Entropy and Correntropy Against Minimum Square Error in Offline and Online Three-Day Ahead Wind Power Forecasting

Ricardo J. Bessa, Vladimiro Miranda, *Fellow, IEEE*, and João Gama

***Abstract*—This paper reports new results in adopting entropy concepts to the training of neural networks to perform wind power prediction as a function of wind characteristics (speed and direction) in wind parks connected to a power grid. Renyi's entropy is combined with a Parzen windows estimation of the error pdf to form the basis of two criteria (minimum entropy and maximum correntropy) under which neural networks are trained. The results are favorably compared in online and offline training with the traditional minimum square error (MSE) criterion. Real case examples for two distinct wind parks are presented.**

***Index Terms*—Correntropy, entropy, neural networks, Parzen windows, wind power forecasting.**

I. Introduction

WHEN wind generation is significant in a power system, wind power forecasting becomes an important factor in defining the operation planning policies to be adopted by a transmission system operator (TSO), namely in accepting high wind penetration [1]. Furthermore, in a market organization of the business, the wind power contribution to the generation pool becomes important in defining the price in the daily or hourly market: variations in the estimated wind power (placed at zero cost as a base generation before the market bidding process) will influence the final clearing price. If an agent has the power to manipulate the wind power prediction, it has the power to manipulate the market. For these reasons, wind power prediction has become a major concern in the European Union and TSOs, generating companies (GENCOs), and regulators all support efforts to develop better, more reliable, robust, and accurate forecasting models. Wind park owners also benefit from better wind power prediction, to support a competitive participation in electricity markets against more stable energy sources [2].

The prediction of power output from a wind park is highly important presently in Europe, where the growing penetration of wind generation has reached heavy percentages (in the range of 5% to 20%) in some countries in recent years, like Germany, Spain, Denmark, or other, and increases in these values are common targets for energy policies defined. For instance, in Portugal by 2010, some 5100 MW of wind generators will be installed, when country peak power consumption is about 8500 MW in 2008. So, we are no longer talking of marginal effects.

Short-term wind power forecasting refers to the prediction of electric power output from wind parks in a range from a few up to 72 h ahead. The basic inputs to a forecasting model are wind speed and direction predictions, usually coming from a numerical weather prediction (NWP) meso-scale model, which must be transformed into power predictions. The quality of predictions will be evaluated in the signal processing sense, i.e., by measuring the adherence of the prediction to real data. This is important for system operators. Other issues such as the economical value of errors, important for wind power producers, will not be discussed.

This paper presents practical results supporting two ideas: a) criteria based on entropy (measure of information content) of the prediction error distribution are more suitable than the traditional minimum square error (MSE) criterion to train accurate wind power prediction models, and b) the entropy-based criteria can be formatted into online self-adaptive models that perform better than offline trained models when using feed-forward neural networks.

II. Brief Overview

A. Main Trends

In [3], a recent overview of the history and state of the art in wind power forecasting is summarized. Two classes of approaches can be found in the literature: statistical and physical. The fundamental idea of the latter is to refine the NWP forecasts through physical considerations about the site, such as surface roughness, orography, obstacles, and stratification of the atmosphere, and by modeling the profile of the local wind possibly accounting for atmospheric stability. Several physical approaches have been developed and some are being used as forecasting tools for electric power system operators or wind power producers, such as the Danish Prediktor model [4], or the German Previento model [5].

The statistical approach is based in one or more models 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 [6],

Manuscript received June 03, 2008; revised January 17, 2009. First published September 15, 2009; current version published October 21, 2009. Paper no. TPWRS-00435-2008.

R. J. Bessa is with INESC Porto, Instituto de Engenharia de Sistemas e Computadores do Porto, Porto, Portugal, and also with FEP, Faculty of Economy of the University of Porto, Porto, Portugal (e-mail: rbessa@inescporto.pt).

V. Miranda is with INESC Porto, Porto, Portugal, and also with FEUP, Faculty of Engineering of the University of Porto, Porto, Portugal (e-mail: vmiranda@inescporto.pt).

J. Gama is with INESC Porto LA/LIAAD, Laboratory of Artificial Intelligence and Decision Support, Porto, Portugal, and also with FEP, Porto, Portugal (e-mail: jgama@liaad.up.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2009.2030291

[7]; and models that use, in addition to the mean electric power time series data, forecasted values from an NWP model corresponding mainly to hourly mean wind speed and direction [8], [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.

It is well known that the wind speed versus power curve of a wind turbine is highly nonlinear. The transformation of wind speed to wind power changes the statistical properties of the errors. This has been shown, for instance, in [10] for six sites in Germany, where error distributions from wind power prediction models were right skewed and had positive excess of kurtosis, meaning that they were asymmetrical, presented a higher frequency of errors to the left of the mean, and were flatter than the Gaussian distribution. The same shape of the error distribution can be found in [11] and [12].

To deal with the non-Gaussian nature of the wind power forecasting error, nonparametric methods have been developed. For example, in [13], a local quantile regression is used to forecast the quantiles of the probability distribution; the kernel density estimation to forecast the pdf is proposed in [14]. Furthermore, alternative cost functions have been developed; for example, in [15], the minimization of a weighted total least squares for the local linear regression is used; in [16], a local polynomial regression based on robust estimation is presented.

Moreover, it has been observed that wind speed behavior exhibits a trait called *concept drift* in the vocabulary of specialists of data streams, meaning that it changes characteristics over time. Therefore, a model trained offline will display, after some time, a pattern of growing error in prediction values.

To deal with stationarity, adaptive models have been developed. In [16], the estimation of the model parameters is based on an exponential weighted adaptive recursive least squares controlled by a forgetting factor. Also in [15], a recursive method for the estimation of the local model coefficients is proposed; the time dependence of the cost function is ensured by exponential forgetting of past observations. Another example is the Spanish model Sipreolico [9], which is being used as forecasting tool by the Spanish TSO (REE). The use of stochastic gradient for online training of neural networks in wind power forecasting was addressed in [17]. The application to local recurrent neural networks of online learning algorithms based on the recursive prediction error is described in [18]. However, even now, some authors still seem to train neural networks for wind power forecasting in an offline mode: one example is presented in [8].

When observing the literature, one realizes that online and offline models usually adopt the MSE as a quality criterion. The applicability of MSE to train a mapper (any model mapping an input-output relation such as neural networks, fuzzy inference systems, time series, or other, with parameters to be learned) is optimal only if the probability distribution function (pdf) of the prediction errors is Gaussian [19]. Minimizing the square error is equivalent to minimizing the variance of the error distribution. Using this criterion, the higher moments (e.g., skewness, kurtosis, etc.) are not captured, but they contain information that should be passed to the parameters (weights) of the neural network instead of remaining in the error distribution.

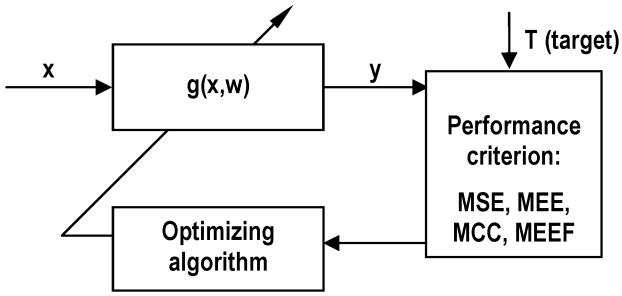

Fig. 1. Basic arrangement of a mapper training procedure identifying its three main modules.

B. Entropy Criteria for Non-Gaussian Prediction Errors

The presence of non-Gaussian distributions has motivated research for techniques that would train mappers 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) [20].

In a first paper devoted to wind power estimation [21], an evolutionary particle swarm optimization (EPSO) algorithm was used to optimize offline the weights of a Takagi–Sugeno Fuzzy Inference System (TS-FIS) to perform wind speed conversion to wind power based on wind speed and wind direction measurements. In that paper, then, a comparison was made between a TS-FIS trained by minimizing the mean square error of predictions and one trained by minimizing the Renyi's quadratic entropy [22] of the error distribution—and the results have shown that a model with higher frequency of errors close to zero was produced by the entropy-based model.

In a more recent paper [23], the authors have engaged in evaluating the performance of neural networks, trained in offline mode, comparing the MSE criterion with three ITL inspired criteria. The conclusion, drawn from the analysis of two real cases of wind parks in Portugal, was unmistakable: in offline training, entropy as a performance criterion leads to better predictions (in terms of higher frequency of errors close to zero and insensitivity to outliers) than adopting minimum square error as a training criterion.

Wind power prediction errors are non-Gaussian (see also Appendix B). This paper brings further contributions to build better wind power forecasting models by extending the principles of information theoretic learning [24]–[27] and the concept of training mappers based on entropy to online trained systems in three-day ahead forecasting. Until know, the online training with entropy had been only tested in artificial data, and no publication has so far come to light about online training of neural networks using correntropy.

III. MSE versus MEE, MCC, and MEEF Criteria

Fig. 1 illustrates a mapper $\mathbf{g}$ being subject to supervised training to produce an output y from an input x by having its weights w adjusted by an optimizing algorithm that is driven by a performance criterion. This picture clearly illustrates that

the choice of criterion is decoupled from the choice of algorithm. The MSE and the three ITL criteria under comparison, optimized by a backpropagation algorithm, are:

1) Minimum square error (MSE). This is the classical criterion that minimizes the variance of the error distribution and has the form

$$MSE(\varepsilon) \Leftrightarrow min \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \varepsilon_i^2 \tag{1}$$

where $\varepsilon = (T_i - y_i)$ is the error of sample i relative to the target value T_i.

2) Minimum error entropy (MEE). This is the fundamental ITL criterion where the minimization of the entropy of the error distribution is equivalent to

$$MEE(\varepsilon) \Leftrightarrow maxV = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G(\varepsilon_i - \varepsilon_j, 2\sigma^2 I) \tag{2}$$

where G is a Gaussian function, in this case with a variance given by a value represented by $2\sigma^2$, for reasons that will be seen in Appendix A. V is called the *information potential*.

3) Maximum correntropy (MCC). This criterion is based on a generalized similarity measure called correntropy [18] and may be translated by

$$MCC(\varepsilon) \Leftrightarrow max \frac{1}{N} \sum_{i=1}^{N} G(\varepsilon_i, \sigma^2 I). \tag{3}$$

4) Minimum error entropy with fiducial points (MEEF). This criterion [18] intends to anchor the error distribution to a zero mean by defining a compromise between minimizing entropy and maximizing correntropy through a cost function

$$MEEF(\varepsilon) \Leftrightarrow max\, \gamma \frac{1}{N} \sum_{i=1}^{N} G(\varepsilon_i, \sigma^2 \mathbf{I}) + (1-\gamma) \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} G(\varepsilon_j - \varepsilon_i, 2\sigma^2 \mathbf{I}) \tag{4}$$

where γ is a weighting constant between 0 and 1.

MEE is an exact criterion in terms of entropy concept while MCC is only an approximation. However, MEE is much more demanding in computing effort. Also, MEE has degenerate minima because it is insensitive to the mean of the error. 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 cost function, leading to the MEEF criterion.

Criteria 2, 3, and 4 are based on a representation of the error pdf by the Parzen window method [28], where the expression of the estimation $\hat{f}_Y$ for the real pdf f_Y of a set of N points (errors) is a summation of individual contributions:

$$\hat{f}_Y(z) = \frac{1}{N} \sum_{i=1}^{N} G(z - \varepsilon_i, \sigma^2 I). \tag{5}$$

This expression assumes that a Gaussian kernel G with variance σ^2 is used to define the Parzen window around each point (kernel bandwidth)—see Appendix A.

These four criteria represent different options and assumptions about the error distribution when training a neural network. Ideally, the best error distribution would have a Dirac function representing its probability density, meaning that all errors would be equal—this would allow one to simply compensate this systematic error with a bias to have a *perfect* mapper producing exact forecasts. The Dirac function has minimum entropy and the MEE criterion seeks precisely that target. As a consequence, it is more effective than the MSE criterion in isolating outliers [29]. However, as the entropy calculation is done over all pairs of errors of the training (or test) set, each evaluation becomes much heavier than with the MSE or the MCC criteria that depend only on the errors.

Correntropy is discussed in [30] and [31]. It has been proven that correntropy is related with a distance measure CIM(X,Y) between two arbitrary scalar random variables X and Y satisfying all the properties of a metric. CIM divides space in three different regions: when the error is close to zero, CIM is equivalent to an L2 norm (Euclidean, similar to the MSE criterion); when the error grows, CIM becomes like an L1 norm (sum of the differences of coordinates); when the error is very large, CIM becomes an L0 norm, the metric saturates and becomes very insensitive to large errors. This property highlights the importance of the definition 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.

When we use correntropy to train adaptive systems, we actually make the system output close to the desired response in the CIM sense. We can use the MCC as a performance function, with the advantage over MSE of being a local criterion of similarity and very useful for cases with nonzero mean, non-Gaussian, with large outliers. It does not require the computing effort of MEE but tends to minimize entropy because it tends to maximize the pdf value at the origin.

All criteria are continuous functions of the errors in a real-valued domain that can be differentiated. This means that one may derive expressions in all cases for the derivative of an error relative to the output of a neural network. These can be used in a chain rule derivative of errors as a function of network weights, and therefore, in all three cases, one may build a gradient-based back-propagation algorithm to optimize weights and train neural networks to perform according to each criterion [30], [32], [33]. The backpropagation algorithm is not a mandatory condition however, to determine the conclusions of this paper, and other optimization algorithms could have been used instead.

IV. Importance of Self-Adaptive Online Training

Offline training of neural networks is a well-known technique. For online training, the methodology adopted when following the MCC or MSE criteria is as follows.

- First, train a neural network using a batch back-propagation approach with the available historical data.
- Then, in the online mode, the neural network makes predictions for time sample $\mathrm{t} + \mathrm{k}$ at the time stamp t.

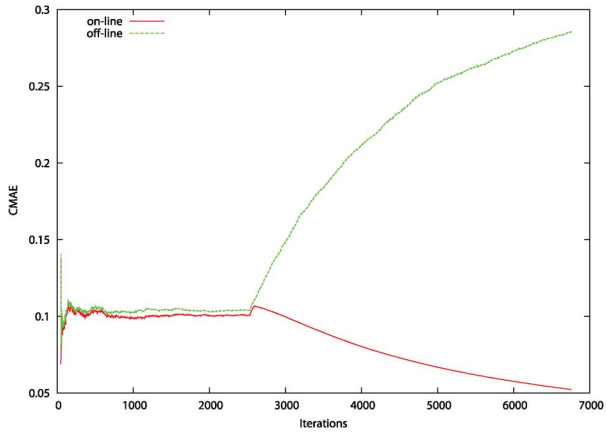

Fig. 2. Comparison between online and offline train for artificial data with concept change.

- When the measured value is known past k time stamps, then the neural network forecasts again for time stamp t and the forecast error (or the correntropy) committed in t (on the new arrived measured value) is computed and back-propagated through the network (weights and bias are updated) only once.

The artificial data set consists on 12 667 examples generated using the function from Friedman 2 problem [34] and 4223 generated using the function from Friedman 3 problem [34]. The white Gaussian error parameters were the same used in the Friedman problems. A preliminary test was performed with such approach in artificial data exhibiting concept drift, to confirm the importance of providing online training when in the presence of this phenomenon.

Two neural networks were trained in offline and online modes, under the MSE criterion. The cumulative mean absolute error (CMAE) of the normalized values was used for evaluating the NN performance over time (iterations), for test purposes (avoiding a bias in performance evaluation). Fig. 2 displays the evolution of CMAE in both cases, when a concept change was introduced in the test data, in iteration 2581. The NN trained online quickly adapts to the new concept, in contrast to the offline model which is unable to change with the concept. The online mode presents a decrease in the error because the values of the second function are lower and so the error also displays lower differences.

The methodology for the self-adaptive online training under the MEE criterion is similar to the one described above. The only difference is that the information potential of the error is then recursively estimated [29]. When a new measure arrives, the prediction error of the neural network is computed and added to a time window with M errors of previous predictions. The information potential of the error is then recursive estimated using the following equation:

$$V_{k+1} = (1-\lambda)V_k + \frac{\lambda}{M}\sum_{i=k-L+\backslash}^{k} k_\sigma(x_i - x_{k+1}) \qquad (6)$$

where λ is a forgetting factor with values between 0 and 1. A window with the M most recent errors was also used. The recursion of the information potential uses the gradient from the previous time step.

TABLE I
SOME CHARACTERISTICS OF THE DATA AVAILABLE

Year	Maximum	Minimum	Mean	Variance
	% r. pw.	% r. pw.	% r. pw.	(MW^2)
Wind park A (~20 MW)				
2005	100.37	-0.36	28.31	44478.37
2006	100.36	-0.42	26.28	42020.23
Wind park B (~35 MW)				
2005	100.06	0.00	23.06	61282.50
2006	101.01	0.00	23.60	63823.15

It is important to choose a small step size because the two gradients $\partial V_k/\partial w_{k+1}$ and $\partial V_k/\partial w_k$ must have a small difference between then.

Although in wind forecasting problems it is difficult to detect the point where concept changes, it is beyond doubt that this effect happens. This concern has been addressed in [9], [17], and [35].

V. CASE STUDIES

This section presents results for two training modes (offline and online) in two real wind power forecasting problems, comparing the performance of neural networks resulting from the adoption of several criteria, comparing the traditional MSE with the set of three ITL criteria.

A. Wind Park Characteristics

Wind park A has a rated power around 20 MW with equal wind turbines, all below 2 MW. The wind park is situated in Iberia, in a complex mountainous region.

Wind park B has a rated power of around 35 MW and comprises a number of equal wind turbines above 2 MW. It is situated in Iberia, in a soft mountainous region near the coast.

B. Data Characteristics

Data collected in the wind parks include SCADA registers with average 10 min power delivered by the wind park to the grid. One has also available forecasts produced for the same period by an MM5 [http://www.mmm.ucar.edu/mm5/] model, for mean wind speed, wind direction, temperature, and atmospheric pressure, for a reference point in the wind park, with forecasting horizons ranging from 0 to 72 h in half-hour intervals. The MM5 model is initialized every day with the predictions of the GFS model (global model) corresponding to the assimilation of atmospheric data at 00:00 GMT. The MM5 forecasts are available at about 07:00 GMT. The first forecasted values corresponded to 07:00 of the present day and the last forecasts to 00:00 three days later.

To organize the tests, the available data were divided into three sets. The first set, with several months of data, was used as a training set. The second set, with one month of data, was used as a validation set. The third set, with the remaining months, was used as a testing set for comparative proposes among the different training modes and training criteria. Table I shows the statistical characteristics of the electric power series data. The negative values indicate electric power consumption of the wind park.

C. Testing Methodology

A neural network using only NWP predictions as inputs is expected to have an unsatisfactory performance when predicting power for the first six hours [17]. This effect is usually compensated by combining with models that only employ time-series models for this first period. The results in this paper are concentrating on power predictions up to 72 h ahead and the compensation for the first six hours is not considered.

Neural networks were used to predict the power $\mathrm{p_{t+k|t}}$ produced by the wind park a look-ahead $t+k$. The wind power prediction was performed for each day of the test data set when new NWP predictions (7:00 GMT) become available. For online training, the arrival of new SCADA measurements were simulated, and between two predictions of consecutive days, some function of the error detected from comparing prediction with measurement was back-propagated.

The inputs used by the neural networks are NWP meteorological forecasted values: forecasted mean wind speed values $(\mathrm{v_{t+k|t}})$, forecasted wind direction values $(\mathrm{d_{t+k|t}})$, and an index m corresponding to the number of past half hours of NWP forecasted values. Due to the cyclic characteristic of wind direction (geography) and the variable m (daily hour), these variables are represented by sine and cosine components. For each day one has three NWP predictions available; these predictions differ on the day they were made (the antiquity). This means a total of five input variables, which were standardized using the min-max method.

A feed-forward neural network was organized with only one hidden layer comprising nine neurons, using hyperbolic tangent activation functions. Determining the best topology of the neural network was out of the scope of this work. The stopping criterion for each cost function was early stopping and maximum number of iterations (800 epochs). The training was stopped when the validation error starts to increase.

Following the analysis performed in [23] and [27], the iRPROP algorithm [36] was used instead of the classic back-propagation, because with a variable learning rate, it can achieve faster convergence and maintains good performance. Improved results were achieved with an adaptive kernel size strategy: in each case one starts with a large kernel size which during training is gradually and slowly decreased. Finally, the use of the batch-sequential approach tested in [23] for this type of problems helped in reducing the large processing time.

In the iRPROP training algorithm, the standard parameter values were chosen for the increase and decrease factors $\eta+ = 1.2, \eta- = 0.5$ because they actually produce the best results independently of the learning problem. The remaining parameters adopted in iRPROP were the range of the weight update values $(\Delta\min = 0, \Delta\max = 50)$ and the initial weight update value $\Delta_0 = 0.0125$. The tuning of these parameters is not critical. The batch-sequential subsets length was 2000 points.

The error measure adopted to evaluate performance was normalized mean absolute error (NMAE). This error average was calculated over all three-day horizon windows available in the test set months. The rationale for this choice is two-fold: it is a criterion often used by researchers working in wind power forecasting [37] and it would not introduce a bias in comparisons (one is adopting a criterion not used in the calculations). The errors are also compared with a classical reference model: persistence, which corresponds to offer as forecast the value of the last known value of the generation time series.

TABLE II
TRAINING, VALIDATION, AND TEST SETS

Wind park	Validation	Training	Test
A	Jan (2005)	Feb-Dec (2005)	Jan-Dec (2006)
B	Feb (2005)	Mar-Dec (2005)	Apr-Dec (2006)

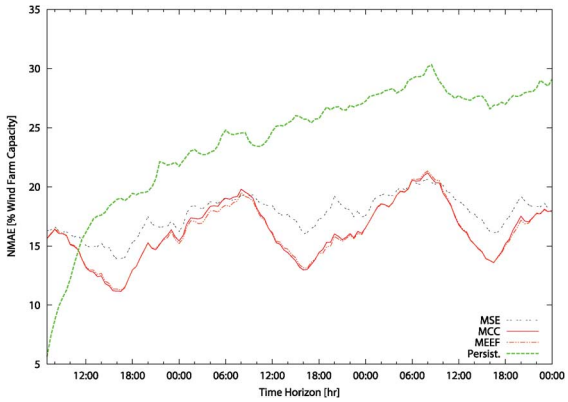

Fig. 3. Wind park A: NMAE errors for offline predictions.

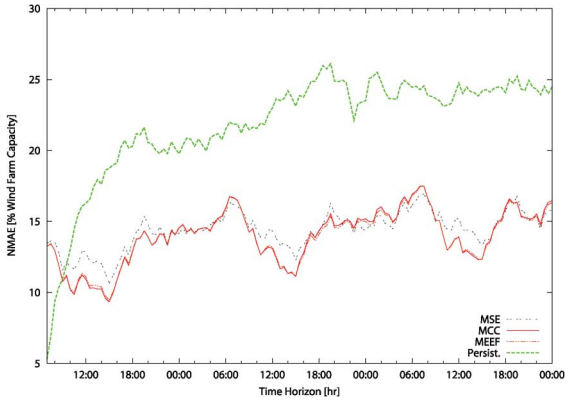

Fig. 4. Wind park B: NMAE errors for offline predictions.

D. Analysis of Offline Prediction Results

This section presents the offline prediction results obtained for the two wind parks when training under MSE, MCC, and MEEF, by showing the NMAE of the power generation forecast errors in forecast horizons from 0.5 until 72 h.

The exclusion of MEE in these tests is justified because in early tests, the error distribution resulted asymmetric and with heavy tails [23]. Modifying the output bias of the neural network to yield zero mean error over the training data set, when applying the simple MEE for problems with nonsymmetric distributions, may not be the best approach.

Table II presents the months used for training, testing, and validation of the neural networks.

In the kernel size annealing strategy used, the initial kernel size for MEEF was 0.3 and for MCC was 0.02. The value of the weighting constant γ in MEEF was set to 0.3.

Figs. 3 and 4 show the NMAE errors for the MCC, MEEF, and MSE models for each hour through the forecast horizons,

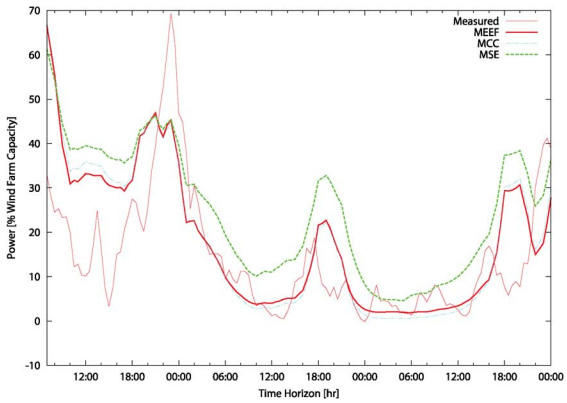

Fig. 5. Forecasted and measured values of the electric power from wind park A obtained with the three cost functions (October 26–28, 2006).

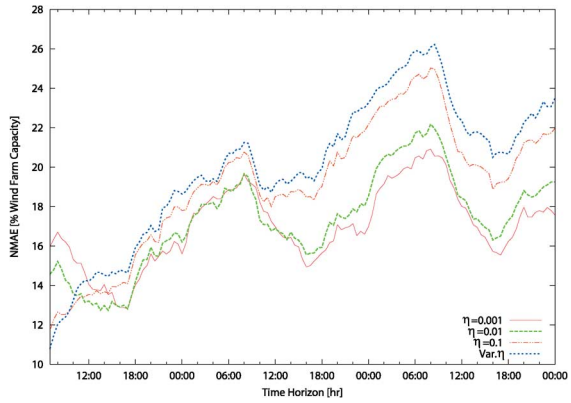

Fig. 6. Wind park A: NMAE for online training with different fixed learning rates and a variable learning rate, for the MSE criterion.

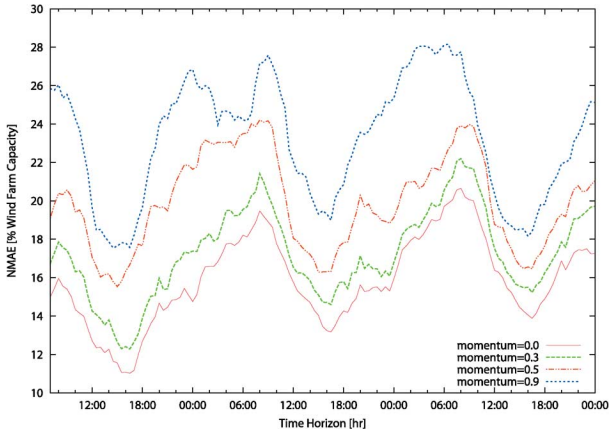

Fig. 7. Wind park A: NMAE for online training with different momentum terms and MCC cost function.

in each wind park (WP), for lead times up to 72 h. Note that the predictions were made only after 7:00 of the first day of the time horizon. The figures include also the predictions from the "persistence" model. These figures make visually evident that the MCC and MEEF criteria presents better results than MSE for all wind parks and for almost every lead time of the forecast horizon. No significant differences exist between MEEF and MCC.

The average difference between the two criteria (MEEF and MCC) and MSE is 1.48% and 0.45%, for wind park A and B, respectively. The average difference between MEEF and MCC is favorable to MEEF for wind farm A in 0.04% and for wind farm B in 0.02%. The NMAE of the persistence model error rises relatively quickly in the first six hours for each wind park. After that, the error stabilizes or increases slowly. As explained, neural network models with only NWP as input display a worse performance than persistence in these first hours.

The use of the persistence is interesting here because it gives a clear indication on how easy (or difficult) predictions may be for distinct parks. For example, neural network predictions for wind park B provided the smallest errors but also the persistence model for this wind park has the best results; so, wind park B is more "predictable".

Fig. 5 compares forecasted and measured values of the 30 min mean electric power for wind park A: the forecast was made at 07:00 on October 26, one of the days of the test data set, with a forecast step of half hour, until 00:00 on October 29, 2006. One may observe that in this case, both criteria follow reasonably well the power values measured from SCADA but the MCC and MEEF models give better results. The forecasted value in both cases followed a similar pattern because the same wind forecast from MM5 was used.

E. Analysis of Self-Adaptive Online Prediction Results

This section presents the prediction results obtained with the self-adaptive online methodology described in Section IV for the two wind parks with the three criteria, as well as an assessment of the impact of each neural network parameter on the online training results.

The online training with MEEF was achieved as a hybrid of the entropy recursive estimation and the stochastic back-propagation of the correntropy. The same value of the weighting constant γ used for the offline training was used.

The first step was to analyze the impact of learning rate and momentum in a real-time neural network training. For this purpose, the same data of the offline training was used. The objective is to change only one parameter at each time. For the self-adaptive online training, the heuristic used in the iRPROP training algorithm was not applicable for this case.

Fig. 6 displays the results obtained for wind park A with different fixed learning rates and with a simple parameter adaptation technique described in [38]; the momentum term was set to zero and the criterion used was MSE.

This figure and other tests made it became clear that for an online training, the effective learning rate must be small. A learning rate with a high value makes the learning process hard to converge while a low value makes the convergence slow and smooth, which is desirable in problems with concept change. For the online methodology (process one new measured value at a time), changing the learning rate based on only the sign of the gradient has not proved to be the best criterion. A more elaborated mechanism with change detection must be developed and tested in the future.

Fig. 7 presents the results for wind park A, for the cost function MCC with fixed learning rate $(\mathrm{lr} = 0.001)$, also a fixed kernel size (0.5) and different momentum terms. The same behavior was verified for wind parks B with MCC, and also with

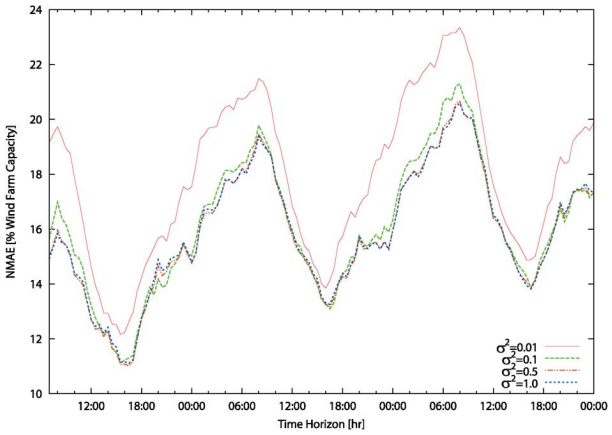

Fig. 8. Wind park A: NMAE for online training with different kernel sizes for the MCC criterion.

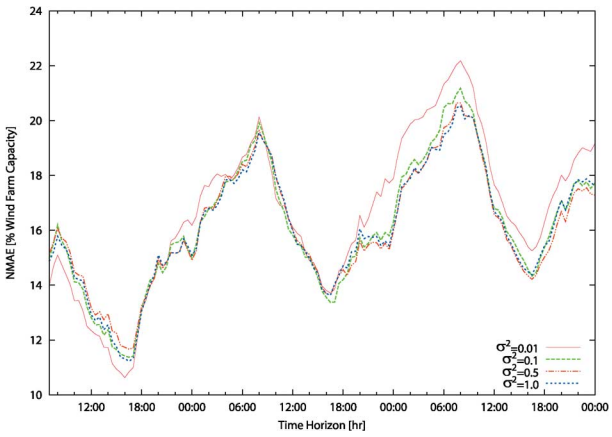

Fig. 9. Wind park A: NMAE for online training with different kernel sizes for the MEEF criterion.

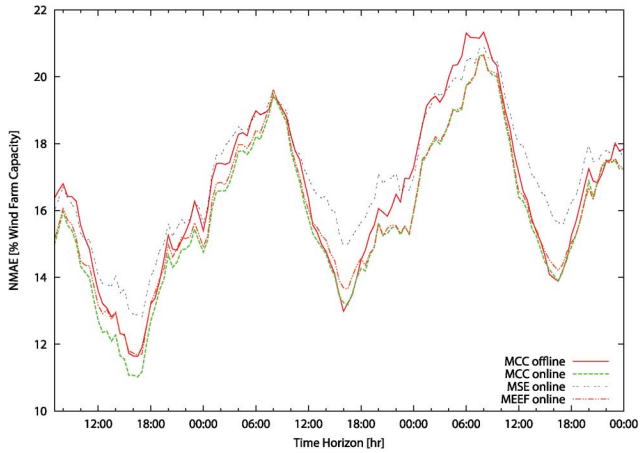

Fig. 10. Wind park A: NMAE for online training.

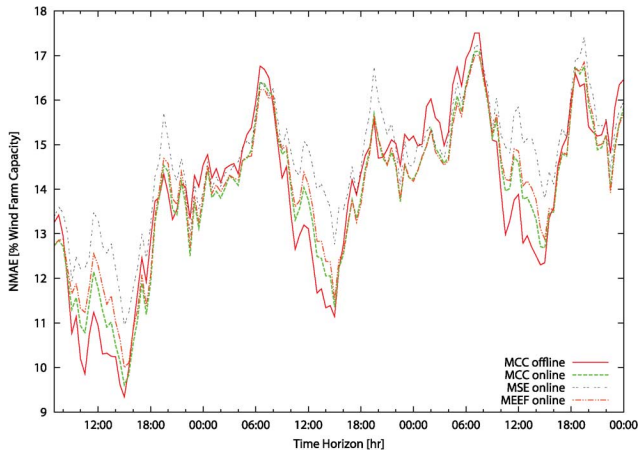

Fig. 11. Wind park B: NMAE for online training.

MSE. A general conclusion from these tests is that the momentum term should have a low or zero value in self-adaptive online training. This result has a simple explanation. In a nonstationary signal, the inputs of the neural network change with time. A known result is that the error surface and the location of the minimum in this surface change with inputs.

If this surface changes during training (not in offline training with epochs), a parameter preventing the learning process from stopping in a shallow local minimum is unimportant. Also, a momentum term with a high value increases the weights' update, which in this type of problem is not desirable.

Fig. 8 presents the results with a fixed learning rate ($\mathrm{lr} = 0.001$), without momentum term and different kernel sizes for wind park A. The best results were obtained with a kernel of 0.5; the results for sizes between 0.1 and 1.2 are very similar.

Fig. 9 shows the results for online training with the MEEF criterion with a fixed learning rate (0.001) and different kernel sizes (the same for MCC and MEE terms) for wind park A. The same behavior with different kernel sizes for both MCC and MEEF was verified for wind park B.

The correct choice of kernel size is a major concern. As expected, a small kernel size makes the neural network very robust to outliers. But these outliers in a problem with concept drift may be points from a new concept, and with a small kernel size, the neural network cannot adapt. More important than having a variable learning rate, for the case of the ITL criteria, is to vary the kernel size as the concept changes. The use of a kernel size that is a function of a change detection mechanism is a strong tool to deal with concept change.

The same data were used for the online training. For wind park B, one could examine the behavior of the method in a demanding circumstance: training data comprised a period when the wind park had a different installed power from the test set; this is one cause behind concept change. The training data for wind park B consisted of months March to June 2005, and the test data consisted of months July to December 2005. The parameters of the neural network adopted for the online training were the best parameters determined above. The forgetting factor for the entropy recursive estimation was 0.08 and the window length M was 2000.

Figs. 10 and 11 compare the use of MSE, MCC, and MEEF criteria for each wind park for an online training mode. The best result of the offline training (MCC) is depicted for comparison with the online training result. Notice how the offline MCC performs better than the MSE online.

In the online case, there are some differences between MEEF and MCC. The average difference between online MEEF and online MCC is 0.1% favorable to MCC for wind farm A and 0.06% favorable to MEEF for wind farm B.

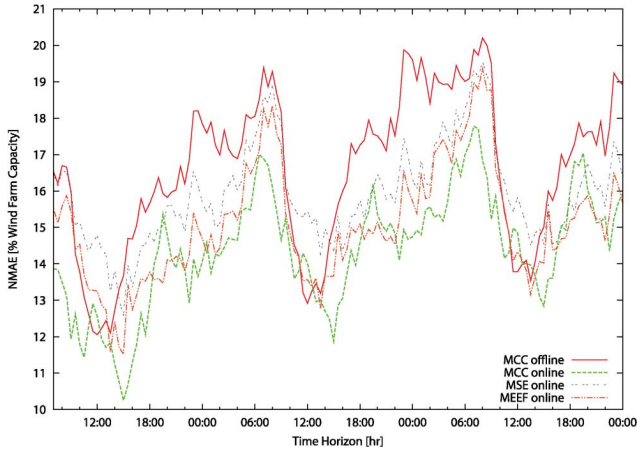

Fig. 12. Wind park B' data: NMAE for online training.

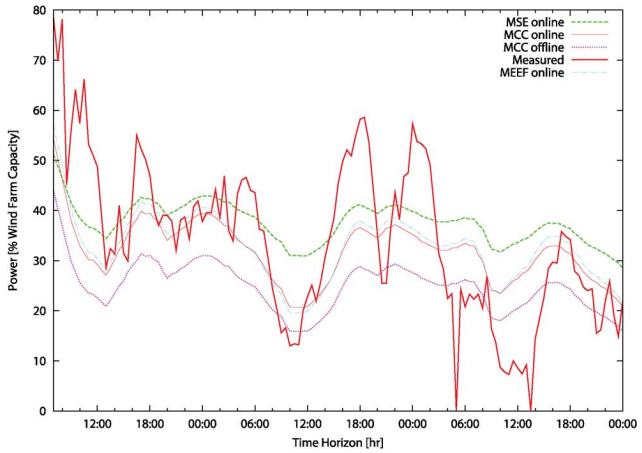

Fig. 13. Wind park B: forecasted and measured values of the electric power obtained with online and offline training (November 12–14, 2005).

The online training presents better results than offline for all cases. The case of wind park B' data (Fig. 12) where there was an addition of installed capacity is an excellent result. It shows that the online learning mode is a good tool to deal with upgrades in wind park capacity without the need to re-train the forecasting model.

Also bear in mind that in cases where a wind park is built close to an existing one, the concept may change.

In Fig. 12, the average difference between online MEEF and online MCC is 0.09% and favorable to MEEF. The average difference between the online training and offline of MEEF is 1.56%. The average difference between online MEEF and online MSE is 1.19%. This illustrates that the theoretical result is achievable in practice: an entropy criterion will always lead to some improvement over MSE, except in the limit case of, by chance, having a specific site Gaussian prediction errors. The actual improvement will depend of the wind park site. Using entropy assures that the best result will be achieved and this is independent of the site.

Finally, Fig. 13 shows one example of forecasted and measured values of the 30 min mean electric power from wind park B obtained with online and offline training. The forecast was made at 07:00 on November 12, 2006, one of the days of the test data set from wind park B, until 00:00 on November 15, 2006. We can visually observe that in this case, the online training model is a better approach to the actual measured value. An analysis of the error probability distribution function would confirm that smaller errors have a higher frequency for the online model.

VI. Conclusions

Wind power prediction is associated with non-Gaussian error probability distributions. It is unfortunate that so many prediction works remain based on criteria optimizing variance (such as MSE) when advantage could be taken from dealing with the full information content of the underlying probability distribution by training models based on entropy concepts.

This paper is a major contribution to the adoption of ITL criteria to wind power prediction. Evidence is given that online training procedures for a 72-h prediction horizon can be organized for entropy-based criteria and that these produce better models than with MSE or in offline training. Entropy combined with a self-adaptive online training mode allows one to adequately deal with data streams in the presence of concept drift or concept changes. These derive from wind behavior and from the practical operation of wind parks, namely, due to the variation in the available generating capacity, either because of maintenance, failure, or simply because of capacity additions.

The paper shows that the direct use of an entropy measure such as the Renyi's quadratic function (the MEE criterion) is possible, but that an approximation based on correntropy (the MCC criterion) has clear computing advantages. The paper also includes useful indications on the algorithmic implementation of the methods.

All statements have been supported with evidence from real cases confirming theoretical expectations. It is important to point out that the objective of this paper was not to build a complete model to predict for time horizons between 24–72 h. But the integration of the methodology described in this paper will be a major improvement in more sophisticated wind power prediction systems: the entropy criteria will always provide a better result than the MSE criterion, in the signal processing sense, whenever errors are not Gaussian-distributed. The use of an independent criterion such as NMAE has served to confirm such assertion.

In all real cases tested, the ITL criteria presented better results than MSE for all wind parks in the online training mode and for almost every lead time of the forecast horizon. Therefore, the paper allows the conclusion 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.

Appendix A

A. Estimation of a Pdf With Parzen Windows

The estimation of the pdf of data from a sample constituted by discrete points $y_i \in R^M, \mathrm{i} = 1, \ldots, \mathrm{N}$ in a M-dimensional space, may be done by the Parzen window method [4]. This technique looks at a point as being locally described by a probability density Dirac function, which is replaced or approximated

by a continuous function (kernel) representing it. If a Gaussian kernel G is used, the expression of the estimation $\hat{f}_Y$ for the real pdf f_Y of a set of N points is

$$\hat{f}_Y(z) = \frac{1}{N}\sum_{i=1}^{N} G(z - y_i, \sigma^2 I) \tag{7}$$

where σ^2I is the covariance matrix (here assumed with independent and equal variances in all dimensions). In each dimension, we have

$$G(z_k - y_{ik}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(z_k - y_{ik})^2}. \tag{8}$$

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. Renyi's Entropy and Its Estimation With Parzen Windows

Renyi's entropy [11] of a discrete probability distribution $\mathrm{P} = (\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_\mathrm{n})$ is defined as

$$H_{R\alpha} = \frac{1}{1-\alpha} log \sum_{k=1}^{N} p_k^{\alpha} \quad \text{with} \quad \alpha > 0, \alpha \neq 1. \tag{9}$$

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

$$H_{R2} = -log \sum_{k=1}^{N} p_k^2. \tag{10}$$

This definition can be generalized for a continuous random variable Y with pdf $\mathrm{f}_\mathrm{Y}(\mathrm{z})$:

$$H_{R2} = -log \int_{-\infty}^{+\infty} f_Y(z)^2 dz. \tag{11}$$

A breakthrough has been achieved with combining Renyi's entropy definition with an estimate of a pdf by the Parzen window method [20], [24]—this has been called ITL. An entropy estimator for a discrete set of data points $\{\mathrm{y}\}$ is

$$H_{R2}(y) = -log \int_{-\infty}^{+\infty} \hat{f}_Y(z)^2 dz = -logV(y) \tag{12}$$

where, using (7)

$$V(y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \int_{-\infty}^{+\infty} G(z - y_i, \sigma^2 I) G(z - y_j, \sigma^2 I) dz. \tag{13}$$

In this expression, we recognize the convolution of Gaussian functions, and the integral of two Gaussians with equal standard deviations is a Gaussian with twice the standard deviation. Then we have the following result:

$$V(y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(yi - yj, 2\sigma^2 I) \tag{14}$$

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.

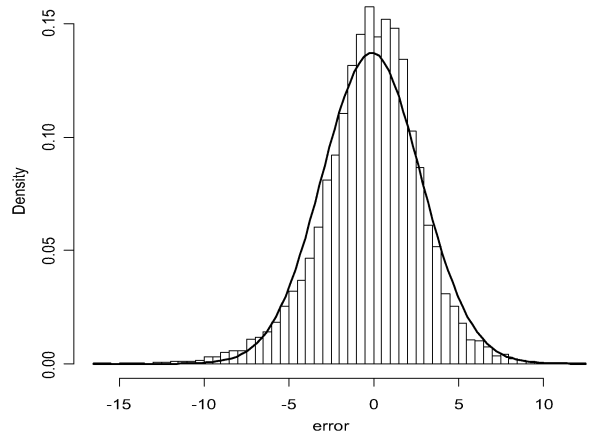

Fig. 14. MM5 wind speed prediction error histogram in one wind park and normal distribution (curve) with mean of -0.10 and standard deviation of 2.90.

APPENDIX B

Real data reinforce the need to develop models that do not rely on the assumption that errors are Gaussian distributed (an assumption behind the MSE criterion if this is thought as the best criterion to train mappers).

Fig. 14 displays a histogram of one year (2005) of 24-h prediction errors from the NWP MM5 meteorological mesoscale wind speed/direction model—Against real wind speed values measured by the metering station at a real wind park.

A *Kolmogorov–Smirnov* test [39] rejected the null hypothesis (that the NWP error could be taken as a Gaussian distribution) for all practical levels of significance with a *p-value* of 2.2×10^{-16}. This non-Gaussian uncertainty will contaminate any model predicting wind power and adds importance to the working hypothesis of non-Gaussian errors, already suggested in [10].

REFERENCES

[1] M. L. Ahlstrom and R. M. Zavadil, "The role of wind forecasting in grid operations & reliability," in *Proc. IEEE/PES Transmission and Distribution Conf. Exhib.: Asia and Pacific*, China, 2005, pp. 1–5.

[2] J. Usaola, O. Ravelo, G. González, F. Soto, M. C. Dávila, and B. Díaz-Guerra, "Benefits for wind energy in electricity markets from using short term wind power prediction tools; A simulation study," *Wind Eng.*, vol. 28, no. 1, pp. 119–127, 2004.

[3] A. Costa, A. Crespo, J. Navarro, G. Lizcano, H. Madsen, and E. Feitosa, "A review on the young history of the wind power short-term prediction," *Renewab. Sustain. Energy Rev.*, vol. 12, no. 6, pp. 1725–1744, 2008.

[4] L. Landberg, "Short term prediction of the power production of wind parks," *J. Wind Eng. Ind. Aerodynam.*, vol. 80, pp. 207–220, 1999.

[5] U. Focken, M. Lange, and H.-P. Waldl, "Previento—A wind power prediction system with an innovative upscaling algorithm," in *Proc. Eur. Wind Energy Conf. (EWEC 01)*, Copenhagen, Denmark, 2001.

[6] T. H. M. El-Fouly, E. F. El-Saadany, and M. M. A. Salama, "Grey predictor for wind energy conversion systems output power prediction," *IEEE Trans. Power Syst.*, vol. 21, no. 3, pp. 1450–1452, Aug. 2006.

[7] I. G. Damousis, M. C. Alexiadis, J. B. Theocharis, and P. S. Dokopoulos, "A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation," *IEEE Trans. Energy Convers.*, vol. 19, no. 2, pp. 352–361, Jun. 2004.
[8] G. Sideratos and N. D. Hatziargyriou, "An advanced statistical method for wind power forecasting," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 258–265, Feb. 2007.
[9] I. Sanchez, "Short-term prediction of wind energy production," *Int. J. Forecast.*, vol. 22, pp. 43–56, 2006.
[10] M. Lange, "On the uncertainty of wind power predictions—Analysis of the forecast accuracy and statistical distribution of errors," *Trans. ASME, J. Solar Energy Eng.*, vol. 2, no. 127, pp. 177–194, May 2005.
[11] H. Bludszuweit, J. A. Dominguez-Navarro, and A. Llombart, "Statistical analysis of wind power forecast error," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 983–991, Aug. 2008.
[12] A. Fabbri, T. Gomezsanroman, J. Rivierabbad, and V. H. Mendezquezada, "Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market," *IEEE Trans. Power Syst.*, vol. 20, no. 3, pp. 1440–1446, Aug. 2005.
[13] J. B. Bremnes, "Probabilistic wind power forecasts using local quantile regression," *Wind Energy*, vol. 7, pp. 47–54, 2004.
[14] J. Juban, L. Fugon, and G. Kariniotakis, "Probabilistic short-term wind power forecasting based on kernel density estimators," in *Proc. Eur. Wind Energy Conf. (EWEC 07)*, Milan, Italy, May 7–10, 2007.
[15] P. Pinson, H. A. Nielsen, H. Madsen, and T. S. Nielsen, "Local linear regression with adaptive orthogonal fitting for the wind power application," *Statist. Comput.*, vol. 18, no. 1, pp. 59–71, 2008.
[16] H. A. Nielsen, P. Pinson, L. E. Christiansen, T. S. Nielsen, H. Madsen, J. Badger, G. Giebel, and H. F. Ravn, "Improvement and automation of tools for short term wind power forecasting," in *Proc. Eur. Wind Energy Conf. EWEC 07*, Milan, Italy, May 7–10, 2007.
[17] G. Kariniotakis, "Contribution to the development of on advanced control system for the optimal management of autonomous wind-diesel systems," Ph.D. dissertation, Ecole Nationale Superieure des Mines de Paris, Centre d'Energetique, Paris, France, Dec. 1996.
[18] T. G. Barbounis and J. B. Theocharis, "Locally recurrent neural networks for wind speed prediction using spatial correlation," *Inf. Sci.*, vol. 177, no. 24, pp. 5775–5797, 2007.
[19] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.
[20] J. C. Principe and D. Xu, "Information-theoretic learning using Renyi's quadratic entropy," in *Proc. 1st Int. Workshop Independent Component Analysis and Signal Separation*, J.-F. Cardoso, C. Jutten, and P. Loubaton, Eds., Aussois, France, 1999, pp. 407–412.
[21] V. Miranda, C. Cerqueira, and C. Monteiro, "Training a FIS with EPSO under an entropy criterion for wind power prediction," in *Proc. Int. Conf. Probabilistic Methods Applied to Power Systems (PMAPS 2006)*, Stockholm, Sweden, Jun. 11–15, 2006.
[22] A. Renyi, "Some fundamental questions of information theory," in *Selected Papers of Alfred Renyi*. Budapest, Hungary: Akademia Kiado, 1976, vol. 2, pp. 526–552.
[23] R. Bessa, V. Miranda, and J. Gama, "Wind power forecasting with entropy-based criteria algorithms," in *Proc. 10th Int. Conf. Probabilistic Methods Applied to Power Systems (PMAPS 2008)*, Rincon, PR, May 2008.
[24] J. C. Principe and D. Xu, "Introduction to information theoretic learning," in *Proc. Int. Joint Conf. Neural Networks (IJCNN'99)*, Washington, DC, Jul. 10–16, 1999, pp. 1783–1787.
[25] J. C. Principe and D. Xu, "Information-theoretic learning using Renyi's quadratic entropy," in *Proc. 1st Int. Workshop Independent Component Analysis and Signal Separation*, J.-F. Cardoso, C. Jutten, and P. Loubaton, Eds., Aussois, France, 1999, pp. 407–412.
[26] D. Erdogmus and J. C. Principe, "Generalized information potential criterion for adaptive system training," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1035–1044, Sep. 2002.
[27] R. A. Morejon and J. C. Principe, "Advanced search algorithms for information-theoretic learning with kernel-based estimators," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 874–884, Jul. 2004.
[28] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.
[29] D. Erdogmus, "Information theoretic learning: Renyi's entropy and its applications to adaptive system training," Ph.D. dissertation, Univ. Florida, Gainesville, FL, 2002.
[30] L. Weifeng, P. P. Pokharel, and J. C. Principe, "Error entropy, correntropy and M-estimation," in *Proc. 16th IEEE Signal Proces. Soc. Workshop Machine Learning for Signal Processing*, 2006, pp. 179–184.
[31] W. Liu, P. Pokharel, and J. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.
[32] D. Erdogmus and J. C. Principe, "Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics," in *Proc. 2nd Int. Workshop Independent Component Analysis and Blind Signal Separation*, P. Pajunen and J. Karhunen, Eds., Helsinki, Finland, 2000, pp. 75–80, Otamedia, Espoo, Finland.
[33] D. Erdgomus and J. C. Principe, "An error-entropy minimization algorithm for supervised training of non-linear adaptive systems," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1780–1786, Jul. 2002.
[34] J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Statist.*, vol. 19, no. 1, pp. 1–67, 1991.
[35] P. Pinson, H. A. Nielsen, and H. Madsen, Robust Estimation of Time-Varying Coefficient Functions—Application to the Modeling of Wind Power Production, Informatics and Mathematical Modelling, Technical Univ. Denmark (DTU), 2007, Tech. Rep.
[36] C. Igel and M. Hüsken, "Improving the Rprop learning algorithm," in *Proc. 2nd Int. ICSC Symp. Neural Computation (NC 2000)*, H. Bothe and R. Rojas, Eds., 2000, pp. 115–121, ICSC Academic Press.
[37] H. Madsen, P. Pinson, G. Kariniotakis, H. A. Nielsen, and T. S. Nielsen, "Standardizing the performance evaluation of short-term wind prediction models," *Wind Eng.*, vol. 29, no. 6, pp. 475–489, Dec. 2005.
[38] L. B. Almeida, T. Langlois, J. D. Amaral, and A. Plakhov, "Parameter adaptation in stochastic optimization," in *On-Line Learning in Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[39] F. J. Massey, Jr., "The Kolmogorov-Smirnov test of goodness of fit," *J. Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, Mar. 1951.

Ricardo J. Bessa received the Licenciado degree in electrical and computer engineering from the Faculty of Engineering of the University of Porto, Porto, Portugal (FEUP), in 2006 and the M.Sc. degree in data analysis and decision support systems from the Faculty of Economy of the University of Porto (FEP) in 2008.

He is also a Researcher at INESC Porto in its Power Systems Unit.

Vladimiro Miranda (M'90–SM'04–F'05) received the Licenciado, Ph.D., and Agregado degrees from the Faculty of Engineering of the University of Porto, Porto, Portugal (FEUP), in 1977, 1982, and 1991, respectively, all in electrical engineering.

In 1981, he joined FEUP and currently holds the position of Professor Catedrático. He is also currently a Director of INESC Porto, an advanced research institute in Portugal. He has authored many papers and been responsible for many projects in areas related with the application of computational intelligence to power systems.

João Gama received the Licenciado degree in electrical and computer engineering from the Faculty of Engineering of the University of Porto, Porto, Portugal (FEUP), and the Ph.D. degree in computer science from the Faculty of Sciences in 2000.

He joined the Faculty of Economy (FEP), where he holds the position of Assistant Professor. He is also a Senior Researcher at the Laboratory for Artificial Intelligence and Decision Aid (LIAAD), a group belonging to INESC Porto LA. He has worked in projects and authored papers in areas related to machine learning, data streams, and adaptive learning systems.

Dr. Gama is a member of the editorial board of international journals in his area of expertise.