Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2017-122-AC2, 2017

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Revisiting the synoptic-scale predictability of severe European winter storms using ECMWF ensemble reforecasts" *by* Florian Pantillon et al.

Florian Pantillon et al.

florian.pantillon@kit.edu

Received and published: 28 July 2017

The manuscript investigates the forecast skill of extreme storms (often called windstorms) in the ECMWF 20-year reforecasts. The ECMWF 20-year reforecasts are found to be skillful and ensemble spread well calibrated up to lead times of 3-5 days. After this the skill drops; storms are found to move too slowly and do not capture the intensity of observed events as measure by a Storm Severity Index. No systematic links between storm properties (size, intensity, etc..) and forecast skill is found. Some skill beyond 3-5 days is found using EFI and SOT indices, suggesting some utility for windstorm warnings at these lead times.

The paper will be of interest to weather forecasting community as it contains some new and interesting results. In general, the paper is clear in its approach and figures are clear. I have a couple of specific comments on the paper (below) which should be addressed. I'd consider these major revisions, although I don't think it would take much to address these comments. Subsequently, I'd recommend the paper for publication provided these comments are fully addressed.

We thank the reviewer for his/her comments on the manuscript.

We will address all the comments below. In particular, we will refer more to the results of earlier studies and emphasize the novelty of ours. We will also clarify the limitations of using the ensemble average for the track and intensity of the storms. We will finally discuss the representation of wind gusts in the ensemble reforecast and in the reanalysis datasets. We hope that these revisions will better support the results of the paper.

Specific Comments

1. Novelty of the study: The paper seems incremental in terms of progressing this area of science, since a lot of what is said in this paper was covered by Froude et al (2007). It would be helpful in terms of highlighting the novelty of this particular study if a) the Froude et al 2007 paper is discussed in the introduction and b) that the novelty of this paper is discussed in the conclusions.

There a two major differences between this paper and that of Froude: (1) Froude investigated extratropical cyclones in general, while this paper focuses on severe storms, which requires a much longer dataset to cover enough events; (2) Froude investigated the track and intensity only, while this paper uses two additional methods for the early warning and for the impact of storms, which both require forecasts of wind gusts.

We will clarify these two points by adding a paragraph in the introduction to discuss the papers of Froude and Pirret – using the same approach but applied to severe storms –

and by explicitly stating the novelty of the paper i.e. the combination of three different methods and the use of a long homogeneous dataset. We will additionally compare our results with those of these and other previous papers in the conclusions to emphasize the novelty of this paper.

2. Page 8. Lines 19 to 33 and figure 6. Figure 6 is very useful as it gives another sense of the utility of the reforecasts. However, I don't agree with some of the statements here about the validity or not using an ensemble mean. The statements seem rather confused to me. For example, we could say that for your MSLP analysis in figure 3 we shall choose a threshold error of 10hPa to indicate a useful forecast, and therefore we shouldn't compute an ensemble mean for when the bias in the ensemble mean went above this. You'd agree that this would sound like a strange and arbitrary thing to do, but this is effectively what you're arguing in this piece of text. This strange argument should be removed. Furthermore, I find it difficult to see how your results make the results of Froude et al 2007 invalid (line 19) as they looked at a different dataset. Could you be clearer here what you mean?

The use of the ensemble average is limited by two factors when the lead time increases: (1) the identification of the storms becomes ambiguous and (2) the number of members containing storms decreases. Both factors may bias the average towards tracks that are close to the analysis and thus overestimate the actual skill of the ensemble forecast. Thus an alternative metric is given as the number of members forecasting the “actual” storm. This obviously depends on how the “actual” storm is defined, but reasonable values suggest that the storms are predicted by almost all members (with high certainty) until day 2–4.

We will first clarify the limitations of the ensemble average due to the two factors mentioned above and then better justify the chosen thresholds with the alternative metric. However, we agree that the 2–4 day limit does not strictly restrict the range of utility of the ensemble average, as it depends on the exact threshold used, and will therefore remove this argument.

3. Page 12. Line 9 and Figure 7a and 7b. "The predicted SSI is thus divided by a factor of 2 for ease of comparison unless stated otherwise" Have you done this for the plots in Figure 7? If so then you will need to redo the plots without this adjustment and revise the text. There's no justification for dividing one dataset by an arbitrary number to make it more comparable to the other. Furthermore, why are the SSI much larger in the reforecasts compared to ERA-I? Further down the page you say (Line 22), "ERA-Interim may also contribute to the cases of overestimation by underestimating the actual SSI due to its limitation at representing the mesoscale structure of some storms." You will need to provide some evidence of this statement (e.g. a reference). How much is ERA-I underestimating the true SSI? If ERA-I is very wrong, why are you using it as your main evaluation dataset? You'll need to address these questions.

The SSI is systematically overestimated by a factor of 2 in the reforecast compared to ERA-Interim, not only for the selected storms but for intense and extreme events in general, as illustrated by the 95th and 99th percentiles of the whole reforecast dataset in Figure 7 (dotted and dashed curves). The overestimation is due to a longer tail of the distribution of wind gusts in the reforecast compared to ERA-Interim, which impacts the SSI although it is calibrated with a local climatological percentile (Equation 1). The overestimation must be accounted for when investigating the SSI of the selected storms; one means of doing this is by calibration of the reforecast by a factor of 2, as would likely be done in an operational context to correct a systematic bias.

However, we agree that the calibration might be confusing here. We will therefore present the results without calibration for the SSI of the storms. Instead, we will state that the overestimation until day 3 could be corrected, because it is systematic in the whole dataset, while the underestimation at longer lead times is specific to the storms and thus indicates a poor predictability. Finally, we will add a paragraph in the methods Section to discuss the representation of wind gusts in ERA-Interim and the reforecast datasets.

Technical Comments

Page 1 Line 5. "... storms are correctly predicted. . ." correctly would mean without any bias. Perhaps "well predicted" or "predicted with only small forecast errors" would be a better expression.

We will reword to "well predicted" as suggested.

Line 9. "However, a large variability is" should be "However, large variability is. . ."

We will correct this.

Line 10. "and does not appear. . .". What is it that does not appear? Do you mean the ". . .and the predictability of storms does not appear. . ."?

We will reword to "large variability is found between the individual storms and the predictability does not appear...".

Line 21 ". . .and of their forecast in numerical weather prediction systems." Perhaps could be better expressed as ". . .and of the ability of numerical weather prediction systems to forecast them." In addition, I don't disagree with the sentence but references need to be added.

We will clarify to "and on the ability of numerical weather prediction systems to forecast them, as detailed below".

Page 2 Line 24-Line 28 and Figure 1. The sentences and the reference to Figure 1 do not belong in the introduction. They should be moved to the methods section.

We will move the references to Figure 1 to the methods section as suggested.

Page 4 Line 6. "In a second step, the mimima of MSLP are connected between subsequent model outputs every 6 h to form tracks, if their displacement velocity remains consistent in time." This second half of the sentence doesn't really make sense. Could you split the sentence and make clear what "their displacement velocity remains consistent in time" means?

We will clarify to "the mimima of MSLP are connected between subsequent model outputs every 6 h, using a predicted velocity based on both the previous displacement and the steering by the environmental flow".

Line 8. "filtered to exclude storms with a weak Laplacian" Can you specify the threshold is?

We will specify "below 0.8 hPa (° great circle)$^{-2}$".

Page 5 Line 24. Do you include SSI values over ocean in your European spatial average? If so this doesn't seem like a good idea – does it make a difference if you use land-only values of SSI?

Indeed, SSI values are also included over adjacent ocean areas. This is to avoid large sensitivities to the predicted position of storms that track close to the coasts. In additon, although the impact of storms is expected over land mostly, including the ocean partially accounts for storm surges, which represent the main impact of some severe storms (e.g. Xynthia).

We will clarify this in the text.

Line 31 "resoved" should be "resolved"

We will correct this.

Page 7 Line 14 and line 24. "Dispersion" often has a very technical meaning. I think here what you mean is "variability". There are other examples of this in the manuscript that should be changed for readability.

We will replace "dispersion" by "variability" as suggested.

Page 8. Line 14. Rephrase "The motion of cyclones was also too slow in the forecast but their MSLP was too deep." As something like "The motion of cyclones was too slow in the forecasts. In addition, but the forecasted MSLP was too deep."

We will substiantially rewrite the paragraph to clarify the interpretation of the results.

Page 9 Line 4. "...on a specific day but anywhere over central..." would be better expressed as "...on a specific day over central..."

We will change this as suggested.

Line 7 to 8. "...the predicted distribution of SSI is overestimated overall." I don't know what this means - what is the predicted distribution, is it the reforecasts? If so state this explicitly. Also state explicitly what the predicted distribution is overestimated relative to.

We will rephrase to "although the SSI is scaled locally with separate model climates, it is systematically overestimated in the reforecast compared to ERA-Interim"

Line 12. Can you add some detail to explain how you select events for the 99th percentile of SSI?

We will precise "the 99th percentile of SSI values in the whole reforecast dataset".

Line 26. "Early Warning". This terms means something very different in different contexts. In some contexts, early warning only means 1-2 day lead time. I'd suggest being specific here in terms of timescale and call this section "Potential for Early Warnings on 5-10 day timescales".

We will rename the section as suggested.

Page 10. Line 3-10. The description of Brier Skill Score should really be in the methods section.

We will move the Brier Skill Score to the methods section as suggested.

Page 11 Line 1. "This value is taken for consistency with the SSI." Can you say explicitly what this means?

We will explain that "The 98th percentile represents the strength at which gusts become

damaging in the SSI (Equation 1)".

Line 19. "...the optimal thresholds need to be levelled up and..." What do you mean by levelled up? Do you mean increased?

Yes, we will change this.

Page 12 Line 2. Should be "...which was noted..."

We will change this as suggested.

Line 27 "The ensemble average is unbiased until day 3 to predict the position and minimum MSLP of the storms on the day of maximum intensity.", would be better expressed as, "The ensemble average has small biases until day 3 in terms of predicting the position and minimum MSLP of the storms on the day of maximum intensity."

We will change this as suggested.

Line 30. Should be "...ensemble members captures the actual storm..."

We will correct this as suggested.

Line 30. "This bias is accompanied by an increase in ensemble spread by a similar magnitude, which suggests that the ensemble is calibrated, but only a minority of ensemble members still captures the actual storm at lead times beyond 3–5 days. This questions the relevance of using the ensemble average at longer lead times. This differs from a classical situation of averaging the ensemble members to smooth the unresolved scales, as the variables of interests are objects here rather than continuous fields" This appears to be a different argument than from earlier, where arbitrary thresholds were used to determine whether the ensemble contained the storm or not. Can you comment on this?

We will clarify that there exists a limit of validity of the ensemble average for the track of the storms, because they are identified as objects, which are not always clearly defined. This contrast with the metrics based on the strength of wind gusts, which are defined even in the absence of storm. We will further revise the paragraph based on the modifications in Section 3.2.

Page 13 Line 17. "The EFI and SOT indices confirm the skill of the reforecast at predicting the area covered by strong wind gusts until day 10 for storms as for the whole dataset." You argued a few paragraphs ago that few of the ensemble members actually predicted storm beyond 3-5 days lead time. If that's the case, how can there be skill at the lead times of up to a week? This needs to be explained in the conclusions.

We will clarify that the EFI and SOT, which emphasize the most extreme members, show a skill for predicting strong gusts until 9–10 days, while an accurate prediction of the position and intensity, which are based on the ensemble average, is limited to the first 2–4 days. We will further add a figure to clarify and summarize the results.

Line 29. "The predictability of the severe storms investigated here may not be linked to common factors but rather be due to characteristics of the individual storms." You've just argued in the previous paragraph that you don't have enough data to make this statement! So how can this statement also be true?

We agree that this statement is not justified and will clarify the limitation of the data for predicting extreme events.

Figures Table 1 "Some particularly high or low values are emphasized in bold." This is a con- fusing thing to do – either remove the bold numbers or decide on a sensible reason for using bold numbers.

We will specify that "The values corresponding to the deepest, most severe and smallest storms cited in the text are emphasized in bold".

Figure 3 and Figure 4. Font used in legends is too small and needs to be substantially larger to be readable.

We will increase the font size in legends as suggested.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2017-122, 2017.