https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

Statistical Postprocessing of Ensemble Forecasts for Severe Weather at Deutscher Wetterdienst

Reinhold Hess
Deutscher Wetterdienst
Correspondence: Reinhold Hess (reinhold.hess@dwd.de)

Abstract. Forecasts of the ensemble systems COSMO-D2-EPS and ECMWF-ENS are statistically optimised and calibrated by Ensemble-MOS with a focus on severe weather in order to support warning decision management at Deutscher Wetterdienst (DWD). Ensemble mean and spread are used as predictors for linear and logistic multiple regressions to correct for conditional biases. The predictands are derived from synoptic observations and include temperature, precipitation amounts, wind gusts and many more, and are statistically estimated in a comprehensive model output statistics (MOS) approach.

This paper gives an overview of DWD's postprocessing system called Ensemble-MOS together with its motivation and the design consequences for probabilistic forecasts of extreme events based on ensemble data. Long time series and collections of stations are used for significant training data that capture sufficient number of cases with observed events, as required for robust statistical modelling. Logistic regression is applied for threshold probabilities and details of its implementation including the selection of predictors with testing for significance are presented.

For probabilities of severe wind gusts global logistic parameterisations are developed that depend on local estimations of wind speed. In this way robust probability forecasts for extreme events are obtained while local characteristics are preserved.

Caveats of Ensemble-MOS, such as model changes and requirements for consistency, are addressed that are known from DWD's operational MOS systems.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

1 Introduction

Ensemble forecasting rose with the understanding of the limited predictability of weather. This limitation is caused by sparse and imperfect observations, approximating numerical data assimilation and modelling and by the chaotic physical nature of the atmosphere. The basic idea of ensemble forecasting is to vary observations, initial and boundary conditions, and physical parameterisations within their assumed scale of uncertainty, and rerun the forecast model with these changes.

The obtained ensemble of forecasts expresses the distribution of possible weather scenarios to be expected. Probabilistic forecasts can be derived from the ensemble like forecast errors, probabilities for special weather events, quantiles of the distribution or even estimations of the full distribution.

In a perfect ensemble system the estimated forecast errors meet the observed errors of the ensemble mean (Wilks, 2011, e.g.). Typically, an optimal spread-skill relationship close to 1 and its involved forecast reliability is much easier obtained for high atmospheric variables, as e.g. 500 h Pa geopotential height, than for screen level variables like 2 m temperature, 10 m wind speed or precipitation (Buizza et al., 2005; Buizza, 2018, e.g.), see also Sect. 2.2.

In order to make best use of the probabilistic information contained in the ensembles, e.g. by relating probabilities for harmful weather events with ecological value in cost-loss evaluations (Wilks, 2001, e.g.), the ensemble forecasts should be calibrated to observed relative frequencies as motivated in Buizza (2018). This calibration should be done in respect to maximise forecast sharpness (Gneiting et al., 2007, e.g.) in order to provide maximal information (the climate mean is calibrated too, however it has no sharpness and is useless as a forecast).

Statistical calibration of ensembles is an additional application for postprocessing systems to statistical optimisation and interpretation of deterministic forecasts. Nevertheless, optimisation is still an issue for ensemble forecasts, as in general the systematic errors of the underlying numerical model are retained (only the random errors are reduced by rerunning the model).

Due to its utmost importance in probabilistic forecasting an abundance of postprocessing methods for various ensemble systems and also for multimodel ensembles exist. In general they differ in purpose of application, they are multivariate or specific to a certain forecast element and they use various approaches of statistical modelling. Length of training data generally depends on the statistical method and the purpose of application. However the availability of data is also often a serious limitation. Some systems perform individual training for different landmarks in order to account for local characteristics, whilst others apply the same statistical model to collections of stations or grid points. Global modelling improves statistical sampling to the costs of orographic and climatologic disparities.

Classical MOS systems tend to underestimate forecasts errors if corrections are applied for each ensemble member individually. In order to maintain forecast variability, Vannitsem (2009) suggests considering observation errors. Gneiting et al. (2005) proposes an ensemble postprocessing method named EMOS, which relies on Gaussian distributions whose mean is a weighted average of the ensemble members and the variance is a linear function of the ensemble spread. Weights and coefficients of the univariate method are trained by minimising the continuous ranked probability score (CRPS). In Bayesian model averaging (BMA) (Raftery et al., 2005; Möller et al., 2013, e.g.) distributions of already bias corrected forecasts are combined as weighted averages using kernel functions.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

Many special systems for individual forecast elements exist, only some are mentioned here. For 24-hourly precipitation Hamill (2012) presents a multimodel ensemble postprocessing based on extended logistic regression and eight years of training data. Hamill et al. (2017) describe a method to blend high-resolution multimodel ensembles by quantile mapping with short training periods of about two months for 6 and 12-hourly precipitations. Systems specialising in wind speed also exist, e.g. Sloughter et al. (2013) uses BMA in combination with Gamma distributions. Fewer methods focus on extreme events of precipitation and wind gusts that are essential for weather warnings. Friederichs et al. (2018) use the tails of generalised extreme-value distributions to estimate conditional probabilities of extreme events. An overview of conventional univariate preprocessing approaches is given in Wilks (2018).

Other approaches create a calibrated ensemble, from which arbitrary calibrated statistical products can be derived in a straightforward way, including area related probabilities. Schefzik et al. (2013); Schefzik and Möller (2018) use ensemble coupula coupling (ECC) and Schaake shuffle-based approaches for temperature, precipitation and wind. The use of a calibrated ensembles provides high flexibility in product generation to the constraint that all ensemble data is accessible.

Here we present a MOS approach that has been tailored to postprocessing ensembles for statistically calibrated extreme and rare events. It uses ensemble mean and spread as predictors in order to avoid underestimation of forecast errors on longer time scales. Apart from motivation and conceptual design, verification results and technical details are also provided, e.g. for the application of logistic regression. Ensemble-MOS is operationally applicable with regard to its robustness and computational costs and runs in trial mode in order to support warning management at DWD. Further outline of the paper is as follows: After the introduction the application of Ensemble-MOS for severe weather is motivated in Sect. 2. Thereafter Sect. 3 describes the design, first the deterministic optimisation and interpretation with linear regression, and afterwards probabilistic forecasting and calibration with logistic regression. At the end of that section a global parameterisation for extreme wind gusts is presented. Finally, Sect. 4 provides a summary and conclusions.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics Discussions
Open Access
EGU

2 Probabilistic forecasts for weather warnings

Postprocessing of ensemble forecasts has been set up at DWD in order to support warning management with probabilistic forecasts of potentially harmful weather events within AutoWARN, see Reichert et al. (2015); Reichert (2016, 2017). Altogether 37 different warning elements exist at DWD, including heavy rain and strong wind gusts, both at several levels of intensity, thunderstorm, snowfall, fog, limited visibility, frost and others.

Special requirements for weather warnings are highlighted in Sect. 2.1, whereas the ensemble systems COSMO-D2-EPS and ECMWF-ENS and their need for postprocessing are introduced in Sect. 2.2.

2.1 Special requirements for weather warnings

Automated warning support focuses on severe weather when warnings are essential. As extreme meteorological events are (fortunately) rare, long time series are required to capture a sufficiently large number of occurred events in order to derive statistically significant estimations. For example, strong precipitation events with rain amounts of more than 15 mm per hour are captured only about once a year at each rain gauge within Germany. Extreme events with more than 40 mm and 50 mm rarely appear, nevertheless warnings are essential when they do. With long time series a significant portion of the data consists of calm weather without relevance for warnings. However, it is problematic to restrict or focus data on severe events, since predictors also need to be detected that are *not* correlated to extreme events in order to control frequency bias (FB) and false alarm ratio (FAR) of probabilistic forecasts.

Ensemble postprocessing substantially derives event probabilities. Thresholds are required to define the level of probability at which meteorological warnings are issued. These warning thresholds may be derived from cost-loss scenarios for specific customers or may be tailored to the public depending on categorical scores such as probability of detection (POD) and FAR. Statistical reliability of forecasted probabilities is considered essential for qualified threshold definitions and the control of warning issuance. Also forecasters are supported to provide probability forecasts that match the real probability of occurrence.

As weather warnings are issued for a certain period of time and a specified region, continuity of probabilistic forecasts in time and space is important. It should be accepted, however, that maps of statistical forecasts do not comply with deterministic runs of numerical models, as probabilistic forecasts are smoothed according to forecast uncertainty. For example, there are hardly convective cells in probabilistic forecasts, but rather areas exist where convection might occur with a certain probability within a given time period.

The use of probabilistic forecasts for warnings of severe weather impacts the way the forecasts need to evaluated. Verification scores like root mean square error (RMSE) or CRPS (Hersbach, 2000; Gneiting et al., 2005, e.g.) are highly dominated by the overwhelming majority of cases when no event occurred. Excellent but irrelevant forecasts of calm weather can pretend good verification results, although the few relevant extreme cases might not be forecasted well. Categorical scores like POD and FAR are considered more relevant for rare and extreme cases, along with other more complex scores like Heidke Skill Score (HSS) or equitable threat score (ETS). Also scatter plots reveal outliers and are sensitive to extreme values, see e.g. Fig. 4.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics Discussions

Long time periods are required not only for statistical modelling, but also for evaluation in order to capture enough extreme and rare events for statistically significant results.

At DWD statistically postprocessed forecasts of the ensemble systems COSMO-D2-EPS and ECMWF-ENS and also of the deterministic forecast models ICON and ECMWF-IFS are combined in order to provide a consistent data set and a seamless transition from very short term to medium range forecasts. This combined product provides a single voice basis for the generation of warning proposals, see Reichert et al. (2015).

2.2 Postprocessing of COSMO-D2-EPS and ECMWF-ENS

Currently, ensemble data of COSMO-D2-EPS and ECMWF-ENS are postprocessed with Ensemble-MOS at DWD. With repeated runs of the numerical model basically only random errors are reduced, while systematic model errors remain. These systematic errors are subject to statistical optimisation just as they are in deterministic forecasting.

For probabilistic forecasting, apart from accuracy, also statistical reliability is a subject of postprocessing, however. Error and probability forecasts have to be related to observed error statistics and relative frequencies, respectively. Especially for short lead times the ensemble spread often underestimates the error of the ensemble mean compared to synoptic observations. The following Sect. 2.2.1 and 2.2.2 briefly describe the ensemble models COSMO-D2-EPS and ECMWF-ENS, respectively, along with examples for underdispersive forecasts that are subject to statistical calibration.

2.2.1 COSMO-D2-EPS and upscaled precipitation probabilities

The ensemble system COSMO-D2-EPS of DWD consists of 20 members of the numerical model COSMO-D2. It provides short term weather forecasts for Germany with runs every three hours (i.e. 00 UTC, . . . , 21 UTC) with forecast steps of 1 h up to 27 h ahead (up to 45 h for 03 UTC). COSMO-D2 was upgraded from its predecessor model COSMO-DE on 15 May 2018, together with its ensemble system COSMO-D2-EPS; the upgrade included an increase in horizontal resolution from 2.8 km to 2.2 km with an adapted orography. Detailed descriptions of COSMO-DE and its ensemble system COSMO-DE-EPS are provided in Baldauf et al. (2011) and Gebhardt et al. (2011); Peralta et al. (2012), respectively. For the ensemble systems initial and boundary conditions as well as physical parameterisations are varied according to their assumed levels of uncertainty.

For the postprocessing of COSMO-D2-EPS, eight years of data have been gathered by the time of writing (including data from the predecessor system COSMO-DE-EPS which has been available since 8 December 2010). Thus, a number of model changes and updates are included in the data, impacts on statistical forecasting are addressed later in Sect. 3.4. Each run of Ensemble-MOS starts two hours after the corresponding run of COSMO-D2-EPS to assure that the ensemble system has finished and the data is available.

Using ensemble systems, probabilities of meteorological events can be estimated as the relative frequency of ensemble members that show the event of interest. Evaluation of this frequency can be evaluated grid point by grid point, which results in probabilities corresponding to areas of the resolution of COSMO-D2 of $2.2 \times 2.2\,\mathrm{km}^2$.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

For near surface elements and short lead times COSMO-D2-EPS is often underdispersive and underestimates forecast errors. This is shown for wind gusts in the scatter diagram Fig. 4d (left) with absolute errors of ensemble mean versus ensemble spread (normalised to absolute error). Figure 1 shows a rank histogram for 1-hourly precipitation amounts of COSMO-D2-EPS. Too often the observations have either less or more precipitation than all ensemble members, which also indicates underdispersivity. Probabilities derived from these relative frequencies are overconfident and result in too many probabilities with values 0 and 1.

Because of the high spatial variability of precipitation, upscaled precipitation products are also derived for COSMO-D2-EPS that refer to areas of 10×10 grid points (i.e. $22 \times 22\,\mathrm{km}^2$). A meteorological event (e.g. that the precipitation rate exceeds a certain threshold) is considered to occur within an area, if the event occurs at any one of its grid points. Area probabilities are estimated straightforward as the relative number of ensemble members with the area event, whereby it is not required that the event takes place at exactly the same grid point for all members.

Certainly, also these purely ensemble based estimates are affected by systematic errors of the numerical model COSMO-D2. Hess et al. (2018) observed a bias of -6.2 percentage points for the upscaled precipitation product of COSMO-DE-EPS for the probability that hourly precipitation rate exceeds $0.1\,\mathrm{mm}$. Verification has been done against gauge adjusted radar observations, since suitable area observations are required, rather than point-based measurements from rain gauges.

2.2.2 ECMWF-ENS and study based on TIGGE-data

The ECMWF-ENS is a global ensemble system based on the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasting (ECMWF). It consists of 50 perturbed members plus one control run and is computed twice a day for 00 UTC and 12 UTC up to $15\,\mathrm{d}$ forecast time (and even further with reduced resolution). Preprocessing of Ensemble-MOS at DWD is based on the 00 UTC-run up to $10\,\mathrm{d}$ forecast time in steps of three hours. ECMWF-ENS is interpolated from its genuine spectral resolution to a regular grid with $28\,\mathrm{km}$ (0.25°) mesh size. Data has been gathered in accordance with the availability of COSMO-DE/2-EPS since 8 December 2010.

A previous study with TIGGE-data, see Bougeault and et al. (2009); Swinbank and et al. (2016), has been carried out in order to demonstrate the benefits of Ensemble-MOS for ECMWF-ENS. Training is based on ensemble data from 2002 - 2012 and corresponding observations, whereas statistical forecasting and verification is performed for 2013, see Hess et al. (2015). Because of the availability from TIGGE, only a restricted set of model variables (2 m-temperature, mean wind, cloud coverage and $24\,\mathrm{h}$ precipitation) is used for multiple regression, as described later in Sect. 3.

Results for 2 m-temperature forecasts are shown in Fig. 2, which illustrates essential improvements of postprocessed forecasts of Ensemble-MOS compared to raw ensemble output. The statistical forecast (blue) not only improves the raw ensemble mean (red), it also outperforms the high resolution ECMWF-IFS (which has not been used for postprocessing). Also the statistical estimation of Ensemble-MOS of its own errors (pink), see Sect. 3.1, is more realistic over the first few days than the estimate of the ensemble mean errors by the ensemble spread (yellow). Improvements of ECMWF-ENS with Ensemble-MOS were also obtained for $24\,\mathrm{h}$ precipitation and cloud coverage.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

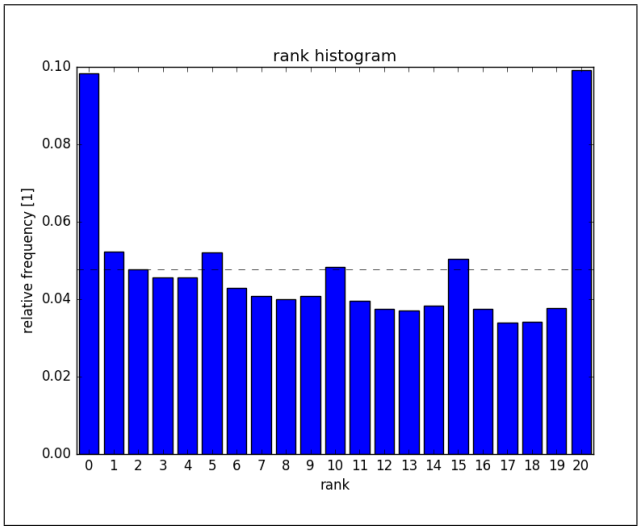


Figure 1. Rank/Talagrand histogram for 1-hourly precipitation amounts of COSMO-DE-EPS (forecast lead time 3 h, data for 18 stations from 2011 to 2017

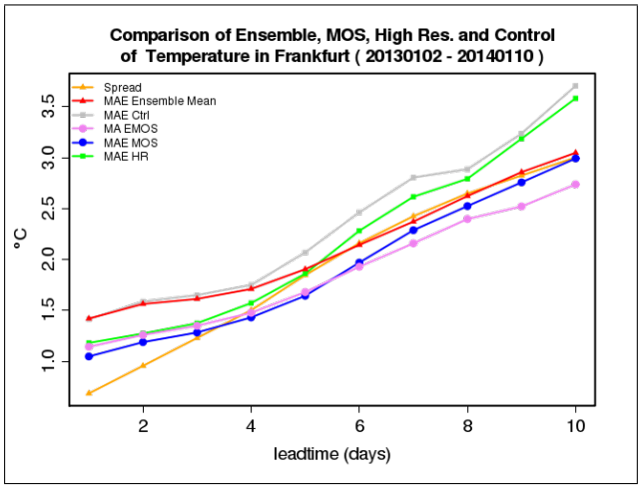


Figure 2. Mean absolute error (MAE) of 2m-temperature forecast and error estimations depending on forecast lead time. Green: MAE of high resolution ECMWF-IFS; red: MAE of mean of ECMWF-ENS; grey: MAE of ECMWF-ENS control run; yellow: spread of ECMWF-ENS (normalised to MAE); blue: MAE of Ensemble-MOS for ECMWF-ENS; pink: Ensemble-MOS forecast of absolute error of Ensemble-MOS

3 Postprocessing by Ensemble-MOS

The Ensemble-MOS of DWD is a model output statistics (MOS) system specialised to process probabilistic information of NWP-ensembles. Currently, it is applied for statistical optimisation and calibration of COSMO-D2-EPS and ECMWF-ENS, but it is expandable to other ensembles in general. The basic concept of Ensemble-MOS is to use ensemble mean, spread, and other ensemble products as predictors in multiple regressions.

The use of ensemble products as predictors prevents difficulties with underdispersed forecasts and underestimated errors on longer forecast lead times as would likely occur if each ensemble member were processed individually. The reason for this underdispersion is that MOS systems tend to converge towards climatology due to the fading accuracy of numerical models and the limited predictability of meteorological events (Vannitsem, 2009, e.g.).

But not only probabilistic products are statistically forecasted by Ensemble-MOS, simultaneously also the corresponding continuous variables are optimised and interpreted, as e.g. precipitation amounts and speed of wind gusts.

Ensemble-MOS is based on a MOS system originally set up for postprocessing deterministic forecasts of the former global numerical model GME of DWD and of the deterministic, high resolution IFS of ECMWF, see Knüpffer (1996). Its concept for optimisation and interpretation using synoptic observations is briefly recapped in Sect. 3.1. Special adaptations for probabilistic

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

forecasting and calibration including logistic regression are described thereafter in Sect. 3.2. For an introduction to MOS in general we refer to Glahn and Lowry (1972); Wilks (2011); Vannitsem et al. (2018).

3.1 Optimisation and interpretation by linear regression

Using the ensemble mean and spread as model predictors allows the MOS approach of Knüpffer (1996) for deterministic NWP-models to be applied to ensembles in a straightforward way. The ensemble mean itself is still a subject for statistical optimisation, since generally the systematic errors of the underlying numerical model are retained in the ensemble systems. Also, interpretation is still required for meteorological elements that are not forecasted numerically (e.g. range of visibility).

The basis of the MOS approach is the use of synoptic observations, from which required predictands for the statistical modelling are derived (for precipitation gauge adjusted radar products can also be used). In principle all meteorological events that are regularly observed are statistically forecasted. This includes temperature, dew point, wind speed and direction, wind gusts, surface pressure, global radiation, visibility, cloud coverage at several height levels and past and present weather. The latter two contain observations of thunderstorm, kind of precipitation, fog and more. Also special predictands for event probabilities exist. They are basically defined as 1 in case the event occurred and 0 if not. For wind gusts and precipitation amounts individual predictands for various reference periods (e.g. 1-hourly, 3-hourly, 6-hourly and longer) are defined, in order to use them as predictors for threshold probabilities.

The probability predictands are smoothed by fitting a monotonically increasing sinusoidal curve between 0 and 1, reaching 0.5 just at the threshold. The smoothing virtually increases the fraction of observed events, since the high number of events below the threshold contribute fractionally to the number of observations above. This stabilises statistical modelling, but affects the reliability of forecasts in combination with subsequent linear regression (as a consequence, logistic regression has been introduced for probability forecasts, see Sect. 3.2). Synoptic observations of more than 300 stations within Germany and its surroundings are used to provide some 150 predictands altogether.

For each predictand the most relevant predictors are selected from a set of independent variables during statistical modelling by multiple stepwise regression. This modelling is performed in general for each predictor, station, season, forecast run, and forecast time individually. For rare events, however, nine zones of similar climatology are defined (e.g. coastal strip, north German plain, various height zones in southern Germany, high mountain areas, etc.) and the stations are clustered together in order to increase the number of observed events and, in turn, the statistical significance of the data sets. Statistical modelling is performed for all stations of a cluster together for those events.

Most independent variables and potential predictors are based on forecasts of the numerical model or ensemble system, which are interpolated to the locations of the observation sites. Besides the model values of the nearest grid point, mean and standard deviation of the 6×6 and 11×11 surrounding grid points are also evaluated and provided as medium and large scale predictors. Additional variables are also derived from the NWP-model fields, as e.g. potential temperature, various atmospheric layer thicknesses, rotation and divergence of wind velocity, dew-point spread and even special parameters, such as

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

convective available potential energy (CAPE) and severe weather threat index (SWEAT). For NWP-ensembles these variables are computed in a straightforward manner from the ensemble means of the required fields.

Further predictors are derived from the latest observations available at the time when the statistical forecast is computed. Generally, the latest observation is an excellent projection for short term forecasts and is therefore added to the set of available predictors. Special care has to be taken to process these predictors for training, however. Only those observations that are available at run time of the forecast can be used. Moreover, these persistence predictors have to be processed in exactly the same way in training and forecasting, which is an issue especially for calibration, in case forecasts are computed for arbitrary locations apart from observation sites, see Sect. 3.2. Due to meteorological persistency the statistical forecasts of the previous forecast step also carry valuable information to the following step and are therefore provided as available independent variables.

Other special predictors exist as well, as e.g. indicators or binary variables that allow jumps to be treated of systematic errors of the numerical models due to model changes, see Sect. 3.4. Altogether, an abundance of more than 300 independent variables is defined, from which up to ten predictors are selected for each predictand during multiple regression.

For continuous variables, such as temperature, precipitation amount, or wind speed, and for their error estimates, linear regression is applied, whereas probability forecasts are modelled by logistic regression, see Sect. 3.2. During stepwise multiple regression, the predictor with the highest correlation with the predictand is first selected from the set of available independent variables. After computing the coefficients of the linear regression, the predictor with the highest correlation with the residuum is identified, and so on. Selection stops if no further predictor can be found with a statistically significant correlation according to a Student's t-test. The level of significance of the test is 0.18 divided by the number of available independent variables. This division is used because of the high number of potential predictors. With a type I error of e.g. 0.05 and a number of 300 available predictors, 15 predictors on average would be selected randomly without containing significant information. The value 0.18 is found to be a good compromise in order to select a meaningful number of predictors and to prevent overfitting in this scenario.

In case of linear regression, the resulting MOS-equation for an estimation $\hat{y}_p$ of a predictand y looks like

$$\hat{y}_p = c_0 + c_1 x_1 + \ldots + c_p x_p \tag{1}$$

with p predictors $x_1, \ldots, x_p$ and $p+1$ coefficients $c_0, \ldots, c_p$. The corresponding error estimation $\hat{y}_p^e$ is defined as the absolute value of the residuum, $\hat{y}_p^e = |\hat{y}_p - y|$. Modelling this error predictand just as well by multiple linear regression provides straightforward error estimations of the current forecast. The absolute value is preferred over the squares of the residuum, since it shows higher correlations to many predefined predictors and a better linear fitting. For Gaussian distributions the absolute error e can be estimated from the standard deviation σ as $e = \sqrt{\frac{2}{\pi}}\sigma \approx 0.8\,\sigma$.

In order to compute statistical forecasts at observation and training sites, the relevant MOS-equations can be evaluated with values of the numerical model and with current observations at these locations, if the latter are used as persistency predictors. For locations apart from the training sites, the equations are linearly interpolated (in case of rare events the corresponding cluster equation is used). Required values of the numerical model for these equations are evaluated for the exact location.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

The required observations need to be interpolated from surrounding sites, however. In this way gridded forecast maps can be obtained as displayed in Fig. 3. For computational efficiency the forecasts are initially computed on a regular grid of 20 km resolution and are downscaled thereafter to 1 km taking into account the various height zones in southern Germany. The details of the downscaling are beyond the scope of the paper.

3.2 Calibration of probabilistic forecasts by logistic regression

The spread of NWP-ensembles is often too small compared to the error of the mean versus observations, especially for near surface weather elements and short lead times (Gebhardt et al., 2011, e.g.) and Figs. 1, 2, 4d. Ensemble-MOS is specialised for calibrated probabilistic forecasts that statistically fit to observations. Logistic multiple regressions are applied for probabilities, such that meteorological events like thunderstorms occur or that continuous variables like precipitation amounts or wind speeds exceed predefined thresholds. Logistic regression (Hosmer et al., 2013, e.g.) is considered state of the art for statistical models of probabilities. Details of the implementation of logistic regression in Ensemble-MOS are presented in the following:

In logistic regression an estimation

$$\hat{y}_p = \frac{1}{1+e^{-(c_0+c_1 x_1+\ldots+c_p x_p)}} \tag{2}$$

of the predictand y based on p predictors $x_1,\ldots,x_p$ and $p+1$ coefficients $c_0,\ldots,c_p$ is determined using a maximum likelihood approach. For this a cost function

$$P(y,c_0,\ldots c_p) = \prod_{i=1}^{n} (\hat{y}_p^i)^{y^i} (1-\hat{y}_p^i)^{1-y^i} \tag{3}$$

is maximised that expresses the probability that y is observed given the the estimation $\hat{y}_p$ via the coefficients $c_0,\ldots,c_p$ (and by now with fixed predictors $x_1,\ldots,x_p$), with n being the time dimension (sample size) and i the time index. It is mathematically equivalent and computationally more efficient to maximise the logarithm

$$\ln(P(y,c_0,\ldots c_p)) = \sum_{i=1}^{n} y^i \ln(\hat{y}_p^i) + (1-y^i)\ln(1-\hat{y}_p^i) \quad . \tag{4}$$

This maximisation is implemented by calling the Routine `G02GBF` of the NAG-Library in FORTRAN 90, see Numerical Algorithms Group (1990). The resulting fit of the estimation $\hat{y}_p$ can be evaluated by the deviance

$$D_p = -2\ln\left(P(y,c_0,\ldots,c_p)\right), \tag{5}$$

which is a measure analogous to the squared sum of residua in linear regression.

The selection of predictors is again performed stepwise. Initially, the coefficient c_0 of the null model $\bar{y} = \frac{1}{1+e^{-c_0}}$ that fits the mean of the predictand is determined and the null deviance $D_0 = -2\ln(P(y,c_0))$ is computed. The coefficient c_0 is often called the intercept. Starting from the null model the predictor that is first selected is that which shows the smallest deviance $D_1 = -2\ln\left(P(y,c_0,c_1)\right)$. The difference $D_1 - D_0$ is χ^2_1-distributed with 1 degree of freedom and is used to check the statistical significance of the predictor. This check replaces the t-test in linear regression and uses the same statistical level.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

If the predictor shows a significant contribution, it is accepted and further predictors are tested based on the new model in the same way. Otherwise the predictor is rejected and the previous fitting $\hat{y}_{p-1}$ is used as the final statistical model.

As a rule of thumb, at least ten events need to be captured within the observation data for each selected predictor (*one in ten rule*) to find stable coefficients. This rule is critical especially for rare events such as extreme wind gusts or heavy precipitation, and sometimes only one predictor is used to fit a global data set, as described in Sect. 3.3 for wind gusts.

Since testing all predictors from the set of about 300 by computing their deviances is very costly, however, the score test (Lagrange multiplier test) is actually applied in Ensemble-MOS. Given a fitted logistic regression with $p-1$ selected predictors, the predictor is chosen next as x_p, that shows the steepest gradient of the log-likelihood function Eq. (4) in an absolute sense when introduced, normalised by its standard deviation σ_{x_p}, i.e.

$$\frac{1}{\sigma_{x_p}} \left| \frac{\partial \ln\big(P(y, c_0, \ldots, c_p)\big)}{\partial c_p} \Big|_{c_p=0} \right| = \left| \sum_{i=1}^{n} \left(y^i - \hat{y}^i_{p-1}\right) \frac{x^i_p}{\sigma_{x_p}} \right| \quad , \tag{6}$$

which results from basic calculus, including the identity $\frac{\partial \hat{y}_p}{\partial c_p} = \hat{y}_p(1-\hat{y}_p)\, x_p$. The right hand side of Eq. (6) is basically the correlation of the current residuum to the new predictor. The score test thus results in the same selection criterion as applied for linear stepwise regression. Once the predictor x_p is selected, the coefficients $c_0, \ldots, c_p$ are updated to maximise Eq. (4).

The latest available observations at the computing time of the statistical forecast have an important impact as persistence predictors for short term forecasts up to about four to six hours. When evaluating the MOS-equations at locations apart from the observations sites, however, special care has to be taken not to affect the statistical reliability of the forecast. At locations other than observation sites, the required values need to be interpolated from surrounding stations. As interpolation generally is a weighted average based on horizontal and vertical distance, it introduces smoothing and, with it, a systematic change in the histogram of observations. If the training is performed for the original observations at the stations and the evaluation is using interpolation, the calibration of the statistical forecasts is deteriorated. As a remedy, Ensemble-MOS uses observations as persistence predictors for training that are interpolated from up to five surrounding stations in exactly the same way as when computing the forecast at arbitrary locations, even if an observation at the correct location was available. Logistic regression in combination with interpolated observations for training allows for well calibrated probabilities on a regular grid, see Figs. 3, 4.

3.3 Statistical parameterisation of wind gust probabilities

Speeds of wind gusts are modelled by linear stepwise regression for each station individually, as described in Sect. 3.1, in order to consider local characteristics. The probabilities that certain thresholds are exceeded, however, are globally parameterised for all stations together in combined logistic regressions that use the statistical forecasts of wind gust speeds as predictors. In this way rare occurrences of extreme events are gathered in order to provide meaningful statistical modelling. Concurrently a certain degree of locality is maintained.

The eight warning thresholds of DWD for wind gusts range from $12.9\,\mathrm{m\,s^{-1}}$ ($25.0\,\mathrm{kn}$, proper wind gusts) up to $38.6\,\mathrm{m\,s^{-1}}$ ($75.0\,\mathrm{kn}$, extreme gales). Statistical modelling is performed for each threshold individually as described in the following:

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

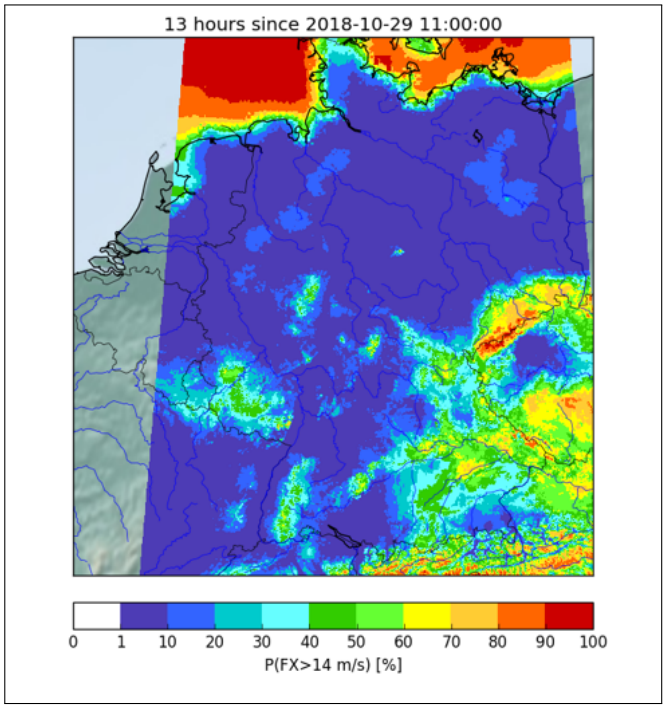


Figure 3. Probabilities for wind gusts higher than $14\,\mathrm{m\,s^{-1}}$ on a regular $1\,\mathrm{km}$ grid over Germany, $13\,\mathrm{h}$ forecast lead time from Ensemble-MOS for COSMO-DE-EPS from 29 October 2018

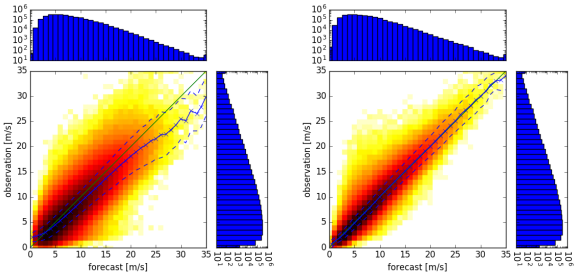
The locally optimised and unbiased speed estimations are excellent predictors for threshold probabilities. Figures 4a (right) and 4b (right) show the statistical fit of wind gust forecasts during a training period of six years of Ensemble-MOS for COSMO-DE-EPS for lead times of $1\,\mathrm{h}$ and $6\,\mathrm{h}$, respectively. The fit is almost bias free, independent of the speed forecast. The raw ensemble means show overforecasting for high wind gusts (same Figs., left). If no overfitting occurs, free forecasts are expected to behave accordingly, which is verified in Fig. 4c (right) for a test period of three months (at least for wind gusts up to about $20\,\mathrm{m\,s^{-1}}$).

According to Eq. (2) a logistic distribution is fitted to the cumulative distribution of observed wind gusts. For wind gust probabilities $p = 1$, the only predictor x_1 is the statistical forecast of wind gust speed, which provides local information and conditional optimisation with its multiple regression approach. Examples for fitting are given in Fig. 5 for threshold t = $13.9\,\mathrm{m\,s^{-1}}$ ($27.0\,\mathrm{kn}$) and forecast lengths of $1\,\mathrm{h}$ and $7\,\mathrm{h}$, respectively. Mean and variance of the fitted logistic distributions are given as $\mu_t = -\frac{c_0}{c_1}$ and $\sigma_t^2 = \frac{\pi^2}{3c_1^2}$, respectively, and are listed for different forecast lead times h in Table 1.

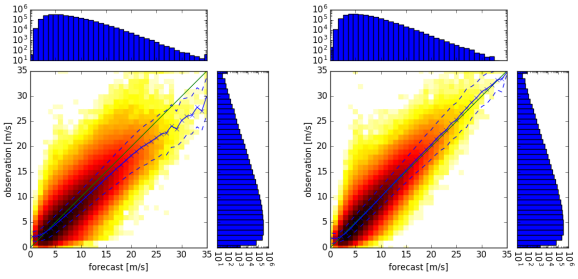
The expectation μ_t is slightly smaller than the threshold t = $13.9\,\mathrm{m\,s^{-1}}$, almost independently of forecast lead time. The reason is that for given statistical forecasts of wind gusts the distribution of observations is almost Gaussian, see Fig. 6, albeit a little left skewed with a small number of very slow wind observations. The standard deviation σ_t increases with forecast lead time reflecting the loss of accuracy of the statistical forecasts. As a consequence, the graph of the cumulative distribution

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

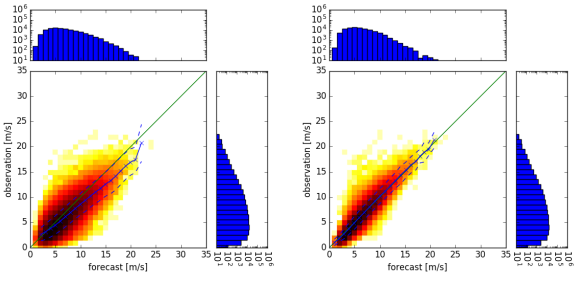
Nonlinear Processes in Geophysics
Discussions
Open Access
EGU



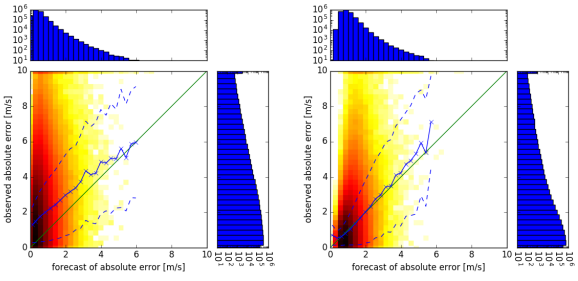
(a) Ensemble means of 3 h forecasts of COSMO-DE-EPS (left) and statistical fits of 1 h forecasts of Ensemble-MOS based on same ensemble data (right), 6 years of data (2011-2016)



(b) As Fig. 4a, but for 8 h and 6 h forecasts of COSMO-DE-EPS and Ensemble-MOS, respectively



(c) As Fig. 4a, but for three months of data (May-July 2016) and statistical forecasts of Ensemble-MOS not using this period for training



(d) 3 h forecasts of absolute errors (estimated as ensemble standard deviations*0.8) of COSMO-DE-EPS versus observed errors of ensemble means (left) and corresponding 1 h error forecasts of Ensemble-MOS versus observed errors of Ensemble-MOS (statistical fit of training period, right). 6 years of data (2011-2016)

Figure 4. Scatter plots of forecasts of COSMO-DE-EPS (left) and statistical optimisation by Ensemble-MOS (right) of 1-hourly wind gusts versus observations, including mean (solid) and mean+/-standard deviation (dashed); number of cases given by histograms

Table 1. Parameters of the fitted logistic distributions as shown in Fig. 5, with forecast lead time h, coefficients of logistic fit c_0 and c_1 and resulting mean μ_t and standard deviation σ_t for threshold t = 13.9 m s^{-1}

h	c_1	c_0	μ_t	σ_t
1	-17.7	1.30	13.7	1.40
4	-13.8	1.01	13.7	1.72
7	-13.3	0.98	13.7	1.86
10	-13.0	0.95	13.7	1.91
16	-12.5	0.92	13.6	1.97

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

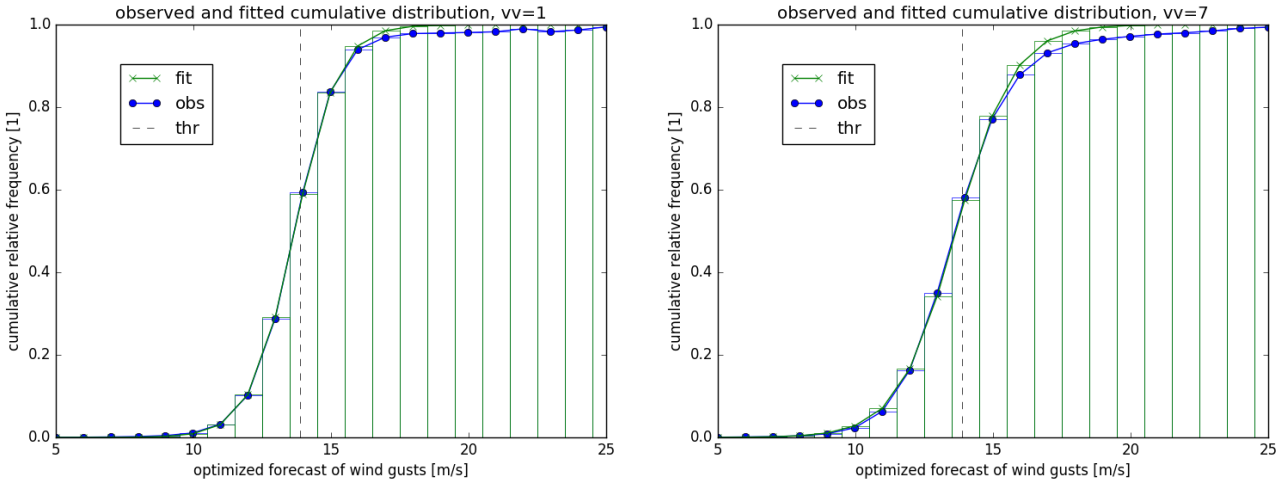


Figure 5. Observed cumulative distributions of wind gusts exceeding threshold 13.9 $\mathrm{m\,s^{-1}}$ (dashed) depending on statistically optimised forecasts of wind gusts (blue) and fit of logistic distribution (green) according to Eq. (2) with $p = 1$ for forecast lead times 1 h (left) and 7 h (right)

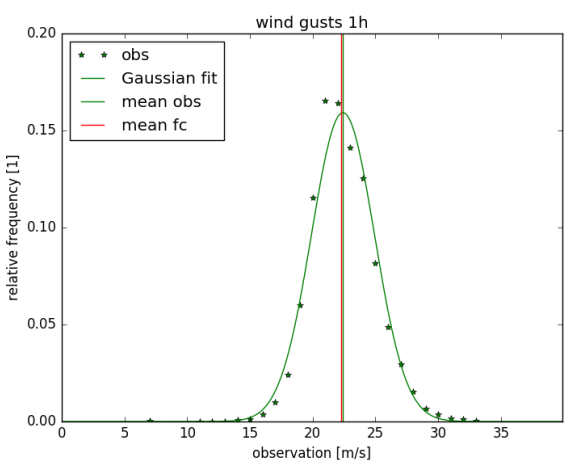


Figure 6. Distribution of wind gust observations from 2011 to 2016 for 178 synoptic stations for cases where statistical forecast (fitted for that period) with 1 h lead time in between 21 $\mathrm{m\,s^{-1}}$ and 25 $\mathrm{m\,s^{-1}}$. Gaussian fit and mean of obs (green), mean of forecasts (red)

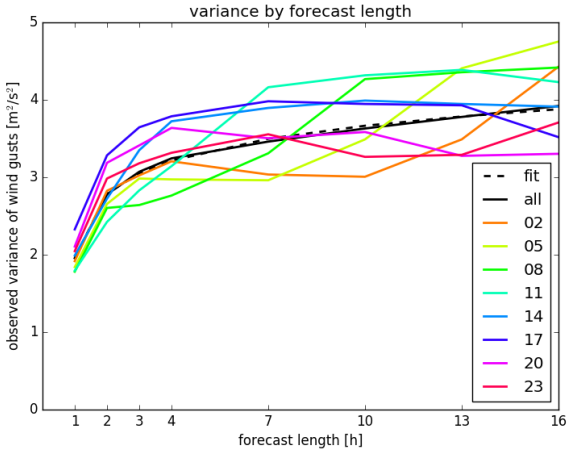


Figure 7. Variances σ_t^2 of logistic distributions fitted to cumulative distributions of observed wind gusts for threshold t = 13.9 $\mathrm{m\,s^{-1}}$ depending on forecast lead time. Individual runs of Ensemble-MOS for COSMO-DE-EPS-MOS starting at 02 UTC, 05 UTC, ..., 23 UTC in colours, mean of all runs in black, fitted parameterisation of variances dashed in black

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics Discussions

function in Fig. 5 is more tilted for 7 h forecast times than for 1 h. Figure 7 shows fitted variances σ_t^2 of the eight individual forecast runs of Ensemble-MOS for COSMO-DE-EPS and their mean depending on lead time. In order to reduce the number of coefficients and to increase consistency and robustness of the forecasts, the variance σ_t^2 is parameterised depending on forecast time h by fitting the function

$$\sigma_t^2(h) = c_t \log(a_t h + b_t) \tag{7}$$

with its parameters a_t, b_t, and c_t for threshold t.

Dependencies of the fitted variances on the time of the day have been found to be weak and are neglected. Therefore, logistic regressions of wind gust probabilities can be expressed by the mean μ_t and parameterisation of σ_t, depending on forecast lead time h, according to Eq. (7) for all individual forecast runs of Ensemble-MOS together.

Even for extreme and very rare gales of 38.6 $\mathrm{m\,s^{-1}}$ more than 130 events are captured using six years of training data and all stations and forecast runs together, which allows for robust statistical estimation by logistic regression. Training for these extreme events is based mainly on coastal and mountain stations, but the parameterisations are applied to more wind sheltered locations as well. Small threshold probabilities will result for those locations in general. However, meaningful estimations will be generated once the statistical forecast of local wind speed rises when induced by the numerical model.

3.4 Specific issues and caveats of MOS

Ensemble-MOS effectively improves, optimises and calibrates ensemble forecasts with the use of synoptic observations. As with any other statistical method it is vulnerable to systematic changes in input data, since it assumes that errors and characteristics of the past persist in the future.

One important part of the input are observations that sometimes change due to changes in the measurement instruments used. It is recommended to use quality checked observations in order to avoid the use of defective values for training. Especially observation sites that are automised need to be checked. Furthermore, numerical models change with new versions and updates, which can affect statistical postprocessing, as further discussed in Sect. 3.4.1.

Although statistical forecasts generally improve the model output in terms of verification against observations, the results are not always consistent in time, space and between forecast variables (as e.g. between temperature and dew point), if they are individually optimised. This issue is addressed in Sect. 3.4.2.

3.4.1 Model changes

Statistical methods like Ensemble-MOS detect systematic errors and deficiencies of NWP-models during a past training period in order to improve topical operational forecasts. Implicitly it is assumed that systematic characteristics of the NWP-model persist. Note that multiple regressions not only correct for model bias, but also for conditional biases that depend on other meteorological variables and are therefore more vulnerable than simple regressions. Systematic changes in NWP-models can affect statistical forecasts, even if the NWP-forecasts are objectively improved as confirmed by verification. Given unbiased

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics Discussions

statistical modelling, any systematic change in NWP-model predictors will reflect in biases in the statistical forecasts. The resulting biases depend on the magnitudes of the changes of the predictors and their weights in the MOS-equations.

One remedy for jumps in input data is the use of indicator (binary) predictors. These predictors are related to the date of the change of the NWP-model and are defined as 1 before and 0 after. When they are selected during multiple stepwise regression, they account for sudden changes and can prevent the introduction of unconditional biases in the statistical forecasts. Conditional biases depending on other forecast variables, however, are not corrected in this way.

In order to process extreme and very rare events for weather warnings, long time series of seven years of data for COSMO-D2-EPS have been gathered at the time of writing, while a number of model changes have taken place. A significant model upgrade from COSMO-DE to COSMO-D2, including an increase of horizontal resolution from 2.8 km to 2.2 km and an update of orography, took place in May 2018. Since reforecasting of COSMO-D2-EPS for more than one year was technically not possible, the existing COSMO-DE-EPS database was used further and extended with reforecasts of COSMO-D2-EPS of the year before operational introduction. However, statistical experiments using these reforecasts of COSMO-D2-EPS (and the use of binary predictors, see above) revealed only small improvements compared to training with data of COSMO-DE-EPS only. For rare events longer time series are considered more important than the use of unaltered model versions.

3.4.2 Forecast consistency

The statistical modelling of Ensemble-MOS is carried out for each forecast variable, forecast lead time and location independently and individual MOS-equations are derived. For rare meteorological events clusters of stations are grouped together that are similar in climatology in order to derive individual cluster equations. This local and individual fitting results in optimal forecasts for the specific time, location and variable as measured e.g. by RMSE compared to observations. However, it does not guarantee that forecast fields are consistent in space, time or between variables.

In forecast time, spurious jumps can appear and variables with different reference periods usually do not match (e.g. the sum of twelve successive one-hourly precipitation amounts would not equal the corresponding 12-hourly amount, if the latter is modelled as an individual predictand). Forecasts of temperature cannot be guaranteed to exceed those of dew point. Forecast maps show high variability from station to station and unwanted anomalies in case of cluster equations (e.g. cluster edges turn up and it may appear that there are higher wind gusts in a valley than on a mountain nearby, in cases where the locations are arranged in different clusters). For consistency in time and space the situation can be improved by using combined equations for several lead times and for larger clusters or by elaborate subsequent smoothing. However, forecast quality for a given space and time will be degraded as a consequence. For consistency between all forecast variables multivariate regressions are required that model the relevant predictands simultaneously.

From the point of view of probabilistic forecasting, however, statistical forecasts are random variables with statistical distributions, although commonly only their expectations are considered *the* statistical forecast. In case forecast consistency is violated from a deterministic point of view, this is not the case if statistical errors are taken into account. The statistical forecasts remain valid as long as the probability distributions of the variables just overlap. As this is a mathematical point of view,

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020


Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

the question remains, how to communicate this nature of probabilistic forecasts to the public or traditional meteorologists in terms of useful and accepted products.

4 Conclusions

This paper describes the Ensemble-MOS system of DWD used to postprocess the ensemble systems COSMO-D2-EPS and ECMWF-ENS with respect to severe weather to support warning management. MOS in general is a mature and sound method and in combination with logistic regression it can provide optimised and calibrated statistical forecasts. Multiple stepwise regression allows for reducing conditional biases depending on the meteorological condition that is expressed by the selected predictors.

The setup of Ensemble-MOS based on ensemble mean and spread as predictors is computationally efficient and simplifies forecasting of calibrated event probabilities and error estimates on longer forecast lead times. The ensemble spread is less often detected as an important predictor in multiple regression as might be expected, however. One reason is that the spread actually carries less information about forecast accuracy as was originally intended. It is often too small and too steady to account for current forecast errors. Another reason is that some forecast variables correlate with their own forecast errors (e.g. precipitation and wind gusts). If the ensemble spread does not provide significant independent information, it is not selected additionally to the ensemble mean during stepwise regression.

Currently, only ensemble mean and spread are provided as predictors for Ensemble-MOS. The implementation of various ensemble quantiles as additional predictors is technically straightforward, however, and could improve the exploitation of the probabilistic information of the ensemble.

Forecasts of wind gust speed are excellent predictors for logistic modelling of threshold probabilities. The same approach could be advantageous for probabilities of heavy precipitation as well, where estimated precipitation amounts would be used as predictors.

An important further step in probabilistic forecasting is the estimation of complete (calibrated) distributions of forecast variables rather than forecasting only discrete threshold probabilities. For wind gusts with Gaussian conditional errors as shown in Fig. 6 this seems possible but certainly requires additional research.

With its inherent linearity (also in the case of logistic regressions there are linear combinations of predictors only) MOS has its restrictions in modelling, but supports traceability and robustness, which are important features in operational weather forecasting. Therefore MOS is considered a possible baseline for future statistical approaches based on neural networks and artificial intelligence that allows for general statistical modelling. Many of the statistical problems, however, will remain, such as e.g. finding suitable reactions to changes in the NWP-models, (deterministic) consistency (see Sect. 3.4.2) and the verification of rare events. In all cases training data is considered of utmost importance, including the NWP-model output, as well as quality-checked historic observations.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

Author contributions. Conceptual design of Ensemble-MOS, enhancement of software for probabilistic forecasting of ensembles (including logistic regression), processing of forecasts, verification, and writing was done by the author.

Competing interests. The author declares that he has no conflicts of interest.

Acknowledgements. Thanks to James Paul for improving the use of English.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

References

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T.: Operational convective–scale numerical weather prediction with the COSMO model: description and sensitivities, Mon. Weather Rev., 139, 3887–3905, 2011.

Bougeault, P. and et al.: The THORPEX interactive grand global ensemble (TIGGE), Bulletin of the AMS, 91, 1059–1072, 2009.

Buizza, R.: Ensemble forecasting and the need for calibration, in: Statistical Postprocessing of Ensemble Forecasts, edited by Vannitsem, S., Wilks, D. S., and Messner, J. W., chap. 2, pp. 15–48, Elsevier, Amsterdam, 2018.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M., and Zhu, Y.: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems, Mon. Weather Rev., 133, 1076–1097, 2005.

Friederichs, P., Wahl, S., and Buschow, S.: Postprocessing for Extreme Events, in: Statistical Postprocessing of Ensemble Forecasts, edited by Vannitsem, S., Wilks, D. S., and Messner, J. W., chap. 5, pp. 128–154, Elsevier, Amsterdam, 2018.

Gebhardt, C., Theis, S. E., Paulat, M., and Ben Bouallègue, Z.: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries, Atmos. Res., 100, 168–177, 2011.

Glahn, H. R. and Lowry, D. A.: The use of model output statistics (MOS) in objective weather forecasting, J. Appl. Meteorol., 11, 1203–1211, 1972.

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

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. R. Statist. Soc: B, 69, 243–268, 2007.

Hamill, T.: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the conterminous United States, Mon. Weather Rev., 140, 2232–2252, 2012.

Hamill, T. M., Engle, E., Myrick, D., Peroutka, M., Finan, C., and Scheuerer, M.: The U.S. national blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount, Mon. Weather Rev., 145, 3441–3463, 2017.

Hersbach, H.: Decomposition of the continuous ranked probabilitxy score for ensemble prediction systems, Wea. Forecasting, 15, 559–570, 2000.

Hess, R., Glashoff, J., and Reichert, B. K.: The Ensemble-MOS of Deutscher Wetterdienst, in: EMS Annual Meeting Abstracts, 12, Sofia, 2015.

Hess, R., Kriesche, B., Schaumann, P., Reichert, B. K., and Schmidt, V.: Area precipitation probabilities derived from point forecasts for operational weather and warning service applications, Q. J. R. Meteorol. Soc., 144, 2392–2403, 2018.

Hosmer, D. W., Lemenshow, S., and Sturdivant, R. X.: Applied Logistic Regression, Wiley Series in Probability and Statistics, Wiley, New Jersey, 3rd edn., 2013.

Knüpffer, K.: Methodical and predictability aspects of MOS systems, in: Proceedings of the 13th Conference on Probability and Statistics in the Atmospheric Sciences, pp. 190–197, San Francisco, 1996.

Möller, A., Lenkoski, A., and Thorarinsdottir, T. L.: Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas, Q. J. R. Meteorol. Soc., 139, 982–991, 2013.

Numerical Algorithms Group: The NAG Library, Oxford, UK, www.nag.com, 1990.

Peralta, C., Ben Bouallègue, Z., Theis, S. E., Gebhardt, C., and Buchhold, M.: Accounting for initial condition uncertainties in COSMO-DE-EPS, J. Geophys. Res., 117, 2012.

https://doi.org/10.5194/npg-2019-64
Preprint. Discussion started: 20 January 2020

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, 2005.

Reichert, B. K.: The operational warning decision support system AutoWARN at DWD, in: 27th Meeting of the European Working Group on Operational Meteorological Workstation Systems (EGOWS), Helsinki, 2016.

Reichert, B. K.: Forecasting and Nowcasting Severe Weather Using the Operational Warning Decision Support System AutoWARN at DWD, in: 9th Europ. Conf. on Severe Storms ECSS, Pula, Croatia, 2017.

Reichert, B. K., Glashoff, J., Hess, R., Hirsch, T., James, P., Lenhart, C., Paller, J., Primo, C., Raatz, W., Schleinzer, T., and Schröder, G.: The decision support system AutoWARN for the weather warning service at DWD, in: EMS Annual Meeting Abstracts, 12, Sofia, 2015.

Schefzik, R. and Möller, A.: Ensemble postprocesing methods incorporating dependence structures, in: Statistical Postprocessing of Ensemble Forecasts, edited by Vannitsem, S., Wilks, D. S., and Messner, J. W., chap. 4, pp. 91–125, Elsevier, Amsterdam, 2018.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T.: Uncertainty quantification in complex simulation models using ensemble copula coupling, Statist. Sci., 28, 616–640, 2013.

Sloughter, J. M., Gneiting, T., and Raftery, A. E.: Probabilistic wind vector forecasting using ensembles and Bayesian model averaging, Mon. Weather Rev., 141, 2107–2119, 2013.

Swinbank, R. and et al.: The TIGGE project and its achievements, Bulletin of the AMS, 97, 49–67, 2016.

Vannitsem, S.: A unified linear Model Output Statistics scheme for both deterministic and ensemble forecasts, Q. J. R. Meteorol. Soc., 135, 1801–1815, 2009.

Vannitsem, S., Wilks, D. S., and Messner, J. W., eds.: Statistical Postprocessing of Ensemble Forecasts, Elsevier, Amsterdam, Oxford, Cambridge, 2018.

Wilks, D. S.: A skill score based on economic value for probability forecasts, Meteorol. Appl., 8, 209–219, 2001.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, San Diego, 3rd edn., 2011.

Wilks, D. S.: Univariate ensemble postprocessing, in: Statistical Postprocessing of Ensemble Forecasts, edited by Vannitsem, S., Wilks, D. S., and Messner, J. W., chap. 3, pp. 49–89, Elsevier, Amsterdam, 2018.