BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 76 (2017) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Zied Ben Bouallegue

Verification and post-processing of ensemble weather forecasts for renewable energy applications

BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 76 (2017) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Zied Ben Bouallegue

Verification and post-processing of ensemble weather forecasts for renewable energy applications

Verification and post-processing of ensemble weather forecasts for renewable energy applications

DISSERTATION
ZUR
ERLANGUNG DES DOKTORGRADES (DR. RER. NAT.)
DER
MATHEMATISCH-NATURWISSENSCHAFTLICHEN FAKULTÄT
DER
RHEINISCHEN FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

vorgelegt von
Dipl.-Ing. Zied Ben Bouallegue
aus
Tunis

Bonn, Juni 2016

Diese Arbeit ist die ungekürzte Fassung einer der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn im Jahr 2016 vorgelegten Dissertation von Zied Ben Bouallegue aus Tunis.

This paper is the unabridged version of a dissertation thesis submitted by Zied Ben Bouallegue born in Tunis to the Faculty of Mathematical and Natural Sciences of the Rheinische Friedrich-Wilhelms-Universität Bonn in 2016.

Anschrift des Verfassers: Address of the author:

Zied Ben Bouallegue
Deutscher Wetterdienst
Frankfurter Straße 135
D-63067 Offenbach am Main

1. Gutachter: Prof. Dr. Andreas Hense, Rheinische Friedrich-Wilhelms-Universität Bonn
2. Gutachter: Prof. Dr. Pierre Pinson, Technical University of Denmark

Tag der Promotion: 08. Februar 2017

Für Anne-Sophie,
Naïm und Maya

"Here come the sun
Aren't you glad to see it
I say, It's all right"

Nina Simone, 1971

Abstract

The energy transition taking place in Germany encourages a large scale penetration of weather-dependent energy sources into the power grid. The grid integration of intermittent sources increases the need for balancing demand and supply in order to ensure the reliability and safety of the power system. In this context, forecasts are essential for the cost-effective management of reserves and trading activities. Solar and wind power forecasts with a time horizon of few hours up to several days are usually based on outputs of numerical weather prediction systems routinely provided by weather centres. At the German Weather Service, the high-resolution ensemble prediction system COSMO-DE-EPS is called to support renewable energy applications which require dealing with the intermittency and uncertainty in the energy production. In this study, ensemble forecast verification and post-processing are addressed focusing on global radiation, which is the main weather variable affecting solar power production.

First, the ensemble forecast performances are assessed from the user's and developer's perspectives. New tools are proposed for the verification of quantile forecasts which are probabilistic products appropriate for many renewable energy applications. Forecast discrimination ability and value are assessed considering users with different aversions to under- and over-forecasting. Moreover, a new measure is introduced in order to summarize the added value of the ensemble approach with respect to a single run approach. The new skill score is conditioned on calibration, that is, statistical consistency between the distributional forecasts and observations. Second, an enhanced framework for the post-processing of ensemble forecasts is proposed. The aim is to provide the users with calibrated consistent scenarios which are required for the optimization of complex decision-making processes. Therefore, a two-step procedure is developed starting with the marginal calibration of the forecasts based on quantile regression and the selection of appropriate predictors. Next, consistent scenarios are generated using a dual ensemble copula coupling approach which combines information from past error statistics and the dependence structure in the original ensemble forecast.

Contents

1. Introduction

Renewable energies are the cornerstone of the energy transition taking place in Germany. Guided by motivations rooted in ecology, geopolitics, and socio-economics, the *Energiewende* aims at more sustainability in a broad sense. Evidence of anthropogenic climate changes has led to the development of effective decarbonized energy supplies, while geopolitical instability concurrently encourages measures to ensure energy security by reducing the dependency on energy imports. Taking into consideration the risk associated with nuclear power and the societal anxiety generated by the disaster of Fukushima, the energy transition is also seen as a chance to stimulate scientific and technical innovations.

During the last two decades, solar and wind capacities installed in Germany have increased exponentially. In 2014, renewable energy as a whole reached $\sim$31% of the net electricity consumption. In particular, photovoltaic (PV) generated power covered on average $\sim$7% of the needs[1], while on sunny weekends, PV power could at times cover up to half of the momentary electricity demand (Wirth, 2015). The variability in electricity production and therefore the ability of renewable energy supplies to meet the demand is directly related to their weather-dependent nature. Figure 1.1 illustrates the power production from conventional and renewable energy sources over two consecutive days. Wind and solar productions exhibit a variability related to the concomitant weather conditions. A fundamental characteristic of renewables is that this variability cannot be known with certainty beforehand.

With a large scale penetration of renewable energies, the electric power system has evolved in order to account for intermittency and stochasticity in power generation (Boyle, 2008; Gross *et al.*, 2008; Morales *et al.*, 2014; Troccoli *et al.*, 2014). In the front line, transmission and distribution system operators, which are responsible for the safety and stability of the power grid, have to ensure that the total capacity available on the grid always meets the demand. The continuous balancing of demand and supply is facilitated by a greater flexibility of the system. Scheduling and dispatch of power units are coordinated and balancing reserves help to deal with unexpected fluctuations. With regards to the energy pricing, the electricity market is influenced by renewable outputs since power units from renewable sources are usually scheduled before conventional ones. Technical innovations are also expected to help face the challenge of variability with, for example, the introduction of storage capacities or the increase of the demand-side flexibility.

In any case, information about the expected power production is required for an optimal integration of intermittent energy sources. In terms of reserve management, forecasts help reduce the need for expensive regulating reserves, thereby reducing costs related to balancing the system (Bird *et al.*, 2013). Forecasts of power ramps, that is abrupt changes in power production, are of particular relevance for the dimensioning of backup energy sources. In terms of energy market operations, power production is contractually committed on the day-ahead market and can be adjusted in the intraday

[1] thanks to over 1.5 million power plants distributed over the country

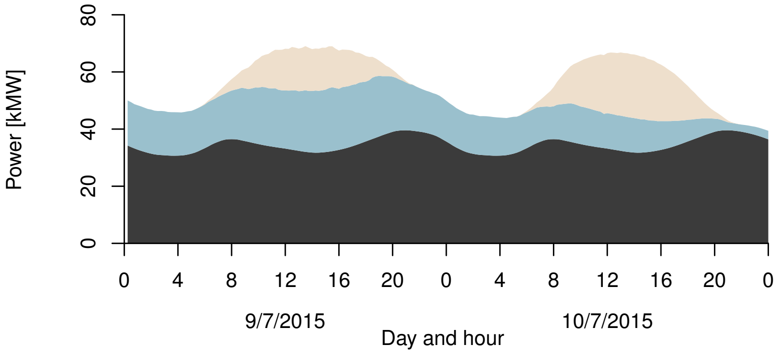

Figure 1.1: Conventional power production (dark grey), wind power production (blue) and solar production (yellow) on July 9 and 10, 2015[2].

markets. Remaining imbalances are treated on the real-time market using balancing power, which has much higher costs than regular trading.

Forecasts of power production at various temporal and spatial scales are therefore essential for an efficient and cost-effective integration of intermittent sources in the power grid. Moreover, the limited *predictability* of outputs from solar and wind sources favours a probabilistic approach targeted at the optimisation of decision-making under uncertainty. Numerous methods have been developed for the prediction of wind and solar power productions (Costa *et al.*, 2008; Pelland *et al.*, 2013). Wind forecasting has a longer tradition than solar forecasting, which has gathered attention more recently, but from the mathematical and modelling point of view, both share a high degree of similarity.

Focusing on solar forecasting, a common classification distinguishes between physical, statistical, and hybrid methods depending on whether the forecast is based on a solar/PV model, historical data, or the combination of both (Saint-Drenan *et al.*, 2015). Typically, a physical PV model describes solar power production based on the characteristics of the PV modules and prediction of solar irradiance at ground and ambient temperature. More generally, solar radiation is the essential component in most PV power prediction systems and the type of auxiliary data used as primary source of information characterize the range of applications (Lorenz *et al.*, 2014). Intra-hour forecasts of solar radiation can be generated by processing sky images at high frequency from ground-based total sky imager or can rely on satellite images (Chow *et al.*, 2011; Hammer *et al.*, 2003; Perez *et al.*, 2010). Based on direct measurements of PV outputs, statistical techniques, such as time series modelling or artificial neural networks, have demonstrated to be competitive for short term and very-short term forecasts (Kalogirou, 2001; Mellit, 2008).

For longer time horizons, from few hours to several days, applications usually require Numerical Weather Prediction (NWP) forecasts (Lorenz *et al.*, 2011; Perez *et al.*, 2013; Zamo *et al.*, 2014a; Thorey *et al.*, 2015). NWP forecasts are computer simulations based on the primitive equations that describe the physical laws of fluid dynamics and thermodynamics (Bjerknes, 1904). Forecasts are derived by integrating numerically the partial differential equations, starting from the observed current weather situation. The initial state of the atmosphere is estimated by data assimilation techniques that combine first guess forecasts and meteorological observations. The resolution of the primitive equations calls a discretisation in the three spatial dimensions and in time, and hence empirical parametrisa-

[2] data source: Energy Exchange Leipzig EEX (*www.eex-transparency.com*).

tions of the processes on the unresolved spatial and temporal scales.

NWP model outputs are routinely provided by weather centres. The German Weather Service (DWD) operationally runs, among others, a high-resolution model centred over Germany called COSMO-DE (Baldauf *et al.*, 2011). With a spatial grid resolution of 2.8 km, the model explicitly represents small scale processes such as deep convection. The increase in renewable energy applications has led to enhanced attention to weather variables relevant for the energy sector. Since 2012, the improvement of the NWP forecast skill focusing on typical weather variables such as wind and global radiation is on the DWD's agenda[3] (Hagedorn *et al.*, 2015).

Despite the continuous improvement of their performances, NWP models remain subject to errors (Bauer *et al.*, 2015). The discretisation of the equations, the parametrisation of the model and the imperfect description of the initial conditions are sources of forecast errors. Beside the epistemic uncertainty, aleatoric uncertainty related to atmospheric chaos limits the predictability skill of the forecast, though the predictability itself is both variable and predictable (Lorenz, 1969; Slingo and Palmer, 2011; Palmer *et al.*, 2014).

Information about the uncertainty associated with a forecast is of high relevance for the users and needs to be assessed. In NWP, running an Ensemble Prediction System (EPS) has become a standard approach for providing a flow-dependent assessment of the forecast uncertainty. An EPS consists in running a NWP model several times with variations that account for model error sources. At DWD, COSMO-DE-EPS is the operational ensemble system based on COSMO-DE (Gebhardt *et al.*, 2011; Peralta *et al.*, 2012). The 20-member ensemble arises from variations in the initial and boundary conditions, and from model perturbations. Probabilistic forecasting requires the interpretation of the ensemble forecast in probabilistic terms. The derived probabilistic products are targeted at users, provided as a basis for their decision-making processes.

Probabilistic forecasts based on NWP forecasts can also be derived by a variety of other approaches. Rooted in model output statistics (Glahn and Lowry, 1972), statistical methods have emerged as powerful tools that draw probabilistic information based on historical data. Appropriate statistical techniques have been successfully applied to wind and solar power forecasting (Bremnes, 2004; Zamo *et al.*, 2014a). Other approaches are pragmatic, like the *neighbourhood method*, which builds a forecast sample from neighbouring forecasts (Theis *et al.*, 2005), and the *lagged average* approach, which gathers forecasts from different starting times but with overlapping verification periods (Hoffman and Kalnay, 1983). Alternatively, *analogue* ensemble forecasts can be composed from past observations whose related past forecasts share analogies with the forecast under focus (Hamill and Whitaker, 2006; Delle Monache *et al.*, 2013).

The different approaches for providing probabilistic information based on NWP forecasts are complementary in most cases, thus ensemble forecasting can be combined with other techniques. Computationally inexpensive, the neighbourhood method and the lagged averaged forecasting allow an increase of the ensemble size and have demonstrated that they bring noticeable improvement in terms of forecast skill (Schwartz *et al.*, 2010; Ben Bouallègue *et al.*, 2013; Raynaud *et al.*, 2015). Analogue-based post-processing of ensemble forecasts has also been explored recently (Junk *et al.*, 2015). More generally, post-processing of ensemble forecasts based on historical dataset enables bias correction and provides reliable probabilistic products. Therefore, post-processing is usually considered a necessary step in order to fully benefit from the ensemble approach (Gneiting *et al.*, 2007).

In this study, the ability of an ensemble prediction system to support renewable energy applications is addressed. More precisely, this work investigates global (direct + diffuse) radiation ensemble

[3] the investigations presented in this study have been performed in the framework of the EWeLiNE project (*http://projekt-eweline.de*)

forecasts from the operational system COSMO-DE-EPS. Today, ensemble forecasting is the state-of-the-art method for providing probabilistic weather forecasts, yet there is a need:

- to assess the value of probabilistic products relevant for the users,
- to assess the benefit of the ensemble approach compared to less expensive techniques.

In this context, probabilistic forecasts based on a single run are considered as a relevant benchmark. Moreover, in order to fully benefit from the ensemble forecast, statistical inconsistencies have to be corrected in order to provide the users with reliable probabilistic products. At the same time, forecast scenarios, which are physically consistent over time, space and weather variables have to be delivered in order to cover a full range of users' applications. Therefore, there is a need:

- to calibrate the ensemble and provide reliable forecasts at each location and forecast horizon,
- to generate consistent scenarios based on the calibrated ensemble forecasts.

In this framework, a computationally effective two-step procedure is followed. Innovative approaches are proposed and here applied to intra-day and day-ahead global radiation forecasts.

In Chapter 2, the concept and motivations leading to ensemble forecasting are recalled. The operational setup of the ensemble system COSMO-DE-EPS is described as well as the observation dataset used in this study. The interpretation of the ensemble forecast in terms of probabilistic products is discussed with an emphasis on quantile forecasts, which are key products for renewable energy applications.

In Chapter 3, fundamental concepts for the verification of probabilistic forecasts are introduced. Proper scoring rules and their decomposition lead to the discussion on the ensemble performance in terms of statistical consistency and information content. Focusing on the latter, quantile forecasts used as decision variables are assessed in a decision making framework. Ensemble-derived forecasts and deterministic forecasts are then compared in terms of forecast *value*. Taking a developer's perspective, the benefit of using the ensemble approach rather than a single forecast is further depicted with a summary measure.

In Chapter 4, statistical methods for the correction of systematic errors based on past data are discussed. Quantile regression is shown to be a well-suited method for the calibration of global radiation ensemble forecasts. Moreover, a *regularization* of the regression scheme is used in order to select adequate predictors from a pool of weather model outputs. The *standard* and *weather-dependent* calibration approaches are compared when applied to ensemble and deterministic forecasts. In the ensemble case, forecast dependence structures existing in the original ensemble can be preserved after calibration. Generation of scenarios based on the original ensemble forecasts and the calibrated marginals is discussed and a new method is proposed and implemented.

Finally, Chapter 5 presents the main findings of this work, while the appendices compile the related original studies entitled:

A. Quantile forecast discrimination ability and value.

B. Assessment and added value estimation of an ensemble approach with a focus on global radiation forecasts.

C. Statistical post-processing of global radiation ensemble forecasts with penalized quantile regression.

D. Generation of scenarios from calibrated ensemble forecasts with a dual ensemble copula coupling approach.

2. Ensemble forecasting

Weather predictability strongly affects management activities in the energy sector. In NWP, predictability is shaped by the uncertainty intrinsically associated with a forecast. Approximations in the description of the initial conditions, the model discretisation in time and space, and the closure of the equations are sources of errors. Moreover, the atmosphere is a chaotic system by nature: small uncertainties in the initial conditions can grow rapidly at small scales and spread to upper scales (Lorenz, 1969). Since no deterministic solution exists for such non-linear systems, a stochastic-dynamic view on the forecasting problem is required: a weather variable is no longer regarded as a deterministic variable but as a random variable with associated stochastic properties described by a probability distribution (Epstein, 1969a). Hence, numerical weather forecasts are regarded as probabilistic both in the production phase, by the model developers, and in the operation phase, with the dissemination of probabilistic products.

The evolution of probability distributions in the phase space can theoretically be described by the Liouville's equation (the continuous equation for probabilities, which accounts for model non-linearities and imperfect initial state, e.g. Ehrendorfer, 1997). However, the dimensionality of a weather prediction system prevents using this approach for the description of the atmospheric state in a probabilistic framework. As an alternative, a Monte Carlo approach can provide a limited sample of realisations that represents the predictive probability distribution. The forecast sample is obtained running several times a numerical model accounting for uncertainty in the initial conditions and physic parametrisations. In NWP, this approach is known as *ensemble* technique, and each single run is called ensemble *member* (Epstein, 1969a; Leith, 1974). Thus, an ensemble system provides a range of possible outcomes which aims at capturing the flow-dependent uncertainty of the forecast (Palmer, 2000; Zhu, 2005).

Based on an ensemble forecast, each user can assess the level of confidence in the final forecast and the related forecast uncertainty can be quantified. For example, the probability of occurrence of any discrete event can be estimated. The quantification of the forecast uncertainty provides a probabilistic guidance to the user. Thereby, a framework is proposed for decision-making where appropriate action can be taken by the users according to their own level of risk and degree of vulnerability (Slingo and Palmer, 2011).

2.1 Ensemble prediction systems

Benefiting from advances in computing sciences and technology, numerical Ensemble Prediction Systems (EPS) have become a state-of-the-art technique in numerical weather forecasting. Tracked back to the 1950s, ensemble systems are operational since the 1990s and run nowadays at various spatial and temporal scales (Lewis, 2005; Tracton and Kalnay, 1993; Houtekamer *et al.*, 1996; Molteni *et al.*, 1996; Bowler *et al.*, 2008; Montani *et al.*, 2011; Lewis, 2014). An ensemble can arise as the combination

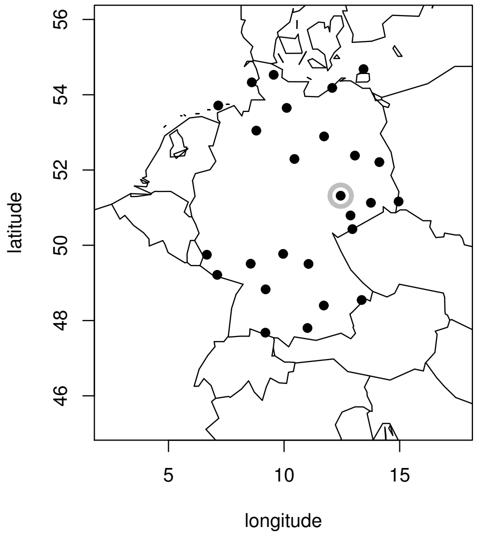

Figure 2.1: Model domain of COSMO-DE and pyranometer stations used for verification and postprocessing purposes. The station *Leipzig* is highlighted with a grey circle.

of different NWP models (multi-model ensembles) or can be based on a single model with different setups. Several techniques allow integrating uncertainty in the initial conditions and in the physic parametrisations (Leutbecher and Palmer, 2008). In contrast to global ensemble systems, limited-area ensemble systems have also to account for uncertainty in the boundary conditions.

At the German Weather Service (DWD), the high-resolution ensemble system COSMO-DE-EPS has been running operationally since May 2012. The Consortium for Small-scale Modeling (COSMO[1]) provides a non-hydrostatic, limited-area model whose flexibility allows for a wide range of applications. The ensemble system is based on the convection-permitting COSMO-DE model, which is operationally run at DWD (Baldauf *et al.*, 2011). The model has a model grid size of 2.8 km and the radiation scheme follows Ritter and Geleyn (1992). The model domain covers Germany and part of the neighbouring countries as shown in Figure 2.1.

Based on a multi-model and multi-parameter approach, COSMO-DE-EPS comprises 20 members with variations of the boundary and initial conditions as well as model perturbations. A detailled description of the ensemble setup can be found in Peralta *et al.* (2012). The combination of model variations is summarised in Table 2.1. Variations of the boundary and initial conditions are obtained from four different global models: the Integrated Forecasting System (IFS) of the European Centre for Medium range Weather Forecast (ECMWF), the global model GME[2] of DWD, the Global Forecast System (GFS) of the National Centres for Environmental Prediction (NCEP), and the Global Spectral Model (GSM) of the Japanese Meteorological Agency (JMA). Additionally, 5 different perturbations are applied, each to forecasts driven by the different global models, and kept constant over the integration time.

In its operational version, COSMO-DE-EPS is run 8 times a day, at 00, 03,..., 18, and, 21 UTC, with a forecast horizon of 27 hours. Targeted to renewable energy applications, the forecast range of the

[1] http://www.cosmo-model.org
[2] ICON since January 2015.

physics variation\driving boundary model	IFS	GME	GFS	GSM
Entrainment rate for shallow convection	1	6	11	16
Critical value for normalized over-saturation	2	7	12	17
Scaling factor boundary layer for heat (<default)	3	8	13	18
Scaling factor boundary layer for heat (>default)	4	9	14	19
Asymptotic mixing length of turbulence	5	10	15	20

Table 2.1: Configuration of the COSMO-DE-EPS setup showing the combination of 4 global models and 5 physic perturbed parameters leading to the 20 ensemble members.

03 UTC run has been extended up a horizon of 45 hours. Thus, day-ahead power forecasts based on COSMO-DE-EPS can be available before the closing of the energy market at 12 UTC. In the remainder of the manuscript, the weather variable under focus is global radiation at ground level, the forecasts of which correspond to the sum of two model outputs: the direct and diffuse short-wave radiations. The innovative techniques introduced in this study can however be applied to other weather variables, as for example wind forecasts (see Appendix D).

In order to assess the performance of the forecast and to apply post-processing techniques based on historical data, high quality observations are required. Observations of solar radiation are provided by quality controlled measurements from pyranometer stations distributed over Germany (Becker and Behrens, 2012). Hourly averaged observations and forecasts are compared with a temporal resolution of one hour. The test period used here for illustrative purposes corresponds to July and August 2015. During this period, 24 stations provided measurements on a regular basis (see Figure 2.1).

2.2 Probabilistic products

Probabilistic products are derived from the ensemble forecast in order to provide uncertainty information along with the forecast. This step is usually referred to as *ensemble interpretation* and consists in considering an ensemble forecast as drawn from a probability distribution (Bröcker and Smith, 2008). Properties of the predictive distribution are estimated and communicated to the users. For example, at a given forecast horizon and location, the ensemble mean is an estimation of the most likely outcome and the ensemble spread, an estimation of the associated uncertainty (Zhu, 2005; Grimit and Mass, 2007; Hopson, 2014). Other functionals of the predictive distribution are estimated focusing on particular events or targeted at users with a specific level of risk adversity. In that case, a forecast takes the form of a threshold exceedance probability, and the form of a quantile, respectively.

For a formal definition of probabilistic products, the following notations are considered. First, the quantity to be forecast is intended as a continuous random variable denoted $Y = \{(y, F_Y(y), y \in \mathbb{R})\}$. The associated cumulative distribution $F_Y(y)$ follows $F_Y(y) = Pr(Y \leq y)$ where Pr denotes the probability. An observed event is defined by a threshold $\psi \in \mathbb{R}$ as $E : Y \geq \psi$. Second, the forecast for Y is assumed to take the form of a predictive cumulative distribution $F_X(x)$. The exceedance probability forecast p_ψ is defined as:

$$p_\psi = 1 - F_X(\psi), \tag{2.1}$$

while the quantile forecast at probability level τ with $0 \leq \tau \leq 1$ is defined as:

$$q_\tau = F_X^{-1}(\tau) = \inf\{x : F_X(x) \geq \tau\}. \tag{2.2}$$

Exceedance probability and τ-quantile correspond to two particular points of the predictive distribution as illustrated in Figure A.1.

Probabilistic products are delivered to users so they can use them in their decision-making processes. In this study, the focus is set on quantile forecasts that are optimal point forecasts for users with an asymmetric loss functions (Koenker and Machado, 1999; Friederichs and Hense, 2007; Gneiting, 2011a). In particular, quantile forecasts are of high relevance for renewable energy applications (Pinson *et al.*, 2007; Pinson, 2013; Morales *et al.*, 2014). Asymmetric loss functions are indeed associated with users with different sensitivity to under- and over-forecasting which is typically the case for applications related to reserve management and trading activities. Scheduling operating reserves has a cost which implies that over-forecasting has to be avoided, but activating reserves short before term is more expensive which means that under-forecasting has a higher cost. Similarly, market participants are penalized differently when committed power is under- or over-estimated. In all these cases, the user's optimal forecast corresponds to a specific quantile of the predictive distribution where the probability level is defined by the user's cost-loss ratio.

In the following applications, the COSMO-DE-EPS global radiation forecast is generally interpreted in terms of quantiles following a non-parametric approach. Considering the ensemble forecasts $(x^{(1)}, x^{(2)}, ..., x^{(M)})$ with M the ensemble size, an ensemble member $x^{(m)}$ can be interpreted as a quantile forecast considering its rank within the sample. Assuming that the sample is ordered, a quantile forecast q_{τ_m} can be estimated as:

$$q_{\tau_m} = x^{(m)}, \tag{2.3}$$

where τ_m the probability level associated with the member of rank $m \in 1, ..., M$ is here defined as:

$$\tau_m = \frac{m - 0.5}{M}. \tag{2.4}$$

This definition of τ_m as a function of the member rank m and the ensemble size M is consistent with the ensemble interpretation applied for the computation of the continuous ranked probability score, a common ensemble verification score (Bröcker, 2012, see Chapter 3). However, the definition of τ_m in Eq. (2.4) is not consistent with a reliability test like for example the rank histogram, which assumes that the probability masses between two consecutive ensemble members as well as below $x^{(1)}$ and above $x^{(M)}$ are all equal. In this case, the probability level associated to a ranked member $x^{(m)}$ follows $\tau_m = \frac{m}{M+1}$. More generally, quantiles at probability levels that are not comprised in the set defined in Eqs (2.4) can be derived for example by interpolation (Hyndman and Fan, 1996).

Figure 2.2(a) shows an example of COSMO-DE-EPS global radiation forecasts and the corresponding station measurements. This example covers 2 consecutive days and illustrates the typical diurnal cycle associated with this weather variable. In Figure 2.2(b), the ensemble is interpreted in terms of quantile forecasts as a function of the forecast horizon. Quantiles are provided at probability levels ranging between 10% and 90% with a 10% interval. The verification of these probabilistic products is discussed in the next chapter. A subjective assessment already suggests that the ensemble variability is not able to fully capture the observation variability. Appropriate techniques can correct for this type of ensemble drawback in order to provide *calibrated* quantile forecasts and to generate *consistent* scenarios as shown in Figure 2.2(c) and (d), respectively. These post-processing techniques are discussed in Chapter 4.

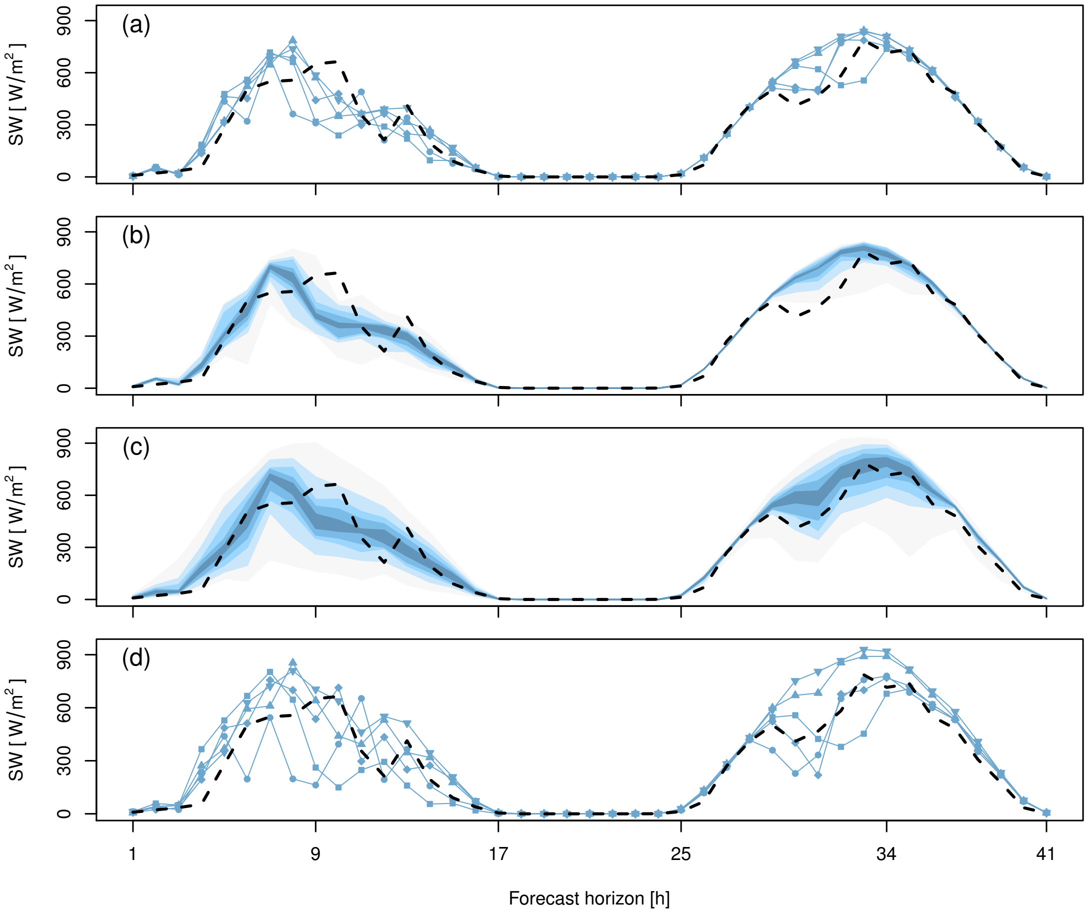

Figure 2.2: Example of ensemble forecasts, ensemble interpretation, and ensemble post-processing, based on COSMO-DE-EPS global radiation forecasts of the 03UTC run on July 9, 2015, valid at station Leipzig: (a) 5 of the 20 ensemble members, (b) the ensemble forecasts interpreted in terms of quantile forecasts at probability levels 10%, 20%, ..., 80%, 90%, (c) quantile forecasts calibrated by penalized quantile regression (see Section 4.2) at the same probability levels, (d) 5 of the 20 calibrated scenarios generated with a dual ensemble copula coupling approach (see Section 4.4) associated to the 5 ensemble forecasts shown in (a). The forecasts are plotted as a function of the forecast horizon. The corresponding pyranometer measurements are represented by a black dashed line.

3. Ensemble verification

Using the words of Atger (2004), the verification of ensemble-based probabilistic forecasts follows two goals: the validation of the system and the evaluation of end-product performance. Here, these two aspects are investigated for global radiation forecasts from COSMO-DE-EPS. The aim is to demonstrate the benefit of the ensemble approach and to provide evidence of the ensemble forecast deficiencies. In terms of products, the ensemble performance is examined with a focus on quantile forecasts, which are relevant probabilistic products for renewable energy applications. A general framework is first defined in order to describe appropriate tools and discuss their interpretation[1].

Forecast verification is here intended as the process of analysing the joint probability distribution of the forecast and the corresponding observation, where both observation and forecast are treated as random variables (Murphy and Winkler, 1987; Murphy, 1993). Considering the observation $Y = \{(y, F_Y(y), y \in \mathbb{R})\}$ and the forecast $X = \{(x, F_X(x), x \in \mathbb{R})\}$, the joint distribution is denoted $F_{YX}(y, x)$. The verification process consists in reducing the analysis of $F_{YX}(y, x)$ to a single dimension. Scoring rules, and more generally verification measures, try to summarise in the form of a single value some aspects of the forecast performance that necessarily implies to concurrently discard information about the joint distribution (Wilks, 2006b).

Particular aspects of the forecast performance, often referred to as forecast attributes, can be drawn from the investigation of the properties of the joint distribution (Murphy and Winkler, 1987). Following the multiplicative law of probability, two manipulations of the joint distribution can be applied for this purpose. The *calibration-refinement* factorisation,

$$F_{YX}(y, x) = F_Y(y \mid X = x) F_X(x), \tag{3.1}$$

conditions the observation on the forecast, whereas the *likelihood-base rate* factorisation,

$$F_{YX}(y, x) = F_X(x \mid Y = y) F_Y(y), \tag{3.2}$$

conditions the forecast on the observation. From these two factorisations, one can estimate attributes of the forecast performance such as *reliability*, *resolution*, and *discrimination*.

Adequate scores and validation tools have been developed for the verification of probabilistic products in the form of a predictive probability distribution, a probability forecast or a quantile forecast. The decomposition of scores based on the calibration refinement factorisation provides a framework for a detailled interpretation of the results in terms of forecast statistical inconsistency and forecast information content, which are the two fundamental aspects of the forecast performance. Based on the likelihood-base rate factorisation, the information content can be further related to the forecast value, which reflects the point of view of the user on the forecast performance (Chen *et al.*, 1987;

[1] the reader is invited to refer also to Wilks (2006b) and Jolliffe and Stephenson (2011)

Buizza, 2001). Finally, relationships between forecast value and scoring rules complete the verification picture (Murphy, 1969; Richardson, 2011). Thereby, the forecast ability to capture and resolve the observations can be assessed in a cascading process from the general to the specific forecast skill.

3.1 Proper scoring rules

Scoring rules are mathematical tools dedicated to the evaluation of probabilistic forecasts. A scoring rule measures the accuracy of a probabilistic prediction assigning a numerical score as a function of the predictive distribution and the event that materialised (Gneiting and Raftery, 2007). Historically, the Brier score is the first scoring rule that has been proposed in the context of probabilistic verification (Brier, 1950). Today, this verification measure is commonly used for the assessment of a forecast expressed in terms of a probability for a discrete dichotomous event.

A fundamental characteristic of a score is its *propriety* (Bröcker and Smith, 2007). A scoring function is called strictly proper when *'its expectation is optimal if and only if the forecast probability represents the true distribution of the target'* (Bröcker, 2009). In other words, a forecaster optimises the expected score by issuing his truth believes, avoiding therefore hedging (Murphy and Epstein, 1967). For example, propriety is a characteristic of the Brier score (Murphy, 1973).

In order to provide a formal definition of a proper score, let's note s a score of a probabilistic forecast $P \in \mathcal{P}$ and the corresponding observation $\omega \in \Omega$. The score s is in the following considered as negatively oriented (the smaller the better) and can therefore be intended as a cost function, which is aimed to be minimized (Bentzien and Friederichs, 2014). Determined by the joint distribution $F(\omega, P)$, the expected overall score is:

$$\mathbf{S} = \int_{\mathcal{P}} \int_{\Omega} s(\omega, P) F(\omega, P) \mathrm{d}\omega \mathrm{d}P \tag{3.3}$$

Applying the calibration-refinement factorisation following Eq. (3.1), the score is written as follows:

$$\mathbf{S} = \int_{\mathcal{P}} \int_{\Omega} s(\omega, P) F(\omega \mid P) F(P) \mathrm{d}\omega \mathrm{d}P \tag{3.4}$$

Denoting $Q(\omega) = F(\omega \mid P)$ the expected distribution of ω for a fixed probabilistic forecast P, Eq. (3.4) is developed as:

$$\begin{aligned} \mathbf{S} &= \int_{\mathcal{P}} \int_{\Omega} s(\omega, P) Q(\omega) F(P) \mathrm{d}\omega \mathrm{d}P \\ &= \int_{\mathcal{P}} S(Q, P) F(P) \mathrm{d}P \end{aligned} \tag{3.5}$$

where

$$S(Q, P) = \int_{\Omega} s(\omega, P) Q(\omega) \mathrm{d}\omega \tag{3.6}$$

is the expected score given a forecast P. By definition (Gneiting and Raftery, 2007), the score function S is proper if:

$$S(Q, Q) \leq S(Q, P) \tag{3.7}$$

for all P. It is strictly proper if equality is given if and only if $P = Q$.

Proper scoring rules also include the Quantile Score (QS), the natural score for the assessment of quantile forecasts (Koenker and Bassett, 1978; Friederichs and Hense, 2007; Bentzien and Friederichs, 2014). QS is based on an asymmetric piecewise linear function called check function and noted

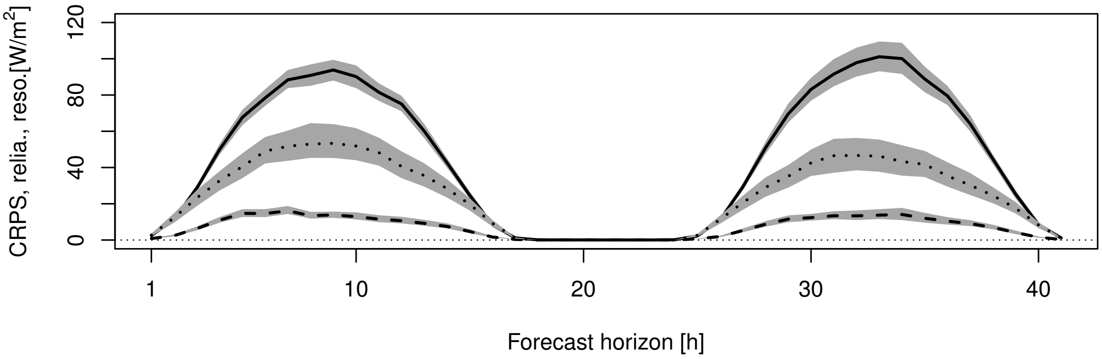

Figure 3.1: CRPS (full line), CRPS reliability (dashed line), and resolution components (dotted line) as a function of the forecast horizon. The confidence intervals are estimated by block bootstrapping. Results for the period July-August 2015.

ρ. Considering the pairs of quantile forecasts $q_{\tau,i}$ at probability level τ and observations y_i of the verification sample with $i \in 1, ..., N$, QS is defined as:

$$\mathrm{QS}_\tau = \frac{1}{N} \sum_{i=1}^{N} \rho_\tau (y_i - q_{\tau,i}) \tag{3.8}$$

with

$$\rho_\tau(u) = \begin{cases} \tau u & \text{if} \quad u \geq 0 \\ (\tau - 1)u & \text{if} \quad u < 0. \end{cases} \tag{3.9}$$

The asymmetry of the loss function is related to the probability level of the quantile forecast under assessment. Both are defined by τ which makes the link between user and probabilistic product explicit.

Each ensemble member can be interpreted as a quantile considering its rank within the ensemble sample (see Section 2.2). Therefore, the assessment of an ensemble forecast as a whole can consist in applying QS to each quantile defined by the ensemble size. This approach leads to the estimation of the continuous ranked probability score (CRPS), which is a common proper scoring rule for the verification of predictive density functions (Matheson and Winkler, 1976; Bouttier, 1994; Hersbach, 2000; Gneiting and Raftery, 2007). In the ensemble case, the CRPS follows:

$$\mathrm{CRPS} = \frac{2}{M} \sum_{m=1}^{M} \mathrm{QS}_{\tau_m} \tag{3.10}$$

where M is the ensemble size and τ_m the probability level defined in Eq. (2.4) (Bröcker, 2012). In the continuous case, the CRPS corresponds to the integral of QS over all probability levels or equivalently to the integral of the Brier score over all possible thresholds.

In Figure 3.1, the performance of COSMO-DE-EPS global radiation forecasts is estimated in terms of CRPS. The verification results summarize the *quality* of the forecast considering all types of events and users. Plotted as a function of the forecast horizon, the score is strongly influenced by the diurnal cycle. However, it appears that the forecast exhibits a larger error in day 2 with respect to day 1. A deeper insight in the forecast performance is provided with the help of a score decomposition based on the calibration-refinement factorisation.

3.2 Score decomposition

The decomposition of a score in terms of reliability, resolution and uncertainty has been first proposed for the Brier score (Murphy, 1973; Murphy and Winkler, 1987). This decomposition becomes a standard tool for the interpretation of verification results based on this score. Moreover, a finer analysis of the reliability component can be performed based on a related graphical tool: the *reliability diagram*. The statistical consistency between forecast probability and observed binary outcomes is represented plotting the relative frequency of observed binary events as a function of the forecast probability class.

Not only can the score decomposition be applied to the Brier score, but also to any proper scoring rule (De Groot and Fienberg, 1982; Bröcker, 2009). For a more detailled description of the decomposition procedure, the terms of *entropy* and *divergence* are first defined (Gneiting and Raftery, 2007). The entropy corresponds to the minimum achievable score considering the set of observations Ω. Following Eq. (3.7), the entropy is noted as follows:

$$e(Q) = S(Q, Q). \tag{3.11}$$

The divergence is defined as the difference between the expected score and the entropy:

$$d(P, Q) = S(P, Q) - S(Q, Q). \tag{3.12}$$

The divergence of a negatively oriented proper score is by definition non negative (see again Eq. 3.7), and is commonly referred to as the *reliability* term. Reliability measures the statistical consistency (or more exactly here the statistical divergence) of the forecast P with respect to the distribution Q, which is the expected distribution of the observation given P. The reliability term is negatively oriented (the lower the better).

Introducing a climatological forecast $\bar{Q}$, estimated as the marginal distribution of the observations, the entropy can be further decomposed:

$$e(Q) = S(\bar{Q}, Q) - d(\bar{Q}, Q) \tag{3.13}$$

where $S(\bar{Q}, Q)$ is the expected score of the climatological forecast, and $d(\bar{Q}, Q)$ is the divergence between the forecast and the climatological forecast. The first term is referred to as *uncertainty* and is a function of the observation only. The second term, called *resolution*, is related to the discrimination ability of the distribution Q under different forecast P and thus is a measure of the forecast information content (DelSole, 2004). The resolution term is positively oriented (the higher, the better).

A proper score S can be finally decomposed as follows:

$$S(P, Q) = d(P, Q) - d(\bar{Q}, Q) + S(\bar{Q}, \bar{Q}) \tag{3.14}$$

where the three terms on the right side of the equation are the reliability, resolution, and uncertainty terms, respectively. Integrating over all possible forecasts P, the 3 components of the overall score are estimated.

Recently, it has been proposed a decomposition of the quantile score which participates to the effort of providing equivalent tools for the verification of quantile forecasts as for the verification of probability forecasts (Bentzien and Friederichs, 2014). Formally, the decomposition of QS is noted:

$$QS_\tau = QS_\tau^{\text{reliability}} - QS_\tau^{\text{resolution}} + QS_\tau^{\text{uncertainty}} \tag{3.15}$$

where $\text{QS}_\tau^{\text{reliability}}$, $QS_\tau^{\text{resolution}}$, and $QS_\tau^{\text{uncertainty}}$ are the reliability, resolution, and uncertainty terms of the quantile score at probability level τ, respectively. Similarly as in the categorical case, a graphical tool called quantile reliability diagram provides a representation of the forecast reliability performance. Conditional quantiles of the observations are plotted as a function of quantile forecast classes. A deviation of the reliability curve from the diagonal is interpreted as a lack of reliability.

The decomposition of the CRPS can directly be derived from the decomposition of QS described in Eq. (3.15). Based on the definition of the CRPS in terms of QS in Eq. (3.10), a natural decomposition of CRPS in the ensemble case follows:

$$CRPS = \frac{2}{M}\sum_{m=1}^{M} QS_{\tau_m}^{\text{reliability}} - \frac{2}{M}\sum_{m=1}^{M} QS_{\tau_m}^{\text{resolution}} + \frac{2}{M}\sum_{m=1}^{M} QS_{\tau_m}^{\text{uncertainty}} \tag{3.16}$$

where the three terms on the right of the equation are the CRPS reliability, resolution, and uncertainty terms, respectively. In the continuous case the weighted sum is replaced by the integral over all probability levels. Similarly, the CRPS decomposition can be based on the integration of the Brier score components over all thresholds (Candille and Talagrand, 2005). This approach is however different from the commonly used CRPS decomposition proposed by Hersbach (2000), which is based on the average interval lengths between two successive ensemble values (Tödter and Ahrens, 2012).

In Figure 3.1, besides the CRPS are plotted the corresponding reliability and resolution components. The uncertainty component is not shown since it is a property of the observations only. The decomposition allows a deeper interpretation of the performance results. The decrease in forecast quality from day 1 to day 2 is mainly explained by a decrease in forecast resolution. The information content of the ensemble forecast tends to decline with the forecast horizon. The reliability term follows a diurnal pattern similar over the two days and significantly contributes to the CRPS estimation. Thus, the ensemble suffers from statistical inconsistencies that significantly deteriorate the reliability of the ensemble forecasts.

Based on QS and its decomposition, a deeper analysis of the forecast statistical properties with respect to the observations is performed at the product level. The graphical representation of the QS reliability component with the help of reliability diagrams evidences with more details the statistical deficiencies of the forecast. In Figure 3.2, quantile reliability diagrams are shown for quantiles at three probability levels: 10%, 50% and, 90%. Negative and positive biases affect low and high probability levels, respectively, while quantile forecasts at intermediate levels are well calibrated. This configuration of quantile forecast biases is typically related to underdispersiveness in the ensemble. Moreover, the asymmetry of the biases at low and high probability levels indicates an overall negative bias of the forecast. Now, a deeper analysis of the information content is discussed based on the concept of forecast value.

3.3 Quantile forecast value

The value of a forecast is examined in a risk-based decision-making framework. The estimation of whether appropriate decisions can be taken based on a forecast is directly related to the concept of forecast discrimination. Indeed, forecast discrimination measures the ability of a forecast to successfully discriminate between two different outcomes (Murphy, 1991). This forecast attribute is estimated based on the second factorisation of the joint probability distribution, that is the likelihood-refinement factorisation defined in Eq. (3.2). Discrimination ability is the convert of forecast resolution, both depending on the forecast information content (Wilks, 2006b; Bröcker, 2014).

Discrimination of probability forecasts for binary outcomes is commonly investigated with a tool originating in signal detection theory: the Relative Operating Characteristic (ROC) curve (Mason, 1982). The ROC curve plots the relationship between two characteristics of a binary forecast, the Hit Rate (HR) and the False Alarm Rate (FAR), as a decision criterion varies. HR and FAR are derived from a contingency table that summarises the joint probability distribution of binary forecasts and observations. Binary observations result from the definition of an event while binary forecasts are

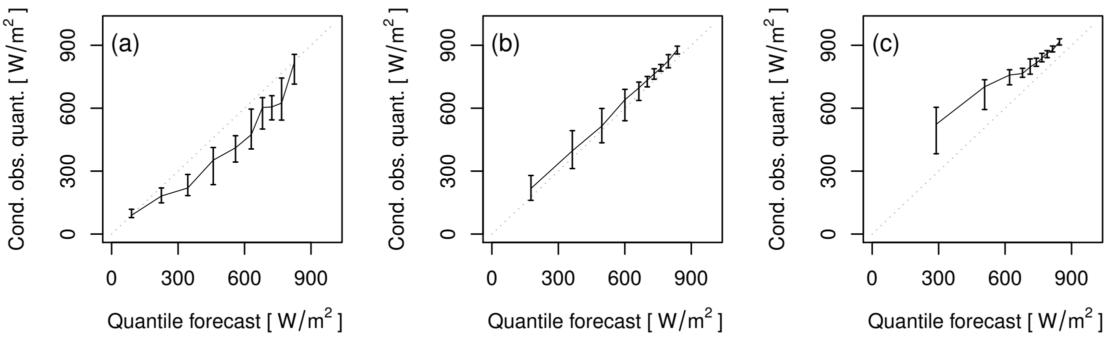

Figure 3.2: Reliability diagrams for quantile forecasts at probability levels 10% (a), 50% (b), and 90% (c). The confidence intervals are estimated by block bootstrapping. Results for the period July-August 2015 and a forecast horizon of 9 hours.

here derived from continuous forecasts applying a *decision criterion*. The area under the ROC curve has been popularised as a summary measure of the forecast discrimination ability when the focus in on a specific event[2].

In Appendix A, an equivalent tool is proposed for the estimation of the forecast discrimination ability focusing on a specific user. The so-called Relative User Characteristic (RUC) curve plots FAR and HR as the event of interest varies and the decision criteria are adjusted according to the level of risk accepted by the user (Section A.4). Indeed, the decision criteria applied to a forecast can be adjusted as a function of the sensitivity of the user to over- and under-forecasting characterised by an asymmetry level τ. Thus, the RUC framework allows scanning the ability of a specific user to take opportune actions based on the forecast for a range of events.

The RUC framework is appropriate to the analysis of forecasts expressed in terms of a quantile since the associated probability level defined exactly the user's risk adversity. A RUC analysis can as well be applied to a continuous deterministic forecast though this forecast is not targeted *a priori* at a specific user. A fair comparison between deterministic and ensemble-derived forecasts can take place in this framework: the full information from continuous forecasts can be assessed in both cases. In the following, the ensemble approach is compared to a *single run* approach, which consists in using as decision variable a single forecast randomly chosen among the 20 ensemble members.

The value of a forecast in a dichotomous decision framework can be derived from the ROC or RUC curve. For this purpose, a decision-making framework is depicted by a simple static cost-loss model (Thompson, 1962; Katz and Murphy, 1997). This model describes a situation of dichotomous decisions: a user has to decide whether or not to take protective action against potential occurrence of an event. The user's decision is based on a *decision variable*, *i.e.* a forecast. Taking action implies a cost C while a loss L is encountered when the event occurs without preventive action. The cost-loss ratio C/L, denoted α, fully characterised the users. This model generalized to the continuous case considering an asymmetric loss function ρ_τ with $\tau = 1 - \alpha$ (see Eq. A.19).

The *relative value* (or *economic value*) V of a forecast is estimated based on this model (Richardson, 2000; Wilks, 2001; Zhu *et al.*, 2002). V is expressed as the reduction of mean expense when using

[2] It has been shown that the ROC area is a particular case of the generalised discrimination score (also known as two-alternative forced choice test), which provides a general measure of the potential usefulness of a forecast (Mason and Weigel, 2009).

a forecast instead of climatological information relative to the case of using a perfect deterministic forecast:

$$V = \frac{\bar{E}_{\text{climate}} - \bar{E}_{\text{forecast}}}{\bar{E}_{\text{climate}} - \bar{E}_{\text{perfect}}}, \tag{3.17}$$

where $\bar{E}_{\text{forecast}}$, $\bar{E}_{\text{perfect}}$, and $\bar{E}_{\text{climate}}$, are the mean expenses when a user takes decisions based on a forecast, on a perfect deterministic forecast, and on climatological information, respectively. Eq. (3.17) can be developed (see Appendix A) and V noted as a function of the user's cost-lost ratio α, the event climatological frequency of occurrence π and the two forecast characteristics HR and FAR:

$$V = \begin{cases} (1 - \text{FAR}) - \left(\dfrac{\pi}{1-\pi}\right)\left(\dfrac{1-\alpha}{\alpha}\right)(1 - \text{HR}) & \text{if} \quad \alpha < \pi \\ \text{HR} - \left(\dfrac{1-\pi}{\pi}\right)\left(\dfrac{\alpha}{1-\alpha}\right)\text{FAR} & \text{if} \quad \alpha \geq \pi. \end{cases} \tag{3.18}$$

Therefore, the value estimates the relative performance of a forecast focusing on a specific event, characterised by a frequency of occurrence π, and simultaneously focusing on a specific user with a risk aversion defined by a cost-loss ratio α.

In Eq. (3.18), FAR and HR can be computed considering that the decision criterion applied to the forecast corresponds to the threshold applied to define the event of interest. In this case, the forecast is said to be taken at *face value*. If the decision criterion is optimised for the considered event/user situation, V is called *potential value*, that is, the maximum achievable value considering the forecast at hand. Value and potential value are identical if the forecast is reliable, thus conditioned on calibration (Richardson, 2011). In other words, reliability is a necessary condition for the optimisation of forecast-based decision processes.

Two options have been proposed for the graphical representation of the forecast relative value, depending on whether the focus is on a specific event or a specific risk level. The *probability value plot* shows the relative value of a forecast for a given event of interest as a function of the user's cost-loss ratio. Alternatively, the *quantile value plot* shows the relative value of a forecast for a given cost-loss ratio as a function of the event of interest. A third option, not explored yet, would consist in summarising the potential value in a contour-plot with respect to the plane defined by the user's cost-loss ratio and the event of interest.

Figure 3.3 shows quantile value plots of COSMO-DE-EPS global radiation forecasts focusing on three probability levels, 25%, 50%, and 75%, corresponding to users with cost-lost ratio 75%, 50%, and 25%, respectively. More specifically, the plots represent the potential value, *i.e.* the value of the forecast conditioned on calibration, for forecasts valid at 12 UTC. By definition, the potential value is positive and the curve reaches its maximum for events with a frequency of occurrence π corresponding to $1 - \tau$. Here, the ensemble-derived forecasts are compared to deterministic forecasts. The results show that the ensemble system outperforms the single forecast approach for the three quantile forecasts and the whole range of events of interest investigated. So, the ensemble provides additional information with respect to deterministic forecasts and the ensemble-derived products appear to be valuable for a wide range of applications.

3.4 Ensemble added value

Relative measure of forecast performance are traditionally estimated by means of *skill scores* (Wilks, 2006b). Based on negatively oriented scoring rules, skill scores have the form of a relative difference.

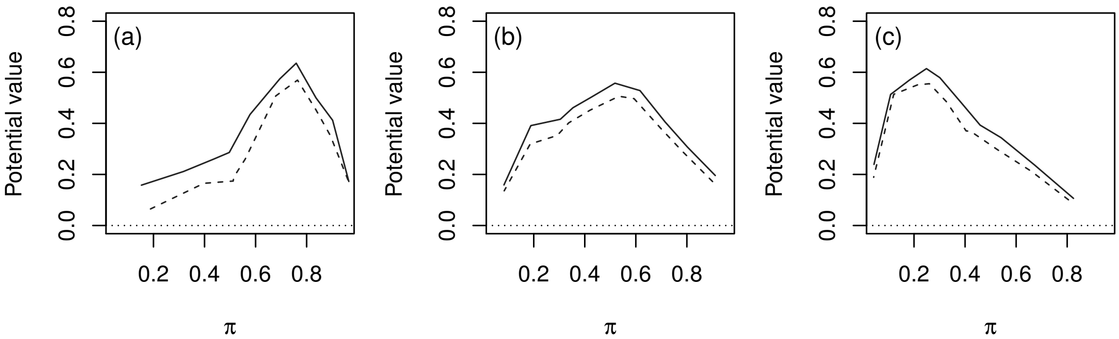

Figure 3.3: Quantile value plots showing the potential value of COSMO-DE-EPS global radiation quantile forecasts as a function of the event of interest expressed in terms of their climatological frequency of occurrence (π). Potential value of the ensemble forecast (full lines) and of a deterministic forecast (dashed lines) for users with cost-loss ratios of 75% (a), 50% (b), and 25% (c). Results for the period July-August 2015 and a forecast horizon of 9 hours.

For example, the continuous ranked probability skill score (CRPSS) is defined as:

$$\text{CRPSS} = \frac{\text{CRPS} - \text{CRPS}^{\star}}{\text{CRPS}^{\diamond} - \text{CRPS}^{\star}} = \frac{\text{CRPS}^{\star} - \text{CRPS}}{\text{CRPS}^{\star}} = 1 - \frac{\text{CRPS}}{\text{CRPS}^{\star}} \tag{3.19}$$

where CRPS, CRPS* and CRPS$^{\diamond}$ are the scores of the forecast under assessment, of a reference forecast and of a perfect deterministic forecast, respectively. The choice of the reference forecast corresponds to the goal of the comparison. Climatological forecasts or persistence forecasts are chosen as reference in order to show the benefit of a forecasting system with respect to a cost-free approach. Skill scores help also to demonstrate the improvement reached by a new numerical model with respect to an older version or to compare two concurrent systems.

Here, the performance of the ensemble forecast is compared to a single forecast. The deterministic forecast is used as benchmark in order to focus on the intrinsic benefit of the computationally expensive ensemble system. A difficulty arises since the two forecasting approaches are by nature different: the first one is probabilistic while the second one is deterministic. Since the CRPS reduces to the Mean Absolute Error (MAE) in the single forecast case, a simple approach could consist in comparing the CRPS of the ensemble forecast and the MAE of a single run. By doing so, the comparison confronts a probabilistic interpretation of a probabilistic forecast on one side, and a deterministic interpretation of a deterministic forecast on the other side. For a fair comparison, both types of forecasts should be interpreted in a probabilistic framework.

For this purpose, consider stochastic optimisation based on a single or an ensemble forecast. In a risk-based decision-making framework, the decision criterion applied to a forecast is adjusted as a function of the user's cost-loss ratio. This can be done independently of the nature of the forecast. Therefore, the results of decisions based on an ensemble-derived forecast or a deterministic forecast can be compared in a fair manner. This approach has been followed for the comparison of quantile forecasts and single run forecasts in terms of potential value as shown in Figure 3.3. Furthermore, the comparison can be extended to more than one single user and one single event using the relationships between relative value and proper scoring rules. Indeed, integrating the forecast value over all cost-loss ratios or over all events leads to the definition of the Brier skill score and quantile skill score, respectively (Murphy, 1969, Section A.6).

The potential value of a forecast is by definition estimated conditioned on calibration. Similarly, performance measures can focus on the information content only. Practically, it consists in considering the entropy of a scoring rule, the minimum achievable score for a given dataset. The integration of the score entropy as defined in Eq. (3.11) over all forecasts leads to the concept of *potential* score. The potential CRPS is defined as the difference between the uncertainty and resolution components, thus:

$$\text{CRPS}_{\text{potential}} = \frac{2}{M}\sum_{m=1}^{M}(\text{QS}_{\tau_m}^{\text{uncertainty}} - \text{QS}_{\tau_m}^{\text{resolution}}) \tag{3.20}$$

where $\text{QS}_{\tau_m}^{\text{uncertainty}}$ and $\text{QS}_{\tau_m}^{\text{resolution}}$ are the uncertainty and resolution components of the QS at probability level τ_m, respectively. Their difference corresponds to the potential QS, the potential value of a forecast for a given user over all thresholds.

Introduced in Appendix B, the *ensemble added value* (EAV) is proposed as a new measure that summarises the benefit of using an ensemble forecasting system rather than a single run. EAV focuses exclusively on the information content: the forecast variability that allows taking adequate decisions is rewarded while the reliability deficiencies are seen as a decision criteria adjustment problem. The ensemble added value is expressed in the form of a skill score as:

$$\text{EAV} = 1 - \frac{\text{CRPS}_{\text{potential}}}{\text{CRPS}^{\star}_{\text{potential}}} \tag{3.21}$$

where $\text{CRPS}_{\text{potential}}$ and $\text{CRPS}^{\star}_{\text{potential}}$ are the scores estimated applying Eq.(3.20) to the ensemble-derived quantile forecasts and to the reference deterministic forecast, respectively.

In Figure 3.4, EAV of COSMO-DE-EPS global radiation forecasts is plotted as a function of the forecast horizon. One forecast is chosen randomly among the 20 ensemble members for each verification day and designated as reference deterministic forecast. The relative potential benefit of the ensemble approach ranges between 5% and 15% and tends to increase with the forecast horizon. Indeed, the benefit over day 2 appears to be higher than over day 1. Despite the decrease in resolution with the forecast horizon noted in Figure 3.1, the resolution of the deterministic forecast tends to zero at a faster rate than the resolution of the probabilistic forecast. In other words, the information content in the ensemble increases with respect to the information content in the single run case, so the use of ensemble forecasts is particularly valuable for long lead times.

The potential benefit of COSMO-DE-EPS global radiation forecasts shown in Figure 3.4 as well as the economic value presented in Figure 3.3 are conditioned on calibration. However, the ensemble system demonstrates not to provide reliable probabilistic forecasts in all cases. COSMO-DE-EPS global radiation forecasts suffer from statistical inconsistencies, *i.e.* biases and ensemble underdispersiveness. The lack of reliability varies as a function of the forecast horizon (see Figure 3.1), of the risk level associated with a probabilistic product (Figure 3.2), and also of the time of the year (see AppendixB). Now, these deficiencies have to be corrected by adequate methods in order to fully benefit from the ensemble approach.

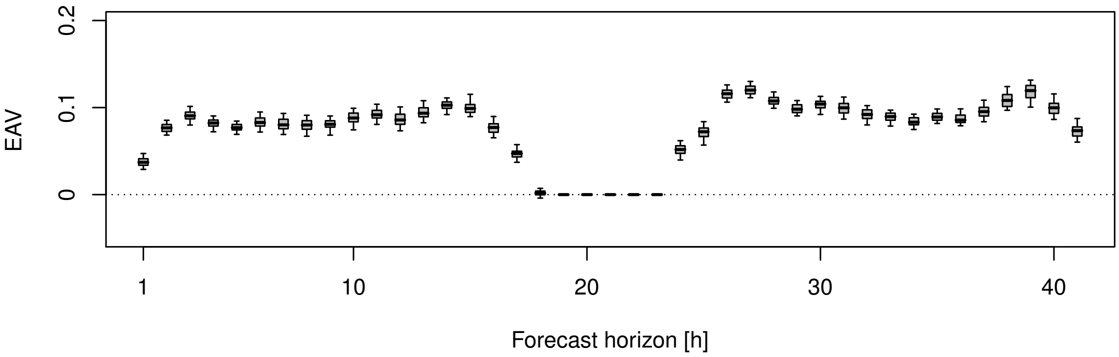

Figure 3.4: Ensemble added value of the global radiation forecasts derived from COSMO-DE-EPS with respect to a single member approach as a function of the forecast horizon. The confidence intervals are estimated by block bootstrapping. Results for the period July-August 2015.

4. Ensemble Post-processing

Post-processing is a fundamental step in the whole process of providing the users with reliable probabilistic forecasts based on ensemble simulations. Reliability is a necessary condition for the optimisation of a decision-making process (see Chapter 3). However, ensemble global radiation forecasts, and more generally ensemble forecasts of surface variables, suffer from biases and spread deficit that affect the reliability of the derived probabilistic products. Adequate calibration techniques that correct for these drawbacks are therefore required.

Based on historical data, statistical post-processing consists in the application of learning algorithms aiming at the correction of systematic forecast deficiencies. Historically, post-processing of ensemble forecasts has first been tackled as a probability bias problem that can be solved with a reassignment of the probability forecasts (Murphy, 1993; Atger, 2004). For example, considering that the frequency of an observed event is, say 25%, for all cases where a system issues a probability forecast of, say 20%, a non-parametric calibration technique consists in assigning the probability 25% when the probability forecast is 20%. Using such an approach for all probability classes, the statistical discrepancy of the probability forecast is corrected and the forecast is said to be reliable. Since the forecast information content is not modified with this simple readjustment, forecast resolution and discrimination ability are not affected by calibration in that case[1].

Numerous ensemble calibration methods have been developed in recent years for a wide range of applications (Gneiting *et al.*, 2005; Wilks, 2006a; Wilks and Hamill, 2007). On the one hand, semi-parametric approaches focus on derived probabilities or on derived quantiles, such as logistic regression (Hamill *et al.*, 2004; 2007) or quantile regression (Bentzien and Friederichs, 2012), respectively. On the other hand, fully parametric approaches are based on the definition of predictive distributions as a whole and require to apply distribution fit (Gneiting *et al.*, 2005), kernel dressing (Bröcker and Smith, 2008) or Bayesian model averaging (Raftery *et al.*, 2005). The frontiers between the different approaches are however often porous. For example, logistic regression can be generalized to a distributive approach by means of an extension of the regression equations, which provides a full description of the predictive distribution (Wilks, 2009; Ben Bouallègue, 2013; Messner *et al.*, 2014b). Today, enhanced techniques are investigated that aim to benefit from the post-processing step in order to increase the discrimination ability of the forecast by correcting biases as a function of the forecast weather conditions (Wahl, 2015).

Calibration methods however usually focus only on a single or few aspects of the ensemble forecast, such as solar radiation at a given location and forecast horizon. The dependence structures of the ensemble forecast across time, space and variables are lost after the statistical adjustment performed for each marginal predictive distribution separately, whereas the variability of the forecast, both its predictable and uncertainty components, is correlated in time, space and between variables. Moreover,

[1] for this reason, it has often been argued that only the reliability (and not the resolution) of a forecast can be improved by calibration (e.g. Toth *et al.*, 2003).

the user's requirements presumably cover a variety of spatial and temporal scales, or a combination of parameters of interest. For example, applications based on global radiation forecasts may focus on intraday variability or may deal with the daily mean of day-ahead forecasts, at local, regional or national scale. Global radiation forecasts are also combined with forecasts of other weather variables such as temperature at ground in order to feed a physical PV power model, or as wind in order to manage intermittent renewable sources as a whole.

The calibration of the multivariate aspect of the forecast in a single step using parametric approaches is computationally expensive and therefore not adapted to the full dimensionality of NWP forecasts (Keune *et al.*, 2014; Feldmann *et al.*, 2015). Alternatively, the generation of marginal predictive distributions and forecast dependencies can be treated sequentially applying the famous Sklar's theorem, which stipulates that a multivariate joint distribution can be described by its univariate marginals plus copula (Sklar, 1959). In particular, dependencies can be modelled based on information in the original ensemble forecast using empirical copula approaches that are computationally efficient.

In this framework, ensemble post-processing is approached as a two-step procedure. First, the marginal calibration adjusts the predictive distributions at each forecast horizon and location. Second, scenarios are generated based on the dependence structures of the original ensemble forecasts. As a result, at the end of the post-processing steps, calibrated scenarios with consistent dependence structures can be delivered to the users.

4.1 Marginal calibration

Learning algorithms applied to ensemble forecasts have the primary goal of correcting predictive distributions for their non conformity with statistical properties of the observations. This first post-processing step, called here marginal calibration, aims to generate reliable predictive marginal distributions or functional of them focusing on probability forecasts or quantile forecasts at the station level. The choice of the calibration method depends mainly on the weather parameter under focus and the end-product of interest (Hagedorn *et al.*, 2007; Hamill *et al.*, 2007; Messner *et al.*, 2014a). The aim here is to provide probabilistic forecasts of global radiation in terms of quantiles in order to feed an empirical copula model in a second post-processing step (see Sections 4.3 and 4.4).

Ensemble calibration has been investigated intensively for weather variables such as temperature, precipitation, or wind, but only recently applied to global radiation forecasts (e.g. Wilks and Hamill, 2007; Junk *et al.*, 2014; Zamo *et al.*, 2014b). In Appendix C, quantile regression (QR) demonstrates to be an appropriate technique for the calibration of ensemble global radiation forecasts. Conveniently, quantile forecasts at nominal probability levels of interest are directly calibrated without assumptions about the form of the underlying probability distribution. Moreover, QR can be indifferently applied to a probabilistic or a deterministic forecast. In the following, this characteristic of the calibration method allows to develop the comparison of probabilistic products derived from ensemble and single forecasts initiated in Chapter 3.

QR is a regression technique proposed by Koenker and Bassett (1978). Applied to the response variable Y, QR estimates quantiles Q_τ of the variable distribution $F_Y(y)$ conditional on a set of predictors based on a linear model. Considering a sample of observations $\{y_1, ..., y_{N'}\}$ with N' the size of the training sample and a probability level of interest τ, the optimisation process consists in minimising asymmetrically weighted absolute residuals:

$$\min_{\theta \in \mathbb{R}} \sum_{j=1}^{N'} \rho_\tau (y_j - \theta) \tag{4.1}$$

where ρ_τ is the check function, the asymmetric loss function defined in Eq. (3.9). Replacing the scalar θ by a parametric linear function, the minimisation problem is written as follows:

$$\arg \min_{(\beta_\tau^0, \boldsymbol{\beta_\tau})} \sum_{j=1}^{N'} \rho_\tau(y_j - \beta_\tau^0 - \boldsymbol{\beta}_\tau \boldsymbol{v}_j) \tag{4.2}$$

where β_τ^0 and $\boldsymbol{\beta_\tau}$ are the so-called regression coefficients and $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_{N'}\}$ the vectors of predictors. Quantile functions are estimated in the following for the probability levels τ_m defined by the ensemble size following Eq. (2.4). Conditioned on the current forecast $\boldsymbol{v_0}$, the derived predictive quantile function corresponds to:

$$\hat{Q}_\tau(y \mid \boldsymbol{v_0}) = \hat{\beta}_\tau^0 + \hat{\boldsymbol{\beta}}_\tau \boldsymbol{v_0} \tag{4.3}$$

where the regression coefficients are estimated separately for each forecast horizon and applied indiscriminately to all locations or model grid points in this study.

The predictors that feed the regression model are usually the variables to be calibrated. For global radiation, a *standard predictor* setup includes three components: the global radiation forecast itself (g), a power transformation of the global radiation forecast (g^2) in order to account for non-linearities, and the radiation at the top of the atmosphere in order to capture the natural cycle associated with the solar geometry. Applied to ensemble forecasts, the main predictor g is the first guess quantile such as $g = x^{(m)}$ where $x^{(m)}$ is the ensemble member interpreted as a quantile forecast at probability level $\tau = \tau_m$ (Eq. 2.3). QR can also be applied to a deterministic forecast. In that case, the predictor g corresponds to the single global radiation forecast for all examined probability levels.

Another important aspect of the first post-processing step is the definition of the historical dataset on which the optimisation process is based. Today, a standard approach consists in using a rolling window covering 4 to 10 weeks preceding the forecast to be calibrated (Gneiting *et al.*, 2005). The dataset is updated regularly in order to follow the seasonal pattern of the forecast-observation statistical characteristics. This approach has the advantage of being appropriate for operational suites with frequent updates of the underlying model. The training dataset is here defined as a rolling window of 45 days updated on a daily basis.

With this setup, QR is applied to COSMO-DE-EPS global radiation forecasts and the impact of calibration on the forecast performance is assessed over a 2 month period. Verification results are summarised by means of the CRPSS. In Figure 4.1(a), calibrated ensemble forecasts are compared to raw ensemble forecasts showing a significant increase of the forecast performance after calibration for all daytime forecast horizons. Ranging between 15% and 40%, the improvement is more important for sunrise/sunset hours but almost similar for day 1 and day 2. A deeper analysis is based on the CRPS and QS decompositions (not shown). The decomposition of the CRPS applied to the calibrated forecasts indicates that the reliability component becomes negligible and the resolution component remains unaffected after the calibration step. Moreover, at the product level, the inspection of quantile reliability diagrams demonstrates that the standard calibration approach is effective in providing reliable probabilistic products.

In a second experiment, probabilistic forecasts are this time derived from deterministic forecasts using the same statistical method (QR) and a training sample of the same size (45 days). The derived calibrated quantile forecasts, based on a single run approach, are used as reference forecasts for the computation of the CRPSS in Figure 4.1(b). Calibrated forecasts appear to be significantly better when derived from the ensemble forecasts rather than from single runs. The forecasts are reliable in both cases as checked by reliability diagrams (not shown). The performance difference lies in the higher information content in the ensemble case compared to the purely statistical approach. The result of this comparison was expected: CRPSS in Figure 4.1(b) and EAV in Figure 3.4 are the practical

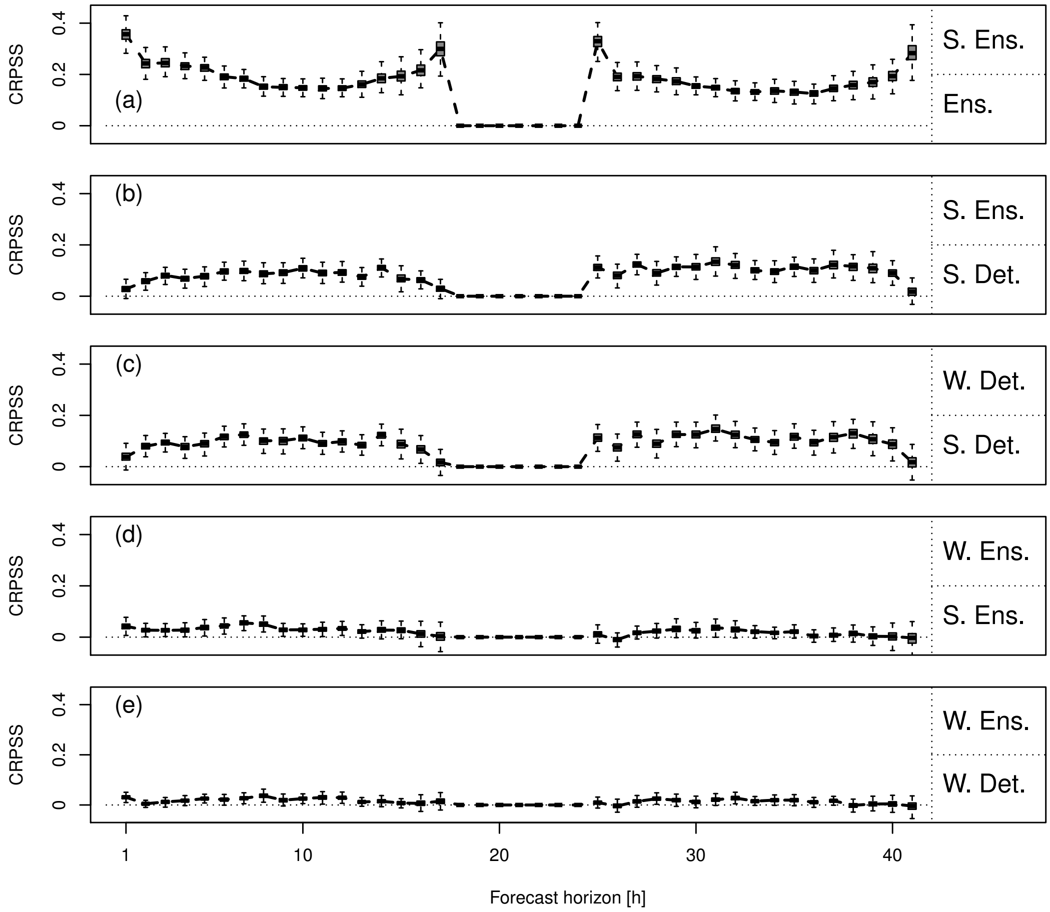

Figure 4.1: Verification results showing the impact of the calibration step in terms of CRPSS as a function of the forecast horizon: (a) standard calibrated ensemble forecast (S. Ens.) against raw ensemble forecast (Ens.), (b) standard calibrated ensemble forecast (S. Ens.) against standard calibrated forecast from a single run (S. Det.), (c) weather-dependent calibration (W. Det.) against standard calibration based on a single run, (d) weather-dependent calibration (W. Ens.) against standard calibration based on the ensemble forecast, and (e) weather-dependent calibrated ensemble forecast against weather-dependent calibrated forecast from a single run. The box plots indicate confidence intervals estimated by block bootstrapping. Results for the period July-August 2015.

and theoretical estimations of the ensemble benefit with respect to a single forecast, conditioned on calibration, respectively.

4.2 Weather-dependent calibration

Weather-dependent calibration aims to go beyond a bias correction, which adjusts the forecast conditioned only on its value and the slowly changing characteristics of the rolling training dataset. Learning processes based on past data can exploit the historical dataset in order to relate weather situations with their influence on bias and spread deficits. In that case, the optimisation of the forecast is not only expected to provide reliable forecasts but also to increase their discrimination ability.

Two different types of approaches have been proposed for this purpose. First, the training dataset can be adapted as a function of the forecast to be calibrated. Post-processing of ensemble forecasts based on the analogue technique follows this approach (Hamill and Whitaker, 2006; Junk *et al.*, 2015). The training dataset is restricted only to past forecasts (and the corresponding observations) sharing a certain degree of similarity with the current one. Therefore, this approach requires a long record of observations and forecasts with a frozen configuration of the numerical model and of the ensemble setup. This is obviously a drawback for regularly updated operational systems.

The second type of approach, which is applied here, consists in enlarging the number of predictors entering the optimisation process. This approach is appropriate for regression techniques such as QR. The training dataset is increased by including forecasts of weather variables that potentially inform about forecast error characteristics. With this approach, training samples can still be defined as rolling windows.

A selection of appropriate predictors is desired for each regression equation. The minimisation problem is solved independently for different forecast horizons and for a range of probability levels. Meaningful predictors in one case could be uninformative in another one. In this context, the regularization of the regression scheme offers an appealing approach for a rigorous and automated predictor selection. Regularization consists in adding a penalty term in the regression equation that constrains a given number of predictors to receive null weights, and consequently activates only the remaining ones. With this technique, the predictor selection avoids problems related to overfitting and colinearities between predictors (Siegert *et al.*, 2011).

The regularization implemented here follows the Least Absolute Shrinkage and Selection Operator (LASSO) approach (Tibshirani, 1996; Wahl, 2015). Formally, the penalty term corresponds to the absolute size of the regression coefficients modulated by a regularization parameter noted λ_r. Applied to QR, the scheme is called penalized quantile regression (PQR), and the regression coefficients $(\beta^0_\tau, \boldsymbol{\beta_\tau})$ are estimated solving the following minimisation problem:

$$\arg \min_{(\beta^0_\tau, \boldsymbol{\beta_\tau}))} \sum_{j=1}^{N'} \rho_\tau(y_j - \beta^0_\tau - \boldsymbol{\beta}_\tau \tilde{\boldsymbol{v}}_j) + \lambda_r \|\boldsymbol{\beta}_\tau\| \tag{4.4}$$

where $\|\cdot\|$ refers to the L_1-norm, and $\{\tilde{\boldsymbol{v}}'_1, ..., \tilde{\boldsymbol{v}}_{N'}\}$ are normalized pre-selected predictors (with zero mean and unit variance). The regularization parameter λ_r is optimised at each calibration step by a leave-n-out score as measure of performance, where the natural score here is QS (Bröcker, 2010, see Section C.5) .

The pre-selected predictors that enter the selection process are derived from direct model outputs. A list of 13 weather variables is proposed in Appendix C (see Table C.1). The *weather predictors* include forecasts of cloud cover, precipitation, temperature, and wind at different levels. Only predictor values at each station are considered in the following while a weather dependent calibration approach could possibly also account for derived weather patterns using for example principal component analysis (Friederichs and Hense, 2007). Considering ensemble forecasts at the station level, several derived quantities are used as predictors, namely the ensemble mean, the ensemble spread, the minimum and the maximum of all members, for each weather variable, at each forecast horizon. The pool of pre-selected predictors $\boldsymbol{v'}$ includes finally the standard predictors (functions of global radiation forecasts and radiation at the top of the atmosphere), the weather predictors, and multiplicative terms combining the weather predictors and the global radiation forecasts (see Section C.3).

PQR based on rolling windows of 45 days is applied to COSMO-DE-EPS global radiation forecasts. An example of the resulting calibrated forecasts is provided in Figure 2.2(c). A subjective assess-

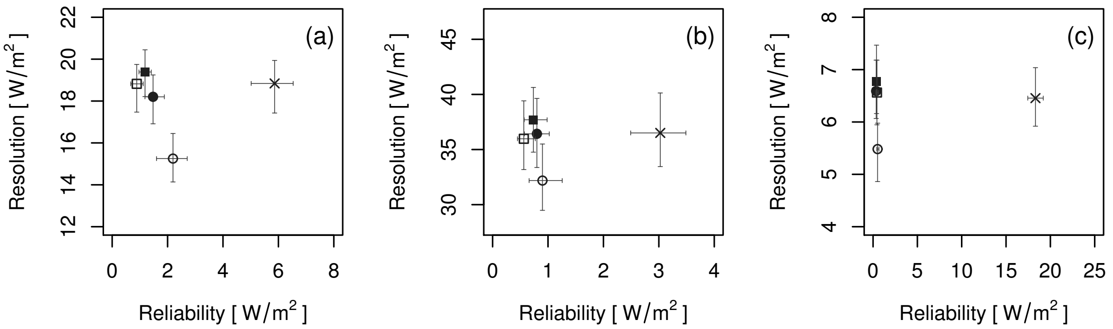

Figure 4.2: Reliability and resolution components of the QS for quantile forecasts at probability levels 10% (a), 50% (b), and 90% (c). Results are shown for the original ensemble (×), calibrated deterministic forecasts with QR (○), calibrated ensemble forecasts with QR (●), calibrated deterministic forecasts with PQR (□), and calibrated ensemble forecasts with PQR (■). The confidence intervals are estimated by block bootstrapping. Results for the period July-August 2015 and a forecast horizon of 9 hours.

ment of this case study suggests that the calibrated forecast with PQR is still relatively sharp but better captures the observation variability than the uncalibrated forecast plotted in Figure 2.2(b). In a parallel experiment, PQR is also applied to deterministic forecasts. First, the impact of the weather-dependent calibration approach is analysed in the single and ensemble cases separately, and then the comparison between single and ensemble cases is discussed at the end of this Section 4.2.

In Figure 4.1(c), the impact on the forecast performance of the weather-dependent calibration with respect to the standard calibration is plotted in terms of CRPSS as a function of the forecast horizon for the single run case. Based on deterministic forecasts, the weather-dependent approach increases significantly the quality of the calibrated forecasts with an improvement of about 10%. Comparing the CRPSS in Figure 4.1(c) and Figure 4.1(b) that share the same reference forecast, weather-dependent calibrated forecasts based on a single forecast appears to be as performant as standard calibrated forecasts based on an ensemble forecast.

In Figure 4.1(d), the impact of the weather-dependent approach is assessed in the ensemble case. Also in that case, calibration based on adequate weather predictors significantly improves the calibrated forecasts. CRPSS is about 5% and the benefit of the weather-dependent calibration appears to be more pronounced for day 1 than day 2. Using now the weather-dependent calibrated forecast based on deterministic forecasts as reference, the CRPSS of the weather-dependent calibrated ensemble is shown in Figure 4.1(e). Considering enhanced calibration techniques at hand, the benefit of the ensemble approach is still significantly positive over day 1 and most hours of day 2 but not exceeding 4%.

In Figure 4.2, the impact of calibration is analysed at the product level showing reliability and resolution components of the QS for quantile forecasts at three probability levels: 10%, 50%, and 90%. The raw ensemble is compared to calibrated forecasts based on the ensemble or on single runs, using a standard calibration approach or a weather-dependent technique. All calibrated quantile forecasts exhibit good statistical consistency and similar reliability terms independently of the predictors and of the approach implemented. The main difference between the calibrated forecasts becomes evident in their information content. Standard calibration of the ensemble forecast does not affect the forecast resolution ability whereas weather-dependent calibration improves it, in particular for intermediate probability levels. Using deterministic forecasts as predictor, the information content of the calibrated forecasts is able to reach the level of the original ensemble when a weather-dependent approach is applied.

The results in Figure 4.2 indicates that the generation of reliable probabilistic products can be achieved using either QR or PQR, based on ensemble or deterministic forecasts. Weather-dependent forecast uncertainty, in a univariate framework, can be assessed with an ensemble system or applying enhanced statistical methods to a single forecast with similar results. The combination of the ensemble approach and complex statistical methods is nevertheless the best option in terms of forecast performance.

Furthermore, the relevance of ensemble forecasts as a basis for post-processing is examined. A characteristic of the ensemble, which has not been explored yet, is that the forecasts provide not only probabilistic information at a single location and forecast horizon but also a complete view of the forecast uncertainty and its dependence structure across time, space and weather variables. The multivariate aspect of the forecast and techniques that allow generating scenarios based on information from the original ensemble are now discussed.

4.3 Generation of consistent scenarios

After the first post-processing step, calibrated quantile forecasts are provided at each forecast horizon and location within the model domain separately. Consequently, information about the dependence structure of the forecast is no more provided as illustrated in the example of Figure 2.2(c). Physically consistent probabilistic forecasts across time and space are however required for complex decision-making situations encountered in renewable energy applications (e.g. Pinson *et al.*, 2009). The second post-processing step aims therefore to reconstruct consistent scenarios based on the ensemble calibrated marginals. This step is performed using the mathematical concept of the *copula* (e.g. Mikosch, 2006; Schölzel and Friederichs, 2008).

The relationship between a multivariate distribution function and its univariate marginal distributions is stated by Sklar's theorem (Sklar, 1959). A multivariate cumulative distribution function $\mathcal{F}(y_1, ..., y_K) = Pr[Y_1 \leq y_1, ..., Y_K \leq y_K]$, with $y_1, ..., y_K \in \mathbb{R}$, can be expressed as:

$$\mathcal{F}(y_1, ..., y_K) = \mathcal{C}(F_1(y_1), F_K(y_K)) \tag{4.5}$$

where $F_k(y) = Pr[Y_k \leq y_k]$ with $k \in \{1, .., K\}$ are the univariate marginals, and $\mathcal{C}$ a copula function. Quantiles of the predictive marginal distributions $F_1, ..., F_K$ are provided by the first post-processing step while the copula $\mathcal{C}$, which models the dependencies, has to be specified. The use of copulas does not modify the problem dimension, but allows to focus sequentially on the marginals and on the dependencies.

The choice of the copula depends on the problem at hand, and in particular on its dimensionality K. Parametric families of copulas based on well-known distributions are adequate for small dimension applications. For example, post-processing based on Gaussian copulas can be applied to the simultaneous calibration of the two components of wind vector forecasts (Schuhen *et al.*, 2012). For high-dimensional problems, non-parametric methods based on empirical copulas appear to be well-suited.

Formally, an empirical approach is based on a discrete reference dataset noted:

$$\boldsymbol{z} = \left\{(z_1^1, ..., z_1^{N_s}), ..., (z_K^1, ..., z_K^{N_s})\right\}, \tag{4.6}$$

with N_s the number of scenarios of dimension K. The rank correlation structure of the *reference template* $\boldsymbol{z}$ is used to generate the multivariate dependence structures of the forecast. The permutations $u_k(n)$ for each dimension $k \in \{1, ..., K\}$ are derived from the univariate rank statistics of z_k^n in $\left\{z_k^1, ..., z_k^{N_s}\right\}$:

$$u_k(n) = rank(z_k^n) = \sum_{i=1}^{N_s} \mathbb{I}[z_k^i \leq z_k^n] \tag{4.7}$$

where $\mathbb{I}[.]$ is an indicator function taking the value 1 if the condition in brackets is true and zero otherwise. Calibrated scenarios are generated applying the permutations $u_k(n)$ to the calibrated quantile forecasts $\boldsymbol{q} = \left\{(q_1^{\tau_1}, ..., q_1^{\tau_{N_s}}), ..., (q_K^{\tau_1}, ..., q_K^{\tau_{N_s}})\right\}$ with $q_k^{\tau_i} = F_k^{-1}(\tau_i)$. The post-processed forecasts noted $\tilde{\boldsymbol{x}}$ finally result from the following rearrangements:

$$\tilde{x}_k^1 = q_k^{\tau_{u_k(1)}}, ..., \tilde{x}_k^{N_s} = q_k^{\tau_{u_k(N_s)}} \tag{4.8}$$

in the dimensions $k \in \{1, ..., K\}$.

The Ensemble Copula Coupling (ECC) approach uses the original ensemble forecast as reference template $\boldsymbol{z}$. The rank structure of the ensemble is conserved after post-processing assuming that the ensemble forecast correctly represents the spatio-temporal dependence structure of the weather variable (Schefzik *et al.*, 2013). This approach is computationally cheap and provides a post-processed ensemble with the same number of members and of the same dimension as the original one. However, ECC can generate non-realistic scenarios when the first post-processing step indiscriminately increases the ensemble spread to a large extent. In that case, non-representative or random rank structures in the raw ensemble are magnified after post-processing and inconsistent forecasts with unrealistically high variability are generated.

4.4 Dual ensemble copula coupling

A new method for the generation of scenarios from calibrated marginals and ensemble information is proposed in Appendix D. Based on ECC, the new empirical method aims to alleviate the generation of unrealistic variability in the scenarios using the assumption of error stationarity in the forecast already adopted in parametric approaches (Pinson *et al.*, 2009; Schölzel and Hense, 2011). The so-called dual ensemble copula coupling (d-ECC) is a semi-parametric approach that combines a *dual* source of information: the original ensemble forecasts and the autocorrelation of the forecast error estimated from past data.

d-ECC is implemented here focusing on scenarios in the form of time series, so the dimension of the multivariate forecast corresponds to the forecast horizon denoted T. The temporal correlation of the forecast error is described by a correlation matrix $\boldsymbol{R_e}$ defined as:

$$\boldsymbol{R_e} = \begin{pmatrix} r_{e_1,e_1} & r_{e_1,e_2} & \cdots & r_{e_1,e_T} \\ r_{e_2,e_1} & r_{e_2,e_2} & \cdots & r_{e_2e,T} \\ \vdots & \vdots & \ddots & \vdots \\ r_{e_T,e_1} & r_{e_T,e_2} & \cdots & r_{e_T,e_T} \end{pmatrix} \tag{4.9}$$

where $r_{e_{t_1},e_{t_2}}$ is the correlation of the forecast error between the forecast horizons t_1 and t_2. The correlation matrix is estimated based on the training sample used for the first post-processing step, which is a rolling window of 45 days updated on a daily basis. Elements of the estimated correlation matrix $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$ are represented in Figure 4.3. Focusing on horizons of 9, 16 and 27 hours, the autocorrelation of the forecast error is plotted as a function of the time lag. In the three examples, the autocorrelation

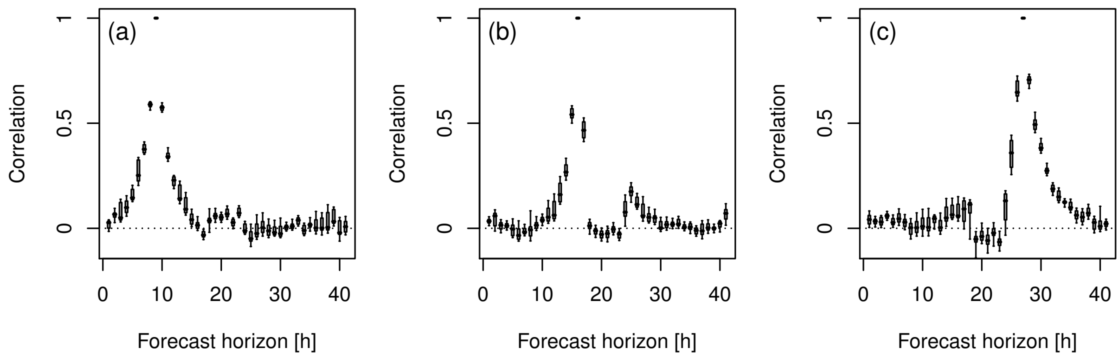

Figure 4.3: Examples of time-lagged correlation coefficients, which are the elements of the forecast error correlation matrix $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$. Error correlation associated with forecasts with a horizon of 9 hours (a), 16 hours (b), and 27 hours (c). The boxplots indicate the variability within the 2 month test period.

of the error is above 0.5 for time lags of ± 1 hour and can still be significant for a time lag up to 10 hours. Moreover, the fairly small boxplots show that the autocorrelation in time is stable over the different training samples.

The d-ECC approach is a process which starts with the application of the ECC method: a post-processed forecast $\tilde{\boldsymbol{x}} = \left\{\tilde{\boldsymbol{x}}^1, ..., \tilde{\boldsymbol{x}}^M\right\}$ is generated using the original ensemble $\boldsymbol{x} = \left\{\boldsymbol{x}^1, ..., \boldsymbol{x}^M\right\}$ as reference template. The difference between the original and the post-processed forecast is computed for each scenario and adjusted using the correlation matrix $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$ assuming stationarity of the forecast error autocorrelation. A new set of reference template $\boldsymbol{z} = \left\{\boldsymbol{z}^1, ..., \boldsymbol{z}^M\right\}$ is derived following:

$$\boldsymbol{z}^m = \boldsymbol{x}^m + \hat{\boldsymbol{R}}_{\boldsymbol{e}}^{\frac{1}{2}}(\tilde{\boldsymbol{x}}^m - \boldsymbol{x}^m). \tag{4.10}$$

with $m \in \{1, ..., M\}$. The second term in Eq. (4.10) corresponds to an adjustment of each member correction that involves the square root of the error correlation matrix. This adjustment resembles a *colouring transformation*, a signal processing technique that allows adapting the covariance of a multivariate random variable. The dependence structure of $\boldsymbol{z}$ and the marginal quantiles $\boldsymbol{q}$ are then used to derive the d-ECC scenarios following Eqs (4.7) and (4.8).

Scenarios derived by d-ECC and ECC are compared when applied to calibrated COSMO-DE-EPS global radiation forecasts. The two scenario-generation techniques are combined with two types of calibration: the standard and the weather-dependent calibration approaches presented in Sections 4.1 and 4.2, respectively. Four sets of calibrated scenarios are thus generated computing the following experiments: (1) standard calibration with QR and scenarios derived by ECC, (2) standard calibration with QR and scenarios derived by d-ECC, (3) weather-dependent calibration with PQR and scenarios derived by ECC, and (4) weather-dependent calibration with PQR and scenarios derived by d-ECC. An example of calibrated scenarios derived by this last combination of post-processing steps is shown in Figure 2.2(d).

The assessment of these multivariate forecasts requires adequate scores. The Energy Score (ES) and the p-Variogram Score (p-VS) are complementary multivariate scores that have demonstrated to be proper (Gneiting *et al.*, 2008; Scheuerer and Hamill, 2015). ES is a generalisation of the CRPS to the multivariate case whereas p-VS is based on the geostatistical concept of variogram. Considering an ensemble with M scenarios $\boldsymbol{x}^m$ with $m \in \{1, ..., M\}$ and an observed scenario $\boldsymbol{y}$, ES is defined as:

$$\text{ES} = \frac{1}{M} \sum_{m=1}^{M} \|\boldsymbol{y} - \boldsymbol{x}^m\| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \|\boldsymbol{x}^m - \boldsymbol{x}^{m'}\| \tag{4.11}$$

where $\|.\|$ represents the Euclidean norm, and and p-VS is defined as:

$$\text{p-VS} = \sum_{t_1 \neq t_2} \mu_{t_1 t_2} \left(\mid y_{t_1} - y_{t_2} \mid^{\text{p}} - \frac{1}{M} \sum_{m=1}^{M} \mid x_{t_1}^m - x_{t_2}^m \mid^{\text{p}} \right)^2 \tag{4.12}$$

where p is the order of the variogram and μ_{kl} are appropriate weights, respectively. The computation of p-VS is performed using the variogram order $\text{p} = 1$ and $\text{p} = 2$, and the weights $\mu_{t_1 t_2}$ are chosen proportional to the inverse squared distance in time:

$$\mu_{t_1 t_2} = \frac{1}{(t_1 - t_2)^2}, \quad t_1 \neq t_2, \tag{4.13}$$

where t_1 and t_2 denote the index of two forecast horizons. Energy skill score (ESS) and p-variogram skill score (pVSS) are defined similarly as the CRPSS, replacing in Eq. (3.19) CRPS and CRPS* by the corresponding multivariate scores applied to the forecast under assessment and to a reference forecast, respectively.

Besides the multivariate scores, the statistical consistency of the multivariate forecasts is investigated by diagnostic tools. The Averaged Rank Histogram (ARH) and the Band Depth Rank Histogram (BDRH) are two variants of rank histograms applied to multi-dimensional fields (Thorarinsdottir *et al.*, 2014). A rank histogram, in the univariate case, assesses the rank of the observations within the ensemble (Hamill, 2001). In the multivariate case, pre-ranks from multivariate forecasts have to be computed. For this purpose, ARH considers the rank, which is averaged over the multivariate aspect. The interpretation of the derived histograms is the following: forecasts leading to a flat rank histogram are interpreted as statistical consistent while histograms with a ∪-shape or a ∩-shape are interpreted as indications of underdispersiveness or overdispersiveness of the forecasts, respectively. On the other side, BDRH assesses the centrality of the observation within the ensemble based on the concept of functional band depth. In that case, a ∪-shape is associated with underestimated correlation, a ∩-shape with overestimated correlation in the ensemble, a skewed rank histogram to bias or dispersion errors and a flat rank histogram to calibrated forecasts.

Figure 4.4 shows the results of the four experiments in terms of the multivariate skill scores ESS and pVSS using the original ensemble forecast as reference. First, the positive impact of post-processing as a whole appears clearly for the different scores and is significant for all experiments. Second, the marginal calibration step has a strong influence on the performance of the final calibrated scenarios. Third, d-ECC clearly outperforms ECC when the first post-processing step is not weather-dependent but the superiority is not statistically significant. The benefit of the dual approach is smaller when the calibration step accounts for weather-dependent biases in the forecast. Weather-dependent calibration improves the resolution of the forecast and thus provides calibrated forecasts which are sharper. In that case, ECC is able to generate more realistic scenarios because the spread is increased to a smaller extent and not applied indiscriminately, as it is the case for a standard calibration.

Figure 4.5 shows the results of the four experiments in terms of the multivariate rank histograms ARH and BDRH. First, Figures 4.5(a) and 4.5(f) confirm that the original ensemble suffers from underdispersiveness and requires calibration. Second, the d-ECC scenarios are better calibrated than the respective ECC scenarios which exhibit remaining biases and a lack of correlation. d-ECC combined with a standard calibration or with a weather-dependent calibration technique provides reasonably well-calibrated scenarios as shown in Figures 4.5(c) and 4.5(h), and Figures 4.5(e) and 4.5(j),

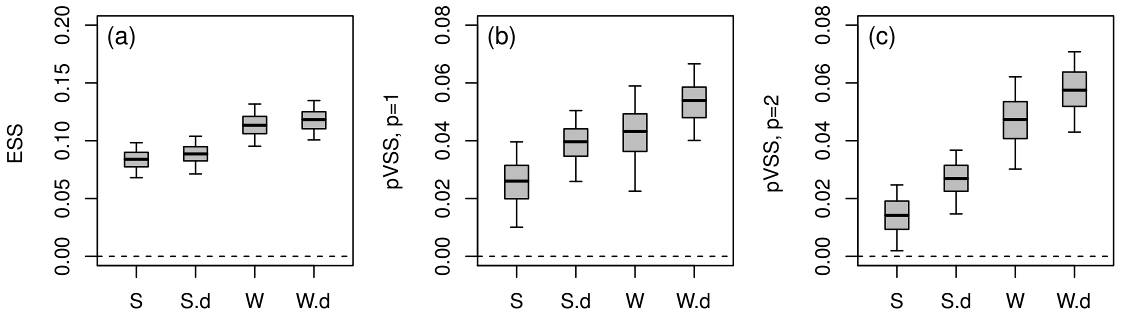

Figure 4.4: Multivariate skill scores of COSMO-DE-EPS global radiation forecasts: energy skill score (a), p-variogram skill score with $p = 1$ (b), and p-variogram skill score with $p = 2$ (c) applied to standard calibrated scenarios combined with ECC (S), standard calibrated scenarios combined with d-ECC (S.d), weather-dependent calibrated scenarios combined with ECC (W), and weather-dependent calibrated scenarios combined with d-ECC (W.d). The reference forecast in all cases is the raw ensemble forecast. The confidence intervals are estimated by block bootstrapping. Results for the period July-August 2015.

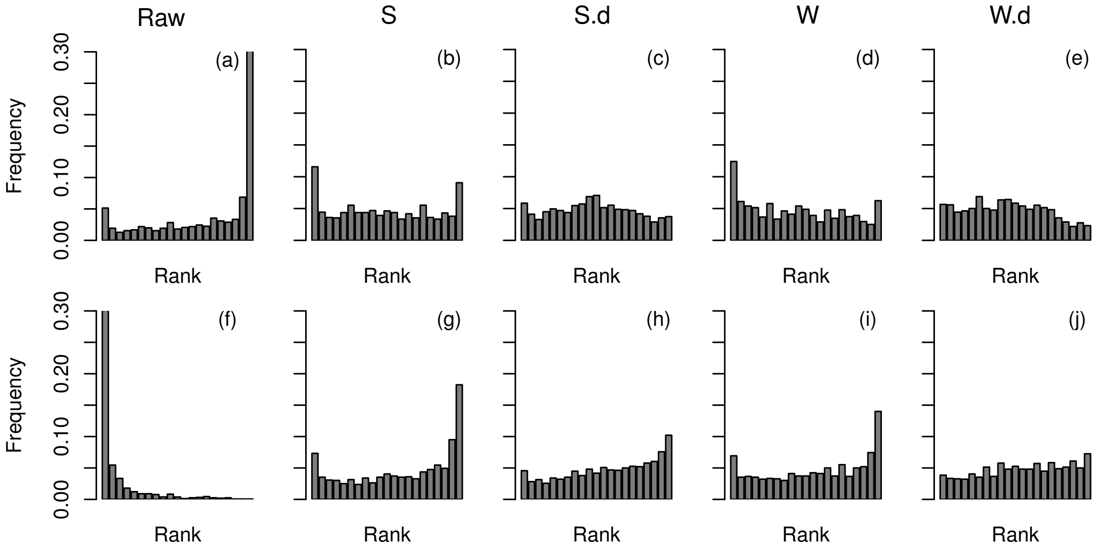

Figure 4.5: Multivariate rank histograms of COSMO-DE-EPS global radiation forecasts: average rank histograms (a,b,c,d,e) and band depth rank histograms (f,g,h,i,j) for time series scenarios. The scenarios correspond to the raw ensemble forecast (a,f), standard calibrated scenarios combined with ECC (b,g), standard calibrated scenarios combined with d-ECC (c,h) weather-dependent calibrated scenarios combined with ECC (d,i), and weather-dependent calibrated scenarios combined with d-ECC (e,j). Results for the period July-August 2015.

respectively. Nevertheless, the combination of PQR and d-ECC provides the best results in terms of multivariate reliability.

5. Summary and conclusion

Power forecasts are essential for an efficient and cost-effective grid integration of renewable energies. For a time horizon of few hours up to several days, renewable energy applications are commonly based on the outputs of numerical weather prediction (NWP) systems. Solar and wind forecasting cover various spatial and temporal scales and, moreover, reserve management and trading activities have to account for users with different risk aversions. Therefore, reliable probabilistic information that is physically consistent across time, space and variables is needed for the optimisation of the users' decision-making processes.

In NWP, ensemble forecasting is a standard method to provide flow-dependent uncertainty information. Possible sources of error in the forecast can be accounted for in the ensemble setup by means of variations in the initial conditions and perturbations of the model physics. At DWD, a high-resolution ensemble system has been running operationally since 2012. COSMO-DE-EPS is a 20-member ensemble based on a multi-analysis and multi-physics setup. First developed focusing on high impact weather events, the ensemble system today aims to be used for renewable applications in order to support the integration of weather-dependent energy sources in the national power grid.

However, probabilistic products directly derived from the ensemble system are usually not fully reliable which is a necessary condition for the optimisation of decision-making processes. The assessment and the statistical post-processing of the ensemble forecast are fundamental in this context. The potential and deficiencies of the ensemble system are evidenced through a verification process while post-processing aims to correct for the statistical deficiencies in order to reveal the full potential of the ensemble forecast.

In this study, COSMO-DE-EPS performances are investigated focusing on global radiation, which is the main weather variable affecting solar power production. The innovative methods proposed and implemented in this manuscript are not specific to solar applications but are applicable to other variables of interest, like wind for example. A two month period of summer 2015 is used as illustrative dataset. Verification results are based on forecasts of the 03 UTC runs with a horizon recently extended to 45 hours. Quality checked pyranometer measurements from 24 stations are used as observations.

Ensemble predictive performance

Quantile forecasts are appropriate probabilistic products for the optimisation of decision making and risk management in many renewable energy applications. The quantile score (QS) is the natural scoring rule for the evaluation of the quantile forecast quality. The assessment of the ensemble forecast as a whole can be performed based on the Continuous Ranked Probability Score (CRPS), which is a common score for the assessment of predictive distributions. The CRPS can be expressed as a weighted sum of QS applied to the sorted ensemble members. The decomposition of the proper scores QS and CRPS allows an interpretation of the verification results in terms of statistical consistency and information content. However, from the user's point of view, adequate verification tools are required for the assessment of the forecast value, which can substantially differ from the forecast quality. From the point of view of the developer, the added value of the ensemble forecast with respect to a single forecast has also been investigated.

In Appendix A, tools for the assessment of quantile forecast discrimination ability and value are proposed. The performance of a forecast is analysed for a specific user in a decision-making framework. The so-called relative user characteristic (RUC) curve and the quantile value plot extend existing concepts for the verification of probability forecasts to their equivalent for the case of quantile forecasts. The relationship between the overall value of a quantile forecast and the respective quantile skill score is also clarified.

In Appendix B, a new metric called ensemble added value (EAV) is proposed in order to assess the potential benefit of the ensemble approach with regards to a single run approach. The comparison relies on a probabilistic interpretation of the ensemble and deterministic forecasts where both forecasts can be optimised as a function of the user's needs. So, the EAV takes the form of a skill score with a focus on the forecast information content.

Based on standard and new tools, the investigation regarding the performance of COSMO-DE-EPS global radiation forecasts can be summarised as follows:

- the ensemble forecast suffers from statistical inconsistencies, which affect the forecasts at all horizons. The decomposition of the QS and the associated reliability diagrams show that negative and positive biases are associated with low and high probability levels, respectively, while quantile forecasts at intermediate levels are well calibrated. Thus, underdispersiveness and biases in the system affect the reliability of the derived probabilistic products.

- in terms of information content, COSMO-DE-EPS demonstrates to bring valuable input for decision-making processes. Results are based on the innovative tools dedicated to the assessment of probabilistic forecast in terms of quantiles from the user's perspective. Conditioned on calibration, the forecast value is higher in the ensemble-derived forecast case than in the case of forecasts based on a single run, for all types of users and all investigated events of interest.

- from the developer's perspective, the ensemble added value of COSMO-DE-EPS global radiation forecasts is significantly positive for all forecast horizons and tends to increase with the lead time. Thus, the ensemble-derived forecasts outperform purely statistical probabilistic products based on a deterministic forecast.

Ensemble calibrated scenarios

Statistical post-processing is required in order to correct for statistical deficiencies, which typically affect ensemble forecasts of surface variables. Learning from past errors, the forecasts are corrected aiming at providing reliable probabilistic products. Statistical methods should also preserve the advantage of the ensemble approach, which provides flow-dependent information with a fully consistent view of the forecast uncertainty across time, space and weather variables.

In Appendix C, quantile regression (QR) is shown to be an appropriate method for the calibration of ensemble global radiation forecasts. A standard setup of predictors includes the first guess quantile forecasts and a variable describing the solar geometry. With an enhanced technique called penalized quantile regression (PQR), adequate meteorological predictors are selected among a pool of ensemble model outputs. The enhanced approach leads to a weather-dependent calibration of the ensemble forecasts which can improve the discrimination ability of the forecasts.

In Appendix D, a new method for the generation of consistent scenarios based on calibrated ensemble forecasts is proposed. While the ensemble copula coupling (ECC) preserves the rank structure of the original ensemble forecasts after post-processing, the new approach called dual ensemble copula coupling (d-ECC) aims to combine information from the original ensemble forecast and from past error statistics. Thus accounting for the autocorrelation of the forecast error, d-ECC allows generating realistic scenarios.

The results regarding the application of these post-processing techniques to COSMO-DE-EPS global radiation can be summarised as follows:

- statistical inconsistencies of the ensemble forecasts can be corrected with QR or PQR. Thereby, the performance of the ensemble forecasts is considerably improved. The CRPSS comparing the forecast before and after calibration with QR is of about 30% and is significantly greater than 10% over all forecast horizons.
- using as benchmark probabilistic products derived by QR applied to deterministic forecasts, calibrated ensemble forecasts are about 10% better and the benefit of the ensemble approach tends to increase with the forecast horizon. The CRPSS emerging from this comparison is the practical equivalent of the theoretical EAV.
- meaningful predictors can be automatically selected by PQR. The weather-dependent calibration increases the information content of the calibrated forecasts and thereby the calibrated forecast value.
- applying d-ECC, statistically consistent time-series scenarios can be generated. The d-ECC approach outperforms ECC and is able to perform well independently of the choice of the marginal calibration method. Thus, a two-step procedure which consists in applying consecutively PQR and d-ECC provides an enhanced framework for the post-processing of ensemble global radiation forecasts.

Appendix A

Quantile forecast discrimination ability and value

The content of this appendix is the submitted author's version of a manuscript published in 2015 in the Quarterly Journal of the Royal Meteorological Society with reference:

> *Ben Bouallègue Z., P. Pinson and P. Friederichs, 2015 : Quantile forecast discrimination ability and value. Q.J.R. Meteorol. Soc. 141: 3415–3424.*

N.B.: mathematical symbols and notations of the original text have been freely adapted in order to be consistent with the rest of the manuscript at hand.

QUANTILE FORECAST DISCRIMINATION ABILITY AND VALUE

ZIED BEN BOUALLÈGUE[a,b], PIERRE PINSON[c], PETRA FRIEDERICHS[b]

[a] Deutscher Wetterdienst, Offenbach, Germany
[b] Meteorological Institute, University of Bonn, Germany
[c] Technical University of Denmark, Denmark

Abstract

While probabilistic forecast verification for categorical forecasts is well established, some of the existing concepts and methods have not found their equivalent for the case of continuous variables. New tools dedicated to the assessment of forecast discrimination ability and forecast value are introduced here, based on quantile forecasts being the base product for the continuous case. The relative user characteristic (RUC) curve and the quantile value plot allow analysing the performance of a forecast for a specific user in a decision-making framework. The RUC curve is designed as a user-based discrimination tool and the quantile value plot translates forecast discrimination ability in terms of economic value. The relationship between the overall value of a quantile forecast and the respective quantile skill score is also discussed. The application of these new verification approaches and tools is illustrated based on synthetic datasets, as well as for the case of global radiation forecasts from the high resolution ensemble COSMO-DE-EPS of the German Weather Service.

A.1 Introduction

Verification of probabilistic weather forecasts is an area of intensive research and growing interest as ensemble forecasting is becoming a standard approach in numerical weather prediction. Ensemble prediction systems (EPS) issue a sample of possible future states of the atmosphere (Lewis, 2005; Leutbecher and Palmer, 2008). The forecasts can be interpreted in the form of a predictive distribution and probabilistic products can be derived in order to support and optimize forecast-based decision-making (Krzysztofowicz, 1983). Appropriate tools for the assessment of probabilistic products from this perspective are therefore essential.

Such tools already exist for probabilistic products expressed in the form a probability forecast for a defined event. The relative operating characteristic (ROC) curve is a common verification tool for the assessment of probability forecasts (Mason, 1982). The ROC curve is related to decision-making analysis and the corresponding fundamental property of the forecast is called *discrimination*. Forecast discrimination assesses whether the forecast can be used to successfully discriminate between the observations (Murphy, 1991) or, said differently, whether appropriate decisions can be taken based on a forecast. Discrimination is translated in terms of *economic value* using a simple cost-loss model that allows the specificity of a user to be taken into account through the definition of a *cost-loss ratio*. The derived quantitative measure is called *value score* or *relative value* and is usually represented in the form of a probability value plot showing the forecast value as a function of the user's cost-loss ratio

(Richardson, 2000; Wilks, 2001; Zhu *et al.*, 2002). The value of a forecast is defined as the benefit to a user as a result of making decisions based on a forecast and has to be distinguished from forecast quality, the overall agreement between forecast and observation (Murphy, 1993). In a verification process, value and quality can be seen as being from the point of view of the forecast user and from the point of view of the forecast provider, respectively. The distinction between the two *types of goodness*, value and quality, is crucial since a non-linear relationship between them can lead to situations where a large improvement in the forecast quality does not imply an increase in the forecast value, or conversely, a small improvement in forecast quality can bring a notable benefit in terms of forecast value (Chen *et al.*, 1987; Buizza, 2001; Pinson, 2013).

Probabilistic products can be expressed in terms of a probability when the focus is on a particular event of interest, but also in terms of a quantile when the focus is on a particular probability level of interest. While a probability forecast first requires the definition of an event, i.e. the categorization of the original information, a quantile forecast is a 'single-valued' forecast expressed in the unit of the variable being forecast. Considering here probabilistic products derived from EPS simulations for continuous variables, such as temperature, wind speed or global radiation, quantile forecasts of the predictive distributions allow one to work with a continuous forecast as the original one by defining a nominal probability level. The choice of a probability level is directly related to the user's loss function: a quantile forecast at a given probability level is the optimal forecast for users with a specific asymmetry in their loss function (Koenker and Machado, 1999; Friederichs and Hense, 2007; Gneiting, 2011a). Asymmetric loss functions find numerous real-world applications, in particular in the renewable energy sector (Pinson *et al.*, 2007; Pinson, 2013). For example, consider solar energy producers who have to agree in advance about the amount of energy to be provided to their customer. If too much energy is produced, they will sell the extra at a reduced price, while if they do not produce enough energy they will be heavily penalised. The asymmetry in the producer's loss function is drawn from the different penalties associated to over- and under-forecasting.

Based on the relationship between user's loss function and quantile forecast level, the quantile score (QS) is the natural scoring rule for assessing the quality of quantile forecasts (Koenker and Machado, 1999; Friederichs and Hense, 2007; Gneiting, 2011a). More recently, the verification of quantile forecasts has benefited from the tradition and concepts stemming from the probability forecast verification framework. It has been shown that QS is a *proper* scoring rule and a decomposition of the score has been proposed (Bentzien and Friederichs, 2014). The QS decomposition provides information about *reliability* and *resolution*, two other fundamental attributes of a probabilistic forecast (Toth *et al.*, 2003).

The aim of the paper at hand is to extend the range of verification methods dedicated to the assessment of quantile forecasts. In particular, the assessment of quantile forecasts from the user's perspective, in a decision-making framework, is explored here. Based on a simple cost-loss model, the concepts of forecast discrimination and forecast value are revisited focusing on a specific user rather than on an specific event. First, a new tool is proposed for the analysis of user-based discrimination. The so-called relative user characteristic (RUC) curve and the associated summary measure are shown to be adequate for the assessment of quantile forecast discrimination ability. Secondly, quantile forecast value is discussed as an application of the value score to quantile forecasts. The quantile value plot, showing the economic value of a forecast as a function of a range of events of interest, is proposed as a new tool for the visualization of quantile forecast performance. Finally, the relationship between quantile forecast value and quantile skill score is discussed in the same vein as the relationship between probability forecast value and Brier skill score (Murphy, 1969). The concepts developed are first illustrated with the help of synthetic datasets and in a second step applied to probabilistic forecasts derived from an EPS.

The manuscript is organized as follows: Section A.2 describes the datasets that are used to illustrate the discussion. Section A.3 introduces definitions and notations and describes the relationship between quantile forecast and forecast user within a cost-loss model framework. Section A.4 discusses the concept of discrimination and Section A.5 the application of the economic value score to quantile forecasts. Section A.6 presents the conclusions.

A.2 Data

Synthetic datasets

In order to illustrate the concepts discussed hereafter, we make use of synthetic and real datasets. The synthetic data are derived from a toy-model based on normal distributions often used to illustrate verification discussions (e.g. Hamill, 2001; Weigel, 2011). The toy-model is kept simple in order to facilitate the interpretation of the results.

We consider a signal h, normally distributed, written $h \sim \mathcal{N}(0, 1)$. We assume that the observations are randomly drawn from a distribution $\mathcal{N}(h, 1)$ and the associated predictive distribution described by $\mathcal{N}(h+b, \sigma)$ where b is the unconditional bias parameter and σ the dispersion parameter. We define the following test-cases:

A_0 : $b = 0$, $\sigma = 1$ (a perfect probabilistic forecast) ,

A_1 : $b = -0.75$, $\sigma = 1$ (a biased forecast),

A_2 : $b = 0$, $\sigma = 1/3$ (an underdispersive forecast),

B : $b = \epsilon_B$, $\sigma = 1$ (a forecast with white noise),

where ϵ_B is derived from a uniform distribution defined on $[-5, 5]$. The first three datasets A_0, A_1 and A_2 differ only in terms of biases while the fourth dataset B corresponds to a forecast with a dynamically disturbed signal. The simulation setup, the values associated to b, σ and ϵ_B, has been chosen in order to make clear the reading of the Figures in Section A.4 and A.5 but does not affect the general interpretation of the simulation results.

COSMO-DE-EPS

Real datasets are provided by COSMO-DE-EPS, a regional ensemble prediction system run operationally at Deutscher Wetterdienst, Offenbach, Germany. The ensemble system is based on a 2.8 km grid resolution version of the COSMO model (Steppeler *et al.*, 2003; Baldauf *et al.*, 2011) with a model domain that covers Germany and parts of the neighbouring countries. The ensemble comprises 20 members including variations in initial conditions, physics parameterisations and boundary conditions (Gebhardt *et al.*, 2011; Peralta *et al.*, 2012).

COSMO-DE-EPS has been first developed focusing on high-impact weather events (Ben Bouallègue *et al.*, 2013; Ben Bouallègue and Theis, 2014) and is planned to be used for energy-applications. The focus in this paper is on global radiation (the sum of direct and diffuse shortwave radiations) which is the main weather variable affecting solar energy forecasts. Verification is applied to the 0300UTC run with a forecast horizon ranging between 5 and 15 hours. Two periods of 3 months are compared: winter (December, January, February) 2012/2013 and summer (June, July, August) 2013. The observation dataset consists of pyranometer measurements from 32 stations distributed over Germany and quality controlled (Becker and Behrens, 2012).

Global radiation forecasts and observations are transformed into clearness index before verification. The clearness index is defined as the ratio between global radiation at ground and global radiation

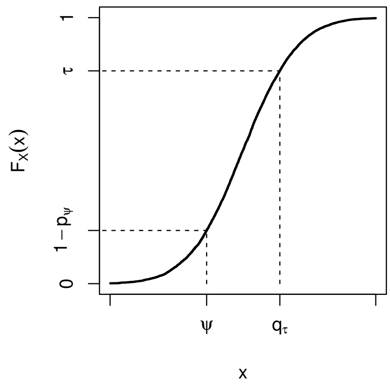

Figure A.1: Example of a predictive cumulative distribution function $F_X(x)$. Probabilistic products are derived either fixing a threshold ψ and deriving the associated probability forecast p_ψ, or fixing a probability level τ and deriving the associated quantile forecast q_τ.

at the top of the atmosphere (Badescu, 2008). This pre-processing of the data allows climatological effects and misinterpretation of the verification results to be avoided (Hamill and Juras, 2006).

A.3 Definitions and framework

Quantile forecast, quantile score, and quantile skill score

We first consider the quantity to be forecast (or *observation*) $Y \in \Re$ that we assume to be a continuous random variable driven by a stochastic process. An observed event E is defined by a threshold ψ as $E : Y \geq \psi$. The base rate π of an event E (or climatological frequency) corresponds to:

$$\pi = Pr(Y \geq \psi). \tag{A.1}$$

Consider now a predictive cumulative distribution $F_X(x)$. The probability forecast p_ψ of event E is defined as:

$$p_\psi = 1 - F_X(\psi). \tag{A.2}$$

The quantile forecast q_τ at probability level τ ($0 \leq \tau \leq 1$) is defined as:

$$q_\tau := F_X^{-1}(\tau) = \inf\{x : F_X(x) \geq \tau\} \tag{A.3}$$

such that the relationship between a probability forecast and a quantile forecast is expressed as:

$$p_{q_\tau} = 1 - \tau. \tag{A.4}$$

Figure A.1 shows an example of a cumulative distribution function $F_X(x)$. A threshold ψ and the associated probability forecast $1 - p_\psi$ as well as a probability level τ and the associated quantile forecast q_τ are shown on the plot.

The quantile score (QS) is the scoring rule applied in order to assess the quality of a quantile forecast. QS is based on an asymmetric piecewise linear function ρ_τ called the check function. The check function was first defined in the context of quantile regression (Koenker and Bassett, 1978):

$$\rho_\tau(u) = u[\tau - I(u < 0)] = \begin{cases} \tau u & \text{if} \quad u \geq 0 \\ (\tau - 1)u & \text{if} \quad u < 0 \end{cases} \tag{A.5}$$

where $I(.)$ is an indicator function having value 1 if the condition in parenthesis is true and zero otherwise. QS results from the mean of the check function applied to the pairs $i = 1, ..., N$ of observation y_i and quantile forecast $q_{\tau,i}$ following

$$QS = \frac{1}{N} \sum_{i=1}^{N} \rho_\tau(y_i - q_{\tau,i}), \tag{A.6}$$

where N is the size of the verification sample. Developing Eq. (A.6) we can write

$$QS = \frac{1-\tau}{N} \sum_{i:y_i<q_{\tau,i}} (q_{\tau,i} - y_i) + \frac{\tau}{N} \sum_{i:y_i \geq q_{\tau,i}} (y_i - q_{\tau,i}) \tag{A.7}$$

The scoring rule consists of penalties $1 - \tau$ and τ per unit of Y associated with under-forecasting and over-forecasting, respectively.

Skill scores are computed in order to measure the relative benefit of using a forecast compared to a reference forecast (Wilks, 2006b). The quantile skill score (QSS) measures the skill of a quantile forecast compared to a reference quantile forecast. Considering the climatology as reference, QSS corresponds to:

$$QSS = \frac{QS_{\text{forecast}} - QS_{\text{climate}}}{QS_{\text{perfect}} - QS_{\text{climate}}} = 1 - \frac{QS_{\text{forecast}}}{QS_{\text{climate}}} \tag{A.8}$$

where QS_{forecast}, QS_{perfect} and QS_{climate} represent the quantile scores of the forecast under assessment, of a perfect deterministic forecast and of a climatological τ-quantile forecast, respectively. QS_{perfect}, by definition, equals 0 and a climatological τ-quantile forecast, noted y_τ, is here defined as the τ-quantile of the observation distribution over the verification sample. Using the sample climatology for the reference forecasts, one should be aware, interpreting the results, that it produces a disadvantage for the forecasts, especially as the sample size is relatively small.

Cost-loss model and optimal decision-making

The framework used to discuss the concept of *use*r and *decision-making* is based on a static cost-loss model (Thompson, 1962; Katz and Murphy, 1997). The cost-loss model describes situations of dichotomous decisions: a user has to decide whether or not to take protective action against potential occurrence of an event E. The decision is made based on a decision variable (or forecast) Λ. A decision criterion λ applied to the decision variable defines an action $A : \Lambda \geq \lambda$. Taking action implies a cost C. In the case of occurrence of the event E without preventive action, a loss L is encountered. The cost-loss ratio is denoted α:

$$\alpha = \frac{C}{L}. \tag{A.9}$$

A user with cost-loss ratio α is called hereafter an α-user. Based on this simple model the optimal decision strategy of an α-user can be discussed (e.g. Richardson, 2011). The problem consists of finding, for a decision variable Λ, the *critical* decision criterion λ_α that minimizes the α-user mean expense if actions are taken when $\Lambda \geq \lambda_\alpha$.

Consider first the case of a probability forecast p_ψ as a decision variable. Based on p_ψ, does the user have to take action or not? In order to answer this question, the average expenses in the cases of positive and negative answers are compared. If the answer is yes, the user encounters a cost C on every occasion, so the average expense $\bar{E}_{\text{yes}}$ is simply

$$\bar{E}_{\text{yes}} = C. \tag{A.10}$$

If the answer is no, the user has no cost but a loss L on each occasion where the event occurs, so on average the user's expense $\bar{E}_{\text{no}}$ is

$$\bar{E}_{\text{no}} = LPr(Y \geq \psi \mid p_\psi), \tag{A.11}$$

where $Pr(Y \geq \psi \mid p_\psi)$ is the probability that the event occurs when the probability forecast p_ψ is issued. So, users with a cost-loss ratio $\alpha < Pr(Y \geq \psi \mid p_\psi)$ should take preventive action, while users with a greater cost-loss ratio should not. The critical decision criterion $p^\star_\psi$ associated with the decision variable p_ψ is thus defined as

$$p^\star_\psi = \{p_\psi \mid \Pr(Y \geq \psi \mid p_\psi) = \alpha\}. \tag{A.12}$$

Thus, the action based on the probability forecast $A : p_\psi \geq p^\star_\psi$ optimizes the user's mean expense in the long term.

If the forecast is reliable, we have by definition $\Pr(Y \geq \psi \mid p_\psi) = p_\psi$: the event actually happens with an observed relative frequency consistent with the forecast probability (Bröcker, 2009). The optimal decision is then to take action if

$$p_\psi \geq \alpha. \tag{A.13}$$

When the probability forecast is compared to the cost-loss ratio in order to decide whether or not to take action (without additional information about forecast reliability), we say that the probability forecast is taken at face value. For example, consider users who have to decide whether or not to take preventive action against precipitation occurrence. If the forecast probability of precipitation is 10%, users with cost-loss ratio lower than 10% take action. If the forecast is not reliable, the critical decision criterion is no longer α but has to be adjusted following Eq. (A.12). Statistical adjustments of the forecast based on past data is usually referred as *forecast calibration* (e.g. Gneiting *et al.*, 2007).

Consider now a quantile forecast q_τ as a decision variable. We apply the same reasoning as for a probability forecast. The critical decision criterion $q^\star_\tau$ associated with q_τ is defined as

$$q^\star_\tau = \{q_\tau \mid Pr(Y \geq \psi \mid q_\tau) = \alpha\} \tag{A.14}$$

such that taking action when $q_\tau \geq q^\star_\tau$ minimizes the user mean expense. By definition, a quantile forecast is reliable if it satisfies

$$Pr(Y \geq \psi \mid q_\tau = \psi) = 1 - \tau, \tag{A.15}$$

i.e. the observed relative frequency of the event defined by the quantile forecast is consistent with the quantile forecast probability level. Eq. (A.14) has a straightforward solution

$$q^\star_\tau = \psi \tag{A.16}$$

when the decision variable is the quantile forecast at probability level τ defined as

$$\tau = 1 - \alpha. \tag{A.17}$$

Taking action when $q_\tau \geq \psi$ with $\tau = 1 - \alpha$ is equivalent to taking action when $p_\psi \geq \alpha$ since the cumulative probability distribution function $F_X(x)$ is by definition monotonically increasing (see e.g. Figure A.1). Hence, a quantile forecast is taken at face value when the user's decision is made based on the comparison of the forecast with the event threshold ψ. In our example, if the 90%-quantile forecast of precipitation is greater than zero, a user with cost-loss ratio $\alpha = 1 - 0.9 = 0.1$ takes preventive action.

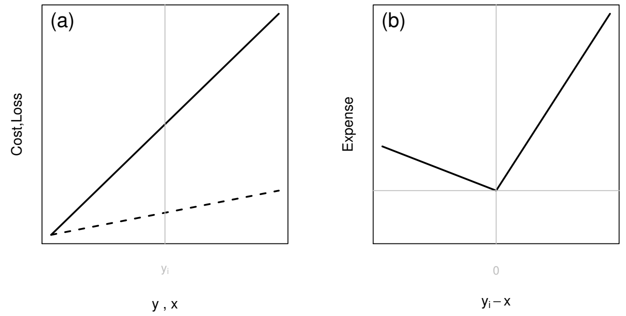

Figure A.2: (a) Cost (dashed line) as a function of the level of protection x and loss (full line) as a function of the observation Y. An observation y_i is represented by a vertical line. (b) Expense as a function of the difference between the observation y_i and the level of protection x. The horizontal line indicates the expense for a perfect level of protection.

In a general form, the critical decision criterion λ_α for an α-user is defined by

$$\lambda_\alpha = \{\lambda \mid Pr(Y \geq \psi \mid \lambda) = \alpha\} \tag{A.18}$$

where the decision variable could equally be the probability forecast p_ψ or the quantile forecast q_τ with $\tau = 1 - \alpha$. Provided that the forecasts are reliable, the critical decision criteria are known and have a simple expression (Eqs (A.13,A.16)). In the following, we say that the decision variable is taken at *face value* when the user applies the decision criterion valid for a reliable forecast, irrespective of whether the forecast is actually reliable or not.

Quantile forecast user

The dichotomous decision problem is extended to a continuous decision problem considering the cost C and the loss L as unitary cost and unitary loss, respectively (Epstein, 1969b; Roulston *et al.*, 2003). The cost of taking protection is a linear function of the level of protection x and the loss without protection is a linear function of the observation Y, as illustrated in Figure A.2. The optimization problem consists of finding the level of protection that minimizes the expected user expense.

Considering a variable defined on $\Re+$ (the generalization to variables defined on $\Re$ is straightforward), the expense associated with a level of protection x corresponds to Cx. If the observation is y, then protection is perfect if $x = y$. But if $x > y$, then there is an unnecessary expense due to a larger level of protection than is actually needed. If the observation y is greater than the level of protection, then a loss $L(y - x)$ is encountered and a cost $C(y - x)$ avoided. Formally, we can write the expense function E as

$$E = \begin{cases} C(x - y) & \text{if } y < x \\ (L - C)(y - x) & \text{if } y \geq x. \end{cases} \tag{A.19}$$

The expense function is represented in Figure A.2. If divided by L, the expense function is an asymmetric loss function equivalent to the check function defined in Eq. (A.5), where the asymmetry is given by $\tau = \frac{L-C}{L}$. Thus the optimal level of protection $x^\star$ which minimizes the user's mean expense corresponds to the $1 - \alpha$ quantile of the true predictive distribution of Y.

This result is not new: quantile forecasts arise as an optimal solution for users with an asymmetric linear loss function (Koenker and Bassett, 1978; Christoffersen and Diebold, 1997). More recently, it has been shown that quantile forecasts are optimal forecasts in a stochastic optimization framework for a more general class of loss functions (Gneiting, 2011b).

Asymmetric loss functions find a number of applications, in particular for operational decision-making problems related to the integration of renewable energies into the electricity grid. For example, asymmetric loss functions can be associated with market participants who want to optimize their bids or system operators who have to optimize their reserves. The different penalties associated to over- or under-forecasting draw the asymmetry in the user's loss function. The user's optimal forecast corresponds then to a specific quantile of the predictive distribution where the probability is defined by the user's cost-loss ratio (Pinson *et al.*, 2007; Pinson, 2013).

A.4 Discrimination

Based on the discussion developed in the previous Section, continuous decision making is seen in the following as a *continuum* of dichotomous decisions. For each threshold ψ of the event spectrum, the question is whether to take action for the next unit of the variable. The adequate decision for a user in order to minimize the expected expense is a function of his (her) cost-loss ratio as defined in Eq. (A.18). Moreover, the relationship between cost-loss ratio and quantile probability level, $\tau = 1 - \alpha$, makes implicit the cost-loss ratio α of a user as soon as the level τ of the quantile forecast used as decision variable is selected.

General verification framework

A general framework for forecast verification is based on the joint distribution of forecasts and observations (Murphy and Winkler, 1987). The overall agreement between forecasts and observations is called quality and is measured by scoring rules, like QS for quantile forecasts. In order to access more information about the forecast performance, two factorizations of the joint distribution, into conditional and marginal distributions, can be applied: the *calibration-refinement* (CR) factorization when conditioning on the forecasts and the *likelihood-base rate* (LBR) factorization when conditioning on the observations. Summary measures based on these two factorizations are associated with attributes, fundamental characteristics of the forecast. Reliability and resolution are derived from the CR factorization while discrimination is derived from the LBR factorization (Murphy and Winkler, 1992).

Here the focus is on discrimination, the key forecast attribute for decision-making processes. A general definition of discrimination is "the ability of a forecasting system to produce different forecasts for those occasions having different realized outcomes" (Wilks, 2006b). Discrimination assumes calibration of a forecast and so does not account for reliability discrepancies. Though biases can strongly affect the skill of a forecast, this can be seen as an advantage since reliability, unlike discrimination, can be improved by recalibration. Investigating discrimination of a forecast means therefore focusing on the *necessary condition for skill* (Jolliffe and Stephenson, 2005).

Discrimination assessment is here discussed in terms of *event* and *action* within the dichotomous decision framework. Regarding the LBR factorization, it is common practice to analyse discrimination in terms of hit rate HR and false alarm rate FAR defined as

$$\text{HR} = Pr(\Lambda \geq \lambda \mid Y \geq \psi) \tag{A.20}$$

and

$$\text{FAR} = Pr(\Lambda \geq \lambda \mid Y < \psi), \tag{A.21}$$

respectively. Actions $A : \Lambda \geq \lambda$ and events $E : Y \geq \psi$ are dichotomous, each presenting two alternatives, so HR and FAR can be easily derived from the construction of a 2×2 contingency table.

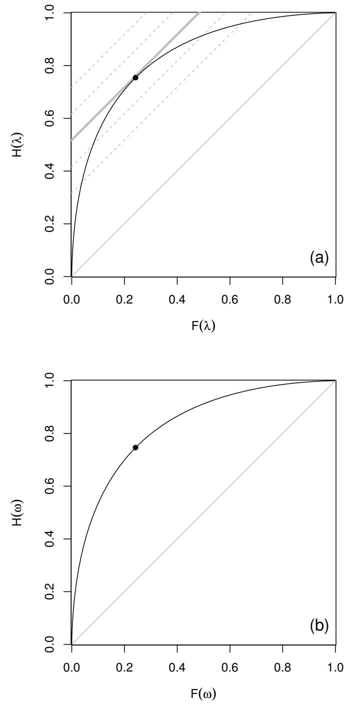

Figure A.3: Discrimination curves for decision variables from the synthetic dataset A_0. The diagonal lines are the no discrimination lines. The points correspond to the (HR,FAR) pair for the event $Y \geq 0$ and the action associated with the 50%-users. (a) ROC curve of the probability forecast p_0 for the event $E : Y \geq 0$, with base rate $\pi = 0.5$, and equi-cost lines (in grey) of slope $\gamma = 1$. (b) RUC curve of the quantile forecast $q_{0.5}$ for the user with cost-loss ratio $\alpha = 0.5$.

No discrimination corresponds to the case where:

$$\text{HR} = \text{FAR} \tag{A.22}$$

for all $\lambda \in \Lambda$ and $\psi \in Y$, meaning that actions and event occurrence are independent (Bröcker, 2014).

Event-based discrimination

We first focus on one particular event defined by a threshold ψ, with event-specific hit rate HR_λ and false alarm rate FAR_λ. A popular way to assess discrimination (Eq. A.22) is to plot the set of points $(\text{FAR}_\lambda, \text{HR}_\lambda)$ for a range of actions with $\lambda \in \Lambda$. The resulting curve is known as the relative operating characteristic (ROC) curve. When action and event occurrence are independent, the ROC curve is a diagonal line. Concavity of the curve indicates a discrimination ability in the forecast and the area under the curve (AUC) becomes a quantitative measure of forecast discrimination (Mason, 1982). Figure A.3 (a) shows an example of a ROC curve for the synthetic dataset A_0. The event of interest is $E : Y \geq 0$ with a base rate $\pi = Pr(Y \geq 0)$ of 0.5. The respective probability forecast $p_0 = 1 - F_X(0)$, the probability that simulation A_0 exceeds 0, is used as decision variable.

The interpretation of the ROC curve can be related to the dichotomous decision model described in Section A.3 as discussed for example in Richardson (2011). In order to describe this relationship, we consider the slope of the ROC curve, defining first the gradient of a line joining two successive ROC points $(\mathrm{FAR}_\lambda, \mathrm{HR}_\lambda)$ and $(\mathrm{FAR}_{\lambda+\Delta\lambda}, \mathrm{HR}_{\lambda+\Delta\lambda})$:

$$\frac{\mathrm{HR}_\lambda - \mathrm{HR}_{\lambda+\Delta\lambda}}{\mathrm{FAR}_\lambda - \mathrm{FAR}_{\lambda+\Delta\lambda}} = \frac{Pr(\Lambda \geq \lambda \mid Y \geq \psi) - Pr(\Lambda \geq \lambda + \Delta\lambda \mid Y \geq \psi)}{Pr(\Lambda \geq \lambda \mid Y < \psi) - Pr(\Lambda \geq \lambda + \Delta\lambda \mid Y < \psi)}. \tag{A.23}$$

The slope of the curve γ is obtained when $\Delta\lambda$ tends to 0:

$$\gamma(\lambda, \psi) = \frac{Pr(\Lambda = \lambda \mid Y \geq \psi)}{Pr(\Lambda = \lambda \mid Y < \psi)} \tag{A.24}$$

where the ratio is also know as the *likelihood ratio* (Bröcker, 2011). Using the Bayes rule and the definition of the critical decision criterion of an α-user in Eq. (A.18), we can write

$$\gamma(\lambda_\alpha, \psi) = \frac{1-\pi}{\pi} \frac{\alpha}{1-\alpha} \tag{A.25}$$

where $\pi = Pr(Y \geq \psi)$ is the base rate of an event $E : Y \geq \psi$ and λ_α the corresponding critical decision criterion of an α-user.

The range of decision criterion λ used to derive the ROC curve $(\mathrm{FAR}_\lambda, \mathrm{HR}_\lambda)$ corresponds to a range of critical decision criteria associated with users with different cost-loss ratios. Each point of the ROC curve is associated with a specific α-user that is identified by the slope of the curve at that point. The slope possibly ranges between 0 and $+\infty$ at the right-top and the bottom-left corners of the ROC plot respectively. Moving along the curve from the top to the bottom consists in varying the cost-loss ratio α between 0 and 1.

For example, consider a user with a cost-loss ratio $\alpha = 50\%$. In Figure A.3, the point of the ROC curve with slope $\gamma = 1$ is highlighted ($\alpha = 0.5, \pi = 0.5$ in Eq. A.25). This point indicates the performance of the forecast in terms of HR and FAR for this particular user. Conversely, the decision criterion applied to obtain this point corresponds to the critical decision criterion for the 50%-user.

The ROC curve applied to a decision variable, then, corresponds to testing whether actions and event occurrence are independent for one event and a range of users with different cost-loss ratios. The ROC curve is an *event specific* but *user unspecific* discrimination tool and is therefore well-adapted to probability forecast discrimination assessment.

User-based discrimination

We focus now on a user with cost-loss ratio α. The critical decision criterion λ_α defines the action of this specific user with respect to an event. We define then the user-specific hit rate HR_ψ and false alarm rate FAR_ψ as in Eqs (A.20) and (A.21) for a fixed α. In order to test Eq. (A.22), the set of points $(\mathrm{FAR}_\psi, \mathrm{HR}_\psi)$ are plotted for a range of events. We call the resulting curve a *relative user characteristic* (RUC) curve because it is a comparison of two user characteristics (FAR_ψ and HR_ψ) as the event definition varies. As for the ROC curve, the no discrimination line corresponds to the diagonal line and concavity of the curve indicates forecast discrimination ability.

Figure A.3 (b) shows an example of a RUC curve valid for a user with cost-loss ratio $\alpha = 50\%$. In this example, the decision variable is the 50%-quantile forecast from the synthetic dataset A_0. Moving along the RUC curve from the bottom left corner to the top right corner involves varying the event under focus, the event's base rate varying from 0 to 1, respectively. The point with slope $\gamma = 1$ corresponds to the event $E : Y \geq 0$ with base rate $\pi = 0.5$. This point is obviously the same as in Figure A.3 (a).

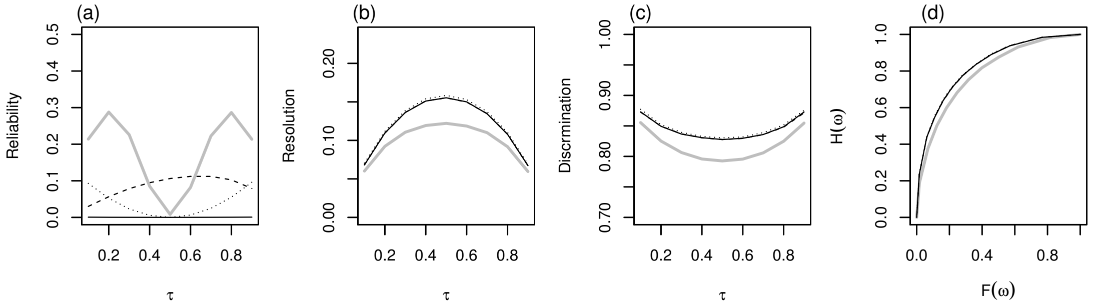

Figure A.4: (a) Reliability, (b) resolution and (c) discrimination as a function of the probability level τ of the τ-quantile forecasts and (d) RUC curves for the 50%-quantile forecasts ($\tau = 0.5$). The results are shown for the simulation test cases A_0 (full lines), A_1 (dashed lines), A_2 (dotted lines) and B (full grey line).

In order to produce a RUC curve, critical decision criteria have to be known for a range of events. They can be estimated resolving Eq. (A.14) numerically. In practice, critical decision criteria can also be estimated by means of a reliability diagram. For example, a reliability diagram for quantile forecasts plots the conditional observed quantile as a function of quantile forecast categories (Bentzien and Friederichs, 2014). With regard to Eq. (A.15), we can deduce that the mean forecast in each forecast category (horizontal axis of the diagram) is an estimation of the critical decision criteria associated with the events defined by the corresponding conditional observed quantile (vertical axis of the diagram).

The RUC curve is user specific (and event unspecific) and therefore well-adapted to quantile forecast discrimination. A summary measure of quantile discrimination ability is obtained mimicking the ROC framework: the area under the RUC curve, noted here AUC', is proposed as a quantitative measure of discrimination for quantile forecasts. Considering n_E events $E_i : Y \geq \psi_i$, $i = 1, ..., n_E$ with increasing base rate, AUC' is estimated by a trapezoidal approximation as

$$AUC' = \sum_{i=0}^{n_E} 0.5(\mathrm{HR}_{\psi_{i+1}} + \mathrm{HR}_{\psi_i})(\mathrm{FAR}_{\psi_{i+1}} - \mathrm{FAR}_{\psi_i}) \qquad \text{(A.26)}$$

with the trivial points $\mathrm{HR}_{\psi_0} = \mathrm{FAR}_{\psi_0} = 0$ (for an event of base rate 0) and $\mathrm{HR}_{\psi_{n_E+1}} = \mathrm{FAR}_{\psi_{n_E+1}} = 1$ (for an event of base rate 1). In order to reduce the biases introduced by the limited number of RUC points, the RUC curve can be fitted under a bi-normal assumption. The procedure involves considering FAR_ψ and HR_ψ as both expressed as integrations of the standard normal distribution (Mason, 1982). The bi-normal model has been shown to be valid in most cases when applied in the ROC framework (Mason and Graham, 2002; Atger, 2004).

The properties of the RUC curve and AUC' are discussed with the help of illustrative examples based on 4 simple simulation test cases (see Section A.2). In Figure A.4, the forecast attributes reliability, resolution and discrimination are shown as a function of the probability level τ of the τ-quantile forecast under assessment. RUC curves for the 50%-quantile forecasts are also shown. Quantile forecast reliability and resolution are estimated using the decomposition of the quantile score (Bentzien and Friederichs, 2014) while discrimination curves and summary measures are estimated based on the bi-normal assumption.

Figures A.4 (a) shows the lack of reliability, which occurs by construction in the simulations A_1, A_2 and B. In Figures A.4 (b) and A.4 (c), resolution and discrimination measures deliver a similar message comparing the different simulations which illustrates the idea that "resolution and discrimination are the two faces of the same coin" (Bröcker, 2014). Resolution and discrimination exhibit however different behaviours as a function of the probability level reflecting the fact that the first takes the forecaster's perspective and the second the user's perspective. Moreover, discrimination ability is identical for the simulations A_0, A_1 and A_2: they are unaffected by biases and dispersion errors. Indeed, AUC' is by construction insensitive to conditional and unconditional biases. In contrast, the forecast derived from simulation B with a perturbed signal presents less discrimination ability than forecasts from the other simulations, in particular for the 50%-quantile forecast. Focusing on users with cost-loss ratio $\alpha = 0.5$ $(\tau = 0.5)$, RUC curves for the 50%-quantile forecasts of simulations A_0, A_1, A_2, and B are shown in Figure A.4 (d). The largest discrepancies between simulations A and B are visible at the centre of the RUC curves, so for events with intermediate base rates, while for events with small or large base rates the RUC curves tend to overlap.

A.5 Value of quantile forecasts

Economic value

The cost-loss model described in Section A.3 has been used to develop the concept of economic value of a probabilistic forecast. The forecast value is assessed considering decision-making made by an α-user about the occurrence of an event. The value of a forecast (also called value score or relative value) is defined as

$$V = \frac{\bar{E}_{\text{climate}} - \bar{E}_{\text{forecast}}}{\bar{E}_{\text{climate}} - \bar{E}_{\text{perfect}}}, \tag{A.27}$$

where the mean expense $\bar{E}$ of an α-user is estimated when decisions are based on a forecast ($\bar{E}_{\text{forecast}}$), on a perfect deterministic forecast ($\bar{E}_{\text{perfect}}$), or on climatological information ($\bar{E}_{\text{climate}}$) (Richardson, 2000; Wilks, 2001; Zhu *et al.*, 2002). V is a measure of the economic gain (or reduction of mean expense) when using a forecast relative to the gain when using a perfect deterministic forecast.

Following *e.g.* Richardson (2011), the mean expense of a forecast user can be written as

$$\bar{E}_{\text{forecast}} = \text{FAR}(1-\pi)C - \text{HR}\pi(L-C) + \pi L, \tag{A.28}$$

where HR and FAR are the hit rate and false alarm rate as defined in Eqs (A.20) and (A.21), respectively, and π the base rate of the event of interest. A user with a perfect deterministic forecast at hand has to face costs only. The user mean expense corresponds in this case to:

$$\bar{E}_{\text{perfect}} = \pi C. \tag{A.29}$$

For a user who bases his (her) decision on climatological information, the optimal mean expense is expressed as

$$\bar{E}_{\text{climate}} = \begin{cases} C & \text{if} \quad \alpha < \pi \\ \pi L & \text{if} \quad \alpha \geq \pi, \end{cases} \tag{A.30}$$

depending on the relationship between cost-loss ratio and base rate. Combining Eqs (A.28)-(A.30), the value of a forecast can finally be written as:

$$V = \begin{cases} (1-\text{FAR}) - \left(\dfrac{\pi}{1-\pi}\right)\left(\dfrac{1-\alpha}{\alpha}\right)(1-\text{HR}) & \text{if} \quad \alpha < \pi \\ \text{HR} - \left(\dfrac{1-\pi}{\pi}\right)\left(\dfrac{\alpha}{1-\alpha}\right)\text{FAR} & \text{if} \quad \alpha \geq \pi. \end{cases} \tag{A.31}$$

So, the economic value V is defined for an event with base rate π and a user with cost-loss ratio α. V depends on the forecast performance in terms of HR and FAR.

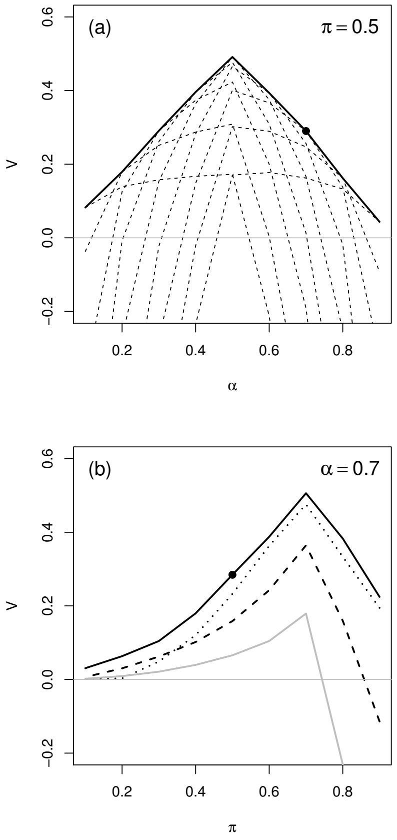

Figure A.5: (a) Potential value V of the probability forecast from simulation A_0 for the event defined as $E : \psi \geq 0$ with base rate $\pi = 0.5$. The dashed lines represent the forecast value when the probability levels $0.1, 0.2, ..., 0.9$ are chosen as decision criterion. The full line represents the envelope of the dashed lines. (b) Value V for users with cost-loss ratio $\alpha = 0.7$ of the 30%-quantile forecasts taken at face value from the 4 synthetic datasets: A_0 (full black line), A_1 (dashed line), A_2 (dotted line) and B (full grey line). The black point is the common point of the two plots: value of the simulation A_0 for the event with base rate $\pi = 0.5$ and a user with cost loss ratio $\alpha = 0.7$.

Applied to a probability forecast, the event's base rate is fixed and the value of a probability forecast is generally represented in the form of a probability value plot showing V as a function of α. An example is provided in Fig. A.5 (a), applied to simulation A_0 considering the event $E : Y \geq 0$. The forecast value curves are plotted for a range of probabilities as decision criterion, then the optimal values for each α-user (the upper envelope of the relative value curves) is selected to represent the value of the probabilistic forecast system (e.g. Richardson, 2000; Wilks, 2001). The probability value plot is related to the ROC framework since the pairs (FAR, HR) of Eq. (A.31) are the ones used to draw the ROC curve. It has also been shown that the overall value of a probability forecast, considering all potential users, corresponds to the Brier skill score of the forecast if the distribution of cost-loss ratio is uniform over all users (Murphy, 1969; Richardson, 2011).

Quantile value plot

Applied to a quantile forecast, so focusing on a α-user, the value score is evaluated for a range of events of interest defined for example by their base rate π. A new tool is therefore proposed for the assessment of quantile forecast performance: the quantile value plot which represents how V varies as a function of π. This is illustrated in Figure A.5 (b). The value of the 30%-quantile forecasts is plotted when the quantile forecasts derived from simulations A_0, A_1, A_2, and B are taken at face value. Taking a quantile at face value means using it as it is, so for each event it implies considering the event threshold as decision criterion (see Section A.3). An alternative is to apply the critical decision criteria, i.e. to use the $(\mathrm{FAR}, \mathrm{HR})$ pairs from the RUC curve to estimate the value in Eq. (A.31). We talk then about *potential* value since it corresponds to the maximum value of the forecast, i.e. the maximum that could be potentially reached if an adequate calibration is applied to the forecast. Indeed, value and *potential* value are by definition identical if the forecast is reliable.

A parallel between probability value plot and quantile value plot can be draw. In a probability value plot, the decision variable is a probability forecast, the base rate π of the event under focus is fixed and the forecast value V is then plotted for a range of cost-loss ratios. The role of α and π are inverted in order to produce a quantile value plot rather than a probability value plot. The cost-loss ratio is defined by the quantile probability level and a range of events of interest are scanned. It results that the cost-loss ratio of the end-user does not appear explicitly in a quantile value plot while it corresponds to the horizontal axis in a value plot for probability forecasts.

The fundamental properties of V are however the same when focusing on one event or on one user. These properties (demonstrations can be found e.g. in Richardson, 2011) are recalled here. First, the forecast value reaches its maximum when $\pi = \alpha$ (or noted differently when $\pi = 1 - \tau$). For instance, a forecast user with a cost-loss ratio of $\alpha = 0.1$ draws a maximum benefit from a forecast if his (her) event of interest has a climatological probability of occurrence of 10%. Secondly, the value of a reliable forecasts (full line in Figure A.5 (b)) is always greater than the value of the same forecast with biases (dashed and dotted lines in Figure A.5 (b)). The value of the reliable forecast corresponds to the potential value of the two other datasets. Finally, the potential value is by definition always non-negative.

A real example

The tools introduced for the assessment of quantile forecast discrimination and value are here applied to a real dataset. Quantile forecasts of global radiation are derived from COSMO-DE-EPS and assessed for two periods of the year 2013. Results for the winter period are shown in Figure A.6 and results for the summer period in Figure A.7. Quantile discrimination is estimated with the area under the RUC curve (AUC') for probability levels $\tau = 0.1, 0.2, ..., 0.9$. A deeper analysis is performed for the 10%-, 50%- and 90%-quantile forecasts with the help of quantile value plots.

The discrimination ability of the EPS quantile forecasts varies as a function of the probability level but is greater than 0.80 which can be interpreted as good performance. For the winter season, discrimination is higher for high and low probability levels than intermediate ones whereas for the summer season, discrimination is approximately constant over the probability levels with a tendency to decrease for high levels. Inspection of the quantile value plot allows a deeper insight into the forecast potential performance. This could be relevant for quantile users with a specific interest in only one part of the event spectrum. For example, consider a user with a non linear loss function, the loss becoming zero if the outcome is below (above) a given threshold T. The user can then interpret the quantile value plot focusing on the range of events below (above) T disregarding the rest of the plot. The potential value and the actual value of the COSMO-DE-EPS quantile forecasts are plotted as a function of event in terms of the clearness index in % to simplify the reading of the plots. An event

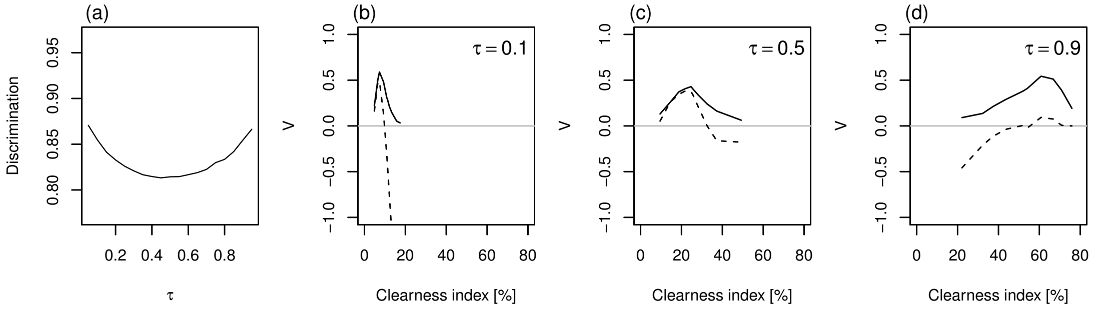

Figure A.6: Verification results for COSMO-DE-EPS global radiation forecasts during winter 2012/2013: quantile discrimination ability (AUC') as a function of the probability level (a), potential value (full line) and actual value (dashed line) of the 10%-quantile forecast (b), 50%-quantile forecast (c) and 90%-quantile forecast (d) as a function of the event of interest defined by thresholds of the clearness index in %.

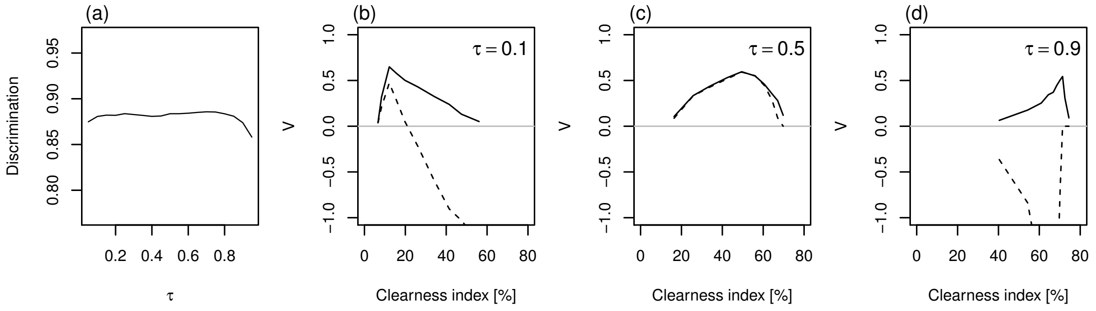

Figure A.7: Same as Figure A.6 but for summer 2013.

has a different base rate for each season which complicates a direct comparison of the quantile value plot in Figures A.6 and A.7. The comparison of the potential and actual values (full lines and dashed lines) shows the impact of the reliability discrepancies on the forecast value.

Overall value and Quantile Skill Score

As a final step in drawing a parallel between probability forecast verification and quantile forecast verification, the relationship between value and skill score with climatology as a reference is explored. It has been shown that the overall value of a probability forecast is equivalent to its Brier Skill Score (BSS) when the users have a uniform distribution of cost-loss ratio (Murphy, 1969; Richardson, 2011). Similarly, we now investigate the relationship between the overall value of a quantile forecast and its QSS.

For this purpose, we extend the cost-loss model to more than two observation categories assuming that the cost C and the loss L of the cost-loss model are the unitary increment of cost and loss per unit of variable, respectively, as discussed in Section A.3. Following Richardson (2011), the overall value is defined as the ratio

$$V_{all} = \frac{T_C - T_F}{T_C - T_P} \tag{A.32}$$

where the total mean expense T of a user is estimated when decisions are based on a climatological forecast (T_C), on a perfect deterministic forecast (T_P) or on a given forecast (T_F) so that Eq. (A.32) is the extension of Eq. (A.27) to all possible events.

The total expense for a perfect deterministic forecast corresponds to the sum of the costs C associated with each observation. The total mean expense T_P can then be expressed as

$$T_P = \frac{1}{N}\sum_{i=1}^{N} C y_i. \tag{A.33}$$

For a climatological quantile forecast y_τ, the total expense corresponds to the sum of the costs associated with y_τ and the losses encountered when the observations are greater than the climatological forecast ($y_i \geq y_\tau$). The total mean expense for a climatological forecast T_C is written as

$$T_C = \frac{1}{N}\sum_{i=1}^{N} C y_\tau + \frac{1}{N}\sum_{i:y_i \geq y_\tau} L(y_i - y_\tau). \tag{A.34}$$

Considering now a sample of quantile forecasts $q_{\tau,i}$ and the corresponding observations y_i, the total expense of a forecasts user corresponds in that case to the sum of the costs associated with each forecast $q_{\tau,i}$ and the losses encountered when $y_i \geq q_{\tau,i}$, given by

$$T_F = \frac{1}{N}\sum_{i=1}^{N} C q_{\tau,i} + \frac{1}{N}\sum_{i:y_i \geq q_{\tau,i}} L(y_i - q_{\tau,i}). \tag{A.35}$$

Combining Eqs (A.33)-(A.35), it is shown in the Appendix that the overall value V_{all} corresponds to QSS (Eq. A.8) with the climatology as a reference based on the assumption of constant cost-loss ratio for all outcomes. In other words, extending the dichotomous event-action framework to a continuous framework allows one to turn back to the ‘classical‘ or ‘natural‘ measure of performance for quantile forecast. Conversely, using the dichotomous framework provides the keys to making a deeper analysis of the quantile performance at the event level.

A.6 Conclusion

Verification measures and tools related to users’ decision-making are provided here for quantile forecasts as decision variables. Drawing a parallel with the verification of probability forecasts, the new verification tools allow the scuite of verification methods for quantile forecasts to be completed. In particular, the concepts of forecast discrimination and forecast value are discussed based on a simple cost-loss model.

First, the RUC curve is shown to be the counterpart of the ROC curve when the focus is on a given user rather than on a given event. The areas under the RUC and ROC curves are summary measures of discrimination adapted to quantile and probability forecasts, respectively. Both measures share the same properties, such as non-sensitivity to calibration.

Second, the translation of discrimination ability into value is explored with the help of the value score. The definition of the forecast value is directly adopted from the probability forecast verification framework. Forecast value and forecast potential value are estimated when the decision variable is a quantile forecast, so focusing on a user with a specific cost-loss ratio. The first is obtained when the forecast is taken at face value and the second when critical decision criteria are applied. The value of a quantile forecast can then be plotted as a function of a range of events of interest, defined for example in terms of base rates. The derived plot is called a quantile value plot and provides a valuable insight into the performance of a quantile forecast. As a real example, the discrimination ability and value of global radiation forecasts from COSMO-DE-EPS are demonstrated over a summer and a winter period.

Finally, it is shown that the overall value of a quantile forecast corresponds to the quantile skill score with climatology as reference when a constant cost-loss ratio for all outcomes is assumed. In the same spirit as the weighted version of the continuous ranked probability score proposed by Gneiting and Ranjan (2011), a weighted version of the quantile skill score could be envisaged in order to take into account specific use of quantile forecasts.

Ackowlegment

This work has been initiated in the framework of the EWeLiNE project (*Erstellung innovativer Wetter- und Leistungsprognosemodelle für die Netzintegration wetterabhängiger Energieträger*) funded by the German Federal Ministry for Economic Affairs and Energy.

Appendix

Overall value and Quantile Skill Score

From Eqs (A.33) and (A.34), the difference in expense between climatological and perfect deterministic forecasts can be written as

$$T_C - T_P = \frac{1}{N}\sum_{i=1}^{N} C(y_\tau - y_i) + \frac{1}{N}\sum_{i:y_i \geq y_\tau} L(y_i - y_\tau) \tag{A.36}$$

Considering the relationship $\tau = 1 - \dfrac{C}{L}$ and setting L equal to 1 in the following demonstration without loss of generality, we obtain

$$T_C - T_P = \frac{(1-\tau)}{N}\sum_{i=1}^{N}(y_\tau - y_i) + \frac{1}{N}\sum_{i:y_i \geq y_\tau}(y_i - y_\tau) \tag{A.37}$$

and with some algebra

$$T_C - T_P = \frac{(1-\tau)}{N}\sum_{i:y_i \leq y_\tau}(y_\tau - y_i) + \frac{\tau}{N}\sum_{i:y_i \geq y_\tau}(y_i - y_\tau) \tag{A.38}$$

This mean expense difference, $T_C - T_P$, corresponds to the definition of the quantile score for a climatological forecast (QS_{climate}).

In the same manner, from Eqs (A.35) and (A.34), the difference between climatological forecast expense and the quantile forecast expense is written as

$$\begin{aligned} T_C - T_F = {} & \frac{1}{N}\sum_{i=1}^{N} C y_\tau + \frac{1}{N}\sum_{i:y_i \geq y_\tau} L(y_i - y_\tau) \\ & -\frac{1}{N}\sum_{i=1}^{N} C q_{\tau,i} - \frac{1}{N}\sum_{i:y_i \geq q_{\tau,i}} L(y_i - q_{\tau,i}) \end{aligned} \tag{A.39}$$

which becomes after some algebra

$$\begin{aligned}T_C - T_F = \frac{(1-\tau)}{N} \sum_{i:y_i \leq y_\tau} (y_\tau - y_i) + \frac{\tau}{N} \sum_{i:y_i \geq y_\tau} (y_i - y_\tau) \\ - \left(\frac{(1-\tau)}{N} \sum_{i:y_i \leq q_{\tau,i}} (q_{\tau,i} - y_i) + \frac{\tau}{N} \sum_{i:y_i \geq q_{\tau,i}} (y_i - q_{\tau,i}) \right)\end{aligned} \tag{A.40}$$

where the first term corresponds to the definition of the quantile score for a climatological forecast (QS_{climate}, Eq. A.38), and the second term to the quantile score (QS_{forecast}, Eq. A.7). With regard to the definition of the quantile skill score and of the overall value (Eqs (A.8) and (A.32), respectively), we end up with:

$$V_{all} = QSS \tag{A.41}$$

Appendix B

Assessment and added value estimation of an ensemble approach with a focus on global radiation forecasts

The content of this appendix is the author's version of a manuscript published in 2015 in the Indian Quarterly Journal of Meteorology, Hydrology & Geophysics (Mausam) with reference:

> *Ben Bouallègue Z., 2015: Assessment and added value estimation of an ensemble approach with a focus on global radiation forecasts. Mausam, 66, 541–550.*

This publication follows the participation to the 6th WMO International Verification Method Workshop in New Delhi and the invitation of Beth Ebert, co-chair of the WWRP/WGNE Joint Working Group on Forecast Verification Research, to contribute to a Mausam special issue on verification methods.

Assessment and added value estimation of an ensemble approach with a focus on global radiation forecasts

Zied Ben Bouallègue

Deutscher Wetterdienst, Offenbach, Germany
Meteorological Institute, University of Bonn, Germany

Abstract

The assessment of the high-resolution ensemble weather prediction system COSMO-DE-EPS is achieved with the perspective of using it for renewable energy applications. The performance of the ensemble forecast is explored focusing on global radiation, the main weather variable affecting solar power production, and on quantile forecasts, key probabilistic products for the energy sector. First, the ability of the ensemble system to capture and resolve the observation variability is assessed. Secondly, the potential benefit of the ensemble forecasting strategy compared to a single forecast approach is quantitatively estimated. A new metric called ensemble added value is proposed aiming at a fair comparison of an ensemble forecast with a single forecast, when optimized to the users' needs. Hourly mean forecasts are verified against pyranometer measurements over verification periods covering 2013. The results show in particular that the added value of the ensemble approach is season-dependent and increases with the forecast lead time.

B.1 Introduction

The German electricity supply is currently being restructured aiming at increasing the integration of sustainable energies (Bartels *et al.*, 2006). Wind and solar energies are expected to play an important role in the ongoing energy transition. However, the intermittency of the power production, due to the weather dependent nature of these energy sources, poses a great challenge to the electricity grid operators (e.g., Boyle, 2008). High quality power forecasts are thus required for management strategies and operation activities in order to ensure the efficiency and safety of the grid as well as for energy trading. In particular, the installed solar capacities in Germany are increasing rapidly and the attention paid to solar forecasting is growing simultaneously.

The use of numerical weather prediction (NWP) models is a common approach for providing power forecasts for a horizon of a few hours to a few days (Costa *et al.*, 2008; Espinar *et al.*, 2010). Weather forecasts are used as input in transformation models that deliver optimized power forecasts to the end users. Therefore, the quality of the power forecasts depends strongly on the quality of the underlying weather forecasts. In this context, NWP models find new applications and efforts are being undertaken in order to improve the forecast quality of weather variables relevant for the energy sector. Forecasting hourly photovoltaic power production based on NWP model outputs has recently been explored by Lorenz *et al.* (2011) and Zamo *et al.* (2014a). The deterministic power forecasts assessed in these studies show variability in skill over different seasons, weather situations and forecast lead times encouraging a probabilistic forecasting approach. Indeed, the limited predictability of weather events implies a need for information about the predictive skill of the forecasts for an optimal use of the prediction systems (Krzysztofowicz, 1983; Richardson, 2000)

Uncertainty about the future state of the atmosphere can be estimated with an ensemble prediction system (EPS). An EPS provides a sample of possible future states of the atmosphere from multiple forecasts (Leutbecher and Palmer, 2008). At Deutscher Wetterdienst (DWD), an operational cloud-resolving EPS that covers Germany has been running operationally since May 2012 (Gebhardt *et al.*, 2011; Peralta *et al.*, 2012). The so-called COSMO-DE-EPS was originally designed to improve the quality of the forecast guidance in cases of high-impact weather events. Forecasts of convection-related weather events such as strong wind gusts and heavy precipitations can today be interpreted in a probabilistic way. The application range of the system is now planned to be extended in order to support the integration of renewable energies into the German electricity grid. Therefore, the performance of the ensemble system has to be assessed focusing on energy relevant weather variables and relevant probabilistic products for the energy sector.

The manuscript at hand deals with forecast verification of global radiation, the main weather variable affecting solar power production. In terms of probabilistic products, the focus is on quantile forecasts, which are key products for many energy applications (Pinson *et al.*, 2007; Morales *et al.*, 2014). Quantile forecasts are optimal point forecasts for users with an asymmetric loss function (Gneiting, 2011b). In other words, users with different penalties associated with under- and over-prediction can optimize their decisions by using quantile forecasts as decision variables.

The quality of the ensemble forecast and of the derived quantile products is estimated by means of proper scoring rules, namely the continuous ranked probability score (CRPS, Hersbach, 2000) and the quantile score (QS, Koenker and Machado, 1999). The decomposition of proper scores provides an estimation of the penalty due to the lack of reliability and reward from resolution (Bröcker, 2009). Reliability measures the ability of the predictive distribution to represent the unknown distribution of the observation conditional on the forecast while resolution measures the forecast ability to distinguish between different subsets of observations (Wilks, 2006b; Bentzien and Friederichs, 2014). Since reliability can be corrected by statistical techniques (Gneiting *et al.*, 2007), resolution, which is related to the forecast information content, is often considered as a more fundamental property that reflects the 'intrinsic value' of a forecasting system (Toth *et al.*, 2003).

Besides the 'traditional' assessment of the ensemble forecast in terms of its statistical attributes (reliability and resolution), the benefit of estimating the forecast uncertainty dynamically using an ensemble system is also quantified. For this purpose, the ensemble forecast is compared to a single (deterministic) forecast. Quantile forecast verification offers an appealing framework for a fair comparison of point forecasts in terms of potential performance. Assuming that statistical adjustments can be applied similarly to any point forecast, the comparison focuses on the forecast information content. The interpretation of the ensemble members in terms of quantiles allows the extension of the comparison to the whole probability distribution described by the ensemble forecast. This approach enables the development of a new metric, which quantitatively estimates the added value of the ensemble forecasting strategy compared to a single forecast approach.

Results are shown for hourly global radiation forecasts from COSMO-DE-EPS against measurements from 32 pyranometer stations distributed over Germany. The results are discussed for verification periods of 90 days in the year 2013. The ensemble added value is also shown and discussed as a function of the forecast lead time for different periods of the year. The manuscript is organized as follows: Section B.2 describes the ensemble and observation datasets, Section B.3 presents the verification methodology, Section B.4 discusses the results and Section B.5 concludes.

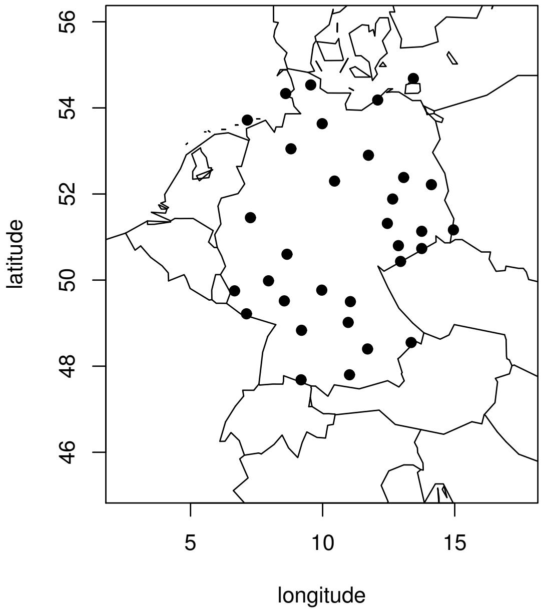

Figure B.1: Approximately the model domain with latitude/longitude axes. The dots indicate the location of the 32 pyranometer stations used in this study.

B.2 Data

COSMO-DE EPS is a regional EPS run operationally at the German Weather Service (DWD). The ensemble system is based on a 2.8 km grid resolution version of the COSMO model (Steppeler *et al.*, 2003; Baldauf *et al.*, 2011) with a model domain that covers Germany and part of the neighbouring countries. The ensemble comprises 20 members with variations in boundary conditions, model physics and initial conditions (Gebhardt *et al.*, 2011; Peralta *et al.*, 2012). COSMO-DE-EPS has been developed focusing on high-impact weather events. Previous studies have discussed the performance of the system in this context (Ben Bouallègue *et al.*, 2013; Ben Bouallègue and Theis, 2014).

Global radiation, the sum of the direct and diffuse radiations, is here the model output variable of interest. Global radiation is defined as the total downward solar radiation (or irradiance) incident on a horizontal surface (Badescu, 2008). The performance of hourly mean ensemble forecasts from the 03 UTC run is explored for the year 2013. During this year, the forecast lead time of the operational ensemble system was 21 hours until March, and then extended to 27 hours. Forecasts associated with solar zenith angles with cosine lower than 0.15 radians are considered as 'night' hours and are excluded from the verification process. Each forecast lead time is investigated separately. We mainly focus on a forecast lead time of 9 hours (12 UTC validity time) corresponding approximately to the daily peak of solar power production. This way, the strong diurnal cycle associated with solar variables does not affect the interpretation of the verification results (Hamill, 2001). The impact on the interpretation of the results due to the natural annual cycle is alleviated by using verification windows of 3 months.

Quality controlled pyranometer measurements are used for the verification process (Becker and Behrens, 2012). Figure B.1 shows the geographical distribution over Germany of the 32 pyranometer stations used in this study. For each observation point, the nearest model grid-point is selected.

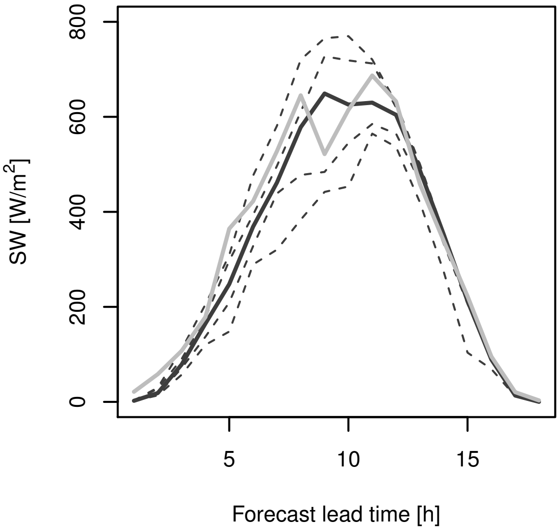

Figure B.2: Global radiation COSMO-DE-EPS ensemble derived forecasts valid on July 5 2013 at Hamburg. The full grey lines correspond to the pyranometer measurements, the full black lines indicates the ensemble median (50%-quantile) and the thin dashed lines indicate quantile forecasts at probability levels 5%, 25%, 75% and 95% (from bottom to top).

If not explicitly specified differently, quantile forecasts are derived from the 20 ensemble members using a linear interpolation of the ensemble empirical cumulative density function. An example of global radiation quantile forecasts from COSMO-DE-EPS valid at Hamburg on July 5 2013 is shown in Figure B.2. Quantile forecasts with probability levels 5%, 25%, 50%, 75% and 95% are plotted as well as the corresponding observations.

B.3 Verification methodology

General framework

The aim is to predict global radiation at specific locations. The ability of COSMO-DE-EPS to achieve this task is questioned: are the forecasts able to capture the observation variability? Are the forecasts able to do it in a valuable way, distinguishing between different subsets of observations? From this perspective, does the ensemble forecast provide additional information compared to a single member approach? The assessment of the ensemble forecast and derived products is performed in order to answer these questions. Starting from well-established scores and their properties, suitable measures are proposed for this purpose.

The verification process is applied to the ensemble forecast as a whole as well as to derived quantile forecasts. A quantile forecast is defined by a nominal probability level: a quantile forecast q_τ with nominal probability level τ indicates that there is a probability τ that the observation will be less than q_τ.

The CRPS is a common tool for evaluating ensemble forecasts in the form of cumulative distributions. The sorted ensemble members are interpreted as quantile forecasts such that the ensemble describes a piecewise constant cumulative distribution function with jumps at the ensemble members (Hersbach, 2000). Quantile forecasts at selected probability levels can be assessed separately. The QS is the natural tool used to evaluate such probabilistic products (Koenker and Machado, 1999; Gneiting, 2011a). QS is based on the so-called check function, an asymmetric loss function, which can ponderate differently errors due to under- and over-prediction.

CRPS and QS are proper scoring rules (Gneiting and Raftery, 2007; Bentzien and Friederichs, 2014) and can be decomposed (Hersbach, 2000; Bentzien and Friederichs, 2014). The decompositions provide reliability, resolution and uncertainty components of the scores. The decomposition of a proper score S can be written as

$$S = S_{\text{reliability}} - S_{\text{resolution}} + S_{\text{uncertainty}}. \tag{B.1}$$

The reliability term reflects the forecast biases and the resolution term is related to the forecast information content. The uncertainty term depends on the observations variability only and so it is not influenced by the forecast (Wilks, 2006b).

The difference between the uncertainty and resolution score components is often denoted as the potential score (e.g. Hersbach, 2000):

$$S_{\text{potential}} = S_{\text{uncertainty}} - S_{\text{resolution}} \tag{B.2}$$

$S_{\text{potential}}$ measures the *potential* performance in the sense that the reliability deficiencies can be potentially corrected. Indeed, deficiencies in terms of reliability can be alleviated using past data and assuming stationarity of the forecast error. The statistical adjustment of forecasts based on past data is usually called *calibration* (Gneiting *et al.*, 2007). The calibration step aims at providing reliable forecasts by correcting systematic biases and spread biases. Since the reliability term in Equation B.1 estimates the statistical consistency between the predictive distributions and the associated observations, $S_{\text{potential}}$ can be interpreted as the score that should be obtained after statistical adjustment of the forecast.

In order to show the pertinence of using a given forecast rather than an other one, skill scores are computed. A skill score measures the relative benefit of using a forecast compared to a reference one (e.g. Wilks, 2006b). We propose also to compute potential skill scores in order to compare forecasts and reference forecasts conditioned on calibration. In general terms, a skill score Sk is defined as:

$$Sk = 1 - \frac{S}{S^{\star}}, \tag{B.3}$$

where $S^{\star}$ is the score of the reference forecast. Applied to the CRPS and QS, Equation B.3 leads to the definition of the continuous ranked probability skill score (CRPSS) and of the quantile skill score (QSS), respectively. Similarly, we define a potential skill score $Sk_{\text{potential}}$ as:

$$Sk_{\text{potential}} = 1 - \frac{S_{\text{potential}}}{S^{\star}_{\text{potential}}} \tag{B.4}$$

where $S^{\star}_{\text{potential}}$ is the potential score of the reference forecast. $Sk_{\text{potential}}$ measures the potential benefit of using a forecast compared to a reference forecast conditioned on calibration. Since the reliability terms of the forecast and of the reference forecast are not taken into account, the potential skill score focuses on the forecast information contents only.

In this study, we consider two different reference strategies: a climatology based forecast and a single forecast approach. The interpretation of the *skill score* and *potential skill score* according to the chosen reference forecast is now discussed.

Sample climatology as reference

As first reference forecast for the computation of skill scores, we consider the sample climatology. For a given verification period, a cumulative probability distribution is derived based on all available observations. The sample distribution and the related quantiles are used as reference forecasts over each verification period.

A climatological forecast, based on the observation sample, is by definition perfectly reliable (Hersbach, 2000). Moreover, a climatological forecast is a constant forecast and therefore has no resolution (Mason, 2004). In that case, $S^\star$ and $S^\star_{\text{potential}}$ are equivalent and correspond to the uncertainty component of S:

$$S^\star = S^\star_{\text{potential}} = S_{\text{uncertainty}}. \tag{B.5}$$

The difference between potential skill score and skill score illustrates the benefit one can expect from calibration. Considering the sample climatology as reference, from Equations B.2, B.3, B.4 and B.5 this difference is given as a simple ratio:

$$Sk_{\text{potential}} - Sk = \frac{S_{\text{reliability}}}{S_{\text{uncertainty}}}. \tag{B.6}$$

The difference between potential skill score and skill score emphasizes the reliability deficiency relative to the observation variability. A difference close to zero indicates a forecast able to capture perfectly the observation variability.

From Equations B.1, B.4 and B.5, we can deduce that the potential skill score with the sample climatology as reference corresponds to:

$$Sk_{\text{potential}} = \frac{S_{\text{resolution}}}{S_{\text{uncertainty}}}. \tag{B.7}$$

$Sk_{\text{potential}}$ can be interpreted as the proportion of observation variability that the forecast is able to correctly resolve. A potential skill score close to 1 indicates perfect resolution of the forecast and a potential skill close to zero indicates no resolution at all.

Using the sample climatology as reference, reliability deficiency and resolution performance of the forecast are analyzed with respect to the variability of the observations. Two important and complementary statistical aspects of the forecast quality, reliability and resolution, can be discussed over different verification periods. The next step consists in evaluating how much of additional information is provided by the ensemble system compared to a single forecast. This step is taken considering a control forecast (or by default an arbitrarily selected member of the ensemble) as a reference forecast.

Single forecast as reference

The CRPS reduces to the mean absolute error (MAE) when applied to a single (deterministic) forecast which allows a direct comparison of ensemble and deterministic forecast performances (Gneiting and Raftery, 2007). However, from the user's perspective, this comparison is often not relevant since two different types of forecasts are compared: a probability distribution and a point forecast. A forecast in the form of a probability distribution is usually transformed into a probabilistic product adapted to the user's need: a probability forecast associated with an event or a quantile forecast for a selected probability level. A comparison based on probability products is not suitable since a basic interpretation of a deterministic forecast in terms of probability reduces its information content by transforming the forecast into a binary outcome.

On the other hand, a direct comparison of deterministic and ensemble derived forecasts based on quantiles seems adequate for including the complete forecast information content. Both types of forecasts, deterministic and quantile forecasts, are point forecasts expressed as continuous variables when dealing for example with global radiation or wind speed. The step of interpreting a point forecast as a quantile does not deteriorate the forecast information content since it only requires the definition of a nominal probability level. This probability level accounts for the user's sensitivity to under- and over-prediction. Any point forecast can potentially be adjusted to the user needs by calibration. A fair comparison of point forecast can then be performed focusing on the information content of the forecasts.

The quantile score (QS) and its decomposition offer an adequate framework for this comparison. QS can be applied similarly to any point forecast, derived from an ensemble or a deterministic forecast, and the QS decomposition provides an estimate of the forecast resolution. A fair comparison of ensemble derived quantiles and deterministic forecasts can be based on the potential quantile score, which reflects the balance between observation variability (uncertainty) and forecast information content (resolution). The reliability terms are disregarded based on the assumption that calibration can be applied similarly to an ensemble derived forecast or to a deterministic forecast. Adequate statistical methods for such calibration exists as for example quantile regression (Koenker and Machado, 1999). However, the step of calibrating the forecasts, which requires sufficient past training data at hand, is not applied here: the verification process relies on the decomposition of the quantile score.

QS is calculated for the derived ensemble forecasts and for the reference forecast assigning the deterministic forecast to the relevant quantiles. The decomposition of the quantile score allows to extract a measure of the usefulness of the forecast for the corresponding user. The components of the quantile score are computed following Bentzien and Friederichs (2014). We denote PQS_τ the potential quantile score of the τ-quantile forecast derived from the ensemble forecast, and $PQS^\star_\tau$ the potential quantile score of the reference forecast for a probability level τ. Following Equation B.4, the potential quantile skill score is computed as:

$$QSS_{\text{potential}} = 1 - \frac{PQS_\tau}{PQS^\star_\tau} \tag{B.8}$$

Choosing a single forecast as reference, the potential QSS becomes a measure of the added value of the ensemble system for a given probability level.

The added value of the ensemble forecast as a whole is derived based on the relationship between CRPS and QS. It has been shown that the CRPS corresponds to a weighted sum of quantile scores applied to the sorted ensemble members (Bröcker, 2012). As already noted, the ensemble members are interpreted as quantile forecasts for the computation of the CRPS. More precisely, the probability level τ_m associated to the sorted member of rank m is defined as:

$$\tau_m = \frac{m - 0.5}{M}, \quad m = 1, ..., M \tag{B.9}$$

with M the ensemble size. Denoting QS_{τ_m} the quantile score at probability level τ_m, the relationship between QS and CRPS is written as:

$$CRPS = \frac{2}{M} \sum_{m=1}^{M} QS_{\tau_m}. \tag{B.10}$$

Based on the potential quantile scores at probability levels τ_m with $m \in \{1, ..., M\}$, we define a potential CRPS as:

$$CRPS_{\text{potential}} = \frac{2}{M} \sum_{m=1}^{M} PQS_{\tau_m}. \tag{B.11}$$

$CRPS_{\text{potential}}$ reflects the potential performance of an ensemble forecast conditioned on the reliability of each sorted member. This assumption is stronger than the one used for the computation of the potential CRPS following Hersbach (2000) where the reliability of the ensemble as a whole is considered. In this latter case, the reliability term of the CRPS decomposition is directly related to the rank histogram (Hersbach, 2000) and can therefore be subject to misinterpretation (Hamill, 2001).

Similarly, the potential quantile scores of the reference forecast are estimated for each probability level defined by the ensemble. The associated potential CRPS is defined as:

$$CRPS^{\star}_{\text{potential}} = \frac{2}{M} \sum_{m=1}^{M} PQS^{\star}_{\tau_m} \tag{B.12}$$

where $PQS^{\star}_{\tau_m}$ is the potential quantile score at probability level τ_m when applied to the reference forecast. This procedure consists consequently in bringing the reference deterministic forecast to the degree of complexity of the ensemble forecast and not the opposite, as it is the case for example when the ensemble mean is computed in order to be compared to a deterministic forecast.

Finally, we define a new metric called ensemble added value (EAV), which takes the form of a potential skill score:

$$EAV = 1 - \frac{CRPS_{\text{potential}}}{CRPS^{\star}_{\text{potential}}} \tag{B.13}$$

EAV is a summary measure of the potential benefit of using the ensemble forecast rather than a single forecast, conditioned on calibration. EAV greater than 0 indicates that the ensemble forecast outperforms the single forecast in terms of valuable information content.

B.4 Results

Verification process

COSMO-DE-EPS global radiation forecasts are assessed over the year 2013. Rolling verification windows are used in order to evaluate the performance for different periods of the year. The size of the verification windows is 90 days and the rolling step is set to 10 days. The results are first shown for forecasts of the 03 UTC run valid at 12UTC, corresponding to a forecast lead time of 9 hours. Afterward, verification results as a function of the forecast lead time are discussed for different seasons.

The statistical significance of the results is estimated by bootstrapping, which is a common resampling technique proposed by Efron and Tibshirani (1986) and popularized in meteorology by Hamill (1999). Confidence intervals of 5% and 95% are attributed to the scores based on 500-member block bootstrap samples. Each day is considered as a separate block of fully independent data such that the score distribution, from which the confidence intervals are drawn, represents the variability of the scores over the verification period and not between locations.

Reliability and resolution

Skill scores of the ensemble forecast and of individual quantile forecasts are shown in Figures B.3 and B.4. The CRPSS and potential CRPSS with the sample climatology as reference are plotted in Figure

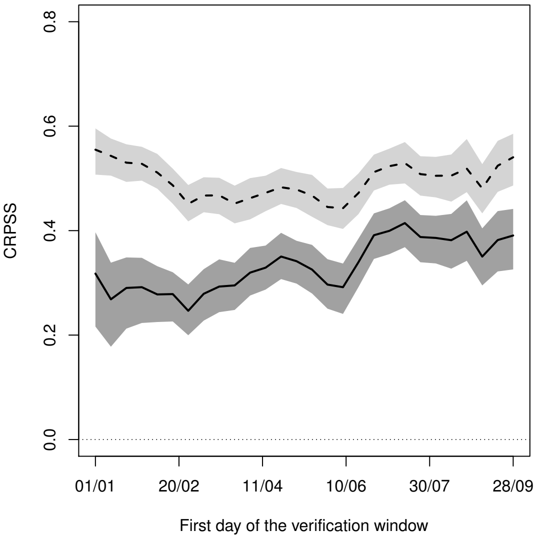

Figure B.3: CRPSS (full line) and potential CRPSS (dashed line) as a function of verification periods of 90 days in the year 2013. The skill scores are computed with the sample climatology as reference. The grey areas indicate the 5% and 95% confidence intervals derived by bootstrapping.

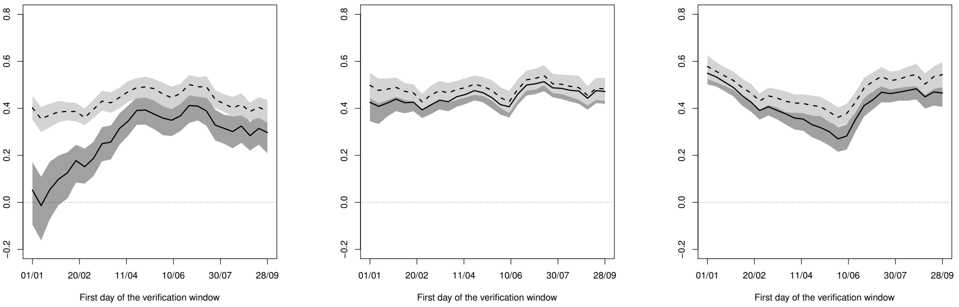

Figure B.4: QSS (full lines) and potential QSS (dashed lines) for the 25%- (left), the 50%- (middle) and the 75%-quantile forecasts (right) as a function of verification periods of 90 days in 2013. The skill scores are computed using quantile climatological forecasts as reference. The grey areas represent the 5%-95% confidence intervals derived by bootstrapping.

B.3. Similarly, QSS and potential QSS with the sample climatology quantile forecasts as references are shown in Figure B.4. The probability levels investigated correspond to the median (50%), the lower and the upper quartiles (25% and 75%).

The reliability of the ensemble forecast is deduced from the difference between potential skill scores and skill scores: a smaller difference indicates more reliable forecasts. Reliability deficiencies occur throughout the year but are stronger during winter. In particular, the need for calibration is more pronounced when looking at the 25%-quantile forecasts. The deficit of reliability at this probability level has a strong negative impact on the forecast skill. On the other hand, the median and upper quartile forecasts show appropriate statistical properties. Their good performance in terms of reliability is confirmed by means of the analysis of quantile reliability diagrams for different verification periods (not shown).

The ability of the forecasts to resolve the observed variability is described by the potential skill scores. The potential CRPSS is not subject to large variations along the year, varying around 50%. Similar results are obtained for the 50%-quantile forecast. The potential QS for the two other probability levels present opposite tendencies as a function of the verification period: in terms of resolution ability, the 25%-quantile forecasts have relatively higher skill in summer than in winter while the 75%-quantile forecasts perform relatively better in winter than in summer, compared to their relative climatological forecasts. This behavior can be explained by the fact that global radiation is a bounded variable: in summer, when clear days dominate, the climatological distribution provides a better (worse) estimation of the upper (lower) quantiles than in winter, where cloudy days dominates, and vice-versa.

In summary, the ability of the ensemble system to capture the observed variability is season dependent, while its ability to distinguish between subsets of events is relatively constant. At the quantile forecast level, reliability deficiencies have a severe impact on the forecast skill for small probability levels while the resolution ability at low and high probability levels exhibits a balancing effect between summer and winter periods.

Ensemble added value

The benefit of the ensemble strategy is now illustrated and discussed. For each verification day, one member from the 20 ensemble members is randomly selected and used as reference forecast. The EAV estimated this way is shown in Figure B.5. Note that similar results are obtained when any of the ensemble members is arbitrarily chosen as reference forecast for the whole verification period. Figure B.6 shows the ensemble added value at specific probability levels. The probability levels investigated here are the ones defined by the ensemble size ($\tau = 2.5\%, 7.5\%, ..., 92.5\%, 97.5\%$, see Equation B.9).

The added value of the ensemble approach is statistically significantly positive all along the year. The potential benefit ranges from 4% in the winter up to 8% during the summer, showing a clear seasonal signal. A similar seasonal cycle is drawn for intermediate probability levels in Figure B.6. The seasonal signal is stronger for low probability levels and has an opposite behavior for high probability levels. The added value for the probability levels corresponding to the members of ranks 18 to 20 exhibit a maximum reached during the winter and a minimum reached during the summer. The low predictability of the upper tail of the predictive distribution during the winter season and of the lower tail during the summer season can explained the clear advantage of using an ensemble system rather than a single forecast in such situations. So, the benefit of the ensemble approach for a user can vary by a factor 10 depending on the season and his/her probability level of interest.

Finally, the analysis of the EAV is extended to all meaningful lead times (excluding the 'night' hours, see Section B.2). Figure B.7 shows EAV as a function of the forecast lead time for four different verification periods of three months: January - February - March (JFM), April - May - June (AMJ), July - August - September (JAS) and October - November - December (OND) of the year 2013. The ensemble added value is statistically significantly greater than zero for all lead times except for 'edge' hours. The first and last verification hours of each season, near sunset and sunrise hours, are affected by the data scarcity that translates into large confidence intervals in the results. Otherwise, EAV is higher during the spring/summer than in the autumn/winter for all forecast lead times. The potential benefit of using the ensemble also shows a tendency to increase as a function of the lead time for all seasons.

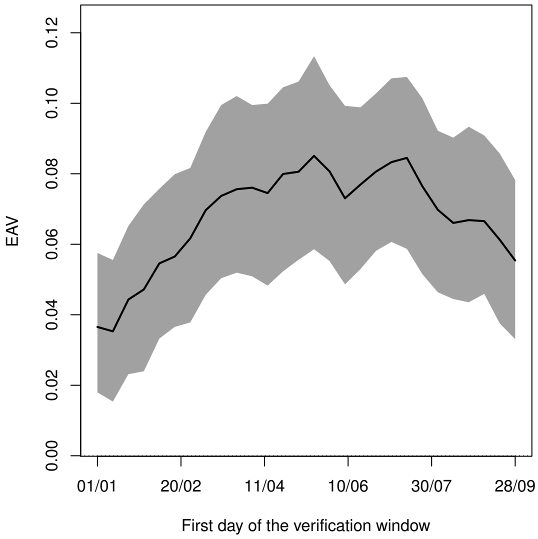

Figure B.5: EAV, added value of the ensemble forecast strategy with respect to a single forecast approach, as a function of verification periods of 90 days in 2013. 5%-95% confidence intervals derived by bootstrapping are represented by the grey area.

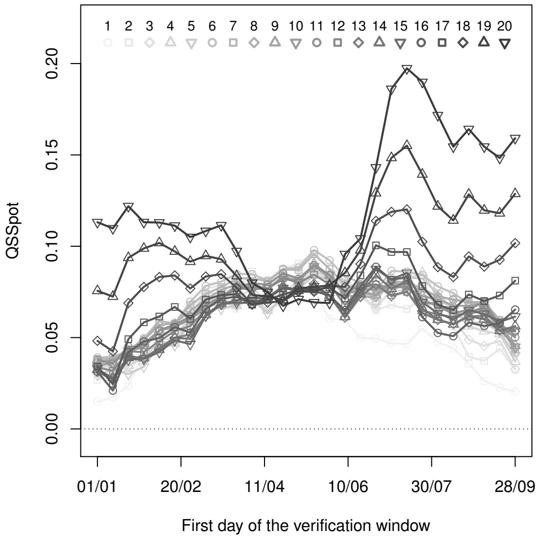

Figure B.6: Potential QSS, with a single forecast as reference, as a function of verification periods of 90 days in 2013. Results for probability levels 2.5%, 7.5%,..., 92.5%, 97.5% (corresponding to the ensemble members of rank 1, 2, ..., 19, 20) are represented with different shades of grey.

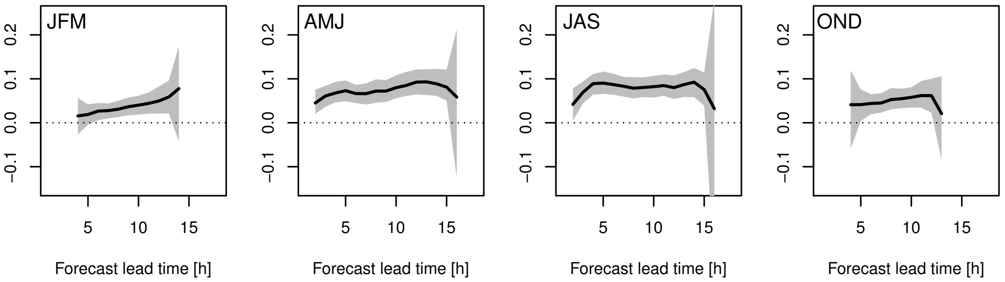

Figure B.7: EAV, added value of the ensemble forecast strategy compared to a single forecast approach, as a function of the forecast lead time. From left to right, the verification period is January - February - March (JFM), April - May - June (AMJ), July - August - September (JAS) and October - November - December (OND) of the year 2013. The grey areas represent the 5%-95% confidence intervals estimated by bootstrapping.

B.5 Conclusion

In this study, global radiation forecasts from the high-resolution ensemble system COSMO-DE-EPS are assessed against hourly measurements from 32 pyranometer stations. The analysis of skill scores and potential skill scores with the sample climatology as reference provides an evaluation of the system performance over the year 2013. The comparison of skill scores and potential skill scores highlights the deficiencies of the ensemble forecast and of the derived quantile products in terms of reliability. The benefit one can expect from calibration is shown to be higher during the winter and autumn seasons and when focusing on low probability levels. The potential skill scores analysis indicates that the resolution ability of the ensemble system is relatively stable along the year, performance at low and high probability levels being balanced for each season.

It is also shown that the ensemble forecasting approach is expected to provide more useful information to weather forecast users than a single forecast approach. The estimation of the ensemble potential benefit is measured by a new metric called ensemble added value (EAV), which aims at a fair comparison between ensemble forecasts and single deterministic forecasts. The EAV computation is based on the decomposition of the quantile score and the ensemble interpretation related to the computation of the continuous ranked probability score. EAV has the form of a skill score and rewards the additional information content provided by the ensemble forecast. For local global radiation forecasts, the benefit of the ensemble approach is statistically significant for all relevant forecast hours and all seasons. The added value of the ensemble is greater during spring-summer periods and increases with the lead time.

Acknowledgment

This work has been realized in the framework of the EWeLiNE project (*Erstellung innovativer Wetter- und Leistungsprognosemodelle für die Netzintegration wetterabhängiger Energieträger*) funded by the German Federal Ministry for Economic Affairs and Energy. Tobias Heppelmann, Richard J. Keane, Kristina Lundgren and two anonymous reviewers are thanked for their valuable comments on draft versions of this manuscript.

Appendix C

Statistical post-processing of global radiation ensemble forecasts with penalized quantile regression

The content of this appendix is the author′s version of a manuscript accepted for publication in the Meteorologische Zeitschrift.

Statistical post-processing of global radiation ensemble forecasts with penalized quantile regression

Zied Ben Bouallègue

Deutscher Wetterdienst, Offenbach, Germany
Meteorological Institute, University of Bonn, Germany

Abstract

Nowadays, ensemble-based numerical weather forecasts provide probabilistic guidance to actors in the renewable energy sector. Ensemble forecasts can however suffer from statistical inconsistencies that affect the forecast reliability. Statistical post-processing techniques address this issue using learning algorithms based on past data. In this study, it is shown that quantile regression is a suitable method for the post-processing of ensemble global radiation forecasts. In a basic approach, conditional quantiles are estimated using the first guess quantile forecasts and a solar geometry variable as predictors. In a more complex approach, adequate meteorological predictors are selected among a pool of ensemble model outputs by means of a regularization scheme. The so-called penalized quantile regression and the basic quantile regression approaches, respectively, are applied to hourly averaged global radiation forecasts of the high-resolution ensemble prediction system COSMO-DE-EPS. Both calibration setups provide reliable probabilistic forecasts at all investigated probability levels, which improves considerably the ensemble forecast skill. Moreover, verification results demonstrate that including rigorously selected predictors in the regression scheme increases the ensemble forecast sharpness and thereby the value of the probabilistic guidance.

C.1 Introduction

At the German weather service (DWD), the numerical weather prediction (NWP) systems are called to enlarge their range of applications in order to support the ongoing energy transition which is taking place in Germany. For the electricity grid operators, the natural variability of solar and wind productions is a challenge that can be addressed and mitigated by accurate weather forecasts (Boyle, 2008). For an optimal forecast's use, information about forecast uncertainty is also essential: the optimisation of management and operation strategies requires indeed probabilistic forecasts (Pinson *et al.*, 2007; Pinson, 2013). Providing reliable probabilistic weather forecasts to actors of the renewable energy sector should therefore contribute to a more efficient integration of intermittent sources in the electricity network (Morales *et al.*, 2014).

Ensemble prediction systems (EPS) are nowadays a standard approach in NWPs that provide the basis for a probabilistic interpretation of weather forecasts (Palmer, 2000; Leutbecher and Palmer, 2008). At DWD, the convection-resolving ensemble system COSMO-DE-EPS has been running operationally since May 2012. It consists of 20 deterministic forecasts based on the COSMO-DE model with variations of the boundary conditions, initial conditions, and model physics (Gebhardt *et al.*, 2011; Peralta *et al.*, 2012). The ability of COSMO-DE-EPS to support the integration of solar energy in the German grid has been investigated and the great potential of the ensemble approach has been demonstrated (Ben Bouallègue, 2015). However, forecast assessment focusing on global radiation, the main weather variable affecting photovoltaic power production, has also shown that the ensemble forecast suffers from a lack of reliability. In order to fully benefit from the ensemble potential, statistical inconsistencies have to be corrected (Ben Bouallègue *et al.*, 2015).

The adjustment of the statistical properties of an ensemble forecast is often referred to as calibration. Following the paradigm of Gneiting *et al.* (2007), the forecast optimisation aims to maximise the forecast sharpness (the concentration of the predictive distributions) subject to calibration (the statistical consistency between the distributional forecasts and observations). Rooting in model output statistics (MOS) for deterministic forecasts (Glahn and Lowry, 1972), various statistical methods have been developed for the calibration of ensemble forecasts (for an overview see Wilks, 2006a; Wilks and Hamill, 2007). Besides variables such as temperature or precipitation, calibration of ensemble wind forecasts, which is of particular interest for renewable energy applications, has been investigated in recent years (see among others Sloughter *et al.* (2010); Pinson (2012); Schuhen *et al.* (2012); and for an intercomparison of the state-of-the-art calibration techniques Junk *et al.* (2014)).

For global radiation forecasts, ensemble calibration is a new area of research and application for existing methods. In this study, we propose to develop approaches based on quantile regression (QR). Initially introduced by Koenker and Bassett (1978), QR has already been successfully applied in order to calibrate ensemble precipitation forecasts (Bentzien and Friederichs, 2012). QR estimates quantile forecasts at a nominal probability level of interest τ where a τ-quantile forecast q_τ indicates that there is τ chance that the observation falls below the quantile q_τ. This semi-parametric method does not require assumptions about the form of the underlying probability distribution and directly provides probabilistic forecasts in the form of quantiles. These quantile forecasts are key probabilistic products in the energy sector (Pinson, 2013; Morales *et al.*, 2014). Moreover, quantile forecasts serve as basic inputs for the application of empirical copulas that allows the retrieval of scenarios from *locally* calibrated forecasts (Wilks, 2014). Scenarios are indeed also of high relevance for renewable applications (Pinson *et al.*, 2009).

In our approaches, particular attention is paid to the definition and selection of adequate predictors. A basic approach of QR applied to global radiation consists in choosing the first guess quantile forecast and a solar geometry variable as predictors. A more complex approach consists of an automated selection of predictors among a pool of ensemble direct model outputs. In that case, the selection is performed based on a penalized quantile regression (PQR) scheme: a penalty term is added in the regression equation following the regularization approach of the least absolute shrinkage and selection operator (lasso, Tibshirani, 1996; Wahl, 2015). The penalty term can be optimised and the predictor selection exploited for forecast error diagnostics (Bröcker, 2010). The proposed statistical post-processing approaches, QR and PQR, are tested on COSMO-DE-EPS forecasts in an operational-like setup, with pyranometer measurements at 32 stations as observation dataset.

The paper is organised as follows: Section C.2 introduces the ensemble system, the observation dataset and the probabilistic products under focus. Section C.3 describes the calibration process, and Section C.4 the verification tools used for the forecast assessment. Section C.5 discusses the calibration setup and the verification results. Section C.6 presents the conclusion.

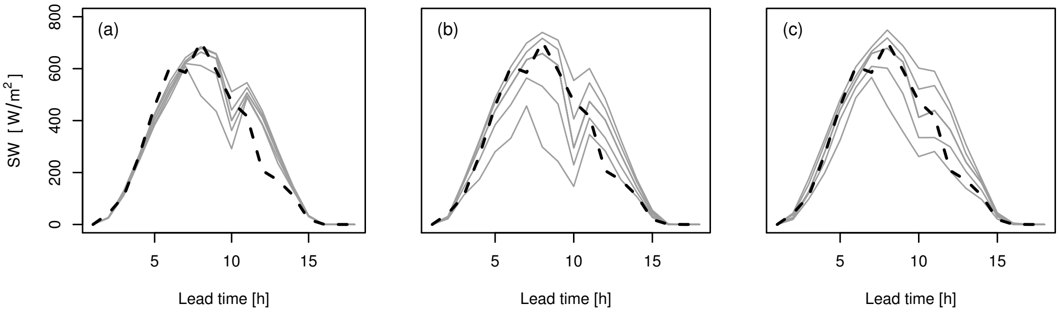

Figure C.1: Example of hourly averaged global radiation COSMO-DE-EPS quantile forecasts for April 18, 2013 at Lindenberg. (a) Raw ensemble derived quantiles, (b) calibrated quantiles with QR_s, (c) calibrated quantile forecasts with PQR_o. The grey lines from bottom to top correspond to quantile forecasts at probability levels $0.1, 0.25, 0.5, 0.75, 0.9$, respectively. The black dashed line represents the corresponding observations.

C.2 Data

Ensemble and observations

COSMO-DE-EPS is the regional ensemble prediction system run operationally at DWD, Offenbach, Germany. COSMO-DE-EPS is based on a 2.8 km horizontal grid resolution version of the COSMO model (Baldauf *et al.*, 2011). The model domain covers Germany and parts of the neighbouring countries, and the ensemble comprises 20 members with variations in the boundary conditions, model physics, and initial conditions (Gebhardt *et al.*, 2011; Peralta *et al.*, 2012). Aiming at solar energy applications, previous studies have explored the performance of the system focusing on global radiation forecasts, the sum of direct and diffuse radiation (Ben Bouallègue, 2015; Ben Bouallègue *et al.*, 2015). In particular, it has been shown that global radiation forecasts suffer from a lack of reliability, which makes the calibration of the ensemble forecast a necessary step.

The observation dataset corresponds to quality controlled pyranometer measurements from 32 stations distributed over Germany (Becker and Behrens, 2012; Ben Bouallègue, 2015). Hourly averaged observations and forecasts are assessed with a temporal resolution of one hour. The chosen verification period is spring (March, April, May) 2013. During this period, the operational ensemble has a forecast horizon up to 21 hours. Based on the 03UTC runs, the forecast lead times of interest range between 1 and 18 hours, excluding night hours. An example of global radiation COSMO-DE-EPS forecasts and the corresponding observations is shown in Fig. C.1(a). The forecasts, valid at station Lindenberg on April 18, 2013, are plotted in the form of quantiles at different probability levels.

Probabilistic products

An ensemble forecast consists of M deterministic forecasts, where M is called ensemble size. Each member is a fully consistent *weather scenario* that describes the state of the atmosphere and its evolution over a period of time. An ensemble allows capturing the forecast uncertainty at specific locations and time horizons as well as across time and space. Information about uncertainty dependence structures are essential for complex decision processes, and therefore often required by users in the renewable energy sector (e.g., Pinson *et al.*, 2009).

Focusing on a given point in space (model grid point) and a temporal window (or forecast horizon), an ensemble forecast is usually interpreted and communicated in the form of a probabilistic product. For energy applications, quantile forecasts are key products required by the end users (e.g., Morales *et al.*, 2014). In general terms, a quantile forecast at a nominal probability level τ is an optimal forecast for a user with an asymmetric loss function whose asymmetry is defined by τ (Roulston *et al.*, 2003; Pinson *et al.*, 2007; Gneiting, 2011a).

Formally, a quantile Q_τ at probability level τ ($0 \leq \tau \leq 1$) is defined as:

$$Q_\tau := F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\} \tag{C.1}$$

where F_Y is the cumulative probability distribution of the random variable $Y \in \Re$ such:

$$F_Y(y) = Pr(Y \leq y). \tag{C.2}$$

Based on the discrete sample drawn by the ensemble forecast, an ensemble member can be interpreted as a quantile forecast by considering its rank within the ensemble. The probability level τ_m associated with the member of rank m is defined as:

$$\tau_m = \frac{m}{M+1},\ m \in 1, ..., M \tag{C.3}$$

with M the ensemble size.

The calibration process described hereafter aims to calibrate the M quantile forecasts at the probability levels defined by the ensemble size following Eq.(C.3), for each grid point of the model domain and each forecast horizon of interest. In order to redraw temporally and spatially consistent scenarios from the calibrated quantile forecasts, the use of empirical copulas can be considered in a second step (Wilks, 2014).

C.3 Calibration process

The statistical adjustment of the ensemble forecasts is based on past data. Pairs of past ensemble forecasts and observations are gathered from a *training dataset*. In a quasi-operational setup, the *training period* is defined as a rolling window covering N_d days before the start of the forecast run, with N_d the length of the training period. Calibration is applied to each forecast lead time separately and the calibration coefficients are updated on a daily basis, moving the rolling window to the following day. In our implementation, the estimated regression coefficients for each lead time are assumed to be valid for all locations (grid points). In this Section, quantile regression (QR) is formally described, the predictors are defined, and their combination, transformation as well as selection are discussed.

Quantile regression

Consider a response variable y (here hourly averaged global radiation) and a set of predictors $\boldsymbol{v}$ (defined in Section C.3). QR focuses on quantiles of the conditional distribution $F_Y(y \mid \boldsymbol{v})$. Originally proposed by Koenker and Bassett (1978), QR has found several applications in meteorology. Examples resemble among others the calibration of wind power forecast (Bremnes, 2004), the downscaling of precipitation forecasts (Friederichs and Hense, 2007), and the calibration of ensemble precipitation forecasts (Bentzien and Friederichs, 2012).

The conditional quantile function for the probability level τ is noted $Q_\tau(y \mid \boldsymbol{v})$. The estimate of the quantile function is provided by the linear model:

$$\hat{Q}_\tau(y \mid \boldsymbol{v}) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}\boldsymbol{v} \tag{C.4}$$

where the τ_{th} quantile coefficients $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ are obtained solving the minimization problem:

$$\arg \min_{(\beta_0, \boldsymbol{\beta})} \sum_{j=1}^{N'} \rho_\tau (y_j - \beta_0 - \boldsymbol{\beta} \boldsymbol{v}_j) \tag{C.5}$$

with ρ_τ the loss function and N' the number of observations in the training dataset. ρ_τ is an asymmetric piecewise linear function, usually referred to as check function, and defined as :

$$\rho_\tau(u) = u[\tau - I(u > 0)] = \left\{ \begin{array}{ccc} \tau u & \text{if} & u \geq 0 \\ (\tau - 1)u & \text{if} & u < 0. \end{array} \right. \tag{C.6}$$

where $I(\cdot)$ is an indicator function having value 1 if the condition in parenthesis is true and zero otherwise. Eq.(C.5) is solved using a set of training data $\{(\boldsymbol{v_1}, y_1), ..., (\boldsymbol{v_n}, y_n)\}$ where $\boldsymbol{v_j}$ is the vector of predictors verifying when y_j is observed.

QR uses least absolute deviations in order to estimate the regression coefficients. The optimisation function minimizes a weighted sum of absolute residuals that can be formulated as a linear programming problem. The algorithm used to compute the fit is the modified version of the Barrodale and Roberts algorithm described in detail in Koenker and d'Orey (1994). In the following, quantile regression is performed using the R package *'quantreg'* (R Core Team, 2013).

Ensemble predictors

First, a basic setup of the QR calibration process is defined and later used as benchmark to evaluate the benefit of a more complex approach. Only a limited number of *a priori* meaningful variables are used as predictors. For each probability level τ of interest, the first guess quantile forecast (fgq_τ), that is the unprocessed ensemble forecast after sorting (see Section C.2), is used as primary predictor for the correction of the corresponding probabilistic forecast. Next to the primary predictor, a power transformation of fgq_τ is included in the regression model. Power transformations are often used in calibration in order to model potential nonlinear connections between response variable and predictor (Hamill *et al.*, 2007; Bröcker, 2010; Ben Bouallègue, 2013). We found that, for global radiation, taking the squared value of fgq_τ appears to be a suitable transformation. Additionally, the radiation at the top of the atmosphere (r_{toa}) is added as predictor in order to account for seasonal and diurnal cycles associated with global radiation without the use of normalized predictors such as the clearness index or clear sky index. r_{toa} reflects the solar geometry and is computed offline as a function of the day of the year, the hour of the day, the longitude, and the latitude of the observation (Badescu, 2008, Chapter 6). For each τ regression equation, the set of predictors that is explicitly defined as:

$$\boldsymbol{v} := \left\{ fgq_\tau, fgq_\tau^2, r_{toa} \right\} \tag{C.7}$$

is in the following referred to as *basic* predictors.

Additional predictors are added in the regression scheme aiming at a *weather dependent* calibration. The pre-selected weather variables along with their acronym used in the text are listed in Table C.1. They correspond to low, medium, high, and total cloud cover, total precipitation, maximum and minimum temperature at 2 meter height above ground within an interval of 6 hours as well as meridional and zonal winds at the pressure levels 500, 850, and 1000 hPa. This pre-selection is suggested by previous studies based on other models (Marquez and Coimbra, 2011) and on heuristic results based on COSMO-DE-EPS. Beside the variables indicated in Table C.1, other predictors have indeed been tested in the regression scheme. Low, medium and high cloud cover have been replaced by relative humidity at different levels. Pressure, temperature and wind components at different model levels have been added to the predictor list. These configurations have however not demonstrated to

Acronym	Model output variable
CLCL	low cloud cover
CLCM	medium cloud cover
CLCH	high cloud cover
CLCT	total cloud cover
PREC	total precipitation
TMIN2M	minimum temperature at 2 meter height
TMAX2M	maximum temperature at 2 meter height
U500	zonal wind at 500 hPa
V500	meridional wind at 500 hPa
U850	zonal wind at 850 hPa
V850	meridional wind at 850 hPa
U1000	zonal wind at 1000 hPa
V1000	meridional wind at 1000 hPa

Table C.1: List of model outputs used to define the additional predictors and their acronym used within the text.

provide any significant improvement of the PQR-based calibrated forecast performance (not shown).

In an ensemble forecast framework, different statistical products can be derived from the ensemble of members. We propose to include the minimum (min) and the maximum (max) of all ensemble members, the ensemble standard deviation (sd) as well as the ensemble mean in the list of predictors, for each variable except total precipitation. For this latest variable, the associated predictor is defined as the probability of precipitation (pop). The predictors, later designated as *weather predictors*, are noted $\left\{v_1^\star, ..., v_{N_p^\star}^\star\right\}$ with $N_p^\star$ the number of additional predictors.

Moreover, we have noticed that the definition of new predictors as a combination of the weather and basic predictors can be beneficial. Indeed, the model presented in Eq. C.4 is based on a linear combination of predictors where each element contributes independently to the regression model. We propose to extend this model including multiplicative terms defined as the combination of the weather predictors with the first guess forecast. The multiplicative terms are usually referred to as interaction terms (Friedrich, 1982). They are formally noted as:

$$\boldsymbol{v'} := \left\{fgq_\tau \cdot v_1^\star, ..., fgq_\tau \cdot v_{N_p^\star}^\star\right\}. \tag{C.8}$$

The additive model for the multilinear regression partially develops into a multiplicative one by using the following set of predictors:

$$\boldsymbol{v} := \left\{fgq_\tau, fgq_\tau^2, r_{toa}, v_1^\star, ..., v_{N_p^\star}^\star, v_1', ..., v_{N_p^\star}'\right\}. \tag{C.9}$$

In that case, the total number of predictors (N_p) corresponds to $N_p = 3 + 2 \cdot N_p^\star$. At this stage, since N_p can be relatively high (here $N_p = 101$) , the application of an efficient predictor selection is essential.

Regularization

A regularization is applied to the regression problem defined in Eq. C.4 in order to operate a selection of adequate predictors among the basic predictors, the weather predictors and the interaction terms. We propose to follow the least absolute shrinkage and selection operator (lasso) approach proposed by Tibshirani (1996).

The regularization takes the form of a penalty for complexity which imposes a compromise between fitting the training data and simultaneously keeping the coefficients small (Siegert *et al.*, 2011). A penalty term based on the absolute size of the regression coefficients is added to the learning algorithm. Penalized quantile regression (PQR) estimates the regression coefficients $(\hat{\beta}_o, \hat{\boldsymbol{\beta}})$ by solving the minimization problem:

$$\arg \min_{(\beta_0, \boldsymbol{\beta})} \sum_{j=1}^{N'} \rho_\tau (y_j - \beta_0 - \tilde{\boldsymbol{v}}_j \boldsymbol{\beta}) + \lambda_r \|\boldsymbol{\beta}\| \tag{C.10}$$

where $\| \cdot \|$ refers to the L_1-norm and λ_r is called the regularization parameter. The predictors $\tilde{\boldsymbol{v}}_j$ correspond to the normalized predictors $\boldsymbol{v}_j$ with zero mean and unit variance.

Increasing the penalty, more and more coefficients are driven to zero. In that case, the impact of the corresponding predictor on the regression is deactivated. Thereby, overfitting is prevented and the forecast accuracy is improved (Gneiting *et al.*, 2005). The strategy adopted for the optimisation of the parameter λ_r is described in Section C.5. Moreover, the remaining non-zero coefficients are manipulated in order to perform forecast error diagnostics as suggested in Bröcker (2010) and as shown in Section C.5.

C.4 Verification process

The verification process focuses on forecasts in the form of quantiles at different probability levels. In this Section, we recall common and innovative tools for the assessment of such probabilistic products: the probability integral transform histogram, the quantile score and its decomposition, its generalization to the continuous ranked probability score, and the relative user characteristic curve. The definition of skill scores and how confidence intervals are drawn is also introduced.

PIT histogram

The reliability of the ensemble of quantile forecasts is assessed by means of the probability integral transform (PIT) histogram (Diebold and Tay, 1998; Gneiting *et al.*, 2007). PIT histograms assess calibration of cumulative predictive distributions checking whether the observations can be considered as random samples of these distributions. A flat histogram is a necessary condition for reliability while a U-shaped histogram is interpreted as an indication of underdispersiveness or conditional biases in the forecast (Hamill, 2001).

Quantile score

The quantile score (QS) is the natural score for the assessment of quantile forecasts. QS is a proper scoring rule based on the check function ρ_τ defined in Eq. C.6 (Bentzien and Friederichs, 2014). Applied to pairs of observations y_i and quantile forecasts $q_{\tau,i}$ with $i \in \{1, ..., N\}$, QS is computed as the mean of the check function over the verification sample following:

$$QS_\tau = \frac{1}{N} \sum_{i=1}^{N} \rho_\tau (y_i - q_{\tau,i}), \tag{C.11}$$

where N is the size of the verification sample.

QS is decomposed into reliability, resolution and uncertainty terms following Bentzien and Friederichs (2014). In particular, the reliability attribute is assessed by means of the reliability diagram for quantile forecasts. The reliability diagram plots the conditional observed quantile as a function of quantile forecast categories. The deviation of the reliability curve from the diagonal is interpreted as a reliability deficiency (Wilks, 2006b; Bentzien and Friederichs, 2014).

Continuous ranked probability score

The continuous ranked probability score (CRPS) is a popular score for the evaluation of ensemble forecasts (Hersbach, 2000; Gneiting and Raftery, 2007). Considering an ensemble $\boldsymbol{e}$ of M members, the CRPS is defined as:

$$\begin{aligned} CRPS(\boldsymbol{x}, y) = \frac{1}{N} \sum_{i=1}^{N} \Big(\frac{1}{M} \sum_{m=1}^{M} \mid x_i^{(m)} - y_i \mid \\ - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \mid x_i^{(m)} - x_i^{(m')} \mid \Big) \end{aligned} \tag{C.12}$$

where $x_i^{(m)}$ and $x_i^{(m')}$ indicate the ensemble member m and m' for the verification pair i, respectively.

Equivalently, the CRPS can be defined as a weighted sum of quantile scores applied to the sorted ensemble members. We can note that the interpretation of the ensemble forecast in terms of equidistant quantiles following Eq. (C.3) does not lead to the optimal ensemble interpretation for the minimization of the CRPS but provides flat PIT histograms conditioned on calibration (Bröcker, 2012). Nevertheless, the CRPS can be interpreted as the overall quality of an ensemble of forecasts, while the QS provides detailed information about the forecast quality at specific probability levels.

Quantile discrimination

In a probabilistic verification framework, forecast discrimination is a forecast attribute related to the user decision making framework (Murphy, 1991). Quantile forecast discrimination can be assessed by means of the relative user characteristic (RUC) curve (Ben Bouallègue *et al.*, 2015). The RUC curve plots two forecast characteristics (the false alarm rate and the hit rate) for a specific user as the event definition varies. The area under the RUC curve, noted AUC, is used as summary measure of the discrimination ability of a quantile forecast for a specific probability level. The RUC framework is the equivalent of the relative operating characteristics (ROC) framework traditionally used for the verification of probability forecasts, and the derived verification measures follow a similar interpretation (Mason, 1982).

Discrimination is not sensitive to systematic inconsistencies and therefore does not reward efforts to provide calibrated forecasts. However, discrimination is directly related to the information content of a forecast and can therefore be interpreted as its "usefulness" in a decision making framework. The direct relationship between the discrimination ability and the *value* of a forecast, makes the estimation of the former an important verification measure from the user's perspective (Richardson, 2000; Ben Bouallègue *et al.*, 2015).

Skill scores and bootstrapping

Skill scores are computed in order to estimate the relative benefit of using a forecast compared to a reference one (e.g. Wilks, 2006b). A skill score Sk is computed as :

$$Sk = 1 - \frac{S}{S^{\star}}, \tag{C.13}$$

where S and $S^\star$ are the scores of the forecast under assessment and of the reference forecast, respectively. Quantile skill score (QSS) and continuous ranked probability skill score (CRPSS) are computed following (C.13).

Since AUC' is positively oriented (the higher the better), the discrimination *gain* (or RUC skill score, $RUCSS$) of a forecast with respect to a reference one is estimated as follows:

$$RUCSS = \frac{AUC}{AUC^\star} - 1 \tag{C.14}$$

where AUC and $AUC^\star$ are the area under the RUC curve of the forecast under assessment and of the reference forecast, respectively.

In order to check whether the benefit of an approach is statistically significant, a block-bootstrapping method is applied. Bootstrapping is a common resampling technique that allows one to estimate the statistical consistency of the results and to draw confidence intervals (Efron and Tibshirani, 1986). Its application to meteorological data has been popularized by Hamill (1999). In the following, we apply a block-bootstrapping method where each block corresponds to a single day of the verification period. Considering each day as a separate block, bootstrapping assesses the score variability over the verification period and not between locations.

C.5 Results and discussion

Two configurations of the calibration process are tested using the COSMO-DE-EPS hourly averaged global radiation forecasts. The first is QR with the basic predictors as defined in Eq.(C.7), hereafter referred to as QR_s. The second is PQR including basic predictors, weather predictors and their interactions as defined in Eq.(C.9), hereafter referred to as PQR_o in an optimised setup. An example of QR_s and PQR_o calibrated quantile forecasts is shown in Fig. C.1(b) and C.1(c), respectively. Before assessing the performance of the derived calibrated forecasts, the optimisation of the penalized regression setup and the interpretation of the predictor selection is discussed.

Regularization setup

The regularization parameter λ_r is the key parameter of the penalized regression scheme (Eq. C.10). In the perspective of an operational implementation, an adequate value of λ_r is aimed to be found for the different probability levels of interest and for each training dataset automatically. A standard approach for the optimisation of λ_r consists in using a leave-n-out score as measure of performance (Bröcker, 2010). In our case, the natural performance measure is QS, and the optimal λ_r is found among a set of predefined values ($\{0., 0.5, 1, 5, 10, 20, 40, 80, 160\}$). $3/4$ of the training dataset is used for the coefficient estimation and $1/4$ for the performance evaluation. The dataset subdivision is performed randomly and the process is reiterated 100 times.

Another crucial calibration parameter, independently of the calibration method, is the length of the training period. Based on the assumption of forecast error stationarity, the calibration process generally rewards small training periods as shown in previous studies (Bentzien and Friederichs, 2012; Ben Bouallègue, 2013). In an operational setting, short training periods are also appreciated, in particular when facing regular model changes. On the other hand, using a large amount of past data allows capturing a greater diversity in terms of weather situations and generally increases the robustness of the coefficient estimation. In the following, the choice of the training length is based on results comparing 45 and 90 day training periods.

The impact of the penalty term and the influence of the training length on the calibrated forecast performance is illustrated in Fig. C.2 and C.3. The results focus on forecasts valid at 12UTC (forecast

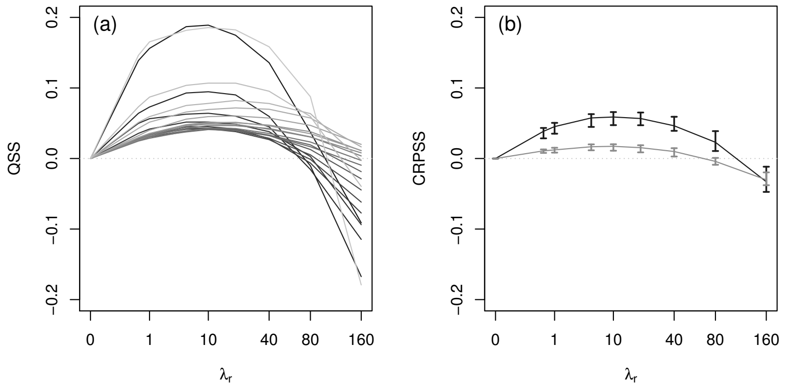

Figure C.2: Impact of the regularization parameter λ_r on the calibrated forecast performance for spring 2013, 03UTC runs, 12UTC validity time. (a) Quantile skill score (QSS) as a function of λ_r when using a training length of 45 days for the probability levels corresponding to the ensemble members of rank 1, 2, ..., 19, 20, represented with shades of grey, from light to dark grey, respectively. (b) Continuous ranked probability skill score (CRPSS) as a function of λ_r when using training lengths of 45 days (black line) and 90 days (grey line). The vertical bars indicate the 5% and 95% confidence intervals estimated by block-bootstrapping. In all cases, the reference forecast is the corresponding calibrated forecast without regularization ($\lambda_r = 0$).

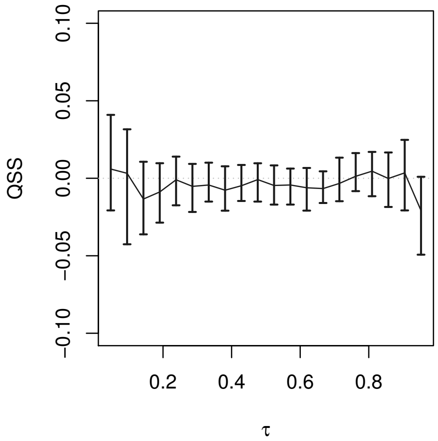

Figure C.3: Impact of the length of the training period, in term of QSS, as a function of the nominal probability level τ. QSS is computed for 45 days training calibrated forecasts with 90 days training calibrated forecasts as reference. The regularization parameter λ_r is optimized in both cases. The vertical bars correspond to the 5%-95% confidence intervals estimated by block-bootstrapping. Results for spring 2013, 03UTC runs, 12UTC validity time.

horizon of 9 hours). QSS of the 20 ensemble quantile forecasts as a function of the regularization parameter is plotted in Fig. C.2(a). In Fig. C.2(b), CRPSS as a function of λ_r is shown for training lengths of 45 and 90 days. The skill scores in Fig. C.2 are computed with the calibrated forecasts without penalization ($\lambda_r = 0$) and a training period of the same length (45 and 90 days, respectively) as reference forecasts. Fig. C.3 shows QSS as a function of the probability level τ comparing performance with training periods of 45 and 90 days.

Using penalized quantile regression, the number of predictors playing a role in the calibration is reduced as λ_r increases. In Fig. C.2, increasing λ_r up to 10 has a positive impact for all probability levels and a significant positive impact in terms of CRPS: overfitting is avoided especially when the training period is small (black line in Fig. C.2(b)). When λ_r becomes large ($\lambda_r > 10$), the number of predictors with non-zero coefficients is further reduced such that the full information content of the training dataset is no more captured. From Fig. C.2, it can be inferred that neither the training length nor the probability level have a decisive influence on the optimal value of λ_r such that choosing $\lambda_r = 10$ in all cases could be considered as an alternative to the computationally expensive parameter optimisation process.

Increasing the training period, the benefit of the regularization is less important (grey line in Fig. C.2(b)). In that case, overfitting is less of a critical issue. However, doubling the training length has no significant impact on the verification results (Fig. C.3), though this statement could be erroneous for extreme probability levels (not considered here) as suggested by the plot. Since similar results have been found for the QR_s approach (not shown), only results of calibration based on a 45 day training period ($N_d = 45$) are shown in the following.

Selected predictors

The regularization of QR is applied in order to avoid overfitting by selecting adequate predictors and deactivating the other ones. The estimated quantile coefficients $\hat{\boldsymbol{\beta}}$ derived with PQR_o are analysed aiming at a diagnosis of the forecast error. For this purpose, the absolute value of the N_p coefficients associated with the N_p predictors are normalized computing:

$$w_p = \frac{\mid \hat{\beta}_p \mid}{\sum_{k=1}^{Np} \mid \hat{\beta}_k \mid} \tag{C.15}$$

for $p \in \{1, ..., N_p\}$ such $\sum_{p=1}^{Np} w_p = 1$. The normalized coefficient w_j corresponds to the weight of the predictor p in the regression model and is therefore interpreted as the predictor information contribution for the derivation of the corresponding calibrated quantile forecast. The number of nonzero coefficients actually corresponds to the number of degrees of freedom of the lasso (Zou *et al.*, 2007)

The regression coefficients are estimated for each verification day based on the corresponding rolling training period. An analysis of the weight distribution is used to determine which predictors are relevant as a function of the probability level of interest. In Fig. C.4, the 10 most influential predictors for the calibration of the 5%, 50%, and 95%-quantile forecasts are shown, respectively. The median weight is used to rank the predictors and the weight variability is assessed by means of the 5% and 95%-quantiles of the weight distributions.

Fig. C.4 shows that, for all probability levels, the most important predictor is the first guess quantile forecast or its power transformation which is an indication of the good performance of the COSMO-DE model. Low cloud cover and total cloud cover forecasts are relevant sources of information in particular for low and intermediate probability levels (Fig. C.4(a) and (b)). These variables appear as predictors often in the form of "extreme" probabilistic products (min, max) and as interaction terms

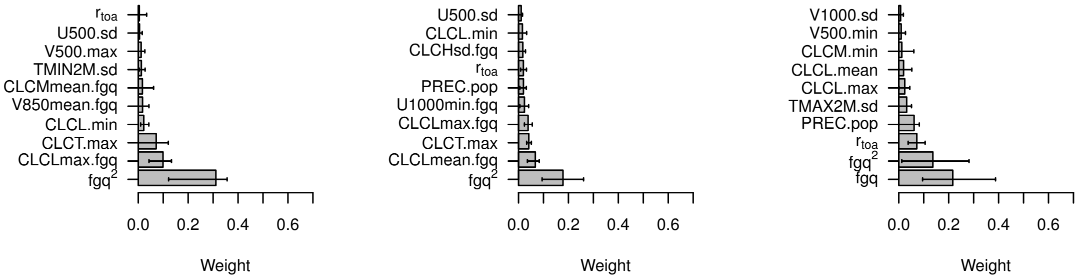

Figure C.4: Most influential predictors for the calibration of quantile forecasts at probability levels (a) 5%, (b) 50% and (c) 95%. The large bars indicate the median and the thin bars the 5%-95% of the predictor weight distribution over spring 2013. Results for the 03UTC forecast runs, 12UTC validity time.

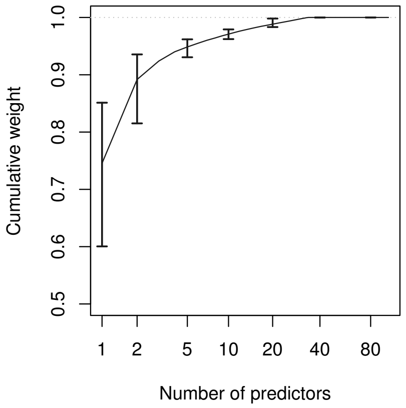

Figure C.5: Cumulative predictor weight as a function of the number of ranked predictors. The bars indicate the 5%-95% variability interval over spring 2013. Results for the 03UTC forecast runs, 12UTC validity time.

(in combination with the first guess quantile). This suggests the importance of the definition of adequate products from the ensemble forecasts on one hand, and of the use of interaction terms on the other hand. For high probability levels, the basic predictors (fgq, fgq^2 and r_{toa}) have the predominant role (Fig. C.4(c)). We can also note that the variability of the predictor weights encourages the use of a short rolling window as training dataset in order to go along with seasonal forecast error characteristics.

In the perspective of an operational use of PQR_o, it is also analysed the number of predictors that are actually relevant for the estimation of the conditional quantiles. For each set of coefficients $\hat{\boldsymbol{\beta}}$, the predictors are ranked as a function of their weight and the derived cumulative weight is used to relate the information content with the model complexity. Fig. C.5 presents statistics of this diagnostic measure. It is shown that a single predictor captures between 60% and 80% of the information, two predictors between 80% and 90% and the five most meaningful predictors at each calibration step around 95%. We can infer that deriving calibrated fields based on only the most significant predictors rather than the complete set should enable to have a computationally effective calibration

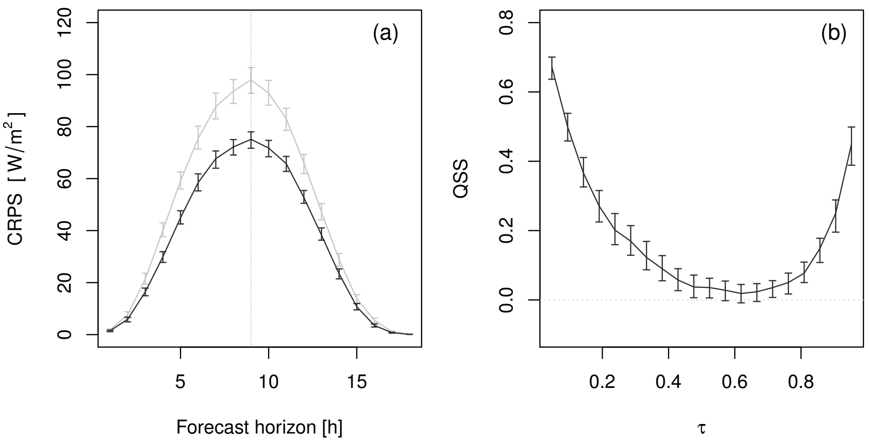

Figure C.6: (a) Continuous ranked probability score (CRPS) of the raw ensemble forecasts (in grey) and of the PQR_o calibrated forecasts (in black) as a function of the forecast horizon. The vertical bars indicate the 5%-95% confidence intervals estimated by block-bootstrapping. Results for spring 2013, 03UTC forecast runs. (b) Quantile skill score (QSS) of the PQR_o calibrated forecasts with the raw ensemble as reference plotted as a function of the probability level τ. The vertical bars indicate the 5%-95% confidence intervals estimated by block-bootstrapping. Results for spring 2013, 03UTC runs, 12UTC validity time.

scheme without significant loss of information content.

Verification results

The general impact of the calibration process is first shown in Fig. C.6(a). The CRPS of the PQR_o calibrated forecasts is compared to the CRPS of the raw ensemble as a function of the forecast horizon. Calibration has a positive impact on the quality of the ensemble for all day-hours with an improvement in terms of CRPSS of about 22%. In the following are shown results focusing only on forecasts valid at 12UTC (9 hour lead time), in order to avoid misinterpretation due to the strong influence of the diurnal cycle of solar radiation on the results. In Fig. C.6(b), QSS is plotted as a function of the probability level. Major improvements affect small and high probability levels and less intermediate ones, but QSS is significantly positive in any case. A deeper insight into the forecast performance is provided investigating now the forecast reliability and discrimination ability.

PIT histograms are shown in Fig. C.7. The histogram for the raw ensemble forecast in Fig. C.7(a) exhibits a clear U-shape which indicates that the ensemble suffers from underdispersiveness. The U-shape of the histogram is also partially interpreted as a combination of conditional biases. After calibration, using either QR_s or PQR_o, the PIT histograms in Fig. C.7(b) and C.7(c), respectively, are flat indicating that the calibrated ensemble forecasts fulfil the necessary condition of reliability. The U-shaped line of the QSS as a function of τ in Fig. C.6(b) can be directly related to the impact of calibration on the PIT histograms. The raw ensemble is strongly underdispersive, and thus the lower/higher quantiles are generally overestimated/underestimated. The higher and lower quantiles hence benefit more from post-processing than the central quantiles. An other way to check the ensemble reliability is to plot a set of quantile reliability curves.

Reliability of the ensemble at the quantile level is explored by means of reliability diagrams as shown in Fig. C.8 for the 5%, 50%, and 95%-quantile forecasts. While the 50%-quantile forecast of the raw

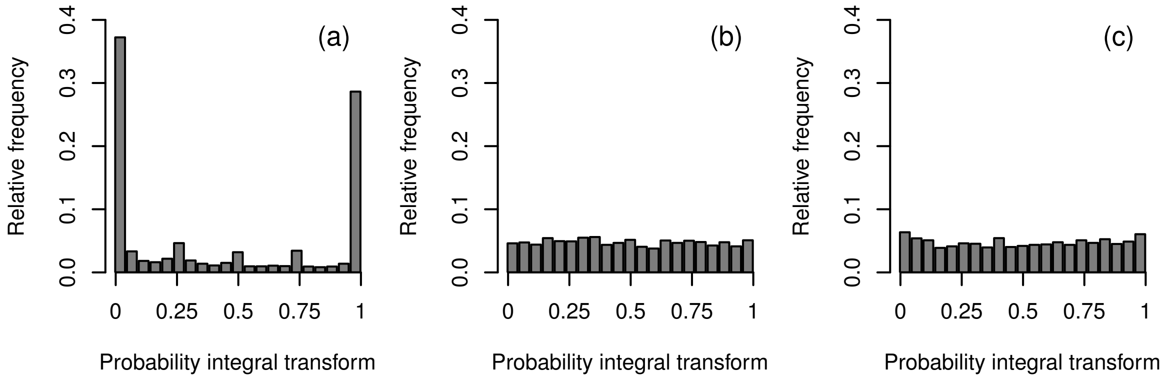

Figure C.7: Probability integral transform (PIT) histograms of (a) the raw COSMO-DE-EPS quantile forecasts, (b) the QR_S calibrated quantile forecasts and (c) the PQR_o calibrated quantile forecasts. Results for spring 2013 , 03UTC forecast runs, 12UTC validity time.

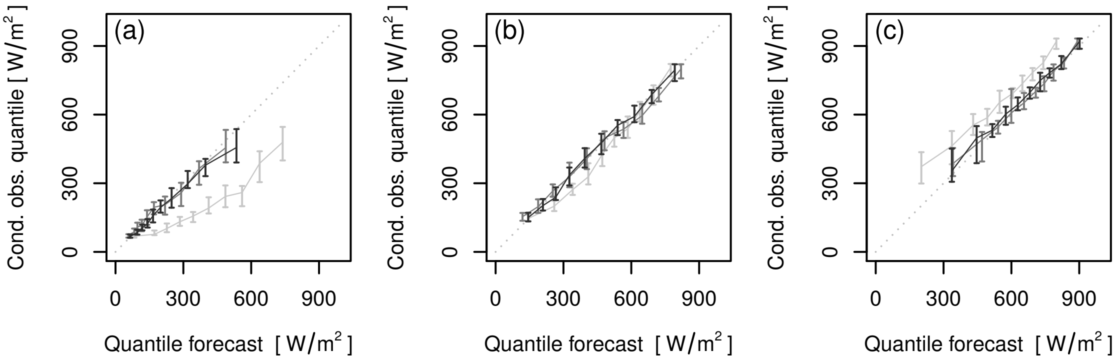

Figure C.8: Reliability diagram of the COSMO-DE-EPS quantile forecasts at probability levels (a) 5%, (b) 50%, and (c) 95%. Reliability curve of the first guess forecasts (light grey lines), of the QR_s calibrated forecasts (dark grey lines) and of the PQR_o calibrated forecasts (black lines). The vertical bars indicate the 5%-95% confidence intervals estimated by block-bootstrapping. Results for spring 2013, 03UTC forecast runs, 12UTC validity time.

ensemble shows reasonably good reliability, the 5% and 95%-quantile forecasts exhibit clear deficiencies as already discussed in a previous study (Ben Bouallègue, 2015). Calibration helps to correct for systematic positive biases for low probability levels and negative biases for high probability levels. Reliability curves after applying quantile regression, in the form of QR_s or PQR_o, include the diagonal lines within their reliability confidence intervals. Those results demonstrate that calibration of global radiation ensemble forecast based on quantile regression is effective and thus allows providing reliable probabilistic guidance to the users.

The benefit of the complex approach (PQR_o) in comparison with the basic one (QR_s) is now assessed. Fig. C.9(a) depicts the QSS of the PQR_o calibrated forecasts, using the QR_s calibrated forecasts as benchmark plotted as a function of the probability level. The use of additional predictors improves the skill of the forecast significantly at nearly all probability levels. The gain is approx. 5% with a peak at intermediate probability levels. In Fig. C.9(b), the focus lies on the discrimination ability of the calibrated forecasts. The results in terms of RUCSS share similarities with the results in terms of QSS: the discrimination ability is significantly improved at intermediate probability levels with the

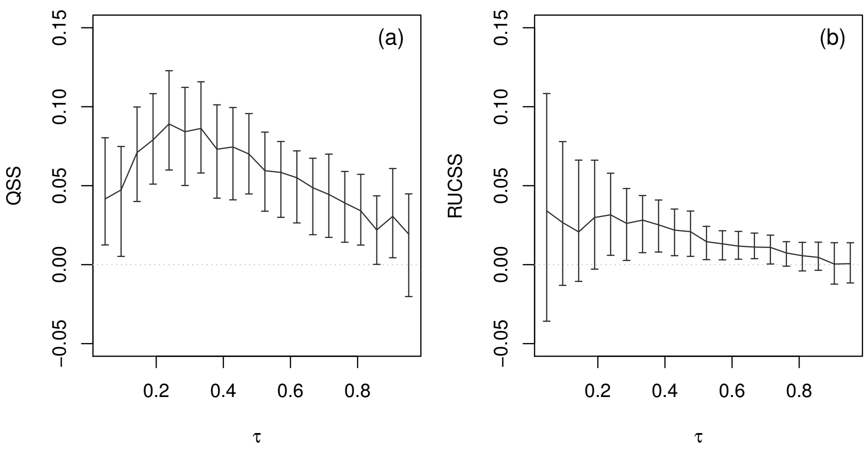

Figure C.9: (a) Quantile skill score (QSS) of the PQR_o against QR_s calibrated forecasts as a function of the probability level τ. The vertical bars indicate the 5%-95% confidence intervals estimated by block-bootstrapping. (b) Discrimination gain ($RUCSS$) comparing PQR_o against QR_s calibrated forecasts as a function of the probability level. The vertical bars indicate the 5%-95% confidence intervals estimated by block-bootstrapping. Results for spring 2013, 03UTC forecast runs, 12UTC validity time.

use of adequate predictors. These results indicate that PQR_o outperforms QR_s thanks to an increase of the forecast information content, and demonstrate that, with comparable reliability performance, PQR_o with respect to QR_s calibrated forecasts have greater or equivalent value.

Finally, a subjective verification based on a case study is proposed. In the example of Fig. C.1, calibrated forecasts are plotted in counterpoint to the raw COSMO-DE-EPS forecasts. Calibrated forecasts with QR_s and PQR_o are shown in Fig. C.1(b) and C.1(c), respectively. We see that the calibration step increases the spread of the ensemble which captures now the observation variability. Moreover, the use of adequate weather predictors with PQR_o improves the sharpness of the forecast: the quantile forecasts are indeed more grouped in Fig. C.1(c) than in C.1(b). This analysis is in perfect accordance with the previous objective verification results.

Discussion

The results presented here, based on a 3-month verification period covering spring 2013, demonstrate that quantile regression performs well for the calibration of ensemble global radiation forecasts. An implementation over a complete year could provide interesting information about the role of different predictors over different seasons. Though the verification period is relatively small, some aspects of the results and of the methodology can be generalized and discussed.

First, using selected predictors in the regression scheme, the improvement of the forecast performance is on average of about 5% in terms of CRPS, and tends to decreases with the lead time (not shown). Similar results which assess the benefit of a weather dependent calibration approach with respect to a standard one has been found focusing on wind forecasts and performing the ensemble calibration based on an *analog* technique Junk *et al.* (2015). Analog ensemble forecasts are drawn from past observations whose related past forecasts share analogies with the forecast under calibration. In that case, a weather dependent calibration consists in finding analogies based not only on the forecast to be calibrated but also on correlated variables. However, we can emphasize that PQR uses a rolling window of limited size as training dataset while the analog approach requires a long record of observations and forecasts with a frozen configuration of the numerical model.

Second, calibration improves drastically the performance of quantile forecasts at high and low probability levels correcting for strong biases. However, the use of additional selected predictors has a limited impact on the forecasts at these probability levels. Indeed, the weather dependent calibration approach mainly improves quantile forecasts at intermediate levels. This result reveals the need of appropriate predictors targeted to the extreme quantile levels or more generally targeted to specific weather phenomena. For example, information about the development of low stratus cloud provided by a physically-based post-processing approach could be included as potential predictor (Köhler *et al.*, 2016).

Third, in our study, calibration is applied homogeneously over space: the coefficients are estimated and applied considering indiscriminately all stations (model grid points) at a given forecast horizon. At each station, calibration clearly improves the forecasts in terms of CRPS (not shown). This result indicates that spatial homogeneous calibration performs appropriately on our dataset. Moreover, a recent study focusing on total precipitation suggests that spatial homogeneous calibration based on selected predictors via lasso might have similar performance as computationally expensive spatial calibration techniques (Wahl, 2015).

Finally, we can note that calibrated quantile forecasts are provided at each lead time and location separately. So, after calibration, the information about the forecast uncertainty across time, space and variables is lost. Dependence structures in the ensemble forecast can be in a second step recovered applying, for example, the ensemble copula coupling (ECC) approach. ECC consists in transferring the rank structure of the original ensemble to the calibrated one expressed optimally in the form of calibrated quantile forecasts (Schefzik *et al.*, 2013). The paper at hand only focuses on the *univariate* calibration step while another study in preparation will explore the generation of consistent scenarios from calibrated ensemble forecasts based on empirical copulas.

C.6 Conclusion

Post-processing of global radiation ensemble forecasts from the convection permitting ensemble system COSMO-DE-EPS is a necessary step in order to correct for systematic statistical inconsistencies. Quantile regression is a suitable method for this purpose. Reliable quantile forecasts are derived with a simple approach using as predictors the first guess quantile forecasts, their power transformation, and a variable related to the solar geometry.

In order to improve the calibrated forecast sharpness, post-processing based on penalized quantile regression with an optimisation of the regularization term is applied. Meaningful predictors are selected from probabilistic products of ensemble direct model outputs and their interaction with the first guess quantile forecasts. In particular, the analysis of the predictor coefficients indicates that low cloud cover and total cloud cover forecasts play an important role for the optimisation of the conditional quantiles.

The low degree of complexity and the computational efficiency of a post-processing based on penalized quantile regression should favour an operational implementation. The rigorous selection of adequate predictors allows deriving reliable and sharp forecasts whose increased discrimination ability should benefit users' decision making processes.

Acknowledgments

This work has been done in the framework of the EWeLiNE project (*Erstellung innovativer Wetter- und Leistungsprognosemodelle für die Netzintegration wetterabhängiger Energieträger*) funded by the German Federal Ministry for Economic Affairs and Energy. Sabrina Wahl (University of Bonn), Carmen Köhler (Deutscher Wetterdienst), and two anonymous reviewers are acknowledged for helpful discussions and/or valuable comments on a previous version of this manuscript.

Appendix D

Generation of scenarios from calibrated ensemble forecasts with a dual ensemble copula coupling approach

The content of this appendix is the author′s version of a manuscript submitted in November 2015 to the Monthly Weather Review.

Generation of scenarios from calibrated ensemble forecasts with a dual ensemble copula coupling approach

Zied Ben Bouallègue^{a,b}, Tobias Heppelmann^{a},
Susanne E. Theis^{a}, Pierre Pinson^{c}

[a] Deutscher Wetterdienst, Offenbach, Germany
[b] Meteorological Institute, University of Bonn, Germany
[c] Technical University of Denmark, Denmark

Abstract

Probabilistic forecasts in the form of ensemble of scenarios are required for complex decision-making processes. Ensemble forecasting systems provide such products but the spatio-temporal structures of the forecast uncertainty is lost when statistical calibration of the ensemble forecasts is applied for each lead time and location independently. Non-parametric approaches allow the reconstruction of spatio-temporal joint probability distributions at a low computational cost. For example, the ensemble copula coupling (ECC) method rebuilds the multivariate aspect of the forecast from the original ensemble forecasts. Based on the assumption of error stationarity, parametric methods aim to fully describe the forecast dependence structures. In this study, the concept of ECC is combined with past data statistics in order to account for the autocorrelation of the forecast error. The new approach, called d-ECC, is applied to wind forecasts from the high resolution ensemble system COSMO-DE-EPS run operationally at the German weather service. Scenarios generated by ECC and d-ECC are compared and assessed in the form of time series by means of multivariate verification tools and in a product oriented framework. Verification results over a 3 month period show that the innovative method d-ECC outperforms or performs as well as ECC in all investigated aspects.

D.1 Introduction

Uncertainty information is essential for an optimal use of a forecast (Krzysztofowicz, 1983). Such information can be provided by an Ensemble Prediction System (EPS) which aims at describing the flow-dependent forecast uncertainty (Leutbecher and Palmer, 2008). Several deterministic forecasts are run simultaneously accounting for uncertainties in the description of the initial state, the model parametrization and, for limited area models, the boundary conditions. Probabilistic products are derived from an ensemble, tailored to specific user’s need. For example, wind forecasts in the form of quantiles at selected probability levels are of particular interest for actors in the renewable energy sector (Pinson, 2013).

However, probabilistic products generally suffer from a lack of reliability, the system showing biases and failing to fully represent the forecast uncertainty. Statistical techniques allow to adjust the ensemble forecast correcting for systematic inconsistencies (Gneiting *et al.*, 2007). This step known as calibration is based on past data and usually focuses on a single or few aspects of the ensemble forecast. For example, calibration of wind forecast can be performed by univariate approaches

(Bremnes, 2004; Sloughter *et al.*, 2010; Thorarinsdottir and Gneiting, 2010) or bivariate methods which account for correlation structures of the wind components (Pinson, 2012; Schuhen *et al.*, 2012). These calibration procedures provide reliable predictive probability distribution of wind speed or wind components for each forecast lead time and location independently. Decision making problems can however require information about the spatial and/or temporal structure of the forecast uncertainty. Examples of application in the renewable energy sector resemble the optimal operation of a wind-storage system in a market environment, the unit commitment over a control zone or the optimal maintenance planning (Pinson *et al.*, 2009). In other words, scenarios that describe spatio-temporal wind variability are relevant products for end-users of wind forecasts.

The generation of scenarios from calibrated ensemble forecasts is a step that can be performed with the use of empirical copulas. The empirical copula approaches are non-parametric and, in comparison with parametric approaches (Keune *et al.*, 2014; Feldmann *et al.*, 2015), simple to implement and computationally cheap. Empirical copulas can be based on climatological records (Schaake Shuffle (ScSh); Clark *et al.*, 2004) or on the original raw ensemble (ensemble copula coupling (ECC); Schefzik *et al.*, 2013). ECC, which consists in the conservation of the ensemble member rank structure from the original ensemble to the calibrated one, has the advantage to be applicable to any location of the model domain without restriction related to the availability of observations. However, unrealistic scenarios can be generated by the ECC approach when the post-processing indiscriminately increases the ensemble spread to a large extent. Non-representative correlation structures in the raw ensemble are magnified after calibration leading to unrealistic forecast variability. As a consequence, ECC can deteriorate the ensemble information content when applied to ensembles with relatively poor reliability as suggested, for example, by verification results in Flowerdew (2014).

In this paper, a new version of the ECC approach is proposed in order to overcome the generation of unrealistic scenarios. Focusing on time series, a temporal component is introduced in the ECC scheme accounting for the autocorrelation of the forecast error over consecutive forecast lead times. The assumption of forecast error stationarity, already adopted for the development of fully parametric approaches (Pinson *et al.*, 2009; Schölzel and Hense, 2011), is exploited in combination with the structure information of the original scenarios. The new approach based on these two sources of information, past data and ensemble structure, is called *dual* ensemble copula coupling (d-ECC). Objective verification is performed in order to show the benefit of the proposed approach with regard to the standard ECC.

The manuscript is organized as follows: Section D.2 describes the dataset used to illustrate the manuscript as well as the calibration method applied to derive calibrated quantile forecasts from the raw ensemble. Sections D.3 and D.4 introduce the empirical copula approaches for the generation of scenarios and discuss in particular the ECC and d-ECC methods. Section D.5 describes the verification process for the scenario assessment. Section D.6 presents the results obtained by means of multivariate scores and in a product oriented verification framework.

D.2 Data

Ensemble forecasts and observations

COSMO-DE-EPS is the high resolution ensemble prediction system run operationally at DWD. It consists of 20 COSMO-DE forecasts with variations in the initial conditions, the boundary conditions and the model physics (Gebhardt *et al.*, 2011; Peralta *et al.*, 2012). COSMO-DE-EPS follows the multi-model ensemble approach, with 4 global models driving each 5 physically perturbed members. The ensemble configuration implies a clustering of the ensemble members as a function of the driving global model when large scale structures dominate the forecast uncertainty.

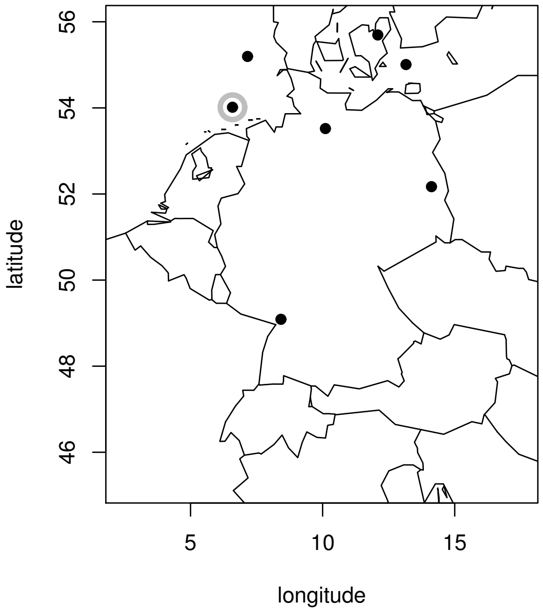

Figure D.1: Map of Germany and neighboring areas (approximately the COSMO-DE domain) with latitude/longitude axes. Location of the 7 wind stations used in this study. The station FINO1 is highlighted with a grey mark.

The focus is here on wind forecasts at 100 meter height above ground. The post-processing methods are applied to forecasts of the 00UTC run with an hourly output interval and a forecast horizon of up to 21 hours. The observation dataset comprises quality controlled wind measurements from 7 stations: Risoe, FINO1, FINO2, FINO3, Karlsruhe, Hamburg and Lindenberg, as plotted in Figure D.1. The verification period covers a 3 month period: March, April and May 2013.

Figure D.2(a) shows an example of a COSMO-DE-EPS wind forecast at hub-height. The forecast is valid on day March 2, 2013, at station FINO1 (see Figure D.1). The ensemble members are drawn in grey while the corresponding observations are drawn in black. In Figure D.2(b), the raw ensemble forecast is interpreted in the form of quantiles.

Formally, a quantile Q_τ at probability level τ (with $0 \leq \tau \leq 1$) is defined as:

$$Q_\tau := F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\} \tag{D.1}$$

where F_Y is the cumulative probability distribution of the random variable $Y \in \Re$:

$$F_Y(y) = Pr(Y \leq y). \tag{D.2}$$

In practice, at each forecast lead time, the member of rank m can be interpreted as a quantile forecast at probability level τ_m:

$$\tau_m = \frac{m}{M+1} \tag{D.3}$$

where M is the number of ensemble members.

In the example of Figure D.2, the raw ensemble is not able to capture the observation variability. Calibration aims to correct for this lack of reliability by adjusting the mean and enlarging the spread of the ensemble forecast.

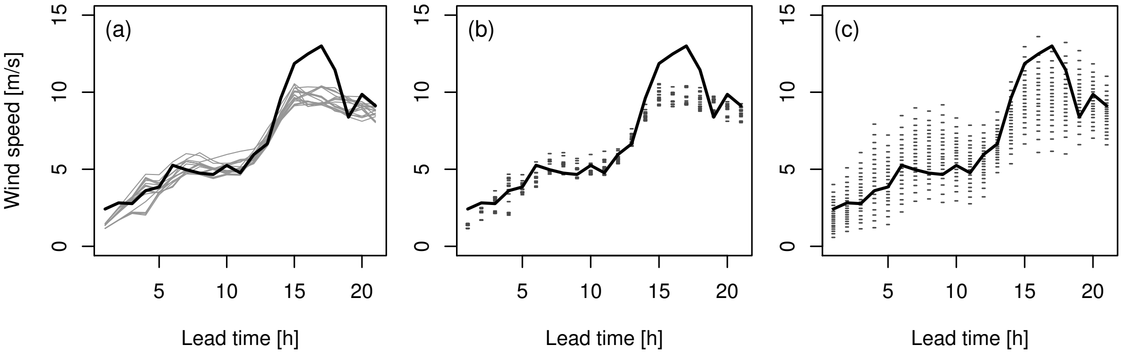

Figure D.2: Wind speed at 100 meter height above ground, on March 2, 2013, at station FINO1: (a) COSMO-DE-EPS forecast (grey lines), (b) raw ensemble forecast in the form of quantiles (sorted members, see text), (c) calibrated quantile forecasts, and the corresponding observations (black lines).

Calibrated ensemble forecasts

Since COSMO-DE-EPS forecasts have shown to suffer from statistical inconsistencies (Ben Bouallègue, 2013; Ben Bouallègue, 2015), calibration has to be applied in order to provide reliable forecasts to the users. The method applied in this study is the bivariate Non-homogeneous Gaussian Regression (EMOS, Schuhen *et al.*, 2012). The mean and variance of each wind component as well as the correlation between the two components characterize the predictive bivariate normal distribution. Corrections applied to the raw ensemble mean and variance are optimized by minimizing the continuous ranked probability score ($CRPS$; Matheson and Winkler, 1976). The calibration coefficients are estimated for each station and each lead time separately (local version of EMOS), based on a training period being defined as a moving window of 45 days.

The final calibrated products considered here are M equidistant forecasts of wind speed estimated for each location and each forecast lead time separately, where the M probability levels associated to the forecast quantiles follow Eq. (D.3). Calibrated quantile forecasts are shown in Figure D.2(c). The spread of the ensemble is increased with respect to Figure D.2(b) and thus the observation variability is now captured by the forecast. From a statistical point of view the calibration method provides reliable ensemble marginal distributions and reliable quantile forecasts as checked by means of rank histograms and quantile reliability plots (not shown). The performance of the applied calibration technique is similar to the one obtained by other methods such as quantile regression (Koenker and Bassett, 1978; Bremnes, 2004).

Information about spatial and temporal dependence structures, which are crucial in many applications, are however not available any more after this calibration step (see Figure D.2(c)). The next post-processing step consists then in the generation of consistent scenarios based on the calibrated samples.

D.3 Generation of scenarios

The generation of scenarios with empirical copulas is here briefly described. For a deeper insight into the methods, the reader is invited to refer to the original article of Schefzik *et al.* (2013), or to Wilks (2014) and references within.

First, consider the multivariate cumulative distribution function (*cdf*) $\mathcal{F}$ defined as:

$$\mathcal{F}(y_1, ..., y_K) = Pr[Y_1 \le y_1, ..., Y_K \le y_K] \tag{D.4}$$

of a random vector $(Y_1, ..., Y_K)$ with $y_1, ..., y_K \in \mathbb{R}$. As in Eq. (D.2), we define F_i the marginals as:

$$F_i(y_i) = Pr[Y_i \le y_i]. \tag{D.5}$$

The Sklar's theorem (Sklar, 1959) states that $\mathcal{F}$ can be expressed as:

$$\mathcal{F}(y_1, ..., y_K) = \mathcal{C}(F_1(y_1), F_K(y_K)) \tag{D.6}$$

where $\mathcal{C}$ is a copula that links an K-variate cumulative distribution function $\mathcal{F}$ to its univariate marginal *cdf*s $F_1, ..., F_K$.

In Eq. (D.6), a joint distribution is represented as univariate margins plus copulas. The problem of estimating univariate distributions and the problem of estimating dependence can therefore be treated separately. Univariate calibration marginal *cdf*s $F_1, ..., F_K$ are provided by the calibration step described in the previous section. The choice of the copula $\mathcal{C}$ depends on the application and on the size K of the multivariate problem. We focus here on empirical copulas since they are suitable for problems with high dimensionality.

We denote $\mathcal{H}$ the empirical copula. $\mathcal{H}$ is based on a multivariate dependence template, a specific discrete dataset $\boldsymbol{z}$ defined in $\mathbb{R}^K$. The chosen dataset is described formally as:

$$\boldsymbol{z} := \left\{(z_1^1, ..., z_1^{N_s}), ..., (z_K^1, ..., z_K^{N_s})\right\} \tag{D.7}$$

consisting of K tuples of size N_s with entries in $\mathbb{R}$. In other words, K is the dimension of the multivariate variable and N_s is the number of scenarios. The rank of z_k^n for $n \in \{1, ..., N_s\}$ and $k \in \{1, ..., K\}$ is defined as:

$$u_k^n := \sum_{i=1}^{N_s} \mathbb{I}(z_k^i \le z_k^n) \tag{D.8}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function taking value 1 if the condition in parenthesis is true and zero otherwise. The empirical copula $\mathcal{H}$ induced by the dataset $\boldsymbol{z}$ is given by:

$$\mathcal{H}(\frac{j_1}{N_s}, ..., \frac{j_K}{N_s}) := \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbb{I}(u_1^i \le j_1, ..., u_K^i \le j_K) \tag{D.9}$$

$$= \frac{1}{N_s} \sum_{i=1}^{N_s} \prod_{k=1}^{K} \mathbb{I}(u_k^i \le j_k) \tag{D.10}$$

for integers $0 \le j_1, ..., j_K \le N_s$.

In practice, N_s equidistant quantiles of F_k with $k \in \{1, ..., K\}$ are derived from the univariate calibration step:

$$\boldsymbol{q} := \left\{(q_1^1, ..., q_1^{N_s}), ..., (q_K^1, ..., q_K^{N_s})\right\} \tag{D.11}$$

with

$$q_k^n := F_k^{-1}(\tau_n); \quad n \in \{1, .., N_s\} \tag{D.12}$$

where τ_n is defined in Eq. (D.3). The sample $\boldsymbol{q}$ is rearranged following the dependence structure of the reference template $\boldsymbol{z}$. The permutations $\pi_k(n) := R_k^n$ for $n \in \{1, .., N_s\}$ are derived from the

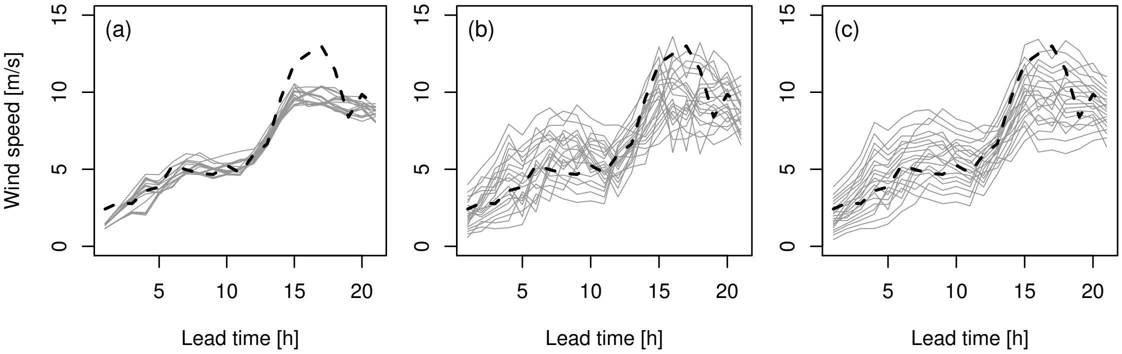

Figure D.3: Same example as in Figure D.2: (a) COSMO-DE-EPS scenarios, (b) ECC derived scenarios, (c) d-ECC derived scenarios, and the corresponding observations (black lines).

univariate ranks $u_k^1, ..., u_k^{N_s}$ for $k \in \{1, .., K\}$ and applied to the univariate calibrated sample $\boldsymbol{q}$. The post-processed scenarios $\tilde{x}_k^1, ..., \tilde{x}_k^{N_s}$ for each margin k is expressed as:

$$\tilde{x}_k^1 := q_k^{\pi_k(1)}, ..., \tilde{x}_k^{N_s} := q_k^{\pi_k(N_s)} \tag{D.13}$$

The multivariate correlation structures are generated based on the rank correlation structures of a sample template $\boldsymbol{z}$. The empirical copulas presented here only differ in the way $\boldsymbol{z}$ is defined. In the following, let $t \in \{1, ..., T\}$ be a lead time and let $L := T$. For simplicity, we consider here a single weather variable.

Ensemble copula coupling

The rank structure of the ensemble is preserved after calibration when applying the standard ensemble copula coupling approach (ECC). The raw ensemble forecast is denoted $\boldsymbol{x}$:

$$\boldsymbol{x} := \left\{(x_1^1, ..., x_1^M), ..., (x_K^1, ..., x_K^M)\right\} \tag{D.14}$$

where M is the ensemble size. ECC applies without restriction to any multivariate setting. The number of scenarios generated with ECC is however the same as the size of the original ensemble ($N_s = M$). The transfer of the rank structure from the raw ensemble forecast to the calibrated one consists then in taking $\boldsymbol{x}$ as the required template in Eq. (D.7).

Based on COSMO-DE-EPS forecasts of Figure D.3(a) (identical to Figure D.2(a)), an example of scenarios derived with ECC is provided in Figure D.3(b). The increase of spread after the calibration step implies a larger step-to-step variability in the time trajectories. Figure D.4 focuses on a single scenario highlighting the difference between the original and post-processed scenarios.

Dynamic ensemble copula coupling

ECC assumes that the ensemble prediction system correctly describes the spatio-temporal dependence structures of the weather variable. This assumption is quite strong and cannot be valid in all cases. On the other side, based on the assumption of error stationarity, parametric methods have been developed focusing on covariance structures of the forecast error (Pinson *et al.*, 2009; Schölzel and Hense, 2011). We propose a new version of the ECC approach which is an attempt to combine both information: the structure of the original ensemble and the error autocorrelation estimated from past data. Therefore, the new scheme is called dual ensemble copula coupling (d-ECC) as the copula relies on a dual source of information.

For this purpose, we denote $\boldsymbol{e}$ the forecast error defined as the difference between ensemble mean forecasts and observations:

$$\begin{aligned} \boldsymbol{e} &:= \{e_1, ..., e_T\} && \text{(D.15)} \\ &= \{y_1 - m(x_1), ..., y_T - m(x_T)\} && \text{(D.16)} \end{aligned}$$

where $m(x_t)$ and y_t are the ensemble mean and the corresponding observation at lead time $t \in \{1, ..., T\}$, respectively. The temporal correlation of the error is described by a correlation matrix $\boldsymbol{R_e}$ defined as:

$$\boldsymbol{R_e} = \begin{pmatrix} r_{e_1,e_1} & r_{e_1,e_2} & \cdots & r_{e_1,e_T} \\ r_{e_2,e_1} & r_{e_2,e_2} & \cdots & r_{e_2,e_T} \\ \vdots & \vdots & \ddots & \vdots \\ r_{e_T,e_1} & r_{e_T,e_2} & \cdots & r_{e_T,e_T} \end{pmatrix} \tag{D.17}$$

where $r_{e_{t_1},e_{t_2}}$ is the correlation coefficient of the forecast error at lead times t_1 and t_2. The empirical correlation matrix $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$ is estimated based on the training samples used for the univariate calibration step at the different lead times. In our setup, $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$ is regularly updated on a daily basis from the moving windows of 45 days defined as training datasets for the EMOS application.

Again here, we aim at constructing a template (Eq. (D.7)) in order to establish the correlation structures within the calibrated ensemble $\boldsymbol{q} := \{(q_1^1, ..., q_1^M), ..., (q_T^1, ..., q_T^M)\}$. In the d-ECC approach, the template is built performing the following steps:

1. Apply ECC with the original ensemble forecast $\boldsymbol{x}$ as reference sample template, in order to derive a post-processed ensemble of scenarios $\tilde{\boldsymbol{x}}$:

$$\tilde{\boldsymbol{x}} := \{(\tilde{x}_1^1, ..., \tilde{x}_1^M), ..., (\tilde{x}_T^1, ..., \tilde{x}_T^M)\}, \tag{D.18}$$

2. Derive the error correction $\boldsymbol{c}^i$ imposed to each scenario i ($i \in 1, ..., M$) of the original ensemble by this post-processing step:

$$\begin{aligned} \boldsymbol{c}^i &:= \{c_1^i, ..., c_T^i\} && \text{(D.19)} \\ &= \{\tilde{x}_1^i - x_1^i, ..., \tilde{x}_T^i - x_T^i\}, && \text{(D.20)} \end{aligned}$$

3. *Transformation step*: Apply a transformation to the correction $\boldsymbol{c}^i$ of each scenario based on the estimate of the error autocorrelation $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$ and its eigendecomposition $\hat{\boldsymbol{R}}_{\boldsymbol{e}} = \boldsymbol{U\Lambda U^{-1}}$ in order to derive the *adjusted corrections* $\breve{\boldsymbol{c}}^i$:

$$\begin{aligned} \breve{\boldsymbol{c}}^i &= \hat{\boldsymbol{R}}_{\boldsymbol{e}}^{\frac{1}{2}} \boldsymbol{c}^i && \text{(D.21)} \\ &= \boldsymbol{U\Lambda^{\frac{1}{2}} U^{-1}} \boldsymbol{c}^i, && \text{(D.22)} \end{aligned}$$

4. Derive the so-called *adjusted ensemble* $\breve{\boldsymbol{x}}$ such:

$$\breve{\boldsymbol{x}} := \{(\breve{x}_1^1, ..., \breve{x}_1^M), ..., (\breve{x}_T^1, ..., \breve{x}_T^M)\} \tag{D.23}$$

where a scenario $\breve{\boldsymbol{x}}^i = \{\breve{x}_1^i, ..., \breve{x}_T^i)\}$ of $\breve{\boldsymbol{x}}$ is defined as a combination of the original member and the adjusted error correction:

$$\breve{\boldsymbol{x}}^i = \boldsymbol{x}^i + \breve{\boldsymbol{c}}^i, \tag{D.24}$$

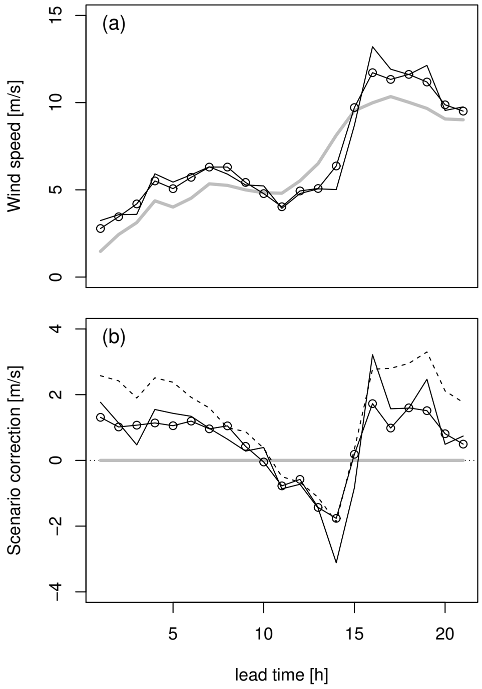

Figure D.4: Illustration of the concept of d-ECC based on the example of Figure D.3 showing (a) one among the 20 scenarios and (b) the correction applied to the original scenario after post-processing. The raw ensemble forecast (here the member 13) is represented in grey, the ECC scenario in black, and the d-ECC scenario in black with dots. The dashed line represents the scenario correction adjusted by the transformation step (see text).

5. Take $\breve{\boldsymbol{x}}$ as reference template in Eq. (D.7) so that the new empirical copula is based on the adjusted ensemble.

The d-ECC reference template $\breve{\boldsymbol{x}}$ combines the raw ensemble structure and the autocorrelation of the forecast error reflected in the adjusted member corrections. The transformation of the scenario corrections in Eq. (D.22) adjusts their correlation structure based on the error correlation matrix $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$. Taking the square root of the correlation matrix (Eq. D.22) resembles a signal processing technique which is described as a *coloring transformation* of a vector of random variables (Kessy *et al.*, 2015).

D.4 Illustration and discussion of d-ECC

Focusing on a single member, the d-ECC steps are illustrated in Figure D.4. First, the correction associated to each ECC scenario with respect to the corresponding original ensemble member is computed (black line in Figure D.4(b), Eq. D.20). This scenario correction is adjusted based on the assumption of temporal autocorrelation of the error (dashed line in Figure D.4(b), Eq. D.22). This adjusted scenario correction is then superimposed on the original ensemble forecast before to draw again the correlation structure of the adjusted ensemble.

The new scheme reduces to the standard ECC in the case where $rank(x_t^i) = rank(\breve{x}_t^i)$ for all $i \in \{1, ..., M\}$ and $t \in \{1, \ldots, T\}$, which means that the additional terms $\breve{\boldsymbol{c}}^i$ do not have any impact on the rank structure of the ensemble. This case occurs if:

- $\hat{\boldsymbol{R}}_{\boldsymbol{e}} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix, which means that there is no temporal correlation of the error in the original ensemble,
- $\boldsymbol{c} = \boldsymbol{0}$ where $\boldsymbol{0}$ is the null vector, which means that the calibration step does not impact the forecast, the forecast being already well calibrated.
- $\boldsymbol{c} = h \cdot \boldsymbol{J}$ where h is a constant and $\boldsymbol{J}$ an all-ones vector, which means that the calibration step corrects only for bias errors and the system is spread bias free.

So the d-ECC typically takes effect if calibration corrects the spread and if this correction is correlated in time at the member level.

Some more insight can be gained by looking at the following equations. Let the observation y_t and the post-processed ensemble members $\tilde{x}_t^i$ be realizations of random variables Y and $\tilde{X}$. Consider the covariance of the forecast error denoted κ and defined as:

$$\kappa_{t_1,t_2} := \mathbb{E}[(Y_{t_1} - m(\tilde{X}_{t_1}))(Y_{t_2} - m(\tilde{X}_{t_2}))] \tag{D.25}$$

where t_1 and t_2 are two lead times and $\mathbb{E}[\cdot]$ the expectation operator. It is assumed that the post-processed ensemble mean $m(\tilde{x}_t)$ is fully bias-corrected so that $\mathbb{E}[Y_t - m(\tilde{X}_t)] = 0$.

After post-processing, the forecast scenarios and observation time series are considered as drawn from the same multivariate probability distribution, so the forecast error covariance can also be expressed as:

$$\kappa_{t_1,t_2} = \mathbb{E}[(\tilde{X}_{t_1} - m(\tilde{X}_{t_1}))(\tilde{X}_{t_2} - m(\tilde{X}_{t_2}))] \tag{D.26}$$

$$= r_{\tilde{x}_{t_1},\tilde{x}_{t_2}} \sigma_{\tilde{x}_{t_1}} \sigma_{\tilde{x}_{t_2}} \tag{D.27}$$

where $r_{\tilde{x}_{t_1},\tilde{x}_{t_2}}$ refers to the correlation between $\tilde{x}_{t_1}$ and $\tilde{x}_{t_2}$ and $\sigma_{\tilde{x}_t}$ refers to the square root of the variances between the members of the calibrated ensemble $(\tilde{x}^1, ..., \tilde{x}^M)$ at lead time t. The corresponding estimators are the following:

$$\hat{\kappa}_{t_1,t_2} = \frac{1}{M-1} \sum_{i=1}^{M} [(\tilde{x}_{t_1}^i - m(\tilde{x}_{t_1}))(\tilde{x}_{t_2}^i - m(\tilde{x}_{t_2}))] \tag{D.28}$$

and

$$\hat{\sigma}_{\tilde{x}_t} = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (\tilde{x}_t^i - m(\tilde{x}_t))^2} \tag{D.29}$$

and

$$\hat{r}_{\tilde{x}_{t_1},\tilde{x}_{t_2}} = \frac{\hat{\kappa}_{t_1,t_2}}{\hat{\sigma}_{\tilde{x}_{t_1}} \hat{\sigma}_{\tilde{x}_{t_2}}}. \tag{D.30}$$

From Eq. (D.20) recall that

$$\tilde{x}_t^i = x_t^i + c_t^i \tag{D.31}$$

so we can rewrite the expression in Eq. (D.27) as

$$r_{\tilde{x}_{t_1},\tilde{x}_{t_2}} \sigma_{\tilde{x}_{t_1}} \sigma_{\tilde{x}_{t_2}} = r_{x_{t_1},x_{t_2}} \sigma_{x_{t_1}} \sigma_{x_{t_2}} + r_{c_{t_1},c_{t_2}} \sigma_{c_{t_1}} \sigma_{c_{t_2}} + \epsilon \tag{D.32}$$

where $r_{x_{t_1},x_{t_2}}$ is the error autocorrelation in the original ensemble, $r_{c_{t_1},c_{t_2}}$ the autocorrelation of the corrections, σ_{x_t} and σ_{c_t} the standard deviation of the original ensemble and the standard deviation of the correction at lead time t, respectively. The term ϵ corresponds to the estimated covariances of x and c, and is considered as negligible assuming that the original forecast and the corrections are

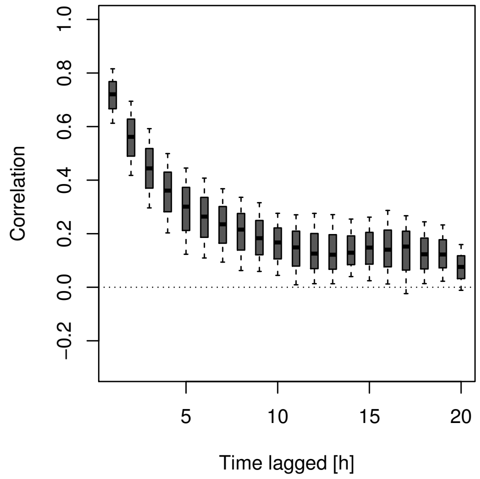

Figure D.5: Temporal lagged correlation coefficients summarizing the error correlation matrix $\hat{\boldsymbol{R}}_e$ used in the d-ECC approach. The boxplots indicate the variability within the 3 month calibration period.

drawn from two independent random processes.

Furthermore, the stationarity assumption of d-ECC implies that the correlation $r_{\tilde{x}_{t_1},\tilde{x}_{t_2}}$ can also be estimated from past error statistics:

$$r_{\tilde{x}_{t_1},\tilde{x}_{t_2}} = \mathbb{E}[\hat{r}_{e_{t_1},e_{t_2}}] \tag{D.33}$$

where the notation $\hat{r}_{e_{t_1},e_{t_2}}$ refers to the elements of the estimated correlation matrix $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$. The stationarity assumption takes effect in the transformation step of d-ECC (Eq. D.22) which modifies the correlation of the scenario corrections $r_{c_{t_1},c_{t_2}}$ and pushes it towards the estimated correlation $\hat{r}_{e_{t_1},e_{t_2}}$. In other words, the transformation affects $r_{c_{t_1},c_{t_2}}\sigma_{c_{t_1}}\sigma_{c_{t_2}}$ (second term in Eq. D.32). We expect d-ECC to have a relevant impact if $r_{c_{t_1},c_{t_2}}\sigma_{c_{t_1}}\sigma_{c_{t_2}}$ dominates the sum in Eq. (D.32). Typically, this is the case when the spread σ_{x_t} of the original ensemble is small compared to the spread $\sigma_{\tilde{x}_t}$ after calibration. In a previous statement, we already noted that d-ECC takes effect if the calibration *corrects* the spread. Regarding Eq. (D.32), we can refine the statement and argue that d-ECC especially takes effect if the calibration *increases* the spread.

Another important aspect of d-ECC is the estimation of the correlation matrix $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$. By means of this matrix, the assumption of error autocorrelation is checked and adjusted. The matrix is estimated from the training datasets used for calibration at the different lead times. Based on the dataset described in Section D.2, Figure D.5 shows the lagged correlation of the forecast error derived from $\hat{\boldsymbol{R}}_{\boldsymbol{e}}$. The correlation is decreasing as a function of the time lag, reaching near zero values for lags greater than 10 hours. However, for short and very short time lags, the correlation is high and stable over the rolling training datasets. In particular, focusing on a time lag of 1 hour, the correlation ranges between 60% and 80%. The correlation variability shown in Figure D.5 is estimated over a 3 month period. Similar results are obtained when checking the variability of the correlation within each training dataset (not shown). The exhibited low variability indicates that the temporal correlation of the forecast error is not flow dependent. As a consequence, d-ECC can be seen as a "universal" approach that does not suffer restriction related to the forecasted weather situation.

Considering again our case study, the scenarios generated with d-ECC based on the COSMO-DE-EPS forecasts are shown in Figure D.3(c). The d-ECC derived scenarios are smoother and subjectively more realistic than the ones derived with ECC in Figure D.3(b). In Figure D.4, focusing on a single scenario, it is highlighted that the difference between the original and the d-ECC time trajectories varies gradually from one time interval to the next one while abrupt transitions occur in the case of the ECC scenario, as in this example between hours 15 and 17.

Note that d-ECC does not give the same result as a simple smoothing of the calibrated scenarios $\tilde{\boldsymbol{x}}$. Smoothing in time would modify the values $\boldsymbol{q}$ of the calibrated ensemble and possibly deteriorate its reliability. Instead, d-ECC affects the time variability of the scenarios by constructing a template (Eq. D.7) based on $\breve{\boldsymbol{x}}$ (Eq. D.24) while preserving the calibrated values $\boldsymbol{q}$.

The discussion and illustration of d-ECC could certainly be extended by idealized studies and a rigorous mathematical framework. This would be welcomed as further research and would add further evidence to the expected behaviour of d-ECC.

D.5 Verification methods

Multivariate scores

Verification of scenarios is first performed assessing the multivariate aspect of the forecast by means of adequate scores. The scores are applied focusing on scenarios in the form of time series. Considering an ensemble with M scenarios $\boldsymbol{x}^{(n)}$ with $n \in \{1, ..., M\}$ and an observed scenario $\boldsymbol{y}$, the energy score (ES; Gneiting *et al.*, 2008) is defined as:

$$ES = \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{y} - \boldsymbol{x}^{(m)}\| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{p=1}^{M} \|\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(p)}\| \tag{D.34}$$

where $\|.\|$ represents the Euclidean norm. ES is a generalization of the $CRPS$ to the multivariate case.

ES suffers from a lack of sensitivity to misrepresentation of correlation structures (Pinson and Tastu, 2013). We consider therefore additionally the p-variogram score (pVS; Scheuerer and Hamill, 2015), which has better discriminative property in this respect. Based on the geostatistical concept of variogram, pVS is defined as:

$$pVS = \sum_{i \neq j} \omega_{ij} \left(| \, y_i - y_j \, |^p - \frac{1}{M} \sum_{m=1}^{M} | \, x_i^{(m)} - x_j^{(m)} \, |^p \right)^2 \tag{D.35}$$

with p the order of the variogram and where ω_{ij} are weights and the indices i and j indicate the i-th and the j-th components of the marked vectors, respectively. In order to focus on rapid changes in wind speed, the weights ω_{ij} are chosen proportional to the inverse square distance in time such:

$$\omega_{ij} = \frac{1}{(i-j)^2}, \quad i \neq j, \tag{D.36}$$

since i and j are here forecast lead time indices.

Multivariate rank histograms

The multivariate aspect of the forecast is in a second step assessed by means of rank histograms applied to multi-dimensional fields (Thorarinsdottir *et al.*, 2014). Two variants of the multivariate rank histogram are applied: the averaged rank histogram (ARH) and the band depth rank histogram ($BDRH$). The difference of the two approaches lies in the way to defined pre-ranks from multivariate forecasts. ARH considers the averaged rank over the multivariate aspect while $BDRH$ assesses the centrality of the observation within the ensemble based on the concept of functional band depth.

The interpretation of ARH is the same as the interpretation of a univariate rank histogram: ∪-shaped, ∩-shaped, and flat rank histograms are interpreted as underdispersiveness, overdispersiveness, and calibration of the underlying ensemble forecasts, respectively. The interpretation of $BDRH$ is different: a ∪-shape is associated to a lack of correlation, a ∩-shape to a too high correlation in the ensemble, a skewed rank histogram to bias or dispersion errors and a flat rank histogram to calibrated forecasts.

Product oriented verification

Besides multivariate verification of time series scenarios, the forecasts are assessed in a product oriented framework. This type of scenario verification follows the spirit of the event oriented verification framework proposed by Pinson and Girard (2012). Probabilistic forecasts that require time trajectories are provided and assessed by means of well-established univariate probabilistic scores.

Two types of products derived from forecasted scenarios are here under focus. The first one is defined as the mean wind speed over a day (here, a day is limited to the 21 hour forecast horizon). The second product is defined as the maximal upward wind ramp over a day, a wind ramp being defined as the difference between two consecutive forecast intervals. For both products, 20 forecasts are derived from the 20 scenarios at each station and each verification day.

The performances of the ensemble forecasts for the two types of products are evaluated by means of the $CRPS$. The $CRPS$ is the generalization of the mean absolute error to predictive distributions (Gneiting *et al.*, 2008), and can be seen as the integral of the Brier score (BS; Brier, 1950) over all thresholds or the integral of the quantile score (QS; Koenker and Bassett, 1978) over all probability levels. Considering an ensemble forecast, the $CRPS$ can be calculated as a weighted sum of QS applied to the sorted ensemble members (Bröcker, 2012). For a deeper insight in the forecast performance in terms of attributes, the $CRPS$ is decomposed following the same approach (Ben Bouallègue, 2015): the $CRPS$ reliability and resolution components are calculated as weighted sums of the reliability and resolution components of the QS at the probability levels defined by the ensemble size (see Eq. D.3), respectively. Formally, we write:

$$CRPS_{\text{reliability}} = \frac{2}{M} \sum_{m=1}^{M} QS_{\tau_m}^{\text{reliability}} \tag{D.37}$$

$$CRPS_{\text{resolution}} = \frac{2}{M} \sum_{m=1}^{M} QS_{\tau_m}^{\text{resolution}} \tag{D.38}$$

where $QS_{\tau_m}^{\text{reliability}}$ and $QS_{\tau_m}^{\text{resolution}}$ are the reliability and resolution components of the QS applied to the quantile forecasts at probability level τ_n, respectively. The QS decomposition is performed following Bentzien and Friederichs (2014). The $CRPS_{reliability}$ is negatively oriented (the lower the better) while the $CRPS_{resolution}$ is positively oriented (the higher the better).

Bootstrapping

The statistical significance of the results are tested applying a block-bootstrap approach. Bootstrapping is a resampling technique which provides an estimation of the statistical consistency and is

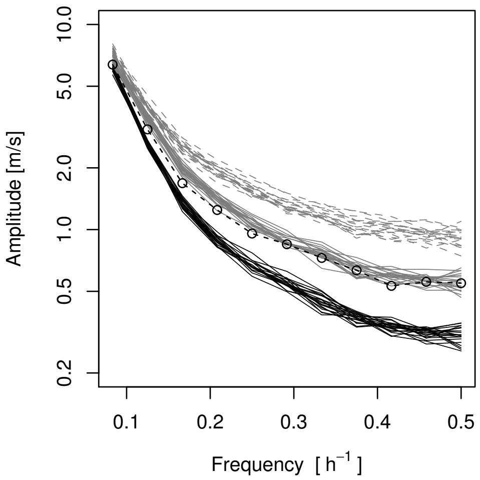

Figure D.6: Spectral analysis of the scenarios from the raw ensemble (black lines), of the scenarios derived with ECC (dashed grey lines) and with d-ECC (grey lines). Each line corresponds to one scenario among the 20. The spectrum of the observed time series is represented by the dashed dotted line.

commonly applied to meteorological datasets (Efron and Tibshirani, 1986).

A block-bootstrap approach is applied in the following which consists in defining a block as a single day of the verification period (Hamill, 1999). Each day is considered as a separate block of fully independent data. The verification process is repeated 500 times using each time a random sample with replacement of the 92 verification days (March, April, May, 2013). The derived score distributions illustrate consequently the variability of the performance measures over the verification period and not between locations. Boxplots are used to represent the distributions of the performance measures, where the quantile of the distributions at probability levels 5%, 25%, 50%, 75 % and 95% are highlighted.

D.6 Results and discussion

Before applying the verification methods introduced in the previous section, we propose to explore statistically the time series variability by means of a spectral analysis, an analysis of the time series in the frequency domain. Such an analysis is useful in order to describe statistical properties of the scenarios but has also direct implications for user's applications (see below; Vincent *et al.*, 2010). A Fourier transformation is applied to each forecasted and observed scenario and the contributions of the oscillations at various frequencies to the scenario variance examined (Wilks, 2006b). In Figure D.6, the mean amplitude of the forecast and observation time series over all stations and verification days is plotted as a function of their frequency components.

As already suggested by the case study, this analysis confirms that the ECC considerably increases the variability of the time trajectories with respect to the original ensemble, in particular at high frequencies. ECC scenario fluctuations are also much larger than the observed ones. Indeed, the amplitude is on average about two times larger at high frequencies in ECC time series than in the observed ones which explains the visual impression that ECC scenarios are unrealistic. Conversely, scenarios derived with the new copula approach do not exhibit such features. While the original ensemble shows a deficit of variability with respect to the observations, the d-ECC approach allows improving this aspect of the forecast. This first result, showing that d-ECC scenarios have a similar mean spectrum as the observation one, is complemented with an objective assessment of the forecasted scenarios based on probabilistic verification measures.

Figure D.7 shows the performance of the forecasted time trajectories by means of multivariate scores. The post-processed scenarios perform significantly better than the raw members in terms of ES (Figure D.7(a)). In terms of pVS, the d-ECC scenarios are better than the ECC ones and significantly better than the raw ones when $p = 0.5$ (Figure D.7(b)). For higher orders of the variogram (here $p = 1$, Figure D.7(c)), the forecast improvement after post-processing is still clear when using d-ECC while the ECC results are slightly worse than the ones of the original forecasts.

Figure D.8 depicts the results in terms of multivariate rank histograms, ARH (upper panel) and $BDRH$ (lower panel). The raw ensemble shows clear reliability deficiencies (Figures D.8(a) and D.8(d)) which motivated the use of post-processing techniques. Forecasts derived with ECC show still underdispersiveness but also too little correlation (Figures D.8(b) and D.8(e)) while forecasts derived with d-ECC are better calibrated according to the rank histograms in Figures D.8(c) and D.8(f). Indeed, both plots indicate good reliability of the d-ECC derived scenarios.

Figure D.9 focuses on two products drawn from the time series forecasts: the daily mean wind speed (upper panel) and the daily maximal upward ramp (lower panel). The performances are assessed in terms of $CRPS$, $CRPS$ reliability and $CRPS$ resolution, from left to right, respectively. Looking at the results in terms of $CRPS$, we note the high similarity of Figures D.9(a) and D.9(d) with Figures D.7(a) and D.7(c), respectively. As for the ES, post-processing significantly improves the forecasts of the daily mean product. As for pVS with $p = 1$, d-ECC improves the ramp product with respect to the original one while ECC does not generate improved products. The $CRPS$ decomposition allows detailing the origin of these performances. We see in Figures D.9(b) and D.9(e) that the $CRPS$ results are mainly explained by the impact of the post-processing on the $CRPS$ reliability components. However, focusing on the results in terms of $CRPS$ resolution in Figures D.9(c) and D.9(f), we note that the resolution of the original and d-ECC products are comparable while ECC deteriorates the resolution of the ramp product with respect to the original one.

Those verification results are interpreted as follows. Calibration corrects for the mean of the ensemble forecast and this is reflected, after the derivation of scenarios, by an improvement of the ES and daily mean product skill. Calibration also corrects for spread deficiencies increasing the variability of the ensemble forecasts. This increase of spread associated with a preservation of the rank structure of the original ensemble, as it is the case in the ECC approach, enlarges indiscriminately the temporal variability of the forecasts and leads to a slight deterioration of the pVS and ramp product results.

The d-ECC approach provides scenarios with a temporal variability comparable to the one of the observation. In that case, the benefit of the calibration step in terms of reliability (at single forecast lead times) persists at the multivariate level (looking at time trajectories) after the reconstruction of scenarios with d-ECC. The multivariate reliability, or the reliability of derived products, is significantly improved after post-processing, though not perfect for specific derived products. Moreover, d-ECC scenarios perform as well as the original ensemble forecast in terms of resolution. So, unlike ECC, d-ECC is able to generate reliable scenarios with a level of resolution that is not deteriorated with respect to the original ensemble forecasts.

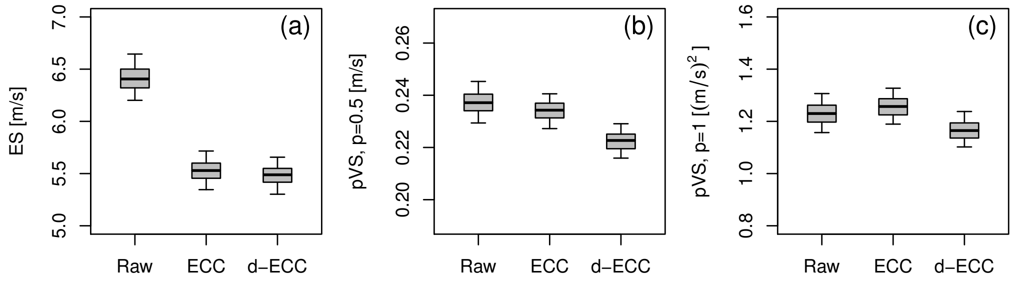

Figure D.7: Multivariate scores of time series: energy score (a) and p-variogram score for $p = 0.5$ (b) and $p = 1$ (c) in the form of box plots drawn from the application of a 500-block bootstrapping.

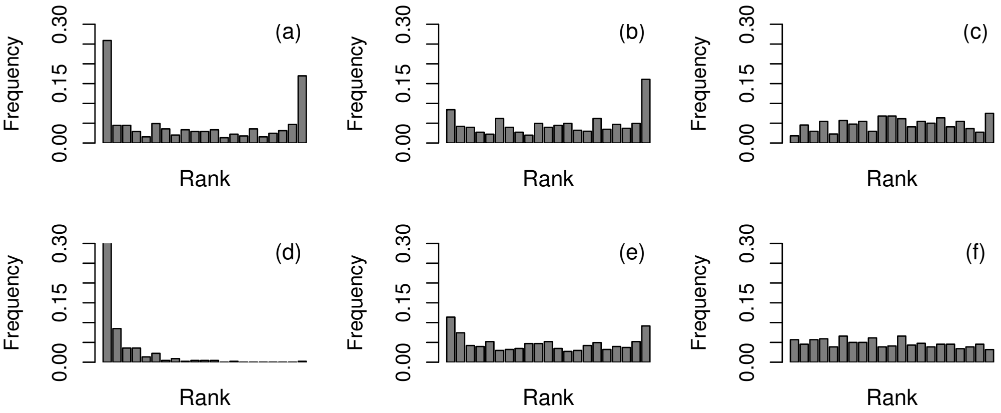

Figure D.8: Multivariate rank histograms: (a,b,c) average rank histograms and (d,e,f) band depth rank histograms for time series from the raw ensemble (a,d) and derived with ECC (b,e) and d-ECC (c,f).

D.7 Conclusion and outlook

A new empirical copula approach is proposed for the post-processing of calibrated ensemble forecasts. The so-called dual ensemble copula coupling approach is introduced with a focus on temporal structures of wind forecasts. The new scheme includes a temporal component in the ECC approach accounting for the error autocorrelation of the ensemble members. The estimation of the correlation structure in the error based on past data allows adjusting the dependence structure in the original ensemble.

Based on COSMO-DE-EPS forecasts, the scenarios derived by d-ECC prove to be qualitatively realistic and quantitatively of superior quality. Post-processing of wind speed combining EMOS and d-ECC improves the forecasts in many aspects. In comparison to ECC, d-ECC drastically improves the quality of the derived scenarios. Applications that require temporal trajectories will fully benefit of the new approach in that case. As for any post-processing technique, the benefit of the new copula approach can be weakened by improving the representation of the forecast uncertainty with more efficient member generation techniques and/or by improving the calibration procedure correcting for conditional biases. Meanwhile, at low additional complexity and computational costs, d-ECC can be considered as a valuable alternative to the standard ECC for the generation of consistent scenarios.

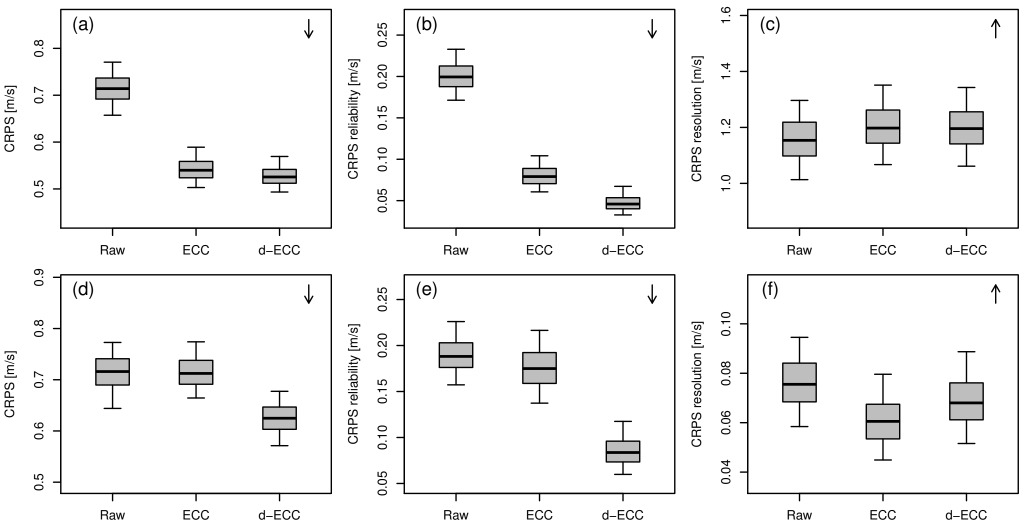

Figure D.9: Product oriented verification of scenarios: (a,b,c) daily means at station, (d,e,f) maximal upward ramps within a day at station. Results are shown in terms of $CRPS$ (a,d), $CRPS$ reliability component (b,e) and $CRPS$ resolution component (c,f). The box plots indicate confidence intervals estimated with block bootstrapping. The arrows in the right corners indicate whether the performance measure is positively or negatively oriented.

Though only the temporal aspect has been investigated in this study, the dual ensemble copula approach could be generalized to any multivariate setting. Further research is however required for the application of d-ECC at scales that are unresolved by the observations. For example, geostatistical tools could be applied for the description of the autocorrelation error structure at the model grid level. Moreover, the mathematical interpretation of the d-ECC scheme developed here would benefit from further theoretical investigations based on idealized case studies.

Acknowledgments

This work has been done within the framework of the EWeLiNE project (*Erstellung innovativer Wetter- und Leistungsprognosemodelle für die Netzintegration wetterabhängiger Energieträger*) funded by the German Federal Ministry for Economic Affairs and Energy. The authors acknowledge the Department of Wind Energy of the Technical University of Denmark (DTU), the German Wind Energy Institute (DEWI GmbH), DNV GL, the Meteorological Institute (MI) of University of Hamburg and the Karlsruhe Institute of Technology (KIT) for providing wind measurements at stations Risoe, FINO1 and FINO3, FINO2, Hamburg and Lindenberg, and Karlsruhe, respectively. The authors are also grateful to Tilmann Gneiting and two anonymous reviewers for helpful and accurate comments on a previous version of this manuscript.

Abbreviations

ARH	averaged rank histogram
AUC	area under the curve
BDRH	band depth rank histogram
CRPS	continuous ranked probability score
CRPSS	continuous ranked probability skill score
d-ECC	dual ensemble copula coupling
DWD	Deutscher Wetterdienst
EAV	ensemble added value
ECC	ensemble copula coupling
EPS	ensemble prediction system
FAR	false alarm rate
HR	hit rate
MAE	mean absolute error
NWP	numerical weather prediction
PQR	penalized quantile regression
PV	photovoltaic
QR	quantile regression
QS	quantile score
QSS	quantile skill score
ROC	relative operating characteristic
RUC	relative user characteristic
UTC	coordinated universal time

Main notations and mathematical symbols

α	cost-loss ratio
β	regression coefficient
λ_r	regularization parameter
π	base rate of an event
ρ	asymmetric loss function
σ	standard deviation of a probability distribution
ψ	threshold used to define an event
C	cost of a preventive action against a potential event
$\bar{E}$	mean expense of a user
$F(X)$	probability distribution function of the random variable X
L	loss when no protective action is taken and an event occurs
M	ensemble size
N	verification size
N'	length of the training period
N_s	number of scenarios
$Pr(X \leq x)$	probability of X smaller or equal to x
p_ψ	exceedance probability forecast for threshold ψ
q_τ	quantile forecast at probability level τ
R_e	correlation matrix of the forecast error
r	correlation coefficient
s	scoring rule function
V	forecast value
v	model output used as predictor
X	forecast variable
$x^{(m)}$	ensemble member denoted m
$\tilde{x}^{(m)}$	ensemble member denoted m after post-processing
Y	observation variable
z	reference template

List of Tables

List of Figures

Bibliography

Atger F. 2004. Estimation of the reliability of ensemble-based probabilistic forecasts. *Q.J.R. Meteorol.Soc* **130**: 627–646.

Badescu V. 2008. *Modelling solar radiation at the earth surface: Recent advances*. Springer-Verlag Berlin Heidelberg, 517 pp.

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

Bartels M, Gatzen C, Peek M, Schulz W, Wissen R, Jansen A, Molly J, Neddermann B, Gerch H, Grebe E, Sassnick Y, Winter W. 2006. Planning of the grid integration of wind energy in Germany onshore and offshore up to the year 2020. *International Journal of Global Energy Issues* **25**: 257–275.

Bauer P, Thorpe A, Brunet G. 2015. The quiet revolution of numerical weather prediction. *Nature* **525**: 47–55.

Becker R, Behrens K. 2012. Quality assessment of heterogeneous surface radiation network data. *Adv. Sci. Res.* **8**: 93–97.

Ben Bouallègue Z. 2013. Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting* **28**: 515–524.

Ben Bouallègue Z. 2015. Assessment and added value estimation of an ensemble approach with a focus on global radiation forecasts. *Mausam* **66**: 541–550.

Ben Bouallègue Z, Pinson P, Friederichs P. 2015. Quantile forecast discrimination ability and value. *Q.J.R. Meteorol. Soc.* **141**: 3415–3424.

Ben Bouallègue Z, Theis SE. 2014. Spatial techniques applied to precipitation ensemble forecasts: from verification results to probabilistic products. *Met. Apps.* **21**: 922–929.

Ben Bouallègue Z, Theis SE, Gebhardt C. 2013. Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift* **22**: 49–59.

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

Bentzien S, Friederichs P. 2014. Decomposition and graphical portrayal of the quantile score. *Q.J.R. Meteorol. Soc.* **140**: 1924–1934.

Bird L, Milligan M, Lew D. 2013. Integrating variable renewable energy: Challenges and solutions. *NREL Technical Report* .

Bjerknes V. 1904. Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteorologische Zeitschrift* **21**: 1–7.

Bouttier F. 1994. *Sur la prévision de la qualité des prévisions météorologiques.* Ph.D. thesis, Université Paul Sabatier, Toulouse, France, 240 pp.

Bowler NE, Arribas A, Mylne K, Robertson KB, Beare SE. 2008. The MOGREPS short-range ensemble prediction system. *Q.J.R. Meteorol. Soc.* **134**: 703–722.

Boyle G. 2008. *Renewable electricity and the grid: the challenge of variability*. Earthscan, 219 pp.

Bremnes JB. 2004. Probabilistic wind power forecasts using local quantile regression. *Wind Energ.* **7**: 47–54.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.* **78**: 1–3.

Bröcker J. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Q.J.R. Meteorol. Soc.* **35**: 1512–1519.

Bröcker J. 2010. Regularized logistic models for probabilistic forecasting and diagnostics. *Mon. Wea. Rev.* **138**: 592–604.

Bröcker J. 2011. Probability forecasts. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB (eds), John Wiley and Sons, pp. 119–140.

Bröcker J. 2012. Evaluating raw ensembles with the continuous ranked probability score. *Q.J.R. Meteorol. Soc.* **138**: 161–1617.

Bröcker J. 2014. Resolution and discrimination - two sides of the same coin. *Q.J.R. Meteorol. Soc.* .

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

Bröcker J, Smith LA. 2007. Scoring probabilistic forecasts: the importance of being proper. *Wea. Forecasting* **22**: 382–88.

Buizza R. 2001. Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Mon. Wea. Rev.* **129**: 2329–2345.

Candille G, Talagrand O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Q.J.R. Meteorol. Soc.* **131**: 2131–2150.

Chen YS, Ehrendorfer M, Murphy AH. 1987. On the relationship between the quality and value of forecasts in the generalized cost-loss ratio situation. *Mon. Wea. Rev.* **115**: 1534–1541.

Chow CW, Urquhart B, Kleissl J, Lave M, Dominguez A, Shields J, Washom B. 2011. Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed. *Solar Energy* **85**: 2881–2893.

Christoffersen PF, Diebold FX. 1997. Optimal prediction under asymmetric loss. *Econometric Theory* **13**: 808–817.

Clark M, Gangopadhyay S, Hay L, Rajagopalan B, 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.

Costa A, Crespo A, Navarro J, Lizcano G, Madsen H, Feitosa E. 2008. A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews* **12**: 1725–1744.

De Groot M, Fienberg S. 1982. Assessing probability assessors: calibration and refinement. *Statistical Decision Theory and Related Topics* **1**: 291–314.

Delle Monache L, Eckel FA, Rife DL, Nagarajan B, Searight K. 2013. Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.* **141**: 3498–3516.

DelSole T. 2004. Predictability and information theory. Part I: Measures of predictability. *J. Atmos. Sci.* **61**: 2425–2440.

Diebold FX, Tay TGA. 1998. Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.* **39**: 863–883.

Efron B, Tibshirani R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* **1**: 54–75.

Ehrendorfer M. 1997. Predicting the uncertainty of numerical weather forecasts: a review. *Meteorol. Zeit.* **6**: 147–183.

Epstein E. 1969a. Stochastic dynamic prediction. *Tellus* **21**: 739–759.

Epstein ES. 1969b. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.* **8**: 985–987.

Espinar B, Aznarte J, Girard R, Moussa A, Kariniotakis G. 2010. Photovoltaic forecasting: a state of the art. *Proceedings 5th European PV-Hybrid and Mini-Grid Conference* : 250–255.

Feldmann K, Scheuerer M, Thorarinsdottir T. 2015. Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.* **143**: 955–971.

Flowerdew J. 2014. Calibrating ensemble reliability whilst preserving spatial structure. *Tellus A* **66**.

Friederichs P, Hense A. 2007. Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.* **135**: 2365–2378.

Friedrich RJ. 1982. In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science* **26**: 797–833.

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

Glahn H, Lowry DA. 1972. The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.* **11**: 1203–1211.

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

Gneiting T. 2011b. Quantiles as optimal point forecasts. *International Journal of Forecasting* **27**: 197–207.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration, and sharpness. *J. Roy. Stat. Soc.* **69B**: 243–268.

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

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

Gneiting T, Ranjan R. 2011. Comparing density forecasts using threshold and quantile weighted proper scoring rules. *Journal of Business and Economic Statistics* **29**: 411–422.

Gneiting T, Stanberry L, Grimit E, Held L, Johnson N. 2008. Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test* **17**: 211–235.

Grimit EP, Mass CF. 2007. Measuring the ensemble spread-error relationship with a probabilistic approach: stochastic ensemble results. *Mon. Wea. Rev.* **135**: 203–221.

Gross R, Heptonstall P, Leach M, Skea J, Anderson D, Green T. 2008. The UK Energy Research Center Review of the costs and impacts of intermitency. In: *Renewable electricity and the grid: the challenge of variability*, Boyle G (ed), Earthscan, pp. 73–94.

Hagedorn R, Hamill TM, S J, Whitaker. 2007. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part I: 2-meter temperature. *Mon. Wea. Rev.* **136**: 2608–2619.

Hagedorn R, Lundgren K, Dobshinski J, Focken U. 2015. Die Energiewende in Deutschland - wie kann der Deutsche Wetterdienst den neuen Herausforederungen begegnen? *promet* **39**: 242–244.

Hamill TM. 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather Forecasting* **14**: 155–167.

Hamill TM. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.* **129**: 550–560.

Hamill TM, Hagedorn R, Whitaker JS. 2007. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.* **136**: 2620–2632.

Hamill TM, Juras J. 2006. Measuring forecast skill: is it real or is it the varying climatology? *Q.J.R. Meteorol. Soc.* **132**: 2905–2923.

Hamill TM, Whitaker JS. 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rew.* **134**: 3209–3229.

Hamill TM, Whitaker JS, Wei X. 2004. Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.* **132**: 1434–1447.

Hammer A, Heinemann D, Hoyer C, Kuhlemann R, Lorenz E, Müller R, Beyer HG. 2003. Solar energy assessment using remote sensing technologies. *Remote Sensing of Environment* **86**: 423–432.

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

Hoffman RN, Kalnay E. 1983. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus A* **35**: 100–118.

Hopson TM. 2014. Assessing the ensemble spread-error relationship. *Mon. Wea. Rev.* **142**: 1125–1142.

Houtekamer PL, Lefaivre L, Derome J, Ritchie H, Mitchell HL. 1996. A system simulation approach to ensemble prediction. *Mon. Wea. Rev.* **124**: 1225–1242.

Hyndman RJ, Fan Y. 1996. Sample quantiles in statistical packages. *American Statistician* **50**: 361–365.

Jolliffe IT, Stephenson DB. 2005. Comments on "Discussion of verification voncepts in Forecast verification: A practitioner's guide in atmospheric science". *Wea. Forecasting* **20**: 796–800.

Jolliffe IT, Stephenson DB. 2011. *Forecast verification: A practitioner's guide in atmospheric science, 2nd edition.* John Wiley and Sons, Chichester, UK, 292 pp.

Junk C, Delle Monache L, Alessandrini S. 2015. Analog-based ensemble model output statistics. *Mon. Wea. Rev.* **143**: 2909–2917.

Junk C, vonBremen L, Kühn M, Späth S, Heinemann D. 2014. Comparison of postprocessing methods for the calibration of 100-m wind ensemble forecasts at off- and onshore sites. *J. Appl. Meteor. Climatol.* **53**: 950–969.

Kalogirou S. 2001. Artificial neural networks in renewable energy systems applications: a review. *Renewable and Sustainable Energy Reviews* **5**: 373–401.

Katz RW, Murphy AH. 1997. Forecast value: prototype decision-making models. In: *Economic Value of Weather and Climate Forecasts*, Katz RW, Murphy AH (eds), Cambridge University Press, pp. 183–217.

Kessy A, Lewin A, Strimmer K. 2015. Optimal whitening and decorrelation. *arXiv:1512.00809* .

Keune J, Ohlwein C, Hense A. 2014. Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts. *Mon. Wea. Rev.* **142**: 4074–4090.

Koenker R, Bassett G. 1978. Regression quantiles. *Econometrica* **46**: 33–50.

Koenker R, Machado J. 1999. Goodness of fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.* **94**: 1296–1310.

Koenker RW, d'Orey V. 1994. Computing regression quantiles. *Applied Statistics* **43**: 410–414.

Köhler C, Steiner A, Saint Drenan YM, UENBs, Ben Bouallègue Z, Metzinger I, Ritter B. 2016. Critical weather situations for renewable energies - Part B: low stratus risk for solar power. *Renewable Energy,* **submitted**.

Krzysztofowicz R. 1983. Why should a forecaster and a decision maker use Bayes theorem. *Water Resour. Res.* **19**: 327–336.

Leith CE. 1974. Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.* **102**: 409–418.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *Journal of Computational Physics* **227**: 3515–3539.

Lewis JM. 2005. Roots of ensemble forecasting. *Mon. Wea. Rev.* **133**: 1865–1885.

Lewis JM. 2014. Edward Epstein's stochastic-dynamic approach to ensemble weather prediction. *Bulletin American Meteorological Society* **95**: 99–116.

Lorenz E, Kühnert J, Heinemann D. 2014. Overview of irradiance and photovoltaic power prediction. *In: Weather Matters for Energy, Editors: A. Troccoli, L. Dubus, S.E. Haupt (Springer 2014).* .

Lorenz E, Scheidsteger T, Hurka J, Heinemann D, Kurz C. 2011. Regional PV power prediction for improved grid integration. *Progress in Photovoltaics* **19**: 757–771.

Lorenz EN. 1969. The predictability of a flow which possesses many scales of motion. *Tellus* **21**: 289–307.

Marquez R, Coimbra C. 2011. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. *Solar Energy* **85**: 746–756.

Mason I. 1982. A model for assessment of weather forecasts. *Aust. Met. Mag.* **30**: 291–303.

Mason SJ. 2004. On using "Climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.* **132**: 1891–1895.

Mason SJ, Graham NE. 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. *Q. J. R. Meteorol. Soc.* **128**: 2145–2166.

Mason SJ, Weigel AP. 2009. A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.* **137**: 331–349.

Matheson JE, Winkler RL. 1976. Scoring rules for continuous probability distributions. *Manage. Sci.* **22**: 1087–1096.

Mellit A. 2008. Artificial intelligence technique for modelling and forecasting of solar radiation data: a review. *International Journal of Artificial Intelligence and Soft Computing* **1**: 52–76.

Messner JW, Mayr GJ, Wilks DS, Zeileis A. 2014a. Extending extended logistic regression: extended versus separate versus ordered versus censored. *Mon. Wea. Rev.* **142**: 3003–3014.

Messner JW, Mayr GJ, Zeileis A, Wilks DS. 2014b. Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.* **142**: 448–456.

Mikosch T. 2006. Copulas: tales and facts. *Extremes* **9**: 3–20.

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The ECMWF ensemble prediction system: methodology and validation. *Q.J.R. Meteorol. Soc.* **122**: 73–119.

Montani A, Cesari D, Marsigli C, Paccagnella T. 2011. Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus A* **63**: 605–624.

Morales J, Conejo A, Madsen H, Pinson P, Zugno M. 2014. *Integrating renewables in electricity markets. operational problems*. International Series in Operations Research & Management Science, Vol. 205.

Murphy AH. 1969. Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which knowledge of the cost-loss ratios is incomplete. *J. Appl. Meteor.* **8**: 863–873.

Murphy AH. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* **12**: 595–600.

Murphy AH. 1991. Forecast verification: its complexity and dimensionality. *Mon. Wea. Rev.* **119**: 1590–1601.

Murphy AH. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting* **8**: 281–293.

Murphy AH, Epstein ES. 1967. A note on probability forecasts and "Hedging". *J. Appl. Meteor.* **6**: 1002–1004.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Mon. Wea. Rev.* **115**: 1330–1338.

Murphy AH, Winkler RL. 1992. Diagnostic verification of probability forecasts. *Int. J. Forecasters* **7**: 435–455.

Palmer TN. 2000. Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.* **63**: 71–116.

Palmer TN, Döring A, Seregin G. 2014. The real butterfly effect. *Nonlinearity* **27**: R123–R141.

Pelland S, Remund J, Kleissl J, Oozeki T, Brabandere KD. 2013. Photovoltaic and solar forecasting: state of the art. *Report IEA-PVPS T14-01* .

Peralta C, Ben Bouallègue Z, Theis SE, Gebhardt C. 2012. Accounting for initial condition uncertainties in COSMO-DE-EPS. *Journal of Geophysical Research* **117**.

Perez R, Kivalov S, Schlemmer J, Jr KH, Renné D, Hoff T. 2010. Validation of short and medium term operational solar radiation forecasts in the US. *Solar Energy* **84**: 2161 –2172.

Perez R, Lorenz E, Pelland S, Beauharnois M, et al. 2013. Comparison of numerical weather prediction solar irradiance forecasts in the US, canada and europe. *Solar Energy* **94**: 305–326.

Pinson P. 2012. Adaptive calibration of (u,v)-wind ensemble forecasts. *Quart. J. Roy. Meteor. Soc.* **138**: 1273–1284.

Pinson P. 2013. Wind energy: forecasting challenges for its operational management. *Statistical Science* **28**: 564–585.

Pinson P, Chevallier C, Kariniotakis G. 2007. Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Trans. on Power Systems* **22**: 1148–1156.

Pinson P, Girard R. 2012. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy* **96**: 12–20.

Pinson P, Papaefthymiou G, Klockl B, Nielsen H, Madsen H. 2009. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energ.* **12**: 51–62.

Pinson P, Tastu J. 2013. Discrimination ability of the energy score. *Technical report, Technical University of Denmark.* .

R Core Team. 2013. R: a language and environment for statistical computing, http://www.r-project.org .

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.* **133**: 1155–1174.

Raynaud L, Pannekoucke O, Arbogast P, Bouttier F. 2015. Application of a Bayesian weighting for short-range lagged ensemble forecasting at the convective scale. *Q.J.R. Meteorol. Soc.* **141**.

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

Richardson DS. 2011. Economic value and skill. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB (eds), John Wiley and Sons, pp. 167–184.

Ritter B, Geleyn JF. 1992. A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. *Mon. Wea. Rev.* **120**: 303–325.

Roulston MS, Kaplan DT, Hardenberg J, Smith LA. 2003. Using medium-range weather forecasts to improve the value of wind power production. *Renewable Energy* **28**: 585–602.

Saint-Drenan YM, Bofinger S, Fritz R, Vogt S, Good GH, Dobschinski J. 2015. An empirical approach to parameterizing photovoltaic plants for power forecasting and simulation. *Solar Energy* **120**: 479–493.

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

Scheuerer M, Hamill TM. 2015. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Mon. Wea. Rev.* **143**: 1321–1334.

Schölzel C, Friederichs P. 2008. Multivariate non-normally distributed random variables in climate research - introduction to the copula approach. *Nonlin. Processes Geophys.* **15**: 761–772.

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

Schuhen N, Thorarinsdottir T, Gneiting T. 2012. Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.* **140**: 3204–3219.

Schwartz CS, Kain JS, Weiss SJ, Xue M, Bright DR, Kong F, Thomas KW, Levit JJ, Coniglio MC, Wandishin MS. 2010. Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting* **25**: 263–280.

Siegert S, Bröcker J, Kantz H. 2011. Predicting outliers in ensemble forecasts. *Q.J.R. Meteorol. Soc.* **137**: 1887–1897.

Sklar M. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**: 229–231.

Slingo J, Palmer TN. 2011. Uncertainty in weather and climate prediction. *Phil. Trans. R. Soc. A* **369**: 4751–4767.

Sloughter J, Gneiting T, Raftery AE. 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.* **105**: 25–35.

Steppeler J, Doms G, Schättler U, Bitzer HW, Gassmann A, Damrath U, Gregoric G. 2003. Mesogamma scale forecasts using the nonhydrostatic model LM. *Met. Atm. Phys.* **82**: 75–96.

Theis SE, Hense A, Damrath U. 2005. Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Met. Apps* **12**: 257–268.

Thompson J. 1962. Economic gains from scientific advances and operational improvements in meteorological prediction. *J. Appl. Meteor.* **1**: 13–17.

Thorarinsdottir T, Gneiting T. 2010. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Statist. Soc. Ser. A* **173**: 371–388.

Thorarinsdottir T, Scheuerer M, Heinz C. 2014. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics,* **in press**.

Thorey J, Mallet V, Chaussin C, Descamps L, Blanc P. 2015. Ensemble forecast of solar radiation using TIGGE weather forecasts and HelioClim database. *Solar Energy* **120**: 232–243.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B.* **58**: 267–288.

Tödter J, Ahrens B. 2012. Generalization of the ignorance score: Continuous ranked version and its decomposition. *Mon. Wea. Rev.* **140**: 2005–2017.

Toth Z, Talagrand O, Candille G, Zhu Y. 2003. Probability and ensemble forecasts. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.*, Jolliffe IT, Stephenson DB (eds), John Wiley and Sons, Chichester, UK, pp. 137–164.

Tracton MS, Kalnay E. 1993. Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting* **8**: 379–398.

Troccoli A, Dubus L, Haupt SE. 2014. *Weather matters for energy*. Springer, 528pp.

Vincent C, Giebel G, Pinson P, Madsen H. 2010. Resolving nonstationary spectral information in wind speed time series using the hilbert-huang transform. *J. Appl. Meteor. Climatol.* **49**: 253–267.

Wahl S. 2015. *Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation*. Bonner Meteorologische Abhandlung Heft 71, 108pp.

Weigel AP. 2011. Ensemble verification. *In: Forecast verification: A practitioner's guide in atmospheric science, 2nd Edition* .

Wilks DS. 2001. A skill score based on economic value for probability forecasts. *Met. Apps* **8**: 209–2019.

Wilks DS. 2006a. Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Met. Apps* **13**: 243–256.

Wilks DS. 2006b. *Statistical methods in the atmospheric sciences*. 2nd Edn. Academic Press, New York, 627pp.

Wilks DS. 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteo. Appl.* **16**: 361–368.

Wilks DS. 2014. Multivariate ensemble model output statistics using empirical copulas. *Quart. J. Roy. Meteor. Soc.* **141**: 945–952.

Wilks DS, Hamill TM. 2007. Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.* **135**: 2379–2390.

Wirth H. 2015. Recent facts about photovoltaics in Germany. *Fraunhofer ISE Report* .

Zamo M, Mestre O, Arbogast P, Pannekoucke O. 2014a. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part I: Deterministic forecast of hourly production. *Solar Energy* **105**: 792–803.

Zamo M, Mestre O, Arbogast P, Pannekoucke O. 2014b. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. *Solar Energy* **105**: 804–816.

Zhu Y. 2005. Ensemble forecast: a new approach to uncertainty and predictability. *Adv. Atmos. Sci.* **22**: 781–788.

Zhu Y, Toth Z, Wobus R, Richardson D, Mylne K. 2002. The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.* **83**: 73–83.

Zou H, Hastie T, Tibshirani R. 2007. On the "degrees of freedom" of the lasso. *Ann. Stat.* **35**: 2173–2192.

Acknowledgments

I would like to thank Andreas Hense, Petra Friederichs and Pierre Pinson for supervising my research and sharing with me their expertise. I thank Patrik Ferrari who kindly accepted to be co-examiner of my thesis. I am very grateful to Susanne Theis and Sabrina Wahl for their valuable advice. I would like also to express my gratitude to Renate Hagedorn and Michael Buchhold for their support when I decided to combine my work at DWD with a PhD at the University of Bonn; to Christoph Gebhardt and Felix Fundel for reading and commenting a previous version of this manuscript; to the "Energy group" and the FE15 colleagues at DWD for a fruitful collaboration; to Tobias Hepelmann, Michael Scheuerer, Richard Keane, Vanessa Stauch and Fayçal Ben Bouallègue, among others, for exciting scientific discussions during the past 3 years; and finally, to my parents, family, and friends for their unstinting support.

BONNER METEOROLOGISCHE ABHANDLUNGEN

Herausgegeben vom Meteorologischen Institut der Universität Bonn durch Prof. Dr. H. FLOHN (Hefte 1-25), Prof. Dr. M. HANTEL (Hefte 26-35), Prof. Dr. H.-D. SCHILLING (Hefte 36-39), Prof. Dr. H. KRAUS (Hefte 40-49), ab Heft 50 durch Prof. Dr. A. HENSE.

Heft 1-59: siehe `http://www.meteo.uni-bonn.de/bibliothek/bma`

Heft 60: ***Christoph Gebhardt***: Variational reconstruction of Quaternary temperature fields using mixture models as botanical – climatological transfer functions. 2003, 204 S. + VIII. € 30

Heft 61: ***Heiko Paeth***: The climate of tropical and northern Africa – A statistical-dynamical analysis of the key factors in climate variability and the role of human activity in future climate change. 2005, 316 S. + XVI. € 15

Heft 62: ***Christian Schölzel***: Palaeoenvironmental transfer functions in a Bayesian framework with application to Holocene climate variability in the Near East. 2006, 104 S. + VI. € 15

Heft 63: ***Susanne Bachner***: Daily precipitation characteristics simulated by a regional climate model, including their sensitivity to model physics, 2008, 161 S. € 15

Heft 64: ***Michael Weniger***: Stochastic parameterization: a rigorous approach to stochastic three-dimensional primitive equations, 2014, 148 S. + XV. open access[1]

Heft 65: ***Andreas Röpnack***: Bayesian model verification: predictability of convective conditions based on EPS forecasts and observations, 2014, 152 S. + VI. open access[1]

Heft 66: ***Thorsten Simon***: Statistical and Dynamical Downscaling of Numerical Climate Simulations: Enhancement and Evaluation for East Asia, 2014, 48 S. + VII. + Anhänge open access[1]

Heft 67: ***Elham Rahmani***: The Effect of Climate Change on Wheat in Iran, 2014, [erschienen] 2015, 96 S. + XIII. open access[1]

Heft 68: ***Pablo A. Saavedra Garfias***: Retrieval of Cloud and Rainwater from Ground-Based Passive Microwave Observations with the Multi-frequency Dual-polarized Radiometer ADMIRARI, 2014, [erschienen] 2015, 168 S. + XIII. open access[1]

Heft 69: ***Christoph Bollmeyer***: A high-resolution regional reanalysis for Europe and Germany - Creation and Verification with a special focus on the moisture budget, 2015, 103 S. + IX. open access[1]

Heft 70: ***A S M Mostaquimur Rahman***: Influence of Subsurface Hydrodynamics on the Lower Atmosphere at the Catchment Scale, 2015, 98 S. + XVI. open access[1]

[1] Available at `http://hss.ulb.uni-bonn.de/fakultaet/math-nat/`

Heft 71: ***Sabrina Wahl***: Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation, 2015, 108 S. open access[1]

Heft 72: ***Markus Übel***: Simulation of mesoscale patterns and diurnal variations of atmospheric CO_2 mixing ratios with the model system TerrSysMP-CO_2, 2015, [erschienen] 2016, 158 S. + II open access[1]

Heft 73: ***Christian Bernardus Maria Weijenborg***: Characteristics of Potential Vorticity anomalies associated with mesoscale extremes in the extratropical troposphere, 2015, [erschienen] 2016, 151 S. + XI open access[1]

Heft 74: ***Muhammad Kaleem***: A sensitivity study of decadal climate prediction to aerosol variability using ECHAM6-HAM (GCM), 2016, 98 S. + XII open access[1]

Heft 75: ***Theresa Bick***: 3D Radar reflectivity assimilation with an ensemble Kalman filter on the convective scale, 2016, [erschienen] 2017, 96 S. + IX open access[1]

Heft 76: ***Zied Ben Bouallegue***: Verification and post-processing of ensemble weather forecasts for renewable energy applications, 2017, 119 S. open access[1]

METEOROLOGISCHES INSTITUT
MATHEMATISCH NATURWISSENSCHAFTLICHE FAKULTÄT
UNIVERSITÄT BONN

Meteorologisches Institut
Universität Bonn