TECHNICAL MEMORANDUM

719

Calibration of Medium-Range Weather Forecasts

Tilmann Gneiting[1]

Forecast Department

[1]Heidelberg Institute for Theoretical Studies and
Karlsruhe Institute of Technology

March 18, 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

Abstract

Statistical postprocessing techniques serve to improve the quality of numerical weather forecasts, as they seek to generate calibrated and sharp predictive distributions of future weather quantities. This document reviews the state of the art in statistical postprocessing, with focus on potential applications to the European Centre for Medium-Range Weather Forecasts (ECMWF)'s Integrated Forecasting System (IFS). At present, a recommended way to proceed is to apply well established, state of the art postprocessing techniques, such as nonhomogeneous regression or Bayesian model averaging, to each univariate weather quantity separately, with training data usefully augmented by reforecast datasets. Areas requiring further research are identified, in particular the suitable size and efficient use of reforecast datasets, and the generation and evaluation of probabilistic forecasts of combined events and spatio-temporal weather trajectories, thereby addressing spatial, temporal and cross-variable dependence structures.

1 Introduction

Numerical weather prediction Weather forecasting has traditionally been viewed as a deterministic problem, drawing on highly sophisticated numerical models of the physics and chemistry of the atmosphere. Equipped with state of the art computational resources including supercomputers, weather centers have run carefully designed numerical weather prediction (NWP) models to generate single-valued forecasts of future states of the atmosphere.

While such is still the case today, weather prediction has been transformed through the operational implementation of ensemble prediction systems (Palmer 2002; Gneiting and Raftery 2005; Leutbecher and Palmer 2008). An ensemble forecast is a collection of NWP model integrations, using slightly different initial conditions and/or model variants, with the output being a probabilistic forecast, providing an estimate of the uncertainty of the forecast, and ideally being interpretable as a random sample from the predictive distribution of future weather states. From users' perspectives, probabilistic forecasts are necessary for optimal decision making (Krzysztofowicz 2001; Palmer et al. 2005/06; Gneiting 2008), and the added economic value of ensemble forecasts has been demonstrated persuasively (Richardson 2000; 2001; Palmer 2002), including but not limited to the case of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble.

The ECMWF's Integrated Forecasting System viewed as a 52-member NWP ensemble The global medium-range forecasting system run by ECMWF comprises both a single run high-resolution forecast (HRES) and a 51-member ensemble of lower-resolution forecasts (ENS), as described by the ECMWF Directorate (2012).

The HRES works at a horizontal resolution of 16 km and provides the ECMWF's most sophisticated and computationally demanding forecasts of future weather states at prediction horizons up to ten days ahead.

The ENS is an ensemble of lower-resolution forecasts that has been available operationally since December 1992 (Buizza and Palmer 1995; Molteni et al. 1996). Its 51 members operate at 32 km horizontal resolution. One forecast, termed the control (CNT) forecast, is run from the operational ECMWF analysis. Fifty further model runs, termed the perturbed (PTB) members, are run from slightly different initial conditions, using initial perturbations generated by the singular vector technique (Buizza and Palmer 1995). To take uncertainties in the model formulation into account, the perturbed members furthermore use slightly distinct model variants in a stochastic setting (Buizza, Miller and Palmer 1999). Twice a

week, the ENS is extended to provide forecast guidance up to 32 days ahead at 64 km horizontal resolution.

Rodwell (2005/06) found that a combined prediction system based on both HRES and ENS performs significantly better than just ENS in forecasting the probability of precipitation occurrence over Europe. These and similar findings strongly support the point of view recently emphasized by the ECMWF Directorate (2012, p. 11), in that the HRES and ENS are to be considered as an "inseparable pair" that ought,

> whenever possible, be used together to provide the most detailed description of future weather and of the associated uncertainties.

Jointly, the HRES and the ENS comprise the ECWMF's Integrated Forecasting System (IFS). In what follows, we think of the IFS as a 52-member ensemble system, consisting of the HRES member, the CNT member, and 50 exchangeable PTB members, with the latter being statistically indistinguishable.

Statistical postprocessing: Bias correction vs. ensemble calibration As individual NWP model runs and ensemble systems are subject to biases and dispersion errors, their predictions can be improved by statistical techniques (Glahn and Lowry 1972; Gneiting and Raftery 2005).

Traditionally, bias correction techniques have been applied to NWP output using regression methods, in an approach commonly known as model output statistics (MOS; Wilks 2011). Pioneered by Glahn and Lowry (1972) and Klein and Glahn (1974), and further developed and studied by Carter, Dallavalle and Glahn (1989), Wilson and Vallée (2002; 2003), Mass et al. (2008) and Glahn et al. (2009a), among others, MOS approaches transform a single-valued NWP output into another single-valued forecast, thereby correcting systematic errors in the central tendencies of model output, either at observation sites or directly on the model grid. Essentially, MOS approaches exploit statistical patterns in the relationships between observations and the outputs of the specific NWP model version at hand. A related, but less successful technique is the perfect prog (PP) method that identifies statistical patterns in the relationships between distinct meteorological variables in observational records and applies them to the output of the NWP model version at hand (Wilks 2011).

Despite being widely applied and being powerful, bias correction techniques have natural limitations, as they aim at single-valued forecasts only.[1] When considering an NWP ensemble, bias correction of individual ensemble members is useful in adjusting errors in central tendencies, but does not readily provide a fully fledged ensemble calibration (Vannitsem 2009; Vannitsem and Hagedorn 2011; Cui et al. 2012). For example, if each member is bias corrected by using a conditional mean, the postprocessed ensemble is bound to be underdispersed. For skewed variables, such as precipitation amounts or wind speeds, Flowerdew (2012, p. 7) summarizes commonly encountered problems by noting that

> [t]he statistical features of precipitation argue against the use of simple bias corrections. Overall additive terms would affect all forecasts of zero precipitation. Corrections which are multiplicative or only affect nonzero forecasts do not help to adjust the frequency of zero precipitation forecasts.

Indeed, if we were to bias correct a single-valued NWP output for a precipitation amount, for which a highly skewed predictive distribution is adequate, would we want to transform into the mean, the

[1] This comment does not apply when a probability forecast of a binary event is sought, such as a precipitation amount or a temperature being above a certain threshold. In these settings, MOS approaches generate probability forecasts, as opposed to single-valued forecasts, with logistic regression being a particularly powerful technique (Applequist et al. 2002).

median, or perhaps another quantile? Following up on the work of Murphy and Daan (1985) in the meteorological literature, Gneiting (2011) analyses and describes the pitfalls and challenges that arise in this type of setting, in that the forecasting task and scoring functions used need to be carefully matched.

As noted, a preferable approach to statistical postprocessing, which avoids these pitfalls and allows for optimal decision making (Palmer et al. 2005/06; Gneiting 2008), is to generate probabilistic forecasts that take the form of probability distributions over future weather quantities and events. In this context, statistical postprocessing techniques aim at generating calibrated and sharp probabilistic forecasts from ensemble output, including the case of a single NWP model integration, with the latter being interpretable as an NWP ensemble with a single member only. Following Murphy and Winkler (1987) and Gneiting, Balabdaoui and Raftery (2007), the goal in probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration. Calibration is a multi-faceted, joint property of the forecasts and the observations; essentially, the forecasts are calibrated if the observations can be reasonably interpreted as drawn from the predictive distributions. Sharpness refers to the concentration of the predictive distributions, and thus is a property of the forecasts only. Evidently, the more concentrated, the more useful the probabilistic forecast, subject to it being calibrated.

Wilks and Hamill (2007), Bröcker and Smith (2008), Schmeits and Kok (2010) and Ruiz and Saulo (2012) review and compare methods for ensemble calibration. State of the art techniques include the Bayesian model averaging (BMA) approach developed by Raftery et al. (2005) and the non-homogeneous regression (NR), or ensemble model output statistics (EMOS), technique of Gneiting et al. (2005), to be described in detail below.

Similar needs for bias correction and calibration arise in applications using NWP output, such as hydrologic and renewable energy forecasts. For example, operational streamflow and water level forecasting systems, including but not limited to the European Flood Awareness System (EFAS) developed by Thielen et al. (2009), are increasingly using precipipation forecasts from NWP ensembles to drive hydrological models and predictions, as reviewed by Cloke and Pappenberger (2009) and in Appendix A in this document. Similarly, NWP ensemble forecasts of wind speed are driving state of the art systems for predicting wind energy power output (Pinson and Madsen 2009; Taylor, McSharry and Buizza 2009; Traiteur et al. 2012; Pinson 2013). In this context, a natural question is whether statistical postprocessing efforts ought to apply to the weather input, or to the application output, or to both. While further research and experimentation will be required to address these questions in full, a natural expectation is that postprocessing efforts undertaken at both levels optimizes the predictive performance, and that if postprocessing can only be done once, then it is most effectively implemented at the application output level.

Current postprocessing efforts for the ECMWF's IFS In the specific case of the ECMWF's IFS, there is a stark contrast between postprocessing efforts implemented by Member and Co-operating States at the national level, and postprocessing techniques studied directly at ECMWF.

Operational implementations in Member and Co-operating States. Table 1 reviews operational postprocessing efforts at the national level based on the most recent available information in the Annual Report on Application and Verification of ECMWF Products in Member and Co-operating States (known as Green Book).[2] While a majority of Member States has implemented postprocessing procedures, these are typically restricted to traditional bias correction of the HRES output, and to key variables such as 2m temperature, precipitation or 10m wind, using some form of model output statistics (MOS) technique, frequently implemented via Kalman filter (KF) algorithms, as described by Crochet (2004), and in some

[2]The Green Book reports are available online at www.ecmwf.int/products/greenbook.

Table 1: Member states (above separator) and co-operating states (below separator) operational statistical post-processing efforts for the IFS, as described in the Annual Report on Application and Verification of ECMWF Products in Member and Co-operating States (known as Green Book). Typically, bias correction is applied to key variables, such as surface temperature and precipitation, and based on the HRES run only, using model output statistics (MOS), the Kalman filter (KF), which is a special type of MOS approach, or the perfect prog (PP) technique, respectively. The sporadic ENS calibration efforts use nonhomogeneous regression (NR), rank based (RB) or CDF matching (CM) methods, respectively. The final column identifies the latest available report, from which the information quoted has been extracted.

Country	Description	Source
Austria	bias correction (MOS, PP)	2012
	ensemble calibration (NR)	2012
Belgium	bias correction (MOS, KF)	2012
Finland	bias correction (KF)	2012
France	bias correction (MOS, KF)	2012
	ensemble calibration (RB)	2012
Germany	bias correction (MOS, PP)	2012
Greece	bias correction (KF)	2012
Iceland	bias correction (KF)	2012
Italy	bias correction (PP)	2012
Netherlands	binary probability forecasts	2010
Norway	bias correction (KF)	2012
Portugal	bias correction (MOS, KF)	2011
Slovenia	bias correction (KF)	2009
Spain	bias correction	2012
	binary probability forecasts	2012
Sweden	bias correction (KF)	2012
	binary probability forecasts	2012
Switzerland	bias correction (KF)	2012
Turkey	bias correction (KF)	2012
United Kingdom	bias correction (KF)	2012
Hungary	ensemble calibration (CM)	2012
Israel	bias correction	2012
Romania	bias correction (MOS, PP)	2012

cases the perfect prog (PP) method. A few member countries have also considered ensemble calibration techniques, including Austria (Kann et al. 2009; Kann, Haiden and Wittmann 2011), France and Hungary (Ihász et al. 2010),[3] and some countries have implemented postprocessing efforts geared at binary probability forecasts, with Schmeits et al. (2008) describing one such application to lightning.

Research studies at ECMWF. At ECMWF, research and development efforts have focused on ensemble calibration, with Hagedorn (2008), Hagedorn, Hamill and Whitaker (2008), Hamill, Hagedorn and Whitaker (2008), Hagedorn (2010), Hagedorn et al. (2012) and Pinson (2012) having studied the application of ensemble calibration techniques for the ECMWF IFS with notable success.[4] For example, Hagedorn (2008) and Hagedorn, Hamill and Whitaker (2008) reported gains in lead time of two to four days for predictions of surface temperature when the nonhomogeneous regression technique of Gneiting et al. (2005) is applied to the ECMWF's ENS, with the improvement generally being stronger at locations where the original forecast skill is low, such as in regions with complex terrain and along coastlines. Hamill, Hagedorn and Whitaker (2008, p. 2630) observed that calibration techniques "dramatically improved" probability of precipitation forecasts from the ECMWF ENS, achieving gains of two days and more in forecast lead. ECMWF has also carried out development work on products to provide forecasters with guidance on the risk of potential extreme events, including the extreme forecast index (EFI; Lalaurette 2003; Petroliagis and Pinson 2012; Zsoter, Pappenberger and Richardson 2014) and the Probability of RETurn (PRET; Prates and Buizza 2011), which build on and post-process reforecast data to accunt for differences between the model and observed climate distributions.

Operational implementations at the ECMWF. Nevertheless, ECMWF has not implemented statistical postprocessing at the medium-range operationally, with the exceptions of the EFI and products for the aforementioned EFAS (Bogner and Pappenberger 2011). In contrast, the ECMWF forecast systems at seasonal and sub-seasonal scales depend considerably on statistical postprocessing (Molteni et al. 2011; Vitart 2013). For example, the new ECMWF seasonal forecast system (System 4) uses hindcasts to generate both bias and variance corrections, as described in Sections 2.4 and 3.2 of Molteni et al. (2011).

2 The state of the art in statistical postprocessing

We turn to a review of the state of the art in ensemble calibration for NWP model forecasts, starting with a general overview, and then discussing nonhomogeneous regression and Bayesian model averaging approaches in detail.

2.1 Overview

Following the pioneering work of Hamill and Colucci (1997; 1998), various types of ensemble calibration techniques have been developed over the past 15 years, with Wilks and Hamill (2007), Bröcker and Smith (2008), Schmeits and Kok (2010), Ruiz and Saulo (2012) and Thorarinsdottir, Scheuerer and

[3] This comment and overview apply to postprocessing efforts specifically for the ECMWF's ENS. In addition, a number of member countries support research and development efforts on statistical postprocessing for their own, nationally developed short- or medium-range systems. For example, the German Weather Service (DWD) and Heidelberg University have a joint project on ensemble calibration for the COSMO-DE ensemble run by the DWD, with results reported by Büermann (2012), Feldmann (2012), Scheuerer (2013) and Scheuerer and Büermann (2013).

[4] Related work on calibrating the ECMWF ENS has also been performed elsewhere, including efforts by Voisin, Schaake and Lettenmeier (2010), Hamill (2012), Schmeits and Kok (2012) and Courtney, Lynch and Sweeney (2013). At ECMWF, Doblas-Reyes, Hagedorn and Palmer (2005) and Molteni et al. (2011) have studied ensemble calibration for seasonal forecasts.

Feldmann (2012) providing reviews and discussions. As noted, ensemble calibration aims to correct for biases as well as dispersion errors in NWP model output, and the general goal is to maximize the sharpness of the postprocessed predictive distributions subject to calibration. Detailed comparisons between the predictive performance of the various types of postprocessing techniques yield rankings that depend on the ensemble system, meteorological variable and prediction horizon at hand.

State of the art techniques include the Bayesian model averaging (BMA) approach developed by Raftery et al. (2005), and the non-homogeneous regression (NR) or ensemble model output statistics (EMOS) technique of Gneiting et al. (2005). For these methods, software for the numerical implementation is available, including the `ensembleBMA` (Fraley et al. 2011) and `ensembleMOS` packages in `R`.[5] All methods involve statistical parameters that need to be estimated from training data, the choice of which we discuss in Section 4.

To fix the idea, let y denote the weather variable of interest, and write $x_1, \ldots, x_M$ for the corresponding M ensemble member forecasts. The NR/EMOS predictive distribution is a single parametric distribution of the general form

$$y \,|\, x_1, \ldots, x_M \sim f(y \,|\, x_1, \ldots, x_M), \tag{1}$$

where the left-hand side refers to the conditional distribution of the future weather quantity y, given the ensemble member forecasts $x_1, \ldots, x_M$ for y.[6] On the right-hand side, f is a parametric density function, with the parameters depending on the ensemble values in ways similar to traditional MOS approaches. For example, f could be a normal or Gaussian density, where the mean is a bias corrected affine function of the ensemble mean and the variance is a dispersion-corrected affine function of the ensemble variance.

The BMA approach employs a mixture distribution of the general form

$$y \,|\, x_1, \ldots, x_M \sim \sum_{m=1}^{M} w_m \, g(y \,|\, x_m), \tag{2}$$

where $g(y|x_m)$ denotes a parametric density that depends on the specific ensemble member forecast x_m in suitable ways, and where the mixture weights $w_1, \ldots, w_M$ are nonnegative and sum to 1. For example, $g(y|x_m)$ could be a normal density, where the mean is a bias corrected affine function of x_m and the variance is fixed at a certain value. The mixture weights $w_1, \ldots, w_m$ reflect the corresponding member's relative contributions to predictive skill over a training period. Viewed slightly differently, and relating closely to the ensemble dressing approach developed by Roulston and Smith (2003), Wang and Bishop (2005) and Fortin, Favre and Saïd (2006), each ensemble member is dressed with a certain kernel $g(y|x_m)$.

In general, the NR/EMOS approach tends to be more parsimonious (i.e., less complex and easier to fit), while the BMA method tends to be more flexible. For example, if f and g are Gaussian, then the NR/EMOS predictive distribution (1) is necessarily unimodal and symmetric, whereas the BMA predictive distribution (2), which is a mixture distribution, allows for multimodality as well as asymmetry. Evidently, the choice of the NR/EMOS predictive density f or the BMA component density g depends critically on the weather variable of interest, and in what follows we focus on temperature, pressure, precipitation and wind.

While we restrict attention to the NR/EMOS and BMA approaches, various other techniques are available, and typically they are similar in spirit to either NR/EMOS or BMA. For example, the ingenious

[5] The `ensembleBMA` and `ensembleMOS` packages are available for download at `www.r-project.org`. The Matlab toolbox `emtool` developed by Jochen Bröcker implements postprocessing techniques described by Bröcker and Smith (2008).

[6] At the expense of added complexity, we can furthermore condition f on observed variables, or on NWP model forecasts for quantities other than y. A similar comment applies to the general BMA equation (2), where the component densities g might also condition on observed variables, or on NWP model forecasts for quantities other than y, as in the approach of Glahn et al. (2009b).

approach of Wilks (2009), which originates in a logistic regression setting, can be interpreted within the NR/EMOS framework (Frei 2012; Roulin and Vannitsem 2012; Scheuerer 2013), and we do so hereinafter. Similarly, the ensemble dressing technique developed by Roulston and Smith (2003), Wang and Bishop (2005) and Fortin, Favre and Saïd (2006), the approach of Bröcker and Smith (2008), the ensemble kernel density MOS (EKDMOS) method of Glahn et al. (2009b), the ensemble regression approach developed by Unger et al. (2009) and the technique of Little, McSharry and Taylor (2009) can be interpreted within the mixture density framework of BMA.

Nonparametric techniques such as the cumulative distribution function (CDF) matching approach of Ihász et al. (2010) and Flowerdew (2012) and analog or similarity approaches as studied by Hamill and Whitaker (2006), and recently in talks by Laurent Descamps and Luca Delle Monache at the International Conference on Ensemble Methods in Geophysical Sciences in Toulouse, France, November 12–16, 2012, provide alternatives to the aforementioned techniques[7] that may gain prominence as reforecast datasets are becoming available. As discussed in the context of the normal quantile transform by Bogner, Pappenberger and Cloke (2012), approaches of this type have limitations, in that the implied predictive distributions do not allow for extrapolations beyond the most extreme extant observations.

2.2 Non-homogeneous regression (NR) or ensemble model output statistics (EMOS)

Overview As noted, the NR/EMOS predictive distribution is a single parametric distribution of the general form

$$y\,|\,x_1,\ldots,x_M \sim f(y\,|\,x_1,\ldots,x_M),$$

where the left-hand side refers to the conditional distribution of the future weather quantity y, given the ensemble member forecasts $x_1,\ldots,x_M$ for y, and where the parameters of the density function f on the right-hand side depend on the ensemble values in ways similar to traditional MOS approaches.

Temperature and pressure For temperature and pressure, Gneiting et al. (2005) employ Gaussian distributions as predictive densities. Specifically, if $\mathcal{N}(\mu,\sigma^2)$ denotes a normal or Gaussian density with mean μ and variance σ^2, the NR/EMOS predictive distribution for temperature or pressure is

$$y\,|\,x_1,\ldots,x_M \sim \mathcal{N}(a_0+a_1x_1+\cdots+a_Mx_M,\, b_0+b_1s^2), \tag{3}$$

where

$$s^2 = \frac{1}{M}\sum_{m=1}^{M}\left(x_m-\frac{1}{M}\sum_{m=1}^{M}x_m\right)^2 \tag{4}$$

denotes the variance of the ensemble values. The bias coefficients $a_0\in\mathbb{R}$ and $a_1,\ldots,a_M\geq 0$ and the spread coefficients $b_0\geq 0$ and $b_1\geq 0$ need to be fitted from training data, with Gneiting et al. (2005) applying a minimum score approach.[8] Hagedorn (2008), Hagedorn, Hamill and Whitaker (2008), Kann et al. (2009) and Kann, Haiden and Wittmann (2011), among others, have applied this approach to calibrate temperature forecasts.

[7] Interpreted broadly, temporally lagged or spatially aggregated ensembles (Theis, Hense and Damrath 2005; Bentzien and Friederichs 2012; Johnson and Wang 2012; Scheuerer 2013) can also be considered as nonparametric postprocessing approaches.

[8] The recently developed Bayesian estimation approach of Richter (2012) and the spatially adaptive approach of Scheuerer and Büermann (2013) offer alternatives.

The above formulation assumes that the ensemble members are statistically distinguishable. For the ECMWF's IFS, we have a group of 50 exchangeable ENS members, and so a modified form is natural. Using obvious notation for the HRES, CNT and ENS members of the 52-member IFS ensemble, which we denote by x_{HRES}, x_{CNT} and $x_1,\ldots,x_{50}$, respectively, we can adapt (3) to this setting, in that

$$y \,|\, x_{\text{HRES}}, x_{\text{CNT}}, x_1, \ldots, x_{50} \sim \mathcal{N}\left(a_0 + a_{\text{HRES}} x_{\text{HRES}} + a_{\text{CNT}} x_{\text{CNT}} + a_{\text{ENS}} \sum_{m=1}^{50} x_m,\ b_0 + b_1 s_{\text{ENS}}^2\right), \tag{5}$$

where

$$s_{\text{ENS}}^2 = \frac{1}{50}\sum_{m=1}^{50}\left(x_m - \frac{1}{50}\sum_{m=1}^{50} x_m\right)^2 \tag{6}$$

denotes the variance of the ENS members. The parameters to be fitted from training data are now the bias coefficients $a_0 \in \mathbb{R}$, $a_{\text{HRES}} \geq 0$, $a_{\text{CNT}} \geq 0$ and $a_{\text{ENS}} \geq 0$, and the spread coefficients $b_0 \geq 0$ and $b_1 \geq 0$, where the nonnegativity constraints can be implemented in ways described by Gneiting et al. (2005) and Thorarinsdottir and Gneiting (2010).

Quantitative precipitation Probability forecasts for the binary event of a precipitation accumulation exceeding (or not exceeding) a certain threshold are frequently obtained using logistic regression (Applequist et al. 2002; Wilks and Hamill 2007) or quantile regression (Bremnes 2004). If a full predictive distribution is required, these methods have difficulties, in that the probability at distinct thresholds may not be compatible with each other, nor does the aforementioned approach for temperature and pressure apply, given that quantitative precipitation is a nonnegative variable, having a point mass at zero, and typically being heavily skewed for positive precipitation accumulations.

Building on a logistic regression approach, Wilks (2009) proposed an elegant solution, which yields an NR/EMOS type of predictive distribution. Generalizing his approach slightly, and describing the predictive distribution in terms of the cumulative distribution function (CDF), $F(y)$, for the future precipitation accumulation, y, and the corresponding ensemble values $x_1,\ldots,x_M$, we obtain $F(y) = 0$ for $y \leq 0$ and

$$F(y) = \frac{\exp(a_0 + a_1 x_1^\alpha + \cdots + a_M x_M^\alpha + b y^\beta)}{1 + \exp(a_0 + a_1 x_1^\alpha + \cdots + a_M x_M^\alpha + b y^\beta)} \quad \text{for} \quad y \geq 0, \tag{7}$$

which is a type of truncated logistic distribution (Frei 2012). The implied probability of no precipitation is

$$F(0) = \frac{\exp(a_0 + a_1 x_1^\alpha + \cdots + a_M x_M^\alpha)}{1 + \exp(a_0 + a_1 x_1^\alpha + \cdots + a_M x_M^\alpha)}.$$

Here, $\alpha > 0$ and $\beta > 0$ are fixed coefficients, with Wilks (2009) selecting $\alpha = 1$ and $\beta = \frac{1}{2}$, and other choices being feasible, based on exploratory analyses, as discussed by Wilks (2009) and Roulin and Vannitsem (2012). The remaining coefficients $a_0 \in \mathbb{R}$, $a_1,\ldots,a_M \geq 0$ and $b > 0$ need to be fitted from training data. Adapted to the 52-member IFS, specifications such as

$$F(y) = \frac{\exp(a_0 + a_{\text{HRES}} x_{\text{HRES}}^\alpha + a_{\text{CNT}} x_{\text{CNT}}^\alpha + a_{\text{ENS}} \sum_{m=1}^{50} x_m^\alpha + b y^\beta)}{1 + \exp(a_0 + a_{\text{HRES}} x_{\text{HRES}}^\alpha + a_{\text{CNT}} x_{\text{CNT}}^\alpha + a_{\text{ENS}} \sum_{m=1}^{50} x_m^\alpha + b y^\beta)} \quad \text{for} \quad y \geq 0 \tag{8}$$

can be employed, where $\alpha > 0$ and $\beta > 0$ are fixed, while the coefficients $a_0 \in \mathbb{R}$, $a_{\text{HRES}} \geq 0$, $a_{\text{CNT}} \geq 0$, $a_{\text{ENS}} \geq 0$ and $b > 0$ are to be fitted from training data. More flexible versions that introduce interaction terms into (7) and (8), respectively, have been proposed by Ben Bouallègue (2013).

The behavior of the predictive distributions at high, extreme threshold values is of particular importance, and the use of extreme value distributions provides a less parsimonious, but interesting and relevant alternative NR/EMOS approach for potentially skewed weather variables, such as precipitation and wind speed (Bentzien and Friederichs 2012; Friederichs and Thorarinsdottir 2012; Scheuerer 2013).

Wind A bivariate wind vector can be represented by its zonal (west-east or u) and meridional (north-south or v) components, or by wind speed and wind direction. In the case of the u and v wind vector components, the Gaussian NR/EMOS specification (3) applies.

If direct probabilistic forecasts of wind speed are sought, the nonnegativity of the variable needs to be properly accounted for. Thorarinsdottir and Gneiting (2010) developed an EMOS approach for wind speed, where the predictive distribution is truncated normal,

$$y \,|\, x_1, \ldots, x_M \sim \mathcal{N}_{[0,\infty)}(a_0 + a_1 x_1 + \cdots + a_M x_m, b_0 + b_1 s^2), \tag{9}$$

with $\mathcal{N}_{[0,\infty)}(\mu, \sigma^2)$ denoting[9] a normal distribution with mean μ and variance σ^2 that is restricted to the positive halfaxis, so that it does not have any mass to the left of zero, while the density on the positive halfaxis is rescaled correspondingly. As usual, the coefficients $a_0 \in \mathbb{R}$, $a_1, \ldots, a_M \geq 0$ and $b > 0$ need to be fitted from training data.

To give an example, Figure 1 shows the postprocessed 48-hour ahead EMOS forecast of the maximum wind speed on June 14, 2003 at The Dalles, Oregon, a wind surfing hub in the Columbia River Gorge in the Pacific Northwest of the United States. The vertical black lines represent the eight members of the University of Washington Mesoscale Ensemble (Eckel and Mass 2005), which take values between 10.1 and 13.3 knots, respectively, with one knot being equal to 0.514 meters per second. The vertical red lines show the EMOS median forecast, at 15.3 knots, and the respective 77.8% central prediction interval for the EMOS density forecast. The blue line represents the verifying observation, at 18 knots. Figure 2 illustrates the postprocessed EMOS forecast distributions at The Dalles for the entire period of June 14 through July 31, 2003.[10] Adapting now to the 52-member IFS, we can employ a specification of the form

$$\begin{aligned} y \,|\, x_{\mathrm{HRES}}, x_{\mathrm{CNT}}, x_1, \ldots, x_{50} \sim \\ \mathcal{N}_{[0,\infty)}\left(a_0 + a_{\mathrm{HRES}} x_{\mathrm{HRES}} + a_{\mathrm{CNT}} x_{\mathrm{CNT}} + a_{\mathrm{ENS}} \sum_{m=1}^{50} x_m,\ b_0 + b_1 s^2_{\mathrm{ENS}}\right), \end{aligned} \tag{10}$$

where s^2_{ENS} is defined in (6), with the coefficients $a_0 \in \mathbb{R}$, $a_{\mathrm{HRES}} \geq 0$, $a_{\mathrm{CNT}} \geq 0$, $a_{\mathrm{ENS}} \geq 0$, $b_0 \geq 0$ and $b_1 \geq 0$ to be fitted from training data.

Thorarinsdottir and Johnson (2011) introduce an extension to wind gusts, while Pinson (2012) and Schuhen, Thorarinsdottir and Gneiting (2012) develop EMOS approaches that apply to bivariate wind vectors directly.

[9] Note that μ and σ^2 refer to the mean and the variance of the original normal distribution prior to truncation. The ensemble variance s^2 is defined in (4).

[10] Figures 1 and 2 are similar to displays presented in Thorarinsdottir and Gneiting (2010), where data from 2008 are used.

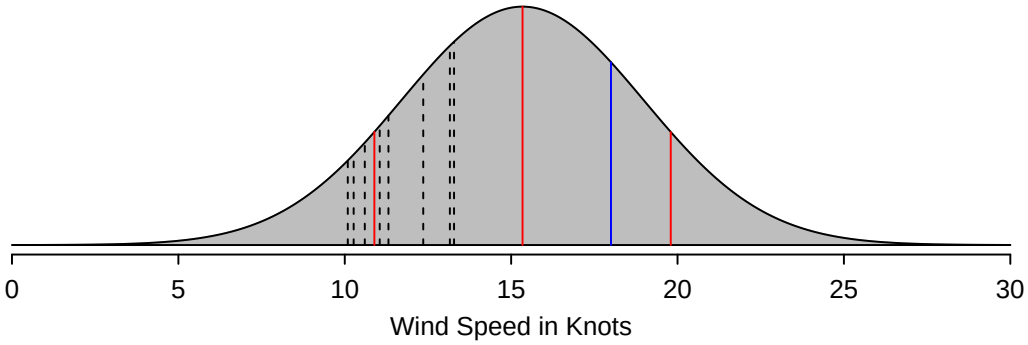

Figure 1: 48-hour ahead NR/EMOS density forecast of maximum wind speed valid June 14, 2003 at The Dalles, Oregon. The broken black lines represent the eight members of the University of Washington Mesoscale Ensemble (Eckel and Mass 2005). The red lines show the EMOS median forecast, at 15.3 knots, and the 77.8% central prediction interval for the EMOS density forecast. The blue line represents the verifying observation, at 18 knots. One knot equals 0.514 meters per second.

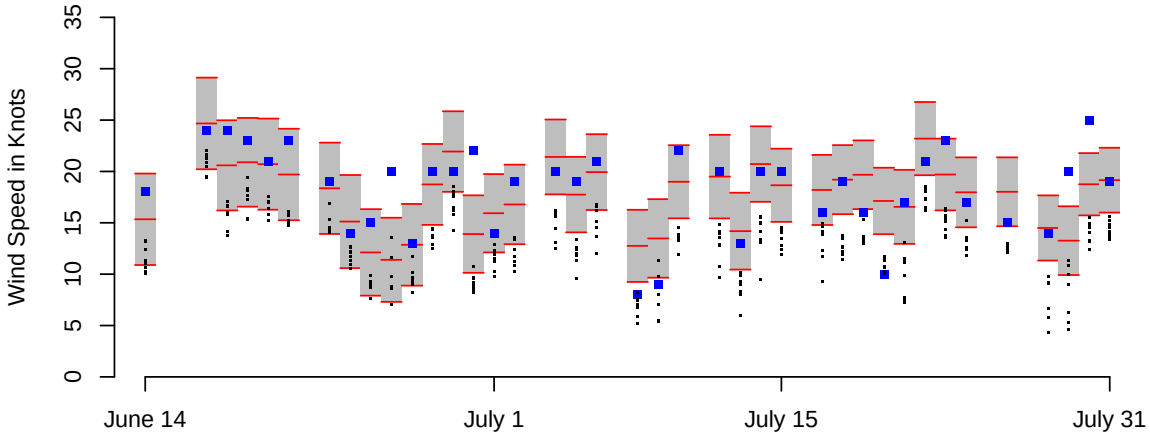

Figure 2: 48-hour ahead NR/EMOS forecasts of maximum wind speed at The Dalles, Oregon valid June 14 through July 31, 2003. The EMOS 77.8% prediction interval is shown as a gray and red box. The small black dots represent the eight members of the University of Washington Mesoscale Ensemble (Eckel and Mass 2005). The large blue dots mark the verifying wind speed.

Table 2: BMA implementations for univariate weather quantities, based on the ensemble values $x_m \in \mathbb{R}$, where $m = 1,\ldots,M$. In the case of precipitation amount, we refer to $y^{1/3} \in \mathbb{R}^+$, because the gamma kernels apply to cube root transformed precipitation accumulations (Sloughter et al. 2007). In the case of wind direction, $f_m \in \mathbb{S}$ is a bias corrected ensemble value on the circle and κ_m is a concentration parameter, for $m = 1,\ldots,M$ (Bao et al. 2010).

Weather Quantity	Range	Kernel	Mean	Variance
Temperature	$y \in \mathbb{R}$	Normal	$a_{0m} + a_{1m} x_m$	σ_m^2
Pressure	$y \in \mathbb{R}$	Normal	$a_{0m} + a_{1m} x_m$	σ_m^2
Precipitation amount	$y^{1/3} \in \mathbb{R}^+$	Gamma	$a_{0m} + a_{1m} x_m^{1/3}$	$b_{0m} + b_{1m} x_m$
Wind components	$y \in \mathbb{R}$	Normal	$a_{0m} + a_{1m} x_m$	σ_m^2
Wind speed	$y \in \mathbb{R}^+$	Gamma	$a_{0m} + a_{1m} x_m$	$b_{0m} + b_{1m} x_m$
Wind direction	$y \in \mathbb{S}$	von Mises	f_m	κ_m^{-1}
Visibility	$y \in [0,1]$	Beta	$a_{0m} + a_{1m} x_m^{1/2}$	$b_{0m} + b_{1m} x_m^{1/2}$

2.3 Bayesian model averaging

Overview As discussed, the BMA predictive distribution is a mixture distribution of the general form

$$y \,|\, x_1,\ldots,x_M \sim \sum_{m=1}^{M} w_m \, g(y \,|\, x_m),$$

where the component or kernel density $g(y \,|\, x_m)$ denotes a parametric distribution that depends on the specific ensemble member forecast x_m in suitable ways, and where the mixture weights $w_1,\ldots,w_M$ are nonnegative and sum to 1. Thus, each mixture component is a parametric probability distribution that is associated with a given individual ensemble member, with the mixture weight reflecting the member's relative contributions to the predictive skill over the training period.

Table 2 sketches BMA implementations for univariate weather variables, including temperature and pressure (Raftery et al. 2005), precipitation (Sloughter et al. 2007), wind speed (Sloughter, Gneiting and Raftery 2010), wind direction (Bao et al. 2010) and visibility (Chmielecki and Raftery 2011).[11] Furthermore, BMA implementations are available for fog (Roquelaure and Bergot 2008), ceiling (Chmielecki and Raftery 2011) and bivariate wind vectors (Sloughter, Gneiting and Raftery 2013).

We now discuss some of the details of the BMA implementations for temperature, pressure, quantitative precipitation and wind.

Temperature and pressure For temperature and pressure, Raftery et al. (2005) propose the BMA specification

$$y \,|\, x_1,\ldots,x_m \sim \sum_{m=1}^{M} w_m \, \mathcal{N}(a_{0m} + a_{1m} x_m, \, \sigma_m^2), \tag{11}$$

[11] Similar tables can be found in Möller, Lenkoski and Thorarinsdottir (2013) and Schefzik, Thorarinsdottir and Gneiting (2013).

so that the kernel density $g(y|x_m)$ is Gaussian with mean $a_{0m}+a_{1m}x_m$ and variance σ_m^2.[12] The BMA weights, $w_1,\ldots,w_M$, the bias parameters $a_{01},\ldots,a_{0M}$ and $a_{11},\ldots,a_{1M}\geq 0$, and the variance parameters $\sigma_1^2,\ldots,\sigma_M^2$ are fitted on training data in ways described by Raftery et al. (2005). Thus, the BMA mixture components are normal densities with a linearly bias corrected mean.

Fraley, Raftery and Gneiting (2010) discuss the adaptation of the basic BMA specification (11) to ensembles with exchangeable members. In the case of the ECMWF'S IFS, their proposals lead to a specification of the form

$$\begin{aligned} y\,|\,x_{\text{HRES}},x_{\text{CNT}},x_1,\ldots,x_{50} \;\sim\; & w_{\text{HRES}}\,\mathcal{N}\left(a_{\text{HRES}}+b_{\text{HRES}}x_{\text{HRES}},\sigma_{\text{HRES}}^2\right) \\ & + w_{\text{CNT}}\,\mathcal{N}\left(a_{\text{CNT}}+b_{\text{CNT}}x_{\text{CNT}},\sigma_{\text{CNT}}^2\right)+w_{\text{ENS}}\sum_{m=1}^{50}\mathcal{N}\left(a_{\text{ENS}}+b_{\text{ENS}}x_m,\sigma_{\text{ENS}}^2\right), \end{aligned} \tag{12}$$

where the various statistical parameters are to be estimated from training data, under the constraints that $w_{\text{HRES}}\geq 0$, $w_{\text{CNT}}\geq 0$ and $w_{\text{ENS}}\geq 0$ with $w_{\text{HRES}}+w_{\text{CNT}}+50w_{\text{ENS}}=1$, $b_{\text{HRES}}\geq 0$, $b_{\text{CNT}}\geq 0$, $b_{\text{ENS}}\geq 0$, $\sigma_{\text{HRES}}^2>0$, $\sigma_{\text{CNT}}^2>0$ and $\sigma_{\text{ENS}}^2>0$.

The BMA approach has been applied to ensemble temperature forecasts over Canada (Wilson et al. 2007) and Iran (Soltanzadeh, Azadi and Vakili 2011), and a real-time implementation over the Pacific Northwest region of the United States, based on the University of Washington Mesoscale Ensemble (UWME; Eckel and Mass 2005), is available to the general public at www.probcast.com. Alternatives to the estimation algorithm proposed by Raftery et al. (2005) include the fully Bayesian approaches of Vrugt et al. (2006) and Di Narzo and Cocchi (2010) and the spatially adaptive technique of Kleiber et al. (2011b) that uses geostatistical methods.

Quantitative precipitation Sloughter et al. (2007) propose a BMA implementation for quantitative precipitation, in which the BMA component density $g(y|x_m)$ is a Bernoulli–gamma mixture. The Bernoulli component provides a point mass at zero via logistic regression, specifying that

$$g(y=0\,|x_m)=\frac{\exp(c_{0m}+c_{1m}x_m^{1/3}+c_{2m}\delta_m)}{1+\exp(c_{0m}+c_{1m}x_m^{1/3}+c_{2m}\delta_m)},$$

where δ_m equals 1 if $x_m=0$ and equals 0 otherwise, under the constraints that $c_{1m}\leq 0$ and $c_{2m}\geq 0$. The continuous part of the kernel density is a gamma density for the cube root transformation, $y^{1/3}$, of the precipitation accumulation, with mean $\mu_m=a_{0m}+a_{1m}x_m^{1/3}$ and variance $\sigma_m^2=b_{0m}+b_{1m}x_m$, where the coefficients are assumed to be nonnegative and need to be fitted using training data.[13] Fraley, Raftery and Gneiting (2010) discuss modifications for ensembles with exchangeable members, which can be applied to the case of the ECMWF's IFS, leading to specifications identical in spirit to that in (12) for the BMA approach with Gaussian components. Finally, supplementing the estimation approach introduced by Sloughter et al. (2007), spatially adaptive techniques have become available (Kleiber, Raftery and Gneiting 2011a).

Wind As noted, a bivariate wind vector can be represented by its zonal and meridional components, or by wind speed and wind direction. In the case of wind components, the BMA approach described above

[12] In the standard implementation of Raftery et al. (2005), the predictive variance is assumed to be constant across the ensemble members. This assumption can readily be relaxed.

[13] In the standard implementation of Sloughter et al. (2007), some of the coefficients are assumed to be constant across the ensemble members. As in the case of temperature, assumptions of this type can readily be relaxed.

for temperature and pressure applies. If direct probabilistic forecasts of wind speed are sought, Sloughter, Gneiting and Raftery (2010) provide a BMA implementation, where the kernel density $g(y|x_m)$ associated with the ensemble member forecast x_m is a gamma density with the mean and the variance both being affine functions of x_m. Baran, Nemoda and Horányi (2013) apply this technique to calibrate the operational Limited Area Model Ensemble Prediction System of the Hungarian Meteorological Service. Wind direction is a circular variable, and Bao et al. (2010) develop a BMA specification where the kernel density $g(y|x_m)$ is a von Mises density on the circle. Sloughter, Gneiting and Raftery (2013) discuss a BMA implementation for bivariate wind vectors that employs mixtures of suitably transformed bivariate normal densities. Again, the adaptations proposed by Fraley, Raftery and Gneiting (2010) for ensembles with exchangeable members can be applied in the case of the ECMWF's IFS.

3 Accounting for structure in combined events and weather trajectories

Combined events and weather trajectories Statistical postprocessing techniques such as nonhomogenous regression (NR) and Bayesian model averaging (BMA) typically apply to a single weather variable at a single location and a single look-ahead time. However, in many applications it is critical that dependencies in combined events and temporal, spatial, or spatio-temporal weather trajectories are properly accounted for. For example, winter road maintenance requires joint probabilistic forecasts of temperature and precipitation (Berrocal et al. 2010), air traffic control calls for joint probabilistic forecasts of spatial wind fields (Chaloulos and Lygeros 2007), renewable energy forecasting depends on spatio-temporal weather trajectories (Pinson et al. 2009; Pinson 2013), and when ensemble forecasts are used to drive hydrologic models, Schaake et al. (2010, pp. 61–62) note that

> relationships between physical variables like, e.g. precipitation and temperature should be respected.

If statistical postprocessing proceeds independently for each weather variable, location and look-ahead time, dependencies between the respective forecast errors are ignored, and it is crucial that they be restored.

Spatio-temporal aspects Special considerations apply when the goal is the joint calibration of ensemble forecasts for temporal, spatial or spatio-temporal trajectories of a given weather variable. In what follows we consider the case of time trajectories; spatial or spatio-temporal scenarios can be handled analogously.

To obtain physically realistic, calibrated ensemble forecasts of time trajectories, it is essential that we address the following issues.

- The postprocessed univariate predictive distributions at distinct lead times need to be consistent among each other – unwarranted, abrupt change is to be avoided.
- The postprocessed ensemble member trajectories need to show physically realistic behavior across lead times.

To address the former aspect, which concerns the univariate marginal distributions only, constraints can be put on the BMA or NR coefficients, so that they vary smoothly across lead times, thereby ensuring the

temporal consistency of the postprocessed predictive distributions. Copula methods are ideally suited to addressing the latter requirement, and we discuss them in what follows.

In this context, a useful distinction is that between EPSgrams and plumes, as described in Sections 5.2 and 3.4.3 of Persson (2011), respectively. EPSgrams serve to illustrate univariate predictive distributions only; thus, if the ensemble calibration effort is aimed at EPSgrams, it suffices to address the former aspect. In contrast, plumes show individual ensemble member trajectories, and if the ensemble calibration effort aims at them, both of the above aspects are to be addressed.

Sklar's theorem Taking a technical perspective momentarily, standard approaches to statistical postprocessing yield a univariate postprocessed predictive cumulative distribution function (CDF), F_l, for each univariate weather quantity, Y_l, where $l = 1,\ldots,L$, with the multi-index $l = (i,j,k)$ referring to weather variable i, location j and look-ahead time k. What we seek is a physically realistic and consistent multivariate or joint predictive CDF, F, which, when restricted to each univariate weather variable, reduces to $F_1,\ldots,F_L$, respectively. By a celebrated theorem of Sklar (1959), every multivariate CDF F with this property can be represented in the form

$$F(y_1,\ldots,y_L) = C(F_1(y_1),\ldots,F_L(y_L)) \qquad \text{for} \qquad y_1,\ldots,y_L \in \mathbb{R},$$

where $C : [0,1]^L \to [0,1]$ is a so-called copula, i.e., a multivariate CDF with standard uniform marginal distributions.

Parametric copula approaches Maintaining the technical point of view, Sklar's theorem demonstrates that standard approaches to statistical postprocessing can accommodate any type of joint dependence structure for combined events, provided that a suitable copula function is specified. If the dimension L is small, or if specific structure can be exploited, such as temporal or spatial structure, parametric families of copulas can be employed, such as those reviewed by Schölzel and Friederichs (2008). For example, the approaches of Gel, Raftery and Gneiting (2004), Berrocal, Raftery and Gneiting (2007; 2008), Pinson et al. (2009), Berrocal et al. (2010), Schuhen, Thorarinsdottir and Gneiting (2012) and Möller, Lenkoski and Thorarinsdottir (2013) invoke Gaussian copulas.[14]

Approaches based on empirical copulas If L is huge and no specific structure can be exploited, one needs to resort to non-parametric approaches, thereby adopting the multivariate rank order structure in historical weather observations or the current ensemble forecast, as embodied in empirical copulas (Schefzik, Thorarinsdottir and Gneiting 2013). Specifically, the Schaake shuffle (Clark et al. 2004; Schaake et al. 2007; Voisin, Schaake and Lettenmeier 2010; Voisin et al. 2011) borrows the rank order structure from suitably chosen past weather records, whereas the ensemble copula coupling approach draws on rank order information supplied by the current ensemble forecast.

Ensemble copula coupling While the origins of the ensemble copula coupling (ECC) technique lie in the work of Bremnes (2007) and Krzysztofowicz and Toth (2008), and in a personal communication by Tom Hamill in summer 2009, the term ECC appeared first in the work of Schefzik (2011). Schefzik, Thorarinsdottir and Gneiting (2013) provide a detailed discussion from both theoretical and

[14] In some cases, this is done explicitly, in others implicitly. A different approach uses mixtures of multivariate normal distributions, such as in Schölzel and Hense (2011) or Sloughter, Gneiting and Raftery (2013). In this latter case, each mixture component can be linked to a Gaussian copula.

applied perspectives, show that techniques developed by Flowerdew (2012), Pinson (2012) and Roulin and Vannitsem (2012) can be interpreted within this framework, and apply ECC to the ECWMF's ENS system. Essentially, ECC is a multi-stage postprocessing procedure that generates a postprocessed ensemble with the same number of members as the original ensemble, and furthermore with the same rank order structure as that in the original ensemble, as follows.

1. Apply state of the art statistical postprocessing techniques to obtain calibrated and sharp predictive distributions for each weather variable, location and look-ahead time individually.

2. Draw a synthetic or statistical ensemble of the same size as the original ensemble from each univariate, postprocessed predictive distribution. Specifically, if the original, dynamic ensemble has M members, and F_l is the postprocessed predictive CDF for the future weather quantity Y_l, the statistical ensemble can taken to comprise equally spaced quantiles of F_l,[15] such as

$$\tilde{x}_{l,m} = F_l^{-1}\left(\frac{m-\frac{1}{2}}{M}\right) \qquad \text{or} \qquad \tilde{x}_{l,m} = F_l^{-1}\left(\frac{m}{M+1}\right) \tag{13}$$

   respectively, where $m = 1,\ldots,M$. This is done for each weather quantity Y_l individually, until all indices $l = 1,\ldots,L$ have been covered.

3. Reassign the postprocessed values $\tilde{x}_{l,1},\ldots,\tilde{x}_{l,M}$ to the M members in such a way that the rank order structure of the raw ensemble is restored.

As Schefzik, Thorarinsdottir and Gneiting (2013) show, this type of procedure is equivalent to adopting the empirical copula of the raw ensemble, thereby justifying the term ensemble copula coupling. The key idea, however, is straightforward, in that the ECC postprocessed ensemble inherits the multivariate rank dependence structure from the raw ensemble, thereby capturing the flow dependence, as explained colorfully by Flowerdew (2012, p. 17):

> The key to preserving spatial, temporal and inter-variable structure is how this set of values is distributed among ensemble members. One can always construct ensemble members by sampling from the calibrated PDF, but this alone would produce spatially noisy fields lacking the correct correlations. Instead, the values are assigned to ensemble members in the same order as the values from the raw ensemble: the member with the locally highest rainfall remains locally highest, but with a calibrated rainfall magnitude.

The defining feature of the ECC approach, namely, the preservation of the rank order structure in the raw ensemble, also sets its limitations, in that ECC is unable to fix any inconsistencies between the postprocessed marginal distributions themselves, and operates under a perfect model assumption for the dependence structures across weather variables, locations and look-ahead times. It is therefore to be expected that in low-dimensional or highly structured settings parametric techniques outperform ECC,[16] provided that a sufficient supply of training data permits the statistical correction of dependence structures. This is a critically important cutting-edge area of interdisciplinary research, with considerable progress likely to materialize over the next decade.

[15] The use of equally spaced quantiles has also been proposed by Hagedorn (2010) and Hagedorn et al. (2012).

[16] The recent study of Schuhen, Thorarinsdottir and Gneiting (2012) on wind vectors, where each site and look-ahead time is considered individually — and thus the dimensionality is $L = 2$ only, corresponding to the two wind vector components — illustrates this type of situation.

4 The choice of training data

Using observations vs. using analyses Statistical postprocessing is based on the idea of exploiting structured patterns seen in past forecast-observation pairs in order to correct any systematic deficiencies in the raw model output. Thus, the use of ensemble calibration techniques depends on the availability of training data, consisting of historic ensemble forecasts and the corresponding verifying values. In particular, the nonhomogeneous regression (NR) and Bayesian model averaging (BMA) techniques described in Section 2 require coefficients and statistical parameters to be estimated from training data. In doing so, one typically uses optimum score approaches (Gneiting et al. 2005; Gneiting and Raftery 2007), with the classical technique of maximum likelihood estimation (Raftery et al. 2005; Wilks 2011, Section 4.6) being a special case thereof.[17]

Either station observations or analyses can serve to provide the verifying values in the training set. Station observations have the appealing property that they are model inpedendent, as opposed to the use of analyses, which bear the danger of missing intricate features in weather field observations. Therefore, station observations are natural choices for training purposes.[18] For example, synoptic observations (SYNOPs) from the Global Telecommunication System (GTS) have been available and in use at the ECMWF (Haiden, Rodwell and Richardson 2012). Such an approach requires that the raw ensemble output is interpolated from the model grid to the station locations. The UK Met Office's Intelligent Grid Point Selection (IGPS) approach along with lapse rate corrections in the case of temperature is a state of the art technique for doing this (Sheridan et al. 2010).

Statistical postprocessing directly on the model grid and over large, potentially global areas, particularly over the oceans or above the surface level, where observational assets are scarce, or for variables not directly measured at weather stations, may necessitate the use of an analysis to provide verifying values in training sets. While this is a viable and attractive option, the above caveat applies, and any such efforts ought to be subject to scrutiny, to avoid undue influences of any NWP model peculiarities on the postprocessing. Considerations and experiments similar to those reported in Box A of Hagedorn (2010) and Section 4.3 of Hagedorn et al. (2012) seem helpful here.

Rolling training periods and adaptive estimation Typically, the training data are taken from a rolling training period consisting of the recent past, including the most recent available ensemble forecasts along with the respective verifying values, with common choices for the length of the training period ranging from 20 to 40 days. In schemes of this type, the training set is updated continually, thereby allowing the estimates to adapt to changes in the seasons and in large scale weather regimes. Clearly, there is a trade-off here, in that larger training periods may allow for better estimation in principle, thereby reducing estimation variances, but may introduce biases due to seasonally varying effects and relationships. More flexible, adaptive estimation approaches such as recursive maximum likelihood techniques have been explored by Pinson and Madsen (2009), Raftery, Kárný and Ettler (2010) and Pinson (2012).

In addition to deciding about the temporal extent of the training sets, choices regarding their spatial composition are to be made. Local approaches use training data from the station location or grid box at hand only, resulting in distinct sets of coefficients that are tailored to the local terrain, while regional approaches composite training sets spatially, to estimate a single set of coefficients that is then used over an entire region (Thorarinsdottir and Gneiting 2010). Recently, flexible spatially adaptive approaches have been developed that estimate coefficients at each station location individually, interpolating them

[17] For alternative approaches see Section 2.

[18] This argument also supports the use of station observations for verification purposes.

to sites where no observational assets are available (Kleiber, Raftery and Gneiting 2011a; Kleiber et al. 2011b; Scheuerer and Büermann 2013).

Using reforecast datasets Introduced by Hamill, Whitaker and Mullen (2004), reforecasts are retrospective weather forecasts, where today's NWP models are applied to past initialization and prediction dates. As the reforecasts are based on the model version that is currently run operationally, the availability of reforecast datasets can result in massive enlargements of training sets for statistical postprocessing. For example, if we seek to calibrate an ensemble forecast in mid June 2013, and reforecasts with the same ensemble system are available since 1980, we can augment the training set by the ensemble reforecasts and the corresponding verifying values for each June, or perhaps each May, June and July, from 1980 through 2012.

In the United States, Hamill, Whitaker and Mullen (2004), Hamill, Whitaker and Wei (2004), Hamill and Whitaker (2006), Whitaker, Wei and Vitart (2006) and Hamill and Whitaker (2007), among others, have successfully used the National Centers for Environmental Prediction (NCEP) ensemble reforecast in statistical postprocessing. For example, Hamill and Whitaker (2006) explored the use of the nonparametric analog technique, which has particularly high demands in terms of the quantity of training data, and becomes a powerful alternative when massive amounts of reforecast data are available.

In Europe, the use of reforecast data to calibrate the ECMWF's ENS has been studied with focus on temperature (Hagedorn 2008; Hagedorn, Hamill and Whitaker 2008; Ihász et al. 2010; Vannitsem and Hagedorn 2011; Hagedorn et al. 2012) and precipitation (Hamill, Hagedorn and Whitaker 2008; Ihász et al. 2010; Hamill 2012; Roulin and Vannitsem 2012). Specifically, Hagedorn (2008) and Hagedorn, Hamill and Whitaker (2008) apply the NR technique of Gneiting et al. (2005) to calibrate ensemble forecasts of surface temperature, using either reforecast training sets or a traditional 30-day rolling training period. At prediction horizons up to three days, the added value of the reforecast data is small; at larger prediction horizons, calibration on reforecast data yields major gains compared to the 30-day training period. As regards precipitation, Hamill, Hagedorn and Whitaker (2008) report major benefits from the use of reforecast data for all lead times, with the improvement in the predictive performance being the most pronounced at higher thresholds. For example, at a threshold of 10 mm, 2-day ahead probability of 12-hourly precipitation forecasts calibrated on reforecast data are as skillful as half-day ahead forecasts calibrated on a 30-day training period.

Design of operational reforecast datasets While the aforementioned results make a persuasive case for the use of reforecast data in ensemble calibration, it is far from obvious how an operational reforecast system ought to be optimally designed and implemented (Hagedorn 2008). As Hamill, Hagedorn and Whitaker (2008) note, three key questions are to be addressed in finding the best possible compromise between costs and benefits, given limited computational resources. How many members should an operational reforecast ensemble comprise? How many years should it go back in time? And at what frequency should the reforecast ensemble be run — every day, every third day, every week, every tenth day, or perhaps every two weeks only?

The empirical studies of Hagedorn (2008), Hagedorn, Hamill and Whitaker (2008) and Hamill, Hagedorn and Whitaker (2008) address these questions for the ECMWF's ENS, using an experimental dataset comprising 15-member ensemble reforecasts for the 20-year period from 1982 through 2001, with reforecasts computed once weekly for the fall season. To summarize the findings, the improvement by the use of daily (as opposed to weekly) reforecasts generally is small, and the inclusion of further years in the reforecast dataset yields higher benefits than the inclusion of further ensemble members. Based on these

experiments and consideration of additional operational requirements and constraints, the current operational ensemble reforecast at the ECMWF comprises five members, including the CNT member and four PTB members, which are run once every week over the past 20 years (Hagedorn 2008; Richardson 2012).

While these choices result in a well balanced compromise between costs and benefits, matters are considerably complex, as exemplified by Hagedorn (2010), who combines simple bias correction techniques for temperature on a 30-day rolling training period with the NR ensemble calibration method trained on reforecast data. Roulin and Vannitsem (2012, p. 886) fit ensemble calibration techniques for quantitative precipitation forecasts based on the ECMWF's ENS on training datasets with different numbers of ensemble members, summarizing that

> As the size of the hindcasts is low compared to the full operational ensemble, the regression parameters should be corrected for biases by using, for instance, the so-called regression calibration method.

These experiences serve to illustrate a critical point, in that the optimal choice of training data, including the design of operational reforecast datasets, and the best choice of statistical postprocessing techniques to be used, depend on each other, thus calling for a feedback loop and a periodic, joint reassessment of operational efforts. For example, at a given computational budget, less frequent reforecasts using all members might be preferable over more frequent reforecasts using a subset of members only, as they avoid the need for the type of corrections discussed by Roulin and Vannitsem (2012).

In addition to empirical studies, there is scope for pertinent, theoretically informed work on these issues, based on considerations of effective sample sizes (Wilks 2011, p. 147; Hagedorn, Hamill and Whitaker 2008) and the asymptotic theory of statistical inference in dependent data settings. As a simple illustration, Appendix B provides a tentative discussion of the desirable size of a reforecast dataset when the goal is the estimation of extreme quantiles of the model climate, which is essential to the use of the ECMWF's extreme forecast index (EFI; Lalaurette 2003). Similar considerations might inform decisions on the tradeoff between costs and benefits when the goal is statistical postprocessing.

5 Recommendations

Recommendations for implementation, research and development State of the art statistical postprocessing processing techniques can significantly improve the skill of medium-range forecasts of surface weather parameters. National and international weather centers worldwide are beginning to move towards the operational implementation of ensemble calibration techniques. For example, the National Research Council (2012, p. 37) recommends that the United States National Weather Service (NWS)

> [...] needs to employ statistical techniques to post-process its ensemble forecasts so as to obtain calibrated probabilistic forecasts — these are now within reach. Other national agencies have not yet started issuing probabilistic forecasts of the main weather and water elements, such as temperature, precipitation, and wind speed, on a regular basis, so this is an area in which the NWS has an opportunity to take the lead globally.

There is a strong case to pursue statistical postprocessing at the medium-range operationally, to ensure that the most powerful methods and techniques available are being applied, resulting in full ensemble calibration rather than just bias correction, to take full advantage of the 52-member IFS rather than just the HRES run — a critical task that has recently been emphasized by the ECMWF Direcorate (2012).

Recommendations for implementation The following summarizes recommendations for an initial implementation of an operational ensemble calibration system using ECMWF forecasts.

1. In order to gain maximum benefit from the ECMWF medium-range NWP forecasts, it is recommended that statistical postprocessing is applied to the IFS output.

2. The calibration could be done, for example, for 2m temperature, precipitation, 10m wind speed and cloud cover as shown on the standard ECMWF EPSgrams, using well established parametric ensemble calibration techniques, such as nonhomogeneous regression (NR) or Bayesian model averaging (BMA), with suitable ramifications that ensure temporal consistency across lead times. Data from synoptic observations (SYNOPs) stations that report via the Global Telecommunication System (GTS), or from local networks, can be used as observations in training sets as well as for verification.

Recommendations for research and development The following identifies cutting edge areas that are in critical need of research and development.

1. In order to augment training sets, it is recommended that operational ensemble reforecasting efforts be continued, expanded and strengthened. As decisions about the best design of reforecast datasets and about the optimal choice of statistical postprocessing techniques depend on each other, a periodic, joint reassessment of ensemble calibration and ensemble reforecast efforts is essential.[19] Both empirical and theoretical work has the potential of informing such decisions.

2. The development of model grid based products can also be considered, possibly using analyses for training purposes.

3. Research and development is called for to devise more sophisticated methods for the statistical postprocessing of combined events[20] and, closely related, for the development of theoretically principled techniques for the evaluation of probabilistic forecasts of multivariate quantities and events.

Another important consideration is that insights from byproducts of statistical postprocessing, such as can be obtained by inspecting patterns and peculiarities in the fitted calibration parameters, serve to inform the design of NWP systems. This has the potential to provide a wealth of new and fruitful feedback on the performance of the raw IFS output, and thus will aid ongoing as well as future model development.

[19] While the details of the schedule will depend on the pace and timing of change in the IFS and statistical postprocessing efforts, a reassessment at the order of every five years seems appropriate and useful.

[20] Given its simplicity of implementation, the ensemble copula coupling (ECC) technique can serve as a benchmark, and potentially as an initial solution, in the consideration of dependencies in combined events and spatio-temporal weather trajectories. For example, ECC can be used to construct physically realistic plumes from calibrated EPSgrams, as alluded to in Section 3 and described in detail by Schefzik, Thorarinsdottir and Gneiting (2013). In a nutshell, ECC restores the raw ensemble's multivariate rank dependence structure. Viewing the ECMWF's IFS as a 52-member ensemble, ECC can be realized by sampling $M \leq 52$ equally spaced quantiles (13) of the univariate postprocessed cumulative distribution functions, and rearranging these values according to the rank dependence structure in a collection of M raw ensemble members. If exemplary future spatio-temporal weather trajectories are to be identified, clustering techniques such as those developed by Ferranti and Corti (2011) can be applied to the postprocessed ensemble member scenarios.

Concluding remarks In closing, a relevant question is that for any risks of ensemble calibration. If inappropriate postprocessing techniques or insufficient training data are used, then indeed statistical postprocessing can deterioriate the forecast skill. Furthermore, leveraging potentially improved observational assets, advancements in the data assimilation cycle, model and ensemble designs, numerical techniques and computing power, the skill of raw NWP ensemble output is expected to increase, whence at some future stage the need for statistical postprocessing might become less pronounced, despite the enormous current benefits. To address these concerns, it is recommended that operational postprocessing and reforecast efforts be reviewed periodically, as noted.

Another potential concern is the univariate character of the majority of postprocessing techniques currently in use. As discussed in Section 3, copula methods are tailored to addressing this issue. If the ensemble copula coupling (ECC) technique is used, the multivariate rank dependence structure of the postprocessed ensemble is identical to that of the raw ensemble, thereby precluding degradation. Anticipating future implementations of truly multivariate ensemble calibration techniques, statistical postprocessing has the potential of correcting NWP model biases in terms of dependence structures, too.

While this report focuses on medium-range forecasts, the use of statistical postprocessing techniques can assist in the transition to seamless prediction (Palmer et al. 2008), thereby addressing the short-range, the medium-range, the extended range, and long-range or seasonal forecasts of average conditions (Palmer et al. 2004; Molteni et al. 2011) simultaneously, using blending techniques such as those developed by Kober et al. (2012).

Acknowledgements

I am indebted to the many colleagues and students at the ECMWF, the University of Washington, Heidelberg University, and elsewhere, who have graciously shared their expertise, intuition and time, both during the production phase of this report, and over the past decade, as ensemble calibration techniques have approached today's maturity. While the opinions and any errors in the report are my own, their contributions have been invaluable. In particular, I appreciate the written feedback provided by Erik Andersson, Sandor Baran, Konrad Bogner, Roberto Buizza, Kira Feldmann, Stephan Hemri, Florian Pappenberger, and Michael Scheuerer, and code supplied by Thordis Thorarinsdottir. Special thanks go to David Richardson, my key contact at the ECMWF during the work on this report, for a wealth of expert comments, including the suggestion of the EPSgram project, as well as his unfailing support and patience.

Appendix A: Calibration of hydrologic forecasts

Krzysztofowicz (2001) made a powerful case for the transition from single-valued forecasts to probabilistic forecasts in hydrologic practice. In doing so, various types of predictive uncertainty need to be addressed, including initial conditions uncertainties, meteorological forcing uncertainties, hydrologic model parameter uncertainties, and model structural uncertainty (Ajami, Duan and Sorooshian 2007; Bourdin, Fleming and Stull 2012), with Zappa et al. (2012) arguing that meteorological forcing uncertainties are dominant, and that the full spread from uncertainty superpositions grows non-linearly.

Operational streamflow and water level forecasting systems are increasingly using precipipation forecasts from NWP ensembles to drive hydrological models and predictions, thereby addressing meteorological

uncertainties. Table 1 in Cloke and Pappenberger (2009) provides a gamut of examples of such systems. To mention just a few case studies of ensemble riverine flow forecast systems of this type, Renner et al. (2009) consider the river Rhine, Addor et al. (2011) the city of Zürich, and Alfieri, Thielen and Pappenberger (2012) a catchment in southern Switzerland.

Hydrologic model uncertainty has been addressed by Ajami et al. (2006), Ajami, Duan and Sorooshian (2007), Vrugt and Robinson (2007), Parrish, Moradkhani and DeChant (2011) and Rings et al. (2012), who consider forecast ensembles for streamflow whose members rely on distinct hydrologic models. In the study of Georgakakos et al. (2004), each such model is driven by radar observations, rather than precipitation forecasts, thereby isolating the effects of model structural uncertainty. Duan et al. (2007) consider model parameter as well as model structural uncertainty, by studying a nine-member ensemble, consisting of three hydrologic models using three parameter sets each.

It seems likely that next generation ensemble streamflow prediction system will consider each of the aforementioned types of uncertainty, possibly in Latin square designs, with initial conditions being sampled, meteorological forcings being supplied by state of the art NWP ensembles, and suitable sets of hydrologic models being used, with model parameters being sampled appropriately. Regardless of the design of the ensemble system, the argument of Cloke and Pappenberger (2009, p. 622) applies, in that

> current [ensemble] based forecasts do not result in true probabilities of flooding, as uncertainties are not treated fully and the assumptions of some of the approaches are violated.

In particular, as in the case of NWP ensembles, hydrologic ensembles tend to be biased, and typically they are underdisperded. Thus, some form of statistical postprocessing is required, with Thielen et al. (2008, p. 33) summarizing community thoughts as follows:

> Postprocessing routines capable of reducing or correcting the uncertainty of the hydrologic ensemble model output were discussed as the best way forward for operational applications.

Toward this end, Hashino, Bradley and Schwartz (2007) apply a simple postprocessing technique that applies a bias correction to each ensemble member individually, while Reggiani et al. (2009) adapt the Bayesian approach of Krzysztofowicz (1999). Fundel and Zappa (2011) apply the extended logistic regression approach of Wilks (2009) and explore the use of reforecast data. Bogner and Pappenberger (2011) devise postprocessing techniques for the European Flood Awareness System (EFAS; Thielen et al. 2009; Bartholmes et al. 2009), using vector autoregressive models with exogeneous input (VARX) for normal quantile transformed ensemble data.

Various authors have explored the use of nonhomogeneous regression or model averaging to calibrate hydrologic ensembles, with Ajami et al. (2006), Ajami, Duan and Sorooshian (2007), Duan et al. (2007), Vrugt and Robinson (2007), Wood and Schaake (2008), Diks and Vrugt (2010) and Strauch et al. (2012) investigating techniques for doing this. Recent developments include the combination of BMA and particle filter techniques (Parrish, Moradkhani and DeChant 2012; Rings et al. 2012), which leads to the use of flexible and adaptive, rather than fixed, types of kernels in the mixture densities that form the predictive distributions.

In hydrologic applications, the calibration of ensemble forecasts of time trajectories that are physically realistic and consistent across lead times is of particular importance. As discussed in Section 3, a possible way of achieving this is by putting constraints on the calibration parameters, so that they vary smoothly across lead times, which ensures the consistency of the postprocessed marginal predictive distributions, and then applying the ensemble copula coupling (ECC) technique, or fitting a parametric copula model, in order to honor the dependence structure across lead times within the time trajectories.

Table 3: The ratio r_p in (14) of the sample size that is needed to estimate the quantile of a normal population at level p to the same accuracy as in the case of the median.

p	r_p
0.500	1.00
0.750	1.18
0.900	1.86
0.950	2.84
0.975	4.54
0.990	8.87
0.995	15.15

Appendix B: Sample size considerations in the estimation of extreme quantiles from reforecast data

This appendix contains initial thoughts on the question for suitable sizes and uses of reforecast datasets, motivated by the setting in Zsoter, Pappenberger and Richardson (2014), where the ECMWF reforecast dataset is used to estimate extreme quantiles of the model climate in the context of the extreme forecast index (EFI; Lalaurette 2003).

To set the stage, let $X_1, \ldots, X_n$ be independent identically distributed random variables from a distribution with a strictly positive probability density function, f. For a positive integer n and any fixed value of $p \in (0,1)$, we let q_p denote the true theoretical quantile of f, and we write $X_{([np])}$ for the respective sample quantile. By classical asymptotic theory, as described by Walker (1968) and the references therein, if n is large the distribution of the sample quantile $X_{([np])}$ is approximately normal with mean q_p and variance

$$V_{n,p} = \frac{1}{n} \frac{p(1-p)}{f^2(q_p)}.$$

If f is unimodal and symmetric, then clearly $V_{n,p}$ is minimal when $p = \frac{1}{2}$, and the ratio

$$r_p = \frac{V_{n,p}}{V_{n,\frac{1}{2}}} = 4\,p(1-p)\frac{f^2(q_{\frac{1}{2}})}{f^2(q_p)}$$

can be interpreted as follows: In order to estimate the quantile q_p with the same accuracy as the median, $q_{\frac{1}{2}}$, we need a sample size that is r_p times larger than in the case of the median. In particular, if f is normal then

$$r_p = 4\,p(1-p)\exp(z_p^2), \tag{14}$$

where z_p denotes the standard normal quantile at level p. Table 3 shows values of the ratio r_p for $p \geq \frac{1}{2}$. For example, if $p = 0.99$ we need a sample that is about nine times larger than in the case of the median. If $p = 0.995$ we need a sample that is about 15 times larger than in the case of the median.

These results are illustrative only, in that the assumptions of Gaussianity and independence are unlikely to hold in practice. However, the asymptotic theory extends to dependent random variables. For a detailed discussion see Wu (2005) and the references therein.

References

Addor, N., Jaun, S., Fundel, F. and Zappa, M. (2011). An operational hydrologic ensemble prediction system for the city of Zurich (Switzerland): Skill, case studies and scenarios. *Hydrology and Earth System Sciences*, **15**, 2327–2347.

Ajami, N. K., Duan, Q. and Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structure uncertainty in hydrologic prediction. *Water Resources Research*, **43**, W01403.

Ajami, N. K., Duan, Q., Gao, X. and Sorooshian, S. (2006). Multimodel combination techniques for analysis of hydrometeorological simulations: Application to distributed model intercomparison project results. *Journal of Hydrometeorology*, **7**, 755–768.

Alfieri, L., Thielen, J. and Pappenberger, F. (2012). Ensemble hydro-meteorological simulation for flash flood early detection in southern Switzerland. *Journal of Hydrology*, **424–425**, 143–153.

Applequist, S., Gahrs, G. E., Pfeffer, R. L. and Niu, X.-F. (2002). Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Weather and Forecasting*, **17**, 783–799.

Baran, S., Nemoda, D. and Horányi, A. (2013). Probabilistic wind speed forecasting in Hungary. *Meteorologische Zeitschrift*, **22**, 273–282.

Bartholmes, J. C., Thielen, J., Ramos, M. H. and Gentilini, S. (2009). The European Flood Awareness System EFAS — Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences*, **13**, 141–153.

Bao, L., Gneiting, T., Grimit, E. P., Guttorp, P. and Raftery, A. E. (2010). Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review*, **138**, 1811–1821.

Ben Bouallègue, Z. (2013). Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting*, **28**, 515–524.

Bentzien, S. and Friederichs, P. (2012). Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Weather and Forecasting*, **27**, 988–1002.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2007). Combining spatial statistical and ensemble information for probabilistic weather forecasting. *Monthly Weather Review*, **135**, 1386–1402.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, **2**, 1170–1193.

Berrocal, V. J., Raftery, A. E., Gneiting, T. and Steed, R. (2010). Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association*, **105**, 522–537.

Bogner, K. and Pappenberger, F. (2011). Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resources Research*, **47**, W07524.

Bogner, K., Pappenberger, F. and Cloke, H. L. (2012). The normal quantile transformation and its application in a flood forecasting system. *Hydrology and Earth System Sciences*, **16**, 1085–1094.

Bourdin, D. R., Fleming, S.-W. and Stull, R. B. (2012). Streamflow modelling: A primer on applications, approaches and challenges. *Atmosphere–Ocean*, **50**, 507–536.

Bremnes, J. B. (2004). Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Monthly Weather Review*, **132**, 338–347.

Bremnes, J. B. (2007). Improved calibration of precipitation forecasts using ensemble techniques. Part 2: Statistical calibration methods. Norwegian Meteorological Institute, Technical Report no. 04/2007, available at met.no/Forskning/Publikasjoner/Publikasjoner_2007/filestore/report04_2007.pdf.

Bröcker, J. and Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus Series A*, **60**, 663–678.

Büermann, L. (2012). Spatially adaptive postprocessing of ensemble forecasts for temperature over Germany. Diploma thesis, Heidelberg University, Germany.

Buizza, R. and Palmer, T. N. (1995). The singular-vector structure of the atmospheric global circulation. *Journal of the Atmospheric Sciences*, **52**, 1434–1456.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. (2005). A comparison of the ECMWF, MSC, and NCEP ensemble prediction systems. *Monthly Weather Review*, **133**, 1076–1097.

Buizza, R., Miller, M. and Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **125**, 2887–2908.

Carter, G. M., Dallavalle, J. P. and Glahn, H. R. (1989). Statistical forecasts based on the National Meteorological Center's prediction system. *Weather and Forecasting*, **4**, 401–412.

Chaloulos, G. and Lygeros, J. (2007). Effect of wind correlation on aircraft conflict probability. *Journal of Guidance, Control, and Dynamics*, **30**, 1742–1752.

Chmielecki, R. M. and Raftery, A. E. (2011). Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review*, **139**, 1626–1639.

Clark, M., Gangopadhyay, S., Hay, L., Rajagalopalan, B. and Wilby, R. (2004). The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, **5**, 243–262.

Cloke, H. L. and Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, **375**, 613–626.

Courtney, J. F., Lynch, P. and Sweeney, C. (2013). High resolution forecasting for wind energy applications using Bayesian model averaging. *Tellus Series A*, **65**, 19669.

Crochet, P. (2004). Adaptive Kalman filtering of 2-metre temperature and 10-metre wind-speed forecasts in Iceland. *Meteorological Applications*, **11**, 173–187.

Cui, B., Toth, Z., Zhu, Y. and Hou, D. (2012). Bias correction for global ensemble forecast. *Weather and Forecasting*, **27**, 396–410.

Diks, C. G. H. and Vrugt, J. A. (2010). Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stochastic Environmental Research and Risk Assessment*, **24**, 809–820.

Di Narzo, A. F. and Cocchi, D. (2010). A Bayesian hierarchical approach to ensemble weather forecasting. *Applied Statistics*, **59**, 405–422.

Doblas-Reyes, F. J., Hagedorn, R. and Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Calibration and combination. *Tellus Series A*, **57**, 234–252.

Duan, Q., Ajami, Q. H., Gao, X. and Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, **30**, 1371–1386.

Eckel, F. A. and Mass, C. F. (2005). Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, **20**, 328–350.

ECMWF Directorate (2012). Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter*, **133**, 11–13.

Feldmann, K. (2012). Statistical postprocessing of ensemble forecasts for temperature: The importance of spatial modeling. Diploma thesis, Heidelberg University, Germany.

Ferranti, L. and Corti, S. (2011). New clustering products. *ECMWF Newsletter*, **127**, 6–11.

Flowerdew, J. (2012). Calibration and combination of medium-range ensemble precipitation forecasts. UK Met Office, Forecasting Research Technical Report 567, available at www.metoffice.gov.uk/media/pdf/h/6/FRTR567.pdf.

Fortin, V., Favre, A.-C. and Saïd, M. (2006). Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quarterly Journal of the Royal Meteorological Society*, **132**, 1349–1369.

Fraley, C., Raftery, A. E. and Gneiting, T. (2010). Calibrating multi-model forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, **138**, 190–202.

Fraley, C., Raftery, A., Gneiting, T., Sloughter, M. and Berrocal, V. (2011). Probabilistic weather forecasting in R. *R Journal*, **3(1)**, 55–63.

Frei, C. (2012). Probabilistic forecasts of precipitation using quantiles. Diploma thesis, Heidelberg University, Germany.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, **23**, 579–594.

Fundel, F. and Zappa, M. (2011). Hydrologic ensemble forecasting in mesoscale catchments: Sensitivity to initial conditions and value of reforecasts. *Water Resources Research*, **47**, W09520.

Gel, Y., Raftery, A. E. and Gneiting, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The Geostatistical Output Perturbation (GOP) method (with discussion). *Journal of the American Statistical Association*, **99**, 575–587.

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J. and Butts, M. B. (2004). Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology*, **298**, 222–241.

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.

Glahn, H. R., Gilbert, K., Cosgrove, R., Ruth, D. P. and Sheets, K. (2009). The gridding of MOS. *Weather and Forecasting*, **24**, 520–529.

Glahn, H. R., Peroutka, M., Weidenfeld, J., Wagner, J., Zylstra, G. and Schuknecht, B. (2009b). MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review*, **137**, 246–268.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society Series A*, **171**, 319–321.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**, 746–762.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, **310**, 248–249.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, **67**, 243–268.

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

Hagedorn, R. (2008). Using the ECMWF reforecast data set to calibrate EPS forecasts. *ECMWF Newsletter*, **117**, 8–13.

Hagedorn, R. (2010). On the relative benefits of TIGGE multi-model forecasts and reforecast-calibrated EPS forecasts.

ECMWF Newsletter, **124**, 17–23.

Hagedorn, R., Hamill, T. M. and Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Monthly Weather Review*, **136**, 2608–2619.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M. and Palmer, T. N. (2012). Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1814–1827.

Haiden, T., Rodwell, M. J. and Richardson, D. S. (2012). Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Monthly Weather Review*, **140**, 2720–2733.

Hamill, T. M. (2012). Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Monthly Weather Review*, **140**, 2232–2252.

Hamill, T. M. and Colucci, S. J. (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, **125**, 1312–1327.

Hamill, T. M. and Colucci, S. J. (1998). Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, **126**, 711–724.

Hamill, T. M. and Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogues: Theory and application. *Monthly Weather Review*, **134**, 3209–3229.

Hamill, T. M. and Whitaker, J. S. (2007). Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. *Monthly Weather Review*, **135**, 3273–3280.

Hamill, T. M., Whitaker, J. S. and Wei, X. (2004). Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, **132**, 1434–1447.

Hamill, T. M., Whitaker, J. S. and Mullen, S. L. (2006). Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society*, **87**, 33–46.

Hamill, T. M., Hagedorn, R. and Whitaker, J. S. (2008). Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly Weather Review*, **136**, 2620–2632.

Ihász, I., Üveges, Z., Mile, M. and Németh, C. (2010). Ensemble calibration of ECMWF's medium-range forecasts. *Időjárás*, **114**, 275–286.

Johnson, A. and Wang, X. (2012). Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Monthly Weather Review*, **140**, 3054–3077.

Kann, A., Haiden, T. and Wittmann, C. (2011). Combining 2-m temperature nowcasting and short-range ensemble forecasting. *Nonlinear Processes in Geophysics*, **18**, 903–910.

Kann, A., Wittmann, C., Wang, Y. and Ma, X. (2009). Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Monthly Weather Review*, **137**, 3373–3387.

Kleiber, W., Raftery, A. E. and Gneiting, T. (2011a). Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *Journal of the American Statistical Association*, **106**, 1291–1303.

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F. and Grimit, E. P. (2011b). Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Monthly Weather Review*, **139**, 2630–2649.

Klein, W. H. and Glahn, H. R. (1974). Forecasting local weather by means of model output statistics. *Bulletin of the American Meteorological Society*, **55**, 1217–1227.

Kober, K., Craig, G. C., Keil, C. and Dörnbrack, A. (2012). Blending a probabilistic nowcast method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138**, 755–768.

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, **35**, 2739–2750.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, **249**, 2–9.

Krzysztofowicz, R. and Toth, Z. (2008). Bayesian processor of ensemble (BPE): Concept and implementation. Workshop slides, 4th NCEP/NWS Ensemble User Workshop, Laurel, Maryland, available at www.emc.ncep.noaa.gov/gmb/ens/ens2008/Krzysztofowicz_Presentation_Web.pdf.

Lalaurette, F. (2003). Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quarterly Journal of the Royal Meteorological Society*, **129**, 3037–3057.

Leutbecher, M. and Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, **227**, 3515–3539.

Little, M. A., McSharry, P. E. and Taylor, J. W. (2009). Generalized linear models for site-specific density forecasting of U.K. daily rainfall. *Monthly Weather Review*, **137**, 1029–1045.

Mass, C. F., Baars, J., Wedam, G. and Steed, R. (2008). Removal of systematic model bias on a grid. *Weather and Forecasting*, **23**, 438–459.

Mass, C., Joslyn, S., Pyle, J., Tewson, P., Gneiting, T., Raftery, A., Baars, J., Sloughter, J. M., Jones, D. and Fraley, C. (2009). PROBCAST: A web-based portal to mesoscale probabilistic forecasts. *Bulletin of the American Meteorological Society*, **90**, 1009–1014.

Möller, A., Lenkoski, A. and Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, **139**, 982–991.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, **122**, 73–119.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T. and Vitart, F. (2011). The new ECMWF seasonal forecast system (System 4). ECMWF Technical Memorandum 656.

Murphy, A. H. and Daan, H. (1985). Forecast evaluation. In *Probability, Statistics and Decision Making in the Atmospheric Sciences*, Murphy, A. H. and Katz, R. W., eds., Westview Press, Boulder, Colorado, pp. 379–437.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–1338.

National Research Council (2012). *Weather Services for the Nation: Becoming Second to None.* National Academies Press, Washington.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, **128**, 747–774.

Palmer, T. N., Doblas-Reyes, J., Weisheimer, A. and Rodwell, M. J. (2008). Toward seamless prediction: Calibration of climate change predictions using seasonal forecasts. *Bulletin of the American Meteorological Society*, **89**, 459–469.

Palmer, T. N., Buizza, R., Hagedorn, R., Leutbecher, M. and Smith, L. A. (2005/06). Ensemble prediction: A pedagogical perspective. *ECMWF Newsletter*, **106**, 10–17.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., Déqué, M., Diez, E., Doblas-Reyes, J., Feddersen, H., Graham, R., Gualdi, S. Guérémy, J.-F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A. P., Orfila, B., Rogel, P., Terres, J.-M. and Thomson, M. C. (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, **85**, 853–872.

Parrish, M. A., Moradkhani, H. and DeChant, C. M. (2012). Toward reduction of model uncertainty: Bayesian model averaging and data assimilation. *Water Resources Research*, **48**, W03519.

Persson, A. (2011). User Guide to ECMWF Products. Available at http://www.ecmwf.int/products/forecasts/guide/user_guide.pdf.

Petroliagis, T. I. and Pinson, P. (2012). Early warnings of extreme winds using the ECMWF extreme forecast index. *Meteorological Applications*, in press, DOI: 10.1002/met.1339.

Pinson, P. (2012). Adaptive calibration of (u,v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1273–1284.

Pinson, P. (2013). Wind energy: Forecasting challenges for its operational management. *Statistical Science*, in press.

Pinson, P. and Madsen, H. (2009). Ensemble-based probabilistic forecasting of wind power at Horns Rev. *Wind Energy*, **12**, 137–155.

Pinson, P., Madsen, H., Nielsen, H. A., Papaefthymiou, G. and Klöckl, B. (2009). From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, **12**, 51–62.

Prates, F. and Buizza, R. (2011). PRET, the Probability of RETurn: a new probabilistic product based on generalized extreme-value theory. *Quarterly Journal of the Royal Meteorological Society*, **137**, 521–537.

Raftery, A. E., Kárný, M. and Ettler, M. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, **52**, 52–66.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Reggiani, P., Renner, M., Weerts, A. H. and van Gelder, P. A. H. J. M. (2009). Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research*, **45**, W02428.

Renner, M., Werner, M. G. F., Rademacher, S. and Sprokkereef, E. (2009). Verification of ensemble flow forecasts for the river Rhine. *Journal of Hydrology*, **376**, 463–475.

Rings, J., Vrugt, J. A., Schoups, G., Huisman, J. A. and Vereecken, H. (2012). Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiment. *Water Resources Research*, **48**, W05520.

Richardson, D. S. (2000). Skill and economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **126**, 649–668.

Richardson, D. S. (2001). Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of sample size. *Quarterly Journal of the Royal Meteorological Society*, **127**, 2473–2489.

Richardson, D. (2012). Changes to the operational forecast system. *ECMWF Newsletter*, **132**, 2.

Richter, D. (2012). Bayesian ensemble model output statistics for temperature. Diploma thesis, Heidelberg University, Germany.

Rodwell, M. J. (2005/06). Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better. *ECMWF Newsletter*, **106**, 17–23.

Roquelaure, S. and Bergot, T. (2008). A local ensemble prediction system for fog and low clouds: Construction, Bayesian model averaging, calibration, and validation. *Journal of Applied Meteorology and Climatology*, **47**, 3072–3088.

Roulin, E. and Vannitsem, S. (2012). Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly Weather Review*, **140**, 874–888, **140**, 2018 (Corrigendum).

Roulston, M. S. and Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus Series A*, **55**, 16–30.

Ruiz, J. J. and Saulo, C. (2012). How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? Part I: Sensitivity to calibration methods. *Meteorological Applications*, **19**, 302–313.

Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X. and Seo, D. J. (2007). Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrology and Earth System Sciences*, **4**, 655–717.

Schefzik, R. (2011). Ensemble copula coupling. Diploma thesis, Heidelberg University, Germany.

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, **28**, 616–640.

Scheuerer, M. (2013). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, in press, DOI: 10.1002/qj.2183.

Scheuerer, M. and Büermann, L. (2013). Spatially adaptive post-processing of ensemble forecasts for temperature. Preprint, arXiv:1302.0883v1.

Schmeits, M. J., Kok, K. J., Daan, H., Vogelezang, P. and van Westrhenen, R. M. (2008). Probabilistic forecasts of (severe) thunderstorms for the purpose of issuing a weather alarm in the Netherlands. *Weather and Forecasting*, **23**, 1253–1267.

Schmeits, M. J. and Kok, K. J. (2010). A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review*, **138**, 4199–4211.

Schölzel, C. and Friederichs, P. (2008). Multivariate non-normally distributed random variables in climate research – introduction to the copula approach. *Nonlinear Processes in Geophysics*, **15**, 761–772.

Schölzel, C. and Hense, A. (2011). Probabilistic assessment of regional climate change in Southwest Germany by ensemble dressing. *Climate Dynamics*, **36**, 2003–2014.

Schuhen, N., Thorarinsdottir, T. L. and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Monthly Weather Review*, **140**, 3204–3219.

Sheridan, P., Smith, S., Brown, A. and Vosper, S. (2010). A simple height-based correction for temperature downscaling in complex terrain. *Meteorological Applications*, **17**, 329–339.

Sklar, A. (1959). Fonctions de répartition à *n* dimensions et leur marges. *Publications de'l Institut de Statistique de'l Université de Paris*, **8**, 229–231.

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135**, 3209–3220.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010). Probabilistic wind forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, **105**, 25–35.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review*, **141**, 2107–2119.

Soltanzadeh, I., Azadi, M. and Vakili, G. A. (2011). Using Bayesian model averaging (BMA) to calibrate probabilistic surface temperature forecasts over Iran. *Annales Geophysicae*, **29**, 1295–1303.

Strauch, M., Bernhofer, C., Koide, S., Volk, M., Lorz, C. and Makeschin, F. (2012). Using precipitation data ensemble for uncertainty analysis in SWAT streamflow simulation. *Journal of Hydrology*, **414–415**, 413–424.

Taylor, J. W., McSharry, P. E. and Buizza, R. (2009). Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*, **24**, 775–782.

Theis, S. E., Hense, A. and Damrath, U. (2005). Probabilistic precipitation forecasting from a deterministic model: A pragmatic approach. *Meteorological Applications*, **5**, 257–268.

Thielen, J., Schaake, J., Hartman, R. and Buizza, R. (2008). Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. *Atmospheric Science Letters*, **9**, 29–35.

Thielen, J., Bartholmes, J. C., Ramos, M.-H. and de Roo, A. (2009). The European Flood Awareness System EFAS — Part 1: Concept and development. *Hydrology and Earth System Sciences*, **13**, 125–140.

Thorarinsdottir, T. L. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Series A*, **173**, 371–388.

Thorarinsdottir, T. L. and Johnson, M. S. (2012). Probabilistic wind gust forecasting using nonhomogeneous Gaussian regression. *Monthly Weather Review*, **140**, 889–897.

Thorarinsdottir, T. L., Scheuerer, M. and Feldmann, K. (2012). Statistical post-processing of ensemble forecasts (in German with English summary). *Promet*, **37 (3/4)**, 43–52.

Traiteur, J. T., Callicutt, D. J., Smith, M. and Somnath, B. R. (2012). A short-term ensemble wind speed forecasting system for wind power applications. *Journal of Applied Meteorology and Climatology*, **51**, 1763–1774.

Unger, D. A., van den Dool, H., O'Lenic, E. and Collins, D. (2009). Ensemble regression. *Monthly Weather Review*, **137**,

2365–2379.

Vannitsem, S. (2009). A unified linear model output statistics scheme for both deterministic and ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **135**, 1801–1815.

Vannitsem, S. and Hagedorn, R. (2011). Ensemble forecast post-processing over Belgium: Comparison of deterministic-like and ensemble regression methods. *Meteorological Applications*, **18**, 94–104.

Vitart, F. (2013). Evolution of ECMWF sub-seasonal forecast skill scores over the past 10 years. ECMWF Technical Memorandum 694.

Voisin, N., Schaake, J.-C. and Lettenmeier, D. P. (2010). Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Weather and Forecasting*, **25**, 1603–1627.

Voisin, N., Pappenberger, F., Lettenmeier, D. P., Buizza, R. and Schaake, J. C. (2011). Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River basin. *Weather and Forecasting*, **26**, 425–446.

Vrugt, J. A. and Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, **43**, W01411.

Vrugt, J. A., Clark, M. P., Diks, C. G. H., Duan, Q. and Robinson, B. A. (2006). Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophysical Review Letters*, **33**, L19817.

Walker, A. M. (1968). A note on the asymptotic distribution of sample quantiles. *Journal of the Royal Statistical Society Series B*, **30**, 570–575.

Wang, X. and Bishop, C. H. (2005). Improvement of ensemble reality with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, **131**, 965–986.

Whitaker, J. S., Wei, X. and Vitart, F. (2006). Improving week-2 forecasts with multimodel reforecast ensembles. *Monthly Weather Review*, **134**, 2279–2284.

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

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*, third edition. Elsevier Academic Press, 676 pp.

Wilks, D. S. and Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, **135**, 2379–2390.

Wilson, L. J. and Vallée, M. (2002). The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Weather and Forecasting*, **17**, 206–222.

Wilson, L. J. and Vallée, M. (2003). The Canadian updateable model output statistics (UMOS) system: Validation against perfect prog. *Weather and Forecasting*, **18**, 288–302.

Wilson, L. J., Beauregard, S., Raftery, A. E. and Verret, R. (2007). Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Monthly Weather Review*, **135**, 1364–1385.

Wood, A. W. and Schaake, J. C. (2008). Correcting errors in streamflow forecast ensemble mean and spread. *Journal of Hydrometeorology*, **9**, 132–148.

Wu, W. B. (2005). On the Bahadur representation of sample quantiles for dependent sequences. *Annals of Statistics*, **33**, 1934–1963.

Zappa, M., Jaun, S., Germann, U., Walser, A. and Fundel, F. (2011). Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research*, **100**, 246–262.

Zsoter, E., Pappenberger, F. and Richardson, D. (2014). Sensitivity of climate configurations and the impact on the extreme forecast index. *Meteorological Applications*, in press.