Response to Reviewer 1

The study assesses the predictability of severe storms over Europe in the most impor- tant season winter using the ECMWF ensemble forecasts. The authors concentrate on 25 events in the period 1995 to 2015 applying different metrics finding that these high impact events are predicted with skill up to 4 days. They also find skill for the area covered by these extreme events up to 10 days which may provide early warn- ing opportunities. Still, the limited sample of only 25 storms shows strong inter-case variability. The small sample is a clear drawback of this study as it limits the reliabil- ity of the deduced skills and the author tend to overemphasize the results. Still the manuscript is nicely written and well structured. It certainly contains new findings, which are fruitful for how to identify predictive skill for extreme events, so I certainly see that the manuscript is suitable for NHESS, if my minor to major comments are treated seriously.

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

We will address all the comments below. In particular, we will clarify that we explore the physical characteristics of some outliers that exhibit a particularly high or low predictability and avoid any suggestion of a systematic link between dynamics and predictability among the sample of storms. We will also detail and 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.

Comments

P1,L9: Please change to 'potential for an early warning'.

We prefer to change to "potential for early warnings".

P2,L1-7: You may add the study of Stucki et al. (2014, Nat. Hazards Earth Syst. Sci.) here.

We will cite the Stucki et al. (2014) paper in the selection of storms, as stated below.

P2,L29: Please change 'manuscript' to 'study'.

We prefer to change "manuscript" to "paper".

P3,L15-16: As wind gusts are an impor- tant metric used in this study, you need to explain how this is derived in the reforecasts and how these gusts compare to observations.

In both the ensemble reforecast and ERA-Interim, the wind gusts are computed from the wind speed on the lowest model level and a turbulent component based on a similarity relation between the variability of the surface wind and the friction velocity. In the ensemble reforecast, which uses a more recent model version, the computation of wind gusts includes an additional component based on the low-level wind shear in convective situations. This additional component is expected to contribute to the strongest wind gusts when convection is embedded

in the cold front. The resolution of the ensemble reforecast and ERA-Interim is known not to be sufficient to capture the strongest gusts due to mesoscale structures such as sting jets and to steep topography. However, as the focus here is on synoptic-scale aspects of winter storms, these limitations are likely rather unimportant. The comparison with ensemble forecasts remains fair, because their horizontal resolution is not sufficient to capture the strongest gusts either, and because the verification of wind gusts is based on values relative to the model climate rather than on absolute values.

We will add a paragraph in Section 2.1 to detail and discuss the representation of gusts in the model data.

P3,L24-25: How do the selected European wind storms compare to the storm cata- logue provided by Stucki et al. (2014, Nat. Hazards Earth Syst. Sci.).

We will discuss the focus of the selection of storms on the United Kingdom due to a fixed threshold for wind gusts above 25 m s^{-1}, which is less often exceeded over continental Europe. We will further mention the Stucki et al. (2014) paper as another catalogue based on alternative criteria for the specific region of Switzerland.

P4,L13: It would be nice to include the publication by Raible et al. (2008) who were the first to inter-compare cyclone tracking methods.

The publication by Raible et al. (2008) will be included.

P4,L16: Please change to 'Neu et al. (2013) emphasized. . .'

We will implement the suggested change.

P5,L11: It remains unclear which level is used for the wind – is it 10-m wind? Another question is whether the authors use wind gusts as v_max or sustained wind. If the authors use wind gusts they need to include a discussion on the parametrization used.

As stated above, we will add a paragraph in Section 2.1 to detail the representation of gusts in the model data.

P5,L17-18: This could also a problem of the wind gust parameterization and not just a problem of the spin-up of the model. Stucki et al. (2016, Tellus) showed this how different gust parameterizations work over complex terrain showing strong changes from one to another parameterization.

We will cite Stucki et al. (2016) in the discussion of Section 2.1 about the representation of gusts over complex terrain. However, the problem seems to be different here, as it occurs in the first 6-h output of the reforecast only and not during the subsequent outputs. This suggests that the problem is due to the model spin-up when the higher-resolution reforecast is initialized from the lower-resolution reanalysis. We will clarify this in the manuscript.

P6, bottom line: This is why it is so important to say something about the gust param- eterization and why the authors shall be encouraged to compare their result to direct observations also on areas with complex terrain.

As stated above, we will add a paragraph in Section 2.1 to detail the representation of gusts in the model data.

P7,L27-29: If I understand the results correctly you only have two cases so such a strong statement that poor predictability is linked to process of extra tropical transition and convective dynamics cannot be derived, so the authors need to weak this state- ment and elsewhere in the manuscript.

We will clarify that the case of ex-Lili emphasizes the poor predictability of the position during extratropical transition due to the difficulty at representing convective dynamics but that it is unique among the selected storms and that

other cases that exhibit strong biases formed over very different regions, as e.g. Patrick over the southeastern United States and Jennifer (1996) over the eastern North Atlantic.

P8,L34: It seems to be a bit awkward that the authors argue a high storm to storm dependency as in the rest of the paper they use all the cases to get some robust conclusion about predictability of severe storms which implies averaging over as much cases as possible, also the dependency to the threshold is expected as it is a matter of statistics that there is dependency to thresholds.

We agree that the results depend on the exact thresholds but we believe that a limit of 2–4 days is realistic for the large majority of storms with a reasonable definition of a useful forecast of the actual storm for an operational forecaster. We will omit the storm-to-storm variability here, which is not necessary at this point of the discussion, and further revise and clarify the choice of thresholds.

P10,L32-33: Change to ' further suggested to maximize . . . optimal threshold is used to predict gusts'

We will change to "further suggested to maximize the Heidke Skill Score to define the optimal threshold".

P11,L5: From Figure 9 I think that the hit rate decrease but the false alarm rate in- crease, correct?

This is indeed confusing. We will clarify that this is due to a different balance between hit rate and false alarm rate.

P11-12, section 4.3: Well single storms are always special so I do not see why there is a need for this section.

The purpose of this section is double. Firstly, it illustrates how the verification of forecasts can be biased by focusing on observed events only. Secondly, it explores possible links between the predictability of the storms and their physical characteristics.

We will clarify the separation between the characteristics of some outliers and the absence of a systematic link dynamics and predictability. We will further extend the discussion of the results and include a comparison with findings of previous studies.

P12,L16-26: Please shorten this part – it is a summary and not a conclusion.

We will shorten this part as suggested.

P13,L8: Please cite the earlier studies and change 'should' to 'shall'.

We will implement the suggested change and cite the earlier studies.

P13,L20: I think the cases to case variability is expected.

Again, we will clarify the separation between the characteristics of some outliers and the absence of a systematic link between dynamics and predictability. We will further discuss the limitation for the verification of extreme events and compare alternative methods.

P13,L21: The conclusion on low predictability for storms of tropical origin only relys on 2 cases so weaken this statement here.

We will clarify that the link with extratropical transition concerns a unique case in the dataset.

References: Please get rid of the numerous errors in the reference list – this is annoy- ing!

There appears to be a problem with the URL of several references. We will therefore omit the URL whenever the DOI is available.

Figs. 5, 6, 7 and 10 needs to have increase axis labels as e.g. Fig. 8 has.

We will increase the axis label in those figures as suggested.

Response to Reviewer 2

The manuscript investigates the forecast skill of extreme storms (often called wind- storms) 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 ($^\circ$ 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 re-forecasts? 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.

Revisiting the synoptic-scale predictability of severe European winter storms using ECMWF ensemble reforecasts

Florian Pantillon, Peter Knippertz, and Ulrich Corsmeier
Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany
Correspondence to: Florian Pantillon (florian.pantillon@kit.edu)

Abstract.

New insights into the synoptic-scale predictability of 25 severe European winter storms of the 1995–2015 period are obtained using the homogeneous ensemble reforecast dataset from the European Centre for Medium-Range Weather Forecasts. The predictability of the storms is assessed with different metrics including (a) the track and intensity to investigate the storms' dynamics and (b) the Storm Severity Index to estimate the impact of the associated wind gusts. The storms are ~~correctly~~ well predicted by the ~~ensemble reforecasts~~ whole ensemble up to 2–4 days ahead~~only, which restricts the use of ensemble average and spread to short lead times~~. At longer lead times, the number of members predicting the observed storms decreases and the ensemble average is not clearly defined for the track and intensity. The Extreme Forecast Index and Shift of Tails are therefore computed from the deviation of the ensemble ~~reforecasts~~ from the model climate. Based on these indices, the model has some skill in forecasting the area covered by extreme wind gusts up to 10 days, which indicates a clear potential for ~~the early warning of storms~~early warnings. However, ~~a~~ large variability is found between the ~~predictability of~~ individual storms and ~~does not appear to be related to the storms' characteristics . This may be due to the limited sample of 25 cases, but also suggests that each severe storm has its own dynamics and sources of forecast uncertainty.~~ the poor predictability of outliers appears related to their physical characteristics such as explosive intensification or small size. Longer datasets with more cases would be needed to further substantiate these points.

1 Introduction

One of the most important natural hazards over Europe arises from winter storms associated with low-pressure systems from the North Atlantic, also referred to as cyclonic windstorms ~~(e.g. Roberts et al., 2014)~~ (Lamb and Frydendahl, 1991) . These storms are therefore the focus of various fields of research involving the weather and climate communities but also the windpower and reinsurance industries. At longer time scales, ~~numerous studies are dedicated to the estimation of the footprint and return period of winter storms and often require a combination of dynamical and statistical models (Della-Marta et al., 2009; Hofherr and Kunz, 2010; De~~ ~~A crucial but disputed~~ a crucial question lies in the trends in frequency and intensity of winter storms in the current and future climate~~, which still differ~~ . This question is disputed due to little agreement between climate models and between identification methods (see Feser et al., 2015, for a review). However, the intensity of storms is not necessarily related to their impact and storm losses are better estimated from the strength of winds or wind gusts exceeding a certain threshold

(Klawa and Ulbrich, 2003) . Numerous studies are therefore dedicated to the estimation of the footprint of strong winds and gusts associated with winter storms as well as their return periods (Della-Marta et al., 2009; Hofherr and Kunz, 2010; Donat et al., 2011; Ha These studies often require a combination of dynamical and statistical models to adequately represent the footprints.

At shorter time scales, most studies concentrate on the detailed investigation of case studies of severe storms and ~~of their forecast in~~ on the ability of numerical weather prediction systems ~~.~~ to forecast them. Although the general lifecycle of extratropical cyclones has been ~~described~~ known for almost one century, the intensification of ~~the~~ storms and the generation of strong winds ~~– responsible for most of the damages created by the storms –~~ involve physical processes of different scales that are still not fully understood (Hewson and Neu, 2015) . Recent advances have resulted from the attention drawn by devastating storms. The damaging winds over southeast England during the "Great Storm" of October 1987, which were observed at the tip of the cloud head bounding the bent-back front, now form the archetypal example of a phenomenon known as the sting jet (Browning, 2004). The destructions caused by storm Lothar over central Europe in December 1999 revealed the importance of diabatic processes in a way similar to a diabatic Rossby wave for the rapid intensification of the storm over the North Atlantic (Wernli et al., 2002). The severe wind gusts observed during the passage of storm Kyrill in January 2007 over central Europe finally emphasized the role of ~~the~~ convection embedded in the cold front ~~and~~ including the formation of cold-season derechoes (Fink et al., 2009; Gatzen et al., 2011).

These historical storms were poorly forecast when they occurred and thus captured an even larger attention in the weather research community, which resulted in a prolific scientific literature on specific storms. In particular, Buizza and Hollingsworth (2002) early recognized the potential of the ensemble prediction system of the European Centre for Medium-Range Weather Forecasts (ECMWF) to ~~forecast~~ predict the storms Anatol, Lothar and Martin in December 1999. They showed that ~~the ensemble forecast offers~~ ensemble forecasts offer a more consistent picture between different ~~initialisations than the deterministic forecast and additionally provides~~ initialisation times than deterministic forecasts and additionally provide early indications of the chance of an intense storm. Lalaurette (2003) further showed that the extremeness of the ensemble ~~forecast~~as a whole, measured by its deviation from the model climate, allows ~~to identify~~ identifying areas of unusually strong winds up to 120 h lead time in the case of Lothar, although ~~it fails~~ failing in the case of Martin. Petroliagis and Pinson (2014) and Boisserie et al. (2016) recently extended this methodology to longer periods with~~either operational or~~ , respectively, operational and retrospective forecasts. While they found constrasting results from case to case, the authors confirmed the potential of ensemble forecasts for the early warning of severe European storms.

~~Following~~ A more statistical approach was proposed by Froude et al. (2007a, b) , who identified storms as objects with a tracking algorithm and systematically compared their position and intensity between forecasts and analyses. They investigated the predictability of a large number of extratropical cyclones in both deterministic and ensemble forecasts and found a slow bias in the track forecasts. For this large sample ranging from shallow to deep cyclones, they further found large errors in the intensity forecasts but constrasting biases that depend on the region and on the model. Pirret et al. (2017) recently applied this tracking approach to severe European storms in operational ECMWF ensemble forecasts and found a negative bias in intensity in addition to the slow bias in track. They further investigated the relative contribution of diabatic and baroclinic processes to the intensification of the storms. Although they succeeded in showing a significant correlation between the track and the

dynamics of the storms, they struggled to find an impact on predictability. Their results were, however, limited by the use of operational forecasts, whose skill improves with updates in the model version and with increases in the horizontal resolution in particular.

Building on these previous studies, the predictability of severe European winter storms is systematically investigated here for a 20-year period in an ensemble prediction system by taking advantage of the recently available ECMWF retrospective ~~forecast (reforecast)~~forecasts (reforecasts; Hagedorn et al., 2008, 2012) . While reforecasts are originally designed for calibrating the operational forecasts, which result in a significant improvement in forecast skill, they also represent a homogeneous dataset that is ideal for comparing historical events (Hamill et al., 2006, 2013). The predictability of severe storms is thus not restricted to single case studies here but encompasses a ~~large~~ number of events that allow a statistical analysis. ~~Three metrics are combined~~ Situations with less severe or no storms are also included to check whether the results are biased by the focus on extreme events. Furthermore, three methods are combined here to assess the predictability ~~in regard to different properties (Figure 1): the dynamics are evaluated with the~~ of the storms. In addition to the track and intensity ~~of the storms and~~ and to the unusually strong winds investigated by other authors, a third, novel approach is added for the impact of the ~~impact is estimated with the strength of wind gusts, while the potential for early warnings is computed from the area of predicted gusts that are unusually strong compared to the model climate. These metrics are further generalized to situations with less severe or no storms to ensure that the results are not biased by the focus on extreme events~~storms measured by their gust footprint. To the best of the authors' knowledge, the impact of severe winter storms has been extensively studied in the climatological community but its predictability has not been documented in the peer-reviewed literature.

The ~~manuscript~~ paper is organized as follows. Section 2 describes the reforecast and ~~reanalyses~~ reanalysis model data and the selection of severe storms, as well as the ~~3 different methods that are~~ three different methods used to assess the predictability of the storms in ~~the~~ these data. Section 3 presents the results obtained for general storm characteristics and using either the ensemble average and spread or individual ensemble members. Section 4 discusses the skill for early warning on the 5–10 day timescale using either selected storms or the whole dataset. Section 5 finally gives the conclusions of the study.

2 Data and methods

2.1 Model data

This study extensively makes use of the ensemble reforecast from the ECMWF (Hagedorn et al., 2008, 2012). The ensemble reforecast is based on the current version of the operational model but with a lighter configuration to reduce computing time. It is initialised from the ~~ERA-Interim reanalysis~~ ECMWF Retrospective Analysis (ERA)-Interim (Dee et al., 2011) and ensemble members are obtained from initial perturbations computed with singular vectors. In contrast to the operational model, stochastic perturbations of physical processes are not applied to the ensemble members. Since mid-May 2015, the ensemble reforecast contains 10 perturbed members in addition to a control member and it is run twice a week – every Monday and Thursday at 00 UTC – for the current date in the past 20 years. Until mid-March 2016, when the model resolution was upgraded, the horizontal grid spacing was approximately 30 km for the first 10 days and was then coarser at longer lead times until 46 days.

All 10-day ensemble reforecasts computed between mid-October 2015 and mid-March 2016 are used here, which represents a homogeneous dataset of nearly 10,000 individual reforecasts for the winter seasons 1995/96 to 2014/15.

The reforecasts are verified against ~~the ECMWF Retrospective Analysis (ERA)-Interim, which is~~ ERA-Interim reanalyses, which are available since 1979 and ~~is~~ are computed with a horizontal grid spacing of approximately 80 km, corresponding to a ~~pre-2006~~ 2006 version of the operational model (Dee et al., 2011). ~~Variables of interest include the~~ The verification is based on the 6-hourly Mean-Sea-Level Pressure (MSLP) ~~output at 00, 06, 12 and 18 UTC and wind gusts output~~ for the track and intensity of the storms and on the daily maximum wind gusts for the other metrics. The wind gusts are available in the ERA-Interim dataset from short-range ~~forecasts initialized~~ reforecasts initialized from the reanalyses at 00 and 12 UTC. ~~Although it~~ They are computed from the wind speed on the lowest model level and a turbulent component based on a similarity relation between the variability of the surface wind and the friction velocity (Panofsky et al., 1977) . In the ensemble reforecast, which uses a more recent model version, the computation of wind gusts includes an additional component based on the low-level wind shear in convective situations (Bechtold and Bidlot, 2009) . This additional component is expected to contribute to the strongest wind gusts when convection is embedded in the cold front.

Although the ERA-Interim dataset has been widely used for climatological studies of winter storms, ~~ERA-Interim~~ it has recently been criticized for underestimating the deepening rate of storms and the strength of ~~winds~~ wind gusts (Hewson and Neu, 2015). In particular, the relatively low horizontal resolution of ERA-Interim is not sufficient ~~at representing~~ to represent the mesoscale structure of the storms. ~~The next generation of ECMWF reanalysis ERA5 may alleviate this limitation thanks to its~~ Capturing sting jets for instance, which are responsible for some of the most damaging wind gusts within storms, would require a horizontal grid spacing of ~~about 30 km but it is still under development.~~ 10–20 km (Hewson and Neu, 2015) . Furthermore, gust parameterizations underestimate the observed strength of wind gusts over complex terrain, even at much higher resolution (Stucki et al., 2016) . As the focus here is on the synoptic-scale aspects of winter storms, these limitations of ERA-Interim ~~is used as a reference, while caution is taken for the interpretation of strong winds that could be related to mesoscale structures.~~ are likely rather unimportant. The comparison with ensemble forecasts remains fair, because their horizontal resolution is not sufficient to capture the strongest gusts either, and because the verification of wind gusts is based on values relative to the model climate rather than on absolute values. Finally, although they rely on parameterizations, modelled wind gusts are preferred to wind speeds, because they are output as maximum values over a certain time period rather than 6-hourly instantaneous values and thus better sample storms with a large displacement velocity.

2.2 Selection of storms

Significant historical storms are selected to investigate their predictability in the ensemble reforecast. The selection is made using the "XWS open access catalogue of extreme European windstorms" provided by Roberts et al. (2014), which contains the 50 most severe storms for the 1979–2012 period. The catalogue is based on ERA-Interim dynamically downscaled with the Met Office Unified Model and recalibrated with observations. ~~It is~~ The majority of the 50 storms affected the UK more than any other European country. This is not surprising, considering the location of the UK at the end of the Atlantic storm track. However, it may be exaggerated by the selection of storms based on wind gusts above a fixed threshold of 25 m s^{-1}, which is

less often exceeded over continental Europe (Roberts et al., 2014) . The selection thus differs from other catalogues based on alternative criteria that may be more relevant for specific regions (e.g. Stucki et al., 2014, for Switzerland) .

The catalogue is available online at http://www.europeanwindstorms.org/ and ~~was~~ has been updated with two additional storms for the winter season 2013/14. Following the time period of the ensemble reforecast, the storms that occurred between mid-October and mid-March from 1995/96 to 2014/15 are selected here. One storm occurring in late March is excluded through the restriction to the winter period, expecting that the mid-October to mid-March time span of the reforecast is relevant for severe storms. The selection results in the 25 storms listed in Table 1. The storm names are those given by the Free University ~~off~~ of Berlin when available, with alternative names in brackets when relevant. They were completed for a few storms with respect to the original catalogue of Roberts et al. (2014).

2.3 ~~Storm tracking~~Evaluation of predictability

Three metrics are combined to assess the predictability of the storms with regard to different properties: the dynamics are evaluated with the track and intensity of the storms (Figure 1a; Section 2.3.1), the impact is estimated with the footprint of wind gusts (Figure 1b; Section 2.3.1) and the potential for early warnings is computed from the area of predicted gusts that are well above the model climate (Figure 1c; Section 2.3.1).

Either the ensemble average or the individual members are used for the verification of the reforecasts of the selected storms based on these metrics. When the whole reforecast dataset is considered, the skill is estimated with appropriate scores. In particular, the Brier Score (Brier, 1950) measures the ability to predict if an event will occur or not. It can be split into reliability, resolution and uncertainty components (Murphy, 1973) . The reliability component measures the ability of the forecast to predict the observed frequency of events. A perfect reliability can be achieved with a climatological forecast and is thus not sufficient to be useful. In contrast, the resolution component measures the ability of the forecast to distinguish between events and non-events, which can not be achieved with a climatological forecast. The uncertainty component finally measures the sampling uncertainty inherent to the events. The Brier Score can further be compared to a climatological forecast to obtain the Brier Skill Score (BSS), which is in turn split into

$$BSS = 1 - B_{rel} - B_{res} \quad (1)$$

with reliability and resolution components B_{rel} and B_{res} (e.g. Jolliffe and Stephenson, 2012) .

2.3.1 Storm tracking

The 25 selected storms are tracked both in ERA-Interim and in the members of the ensemble reforecast, using the algorithm described by Pinto et al. (2005) and originally developed by Murray and Simmonds (1991). In a first step, maxima are identified in the Laplacian of MSLP interpolated on a polar stereographic grid then minima in MSLP are looked for in their vicinity. The Laplacian of MSLP is closely related to the quasi-geostrophic vorticity; thus the algorithm is similar to tracking maxima in low-level vorticity. 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~~, using a predicted velocity based on both the previous

displacement and the steering by the environment. As the focus is on severe storms here, the obtained tracks are filtered to exclude storms with a weak Laplacian of MSLP below 0.8 hPa (° great circle)$^{-2}$ or with a duration of less than 24 h. However, the algorithm ~~is applied hemisphere-wide and thus~~ results in a large number of tracks, among which the storms of interest need to be identified.

Identifying the storms in ERA-Interim is straightforward, because the selection of severe storms is based on the same dataset. For each of the 25 storms, the reference time and position of minimum MSLP given by Roberts et al. (2014) are searched for in the tracks obtained from the algorithm. The closest track is unambiguously identified this way and matches the reference track, although differences may arise, particularly at the beginning and end. As ~~shown by~~ suggested by Raible et al. (2008) and generalized by Neu et al. (2013), such differences are a common issue when comparing storm tracking algorithms, which usually agree well for the mature phase of deep cyclones but differ during the phases of cyclogenesis and cyclolysis. In particular, the algorithm of Pinto et al. (2005) tends to identify the cyclones earlier than others. Neu et al. (2013) emphasize that there is no best way of tracking storms, because there is no single definition of extratropical cyclones. As the same algorithm is applied here to both ERA-Interim and the reforecasts, potential biases due to the tracking method would likely cancel out.

In the reforecast, identifying the storms is less straightforward even at short lead times and quickly becomes ambiguous, because the tracks diverge from ERA-Interim when the lead time increases. In earlier studies, Froude et al. (2007a, b) applied strict criteria in the location, timing and duration of tracks to identify storms in forecasts. While such criteria may be required for statistical studies, they would reject too many ensemble members for the sample of storms considered here, in particular at long lead time, and thus would bias the results towards "good" members~~only~~. Instead, the track closest to ERA-Interim is identified in each ensemble member without arbitrary criteria, based on the great-circle distance averaged over a 24-h period. Two methods are compared for the definition of the 24-h period. In the first method, the period is defined as the first 24-h overlap between the track in the ensemble member and in ERA-Interim. If the track is not present at the time of initialization, it is further constrained to start in the ensemble member within 48 h of its first occurrence in ERA-Interim. In the second method, the period is simply defined as the day of maximum intensity.

The two methods are illustrated for the 7-day reforecast of the storm that hit the British Isles on 28 October 1996 (~~"u19961028",~~ Table 1). The storm took its origin in Hurricane Lili, which reached Europe after crossing the North Atlantic and undergoing extratropical transition (Browning et al., 1998). With the first method, the identified tracks start from the same location, because the storm is present in the reforecast at the time of initialization (Figure 2a). They later diverge and only two of them reach Europe, whereas the others remain over the central North Atlantic. With the second method in contrast, the identified tracks all reach Europe, as expected from the identification on the day of maximum intensity (Figure 2b). However, they start from different regions spreading from the western to the eastern North Atlantic. In particular, no single track takes its origin in Hurricane Lili, i.e. the two methods do not show any common track. Although this case of extratropical transition is unique among the selected storms, it illustrates the difficulty of identifying storms in the reforecast. The most relevant method depends on the aims of the analysis; the first method focusing on the dynamics of the storm and the second one on its impact. Both methods are therefore used here.

2.4 ~~Storm Severity Index~~

2.3.1 Storm Severity Index

While the intensity of a storm is commonly measured with its minimum MSLP, its severity mostly depends on the strength of the wind gusts, which is also controlled by the pressure gradient at the synoptic scale and by additional factors at the mesoscale and turbulent scale. In particular, insured losses have been shown to scale with the third power of the strongest wind gusts. Following Klawa and Ulbrich (2003) ~~and ?~~ for observations and Leckebusch et al. (2007) for model data, a Storm Severity Index (SSI) is therefore defined as

$$SSI = \left(\frac{v_{max}}{v_{98}} - 1 \right)^3 \tag{2}$$

if $v_{max} > v_{98}$ and $SSI = 0$ otherwise, with v_{max} the daily maximum wind gust and v_{98} its local 98th climatological percentile. The scaling with v_{98} accounts for the local adaptation to wind gusts, whose impact on infrastructure is weaker in exposed areas such as coasts and montains than in the continental flatlands for the same absolute wind speed (Klawa and Ulbrich, 2003). The climatology of wind gusts is computed separately for ERA-Interim and the reforecast but for the same period of interest, i.e. mid-October–mid-March 1995/96–2014/15. The resulting values of v_{98} are higher in the reforecast, likely due to the higher model resolution ~~. In particular~~but possibly also due to other changes to the ECMWF model. In addition, wind gusts are abnormally high over ~~the~~ topography in the first 6-h output of the reforecast~~, which suggests a problem with~~ . As this does not appear in subsequent outputs, it is likely related to the spin-up of the model when the higher-resolution reforecast is initialized from the lower-resolution reanalysis. The first ~~6 h are~~ 6-h output of the reforecast is thus omitted for computing both v_{max} and v_{98}. ~~Wind gusts are also subject to caution in ERA-Interim but are still preferred to the wind speed (used by ?) , because they represent maximum values over a certain time period rather than instantaneous values and thus better sample storms with a large displacement velocity.~~

~~The~~ As an example, the daily maximum gusts and the resulting SSI in ERA-Interim are shown in Figure 3a and ~~the resulting SSI in Figure 3b~~b, respectively, for storm Lothar on 26 December 1999. The strongest gusts are found over the Bay of Biscay but the highest SSI is found over southern Germany due to the lower values of the local model climatology. The SSI is then averaged over central Europe (defined as 40°N–60°N and 10°W–30°E; corresponds to the map shown in Figure 3) to give a single value for the total severity of the storm, which can then be compared with the reforecast. This method is equivalent to the ~~area~~ SSI defined by ~~?~~ Leckebusch et al. (2007) . It is preferred to including the SSI along the track of the storm only~~(event SSI in ?) ,~~ , as e.g. in (Roberts et al., 2014) , because of the ambiguous identification of the tracks in the reforecast. Among the 25 investigated storms, Lothar exhibits the highest averaged SSI in ERA-Interim, followed by Klaus, Martin and Kyrill (Table 1). These four storms are responsible for the four highest insurance losses during the period of interest (Roberts et al., 2014), which suggests that the averaged SSI in ERA-Interim is a relevant measure of the severity of storms. Inaccuracies are still expected and attributed to mesoscale features that are not ~~resoved~~ resolved by ERA-Interim and by non-meteorological factors such as the density of population and the insured capital. Finally, although the impact of storms is expected from wind gusts over land mostly, the adjacent ocean areas are also included in the calculation of the SSI here to

avoid large sensitivities to the predicted position of storms that track close to the coasts. Including the ocean also accounts at least partially for storm surges, the main impact of some severe storms (e.g. Xynthia, Ludwig et al., 2014) .

2.4 ~~Extreme Forecast Index and Shift of Tails~~

2.3.1 Extreme Forecast Index and Shift of Tails

Forecasting extreme events is a challenge in numerical weather prediction, because predicted extremes tend to underestimate the magnitude of actual events. Lalaurette (2003) therefore introduced the Extreme Forecast Index (EFI), which measures the extremeness of an ensemble forecast as compared to the model climate rather than to the observed climate. The original formulation of the EFI was revised by Zsótér (2006), who included a weighting function to emphasize the tails of the distribution and obtained

$$EFI = \frac{2}{\pi} \int_0^1 \frac{p - F_f(p)}{\sqrt{p(1-p)}} dp \tag{3}$$

with $F_f(p)$ the proportion of ensemble members lying below the p quantile of the model climate. The EFI quantifies the deviation of an ensemble forecast from its climatological distribution with a unitless number between -1 (all members reach record-breaking low values) and +1 (record-breaking high values).

Zsótér (2006) also introduced the Shift of Tails (SOT) as an additional index that focuses even more on the tail of the distribution

$$SOT(p) = -\frac{Q_f(p) - Q_c(p_0)}{Q_c(p) - Q_c(p_0)} \tag{4}$$

with $Q_f(p)$ and $Q_c(p)$ the p quantiles of the ensemble forecast and of the model climate, respectively. The SOT indicates if a fraction of the ensemble members predicts an extreme event, even if the rest of the members do not. Following Zsótér (2006), p is taken as the 90th percentile, i.e. the top two members of the 11-member ensemble reforecast. As in the operational ECMWF configuration, p_0 is taken as the 99th percentile of the model climate, which is smoother than the 100th percentile (maximum) used by Zsótér (2006). A positive value of SOT thus means that at least two members predict an extreme event that belongs to the top percent of the model climate.

Both EFI and SOT are computed here for daily maximum wind gusts. For consistency with the SSI, the model climate is defined from the period mid-October to mid-March 1995/96–2014/15. This contrasts with the operational ECMWF configuration, where the model climate is defined for each forecast within a one-month window centred around the initialization time. As the focus is on winter storms here, a seasonal model climate is preferred to avoid storms to be considered as more or less extreme depending on when they occur during the season. A longer period is also preferred to improve the representation of the 99th percentile of the model climate, as the length of the operational configuration has been validated for precipitation and temperature but not for wind gusts (Zsoter et al., 2015). Finally, as in the operational configuration, the model climate is computed separately at each lead time to compensate for any drift of the reforecast.

Figure 4 illustrates the EFI and SOT for the 6-day reforecast of storm Lothar. High values of EFI spread over a broad region from the Atlantic Ocean to eastern Europe and exhibit stripes further eastward (Figure 4a). Positive values of SOT also spread over a similar, broad region but the highest values are more concentrated (Figure 4b). This is due to the stronger emphasis on the tail of the distribution based on 2 members in SOT rather than on the whole ensemble in EFI. A comparison with ERA-Interim in Figure 3a indicates a skill of both EFI and SOT in predicting the strong gusts over parts of France, Switzerland and Germany. However, it also shows a discrepancy between high EFI or SOT and weaker gusts over other regions. This suggests a potential for warnings but with possible false alarms, as already noted by Lalaurette (2003). The use of EFI and SOT thus requires an appropriate balance between hit ~~rate~~ and false alarm ~~rate~~ rates (Petroliagis and Pinson, 2014; Boisserie et al., 2016).

3 Predictability of storm characteristics

3.1 Position and intensity

The predictability of the selected storms is first evaluated for the position and intensity obtained from the storm tracking algorithm. The storms are identified in the reforecast at the time of first occurence and compared with ERA-Interim at the time of maximum intensity. As the 10-day reforecasts are computed every Monday and Thursday, three lead times are available for most storms but only two for those which occurred on a Sunday. The average bias and spread are computed for each storm and lead time with the median and median absolute deviation, respectively, which are preferred to the mean and standard deviation to ensure robust statistics despite the small number of ensemble members.

On average over all storms, the predicted MSLP remains close to ERA-Interim until day 4, but exhibits a clear positive bias, i.e. it underestimates the intensity of storms from day 5 onwards (black curve in Figure 5a). The predicted MSLP also exhibits a large ~~dispersion~~ variability between the storms, which increases with increasing lead time (symbols in Figure 5a). The most striking outlier is storm Gero (red triangle), which shows the ~~strongest~~ largest positive biases with more than 60 and 40 hPa on days 5 and 8, respectively. Gero experienced an explosive cyclogenesis of 40 hPa in 24 h to reach 948 hPa on 11 January 2005, the deepest MSLP of the sample of storms (Table 1). ~~This suggests an impact of the storm intensity on its predictability, although no systematic link is found in the sample of storms. For instance, the~~ The second and third deepest storms Oratia and ~~Stephen~~Silke, which also experienced an explosive cyclogenesis, show contrasting positive and negative biases in MSLP depending on the lead time (green triangle and blue circle in Figure 5a). ~~The~~ Surprisingly, the predicted MSLP of Gero ~~also~~ exhibits a negative bias on day 1, although this may be due to ERA-Interim underestimating the actual intensity due to its coarse horizontal resolution.

Concerning the position, the predicted longitude exhibits a negative bias on average, i.e. the storms are too slow in the reforecast from day 4 onwards (black curve in Figure 5b). A weak positive bias is present in the reforecast of the latitude but it does not appear to be significant (not shown). Similar to the predicted MSLP, the predicted longitude also exhibits a large ~~dispersion~~ variability between the storms, which increases with increasing lead time (symbols in Figure 5b). Storm Gero is again an outlier with strong negative biases at days 5 and 8 ~~but the strongest biases are shown by ex-Lili~~ (red triangles) but an even larger bias is found at day 7 for Lili (blue square)~~and Dagmar at day 10 (blue cross). These two storms formed remotely~~

from Europe, the former . This storm formed in the tropics (Browning et al., 1998, see also Figure 2)and the latter over the southeastern United States. This suggests a link between the , which is consistent with the poor predictability of the position and the difficulty at representing convective dynamics , especially during extratropical transition during extratropical transition due to the difficulty to represent convective dynamics (e.g. Pantillon et al., 2013). However, storm "u19960207" shows a strong negative bias in longitude at day 7 (green square) though it developed this case is unique among the selected storms. Other cases that exhibit strong biases formed over very different regions, as e.g. Patrick over the southeastern United States (blue cross at day 10) and Jennifer (1996) over the eastern North Atlantic (green square at day 7). This emphasizes that how single factors can influence the predictability of specific stormsbut do not necessarily have a systematic impact.

As expected, the spread between the ensemble members increases regularily with the with increasing lead time on average, both for the intensity (solid black curve in Figure 5c) and the position (solid black curve in Figure 5d). The spread is consistent with the median absolute error (dashed curve), which suggests that the ensemble reforecast is properly calibrated. However, a large dispersion is again found the spread also shows a large variability between the storms and the spread does not it does not necessarily match the error for individual storms. The For instance, the storms with a strong bias mentioned above tend to exhibit a small spread, i. e. their reforecast is overconfident. Inversely, other storms that have a small bias exhibit a large spread, i.e. their reforecast is overdispersive. For instance, . Inversely, the predicted MSLP of Joachim was very uncertain (green crosses at days 7 and 10 on Figure 5c) due to the sensitivity to the phasing of the storm with a Rossby wave train over the western North Atlantic (Lamberson et al., 2016, green crosses at days 7 and 10 on Figure 5c) (Lamberson et al., 2016) . The large uncertainty in the MSLP of Xynthia at day 3 (red plus) may be due to the sensitivity of its intensification to latent heat release during its unusual track over the subtropical North Atlantic (Ludwig et al., 2014) but this is not consistent with does not show for longer lead times. This again emphasizes the difficulty at pointing out a systematic link between physical factors and the predictability of storms.

3.2 Ensemble average and individual members

These results partly agree with findings of Froude et al. (2007a, b) from a systematic evaluation of the track of extratropical cyclones in previous studies using earlier versions of the operational ECMWF ensemble forecast system. The motion of cyclones was also too slow in the forecast but their MSLP was too deep. Beyond the model version, these differences emphasize the dependency on the selection of cyclones. The underestimation of the intensity and speed of storms shown here may thus not be systematic in the ensemble reforecast but rather be related to the selection of deep cyclones that reach Europe. Froude et al. (2007b) further Froude et al. (2007b) also found a slow bias in a systematic evaluation of the track of extratropical cyclones, while Pirret et al. (2017) further found a low bias in intensity for severe European storms. This suggests that the speed is systematically underestimated for extratropical cyclones in general, while the intensity is underestimated for deep cyclones only. Despite these biases, Froude et al. (2007b) found a higher skill of the ensemble mean compared to the control forecast to predict the track and intensity of cyclones . Although not tested here, this from day 3 onwards. This result raises the question of the meaningfulness of limit of validity of the ensemble mean at lead times beyond a few days, when for the track of the storms, as the identification of storms becomes ambiguous. In particular, more ambiguous and the number

of members ~~still containing the storm~~ containing the observed storms decreases when the lead time increases~~, which biases the ensemble mean~~. Both factors may bias the ensemble average towards the tracks that are closer to the analysis and thus overestimate its skill. In the extreme case of ~~ex-Lili for instance,~~ Lili, this metric even becomes meaningless, because all members of the 10-day reforecast valid on the day of maximum intensity have lost track of the storm on the day it reaches Europe~~,~~ ~~making this metric meaningless~~.

~~Using the alternative identification method focusing~~ An alternative measure of predictability is proposed by counting the number of members that forecast the actual storm when it reaches Europe. The closest tracks are identified on the day of maximum intensity ~~ensures that a storm~~ to ensure that a track is identified in each member of the ensemble. ~~The predictability can then be measured by the number of members that match the actual stormwithin certain thresholds in position and intensity. Using moderate~~ Thresholds in position and intensity are combined to define the actual storm, with a 1:2 ratio between the thresholds that roughly corresponds to the ratio between the two median absolute errors (Figure 5d and c, respectively). Using rather generous thresholds of 10° great circle in distance and ~~10~~ 20 hPa in MSLP bias~~(Figure 6a), the~~ , which select about two third of the reforecasts, all 25 storms are captured by almost all 11 members ~~on the first day of lead time only. This number~~ until day 4 (Figure 6a). The proportion of members then decreases and passes below the majority ~~of members beyond day 5. Albeit arbitrary, the thresholds express reasonable criteria for the definition of the actual storm and roughly correspond to the median value of both bias and spread in position and intensity among all storms and lead times.~~ beyond day 8. Using more restrictive thresholds of 5° great circle in distance and ~~5~~ 10 hPa in MSLP bias~~(Figure 6b), the storms are still captured by~~ , which select about one third of the reforecasts, the storms are captured by almost all 11 members ~~at day 1 but~~ until day 2 only and are missed by the majority of members beyond day 3 already ~~. These~~ (Figure 6b). Albeit arbitrary, these combinations of thresholds express a reasonable range of criteria for a useful definition of the actual storm. While the exact number of members forecasting the storm will depend on the precise thresholds, these results suggest that the ~~use of the ensemble mean to predict storms should be restricted to the first~~ storms are forecasted with high certainty until day 2–4~~days of lead time, although the exact limit depends on the thresholds and varies from storm to storm~~. At longer lead times, the certainty decreases but some members still forecast the storms beyond one week in advance, as was already mentioned by Froude et al. (2007b) . The use of ~~single members~~ subsets of the ensemble for early warnings is discussed in ~~the next Section~~ Section 4.

3.3 Storm impact

The predictability of the selected storms is further evaluated ~~for~~ with respect to the impact of the wind gusts estimated from the SSI. Only the daily, spatially averaged SSI is evaluated here, without considering geographical information on where the storm occurred exactly. The reforecast is therefore evaluated for its ability to predict a severe storm on a specific day ~~but anywhere~~ over central Europe. It is compared to ERA-Interim as a logarithmic difference, because the SSI is highly nonlinear (Equation 2) and spans several orders of magnitude between the least and the most severe storms of the selection (Table 1). ~~Finally, although the SSI is scaled locally with separate model climates between the reforecast~~ As illustrated by the 95th and ~~ERA-Interim, the predicted distribution of SSI is overestimated overall. The overestimation is strongest for the low quantiles of the distribution then decreases to~~ 99th percentiles of the model climate, the reforecast systematically overestimates the SSI

of intense and extreme events by a factor of about 2 ~~in the higher quantiles. The predicted SSI is thus divided by a factor of 2 for ease of comparison unless stated otherwise.~~ compared to ERA-Interim (dotted and dashed curves in Figure 7a). This is explained by a longer tail of the distribution of wind gusts in the reforecast compared to ERA-Interim, which impacts on the SSI despite the scaling with separate model climates between the reforecast and ERA-Interim (Equation 2). This systematic overestimation must be taken into account to evaluate the predictability of the selected storms.

On average over all storms, the reforecast ~~is close to ERA-Interim~~ overestimates the SSI until day 3 compared to ERA-Interim, but then drops by one order of magnitude and ~~thus strongly~~ underestimates the SSI at longer lead times (solid curve in Figure 7a). ~~This drop is specific to the sample of severe storms and is not due to a systematic drift in the reforecast, which is illustrated by the 99th percentile of predicted SSI remaining almost constant~~ In contrast, the overestimation of SSI in the whole dataset does not exhibit such a drift with lead time (~~dashed curve).~~ dotted and dashed curves). The overestimation for the storms until day 3 could therefore be corrected, as it results from a systematic bias in the dataset, while the drop on day 4 is specific to the sample of severe storms. The reforecast thus strongly underestimates the severity of the storms beyond day 3. In addition, the average spread in SSI between ensemble members increases until day 3 only, before it decreases again when the average SSI drops (not shown). The reforecast is thus underdispersive at longer lead time. As for the ~~MSLP, however,~~ track and intensity, the predicted SSI shows a large ~~dispersion~~ variability between the storms (symbols). For instance, the deep storms Gero and Oratia are again outliers with strong negative biases at days 5, 8 and 9, respectively, whereas a few other storms even exhibit a positive bias.

These results are confirmed by measuring the number of members that predict at least the SSI of ERA-Interim, which also ~~drop~~ drops at day 4 (Figure 7b). Note that this is a rather ~~pessimistic~~ optimistic estimation, as the predicted SSI is ~~divided by a factor of 2. Before the drop at day 4, the number of members is further separated into two groups with either a large majority or a small minority capturing the storms. This suggests that the reforecast systematically over- or underestimates the severity of individual storms. 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. Beyond day 3, the reforecasts show a systematic underestimation of the SSI for almost all storms.~~ systematically overestimated. However, at least one ensemble member on average still predicts the ERA-Interim value of SSI of the storms until day 7, which suggests a potential for early ~~warning based on individual members~~warnings.

4 Skill for early warnings on the 5–10 day timescale

4.1 ~~Intense~~ Top 5% and ~~extreme~~ 1% SSI events

The results above show that even though the ~~use of the ensemble average is restricted to the first 3 days of lead time, single members are able to predict the storms~~ storms are well predicted by the whole ensemble a few days ahead only, they are forecast by single members up to one week in advance or even beyond~~, as was already mentioned by Froude et al. (2007b)~~. However, these results are biased by the focus on the prediction of observed events ~~(hits)~~ without considering events that are predicted but not observed~~(false alarms)~~. In the following, the skill of the reforecast is investigated not only for the selected

storms but for the whole mid-October–mid-March 1995/96–2014/15 dataset, in order to include days both with and without storms. It is ~~computed~~ measured with the Brier ~~Score (Brier, 1950) , which measures the ability of the reforecast to predict if an event will occur or not.~~

~~The Brier Score can be~~ Skill Score split into reliability ~~, resolution and uncertainty components (Murphy, 1973) . The reliability component measures the ability of the forecast to predict the observed frequency of events. A perfect reliability can be achieved with a climatological forecast and is thus not sufficient to be useful. In contrast, the resolution component measures the ability of the forecast to distinguish between events and non-events, which can not be achieved with a climatological forecast. The uncertainty component finally measures the sampling uncertainty inherent to the events. The Brier Score is further compared to a climatological forecast to obtain the Brier Skill Score (BSS), i. e. the actual skill of the reforecast, which is in turn split as~~

$$BSS = 1 - B_{rel} - B_{res}$$

~~into reliability and resolution components B_{rel} and B_{res} (e.g. Jolliffe and Stephenson, 2012) .~~ and resolution components (Equation 1).

The skill of the reforecast is first investigated for intense events defined as the top 5% of the SSI, which contain the 7–8 most severe storms per ~~year~~ winter on average. Percentiles are preferred to absolute values, because of the systematic overestimation of the reforecast compared to ERA-Interim~~and the reforecast exhibit different distributions of SSI~~. The frequency of intense events is then by definition the same (5%) in the reforecast than in ERA-Interim and thus the reliability component remains close to zero (perfect skill, Figure 8a). The non-zero ~~value reflects~~ values reflect the sampling uncertainty. In contrast, the resolution component increases ~~regularily~~ steadily with lead time to approach ~~1~~ one (no skill). Therefore, the Brier Skill Score follows – with inversed sign – the evolution of the resolution component and decreases ~~regularily~~ steadily until it vanishes (no skill) at day 9. The reforecast thus clearly exhibits positive skill, albeit small, at predicting intense events until day 8.

The skill is less clear for extreme events defined as the top 1% of the SSI. These contain the 30 most severe storms of the whole dataset and approximately match the 25 selected storms in ERA-Interim. Surprisingly, the reforecast does not show any skill at day 1 (Figure 8b). This is linked to a high value of the resolution component (low skill) and may again be due to a problem with the spin-up of the model. The resolution component then ~~regularily~~ steadily increases with increasing lead time as expected. In contrast, the reliability component shows an irregular evolution with lead time and large values reflecting a large sampling uncertainty. This emphasizes that the dataset is too limited to investigate extreme events, which on average represent 8.2 events per lead time only. As a result, the Brier Skill Score suggests that the reforecast exhibits some skill ~~at~~ in predicting extreme events until day 6 but it suffers from the same irregular evolution with lead time.

4.2 EFI and SOT for gusts above the 98th percentile

4.3 ~~Area covered by damaging gusts~~

The potential for early warnings of strong gusts is further investigated with the EFI and SOT, which are both designed for this purpose by highlighting the behaviour of the most extreme ensemble members. As noted by Lalaurette (2003) already, the EFI

gives useful warnings of extreme events but also frequent false alarms. Petroliagis and Pinson (2014) therefore suggested the use of an optimal threshold to balance between ~~hits and false alarms~~hit rate (H) and false alarm rate (F), a higher ~~(lower)~~ ~~threshold increasing (decreasing) both the hits and the false alarms~~or lower threshold increasing or decreasing both H and F. Boisserie et al. (2016) further ~~suggest~~ suggested to maximize the Heidke Skill Score ~~(Heidke, 1926) as a trade-off between hit rate and false alarm rate~~(HSS Heidke, 1926) to define the optimal threshold. Following these authors, an optimal threshold is ~~looked for~~ determined to predict gusts that exceed the local 98th climatological percentile in ERA-Interim. ~~This value is taken for consisteney with the SSI~~The 98th percentile represents the strength at which gusts become damaging in the SSI (Equation 2). In contrast with the previous studies, however, which focused on specific storms or storm ~~intensities~~categories, an optimal threshold is first computed for the whole dataset and only then applied to the selected storms. This ensures that the result is not biased by verifying the forecast with extreme events only.

As shown in Figure 9a, the optimal threshold in EFI decreases with lead time, ~~because both hit rate and false alarm rate decrease with lead time for a given threshold~~as do the corresponding H and F. In contrast, the optimal threshold in SOT is stable until day 6 and decreases at longer lead times only (Figure 9b). This ~~is due to the increase in false alarm rate with lead time for a given threshold in this case, which compensates for the decrease in hit rate (not shown)~~reveals a different balance between H and F for the two indices. A constant threshold is thus only suitable for the SOT and in the early range~~only. The~~ . For all other types of warnings, the dependency of the optimal thresholds on ~~the~~ lead time should ~~else~~ be taken into account~~for warnings~~. The optimal thresholds ~~further show~~ display seasonal and regional variability (not shown), which could also be included to improve warnings. For the sake of simplicity, however, they are not considered here.

Although the optimal threshold exhibits a different evolution with lead time between the ~~EFI and the SOT~~two indices, the corresponding ~~Heidke Skill Score~~ HSS is very similar, with a slightly higher value for the EFI. ~~It decreases regularly~~ The skill decreases steadily with increasing lead time but remains ~~above zero (no skill)~~ positive until day 10, the longest lead time investigated here. The decrease tightly follows ~~the hit rate, while the false alarm rate~~ H, while F slowly increases but remains small due to the rarity of events by definition of the local 98th climatological percentile. Note that ~~the false alarm rate~~F, which is conditioned by the events that are not observed, should not be confused with the false alarm ratio (FAR), which is conditioned by the events that are not forecast. These results demonstrate the actual potential of both EFI and SOT for ~~the early warning~~ early warnings of strong gusts. If the local 99th climatological percentile is preferred to ~~defined~~ define extreme events, as in early studies, the optimal thresholds need to be ~~levelled up~~ increased and the resulting skill becomes lower but it also remains positive until day 10 (not shown).

4.3 ~~Application to~~ EFI and SOT for the ~~selected~~ 25 severe storms

The optimal thresholds described above are applied to the EFI and SOT for the selected severe storms in the reforecast. The ~~Heidke Skill Score~~ HSS is again used as a trade-off between ~~hit rate and false alarm rate~~H and F. It is computed for the prediction of gusts over the central European domain on the day of maximum intensity of each storm. As for the whole dataset, the EFI (Figure 10a) and the SOT (Figure 10b) exhibit ~~similar Heidke Skill Score~~ a similar HSS on average, which lies around 0.8 during the first two days (high skill) and then decreases with increasing lead time until vanishing at day 10 (no skill). In

particular, before day 10, the ~~Heidke Skill Score~~ HSS is higher for the storms (solid curves) than for the whole dataset (dashed curves). It is related to higher ~~hit rates~~ H for the storms, which enhance the skill despite higher ~~false alarm rates~~ F (not shown). This does not necessarily mean that the reforecast is more skillful at predicting the presence than the absence of storms but rather emphasizes how focusing on observed events can bias the verification.

Beyond these average properties, the reforecasts of the storms exhibit contrasting skill from case to case. The ~~dispersion~~ variability between the storms quickly increases with increasing lead time and the ~~Heidke Skill Score~~ HSS of some storms approaches zero or becomes negative from day 6 onwards (symbols on Figure 10). A poor skill is found in both EFI and SOT for storms Lili at day 7 (blue square) and Gero at day 8 (red triangle) in association with a low ~~hit rate~~H, as well as for storm Joachim at day 7 (green cross) in association with a high ~~false alarm rate~~F. This is consistent with the large biases in MSLP and longitude and the large spread in MSLP, respectively, found for these storms (Figure 5). Other storms contrast between poor skill in EFI and good skill in SOT, as Yuma at day 4 (pink square in Figure 10), which was ~~remarked~~ noted for its difficult forecast as it occurred (Young and Grahame, 1999), and Xynthia at day ~~6.~~ 6 (red plus). The higher skill in these cases could be due to the ~~high hit rate~~ higher H of the SOT compared to the EFI, as suggested by Boisserie et al. (2016)~~. However,~~ , although no difference is found here on average in the whole sample.

~~Storm Yumahas the lowest~~ Interestingly, storms Yuma, Lili, Gero and Xynthia mentioned above for their poor skill in EFI have the smallest area of strong gusts of the whole dataset ~~, followed by Lili, Gero and Xynthia~~ (Table 1)~~, which suggest a link between storm size and predictability~~. Similarly, a better performance for Anatol than for the relatively smaller storms Lothar and Martin was previously noted by Buizza and Hollingsworth (2002) in the operational ECMWF ensemble forecast. However, ~~such a link is not systematic, as shown by storm Xaver, which exhibits almost no skill at day 6 in both EFI and SOT though one of the largest area of the dataset.~~ neither this better performance or differences between the predictability of specific storms found by other authors using the EFI (Lalaurette, 2003; Petroliagis and Pinson, 2014; Boisserie et al., 2016) are confirmed here. This suggests a sensitivity to the ensemble prediction system and to the type and region of the reference data used for their validation, which vary from study to study.

Finally, storm Xynthia exhibits a surprisingly high skill at day 10 in both EFI and SOT thanks to a high ~~hit rate~~H. This constitutes an outlier compared to all other storms, which show no skill at that lead time. However, none of the ensemble members predicts the ~~actual~~ observed development of Xynthia over the subtropical North Atlantic (Ludwig et al., 2014). Instead, several members predict a storm forming over the central North Atlantic but reaching the Iberian Peninsula on the same day as Xynthia. Although this successful reforecast could be due to chance rather than to the actual skill of the model, it illustrates how predicting individual storms becomes ambiguous at long range but suggests a potential for predicting an environment favorable to storm development.

5 Conclusions

The synoptic-scale predictability of 25 severe historical winter storms over central Europe is revisited by taking advantage of the ECMWF ensemble retrospective forecast (reforecast), which offers a homogeneous dataset over 20 years with a state-of-

the-art ensemble prediction system. The ~~winter 2015/16 model version is used here and contains 11 ensemble members with a horizontal grid spacing of 30 km up to 10 days lead time that are computed twice weekly for the mid-October–mid-March 1995/96–2014/15 period. The~~ predictability of the storms is investigated with ~~different metrics to include their dynamics, severity and spatial extension. A storm tracking algorithm delivers the position and intensity of the storms~~ three different metrics for their track and intensity (Figure 1a), ~~which are identified in the reforecast either at the time of first occurence or at the time of maximum intensity. The Storm Severity Index (SSI) estimates the actual impact of the storms (~~the strength of wind gusts (Figure 1b) ~~using the strength of wind gusts exceeding the local 98th climatological percentile. The Extreme Forecast Index (EFI) and Shift of Tails (SOT) finally predict~~ and the area covered by strong gusts (Figure 1c)~~by measuring the deviation of the ensemble forecast from the model climate~~. The metrics are combined to assess the reforecast against the ECMWF ~~retrospective analysis (reanalysis)~~ reanalysis ERA-Interim.

~~The ensemble average is unbiased until day 3 to predict~~ For lead times until 3–4 days, the ensemble average has small biases in terms of predicting the position and minimum MSLP of the storms on the day of maximum intensity. At longer lead times, ~~however,~~ it systematically underestimates the speed of motion and the depth of the storms. ~~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~~Previous studies also found a slow bias in the track forecasts of extratropical cyclones in general (Froude et al., 2007b) and a negative bias in the intensity forecasts of severe storms only (Pirret et al., 2017) . This suggests that the underestimation of the speed of motion is systematic but that of the depth is specific to deep cyclones. The ensemble average further underestimates the SSI of the storms ~~at lead times longer than 3 days and the relative error reaches several orders of magnitude. The ensemble spread drops by orders of magnitude, which shows that the SSI of the storms is systematically underestimated. In contrast, there is no general drift in the reforecast, where severe events are present up to 10 days. Similarly, the biases~~ by at least one order of magnitude beyond day 3. Along with the biases with increasing lead time, the identification of storms becomes ambiguous and the number of members containing the observed storms decreases, which questions the limit of validity of the ensemble mean for the track of the storms. This limit is due to the identification of storms as objects, which are not always clearly defined, in contrast to the metrics based on the strength of wind gusts, which are defined even in the absence of a storm. The predictability is further measured by the number of members that forecast the observed storm – within combined thresholds in position and intensity ~~may not be systematic but rather be due to the focus on intense storms that reach Europe~~– on the day it reaches Europe. Although the result depends on the exact thresholds, reasonable values show that the storms are well forecasted until day 2–4 only. These results suggest that relevant predictions of storm properties are restricted to the first ~~2–4 days of lead time. This suggestion is supported by the ambiguousness at identifying the storms at longer lead times in the reforecast~~few days of the forecast.

A different methodology is therefore required ~~at~~ for lead times longer than the 2–4 days horizon. ~~Although they are missed by the ensemble average, the~~ The position, intensity and severity of the storms are captured by some members ~~up to~~ beyond one

week in advance~~or even beyond. As suggested by earlier studies, the~~ , which suggests potential for early warning. The whole distribution of the ensemble ~~should~~ shall thus be used by shifting the focus from the average ~~and spread~~ to individual members for the prediction of extreme events (Buizza and Hollingsworth, 2002; Lalaurette, 2003; Petroliagis and Pinson, 2014; Boisserie et al., 201
The danger with this approach, however, is to verify the predictions with regard to ~~observed events only, i.e. by concentrating on the hit rate~~ their ability to forecast observed events without accounting for ~~the false alarms~~events that are forecast but not not observed. The predictability is therefore investigated here in the whole dataset of 20 winter seasons including both stormy and non-stormy days. ~~Tracking is not used, because it can be applied to cyclones only and becomes ambiguous at long lead times. For~~ Using the EFI and SOT indices, which highlight the most extreme ensemble members, the reforecast shows skill in predicting the area covered by strong gusts until day 9–10. It is also skillful until day 8 to predict the occurence of intense events defined as the top 5% of the SSI ~~, which span~~ (spanning on average 7–8 days per winter~~, the reforecast exhibits a positive Brier Skill Score that regularily decreases until vanishing at day 9. For~~). However, for extreme events defined as the top 1% of the SSI ~~, which approximately correspond~~ (approximately corresponding to the 25 ~~historical storms, the reforecast appears to exhibit a similar skill but suffers from a~~ selected storms), no meaningful results can be obtained due to the large sampling uncertainty~~at longer lead time. The EFI and SOT indices~~ . Despite this limitation for the most extreme events, the results confirm the skill ~~of the reforecast at predicting the area~~ for early warnings of storms beyond one week ahead.

These results are summarized in Figure 11 from long to short forecast lead times separated in three phases. A first phase of early warning starts 8–10 days before a storm occurs. At this point, a few members may already predict the storm, which gives indications of the possibility of a severe event based on the SSI, as well as hints for the area that might be covered by strong ~~wind gusts until day 10 for storms as for the whole dataset. These results highlight the potential for early warnings of storms but also~~ gusts given by the EFI and SOT. In a second phase, the number of members predicting the storm increases but biases are present in the speed of motion and in the intensity measured as the MSLP of the storm, which are both systematically underestimated. The severity of the storm measured by the SSI is also underestimated by one or more orders of magnitude. The certainty then increases until the third phase of accurate forecast, starting 2–4 days before the storm occurs. Most members predict the storm and without systematic bias at this point, which allows a calibrated forecast for the position and intensity of the ~~difficulty at verifying the forecast of extreme events, even with the extended dataset used here.~~ storm and a realistic estimate of its severity. These three phases in the expected skill of an ensemble prediction system may serve as a reference to forecast severe European winter storms in an operational context.

~~While the metrics agree on average, they exhibit a high case-to-case variability. The predictability is particularly low for a few~~ Among the sample of 25 severe storms, some outliers exhibit a particularly low predictability. These are storms involving an explosive cyclogenesis ~~, a tropical origin or~~ or extending over a small area~~. However, no systematic pattern is found among the sample of storms and their predictability partly lacks consistency between lead times. A possible explanation lies in the paucity of data at single lead times and for each storm , as the reforecast is computed twice a week and contains 11 members only. A more frequent initialisation and a larger amount of members may thus prove better ability at identifying~~ , as well as a storm undergoing extratropical transition. Unfortunately, the sample is too small and the number of forecasts per storm is too limited for any robust statistics. The NOAA ensemble reforecast could help to identify systematic links between the dynamics

and predictability of storms~~. The NOAA ensemble reforecast~~ , as it covers a longer period and offers a daily initialization (Hamill et al., 2013)~~but~~ . However, this dataset appears not to perform as well as its ECMWF counterpart for predicting wind over central Europe (Dabernig et al., 2015). ~~Furthermore, even in the~~ The operational ECMWF ensemble forecast ~~initialised every day and containing~~ is initialised twice a day and contains 50 members ~~, ?~~ but Pirret et al. (2017) struggled to find a relation between the predictability and the intensity, track or physical processes of storms~~.~~

~~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. This suggests a fundamental limitation due to the nature~~ , because of the steady increase in skill with more recent model versions. This illustrates the difficulty to systematically investigate the predictability of severe storms, ~~which are extreme events and often do not follow standard patterns.~~ even with an extended dataset such as the 20-year reforecast used here.

More case studies are thus needed to better understand the predictability of specific storm features at different scales. ~~They should be eased by the new generation of global and regional reanalyses that become available with a high horizontal resolution able to better represent the storms. Alternatively~~At larger scale, the focus of the predictability could be shifted from the storms to the ~~large-scale~~ conditions that favour their development (e.g. Pinto et al., 2014)~~, in particular at longer~~ . This could be particularly relevant at long lead times, when the identification of storms ~~is~~ becomes ambiguous among the ensemble members. At smaller scale, the use of available high-resolution model data should help to better understand the structure of the storms. For instance, the next generation of ECMWF reanalysis ERA5, which is currently in production, will reach a horizontal grid spacing of about 30 km and improve the representation of synoptic-scale features. Regional models are required to represent mesoscale features such as sting jets or convection embedded in the cold front, while the accurate representation of wind gusts stays beyond the resolution of operational models and relies on large-eddy simulations or observations at the turbulent scale. Alternatively, dynamical and statistical downscaling can be combined to obtain skillful forecasts at the local level, as demonstrated by Pardowitz et al. (2016) for storm losses, who further took both meteorological and damage model uncertainties into account. These different approaches should be considered to allow advances in the predictability of severe European winter storms.

Author contributions. Florian Pantillon, Peter Knippertz and Ulrich Corsmeier defined the scientific scope of the study. Florian Pantillon performed the data analysis and wrote the paper. All authors discussed the results and commented on the paper.

Acknowledgements. ~~Ensemble~~ ECMWF is acknowledged for providing the ensemble reforecast and ERA-Interim ~~data provided courtesy ECMWF~~reanalysis datasets. The authors thank Philippe Arbogast, Dale Durran, Tim Hewson and Joaquim Pinto for discussions about the interpretation of the results, as well as two anonymous reviewers for comments that helped improving the manuscript. The research leading to these results has been done within the subproject C5 “Forecast uncertainty for peak surface gusts associated with European cold-season cyclones” of the Transregional Collaborative Research Center SFB / TRR 165 “Waves to Weather” funded by the German Research Foundation (DFG).

References

Bechtold, P. and Bidlot, J.-R.: Parametrization of convective gusts, ECMWF Newsl., 119, 2009.

Boisserie, M., Descamps, L., Arbogast, P., Boisserie, M., Descamps, L., and Arbogast, P.: Calibrated Forecasts of Extreme Windstorms Using the Extreme Forecast Index (EFI) and Shift of Tails (SOT), Weather Forecast., 31, 1573–1589, doi:10.1175/WAF-D-15-0027.1, –2016.

Brier, G. W.: Verification of Forecasts Expressed in Terms of Probability, Mon. Weather Rev., 78, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, –1950.

Browning, K. A.: The sting at the end of the tail: Damaging winds associated with extratropical cyclones, Q. J. R. Meteorol. Soc., 130, 375–399, doi:10.1256/qj.02.143, –2004.

Browning, K. A., Panagi, P., and Vaughan, G.: Analysis of an ex-tropical cyclone after its reintensification as a warm-core extratropical cyclone, Q. J. R. Meteorol. Soc., 124, 2329–2356, doi:10.1002/qj.49712455108, –1998.

Buizza, R. and Hollingsworth, A.: Storm prediction over Europe using the ECMWF Ensemble Prediction System, Meteorol. Appl., 9, 289–305, doi:10.1017/S1350482702003031, 2002.

Dabernig, M., Mayr, G. J., and Messner, J. W.: Predicting Wind Power with Reforecasts, Weather Forecast., p. 151008123600008, doi:10.1175/WAF-D-15-0095.1, –2015.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137, 553–597, doi:10.1002/qj.828, –2011.

Della-Marta, P. M., Mathis, H., Frei, C., Liniger, M. A., Kleinn, J., and Appenzeller, C.: The return period of wind storms over Europe, Int. J. Climatol., 29, 437–459, doi:10.1002/joc.1794, –2009.

Donat, M. G., Pardowitz, T., Leckebusch, G. C., Ulbrich, U., and Burghoff, O.: High-resolution refinement of a storm loss model and estimation of return periods of loss-intensive storms over Germany, Nat. Hazards Earth Syst. Sci., 11, 2821–2833, doi:10.5194/nhess-11-2821-2011, –2011.

Feser, F., Barcikowska, M., Krueger, O., Schenk, F., Weisse, R., and Xia, L.: Storminess over the North Atlantic and northwestern Europe-A review, Q. J. R. Meteorol. Soc., 141, 350–382, doi:10.1002/qj.2364, –2015.

Fink, A. H., Brücher, T., Ermert, V., Krüger, A., and Pinto, J. G.: The European storm Kyrill in January 2007: synoptic evolution, meteorological impacts and some considerations with respect to climate change, Nat. Hazards Earth Syst. Sci., 9, 405–423, doi:10.5194/nhess-9-405-2009, –2009.

Froude, L. S. R., Bengtsson, L., and Hodges, K. I.: The Predictability of Extratropical Storm Tracks and the Sensitivity of Their Prediction to the Observing System, Mon. Weather Rev., 135, 315–333, doi:10.1175/MWR3274.1, 2007a.

Froude, L. S. R., Bengtsson, L., and Hodges, K. I.: The Prediction of Extratropical Storm Tracks by the ECMWF and NCEP Ensemble Prediction Systems, Mon. Weather Rev., 135, 2545–2567, doi:10.1175/MWR3422.1, 2007b.

Gatzen, C., Púčik, T., and Ryva, D.: Two cold-season derechoes in Europe, Atmospheric Research, 100, 740–748, doi:10.1016/j.atmosres.2010.11.015, 2011.

Haas, R. and Pinto, J. G.: A combined statistical and dynamical approach for downscaling large-scale footprints of European windstorms, Geophys. Res. Lett., 39, n/a–n/a, doi:10.1029/2012GL054014, –2012.

Hagedorn, R., Hamill, T. M., Whitaker, J. S., Hagedorn, R., Hamill, T. M., and Whitaker, J. S.: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures, Mon. Weather Rev., 136, 2608–2619, doi:10.1175/2007MWR2410.1, ~~,~~2008.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and Palmer, T. N.: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts, Q. J. R. Meteorol. Soc., 138, 1814–1827, doi:10.1002/qj.1895, ~~,~~2012.

Hamill, T. M., Whitaker, J. S., Mullen, S. L., Hamill, T. M., Whitaker, J. S., and Mullen, S. L.: Reforecasts: An Important Dataset for Improving Weather Predictions, Bull. Am. Meteorol. Soc., 87, 33–46, doi:10.1175/BAMS-87-1-33, ~~,~~2006.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., Zhu, Y., and Lapenta, W.: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset, Bull. Am. Meteorol. Soc., 94, 1553–1565, doi:10.1175/BAMS-D-12-00014.1, ~~,~~2013.

Heidke, P.: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst, Geografiska Annaler, 8, 301–349, doi:10.2307/519729, ~~,~~1926.

Hewson, T. D. and Neu, U.: Cyclones, windstorms and the IMILAST project, Tellus, Ser. A Dyn. Meteorol. Oceanogr., 6, 1–33, doi:10.3402/tellusa.v67.27128, 2015.

Hofherr, T. and Kunz, M.: Extreme wind climatology of winter storms in Germany, Climate Research, 41, 105–123, doi:10.3354/cr00844, 2010.

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification : a practitioner's guide in atmospheric science, Wiley-Blackwell, 2012.

Klawa, M. and Ulbrich, U.: A model for the estimation of storm losses and the identification of severe winter storms in Germany, Nat. Hazards Earth Syst. Sci., 3, 725–732, 2003.

Lalaurette, F.: Early detection of abnormal weather conditions using a probabilistic extreme forecast index, Q. J. R. Meteorol. Soc., 129, 3037–3057, doi:10.1256/qj.02.152, ~~,~~2003.

Lamb, H. H. and Frydendahl, K.: Historic storms of the North Sea, British Isles and Northwest Europe, Cambridge, England : Cambridge University Press, 1991.

Lamberson, W. S., Torn, R. D., Bosart, L. F., and Magnusson, L.: Diagnosis of the Source and Evolution of Medium-Range Forecast Errors for Extratropical Cyclone Joachim, Weather Forecast., 31, 1197–1214, doi:10.1175/WAF-D-16-0026.1, ~~,~~2016.

Leckebusch, G. C., ~~Renggli, D., and~~ Ulbrich, ~~U.~~U., Fröhlich, L., and Pinto, J. G.: ~~Development and application of an objective storm severity measure for the Northeast Atlantic region~~Property loss potentials for European midlatitude storms in a changing climate, ~~Meteorol. Zeitschrift, 17, 575–587, , 2008.~~ Geophys. Res. Lett., 34, doi:10.1029/2006GL027663, 2007.

Ludwig, P., Pinto, J. G., Reyers, M., and Gray, S. L.: The role of anomalous SST and surface fluxes over the southeastern North Atlantic in the explosive development of windstorm Xynthia, Q. J. R. Meteorol. Soc., 140, 1729–1741, doi:10.1002/qj.2253, 2014.

Murphy, A. H.: A New Vector Partition of the Probability Score, J. Appl. Meteorol., 12, 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2, ~~,~~1973.

Murray, R. and Simmonds, I.: A numerical scheme for tracking cyclone centres from digital data. Part I: development and operation of the scheme, Aust. Met. Mag., 39, 155–166, http://www.bom.gov.au/amm/docs/1991/murray1.pdf, 1991.

Neu, U., Akperov, M. G., Bellenbaum, N., Benestad, R., Blender, R., Caballero, R., Cocozza, A., Dacre, H. F., Feng, Y., Fraedrich, K., Grieger, J., Gulev, S., Hanley, J., Hewson, T., Inatsu, M., Keay, K., Kew, S. F., Kindem, I., Leckebusch, G. C., Liberato, M. L. R., Lionello, P., Mokhov, I. I., Pinto, J. G., Raible, C. C., Reale, M., Rudeva, I., Schuster, M., Simmonds, I., Sinclair, M., Sprenger, M., Tilinina, N. D.,

Trigo, I. F., Ulbrich, S., Ulbrich, U., Wang, X. L., and Wernli, H.: IMILAST: A Community Effort to Intercompare Extratropical Cyclone Detection and Tracking Algorithms, Bull. Am. Meteorol. Soc., 94, 529–547, doi:10.1175/BAMS-D-11-00154.1, ~~,~~2013.

Panofsky, H. a., Tennekes, H., Lenschow, D. H., and Wyngaard, J. C.: The characteristics of turbulent velocity components in the surface layer under convective conditions, Boundary-Layer Meteorol., 11, 355–361, doi:10.1007/BF02186086, 1977.

Pantillon, F., Chaboureau, J.-P., Lac, C., and Mascart, P.: On the role of a Rossby wave train during the extratropical transition of hurricane *Helene* (2006), Q. J. R. Meteorol. Soc., 139, 370–386, doi:10.1002/qj.1974, ~~,~~2013.

Pardowitz, T., Osinski, R., Kruschke, T., and Ulbrich, U.: An analysis of uncertainties and skill in forecasts of winter storm losses, Nat. Hazards Earth Syst. Sci., 16, 2391–2402, doi:10.5194/nhess-16-2391-2016, 2016.

Petroliagis, T. I. and Pinson, P.: Early warnings of extreme winds using the ECMWF Extreme Forecast Index, Meteorol. Appl., 21, 171–185, doi:10.1002/met.1339, 2014.

Pinto, J. G., Spangehl, T., Ulbrich, U., and Speth, P.: Sensitivities of a cyclone detection and tracking algorithm: individual tracks and climatology, Meteorol. Zeitschrift, 14, 823–838, doi:10.1127/0941-2948/2005/0068, 2005.

Pinto, J. G., Gómara, I., Masato, G., Dacre, H. F., Woollings, T., and Caballero, R.: Large-scale dynamics associated with clustering of extratropical cyclones affecting Western Europe, J. Geophys. Res. Atmos., pp. 704–719, doi:10.1002/2014JD022305.Received, 2014.

Pirret, J. S. R., Knippertz, P., and Trzeciak, T. M.: Drivers for the deepening of severe European windstorms and their impacts on forecast quality, Q. J. R. Meteorol. Soc., 143, 309–320, doi:10.1002/qj.2923, ~~, 2016.~~ 2017.

Raible, C. C., Della-Marta, P. M., Schwierz, C., Wernli, H., Blender, R., Raible, C. C., Della-Marta, P. M., Schwierz, C., Wernli, H., and Blender, R.: Northern Hemisphere Extratropical Cyclones: A Comparison of Detection and Tracking Methods and Different Reanalyses, Mon. Weather Rev., 136, 880–897, doi:10.1175/2007MWR2143.1, 2008.

Roberts, J. F., Champion, A. J., Dawkins, L. C., Hodges, K. I., Shaffrey, L. C., Stephenson, D. B., Stringer, M. A., Thornton, H. E., and Youngman, B. D.: The XWS open access catalogue of extreme European windstorms from 1979 to 2012, Nat. Hazards Earth Syst. Sci., 14, 2487–2501, doi:10.5194/nhess-14-2487-2014, 2014.

Seregina, L. S., Haas, R., Born, K., and Pinto, J. G.: Development of a wind gust model to estimate gust speeds and their return periods, Tellus A, 66, doi:10.3402/tellusa.v66.22905, 2014.

Stucki, P., Brönnimann, S., Martius, O., Welker, C., Imhof, M., von Wattenwyl, N., and Philipp, N.: A catalog of high-impact windstorms in Switzerland since 1859, Nat. Hazards Earth Syst. Sci., 14, 2867–2882, doi:10.5194/nhess-14-2867-2014, 2014.

Stucki, P., Dierer, S., Welker, C., Gómez-Navarro, J. J., Raible, C. C., Martius, O., and Brönnimann, S.: Evaluation of downscaled wind speeds and parameterised gusts for recent and historical windstorms in Switzerland, Tellus A Dyn. Meteorol. Oceanogr., 68, 31 820, doi:10.3402/tellusa.v68.31820, 2016.

Wernli, H., Dirren, S., Liniger, M. A., and Zillig, M.: Dynamical aspects of the life cycle of the winter storm 'Lothar' (~~24?26~~ 24–26 December 1999), Q. J. R. Meteorol. Soc., 128, 405–429, doi:10.1256/003590002321042036, 2002.

Young, M. V. and Grahame, N. S.: Forecasting the Christmas Eve storm 1997, Weather, 54, 382–391, doi:10.1002/j.1477-8696.1999.tb03999.x, ~~,~~1999.

Zsótér, E.: Recent developments in extreme weather forecasting, ECMWF Newsletter, 107, 8–17, http://old.ecmwf.int/publications/newsletters/pdf/107.pdf, 2006.

Zsoter, E., Pappenberger, F., and Richardson, D.: Sensitivity of model climate to sampling configurations and the impact on the Extreme Forecast Index, Meteorol. Appl., 22, 236–247, doi:10.1002/met.1447, ~~,~~2015.

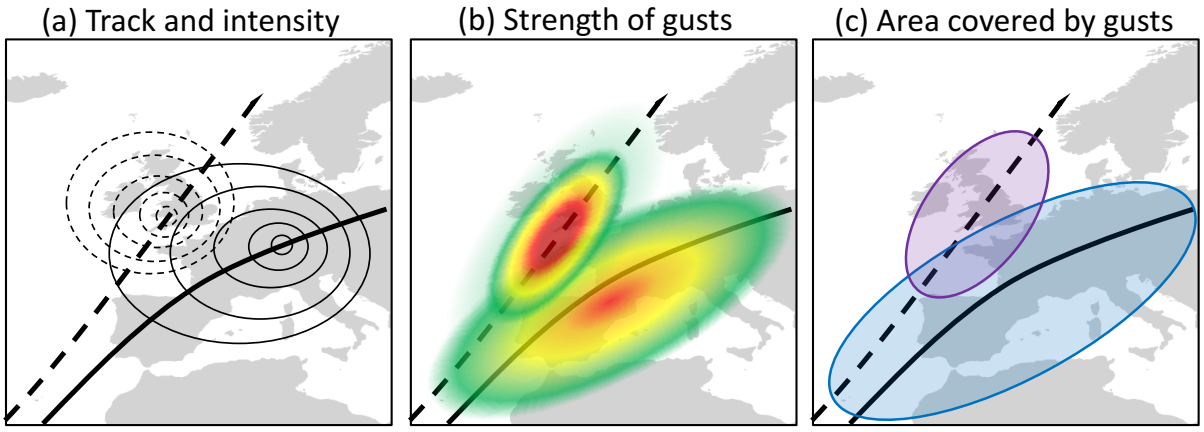

Figure 1. Schematic depiction of the three metrics used to evaluate the predictability of storms: based on the track and intensity of the storms (a), based on the strength of wind gusts (b) and based on the area covered by unusually strong gusts (c). See text for details.

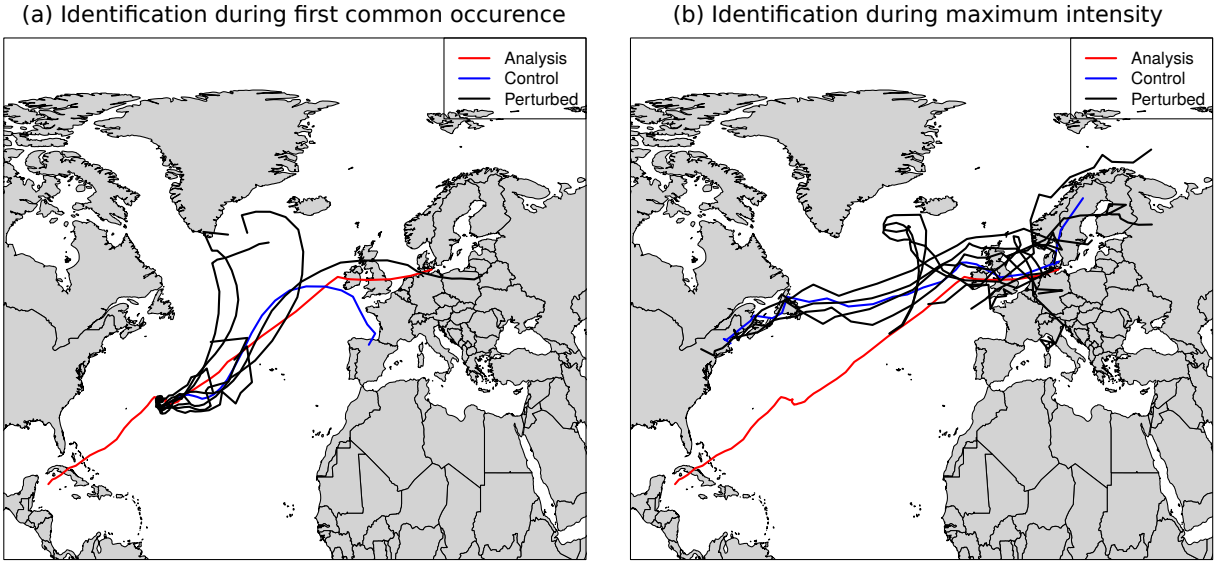

Figure 2. Example of the identified tracks of ex-hurricane Lili in the 6-day ensemble reforecast initialized on 22 October 1996 closest to ERA-Interim during the 24-h period of first common occurrence on 22 October (a) and of maximum intensity in ERA-Interim on 28 October (b).

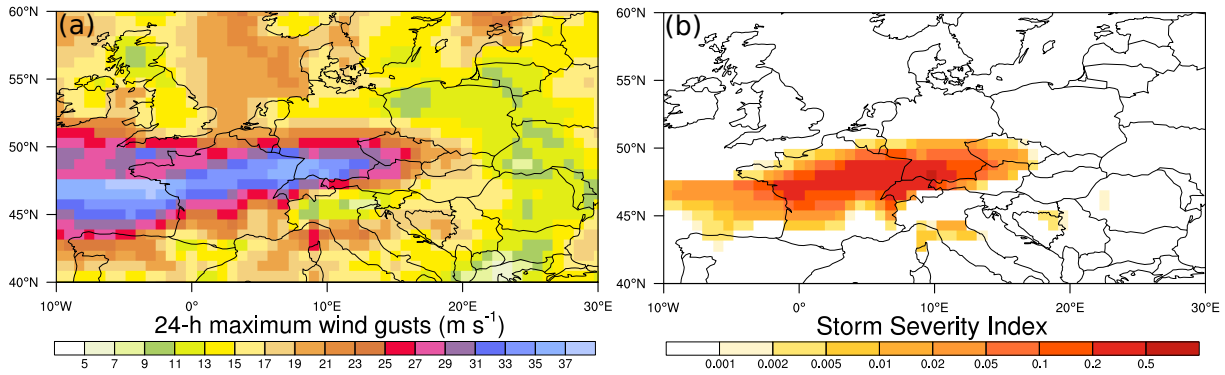

Figure 3. Example of the daily maximum wind gusts (a) and daily Storm Severity Index (b) for storm Lothar in ERA-Interim on 26 December 1999.

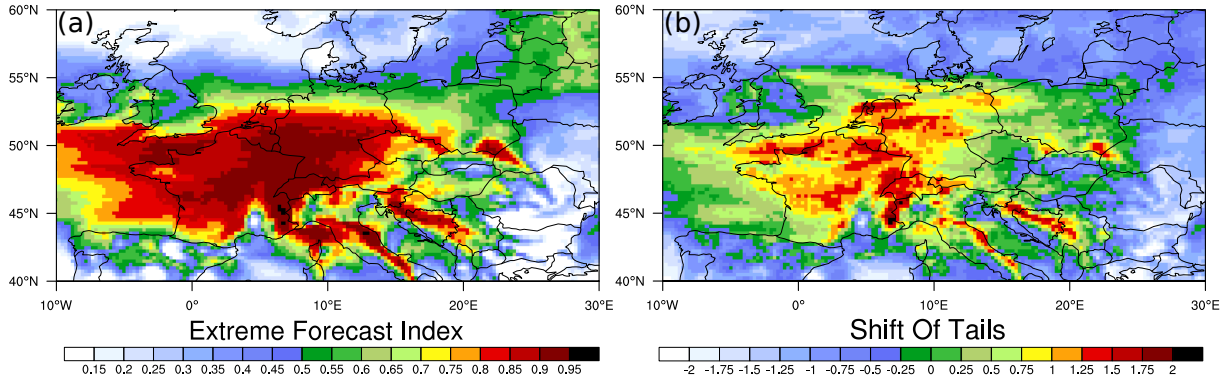

Figure 4. Example of the Extreme Forecast Index (a) and Shift of Tails (b) of daily maximum wind gusts for storm Lothar in the 6-day ensemble reforecast initialized on 21 December 1999 and valid on 26 December.

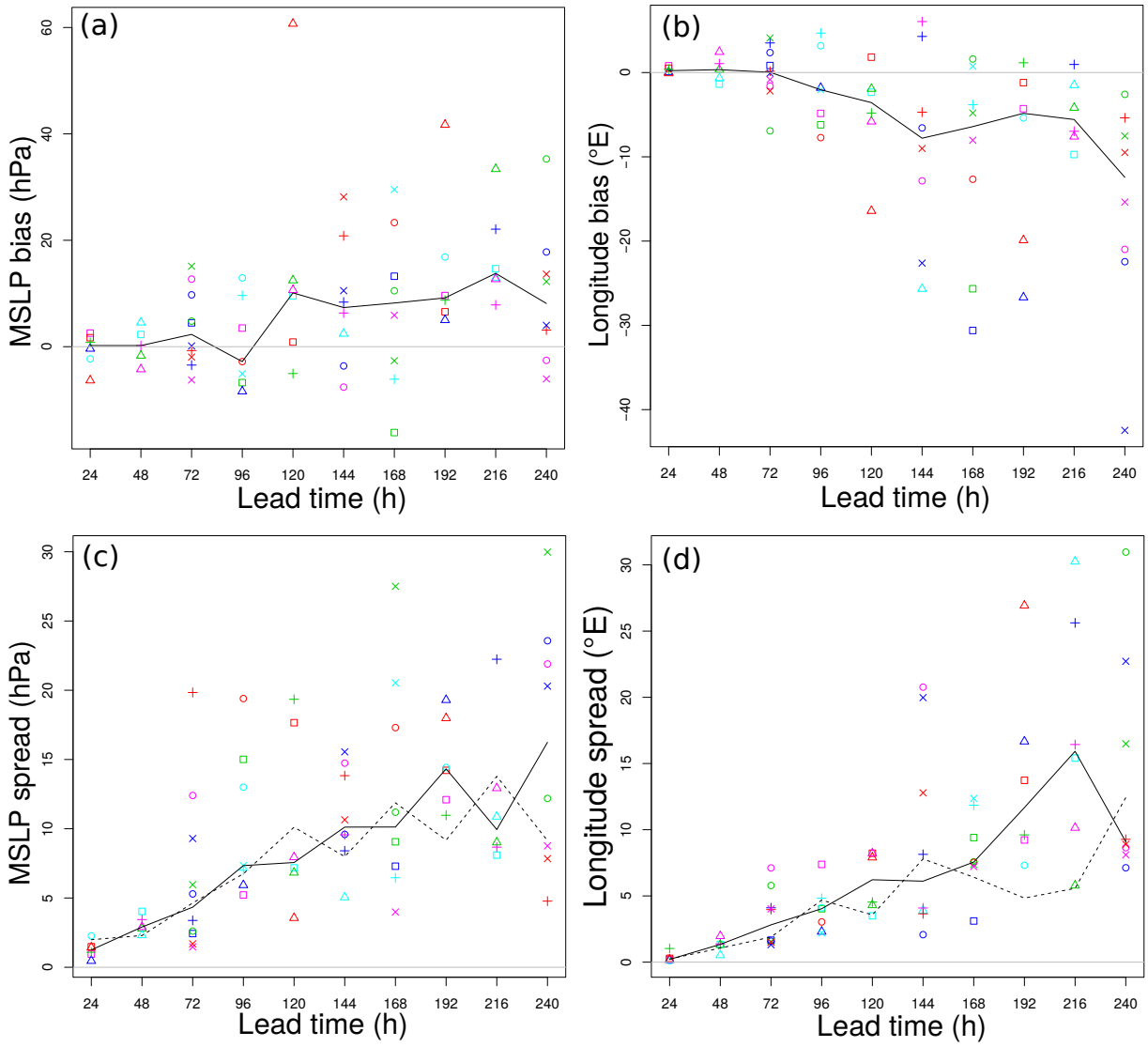

Figure 5. Position and intensity of the storms in the ensemble reforecast as identified at the time of first occurrence and compared on the day of maximum intensity: difference between the ensemble median and ERA-Interim (a, b) and median absolute deviation of the ensemble (c, d) in MSLP (a, c) and longitude (b, d). The symbols represent the storms as given in Table 1 and the solid black curve shows the median of the storms per lead time, while the dashed black curve in (c, d) further shows the median absolute error of the storms per lead time.

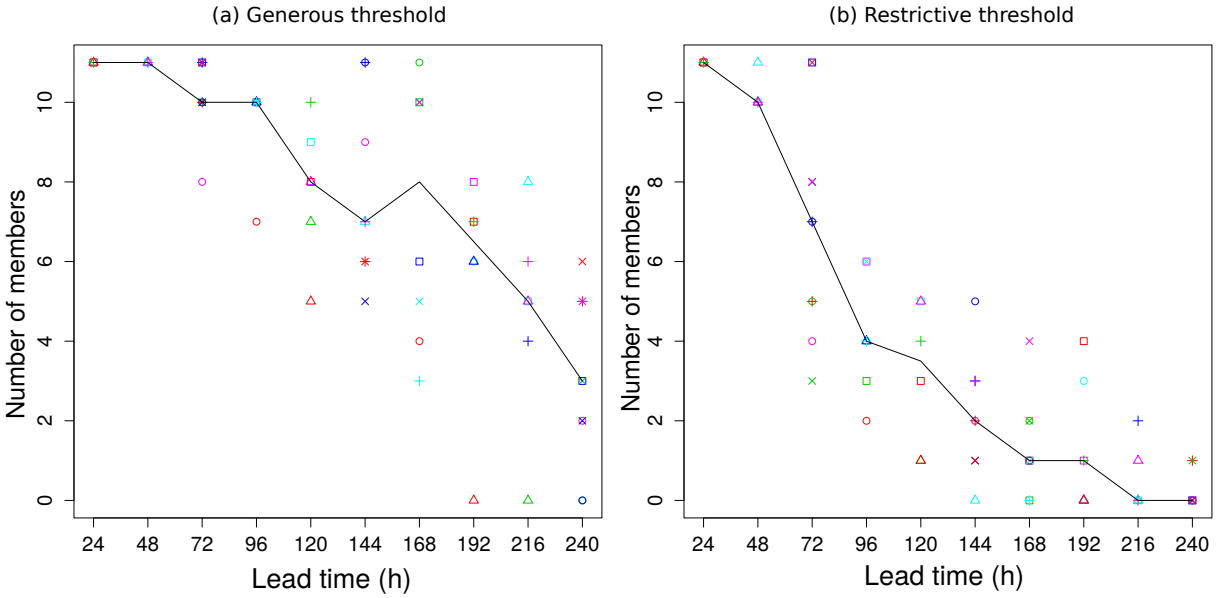

Figure 6. Position and intensity of the storms in the ensemble reforecast as identified and compared on the day of maximum intensity: number of ensemble members predicting the storm within 10 hPa and 10° great circle (a) or 5 hPa and 5° great circle (b) as compared to ERA-Interim in minimum MSLP and position, respectively. The symbols represent the storms as given in Table 1 and the black curve shows the median of the storms per lead time.

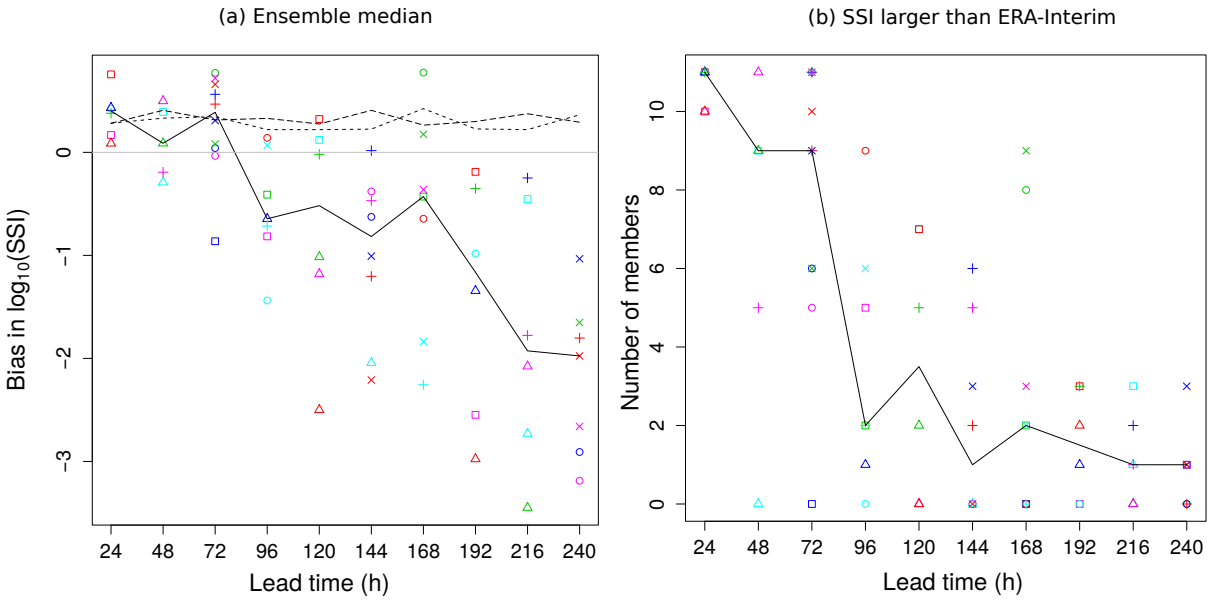

Figure 7. Severity of the storms in the ensemble reforecast on the day of maximum intensity: ~~ensemble median of SSI as~~ logarithmic difference ~~with~~ of SSI between the ensemble median and ERA-Interim (a) and number of members reaching the SSI of ERA-Interim (b). The ~~predicted SSI is divided by a factor of 2 for ease of comparison. The~~ symbols represent the storms as given in Table 1 and the solid black curve shows the median of the storms per lead time~~, while the~~ . The dotted and dashed black ~~curve~~ curves in (a) further ~~shows~~ show the logarithmic difference of the 95th and 99th ~~percentile~~ percentiles of SSI~~compared between~~ , respectively, in the model climates of the reforecast and ERA-Interim.

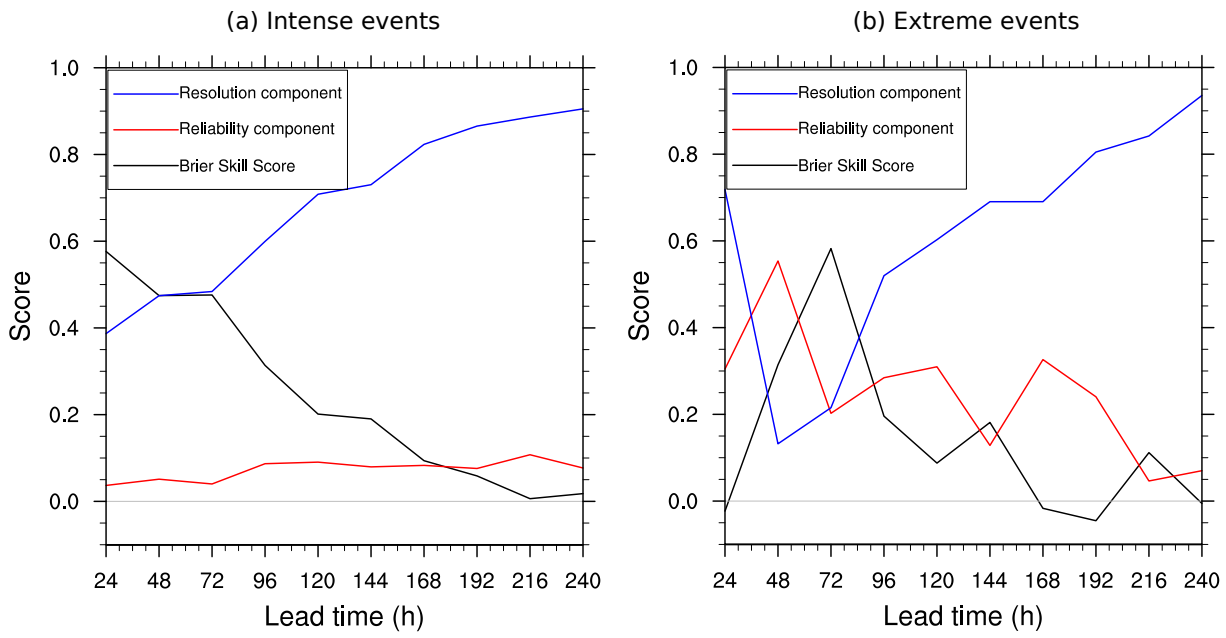

Figure 8. Brier Skill Score as a function of lead time for the SSI exceeding the 95th (a) and 99th percentiles of the model climatology (b). The Brier Skill Score is decomposed into resolution and reliability components (see Equation 1).

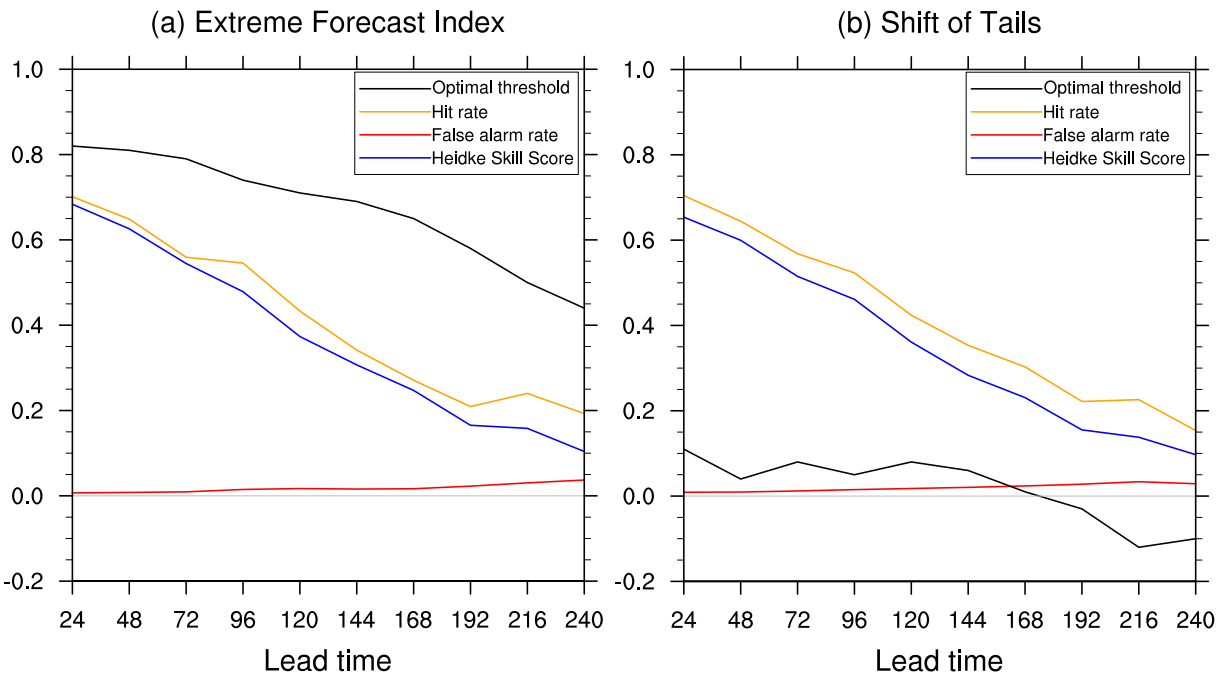

Figure 9. Optimal threshold and corresponding hit rate (H), false alarm rate (F) and Heidke Skill Score (HSS) for the Extreme Forecast Index (EFI; a) and the Shift of Tails (SOT; b) to predict gusts exceeding the local 98th percentile in ERA-Interim.

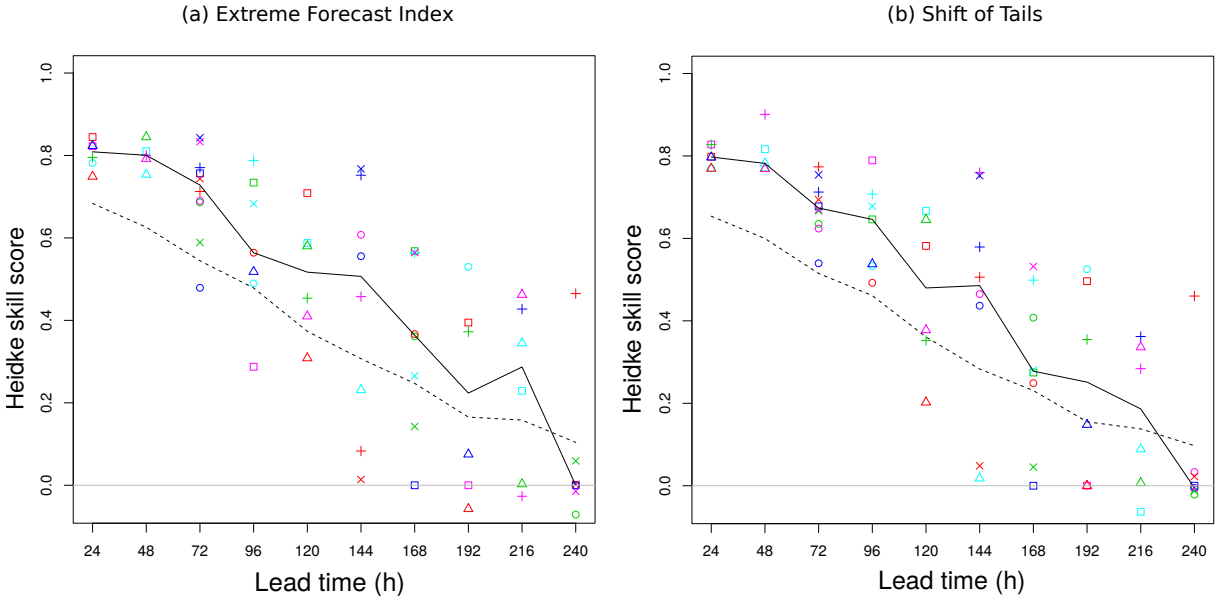

Figure 10. Heidke Skill Score (HSS) for predicting gusts exceeding the local 98th climatological percentile of ERA-Interim using the Extreme Forecast Index (EFI; a) and the Shift Of Tails (SOT; b). The symbols represent the storms as given in Table 1 and the black curve shows the median of the storms per lead time, while the dashed curves illustrate the whole dataset for reference as in Figure 9.

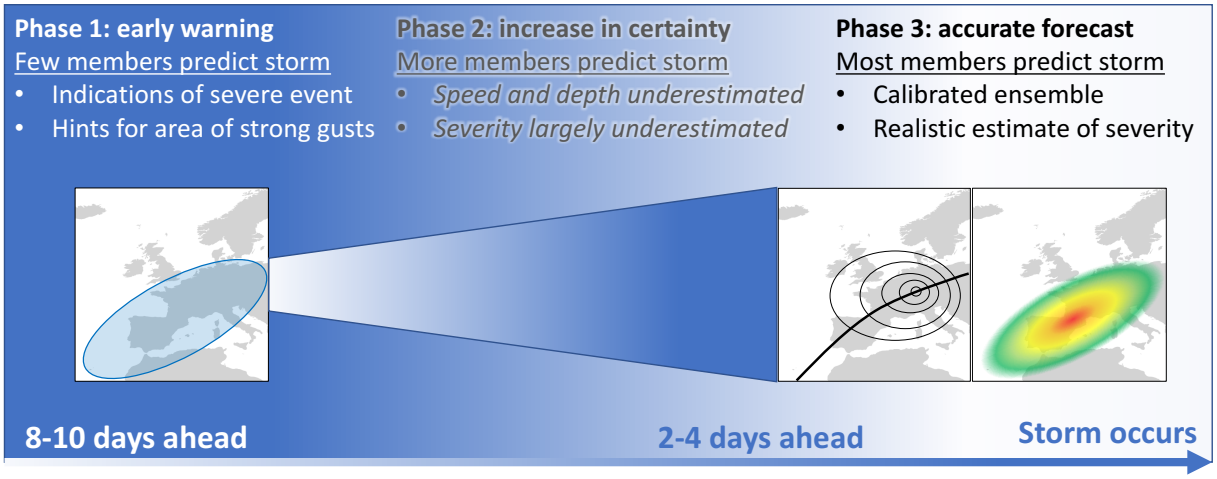

Figure 11. Summary of the expected skill of an ensemble prediction system for the forecast of severe European winter storms, from long to short forecast lead times separated in three phases. The schematic refers to the three methods depicted in Figure 1.

Table 1. Chronological list of the 25 investigated storms with their characteristics in ERA-Interim on the day of maximum intensity: minimum Mean Sea Level Pressure (MSLP), Storm Severity Index (SSI) and area of central Europe covered by gusts exceeding the local 98th percentile. ~~Some particularly high or low~~ The values corresponding to the deepest, most severe and smallest storms cited in the text are emphasized in bold.

Symbol	Name	Date	MSLP (hPa)	SSI ($\times 10^{-3}$)	Area (%)
□	~~u19960207~~ Jennifer (1996)	07 Feb 1996	976	3.0	11.1
□	~~u19961028 (ex-Lili)~~ Lili	28 Oct 1996	970	0.40	**7.3**
□	~~u19961106~~ Romy	06 Nov 1996	960	0.48	20.8
□	Yuma	24 Dec 1997	974	0.35	**5.8**
□	Fanny	04 Jan 1998	966	2.0	16.6
○	Xylia	28 Oct 1998	966	0.64	28.3
○	~~Stephen~~ Silke (Stephen)	26 Dec 1998	**950**	2.4	21.0
○	Anatol	03 Dec 1999	956	5.1	28.7
○	Lothar	26 Dec 1999	976	**15**	23.7
○	Martin	27 Dec 1999	969	**9.7**	20.5
△	Oratia (Tora)	30 Oct 2000	**949**	2.8	24.8
△	Jennifer (2002)	28 Jan 2002	956	1.7	28.1
△	~~Jeanette~~ Jeanett	27 Oct 2002	975	3.8	26.1
△	Erwin (Gudrun)	08 Jan 2005	961	6.4	33.0
△	Gero	11 Jan 2005	**948**	1.9	**7.9**
+	Kyrill	18 Jan 2007	963	**6.7**	35.5
+	Emma	01 Mar 2008	960	2.4	34.7
+	Klaus	24 Jan 2009	966	**13**	12.8
+	Quinten	09 Feb 2009	976	0.59	9.2
+	Xynthia	27 Feb 2010	968	2.7	**8.7**
×	Joachim	16 Dec 2011	966	3.5	31.0
×	~~Dagmar (Patrick~~ Patrick (Dagmar)	26 Dec 2011	965	0.35	10.1
×	Ulli	03 Jan 2012	955	1.6	27.7
×	Christian (St Jude)	28 Oct 2013	969	0.91	18.7
×	Xaver	05 Dec 2013	962	2.3	34.9