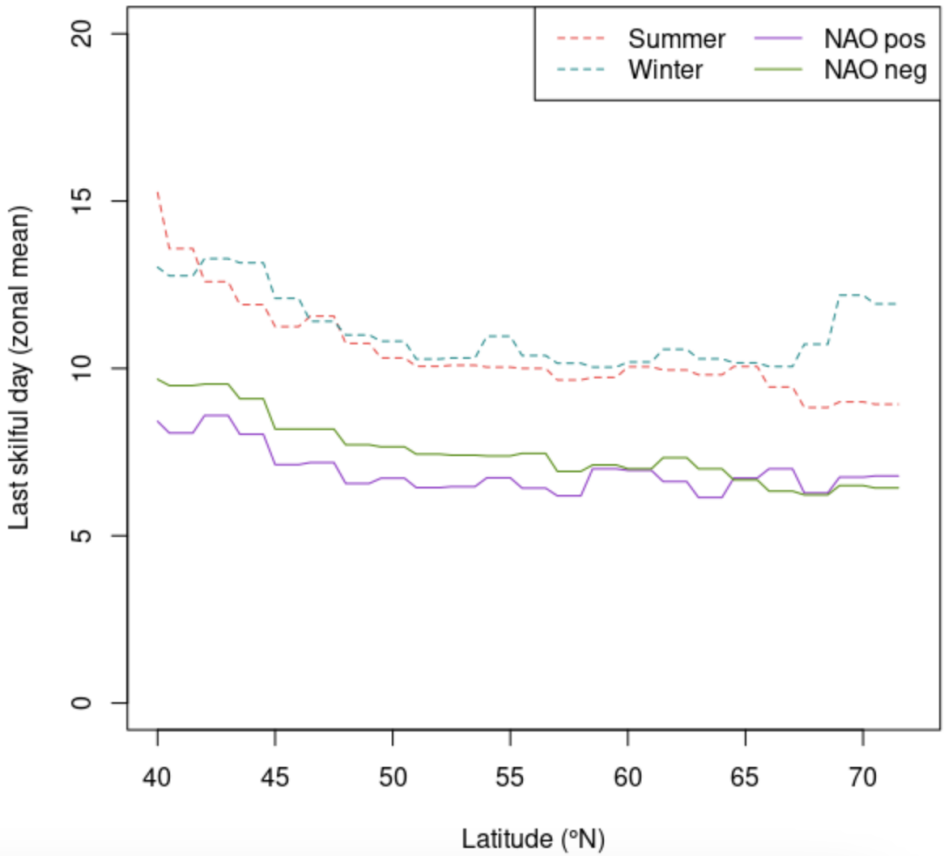

Figure 2_r2: Zonal mean of the last skilful day for the BLF, during the different NAO phases, and during extended summer and extended winter, for comparison (with spatial aggregation). The smaller last skillful day for NAO phases compared to summer and winter is due to shorter time series.

I have mixed feelings on the final judgment of the work. Since the title is Assessment of S2S ensemble extreme precipitation forecasts over Europe, I was expecting a more in depth discussion on practical predictability limit of precipitation extremes. For this reason I finally opted for major revision because I think the material is insufficient for this topic. But if you just wanted to present a new method to score extreme forecast, as you say in the abstract *" The goal of this article is to introduce a new methodology to assess the skill of rare events"*, then the material could be sufficient but text and title needs to be restructured to put the accent on method. In the latter case a more in depth comparisons with results obtained with other scores is also needed.

Thank you for this insightful comment. As specified above, we slightly modified the title, adding of the word "skill" for more precision. We also modified the abstract, refining the goal formulation: "The goal of this article is to assess the forecast skill of rare events, here extreme precipitation, in S2S forecasts, using a metric specifically designed for extremes".

Additionally, following the remark from Reviewer n.1, we justified further the use of the Binary loss index, anchored in extreme value theory (see the new version of the method section "2.3.2 Binary loss index"). The comparison with the commonly used Brier score serves as a reference. We added a sentence about the limits of the Brier score in the method section 2.3.2 ("The commonly-used Brier score rather assesses the average behavior, with a very weak penalty for under-represented classes. Because all

days are compared, the assessment of rare extreme events (missed, false alarm or hit) by the Brier score is lost among the huge amount of correctly predicted 0s.”), to justify further the specific focus on the BLI.

Minor comments:

line 41 reference WCS(2021) looks weird. What is WCS ?

Thank you for pointing this out. We now refer to the World Climate Service website by “(World-Climate-Service, 2021)”.