Information Theoretic Learning applied to Wind Power Modeling

Ricardo J. Bessa, V. Miranda, *Fellow, IEEE,* Jose C. Principe, *Fellow, IEEE,*
A. Botterud, *Member, IEEE*, and J. Wang, *Member, IEEE*

***Abstract*— This paper reports new results in adopting information theoretic learning concepts in the training of neural networks to perform wind power forecasts. The forecast "goodness" is discussed under two paradigms: one is only concerned in measuring the deviation between the forecasted and realized values, the other is related with the value of the forecast in the electricity market for different agents. The results and conclusions are supported by a real case example.**

I. Introduction

A large scale is a template A large-scale integration of wind power poses a number of challenges for electricity markets and power system operators who will have to deal with the variability and uncertainty in wind power generation when making their scheduling and dispatch decisions. Wind power forecasting (WPF) is recognized as an important tool to address the variability and uncertainty in wind power and to efficiently operate power systems with large wind power penetration [1].

Moreover, in a market environment, the wind power contribution to the generation portfolio becomes important in determining the daily and hourly prices, as variations in the estimated wind power will influence the clearing prices for both energy and operating reserves [2].

With the increasing penetration of wind power, WPF is quickly becoming an important topic for the electric power industry in the United States and Europe. System operators (SO), wind generating companies (WGENCO) and regulators all support efforts to develop better, more reliable and accurate forecasting models. Wind farm owners and operators also benefit from better wind power prediction to support competitive participation in electricity markets against more stable and dispatchable energy sources. In general, WPF can be used for a number of purposes, such as: generation and transmission maintenance planning, determination of operating reserve requirements [3], unit commitment [4], economic dispatch, energy storage optimization (e.g., pumped hydro storage [5]), and electricity market trading [6].

This paper presents practical results supporting two ideas; a) criteria based on information theoretic learning (ITL) – entropy and correntropy of the prediction error distribution – are more suitable than the traditional Minimum Square Error (MSE) criterion to train wind power prediction models under a *forecaster paradigm*, and b) the forecast error and the consequences of adopting several criteria for model training should be analyzed carefully by considering the different (and probable conflicting) objectives of the forecast users.

A brief overview of the state-of-the-art in WPF is given in section II, and also the motivation to use training functions more suitable for non-Gaussian error distributions. Section III presents aspects of the training of neural networks with ITL criteria for WPF. The *forecaster paradigm* is presented in section IV and results for a real test case are also presented. The *forecast user paradigm* is described in section V and results are presented for the Iberian (Portugal + Spain) electricity market. Section VI presents the final conclusions.

II. A Brief Overview of the State-of-the-art

This section provides a brief overview of techniques and concepts on the topic of WPF. For a more detailed review of the current state-of-the-art in WPF readers should refer to a recent report by Argonne National Laboratory [7].

A. Short-term Wind Power Forecasting

Wind Power Forecasting (WPF) consists in forecasting (at time instant t for a look-ahead time t+k) the average wind generation the wind farm is expected to generate during the considered period of time (e.g., 1 hr). Forecasts are made for a time horizon, indicating the total length of the forecast period (e.g., 72 hr in the future). Generally, this is called point forecast (or spot forecast) because it is only a single value. Other types of forecasts are currently being made, e.g. probability distributions forecasted for every look-ahead time.

Two classes of approaches can be found in the literature: statistical and physical. The fundamental idea of the latter is to refine Numerical Weather Predictions (NWP) through physical considerations about the site, e.g surface roughness or orography, and by modeling the profile of the local wind possibly accounting for atmospheric stability.

The statistical approach is based on one or more models

Manuscript received February 21, 2010. The author Ricardo J. Bessa acknowledges Fundação para a Ciência e a Tecnologia (FCT) for PhD Scholarship SFRH/BD/33738/2009.

Ricardo J. Bessa and V. Miranda are with INESC Porto - Instituto de Engenharia de Sistemas e Computadores do Porto, and with the Faculty of Engineering of the University of Porto, Portugal (emails: rbessa@inescporto.pt; vmiranda@inescporto.pt).

Jose C. Principe is with the Computational NeuroEngineering Laboratory, Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: principe@cnel.ufl.edu).

A. Botterud and J. Wang are with Argonne National Laboratory, USA (emails: abotterud@anl.gov, jianhui.wang@anl.gov).

that establish a relationship between historical values of generation and forecasted weather variables. These models can be divided in two groups: models that only employ time series data and predict future values taking into account the past history [8]; and models that use, in addition to the mean electric power time series data, forecasted values from a NWP model corresponding mainly to hourly mean wind speed and direction [9]. The published results obtained with these models reveal an important improvement with respect to the results obtained with the models in the first group, but only when the forecast horizon is beyond a few hours (around 6 hrs).

In this paper only the second group that uses NWP as input will be studied. This extrapolation of NWP to power will be referred to in this document as a "wind to power" (W2P) model.

B. *Machine Learning Techniques Applied to WPF*

The W2P model can include one or several statistical linear and nonlinear models of different types, which include most of the machine learning based models.

Fugon et al [10] presented a survey on the performance of different data-mining models in WPF. Two versions of linear regression were examined: one is a simple regression model used as reference, and the other consists of combining the input variables to create extra variables. The analyzed nonlinear models were Neural Networks (NN), Support Vector Machines (SVM), regression trees with bagging and random forests for regression. Jursa [11] compares different techniques, such as a classical MLP NN, mixture of experts, SVM and nearest neighbor search with a Particle Swarm Optimization (PSO) algorithm for feature selection. Kusiak et al. [12] tested five data-mining models to produce forecasts for very short-term horizons (1 to 12 hr ahead) and short/long-term horizons (3 to 84 hr ahead): SVM, MLP NN, RBF NN, regression trees, and random forests. Barbounis et al [13] employed locally recurrent neural networks to forecast wind power of three types: (i) the infinite impulse response multilayer perceptron; (ii) the local activation feedback multilayer network; (iii) and the diagonal recurrent neural network. Negnevitsky et al. [14] addressed the combined use of neural networks and fuzzy logic in WPF. This is a hybrid approach called Adaptive Neural Fuzzy System (ANFIS). Sideratos et al [9] described a combination of the RBF NN with the fuzzy logic model in order to optimize the use of the NWP predictions. Fan et al. [15] describe and test a two-stage hybrid method with Bayesian Clustering by Dynamics (BCD) and SVM.

C. *Non-Gaussian Forecast Errors of WPF*

From the literature it is clear that, it is possible to understand that, one way or another, models depend on a training process and usually adopt the Minimum Square Error (MSE) as a training criterion. The applicability of MSE to train a mapper (any model mapping an input-output relation, e.g. neural network (NN), with parameters to be learned) is only optimal if the probability distribution function of the prediction errors is Gaussian.

It is well known that the wind speed vs. power curve of a wind turbine is highly nonlinear. The transformation of wind speed into wind power changes the statistical properties of the errors. This result has been shown, for instance, in [16] for six sites in Germany, where error distributions from wind power prediction models were skewed right and had a positive excess of kurtosis.

The presence of non-Gaussian distributions has motivated research for techniques that would train wind-to-power (W2P) models based on minimizing the information content of the error distribution instead of minimizing its variance (MSE). A measure of information content is entropy and incorporating entropy as a cornerstone concept in the training of mappers has been the object of Information Theoretic Learning (ITL) [17][18].

In two recent papers [19] [20] devoted to WPF the authors have engaged in evaluating the performance of neural networks, trained in offline and online mode, comparing the MSE criterion with three ITL inspired criteria.

III. Training W2P Models with ITL Criteria

A. *Entropy and Parzen pdf Estimation*

Renyi's entropy [21] of a discrete probability distribution $P = (p_1, p_2, \ldots, p_n)$ is defined as

$$H_{R\alpha} = \frac{1}{1-\alpha}\log\sum_{k=1}^{N} p_k^{\alpha} \quad with \quad \alpha > 0, \alpha \neq 1 \qquad (1)$$

Renyi's entropy is a family of functions $H_{R\alpha}$ depending on a real parameter α. When α = 2, we have what is called quadratic entropy

$$H_{R2} = -\log\sum_{k=1}^{N} p_k^2 \qquad (2)$$

This definition can be generalized for a continuous random variable Y with pdf $f_Y(z)$:

$$H_{R2} = -\log\int_{-\infty}^{+\infty} f_Y(z)^2\, dz \qquad (3)$$

The estimation of the pdf of data from a sample constituted by discrete points $y_i \in R^M$, i=1,…,N in a M-dimensional space may be done by the Parzen Window method [22].

This technique uses a kernel function centered on each point; it looks at a point as being locally described by a probability density Dirac function, which is replaced or approximated by a continuous set whose density is represented by the kernel.

If a Gaussian kernel is used, the expression of the estimation $\hat{f}_Y$ for the real pdf f_Y of a set of N points is a summation of individual contributions:

$$\hat{f}_Y(z) = \frac{1}{N}\sum_{i=1}^{N} G\left(z - y_i, \sigma^2 I\right) \qquad (4)$$

where G(...) is the Gaussian kernel and $\sigma^2 I$ is the covariance matrix (here assumed with independent and equal variances in all dimensions). In each dimension *k* we have

$$G\left(z_k - y_{ik}, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(z_k - y_{ik}\right)^2} \qquad (5)$$

It is easy to understand that the "size" of the window, here defined by the value of σ, is important in obtaining a smoother or more "spiky" estimate for f_Y.

B. Information Theoretic Learning

A breakthrough has been achieved by signal processing researchers when they proposed combining Renyi's entropy definition with an estimate of a pdf by the Parzen window method [17] [18] [23] [24] – this has been called Information Theoretic Learning. Several mapper training criteria have been studied.

1) Minimum Error Entropy (MEE)

An entropy estimator for a discrete set of data points {y} in one dimension (*k=1*) is

$$H_{R2}(y) = -\log \int_{-\infty}^{+\infty} \hat{f}(z)^2 dz = -\log V(y) \qquad (6)$$

where, using (4),

$$V(y) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{-\infty}^{+\infty} G\left(z - y_i, \sigma^2\right) G\left(z - y_j, \sigma^2\right) dz \qquad (7)$$

In this expression we recognize the convolution of Gaussian functions. We have thus the following result:

$$V(y) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G\left(y_i - y_j, 2\sigma^2\right) \qquad (8)$$

which allows the practical evaluation of entropy by simply calculating the Gaussian function values of the vector distances between pairs of samples. V(y) is called the information potential (IP) of the data set.

As the objective in supervised training is to minimize H_{R2} of the errors *e (Output-Target)*, one can instead maximize the information potential V(e). So, max V(e) becomes the training criterion for optimizing a mapper with minimum output entropy [23] – the MEE criterion, for Minimum Entropy Error.

2) Maximum Correntropy Criterion (MCC)

Correntropy [25] is a generalized similarity measure between two arbitrary scalar random variables X and Y defined by:

$$J_\sigma(X, Y) = E\left[k_\sigma(X - Y)\right] \qquad (9)$$

where k_σ is the kernel function (usually Gaussian).

In real problems, the joint *pdf* is unknown and we only have available a finite number of points. The estimator is:

$$J_\sigma(X, Y) = \frac{1}{N} \sum_{i=1}^{N} k_\sigma\left(x_i - y_i\right) \qquad (10)$$

Correntropy is directly related to the probability of how similar two random variables are in a neighborhood of the joint space defined by the kernel bandwidth, and provides the probability density of the event p(X=Y). Using Parzen windows, the bandwidth controls the observation window in which the similarity is assessed but makes one unable to assess similarity in the whole joint space.

In [25] one may find a discussion on the properties of correntropy. It is proved that for the Gaussian kernel CIM(X,Y) = k(0) - V(X,Y) related with correntropy satisfies all the properties of a metric. CIM (Correntropy Induced Metric) may be divided in three different regions: when the error is close to zero CIM is equivalent to L2 norm (Euclidian norm); when the error grows CIM becomes a L1 norm (sum of the absolute differences of the points coordinates); when the error is very large CIM becomes a L0 norm, the metric saturates and becomes very insensitive to large errors.

This property shows the robustness of CIM and the importance of kernel bandwidth. A small kernel size leads to a small Euclidean zone while a large kernel size will increase the Euclidean region where the metric behaves like the MSE criterion.

We can use the MCC as a new performance function, with the advantage over MSE of being a local criterion of similarity and very useful for cases with non-zero mean, non-Gaussian, with large outliers. It does not require the computing effort of MEE but also includes knowledge of the error statistics (attempts to to maximize the pdf value at the origin).

The MCC training criterion becomes

$$\max_w J(e) = \frac{1}{N} \sum_{i=1}^{N} G\left(g(w, x_i) - T_i, \sigma\right) \qquad (11)$$

where $g(w,x_i)$ represents the mapper producing $y_i = g(w,x_i)$ responses from input x_i as a function of weights w, and T_i represents the target values.

3) Minimum Error Entropy with Fiducial Points (MEEF)

MEE does not constrain the mean value of the error, as is normally required in function approximation. There are methods to deal with the problem.

The first method is to correct the MEE result by properly modifying the output bias of the neural network to yield zero mean error over the training data set just after training ends. The other way is to add a so-called MCC term to the MEE training criterion, leading to the Minimum Error Entropy with Fiducial Points (MEEF) criterion.

This criterion [26] intends to anchor the error distribution to a zero mean by defining a compromise between minimizing entropy and maximizing correntropy through a training criterion

$$\begin{aligned} MEEF(e) \Leftrightarrow \max\ & \gamma \cdot \frac{1}{N} \sum_{i=1}^{N} G\left(e_i, \sigma^2\right) + \\ & + (1-\gamma) \cdot \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} G\left(e_j - e_i, 2\sigma^2\right) \end{aligned} \qquad (12)$$

where γ is a weighting constant between 0 and 1.

4) Maximum Parametric Correntropy Criterion (MPCC)

An alternative to the correntropy function described in (9) is the parametric correntropy described in [27]. The idea consists of comparing X and Y along the line $aX+b=y$ where a and b are parameters. The Maximum Parametric Correntropy Criterion (MPCC) is given by

$$MPCC(e,a) \Leftrightarrow \max \frac{1}{N}\sum_{i=1}^{N} G(a \cdot y_i + b - T_i, \sigma) \quad (13)$$

where a and b could be parameters defined by the forecaster or estimated from the data.

This function can be useful in associating economic meaning to the training criterion, such as the forecast error penalization costs in the electricity market (problem addressed in section V). This allows producing forecasts with a bias which gives a different value for the error sign.

IV. The Forecaster Paradigm

A. What is a Good Forecast?

There is an important discussion in the forecasting community, and particularly in WPF, about what a "good" forecast is. The evaluation of the forecasts according to a *forecaster paradigm* means that the only concern is to try to generate a prediction that matches the observed values. This paradigm is what is generally employed by forecasters both in the academic and operational environment. For instance, in [28] a WPF evaluation protocol in agreement with this paradigm is described, where measures like the mean absolute error and bias are used to evaluate the forecasts *quality*.

B. Training and Testing Characteristics

The W2P model used in the following discussion is a feed forward MLP neural network with only one hidden layer comprising 7 neurons, using a hyperbolic tangent activation function. The inputs are NWP meteorological forecasted values: forecasted mean wind speed, forecasted wind direction and an index m corresponding to the number of past half hours of NWP forecasted values. Due to the cyclic characteristic of wind direction and the variable m (daily hour), these variables are represented by sine and cosine components. This means a total of 5 input variables, which were standardized using the min-max method. Finding the best neural network topology and input variables is out of the scope of this paper.

An online training described in [20] was used for training with the following criteria: MSE, MEEF and MCC. The training algorithm was the backpropagation and more details about the training process can be found in the abovementioned publication. The kernel size σ used by MCC is the same used in previous publications [19] and [20].

The neural network was used to predict the power $p_{t+k|t}$ produced by the wind farm at time stamp t (when new NWP predictions become available) for each look-ahead $t+k$ of the next day. The wind power prediction was performed for each day of the test data set.

To validate the results we adopted a Monte Carlo procedure running 25 simulations with distinct networks, generating randomly initial weights in each case by a uniform distribution in the interval [-1,1]. The final conclusions derive from the average of all simulations. When comparing criteria, the same weights were used in all cases.

C. Forecasting Results

This section presents results for a power forecasting problem for a wind farm in Europe, comparing the performance of neural networks resulting from the adoption of several criteria: the traditional MSE and the set of two ITL criteria (MEEF and MCC). The training set is composed of data from ten months of one year, the validation set consists in two months and the test dataset is one year of data.

Fig. 1 compares the *pdf* of errors obtained with MEEF, MCC and MSE. The *pdf* of errors is more centered at zero error when the NN is trained with the two ITL criteria.

The results obtained with MSE confirm that the prediction errors are not Gaussian: it was possible to obtain narrower *pdf* with entropy and correntropy criteria than with a variance-based criterion. In fact, if the error *pdf* were Gaussian, the MSE criterion would perform as well as an entropy-based criterion, but this was not the case.

Therefore, in agreement with the theory, it was possible to design a mapper that would produce a predictor with a higher peak close to zero, a characteristic associated with smaller entropy of the *pdf*. The MCC criterion presents higher probability density near zero than MEEF, but we have larger errors in the negative part of the *pdf*.

The Normalized Mean Absolute Error (NMAE) is adopted to evaluate performance over the whole time horizon (from 7:00 AM to 0:00 AM of the third day) in the test set months. The rationale for this choice is two-fold: it is a criterion often used by researchers working in WPF [28] and it would not introduce a bias in comparisons (one is adopting a criterion not used in the calculations).

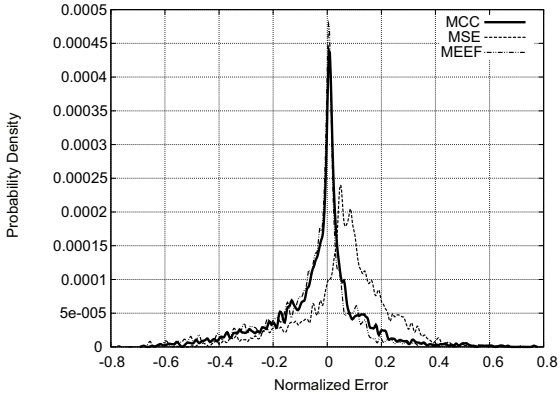

Fig. 1. Comparison of the error *pdf* generated by NN trained with MCC, MSE and MEEF criteria

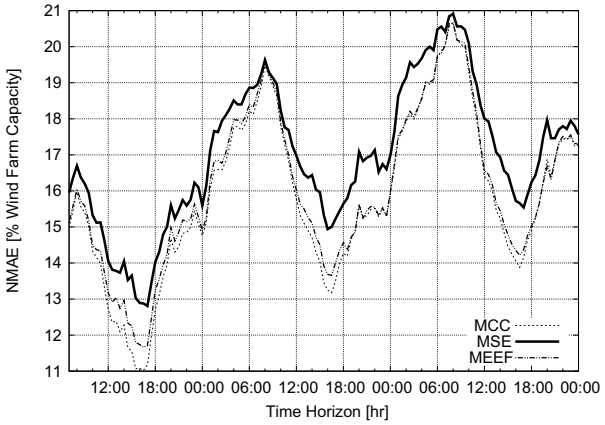

Fig. 2. NMAE errors of MCC, MSE and MEEF

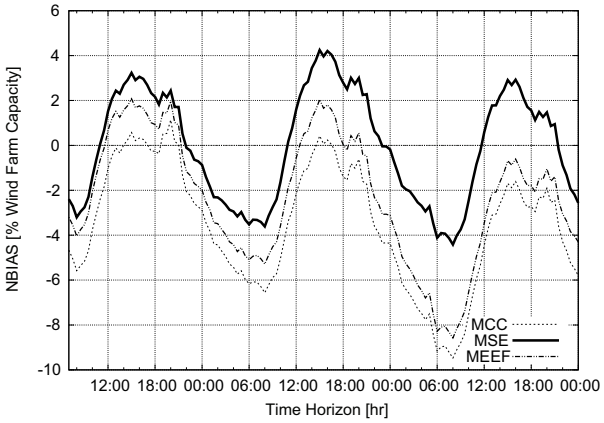

Fig. 3. NBIAS errors of MCC, MSE and MEEF

Fig. 3 depicts the Normalized Bias (NBIAS) for the MCC, MEEF and MSE models for each hour.

Fig. 2 depicts the NMAE errors for the MCC, MEEF and MSE models for each hour through the forecast horizons, for lead times up to 72 hours.

The figure make visually evident that the MCC and MEEF criteria presents better results (in terms of the "forecaster paradigm") than MSE for almost every lead time of the forecast horizon. No significant differences exist between MEEF and MCC.

The average difference between the two ITL criteria (MEEF and MCC) and MSE is 1.48%. The average difference between MEEF and MCC is favorable to MEEF in 0.04 %.

Fig. 3 depicts the Normalized Bias (NBIAS) for the MCC, MEEF and MSE models for each hour.

The information provided by the NBIAS is related with the systematic errors and identifies if the model is providing under or over predictions. Despite being the same data, each training criterion yields totally different results in terms of bias, with the ITL criteria providing systematically underforecasts. As a general trend, a system trained with MSE tends to produce forecasts that slightly overestimate the actual values while the Entropy criteria lead on average to lower values.

V. The Forecasters Consumer Paradigm

A. What is a Good Forecast?

The *forecaster paradigm* presented in section IV is mostly concerned with minimizing the forecast error and no importance is given to the nature and consequences of errors.

The definition of a *good forecast* is strongly connected with the use of the forecast, and an appealing idea is that a forecast should be oriented by user interests. Hence, the nature of the forecast error may be such that no particular forecast is optimal for multiple applications such as electricity market participation and power system operation.

In weather forecast research there are several publications that discuss what a good or bad forecast is. In [29], three different types of *goodness* in forecasts are defined: type 1) correspondence between forecasts and forecaster judgment (*consistency*); type 2) correspondence between forecasts and observations (*quality*); type 3) incremental benefits (economic/or other) when employed by users as an input into their decision-making processes (*value*). The *consistency* in WPF is assumed to be included in the model during the development phase where the forecaster, based on his judgment, builds and adjusts the model.

To assess the WPF *quality*, the standard approach is based on statistical error measures (e.g. mean absolute error) as presented in [28] (and in accordance with the paradigm presented in section IV). However, this approach does not guarantee that the forecast with better score is the one with better *quality*.

An alternative approach, called *verification framework*, is described in [30]. The approach is based on the joint distribution of forecasts and observations and factorization of this distribution into conditional and marginal distributions. The third type of goodness, forecast *value*, is often very difficult to quantify, since in some cases it is related to non-economic factors. However some decision-makers may give a high relevance to this criterion.

With respect to the WPF *value*, some research reported the incremental economic benefit from the use of WPF in the participation of wind power in electricity markets (see for instance [31]).

One fundamental idea is that the relation between *quality* and *value* is nonlinear and differs from problem to problem, as well from user to user. For instance, in [31] a point forecast with 40.55% of imbalances (w.r.t. produced energy) achieves an income of 1145.7 10^3 € over one year, and an advanced strategy based on probabilistic forecast obtained 55.46% of imbalances and an income of 1212.6 10^3 €. This supports the idea that a forecast with higher economic *value* is not necessarily the one with lower forecast error (in accordance with the *forecaster paradigm*).

Hence, it is important to recognize that there are different users of WPF, the most important groups being wind generation companies (WGENCO) that sell in the market and system operators (SO) that must ensure the security of

the system. These user groups may have conflicting interests. For instance, SO are interested in maximizing security and minimizing operational costs while WGENCO are only interested in maximizing their profits in the market. In some conditions, the *good* forecasts are not of greater value to all users.

From the WGENCO viewpoint, the *value* of WPF in the electrical market is measured by the increase in the market revenue for such forecasts over the revenue obtained with an alternative method.

The viewpoint of a SO must be analyzed from the perspective of energy offered in the electricity market by the agents and of maintaining a desired reliability standard. This is related to the management of generation and consumption variability (generation must equal consumption in each time instant), and to reduce the risk of not satisfying load due to forecast errors. Since the load and generation are met by the market mechanism, and assuming that WGENCO participate with bids in the market, the strategic behavior followed by WGENCO may influence significantly the needs for regulation and also the regulation costs. More details about this problem can be found in [3].

Therefore, from the SO viewpoint, a good forecast is the one that leads to: i) less regulation, in particular upward regulation (overforecast errors); ii) less expected maximum forecast error; iii) deviations in favor of the system deviation. Please note that in this case the *value* is not linked with how the SO should make the forecasts, instead it is related to the impact on the operational policy from the economic oriented market bids provided by the WGENCO.

B. Electricity Market

1) Remuneration Model

In an electricity market where WGENCO present selling bids, forecast errors (supplying an amount of energy different from the market bid) have a cost or penalty associated with them.

The market remuneration mechanism adopted in this paper is based on the formulation described in [31]. This formulation is generally valid for most European Markets (e.g. Iberian Market) but in the case of United States there are usually no penalization costs due to forecast errors [2].

In this mechanism, wind power producers offer energy bids (E^b) at day D (typically till noon) for every hour of day $D+1$ in the day-ahead market and are paid at the market clearing price (p^s).

Then, if the wind generation results above or below their offer, they will be subject to a penalty:

- if the actual generation (E^p) is above the market bid, the excess is paid at a discounted price ($p_{surplus}$);
- if the actual generation remains below the market bid, a penalty is applied according to the price ($p_{shortage}$) of the generation the SO has to purchase to compensate for the lack of generation.

These deviation costs are very much dependent on the rules of the market and no general conclusion can be offered. However, one can clearly state that if the penalties are asymmetric then an opportunity arises for a gambler to benefit from using a biased forecast instead of a neutral forecast.

The income (I) of a wind farm from selling for a given look-ahead time $t+k$ can be formulated as:

$$I_{t+k} = \begin{cases} p^s_{t+k} \cdot E^p_{t+k} + c^{down\_reg}_{t+k} \cdot \left(E^p_{t+k} - E^b_{t+k}\right), E^p_{t+k} > E^b_{t+k} \\ p^s_{t+k} \cdot E^p_{t+k} - c^{up\_reg}_{t+k} \cdot \left(E^b_{t+k} - E^p_{t+k}\right), E^p_{t+k} < E^b_{t+k} \end{cases} \quad (14)$$

where c_{down_reg} and c_{up_reg} are regulation costs given by p_s-$p_{surplus}$ and $c_{up\_reg} = p_{shortage}$-$p_s$.

2) Iberian (Portugal + Spain) Market

Data from the Iberian electrical market was used in the following study. The data consists of hourly spot and deviation prices from two years, 2007 and 2008, and was collected from http://www.esios.ree.es.

Table II shows the annual mean energy prices in 2007 and 2008 for the Spanish market (in €/MWh). The regulation prices are very asymmetric. The down-regulation is more expensive (it is better to overforecast), which is the general pattern, according to data used by the authors of [31][32].

TABLE I
ANNUAL MEAN MARKET PRICES

	2007	**2008**
Spot Price (p^s)	39.34	64.47
Up Reg. (c_{up_ref})	2.97	4.04
Down Reg. (c_{down_reg})	7.99	8.02

C. Training and Testing Characteristics

In this section one considers MSE, MCC, MAE (Mean Absolute Error) and MPCC as alternative training criterion, as well as the Minimum Penalization Costs (MPC) proposed by the author of [32]:

$$\min \sum_{J^{down}} w_{down\_reg} \cdot \left|\hat{P}_{t+k} - P_{t+k}\right| + \sum_{J^{up}} w_{up\_reg} \cdot \left|\hat{P}_{t+k} - P_{t+k}\right| \quad (15)$$

which is divided in overestimation J^{up} ($\hat{P}_{t+k} > P_{t+k}$) and underestimation J^{down} ($\hat{P}_{t+k} < P_{t+k}$); w_{down_reg} and w_{up_reg} are penalization factors related with the market regulation costs.

These penalization factors are the average regulation costs (c_{down_reg} and c_{up_reg}) over each hour of the year 2007. Regarding the parameters a and b of the MPCC criterion, b is set to zero since the bias of the forecast can be introduced only with a without increasing the complexity of the problem. For training with MPCC, the methodology adopted was:

First, in the beginning of each epoch estimate the parameter a using the available historical data and hourly averages of the market prices of one year as penalization factors of under- and overestimation. The training criterion is

$$\max_a \begin{cases} k_{\sigma'}\left(\left(a \cdot y_i - T_i\right) \cdot w_{down\_reg}\right), y_i < T_i \\ k_{\sigma'}\left(\left(a \cdot y_i - T_i\right) \cdot w_{up\_reg}\right), y_i > T_i \end{cases} \quad (16)$$

where w_{down_reg} and w_{up_reg} are penalization factors equal to the hourly mean of the regulation costs, and σ' is the kernel

bandwidth. This bandwidth depends on the size of the penalization factors.

Second, when the parameter *a* is known, the following training criterion is maximized

$$\max_{w} k_{\sigma}\left(a' \cdot y_i - T_i\right) \quad (17)$$

where σ is a kernel bandwidth much smaller than σ' and a' is equal to 1/a.

The kernel size σ in MPCC was found in the validation set to be equal to 0.1 and σ' equal to 3.

Applying these two training criteria will introduce a bias towards the side with more economic interest, which can increase the forecast error, but also increases the revenue from the market.

Then, the WGENCO participation of a wind farm in Europe in the electricity market in the year 2008 has been simulated, offering the power prediction coming from NN trained with different training criteria and computing the income with the hourly prices time series.

D. Forecasting and Electricity Market Results

Tables II summarizes the results obtained from the participation in the Spanish market with bids equal to the forecasts provided by each training criterion.

The deviation against the system are the percentage of forecast error that helped the system balance; the total deviations are the forecast error divided by the annual produced energy; $Q_{95\%}$ is the quantile 95% of the overestimation errors distribution, which means that the probability of having an overforecast deviation larger or equal to this value is 5%; CTE is the conditional tail expectation of this quantile, and can be seen similar to the expected maximum forecast deviation; the total income was obtained from simulating the participation of the wind farm over one year with hourly prices.

TABLE II
SIMULATION OF THE PARTICIPATION IN THE IBERIAN MARKET

	MSE	MCC	MAE	MPC	MPCC
Surplus [% of produced energy]	33.79	35.55	39.08	19.71	30.89
Shortage [% of produced energy]	33.74	29.64	24.37	58.07	35.56
Total deviations [% of produced energy]	67.54	65.21	63.47	77.79	66.46
Deviations against the system [% of produced energy]	30.09	28.68	28.21	34.97	29.02
$Q_{95\%}$ *[% of rated power]*	41.74	41.70	43.58	61.50	48.30
CTE [% of rated power]	53.67	58.05	57.89	73.28	63.26
Total Income (deviation from the best) [k€]	-10.16	-7.61	-12.47	-4.1	0.0

1) WGENCO Viewpoint

The first three training criteria are not economically oriented; therefore they achieved lower revenue when compared with MPC and MPCC. MCC and MAE underestimate the wind generation, but despite this underestimation MCC achieved a higher income when compared with MSE. Comparing the results from MPC with the first three training criteria showed that a worst *quality* in forecast does not mean a worst income for a WGENCO. Comparing MCC with MPCC showed that a small increase in the error means an increase around 7600 € in the revenue. For this wind farm the forecasts provided by the MPCC presents the highest profit, around 4100 € more than MPC.

From these results it became clear that with MPCC is possible to increase the income without increasing significantly the deviations due to forecast errors.

2) System Operator Viewpoint

MCC and MAE give forecasts with lower total deviations and deviations against the system. A good property in MSE, from the SO viewpoint, is the importance given to high errors; therefore the CTE presents the lower value of all training criteria.

An appealing result is that with MPCC it was possible to achieve higher revenue without increasing significantly the total deviations and the deviations against the system. One the other hand, the MPC puts a higher bias towards the side with lower regulation costs (shortage of generation) and this is contrary to the interests of the SO in terms of security of the system. MPCC allows increasing the income without a major increase in the total deviations. The comparison of MPC and MPCC in the quantile 95% and CTE values also shows that MPCC is better for the SO when compared with MPC.

It is important to note that the difference in total income between the training criteria were not significant, however the differences in the deviations are significant. Therefore, the adoption of a training criterion is more important to the SO viewpoint.

VI. CONCLUSION

Looking only from the forecaster paradigm point of view, the ITL criteria presented better results (in terms of higher frequency of errors close to zero, insensitivity to outliers and observance of the prediction to real data) than adopting minimum square error as a training criterion for almost every lead time of the forecast horizon. Therefore, one conclusion from the paper is that ITL criteria cannot be ignored when building robust wind power prediction models and whenever the quality criterion is linked with a best fit of the predictions to the actual values verified. However, when value is considered, things are not so straight.

Although a very simple forecasting system and a decision-making problem were considered in this paper, the issues discussed – what a good forecast is for a WGENCO and SO, and the relation between forecast error and market profit – also occur in situations involving more complex forecasting system and electrical markets.

What are the importance of the results shown in this paper for future forecasting practices and market policies? These results reveal the importance of choosing a training function for a forecasting system. And contradicting some opinions,

the work reported shows that it is not necessarily true that it is impossible to simultaneously achieve a high remuneration in the electricity market and maintain the forecast error at an acceptable level: information theoretic learning criteria may achieve this best-of-two-worlds objective.

ACKNOWLEDGMENT

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (''Argonne''). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up non-exclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

REFERENCES

[1] Mark L. Ahlstrom, L. Jones, R. Zavadil, and W. Grant, "The future of wind forecasting and utility operations," *IEEE Power Energy Magazine - Special Issue: Working with Wind; Integrating Wind into the Power System*, vol. 3(6), pp. 57-64, Nov.–Dec. 2005.

[2] A. Botterud, J. Wang, V. Miranda, and R. Bessa, "Wind Power Forecasting in U.S. Electricity Markets," *The Electricity Journal*, vol. 23(3), pp. 71-82, April 2010.

[3] M.A. Matos and R. J. Bessa, "Operating reserve adequacy evaluation using uncertainties of wind power forecast," in *Proceedings of the IEEE PowerTech Conference*, Bucharest, Romania, June 28–July 2, 2009.

[4] J. Wang, A. Botterud, V. Miranda, C. Monteiro, and G. Sheble, "Impact of wind power forecasting on unit commitment and dispatch," *8th Int. Workshop on Large-Scale Integration of Wind Power into Power Systems*, Bremen, Germany, Oct. 2009.

[5] E.D. Castronuovo and J. A. Peças Lopes, "On the optimization of the daily operation of a wind-hydro power plant," *IEEE Transactions on Power Systems*, vol. 19(3), pp. 1599–1606, 2004.

[6] A. Botterud, J. Wang, R. J. Bessa, and V. Miranda, "Risk management and optimal bidding for a wind power producer," in *Proceedings of the IEEE PES General Meeting 2010*, Minneapolis, Minnesota, USA, 25-29 July 2010.

[7] C. Monteiro, R. J. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, "Wind power forecasting: state-of-the-art 2009," Report ANL/DIS-10-1, Argonne National Laboratory, Nov. 2009. Available: http://www.dis.anl.gov/projects/windpowerforecasting.html

[8] C.W. Potter and M. Negnevistky, "Very short-term wind forecasting for Tasmanian power generation," *IEEE Transactions on Power Systems*, vol. 21(2), pp. 965–972, 2006.

[9] George Sideratos and Nikos D. Hatziargyriou, "An Advanced Statistical Method for Wind Power Forecasting," *IEEE Transactions on Power Systems*, vol. 22(1), Feb. 2007.

[10] Lionel Fugon, Jérémie Juban, and G. Kariniotakis, "Data mining for Wind Power Forecasting," in *Proceedings of the European Wind Energy Conference EWEC'08*, Brussels, Belgium, April 2008.

[11] R. Jursa, "Wind power prediction with different artificial intelligence models," in *Proceedings of the European Wind Energy Conference EWEC'07*, Milan, Italy, May 2007.

[12] A. Kusiak, H.-Y. Zheng, and Z. Song, "Wind Farm Power Prediction: A Data-Mining Approach," *Wind Energy*, Vol. 12(3), pp. 275–293, 2009.

[13] T.G. Barbounis and J.B. Theocharis, "Long-term wind speed and power forecasting using local recurrent neural network models," *IEEE Transactions on Energy Conversion*, vol. 21(1), pp. 273–284, 2006.

[14] M. Negnevitsky, S. Santoso, and N. Hatziargyriou, "Data mining and analysis techniques in wind power system applications: abridged," in *Proceedings of the IEEE Power Engineering Society General Meeting*, 2006.

[15] Shu Fan, James R. Liao, Ryuichi Yokoyama, and Luonan Chen, "Forecasting the Wind Generation Using A Two-stage Hybrid Network Based on Meteorological Information," Information and Communication Engineering, Osaka Sangyo University, 2006.

[16] M. Lange, "On the uncertainty of wind power predictions – Analysis of the forecast accuracy and statistical distribution of errors," *Transactions of the ASME, Journal of Solar Energy Engineering*, vol. 2, no. 127, pp. 177–194, May 2005.

[17] J. C. Principe, D. Xu, J. Fisher, *Information theoretic learning*, in Unsupervised Adaptive Filtering, New York: Simon Haykin Editor, pp. 265-319, 2000.

[18] J. C. Principe (Editor), Information theoretic learning: Renyi's entropy and Kernel perspectives, Springer Verlag, 2010.

[19] R. J. Bessa, V. Miranda and J. Gama, "Wind power forecasting with entropy-based criteria algorithms," in *Proceedings of PMAPS 2008 – 10th International Conference on Probabilistic Methods Applied to Power Systems*, Puerto Rico, May 2008.

[20] R. J. Bessa, V. Miranda and J. Gama, "Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting," *IEEE Transactions on Power Systems*, vol. 24(4), pp. 1657-1666, 2009.

[21] A. Renyi, "Some fundamental questions of information theory," Selected Papers of Alfred Renyi, vol. 2, pp. 526-552, Akademia Kiado, Budapest, 1976.

[22] E. Parzen, "On the estimation of a probability density function an the mode," *Annals Math Statistics*, vol 33, 1962.

[23] D. Erdgomus and J.C. Principe, "An error-entropy minimization algorithm for supervised training of non-linear adaptive systems," *IEEE Transactions on Signal Processing*, vol. 50(10), Jul. 2002.

[24] D. Erdogmus, J.C. Principe, "From linear adaptive filtering to non-linear signal processing," *IEEE Signal Processing Magazine*, vol. 23, pp 14-33, 2006.

[25] W. Liu, P. Pokharel, and J.C. Principe, "Correntropy: properties and applications in non-Gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55(11), pp. 5286-5298, Nov. 2007.

[26] Liu Weifeng, P.P. Pokharel, J.C. Principe, "Error Entropy, Correntropy and M-Estimation", *16th IEEE Signal Processing Society Workshop on Machine learning for Signal Processing*, pp. 179–184, 2006.

[27] Jianwu Xu, "Nonlinear signal processing based on reproducing Kernel Hilbert space," PhD thesis, University of Florida, USA, 2007.

[28] H. Madsen, P. Pinson, G. Kariniotakis, Henrik Aa. Nielsen, and Torben Skov Nielsen, "Standardizing the performance evaluation of short-term wind prediction models," *Wind Engineering*, vol. 29(6), pp. 475-489, Dec. 2005.

[29] A. H. Murphy, "What is a good forecast? an essay on the nature of goodness in weather forecasting," *Weather and Forecasting*, vol. 8(2), pp. 281-293, Jun. 1993.

[30] A. H. Murphy and R. L. Winkler, "A general framework for forecast verification," *Monthly Weather Review*, vol. 115, pp. 1330–1338, 1987.

[31] P. Pinson, C. Chevallier, and G. Kariniotakis, "Trading wind generation with short-term probabilistic forecasts of wind power," *IEEE Transactions on Power Systems*, vol. 22(3), pp. 1148–1156, 2007.

[32] Hans F. Ravn, "Short term wind power prognosis with different success criteria," in *Proceedings of International Conference on Probabilistic Methods Applied to Power Systems - PMAPS 2006*, pp. 1-5, 11-15 June 2006.