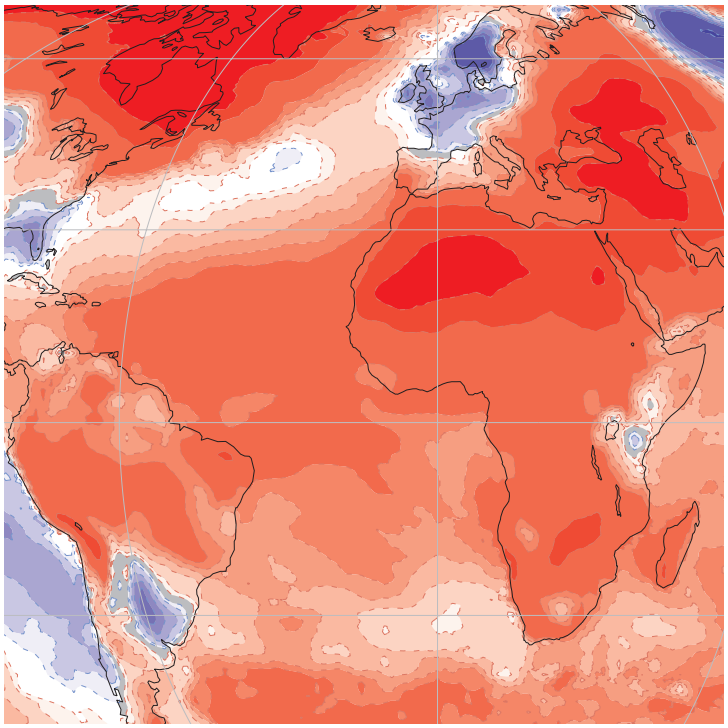
ECMWF Feature article

from Newsletter Number 139 – Spring 2014

METEOROLOGY

Statistical evaluation of ECMWF extreme wind forecasts



www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/ezc4ydat

This article appeared in the Meteorology section of ECMWF Newsletter No. 139 – Spring 2014, pp. 29–33.

Statistical evaluation of ECMWF extreme wind forecasts

Thomas Haiden, Linus Magnusson, David Richardson

An article starting on page 22 of this edition of the *ECMWF Newsletter* analyses forecast skill for two major windstorms which hit Europe in 2013. Case studies such as these are an important part of model evaluation as they allow a detailed diagnosis of model errors associated with specific types of severe weather. However, in order to determine to what extent such findings can be generalized, they need to be complemented by verification over a larger number of cases. The increased generality of results comes at a cost, since one has to include cases that are less extreme in order to obtain robust statistics. Nevertheless, the statistical assessment does provide a framework for the quantification of model deficiencies and the monitoring of forecast improvements.

Here we evaluate the skill of the ECMWF forecasting system in predicting high wind events over a large sample. Events can be defined based on absolute thresholds (e.g. gale-force winds) or the degree of severity compared to climatology (e.g. wind speeds above the 99th percentile). While the absolute value may be more relevant with respect to damage, the percentile-based definition is useful for producing spatially or seasonally aggregated scores, since by definition the number of events becomes comparable between different regions and seasons. An additional reason for choosing a percentile threshold is that the actual impact of an event of given absolute intensity in a certain region will depend on how often it occurs in that location, as this will influence the degree to which the natural environment, buildings and infrastructure are adapted to it. In any case, the choice of specific thresholds involves a compromise. A high threshold is more targeted to rare events but at the cost of a small sample, while a low threshold may provide more reliable statistics but fails to distinguish the skill in forecasting extreme weather from the more general skill of the forecast.

By verifying wind speed forecasts against SYNOP observations, we will show that predictions of severe wind events have benefited from improvements in the forecasting system as much as more 'normal' weather as suggested by improvements in standard skill scores.

Verification method

A basic measure of forecast quality is whether the model is able to simulate the events of interest with the correct frequency. This aspect is evaluated using the frequency bias which is the ratio of the number of forecast and observed events. Here the local conditions (e.g. orography and surface characteristics) at the observation station play a role, as the direct model output is representative of the grid scale rather than a specific location. To evaluate the skill of the forecasts we use the symmetric extremal dependence index (SEDI) which was developed by *Ferro & Stephenson* (2011).

In this investigation we verify both the high-resolution forecast (HRES) and ensemble forecast (ENS) including the ensemble control forecast (CTRL). They are based on the same data assimilation and forecast model but at different resolutions (currently T1279, or 16 km, for the HRES and T639, or 32 km, for the ENS and CTRL). Results from the HRES and CTRL are also compared to those based on forecasts from the ERA-Interim reanalysis.

ERA-Interim uses the forecasting system which became operational in September 2006, but at a different resolution (*Dee et al.*, 2011). The horizontal resolution of ERA-Interim is T255, corresponding to a grid spacing of 80 km. It uses 60 levels in the vertical, compared to 137 for HRES, and 91 for CTRL and ENS. One benefit of a 'frozen' forecasting system such as the one used for ERA-Interim is that it provides a benchmark for operational forecasts and allows the effect of atmospheric variability on the scores to be taken into account.

To calculate a reference model climate, we use the reforecast dataset for the ensemble system which has been operationally produced since 2008. It consists of one unperturbed and four perturbed ensemble members and is run once a week for initial dates in the past 20 years (18 years before 2012). The sensitivity of the resulting model climate to choices in the reforecast configuration, and their effect on the Extreme Forecast Index (EFI), are discussed in *Zsótér et al*. (2014). An important property of the reforecasts is that they are always produced with the latest model cycle.

 doi:10.21957/ezc4ydat

In this study we focus on the verification of wind speed against SYNOP observations in Europe (defined here as 35°–75°N, 12.5°W–42.5°E) where the overall station density is high. For 10-metre wind speed about 1,600 stations were available. A weighting function is used to account for geographical variations in station density (*Rodwell et al.*, 2010). The station climatology is calculated separately for each calendar month based on observations from the 30-year period 1980–2009.

Forecasts can be verified against analyses and observations. A drawback of using analyses in surface verification is that they share some of the systematic errors of the forecasts. On the other hand, conventional observations such as SYNOPs are more or less point measurements and do not represent the same scales as the model. This representativeness mismatch is particularly relevant for severe weather events that are small-scale (e.g. convective precipitation and wind gusts). Another issue is the quality control of observations, which becomes more important as the observations reach more extreme values and sample sizes get smaller. The evaluation presented here uses the simple nearest neighbour method to match forecasts and observations, and employs just a basic quality control. The results should therefore be regarded as a conservative estimate of forecast skill.

Verification scores

The verification of severe weather is commonly based on binary events defined as either exceeding a specific absolute value of a physical quantity or exceeding a percentile of the climate distribution of that quantity. Paired with the observation, the forecasts represent four types of outcome (hits, misses, false alarms and correct negatives) forming a 2x2 contingency table.

	Observed	**Not observed**
Forecast	*a* (hits)	*b* (false alarms)
Not forecast	*c* (misses)	*d* (correct negatives)

Figure 1shows hits, misses and false alarms of the three-day forecast for 10-metre wind speed valid at 12 UTC on 28 October 2013 (storm ‘Christian’, though it also has a variety of other names including ‘St Jude’ and ‘Simone’). The threshold of 16 ms^{-1} approximately corresponds to the 98th percentile of the model climate over the North Sea. This particular forecast underestimated the speed of propagation of the storm system. The timing error leads to false alarms to the west, and misses in the east. Use of the absolute value of 16 ms^{-1}(Figure 1a) leads to a restriction of the event mainly to the sea, while the definition relative to the model climate (Figure 1b) gives signals also over land. Because of this, and because of the need to aggregate over climatologically diverse areas, we use relative thresholds in this study. We specifically focus on the 98th percentile of the climate distribution as a compromise between sample size and rarity of the event.

A common problem of standard scores which are based on a 2x2 contingency table, such as the equitable threat score or the Peirce skill score, is that they degenerate to trivial values (0 or 1) for rare events because the correctly forecast non-events (i.e. correct negatives) dominate the score. Consequently, *Ferro & Stephenson* (2011) introduced the symmetric extremal dependence index (SEDI) to address this problem – see Box A.

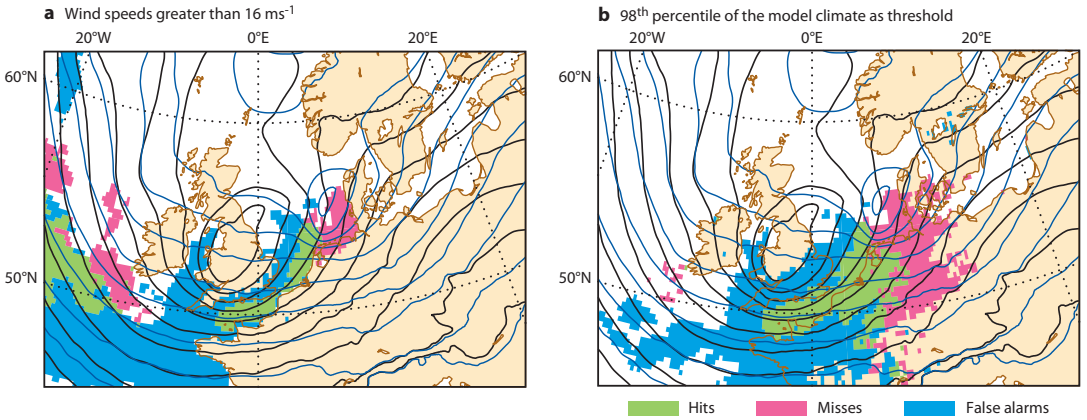


Figure 1 Example of the spatial distribution of hits (green), misses (red) and false alarms (blue) for (a) wind speeds greater than 16 ms^{-1} and (b) use of the 98th percentile of the model climate as a threshold for the three-day forecast valid at 12 UTC on 28 October 2013, verified against the model analysis. Also shown is the mean-sea-level pressure of the forecast (black contours) and the analysis (blue contours).

The SEDI score has a number of desirable properties such as: no explicit dependence on the base rate (climatological frequency of occurrence), robustness to hedging (the score cannot be improved by making unskilful modifications to the forecast), and symmetry with respect to events and non-events. However, as pointed out in *Ferro & Stephenson* (2011), forecasts still need to be calibrated in order to obtain a fair comparison between different forecasting systems. It means that the results indicate potential rather than actual skill.

The calibration is performed for each threshold independently, over a three-month (i.e. seasonal) verification period. Data from all stations in the verification domain is pooled, which is necessary to get a sufficiently large sample; this is made possible by the use of percentile thresholds. The actual calibration is carried out iteratively by varying the percentile threshold applied to the forecast until the frequency bias (see Box A) gets as close as possible to 1, which means that the number of misses and false alarms become (almost) equal.

A contingency-table based score which measures actual skill is the potential economic value *V* (*Richardson*, 2000). This score is based on a simple cost-loss model, where an event is connected to a loss that could be avoided by taking an action which is associated with a cost – see Box A. A zero value of *V* means there is no benefit in using the forecast rather than climatology as a basis for action, while *V*=1 means that one always makes the correct decision (perfect forecast). For ensemble forecasts *V* is calculated for a set of probability thresholds (e.g. action is taken if 10% of the members predict the event), and the maximum *V* for the ensemble is determined for each cost-loss ratio.

Verification scores used in the investigation **A**

Frequency Bias (FB)

Referring to the 2x2 contingency table, the frequency bias of an event is defined as the ratio of the number of forecasts and the number of observations.

$$FB = \frac{a+b}{a+c}$$

Values larger (smaller) than 1 indicate the event is over-forecast (under-forecast).

Symmetric Extremal Dependence Index (SEDI)

$$SEDI = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}$$

where *H* and *F* are the hit and false alarm rates given by:

$$H = \frac{a}{a+c} \quad F = \frac{b}{b+d}$$

Potential Economic Value (V)

$$V(r) = \frac{\min(r,B) - Fr + H(1-r) - B}{\min(r,B) - Br}$$

Where *r* is the cost-loss ratio, *B* is the base rate of the event, and *H* and *F* are the hit and false alarm rates defined above.

Verification results – systematic errors in the forecast climatology

Before we evaluate the predictability of extreme events we investigate systematic errors in the forecast climatological distribution of such events. By climatology we refer to the full probability density function (PDF) for each point (observation station or model grid point) in a given month. The PDF will mainly be evaluated in its cumulative form (CDF), where the phrasing '98th percentile' refers to a value which is not exceeded 98% of the time. Hence, evaluating daily data, values above the 98th percentile will on average occur once in 50 days at each grid point. Figure 2 shows the 98th percentile for 10-metre wind speed of the model climate (shaded) and the observed climatology at individual stations (circles). Plots such as this help to highlight differences between modelled and observed climatologies.

Over the Alps the model gives very low values of the 98th percentile. Observed values show a much large variation in this region than those generated by the model. There are stations with more than 15 ms^{-1} as observed for the 98th percentile, while the model climatology gives values less than 6 ms^{-1}. The stations with high extreme winds are typically mountain stations, whereas nearby stations that have a wind-speed climatology similar to the model are usually located in valleys. Along the coasts the model underestimates the 98th percentile at many stations, for example along the North Sea coast. Here the climatology is sensitive to the land-sea mask in the model. It is another example of a representativeness mismatch between the model and observation scales. Nevertheless, in the evaluation performed here we have included both mountain and coastal stations.

 doi:10.21957/ezc4ydat

While Figure 2 refers to the most recent model configuration, Figure 3 shows the longer-term evolution of frequency bias for the 98th percentile of 10-metre wind speed in the operational forecast. All data is valid for 12 UTC. The figure includes results for HRES, CTRL and ERA-Interim for one-day, four-day and seven-day forecasts. In the absence of model drift the frequency bias should be approximately constant with forecast range and, optimally, it should also be close to 1. The reasons for a frequency bias could be representativeness (model resolution) and/or model errors. As already discussed, large representativeness errors may occur in the presence of steep orography for wind speed, but also surface characteristics (e.g. closeness to sea and surface roughness) around the station play a significant role. ERA-Interim is using the same forecasting system throughout this period; hence its variability with respect to the frequency bias mainly reflects atmospheric variability.

As shown in Figure 3, in terms of the frequency bias for the 98th percentile, HRES, CTRL, and ERA-Interim over-forecast the extreme winds. The frequency bias was similar for all three forecasts around 2007, when HRES and CTRL used the same model physics as ERA-Interim. In June 2011, the roughness length was modified, targeting the positive wind bias; this led to a marked improvement of the frequency bias in HRES and CTRL. For both forecasts the frequency bias is similar for different lead times, indicating no severe model drift with regard to wind speed.

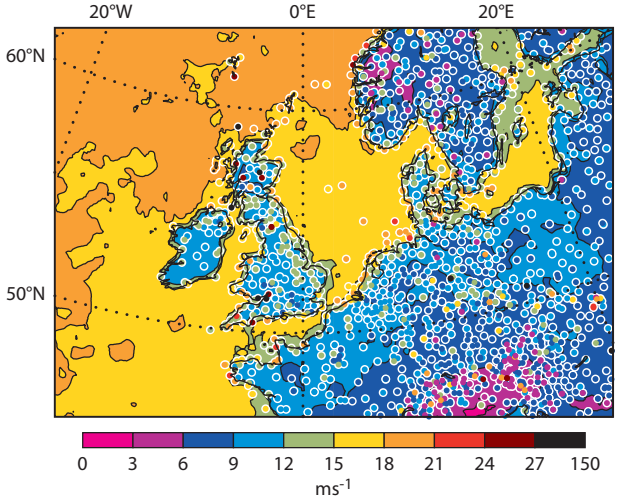


Figure 2 Value of the 98th percentile for 10-metre wind speed in October for the model climate (shaded) and observed climatology (circles).

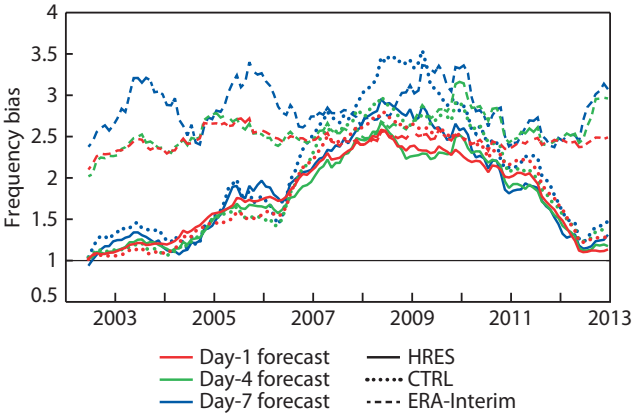


Figure 3 Time series for 2002–2013 (one-year running mean) of frequency bias for the 98th percentile over Europe for HRES, CTRL, and ERA-Interim for day-1, day-4 and day-7 forecasts.

doi:10.21957/ezc4ydat

Verification results – prediction of extreme events

We now consider the ability of the forecasting system to predict extreme events and how the forecast skill has varied with time.

Figure 4a shows the SEDI score for four-day forecasts as a function of the evaluated percentile for HRES, CTRL and ERA-Interim. As described above, SEDI is designed to not explicitly depend on the base rate. Therefore a change in SEDI for higher percentiles reflects an actual change in the ability of the forecasting system in predicting such events. In general SEDI decreases for more extreme events, and it does so more rapidly for percentiles above the 95th. As expected, HRES generally scores higher than CTRL and ERA-Interim but the differences do not seem to increase for more extreme events.

As can be seen in Figure 4b, the skill of the forecasts decreases with increasing lead time. The loss in skill from day 1 to day 4 for the 98th percentile is about the same as the loss in skill from the 80th to the 98th percentile on day 4. Nevertheless, positive skill for the 98th percentile is present even at day 10 in all three forecasts.

The results displayed in Figure 4 indicate that the skill is higher for HRES than for CTRL, showing the benefit of the higher resolution. As expected, the difference between HRES and ERA-Interim is much larger, indicating the importance of both increased resolution and model changes for the prediction of severe wind events.

Figure 5 illustrates to what extent forecast skill has improved over time. It shows time series from 2002 to 2013 of the difference in SEDI between HRES and ERA-Interim for three percentiles (50th, 80th and 98th). These three percentiles represent the change in skill for the median, one-in-five-day events, and one-in-fifty-day events. A positive value indicates that HRES is better than ERA-Interim. In general the scores are better for HRES than ERA-Interim for all years (because of the higher resolution), and the operational forecasts improve over time compared to ERA-Interim due to increasing resolution and model improvements.

Any trends in the difference between HRES and ERA-Interim are superimposed on considerable inter-annual variability which increases with lead time and percentile. A general conclusion from these plots, although the results are noisy, is that over the past ten years SEDI has improved by about the same amount for the 50th, 80th and 98th percentiles. This is an important result since it suggests that (a) forecasts of extremes benefit from the general model improvement and (b) one may not need to specifically verify extremes when evaluating model changes.

Figure 5 also shows that for the 50th and 80th percentiles the difference in skill between HRES and ERA-Interim is slightly higher at day 4 than at days 1 and 7. This can be explained by the constraining effect of the analysis on the forecast at short lead times and the asymptotic approach towards the model climatology at longer lead times.

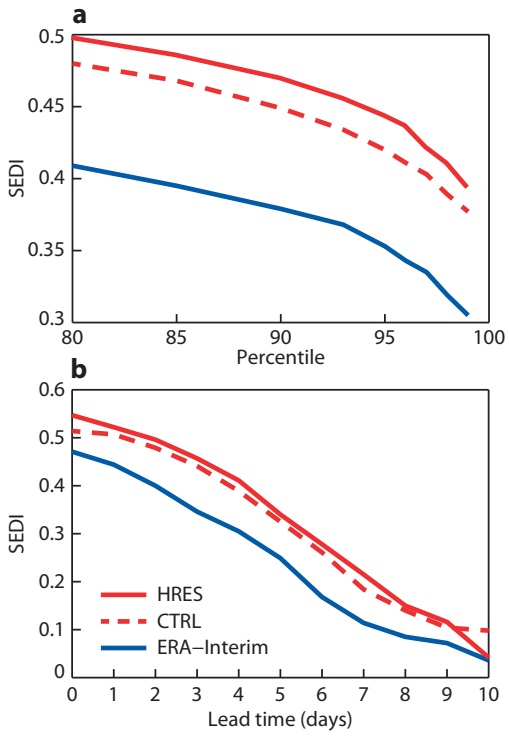


Figure 4 SEDI score over the one-year period July 2011 to June 2012 as (a) a function of percentile for the four-day forecast and (b) as a function of forecast lead time for the 98th percentile.

doi:10.21957/ezc4ydat

A user-oriented measure of severe wind forecast skill is the potential economic value; this is shown in Figure 6 for the 98th percentile. The benefit to the user critically depends on their specific cost-loss ratio. At forecast day 1 the users with cost-loss ratios up to about 0.2 can benefit from the forecast. With increasing lead time this range diminishes. As for SEDI, the skill of HRES exceeds CTRL and ERA-Interim. Due to the additional degree of freedom provided by the choice of probability threshold, the ensemble forecast has considerably higher skill than HRES for users in a certain cost-loss range. This is most apparent in the intermediate forecast range at day 4.

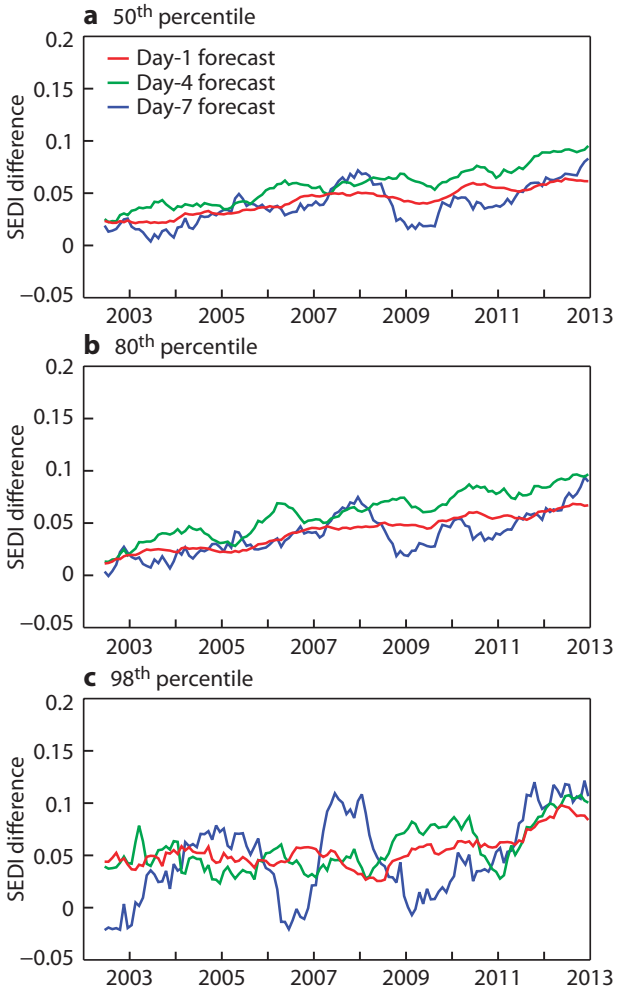


Figure 5 Time series from 2002 to 2013 of the difference in SEDI between HRES and ERA-Interim for 10-metre wind speeds above (a) 50th, (b) 80th and (c) 98th percentiles for day-1, day-4 and day-7 forecasts.

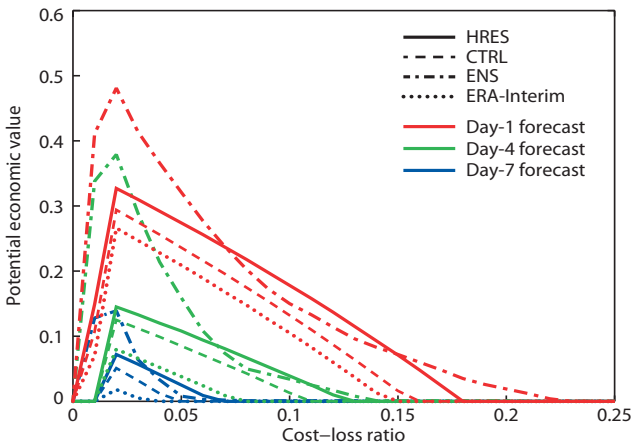


Figure 6 Potential economic value of forecasts of 10-metre wind speed exceeding the 98th percentile for day-1, day-4 and day-7 forecasts for July 2011 to June 2012.

Summary and outlook

We have evaluated the forecast performance for extreme events of wind speed. However, verification of extreme events is not straightforward as sample sizes are small and scores need to be designed to be applicable to rare events. With respect to the threshold for event definition we focus on the 98th percentile of the climate distribution, as a compromise between sample size and rarity of the event. On average such an event occurs once every 50 days and can therefore not be regarded as extreme. High-impact events such as the storm Christian (see the article by Tim Hewson and others in this edition of the *ECMWF Newsletter*) have return periods of several years.

One aspect of forecast performance is whether a model can produce events with a frequency similar to that observed. Such an evaluation is useful to find systematic model issues and to recognize limitations due to resolution in simulating extreme events. By studying maps of frequency biases for the 98th percentile, potential sources for biases of extreme events can be identified, such as orographic and coastal effects.

We have quantified forecast skill using the recently-developed SEDI score. For a fair comparison of different forecasts, they have to be calibrated before calculation of the score. The calibration adds complexity to the verification and removes part of the systematic error such that the result needs to be interpreted as potential skill.

With respect to the long-term evolution of the SEDI score, we found that SEDI for the 98th percentile has improved over the past ten years by about as much as the 50th and 80th percentiles. This indicates that the prediction of extremes has benefitted from improvements in the forecasting system (data assimilation and model) as much as the forecasts of more 'normal' weather.

Apart from the removal of frequency bias required in the computation of SEDI, no calibration has been performed. We expect that forecast calibration will improve forecast skill. Work is being carried out at ECMWF to explore this topic.

We have focused on scores based on hit and false alarm rates. Future work will include more probabilistic verification. One possibility is to use a modified version of the continuous ranked probability score (CRPS), where a function is applied to give more weight to extreme events.

In this study we used SYNOP observations for verification and employed only the most basic quality control by filtering out obviously unphysical values. In order to be able to extend the evaluation to higher percentiles a more sophisticated quality control is required. Finally, we need to acknowledge that for events with return periods of several years, such as the storm Christian, a robust statistic is difficult to achieve even with a very good quality control process. This is why case studies will remain an important tool in the evaluation of forecasts of extremes.

Further reading

Dee, D.P. & **co-authors,** 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.,* **137,** 553–597.

Ferro, C.A.T. & **D.B. Stephenson,** 2011: Extremal Dependence Index: Improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting,* **26,** 699–713.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.,* **126,** 649–667.

Rodwell, M.J., D.S. Richardson, T.D. Hewson & **T. Haiden,** 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc.,* **136,** 1344–1363.

Zsótér, E., F. Pappenberger & **D.S. Richardson**, 2014: Sensitivity of model climate to sampling configurations and the impact on the Extreme Forecast Index. *Meteorol. Appl.,* **21,** (in press).

 doi:10.21957/ezc4ydat