Calibrating Ensemble Forecasts to Produce More Reliable Probabilistic Extreme Weather Forecasts

Kaisa Ylinen
Finnish Meteorological Institute
kaisa.ylinen@fmi.fi

Juha Kilpinen
Finnish Meteorological Institute
juha.kilpinen@fmi.fi

ABSTRACT

Accurate predictions of severe weather events are extremely important for society, economy, and environment. Due to the fact that weather forecasts are inherently uncertain, it is required to give information about forecast uncertainty to all users providing weather forecasts in probabilistic terms utilizing ensemble forecasts. Since ensemble forecasts tend to be under dispersive and biased, they need to be calibrated with statistical methods. This paper presents a method for the calibration of temperature forecasts using Gaussian regression, and the calibration of wind gust forecasts with a box-cox t-distribution method. Statistical calibration was made for the operational European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system (ENS) forecasts for lead times from 3 to 360 hours. The verification results showed that calibration improved both temperature and wind gust ensemble forecasts. The probabilistic temperature forecasts were better after calibration over whole lead time scale, but the probabilistic wind gust forecasts up to 240 hours.

Keywords

Weather forecasts, probabilistic forecasting, statistical calibration, high impact weather events.

INTRODUCTION

Due to climate change, extreme weather events are expected to become more frequent, more extreme and longer lasting. Greater evaporation will lead to increased water vapor in the atmosphere, producing more intense precipitation. This, together with rapid snow melting, increases the likelihood of floods. Also, higher temperatures will increase the frequency of wildfires and heat waves as well as other natural disasters.

Current systems for risk management are still limited in their effectiveness. Despite technological progresses and the availability of large amounts of data, there is no platform that integrates and analyses in real time all the useful data to manage emergencies for improving prediction and management of extreme climate and weather events. At the moment, there are several on-going EU projects (http://www.i-react.eu/, http://beaware-project.eu/ and http://anywhere-h2020.eu/) that are aiming to solve this problem. All these project have the similar objective to establish a multi-hazard platform to integrate emergency management data coming from multiple sources, including that provided by citizens. This way, the necessary information from different sources can be produced faster for citizens and civil protection agencies to enhance prevention and reacting against natural disasters.

One objective of the I-REACT project, is to improve European level high-impact weather event detection. Finnish Meteorological Institute (FMI) is responsible for developing and providing the forecasted occurrence risk maps for the system. These risk maps provide information in terms of probabilities based on relevant thresholds of precipitation amount, wind speed, and temperature phenomena that can have serious societal impact. FMI's focus in the project is in the medium-range weather forecasts which consist the lead times from few hours to two weeks.

Nowadays, Numerical Weather Prediction (NWP) models are the main tool to provide weather forecast guidance to stakeholders. Different NWP models support weather forecasting from daily to weekly forecast scales and also from monthly to seasonal scales (e.g. Molteni et.al. 1996; Willet et.al. 2009). The initial state for NWP models is made using data-assimilation. In data-assimilation process atmospheric physical variables represented by weather observations are combined with short range forecasts fields of previous forecast cycle to make the best estimate for atmospheric state. NWP models can simulate most large scale physical processes in the atmosphere but some

processes are totally excluded or inadequately modelled. The initial analysis has also errors and together with modelling errors the forecast will also be erroneous. Due to the fact that the atmosphere is a chaotic system, the error is typically small in short forecast ranges and it will increase towards longer forecast ranges. Therefore, rather than integrating a single forecast from a supposedly best guess of the initial state, it has been shown that a better approach would be to provide multiple separate forecasts with slightly different initial conditions, or different formulations of a forecast model (Palmer, 2000). This approach is called ensemble prediction and is used to estimate the forecast uncertainty. The probabilistic output of ensemble forecast system is a useful tool for end-users to base the decisions on, e.g. the impacts of hazardous extreme weather events.

In the ideal situation, the ensemble spread would equal the true probability distribution of forecast uncertainty. In reality, medium-range ensemble forecasts suffer from local forecast biases and also in most cases from too high forecast confidence compared to measured forecast uncertainty. Due to this, post-processing for ensemble forecast output is necessary to guarantee optimal forecast reliability and accuracy. In this process statistical post-processing methods are applied to the raw model output. Most of the recently used statistical methods share a general approach of correcting the current forecast by using past forecast errors, as has been done for deterministic forecasts in the so-called Model Output Statistics (MOS) procedure introduced originally by Glahn and Lowry (1972). This process makes use of information from prior forecasts and observations to produce probabilistic forecasts or to improve their reliability. The method providing the best outcome is dependent on the weather variable being forecasted. The calibration has been shown to be useful at a variety of forecast time scales including the lead times from few hours (e.g. Lugt, 2013) to two weeks (e.g. Wilks, 2009). Different statistical methods and approaches have also been widely used (Hansen and Emanuel. 2003, Rosgaard et. al. 2016, Wilks 2015).

This paper describes the calibration of European Centre for Medium-Range Weather Forecast (ECMWF) ensemble prediction system (ENS) forecasts. The calibration methods that were used are Gaussian distribution for temperature and Box-Cox t-distribution for wind gust forecasts. These calibration methods are currently in use of The Grand Limited Area Model Ensemble Prediction System, GLAMEPS. The GALMEPS model (Iversen et.al. 2011) and calibration of the model are developed by the HIRLAM-C consortium which consists of several European meteorological institutes. Compared to global models (like ECMWF), limited area models are able to produce finer spatial resolution output because computational resources are focused on a specific area instead of covering the whole globe. Therefore, these models can usually produce better forecasts, but the forecast length of limited area models is only two or three days which is much less compared to global models which provide forecasts up to 10-15 days ahead. In this study, our goal was to investigate whether the calibration methods that are used for calibration of GLAMEPS model, could be applied also for longer range ECMWF ENS forecasts in European domain.

In the I-REACT project, we utilize both the short term GLAMEPS forecasts and the longer term ECMWF ENS forecasts to provide reliable probabilistic forecasts for high-impact weather monitoring. Provided early warning products estimate the probability of occurrence of heavy rainfall, strong wind gusts, and extreme high/low temperatures, and are routinely produced across Europe. The forecast probability of the occurrence of different severe weather events is defined based on specific pre-defined thresholds which can be modified based on the end-users needs. In addition, the calibrated ensemble forecasts are provided as an input for weather dependent impact variables such as fire weather index and heat wave forecast.

FORECAST AND OBSERVATION DATA

The ECMWF operational ensemble forecast system (ENS) consists of 51 members computed twice a day at 0000UTC and 1200UTC. Temporal resolution of the forecasts is 3 hours for lead times up to 144 hours, and 6 hours for lead times from 150 to 360 hours. The ECMWF model covers the entire globe but for this study we only used the European domain. Forecasts were extracted on a 0.2° grid from 32.60°N to 73.40°N and from 26.00°W to 42.40°E (Figure 1). For calibration coefficient calculation, 2 meter temperature and 10 meter wind gust forecasts were interpolated to the observation station points.

Temperature and wind gust observations were gathered from European SYNOP (surface synoptic observations) stations. A plot of these station locations that where operationally available for calibration is provided in Figure 1. Also, the information of the station elevation was gathered for the calibration model definition.

Figure 1: Forecasting domain and station locations (dots) for the calibration and verification of ECMWF ensemble forecasts.

METHODS

Good (and useful) ensemble forecasts should produce no mean errors (bias); otherwise, the probabilities will be biased as well. The ensemble forecasts should also have the ability to span the full climatological range otherwise the probabilities will either over- or under-forecast the risks of anomalous or extreme weather events. This can be validated by comparing the spread-skill relationship, which should be positively correlated on average in well-calibrated ensemble forecasts. The calibration method that is most suitable for specific variable is mostly dependent on the error distribution that forecasts tend to have.

Calibration Methods

For 2 meter temperature, a Gaussian regression (e.g. Gneiting et.al. 2005) was used with the following parameters:

$$mean = a + b * ensemble\_mean + c * model\_elevation$$
$$stdev = exp(d + e * log(ensemble\_std) + f * log(max(1,model\_elevation))),$$

where *ensemble_mean* denotes the mean value and *ensemble_std* the standard deviation of the 51 ensemble members; the variable *model_elevation* indicates the average elevation (meters above the mean sea level) of each grid point in the ECMWF model.

The regression coefficients *a, b, c, d, e,* and *f* were estimated by using point observations and forecasts for the last 30 days and are updated once a week. Separate values were calculated for each lead time and for each forecast cycle independently. The coefficients were determined to be common for the entire region to obtain a large enough training data set. Another benefit of using common calibration coefficients for the entire domain was that forecast maps do not produce inconsistencies between different calibration areas.

For 10 meter wind gust, a box-cox t-distribution (Rigby and Stasinopoulos, 2006) was used with the following parameters:

*mu = a + b * ensemble_mean + c * model_elevation*
*sigma = exp(d + e * log(ensemble_std) + f * log(max(1,model_elevation)))*
*nu = g + h * ensemble_mean*
tau = exp(i),

where *mu* denotes the median of the distribution, *sigma* the variance, *nu* the skewness, and *tau* the kurtosis.

The above-mentioned model distribution was created for each grid point, and converted back to ensemble members by extracting evenly spaced quantiles from the distribution. The rank of the each member was maintained by reordering the members to same order as they have in the raw ensemble. The conserving of the members' order is essential if the ensemble forecasts are used as input for indexes that utilize more than one variable.

Verification Metrics

The reliability characteristics (bias and dispersion) of the probabilistic forecasts were defined with rank histograms (e.g. Wilks, 2006). The rank histogram illustrates in which bin of the ensemble forecast the observation falls. The rank histogram shows how well the ensemble spread of the forecast represents the true variability of the observations. In an ensemble with perfect spread, each member represents an equally likely scenario and the form of rank histogram is flat. A U-shaped form of the rank histogram indicates that the ensemble spread is too small and many observations fall outside the extremes of the ensemble. If a U-shaped rank histogram is asymmetric, the ensemble contains bias.

The Brier Skill Score (BSS) was used to diagnose the quality of the probabilistic forecasts predicting higher temperature and wind gusts. The Brier Score (Brier, 1950) measures the magnitude of the probabilistic forecasts errors computed over the verification sample. The Brier Score (BS) is calculated with the following formula:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(y_\mathrm{i} - o_\mathrm{i})^2.$$

The value of the probability forecast, y_i, is between 0 and 1 and the value of observation, o_i, is either 0 (no occurrence) or 1 (occurrence). The Brier Score ranges between 0 and 1 and the perfect score is 0.

The Brier Skill Score (e.g. Wilks, 2006) measures the relative skill of the probabilistic forecast over the sample climatology (reference forecast). The Brier Skill Score is calculated with following formula:

$$\mathrm{BSS} = 1 - \frac{BS}{BS\_ref}.$$

The Brier Skill Score ranges between -∞ and 1. When the BSS results in positive values, the model is more accurate than the reference. A value of 0 indicates no skill over the reference model, while negative values indicate that the reference model is more accurate.

RESULTS

Figures 2 and 3 show the rank histograms for temperature and gust forecasts with shorter (upper) and longer (lower) lead times. As can be seen from Figure 2, the temperature forecasts tend to be underdispersive and negatively biased before calibration (red bars). Underdispersion is especially a problem in the beginning of medium-range forecasts which can be seen when comparing the amount of outliers of the upper and lower histograms. After calibration, the histogram (orange bars) is almost flat which indicates that calibration can improve the ensemble forecasts. The same kind of feature can be seen from the wind gust forecasts (Figure 3). However, contrary to the temperature forecasts, the wind gust forecasts exhibit positive bias.

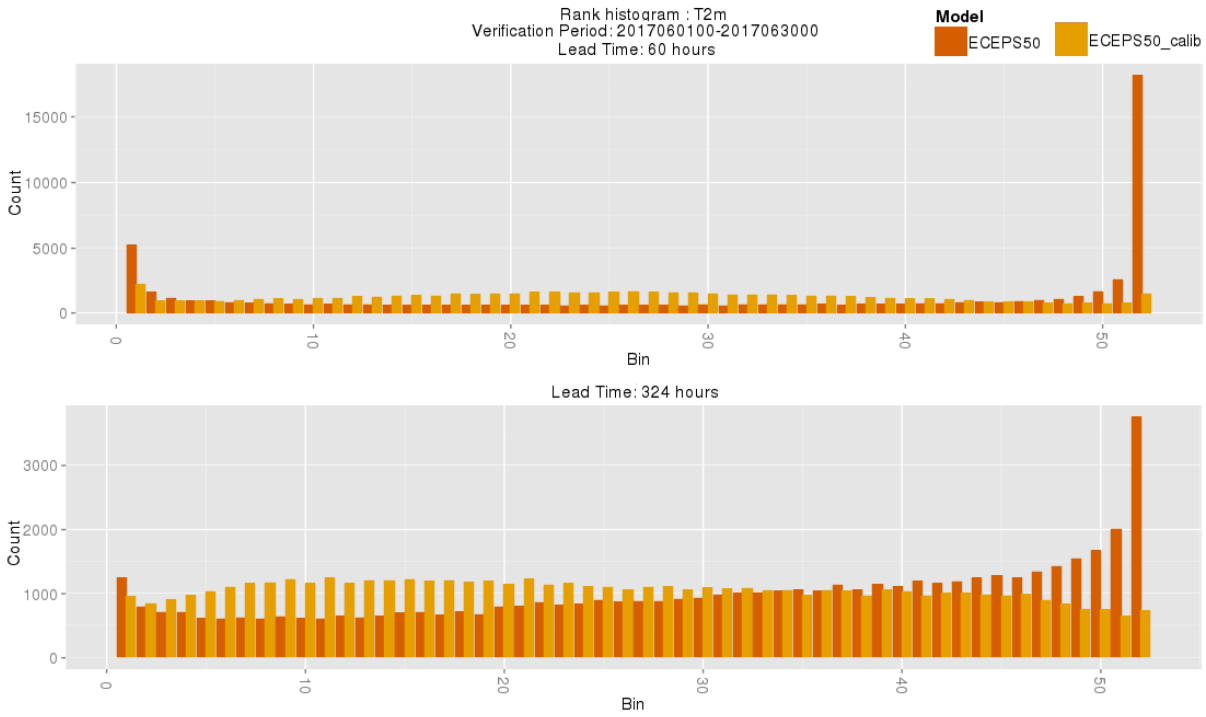

Figure 2: Rank histograms illustrating the under dispersion and negative (cold) bias of raw temperature forecasts with lead times of 60 (upper) and 324 hours (lower). (ECEPS50 = ECMWF-ENS)

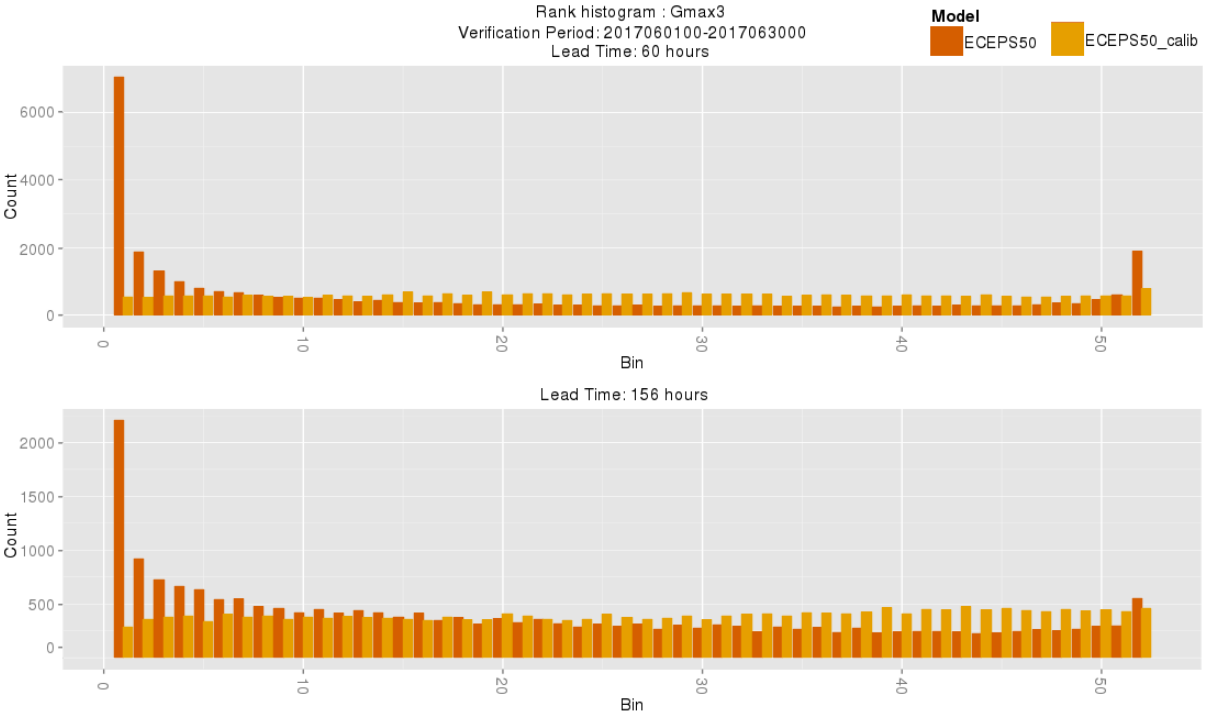

Figure 3: Rank histograms illustrating the under dispersion and positive bias of raw wind gust forecasts with lead times of 60 (upper) and 156 hours (lower). (ECEPS50 = ECMWF-ENS)

In Figure 4, the BSS for probabilistic temperature forecasts with a threshold temperature of 25℃ is shown.The verification results are plotted only for lead times when the forecast time is 12 UTC and the daily temperature is usually the highest. The sample climatology of the verification period is used as a reference model when calculating the BSS. Figure 4 shows that the BSS is better in the calibrated forecasts for all lead times. This result proves that our calibration model is capable to improve ensemble forecasts also in the higher part of temperature

distribution.

The BSS for probabilistic wind gust forecasts with a threshold of 15 m/s are shown in Figure 5. The verification results indicate that wind gust forecasts can be improved with calibration, especially with lead times of up to 240 hours. However, for lead times above 240 hours the raw and calibrated forecasts are generally equally good.

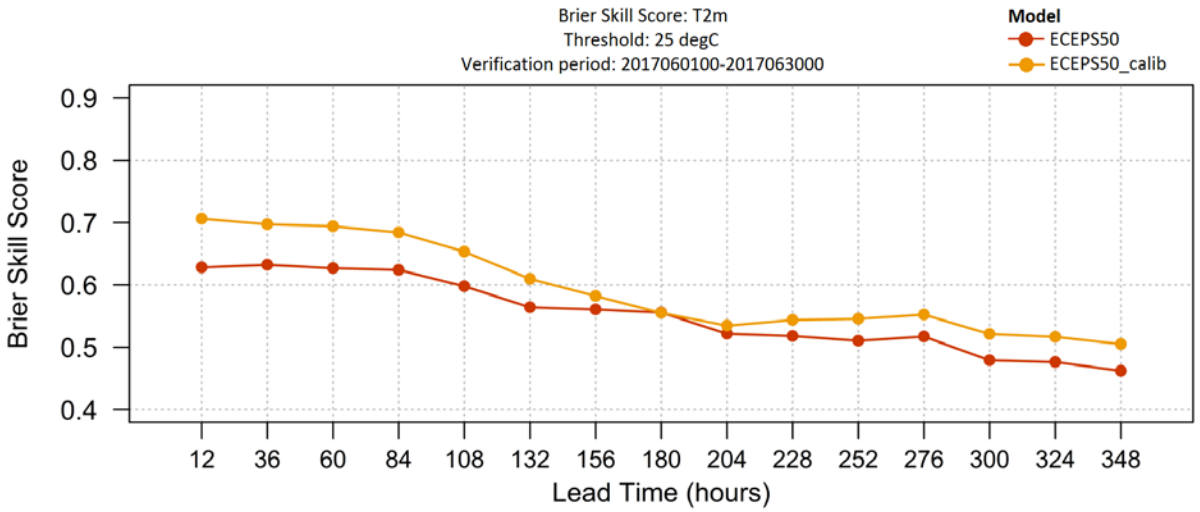

Figure 4: The Brier Skill Score for the probabilistic temperature forecasts (threshold 25°C)

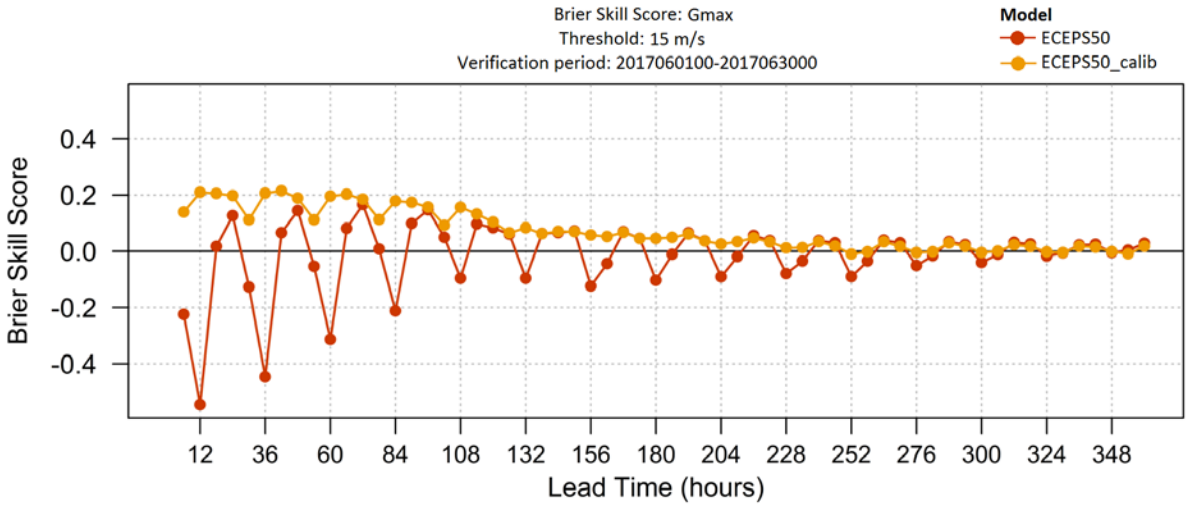

Figure 5: The Brier skill score for probabilistic wind gust forecasts (threshold 15 m/s)

Figures 6 and 7 illustrate the probabilistic weather forecasts for high temperature and strong gust respectively. Forecasts are based on the calibrated ECMWF ENS forecasts. The threshold for probabilistic temperature forecast is 37°C, and for probabilistic gust forecast 17 m/s.

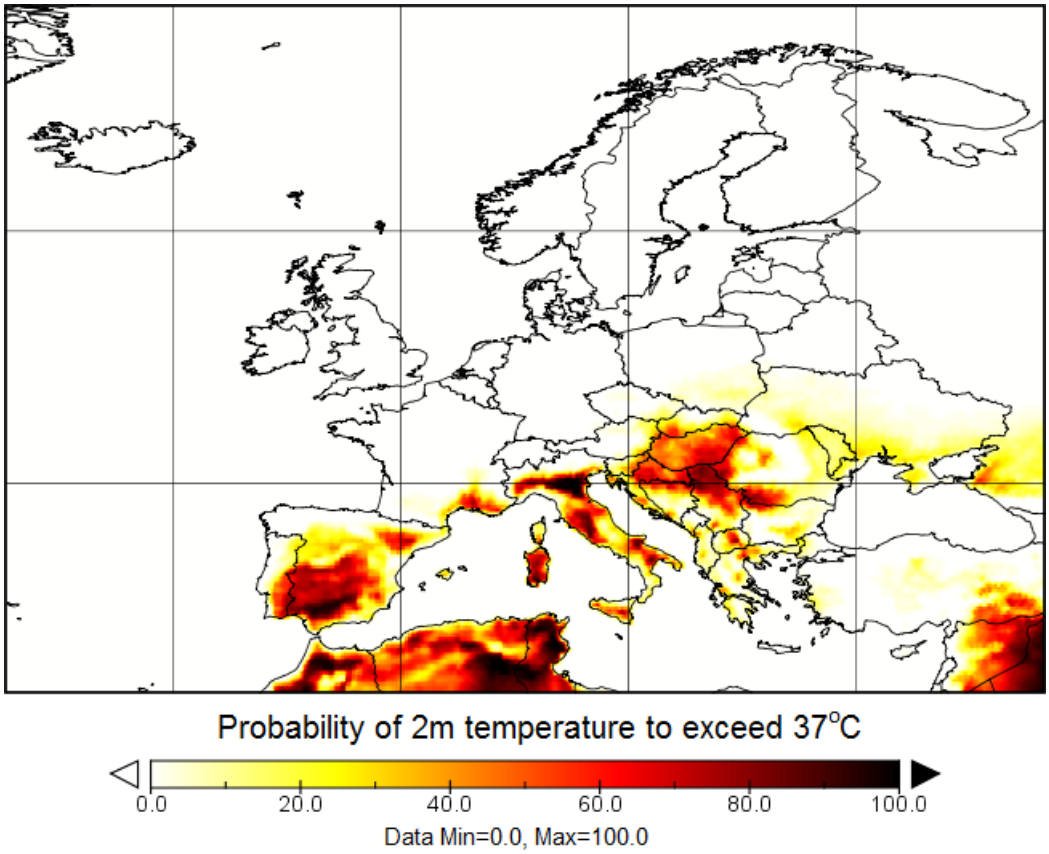

Figure 6: Probabilistic forecast for extreme high temperature, forecast valid time 4th August 2017, 15 UTC

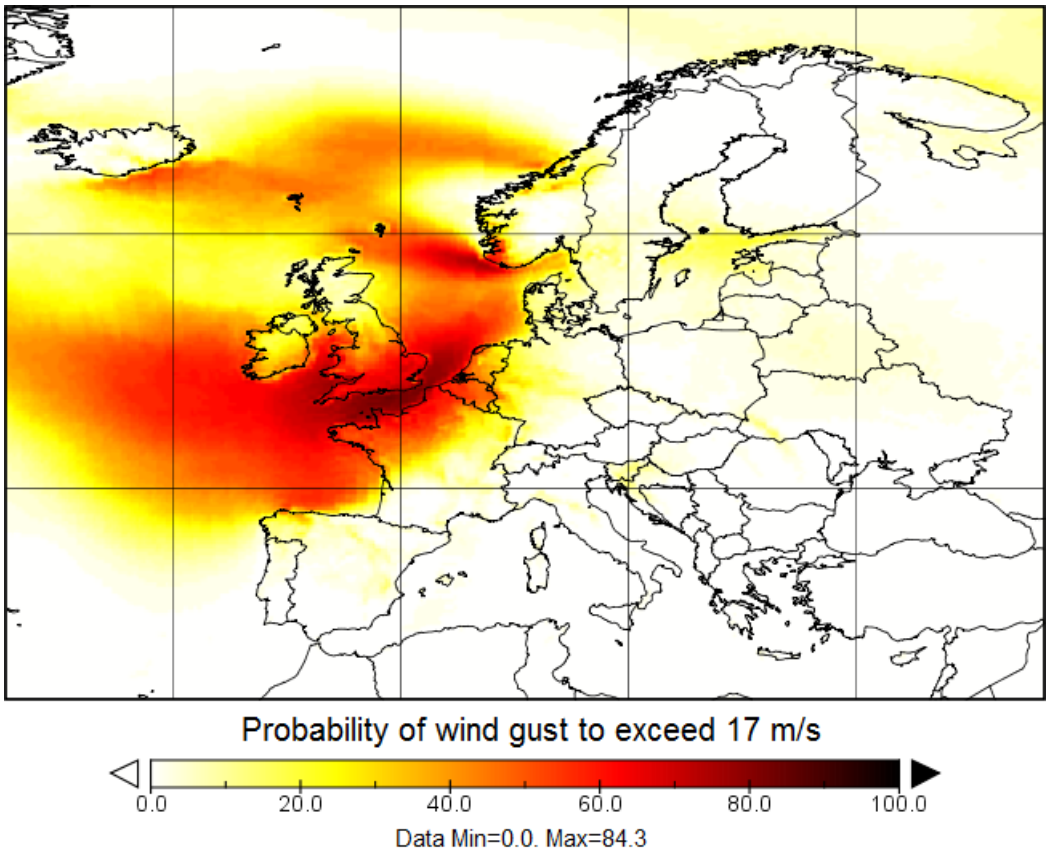

Figure 7: Probabilistic forecast for high wind gusts, forecast valid time 21st October 2017 12 UTC

CONCLUSION

In this paper we examined the effect of calibration for ECMWF ENS temperature and wind gust forecasts. The Gaussian distribution was used for temperature and a box-cox t-distribution for wind forecasts. The calibration was tested as an operational part of forecast post-processing. Therefore, calibration coefficients were calculated by using the past 30 day's forecast and observations pairs instead of huge amount of historical data which would be more time consuming. The calibration coefficient were updated once a week, and separate coefficients were estimated for each lead time independently.

The quality of these calibrated forecasts, compared to raw forecasts, was estimated using rank histogram and Brier Skill Score (BSS). The verification results showed that tested calibration methods improved the overall ensemble forecasts (whole range of values) and also the probabilistic forecast based on some higher threshold. The probabilistic temperature forecasts were better after calibration over whole lead time scale, but the probabilistic wind gust forecasts just up to 240 hours. This result is in line with the previous studies (e.g. Hagedorn, 2008) which have shown that with longer lead times the historical training period should be longer than 30 days. This seemed to be more common for variables whose distribution of the forecast error does not follow the normal distribution.

The purpose of these developed weather forecast products is to provide information of the upcoming extreme weather events early enough, so that necessary preparedness and mitigation actions can be made by the authorized experts. When providing calibrated ensemble forecasts instead of raw ones, we can provide more accurate probabilistic weather forecasts to the end users.

Since the verification results of calibration were promising for both temperature and wind forecasts, next steps would definitely be to test the calibration of the other parameters. The method that will be investigated for precipitation is the zero-adjusted gamma distribution that has been tested for GLAMEPS model.

ACKNOWLEDGMENTS

This work was partially founded by the European Union through the I-REACT project (H2020-DRS-1-2015), grant agreement No.700256.

The calibration methodology has been developed in co-operation of HIRLAM-C (Hirlam.org).

REFERENCES

Brier G. W. (1950) Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* 78:1–3.

Glahn, H. R. and Lowry D. A. (1972) The use of model output statistics (MOS) in objective weather forecasting, *Journal of Applied Meteorology*, 11, 1203-1211.

Gneiting T, Raftery A. E., Westveld A. and Goldman T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, 133, 1098–1118.

Hagedorn R., Hamill T. M. and Whitaker J. S. (2008) Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures. *Mon. Wea. Rev.*, 136, 2608–2619.

Hansen J.A. and Emanuel K. A. (2003) Forecast 4D-Var: Exploiting Model Output Statistics, *Q. J. R. Meteorol. Soc.* 129, 1255–1267.

Iversen, T., Deckmyn A., Santos C., Sattler K. A. I., Bremnes J. B., Feddersen H. and Frogner I.-L. (2011) Evaluation of 'GLAMEPS'—A proposed multimodel EPS for short range forecasting. *Tellus*, 63A, 513–530.

Lugt D. (2013) Improving GLAMEPS wind speed forecasts by statistical postprocessing. *Available:* http://www.knmi.nl/bibliotheek/knmipubIR/IR2013-03.pdf

Molteni F., Buizza R., Palmer T. N. and Petroliagis T. (1996) The ECMWF Ensemble Prediction System: Methodology and validation, *Q. J. R. Meteorol. Soc.* 122, 73-119.

Palmer, T. N. (2000) Predicting uncertainty in forecasts of weather and climate, *Rep. Prog. Phys.*, 63, 71-116.

Rigby R. A. and Stasinopoulos D. M. (2006) "Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis." *Statistical Modelling*, 6, 209–229.

Rosgaard M. H., Nielsen H. A., Nielsen T. S. and Hahmann A. N. (2016) Probing NWP model deficiencies by statistical postprocessing, Quarterly Journal of the Royal Meteorological Society *Q. J. R. Meteorol. Soc.* 142: 1017–1028.

Wilks, D. S. (2006) Statistical Methods in the Atmospheric Sciences. *3rd ed. Academic Press*, pp.

Wilks, D. S. (2009) Extending logistic regression to provide full probability-distribution MOS forecasts. *Meteorological Applications* 16: 361-368.

Wilks D. S. (2015) Multivariate ensemble Model Output Statistics using empirical copulas, *Q. J. R. Meteorol. Soc.* 141: 945–952.

Willet M. R., Bechtold P., Williamson D. L., Pecht J. C., Milton S. F. and Woolnough S. J. (2009) Modelling suppressed and active convection: Comparisons between three global atmospheric models, *Q. J. R. Meteorol. Soc*., 134, 636, Oct 2008 Part A, 1881–1896.