TECHNICAL MEMORANDUM

731

Verification of extreme weather events: Discrete predictands

Linus Magnusson, Thomas Haiden and David Richardson

Forecast Department

September 2014

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Abstract

Forecasting severe weather is one of four main goals in the ECMWF strategy 2011-2020. In this report we evaluate the forecast performance for extreme events of temperature, wind speed and precipitation, all verified against SYNOP observations. We compare the high-resolution forecast (HRES), the ensemble control forecast (CTRL)and the ensemble forecast (ENS). We focus on the frequency bias, the SEDI score and potential economical value.

Comparing the evolution of the SEDI score for three different percentiles since 2002, we found that SEDI for the 98th percentile has improved more over the past 10 years than the 50th and 80th percentile, with the clearest result for 7-day temperature forecasts. This indicates that forecasting extremes have benefited even more from improvements in the forecast system (data assimilation and model) than the forecasting of more average weather. We acknowledge that for many places and parameters the 98th percentile cannot be classified as severe. However, sample size for higher thresholds is not sufficient to give robust results. For really extreme/severe events (with return periods of several years), we believe that one has to look into each specific case.

1 Introduction

Forecasting severe weather is one of four main goals in the ECMWF strategy 2011-2020. Severe weather appears in many shapes such as wind storms, precipitation, thunderstorms, heat waves and cold spells and these meteorological conditions lead to floodings, forest fires, low air quality events etc. To evaluate the usefulness of forecasts for severe weather, suitable verification measures are needed as well as reliable data of the outcome of the event (observations).

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.

In 2010 a report about verification measures was delivered to the Technical Advisory Committee (TAC). The report discussed, among others, verification of severe weather and it was noted that WMO does not provide guidelines for the verification of severe weather events. The report recognised the issue that temporal and spatial resolution of observations may be insufficient for routine verification of severe events. The project also discussed different metrics based on hits, misses and false alarms. A variety of scores was available (see e.g. Stephenson et al. (2008) for a review), but none of the measures available at the time satisfied all of the outlined requirements for verification of rare events. However, since the TAC report, the symmetric extremal dependency index (SEDI) has been developed by Ferro and Stephenson (2011) to address some of the shortcomings of previous scores (hedging and base rate dependence). It has subsequently been used to verify precipitation forecasts from UK Metoffice and ECMWF (North et al., 2013).

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 of the events. Here the local conditions (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.

The current (supplementary) headline score for severe weather is the ROC area calculated for the Extreme Index Forecasts (EFI). For the verification, the event is defined as extreme if the observed values is above the 95th percentile. The evolution of the ROC area since 2004 can be found in Figure 34 in Richardson et al. (2013). One drawback of this verification is that the EFI is an index and is not supposed to directly correspond to a specific percentile. In Ghelli and Primo (2009), precipitation forecasts from ECMWF were evaluated using the extreme dependency score (EDS). In this report we extend the investigation to temperature and wind speeds, using the SEDI score as well as the Potential Economic Value (PEV) (Richardson, 2000), together with an evaluation of the frequency bias for extreme events.

2 Definition of verification scores

Many scores that are considered for severe weather are based on binary events, by simulating a decision made from a forecast. The event could be defined either as exceeding a specific absolute value of a physical quantity, or exceeding a percentile of the climate distribution. The binary decision (yes/no) could come from a deterministic forecast or use a probabilistic forecast (action is taken if the probability is higher than a specified value) as a basis for the decision. Paired with the observed outcome, the forecasts represent four types of verification events (hits, misses, false alarms and correct negatives), forming a 2x2 contingency table (Table 1).

	Obs. Yes	Obs. No
Fc. Yes	a	b
Fc. No	c	d

Table 1: 2x2 Contingency table

From Table 1, the frequency bias is defined as

$$FB = \frac{a+b}{a+c} \tag{1}$$

which is the ratio between the number of forecasted and observed events. A value larger than 1 means the event is overforecast, and a value below 1 means it is underforecast.

2.1 SEDI score

Several scores to evaluate the forecast quality based on Table 1 are available (threat score, equitable threat score, Peirce skill score, Heidke skill score, etc.). A common problem for many of these measures is that they degenerate to trivial values for rare events (either converging to 0 or 1), when the correct negative term dominates the outcome. For a review of the properties of different scores, see Stephenson et al. (2008). In Stephenson et al. (2008) the extreme dependency score or EDS was introduced to overcome the convergence problem. However, it was found that EDS encourages hedging (yielding a better score for a system that over-forecasts an event) (Ghelli and Primo, 2009; Primo and Ghelli, 2009).

In Ferro and Stephenson (2011) the symmetric extremal dependence index (SEDI) score was introduced. The score is defined as

$$SEDI = \frac{logF - logH - log(1-F) + log(1-H)}{logF + logH + log(1-F) + log(1-H)} \tag{2}$$

where H is the hit rate defined as

$$H = \frac{a}{a+c} \tag{3}$$

and F is the false alarm rate

$$F = \frac{b}{b+d} \tag{4}$$

The SEDI score fulfils most of the desired properties. However, as pointed out in Ferro and Stephenson (2011), the forecasts have to be calibrated to a frequency bias equal to 1 before verification to produce a score suitable for comparing different forecasting systems. This means that the results need to be interpreted as potential rather than actual skill. We complement this metric with the potential economic value, where such calibration is not required.

The calibration of SEDI is performed for each threshold independently, over 3-month (i.e. seasonal) verification periods. Data from all stations in the verification domain (Europe) is pooled, which is necessary to get a large enough sample, and made possible by the use of percentile thresholds. The actual calibration is carried out iteratively by varying the percentile threshold applied to the forecast such that the misses and false alarms (the off-diagonal elements of the contingency table) become equal.

2.2 Potential economic value

For ensemble forecasts one can construct 2x2 contingency tables for a number of discrete probability thresholds. In decision making, the probability threshold should be dependent on the cost-loss ratio of the application. If the probability for the forecast exceeds this threshold, the forecast is considered as a yes. One way to summarise the results from several probability thresholds is the potential economical value (Richardson, 2000). The score is building on a simple cost/loss model, where the event is connected to a loss(L) that could be avoided bytaking an action which is associated with a cost(C).

$$PEV(\frac{C}{L}) = \frac{min(\frac{C}{L},\bar{o}) - F\frac{C}{L}(1-\bar{o}) + H\bar{o}(1-\frac{C}{L}) - \bar{o}}{min(\frac{C}{L},\bar{o}) - \bar{o}\frac{C}{L}} \tag{5}$$

The PEV is a function of the cost/loss ratio ($\frac{C}{L}$) and is determined by the climatological occurrence of the event ($\bar{o}$), the false alarm rate F (action was taken but proved unnecessary) and hit rate H (the loss was prevented by the action). A zero value of PEV means that there is no benefit (relative to climatology) in using the forecast as a basis for action, while a PEV equal to 1 means that one always takes the correct decision about the action (perfect forecast).

The PEV is a function of the cost/loss ratio of an application. If the cost of preventing a loss is close to the loss suffered in case of an event, it is rarely an advantage to use the information. In contrast, if the

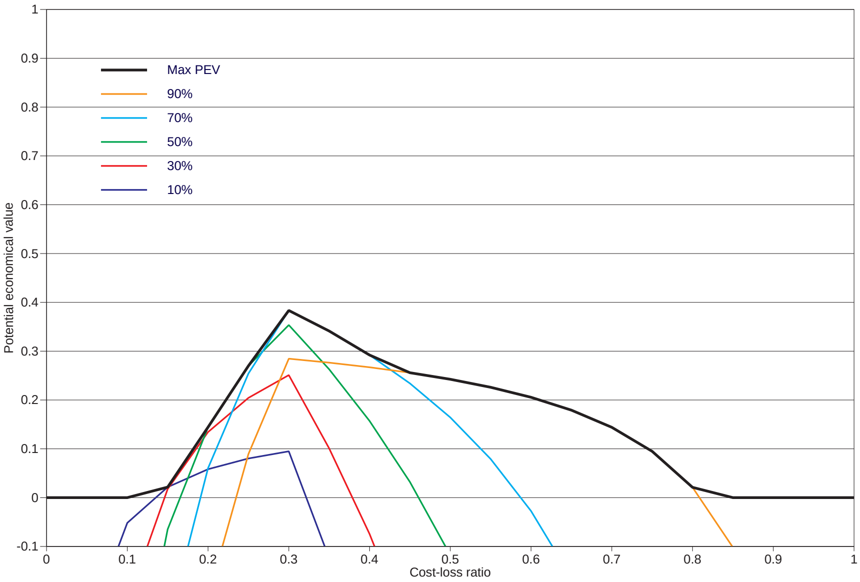


Figure 1: Potential economical value (PEV) for synthetic data for different probability thresholds (coloured lines) and the maximum PEV (black line).

cost is small compared to the loss, the preventing action can be taken almost regardless of the forecast. However, in between these extremes the forecast information can play a crucial role in optimising the decision process. In many cases the cost/loss ratio will be a function of the forecast lead time, since the cost of an early action may be different (typically less) than that of a late action. Early warnings could therefore have a lower cost/loss ratio than forecasts issued closer to the event. By using an ensemble system, which is able to produce different probabilities for an event, one can use different probability levels for applications with different cost/loss ratio. However, the difficulty with PEV is for the user to know the cost and the loss of their application. Another limitation is that the cost or loss may be dependent on the decision history (e.g. too frequent cancellation of trains leading to a change in travel habits). In spite of these limitations of the cost-loss model we nevertheless consider the PEV very useful for quantifying the benefit of the forecasts for decision making.

For ensemble forecasts, the PEV is calculated for a set of probability thresholds (e.g. action is taken if 10%, 20%, .. 90% of the members have the event). The maximum PEV for the ensemble is determined from the probability threshold with the highest PEV for each specific cost/loss ratio. This is illustrated in Figure 1 for synthetic data and an observed occurance ($\overline{o}$) of 30%. For single, deterministic forecasts the maximum PEV is equal to the PEV curve determined by the 2x2 contingency table for yes/no of the binary event.

The procedure of maximising PEV from different probability thresholds also leads to an auto-calibration of the frequency bias for the forecasts. If the model under-forecasts an event, the optimum probability threshold is higher than for an unbiased system. In the case of a deterministic forecast, a frequency bias will lead to shift in the PEV (towards lower cost-loss ratio for over-forecasting, and vice versa).

3 Model description

In this report we verify the high-resolution foreacast (HRES) and ensemble forecast (ENS). Both are based on the same data assimilation but the forecast models are using different resolutions.

Year	Change
Nov 2000	T_L511/L60 (HRES), T_L255/L40 (ENS)
Jan 2003	Major data assimilation, cloud physics and convection changes
Apr 2005	New moist boundary layer scheme
Feb 2006	T_L799/L91 (HRES),T_L399/L62 (ENS)
Sept 2006	Revised cloud and surface drag scheme
Nov 2007	Major physics changes (see text)
Feb 2010	T_L1279/L91 (HRES),T_L639/L62 (ENS)
June 2011	Ensemble of Data Assimilations (DA variances and ENS pert.)
Nov 2011	New cloud physics and roughness length

Table 2: Major upgrades affecting the forecasting system since the year 2000.

Table 2 shows a list of major upgrades (not exhaustive) in the forecasting system at ECMWF since 2000. A comprehensive description of the changes between 2005 and 2008 is given in Jung et al. (2010). The changes in November 2007 (cy32r3) included a new formulation of convective entrainment, reduced vertical diffusion, and modifications to the gravity-wave drag scheme and is further documented in Bechtold et al. (2008).

Results from the operational forecasts will be compared to results from ERA Interim (ERA-I). The ERA-I reanalysis uses the forecasting system (both assimilation and model) that became operational in September 2006, but with different resolution (Dee et al., 2011). The horizontal resolution is T_L255, and it has 60 vertical levels. One benefit of a ’frozen’ forecasting system like ERA-I 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 Zsoter et al. (2014). An important property of the reforecasts is that they are always produced with the latest model cycle.

4 Verification data

For all forecast verification, a proxy for the true outcome is essential. It could be an analysis originating from the data assimilation system, or observations. While the analysis is a gridded data set covering the whole globe and easy to use, the major drawback is that the errors in the analysis are correlated with errors in short-range forecasts and the analysis system shares some of the systematic error characteristics of the forecasts (provided analysis and forecast are from the same model system). This is in particular a problem for weather parameters near the surface (2-metre temperature, precipitation, 10-metre wind speed), which are model generated quantities, as the forecast model and data assimilation model use the same parameterisations.

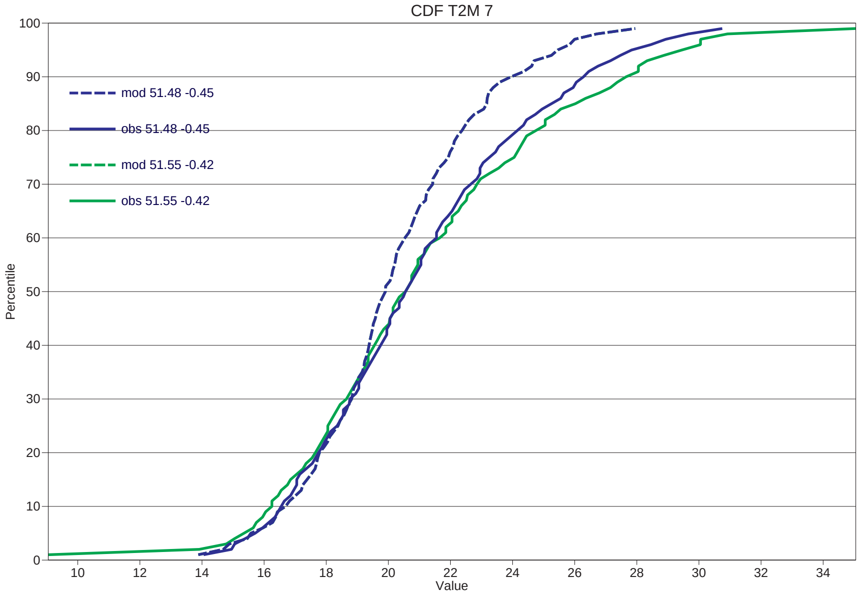


Figure 2: Cumulative distribution function of 2-metre temperature in July for the model (dashed) and observed (solid) for Heathrow Airport (blue) and RAF Northort (green), both in western London, UK.

Ultimately the forecasts should be evaluated against observations. However, conventional observations (SYNOP, radiosondes, aircraft) are point measurements and do not represent the same scales as the model, which generates output as an average over the grid box. This is evident for severe weather events that are small-scale (e.g. convective precipitation) and leads to representativeness errors when comparing model data with observations. Another issue is to find a good match between the observed and modelled quantities, e.g. for wind gusts. A third issue is the quality control of observations, which becomes especially relevant for severe weather where the observations take extreme values. In data assimilation advanced methods are used for observation quality control based on the difference between short forecasts and observations (see e.g. Andersson and Järvinen (1998)). However, for verification purposes, especially for surface variables, the risk is to discard correct extreme observations in cases of bad forecasts or at stations where the representativeness mismatch is especially strong.

In order to illustrate the problem of observation quality control and representativeness, Figure 2 shows the cumulative distribution function (CDF) for 2-metre temperature at 12 UTC in July. The two stations plotted here are Heathrow airport (blue) and the Royal Air Force base Northort (green), both located in western London. In the figure also the model climates for each station derived from reforecasts are also plotted. The model climate here is calculated from the reforecast data set, where we have used 4 ensemble members and 4 dates in July every year in a 18-year period (in total 288 data points). The model climate is valid for a 132-hour forecast and at 12UTC on the day. The climatology of the observations is calculated for July 1980 to 2009 (about 900 data points in total), and the same time of the day.

In this case the model climatologies are identical as both stations have the same nearest gridpoint. For both stations the model underestimates the temperature during warm days; the 90th percentile for the model is 23.5°C, while for observations in Heathrow it is 26°C. This could either be due to a model bias, or limited representativeness of the observations; Heathrow has large airfields which probably warm up more during sunny days than the grid box average. For Northort the tails (both on the warm and the cold sides) are much longer than for Heathrow. This is unexpected as the stations are close to each other and

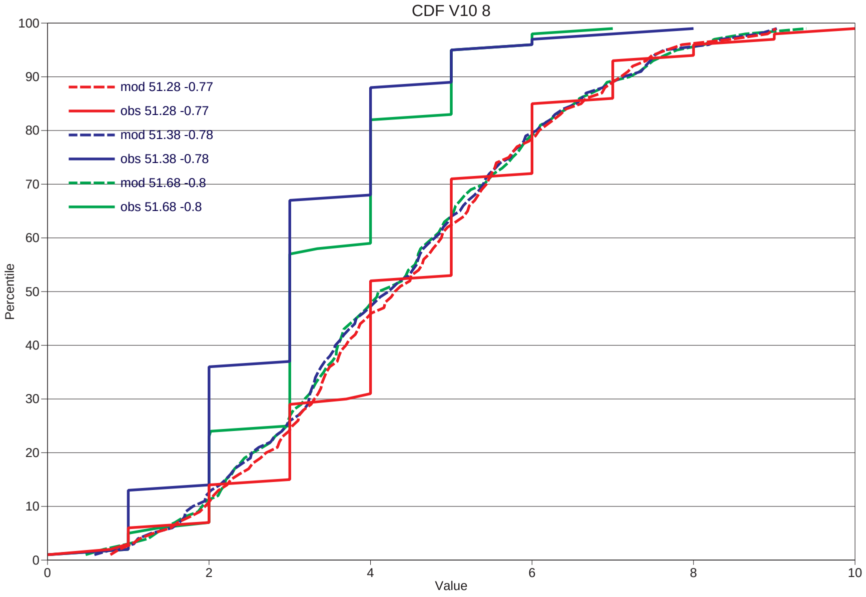


Figure 3: Cumulative distribution function of 10-metre wind speed in August for the model (dashed) and observed (solid) for three stations close to Reading, UK.

both are airfields. Therefore one can suspect some errors in the observations contributing to the observed climate for Northort. However, such a quality control is not trivial as the values are not unrealistic. The SYNOP data available for this study has been subjected to a very simple quality control only, by excluding values which are clearly unphysical.

As a second example, Figure 3 shows the CDF for August from observations of the 10-metre wind speed for 3 stations around Reading, UK, and the model climate for the nearest model gridpoint to the stations. The first striking feature in the plot is the step-like function of the observed CDF. This is because wind speeds are reported as integer numbers in SYNOP messages. This puts a limit on how accurately the forecasts can be verified, but also leads to inaccuracies in the CDF. One solution (for the calculation of the CDF) is to add a random number to each observation in order to smooth the distribution, but it has not been used in this study.

The observed CDF differs among the three stations, while the model CDFs are much more similar. This is mainly caused by local conditions around the observations stations. In this example, the station plotted in red is Farnborough, which is located next to an airfield and is probably a less sheltered station than the other two (Bracknell - blue, High Wycombe - green). The model CDF seems to agree better with Farnborough than with the other ones. This likely implies an over-estimation of the grid box average it is supposed to represent.

In this report we focus on the verification of precipitation, wind speed and temperature against SYNOP observations. While wind speed and temperature are instantaneous values the precipitation is accumulated over 24-hours. We define the lead time as the end of the 24-hour window (a 5-day forecast is accumulated between day 4 and day 5). We will only discuss verification results from Europe (defined as *35N-75N, 12.5W-42.5E*), where station density is high, but the same methodology can be applied to other parts of the world with a sufficiently dense observation network. For 2-metre temperature and 10-metre wind speed about 1600 stations were available, for 24-h precipitation 1100 stations. A weighting

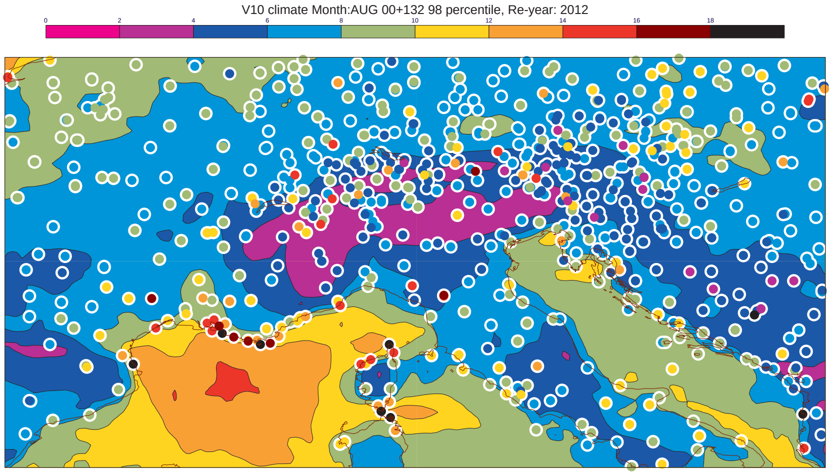


Figure 4: Value of the 98th percentile for 10-metre wind speed in August for southern Europe. Model climate (shade) and observed climatology (dots).

function is used to account for geographical variations in station density (Rodwell et al., 2010). The verification methodology follows the one used in Haiden et al. (2012). The station climatology is calculated separately for each calendar month.

5 Climatology of extreme events

Before we evaluate the predictability of extreme events we investigate systematic errors in the forecasted 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). To sample the tails of the PDF, a large number of observations (or forecasts) are necessary. To consider the seasonality in the PDF, the climatology is estimated for each calendar month. The PDF will mainly be evaluated in its cumulative form (CDF), where the cumulative frequency of the event below a certain threshold is considered. The phrasing ”98th percentile” therefore referes to the value that is not exceeded 98 % of the time. Hence, evaluating daily data, values above the 98th percentile will on average appear every 50th day in each grid point.

Figures 4 to 6 show the 98th (2nd in the case of cold temperatures) percentile from the model climate (shaded) and observed climatology (dots). These plots highlight areas of differences between the modelled and observed climatology. Obviously there are other regions in Europe which are of interest and other times of the year that need further investigation. These plots merely serve as examples.

Figure 4 shows the 98th percentile for 10-metre wind speed in August. Over the Alps the model gives very low values of the 98th percentile. Observed values show a much larger variations in this region than for the model. There are stations with more than 14 m/s as observed for the 98th percentile, while the model climatology gives values less than 4 m/s. The stations with high extreme winds are typically

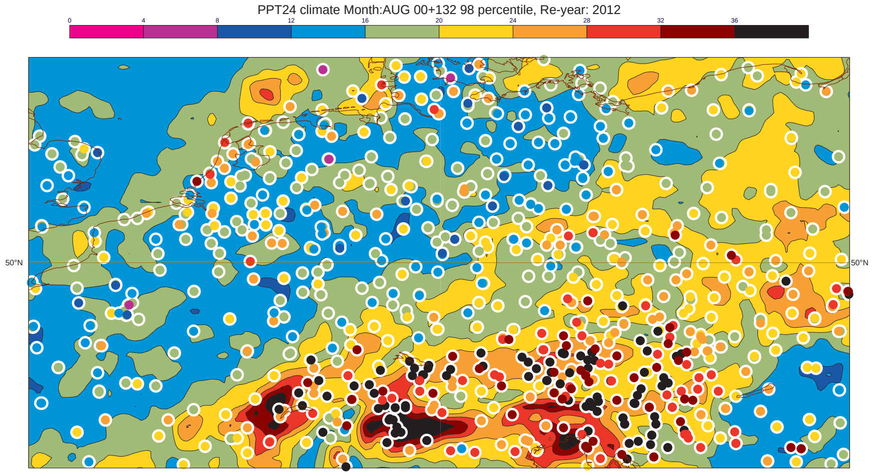


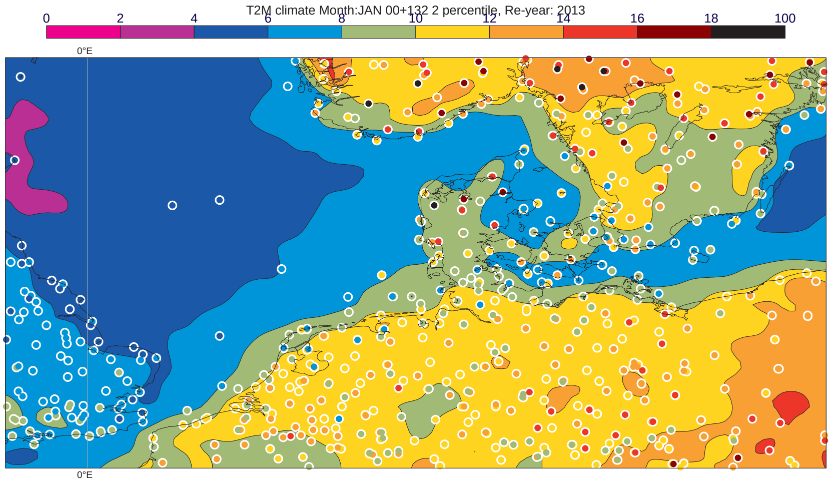
Figure 5: Value of the 98th percentile for 24-hour precipitation in August for central Europe. Model climate (shade) and observed climatology (dots).

mountain stations, whereas nearby stations with 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 French Mediterranean 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 included both mountain and coastal stations.

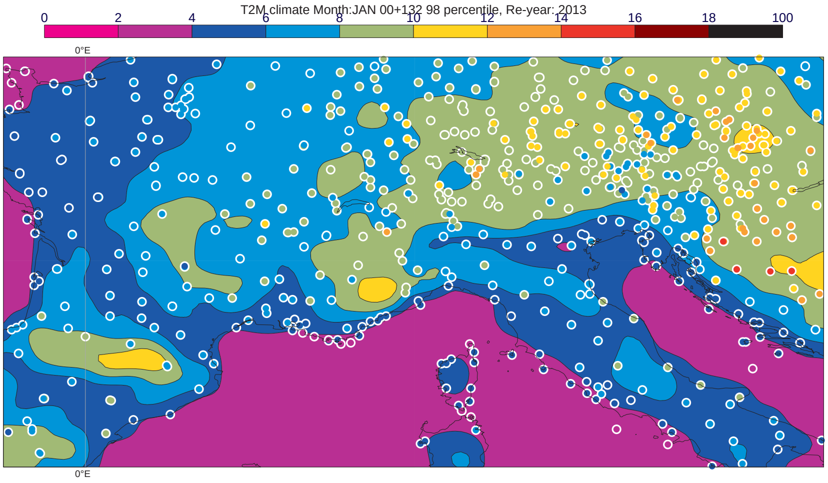
Figure 5 shows the same as Figure 4 but for 24-hour precipitation in August. Here the model generally underestimates the extreme values of precipitation (still 98th percentile) over central Europe. This result is somehow expected as the model output is integrated over a grid box, while the observation is a point measurement. This creates a discrepency during localized convective events. Another area where the model underestimates the values is the northern side of the Alps, indicating an underestimation of orographic rainfall enhancement in upslope-flow situations. During the floods in Central Europe in June 2013 this was the case and is discussed in Haiden et al. (2014).

For extreme values of temperature we have to consider both tails of the distribution (warm and cold events). Figure 6(a) shows the difference between the median (50th percentile) and the 2nd percentile for 2-metre temperature in January, verified at 12 UTC. The reason for plotting the difference to the median is the large difference in mean values between northern and southern Europe. In general, the difference between the median and the extreme percentiles is much less over sea than over land. For the cold extremes the model in general underestimates the width of the cold tail of the distribution, especially over southern Scandinavia. This is an effect of strong inversions over snow, which are insufficiently captured by the model, and which persist during the day at northern latitudes winter time because of the lack of sunlight.

Figure 6(b) shows the the difference between the 98th and 50th percentile of the temperature for southern Europe in January. North of the Alps (and over the Balkans) the warm extremes are underestimated in the model. This bias could be due to a weaker Foehn effect in the model compared to reality, but this



(a) 50th-2nd percentile (northern continental Europe and southern Scandinavia)



(b) 98th-50th percentile (southern Europe)

Figure 6: Difference in values for extreme (2nd in the upper and 98th percentile in the lower panel) to the median (50th percentile) for 2-metre temperaure in January. Model climate (shade) and observed climatology (dots).

needs further investigation.

Figure 7 shows the same as Figure 6 but for July. Here the model underestimates the warm anomalies during the summer while the cold anomalies are in better agreement. The problems with the warm anomalies seem to be worst over southern Scandinavia where the modelled difference to the median is 6-8 degrees while the observed differences for many stations are 10-12 degrees. Similar errors are present over France and south-eastern England, where the observed variability is similar to the variability over eastern Europe where the weather is more dominated by continental air masses.

Figure 8 shows the evolution of frequency bias from 2002 to 2013 for the 98 percentile of 10-metre wind speed (a), 24-hour precipitation (b) and 2-metre temperature (c). All data is valid for 12 UTC. In the figure results are included for HRES (solid), CTRL (dotted) and ERA Interim (dashed) and 1-day (red), 4-day (green) and 7-day (blue) 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 a fixed forecasting system throughout this period; hence its variability with respect to the frequency bias mainly reflects atmospheric variability. It should be noted that 10-metre wind speed and 2-metre temperature are not prognostic model variables but are interpolated to represent the measurement height (2 and 10 metres respectively), which is an additional source of uncertainty.

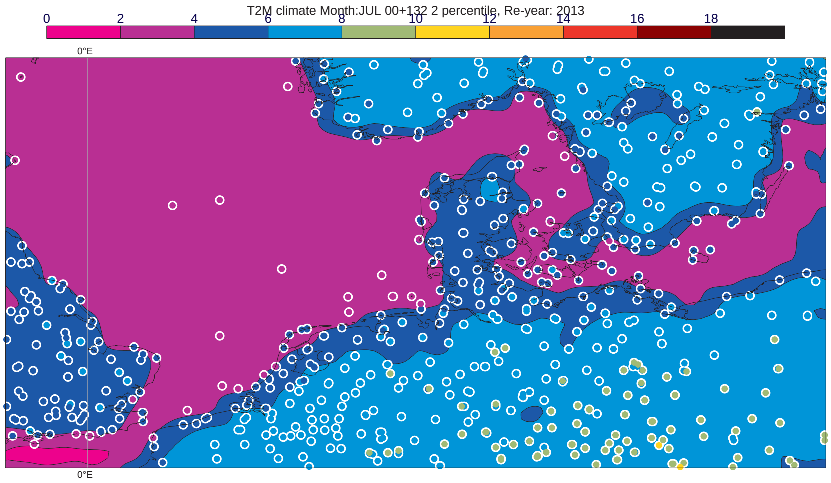
For 10-metre wind speed (Figure 8a), HRES, CTRL, and ERA-Interim all over-forecast the extreme winds for the main part of the time-series. The frequency bias was similar for all three forecasts around 2007, when HRES and CTRL used the same model physics as ERA-Interim. In June 2011, the roughness length was modified in the model, 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.

In the case of 24-hour precipitation, the frequency bias is generally negative for the 98th percentile. An interesting feature here is the spin-up in the model during the first days. For ERA-I the frequency of the event during the first day of integration is much lower that for longer lead times. For HRES the opposite holds true since 2009, the event is more frequent during the first day and then the frequency decreases. For the 95th percentile (not shown), the precipitation is over-forecast during the first day and for longer lead-times the bias is small. Such a spin-up in the model can be due to the interaction between the data-assimilation and the forecast model.

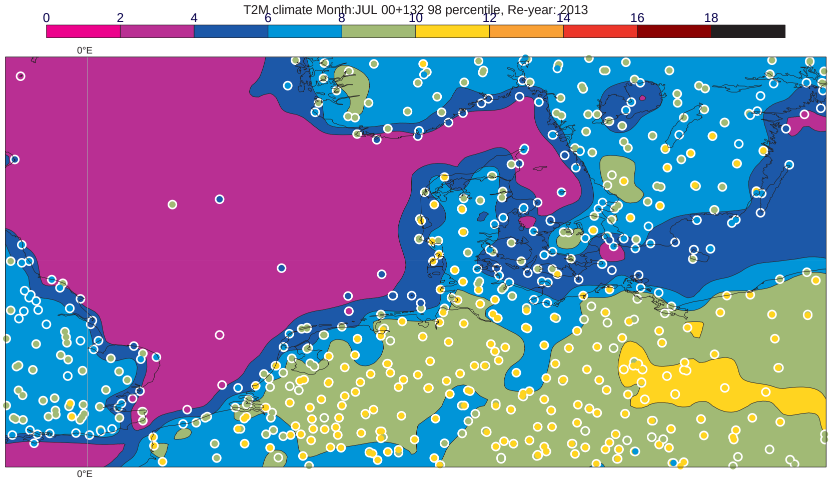
For 2-metre temperature a large variability from year to year is present. The variability is also present for ERA-I, suggesting that it is due to the interannual variability, where e.g a mild winter could have a different bias from a hot summer, even if both falls into the same percentile.

6 Forecast skill

Before we start to discuss longer-term verification results, we will give an example of forecast verification for one particular day. Figure 9 shows hits (green), misses (red) and false alarms (blue) for 10-metre wind speed valid 28 October 2013 12 UTC. The figure also includes the mean-sea-level-pressure for the forecast (black contours) and the analysis (blue contours). The forecast was initialised 2.5 days before (+60h). The forecast is here, for the purpose of illustration, verified against the analysis. The top panel shows verification of an event defined as wind speed greater than 16 m/s, while the bottom panel uses

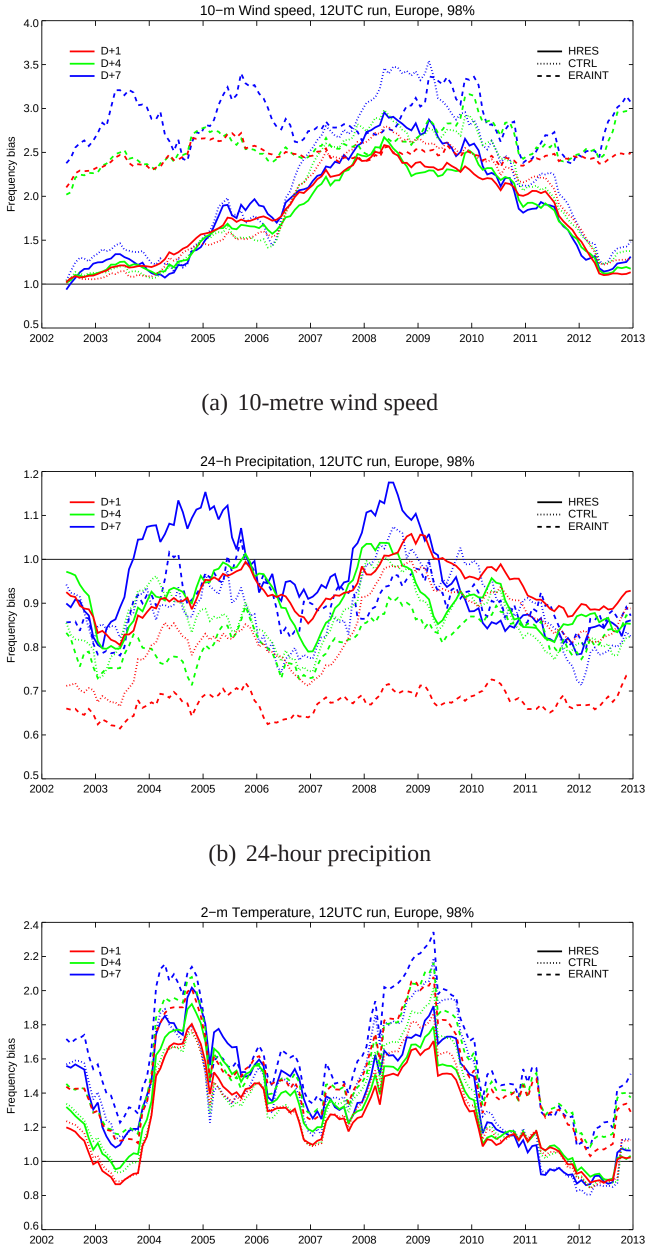


(a) 50th-2nd percentile (northern continental Europe and southern Scandinavia)



(b) 98th-50th percentile (northern continental Europe and southern Scandinavia)

Figure 7: Difference in values for extreme (2nd in the upper and 98th percentile in the lower panel) to the median (50th percentile) for 2-metre temperaure in July. Model climate (shade) and observed climatology (dots).

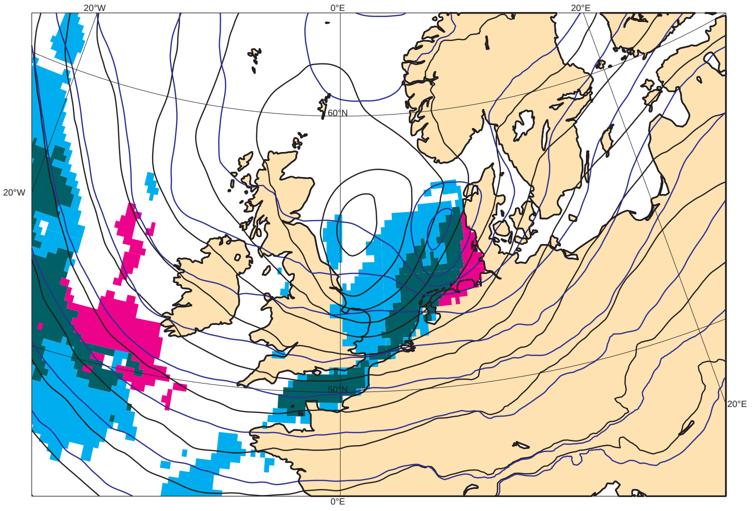


(a) 10-metre wind speed

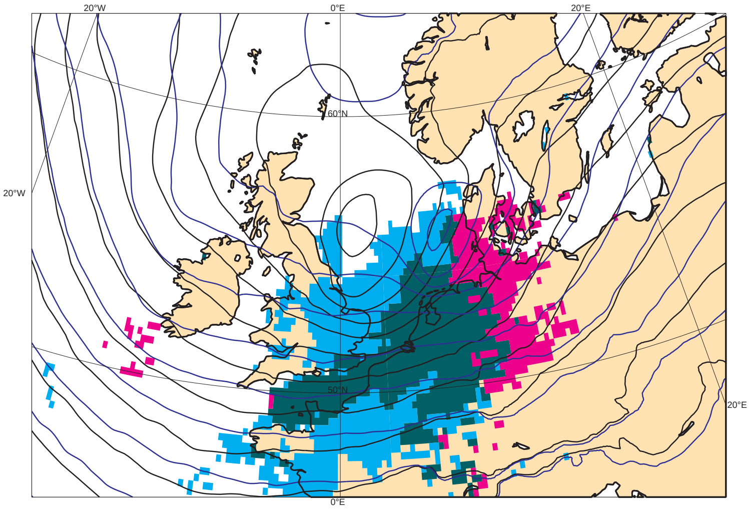
(b) 24-hour precipition

(c) Temperature (warm)

Figure 8: Time-series for 2002-2013 (1-year running mean) of frequency bias for 98th precentile over Europe. HRES(solid), CTRL(dotted) and ERA Interim (dashed). Different lead times in different colors.



(a) +60h wind speed over 16 m/s



(b) +60h wind speed over 98th percentile

Figure 9: Example of hits (green), misses (red) and false alarms (blue) for forecasts valid 28 October 2013 12z. The forecasts are verified against the own analysis.

the 98th percentile of the model climate derived from the reforecast data set for October. The threshold 16 m/s corresponds to the 98th percentile of the model climate over the North Sea.

On this day, a severe storm (Christian) hit the countries around the North Sea. In the forecast, the centre of the cyclone was somewhat shifted to the west, leading to false alarms to the west and misses to the east. Comparing the two different events (16 m/s vs. 98th percentile), using a fixed absolute threshold (16 m/s) leads to exceedence of the threshold mainly over sea, while defining the event relative to the model climate gives signals also over land. Because of this property, we will use the relative threshold in the rest of the report.

6.1 SEDI score

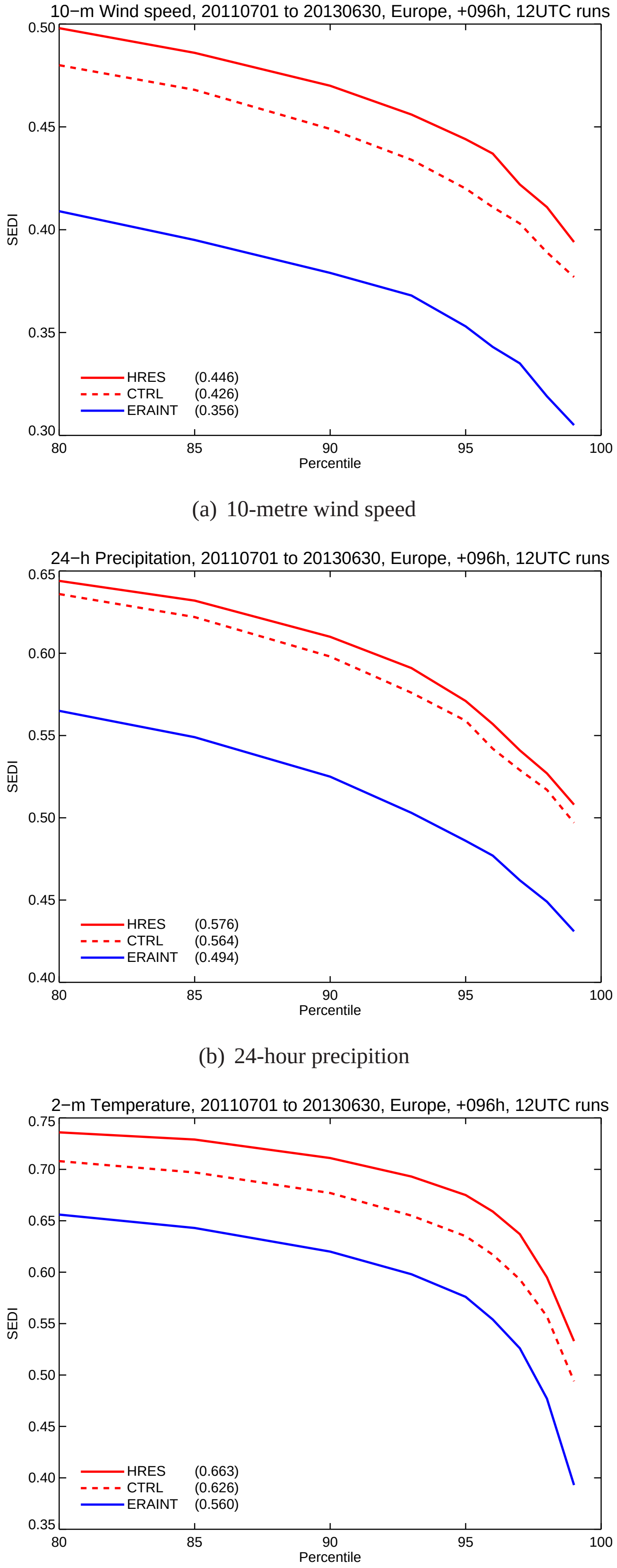
Figure 10 shows the SEDI score as a function of the evaluated percentile, for 4-day forecasts. As described above, SEDI is designed to not explicitly depend on the base rate. Therefore a change in SEDI for higher percentiles reflects an actual change in the ability of the forecasting system in predicting such events. However, we have to bear in mind that the uncertainty in the score increases with decreasing sample size. In general, the SEDI decreases with more extreme events (higher percentiles), and it does so more rapidly for percentiles above the 95th. It seems like the behaviour for the 95th percentile is still in the ”comfort zone” of the forecast system, while for higher percentiles the skill deteriorates considerably. This is especially true for 2-metre temperature. In order to verify the forecasts for extreme values outside this comfort zone, we will subsequently mainly focus on the 98th percentile.

In the figure, results for the HRES (red-solid), ENS control (red-dashed) and ERA-I (blue-solid) are included. Comparing HRES and CTRL, HRES is noticeably better, as expected for all parameters and percentiles, but the difference does not seem to increase with more extreme events. The difference between those two (which is solely due to difference in model resolution), is smallest for precipitation and largest for temperature, which is a somewhat unexpected result. Comparing HRES and ERA-I, the difference between the two increases with higher percentiles for temperature, but is close to constant for the other two parameters.

Figure 11 shows the SEDI score for the 98th percentile for July 2011 to June 2013 as a function of lead time. As expected the skill of the forecasts decreases with incresing lead time. The skill is worst for 10-metre wind speed, while for short-range forecasts the precipitation has the highest SEDI and for long lead times the best forecasts are for 2-metre temperature. The reason for the best performance for precipitation for short forecasts is probably because the precipition is an accumulated quantity over 24-hours while the wind speed and temperature are instantaneous values. For temperature, all three forecasts have some skill also for 10-day forecasts (the score is 0.3 for HRES), which is a sign of long-range predictability for heat-waves.

Comparing HRES and CTRL, the largest impact of the resolution is seen for 2-metre temperature (as also seen in Figure 10). This holds true also for ERA-I compared to HRES and CTRL. Having the largest resolution dependence for temperature forecasts is again somewhat unexpected, but could be due to how well coastlines and mountain areas are resolved.

Comparing the scores for a lead time of 0 days, we can compare the difference between the operational HRES analysis and the ERA-I reanalysis. The difference is smallest for precipitation where the ERA-I forecast for the first 24 hours is as good as HRES forecast accumulated between 24-48 hours. As seen in the frequency bias, HRES and ERA-I have different kind of spin-up biases for precipitation, with too little precipitation for ERA-I and relatively too much for HRES during the first forecast day. For wind speed, the ERA-I reanalysis has the same SEDI score as a 3-day forecast from HRES. For 2-metre

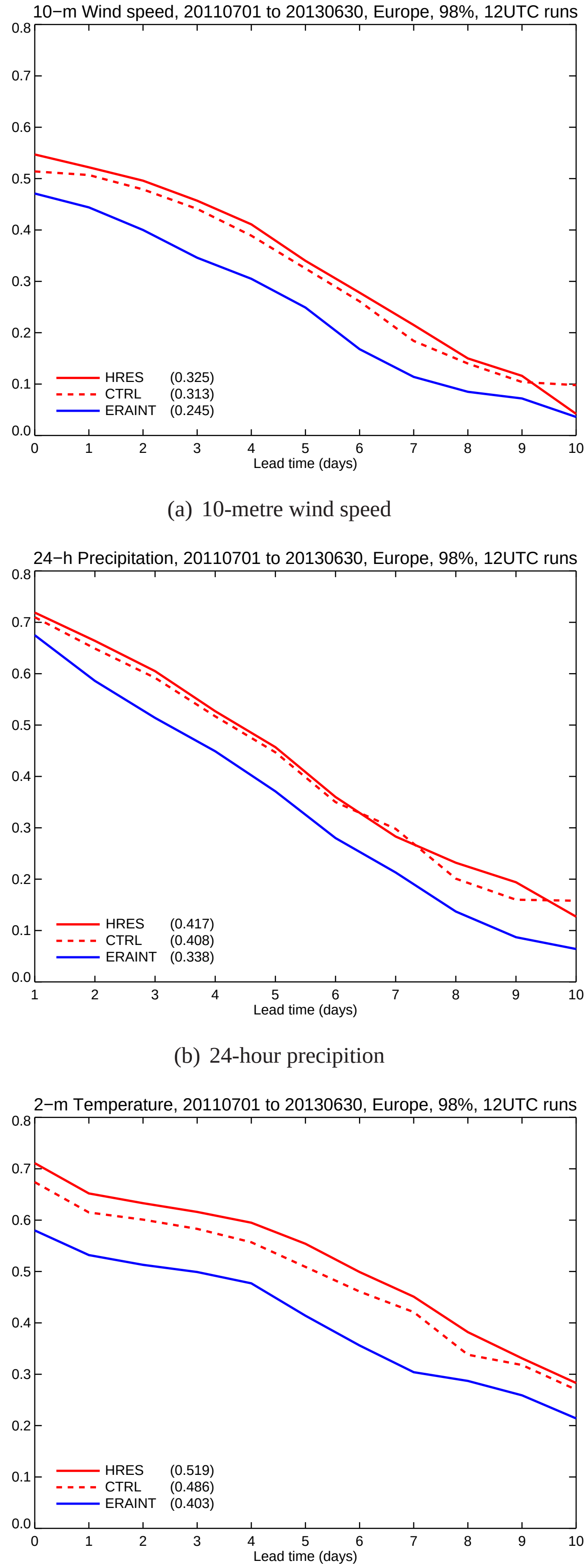


(a) 10-metre wind speed

(b) 24-hour precipition

(c) Temperature (warm)

Figure 10: SEDI score from July 2011 to June 2013 as function of verified percentile for 4-day forecasts. HRES (red-solid), CTRL (red-dashed) and ERA-I (blue-solid).



(a) 10-metre wind speed

(b) 24-hour precipition

(c) Temperature (warm)

Figure 11: SEDI score from July 2011 to June 2013 as function of forecast lead time for the 98th percentile.

temperature, the difference is largest and the ERA-I reanalysis is not better than a 4 day forecast from HRES.

Figures 12-14 illustrates to what extent forecast skill has improved over time. For the long-term development of the upper-air scores and the other scores for surface parameters, we refer to Richardson et al. (2013) and Magnusson and Källén (2013). The panels show 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 improves 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.

For wind speed, we see a large interannual variability for the difference in SEDI for day 7, which is not present for day 1 and day 4. The feature is present for all three percentiles but is most apparent for the 98th. The feature is strongest in 2007-2008. For these years, the observed frequency was higher than normal (not shown), which could affect the predictability.

Figure 13 shows the same as Figure 12, but for 24-hour precipitation. Here the improvement with time is less than for the 10-metre wind. In general the difference to ERA-I is smallest for the short-range forecasts. One exception is in 2010 for the 98th percentile when the 7-day forecast shows the lowest difference. This coincides with a temporarily higher frequency for the event.

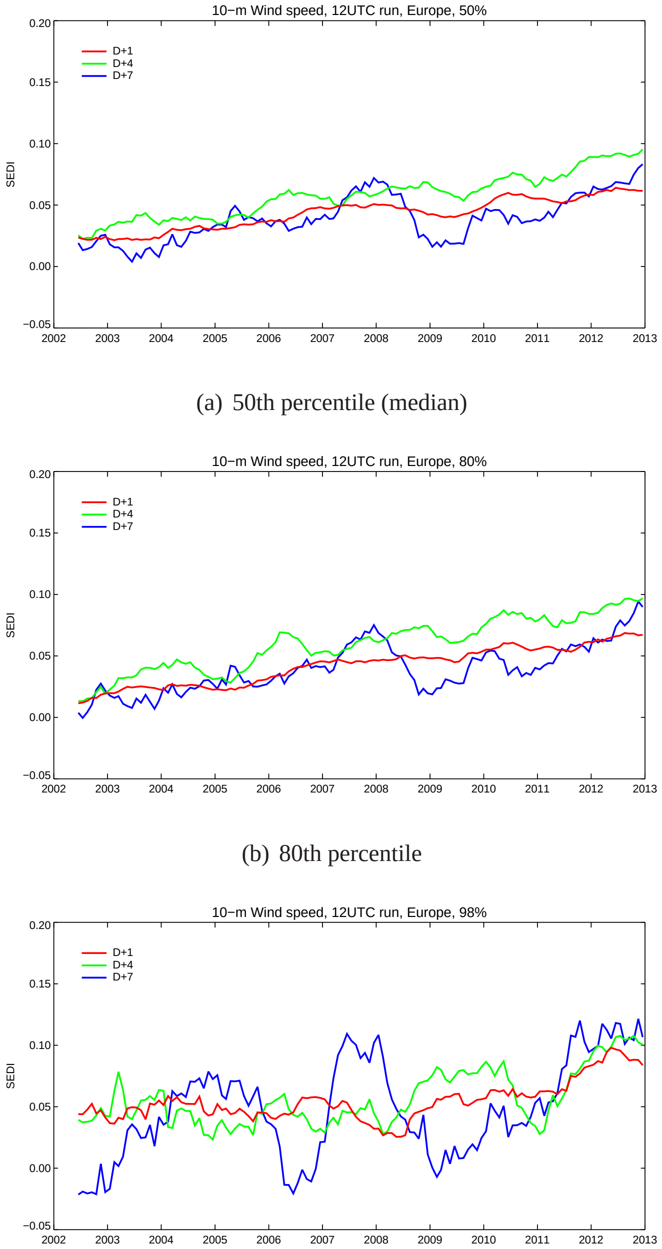
For temperature (Figure 14), the first 3 years of the time-series show no difference for HRES and ERA-I for 7-day forecasts, but a positive difference for the earlier lead times. However, during the last year (2012) the 7-day forecast shows the largest difference to ERA-I and has hence experienced the greatest improvement over the last 10 years.

A general conclusion from these plots, although the results are noisy, is that over the past ten years SEDI has improved faster for the 98th percentile than the 50th and 80th percentile. Hence, the forecasts for extremes seem to have benefitted at least as much from the model and data assimilation changes as the general forecast.

6.2 Results on potential economic value

As a more user-orientated verification, we will in this section show the potential economical value (PEV). In the end of the section we present a comparison between different centres for this measure. The PEV is calculated for each cost/loss ratio and is plotted as a function of the ratio. Having a PEV value above zero means that there is an economic gain in using the forecast instead of knowledge about the climatology of the event, for a certain cost/loss ratio. A value of 1 corresponds to the economical gain if the correct decision was always taken. For a deterministic forecast, the highest PEV is expected to occur at the cost/loss ratio which corresponds to the base rate of the event. Hence, for a rare event, the cost of action has to be much less than the potential loss, in order to gain anything from the forecasts (otherwise numerous false alarms will cause large costs).

By using PEV as a measure we can directly compare the value of an ensemble forecast with that of a single, deterministic forecast. By comparing HRES and CTRL the additional value of higher resolution can be evaluated. Comparing CTRL and ENS the effect of probabilistic forecasts compared to deterministic forecasts is shown and comparing HRES and ENS the we have both effects. One question here is if the

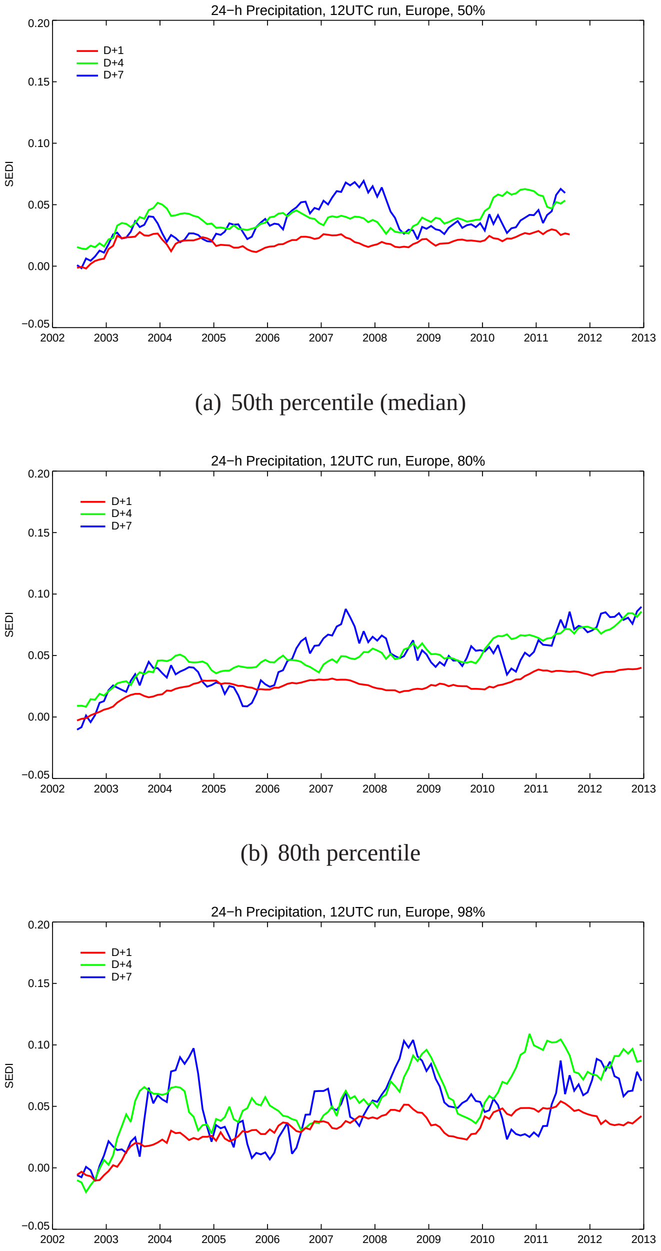


(a) 50th percentile (median)

(b) 80th percentile

(c) 98th percentile

Figure 12: Time-series from 2002 to 2013 of the difference in SEDI between HRES and ERA-I for 10-metre wind speeds.

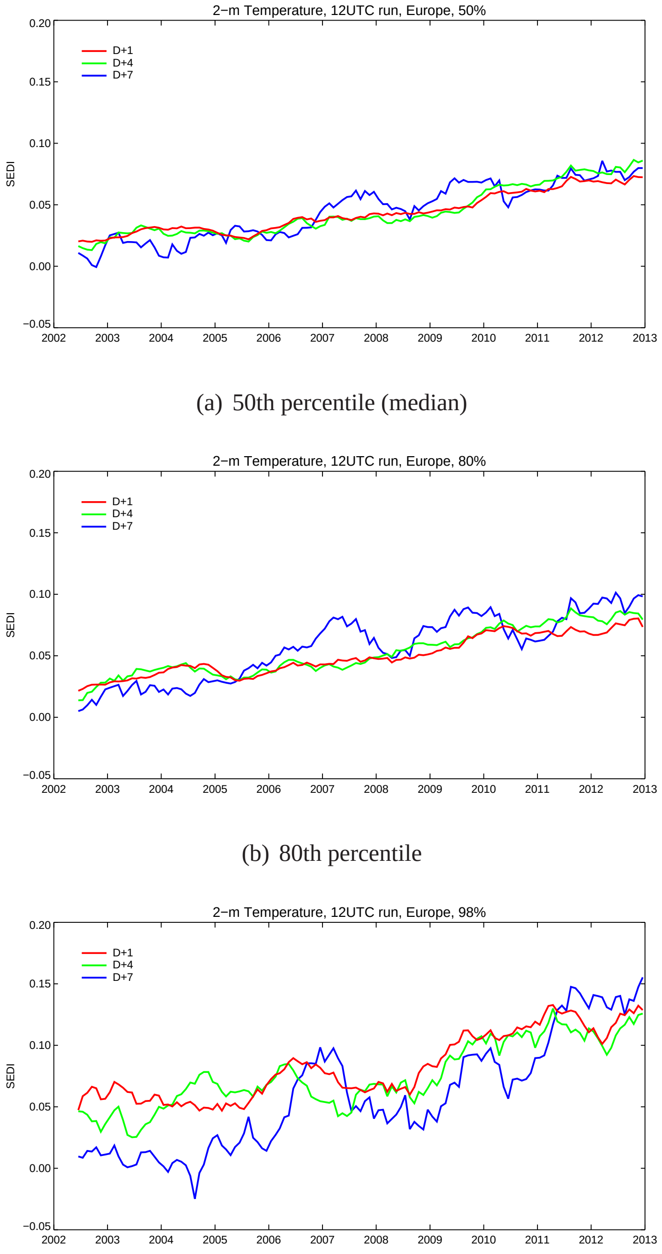


(a) 50th percentile (median)

(b) 80th percentile

(c) 98th percentile

Figure 13: Time-series from 2002 to 2013 of the difference in SEDI between HRES and ERA-I for 24-hour precipitation.

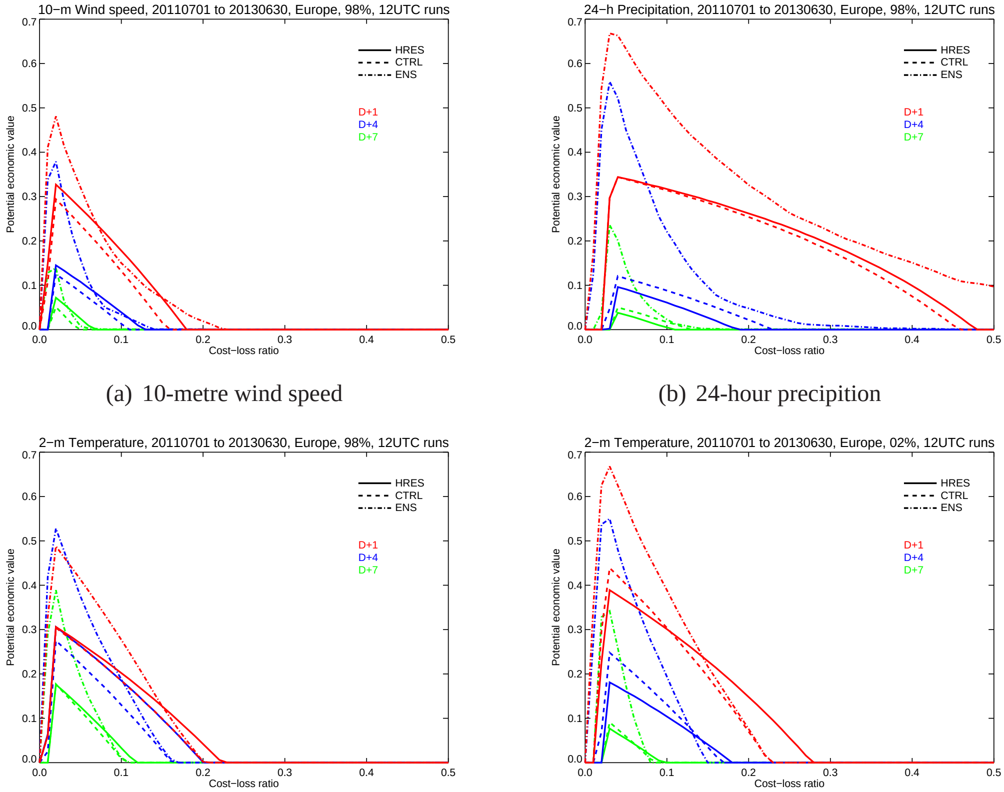


(a) 50th percentile (median)

(b) 80th percentile

(c) 98th percentile

Figure 14: Time-series from 2002 to 2013 of the difference in SEDI between HRES and ERA-I for 2-metre temperature.



(a) 10-metre wind speed

(b) 24-hour precipition

(c) Temperature (warm)

(d) Temperature (cold)

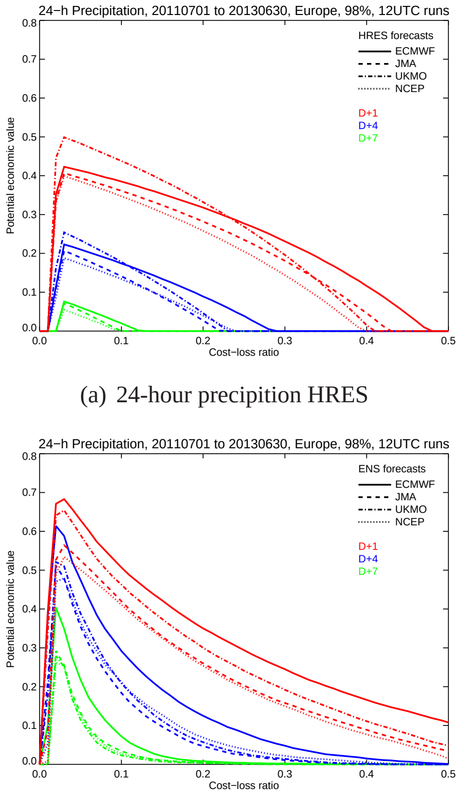
Figure 15: Potential economical value for the 98th percentile (2nd percentile for cold temperature) between July 2011 and June 2013. The plotted forecast lead times are 1 (red), 4 (blue) and 7 (green) days. The plotted forecasts are HRES (solid), CTRL (dotted) and ENS (dash-dotted).

gain in using an ensemble is larger than the loss due to its lower resolution.

Figure 15 shows the potential economic value for the 98th percentile (2nd percentile for cold temperature event) for forecasts from July 2011 to June 2013. In these plots we compare HRES, CTRL and ENS. For all variables and lead times the PEV benefits from the ensemble information (ENS compared to CTRL). The difference is smallest for temperatures and largest for precipitation. One unexpected result for the precipitation verification is the better result for CTRL compared to HRES for day 4 and 7. One could speculate that the higher resolution brings more details in the forecast that might decrease the score for long lead times.

For temperature, the PEV is better for warm extremes than cold extremes at day 7. The reason for this has to further investigated but one would expect that issues with strong inversions during wintertime could be a reason for the lower skill with respect to cold extremes. For both temperature extremes the HRES has a non-zero PEV for higher cost-loss ratios than ENS. The ENS does not improve the upper cost/loss limit for temperature as it does for wind and precipitation.

Figure 16 shows PEV for three lead times for 4 different forecasting centres (ECMWF - solid, JMA - dashed, UKMO - dash-dotted and NCEP - dotted). The upper panel shows the results for HRES forecasts. For Day 1 and Day 4 UKMO has a higher PEV than ECMWF for low cost/loss ratios while the opposite holds true for higher ratios. One explanation for the difference is the discrepancy in frequency biases between the centres. As seen in Figure 8, ECMWF has a negative frequency bias for the 98th percentile.



(a) 24-hour precipition HRES

(b) 24-hour precipition ENS

Figure 16: Potential economical value for different forecast centres for the 98th percentile between July 2011 and July 2013. The plotted forecast lead times are 1 (red), 4 (blue) and 7 (green) days.

For UKMO the bias is positive. Overforecasting of an event favours low cost/loss ratios where the cost for an action is low and one can afford many false alarms, while underforecasting is favoring high cost/loss ratios where few false alarms are desirable.

The lower panel in Figure 16 shows the same as the upper panel but for the ensemble forecasts. Here ECMWF has the best PEV for all cost/loss ratios and lead times. Here the effect of the frequency bias is compensated by the choice of optimal probability threshold for a cost/loss ratio (a negative frequency bias is compensated by having a higher probability threshold).

7 Conclusions

In this report we have evaluated the forecast performance for extreme events of temperature, wind speed and precipitation. Verification of extreme events is not straightforward as the sample size is small, and scores have to be carefully chosen to be applicable to rare events. For verification of extreme events one needs to define a threshold which could either be an absolute value or a percentile value relative to climatology. In this report we have chosen the latter definition to allow the event to appear at all grid points. We have mainly focused on the 98th percentile of the climate distribution, as a compromise between sample size and degree of extremeness. However, it can be argued that the 98th percentile can not be considered extreme, since on average such an event should occur once every second month at each grid point (high impact events have rather a return period of more than 10 years). For wind speeds, for example, the 98th percentile represents approximately 8-10 metres per second for the main part of Europe in November. Such a low threshold is needed in order to obtain robust statistics in our verification.

We have compared the high-resolution forecast (HRES), the ensemble control forecast (CTRL) and the ensemble forecast (ENS). We have also included forecasts run in the reanalysis project (ERA Interim). While HRES, CTRL and ENS forecasts use the most recent model version (with lower resolution for CTRL and ENS), ERA-I uses a frozen forecasting system from 2006 together with a lower resolution.

In this report we have focused on verification against SYNOP observations. However, observations contain errors, originating from measurement errors (typing errors or instrument errors) and representiveness errors. To detect measurement errors one needs to have some sort of quality control of the observations. In this report we have only used a simple quality control and we would like to highlight this as an area for further work. The main challenge when verifying against observations that are point measurements is the difference in scales represented by the observation and the model. The model is an area average over a grid box, and with coarse model resolution large representativeness errors can occur.

As a zero-order metric of the ability to forecast extreme weather we have evaluated the frequency of events. In order to predict an event it is crucial that the model can produce such an event with a frequency similar to the observed one. This evaluation is useful to find systematic model issues and to recognize current limitations in simulating extreme events. By studying maps of frequency biases for the 98th percentile we highlighted a few potential sources for biases for extreme events. This type of map is useful for evaluating local biases. The most problematic area in Europe is in the Alps, for all three variables. Steep orography strongly affects the wind speed measured at ground as well as precipition rates and the Foehn effect on the lee side of mountain ranges. Another issue is the underestimation of extreme precipitation amounts from convective events.

Several metrics (scores) are available for verification. One category of scores are based on hits, misses, false alarms and correct negatives of a particular event. Scores are designed to combine these four categories into meaningful metrics. We have focused on the recently developed SEDI score (after having looked previously at the Peirce skill score and the equitable threat score). The advantage with SEDI is that the score remains well-behaved for extreme events, when other scores tend to degenerate. However, the main issue with the SEDI score is that if it is used on uncalibrated forecasts it favours a positive frequency bias. 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 error such that the result needs to be interpreted as potential skill.

We have also used the potential economic value, PEV (Richardson, 2000), which is based on a simple decision model. The PEV is a function of the specific cost/loss ratio for an application. If the cost for preventing a loss is close to the loss of the weather event, it is rarely an advantage to use the information. In contrast, if the cost is small compared to the loss, a preventing action can be taken regardless of the forecast. However, inbetween these extremes the forecast information can play a role in optimising the decision process. The cost/loss ratio for a user could be a function of the forecast lead time, if the cost for an early action is different (in many cases less) than a late action. By using a ensemble system, which is able to produce different probabilities for an event, one can use different probability levels for different lead times.

Comparing different resolutions, our results show the largest impact of resolution on 2-metre temperature. One could speculate that the higher resolution improves land-sea contrast and the agreement between model topography and station height. Recent findings pointing towards filtering of the model orography might also play a role here (Sandu et al., 2014). The difference in scores between the resolutions also impact the PEV scores for ENS, which for some cost-loss ratios are lower than HRES.

The scores are in general lowest for 10-metre wind speed. For this parameter we have seen that local conditions around the stations have a large influence and the representativeness uncertainty is high. For short forecasts (1-3 days) the highest scores are found for precipitation while for medium-range forecasts (day 3-10) the temperature has the highest scores. One reason for this could be that precipitation is accumulated over 24 hours and therefore the timing error is to some degree filtered out. A smaller contributing factor is that what we call a 1-day forecast is accumulated over the first 24 hours of the forecast, making the effective lead time of the forecast shorter.

Comparing the evolution of the SEDI score for three different percentiles (50th, 80th and 98th), we found that SEDI for the 98th percentile has improved more over the past 10 years than the 50th and 80th percentile. This indicates that forecasting extremes have benefitted even more from improvements in the forecast system (data assimilation and model) than the forecasting of more average weather.

We have also compared PEV for forecasts from different centres. We found for precipitation and the deterministic models, that UKMO had higher PEV for low cost/loss ratios, while ECMWF had higher PEV for high cost-loss ratios. This characteristic can be traced back to different frequency biases. The PEV is a hedgable score, and overforecasting an event (positive bias) is favourable at low cost/loss ratios while underforecsting is beneficial at high cost/loss ratios.

As PEV is hedgeable (not a proper score), it is not suitable for use without any additional score. The strenght of the score is that it demostrates the usefulness of the forecast for different types of applications. The SEDI score would also be hedgeable if one would not apply a bias correction. Used on calibrated forecasts it becomes a score for potential skill and does not contain information about systematic over- or underforecasting. So for both scores the frequency bias is a necessary and useful complement. Ideally one would like to have a score which does not degenerate for extreme events and at the same time is not hedgeable.

Good predictability at longer lead times for temperature is expected as it is connected to large-scale patterns. Future work will focus on predictability of heat waves and cold spells, which are events that appears over extended periods (verification of several days together is of interest).

In the time-series of the difference in SEDI between HRES and ERA-I we saw that some of the evolution in difference (especially for day 7), were apparently dependent on the observed frequency of the event. For periods with a higher than normal frequency, the 7-day forecasts scored better compared to ERA-I. One could ask the question whether this is a property of the score or whether the HRES model is a better model under such circumstances.

Except from the removal of frequency bias used for the SEDI score, we have not attempted to do any calibration. We expect that a more sophisticated calibration will improve the forecast skill. This could be done by using the reforecast data set. Ongoing work at ECMWF will explore this topic.

In this report we have focused on scores based on hit and false alarm rates. Future work will include more probabilistic verification. One possibility here 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.

Finally, we acknowledge that for many places and parameters the 98th percentile can not be classified as severe. For really extreme/severe events (with return periods of many years), we believe that one has to look into specific cases to investigate the model characteristics. Hopefully, such case studies would bring more general model issues to light.

Acknowledgements

We would like to acknowledge Erland Källén and Florence Rabier for valuable discussions during the preparation of this report.

References

Andersson, E. and H. Järvinen, 1998: Variational quality control. Technical Memorandum 250, ECMWF.

Bechtold, P., M. Kohler, T. Jung, F. Doblas-Reyes, M. Leutbecher, M. J. Rodwell, F. Vitart, and G. Balsamo, 2008: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Q. J. R. Met. Soc.*, **134**, 1337–1351.

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

Ferro, C. A. T. and D. B. Stephenson, 2011: Extremal Dependence Index: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events. *Weather and Forecasting*, **26**, 699–713.

Ghelli, A. and C. Primo, 2009: One the use of the extreme dependency score to investigate the performance of a NWP model for rare events. *Meteorol. Appl.*, **16**, 537–544.

Haiden, T., M. J. Rodwell, D. S. Richardcon, A. Okagaki, T. Robinson, and T. Hewson, 2012: Intercomparison of Global Model Precipitation Forecast Skill in 2010/11 Using the SEEPS Score. *Mon. Wea. Rev.*, **140**, 2720–2733.

Haiden, T., et al., 2014: ECMWF forecast performance during the June 2013 flood in Central Europe. Technical Memorandum 723, ECMWF.

Jung, T., et al., 2010: The ECMWF model climate: Recent progress through improved physical parametrizations. *Q. J. R. Met. Soc.*, **136**, 1145–1160.

Magnusson, L. and E. Källén, 2013: Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Wea. Rev.*, **141**, 3142–3153.

North, R., M. Trueman, M. Mittermaier, and M. Rodwell, 2013: An assessment of the SEEPS and SEDI metrics for the verification of 6-hour forecast precipitation accumulations. *Meteorol. Appl.*, **20**, 164–175.

Primo, C. and A. Ghelli, 2009: The affect of the base rate on the extreme dependency score. *Meteorol. Appl.*, **16**, 533–535.

Richardson, D., J. Bidlot, L. Ferranti, T. Haiden, T. Hewson, M. Janousek, F. Prates, and F. Vitart, 2013: Evaluation of ECMWF forecasts, including 2012-2013 upgrades. Technical Memorandum 710, ECMWF.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Q. J. R. Met. Soc.*, **126**, 649–668.

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

Sandu, I., N. Wedi, A. Bozzo, P. Bechtold, A. Beljaars, and M. Leutbecher, 2014: On the near surface temperature differences between HRES and CTL ENS forecasts. RD Memorandum RD14-315, ECMWF.

Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorol. Appl.*, **15**, 41–50.

Zsoter, E., F. Pappenberger, and D. Richardson, 2014: Sensitivity of model climate to sampling configurations and the impact on the Extreme Forecast Index. *Met. Apps.*, **Accepted**, doi:10.1002/met.1447.