NEURAL COMPUTING
NCRG
RESEARCH GROUP

Neural Computing Research Group
Dept of Computer Science & Applied Mathematics
Aston University
Birmingham B4 7ET
United Kingdom
Tel: +44 (0)121 333 4631
Fax: +44 (0)121 333 4586
http://www.ncrg.aston.ac.uk/

Neural Network Based Wind Vector Retrieval from Satellite Scatterometer Data

Dan Cornford, Ian T. Nabney, Christopher M. Bishop[a]
d.cornford@aston.ac.uk

Technical Report NCRG/99/003 January 26, 1999

Accepted *Neural Computing and Applications*

[a]Microsoft Research, 1 Guildhall Street, Cambridge CB2 3NH, UK

Abstract

Obtaining wind vectors over the ocean is important for weather forecasting and ocean modelling. Several satellite systems used operationally by meteorological agencies utilise scatterometers to infer wind vectors over the oceans. In this paper we present the results of using novel neural network based techniques to estimate wind vectors from such data. The problem is partitioned into estimating wind speed and wind direction. Wind speed is modelled using a multi-layer perceptron (MLP) and a sum of squares error function. Wind direction is a periodic variable and a multi-valued function for a given set of inputs; a conventional MLP fails at this task, and so we model the full periodic probability density of direction conditioned on the satellite derived inputs using a Mixture Density Network (MDN) with periodic kernel functions. A committee of the resulting MDNs is shown to improve the results.

Keywords: Conditional Probability Density Estimation, Mixture Density Network, Multi Layer Perceptron, Periodic Variables, Wind vectors, Scatterometer.

1 Introduction

Obtaining wind vectors over the ocean is important to Numerical Weather Prediction (NWP) since the ability to produce a forecast of the future state of the atmosphere depends critically on knowing the current state accurately [Haltiner and Williams, 1980], particularly since the system is non-linear. However, the observation network over the oceans (especially in the southern hemisphere) is very limited [Daley, 1991]. Thus it is hoped that the global coverage of ocean wind vectors provided by satellite borne scatterometers will improve the accuracy of numerical weather forecasts by providing better initial conditions [Harlan and O'Brien, 1986; Lorenc *et al.*, 1993]. The scatterometer data also offers the ability to improve wind climatologies over the oceans [Levy, 1994] and the possibility of studying, at high resolution, interesting meteorological features such as cyclones [Dickinson and Brown, 1996].

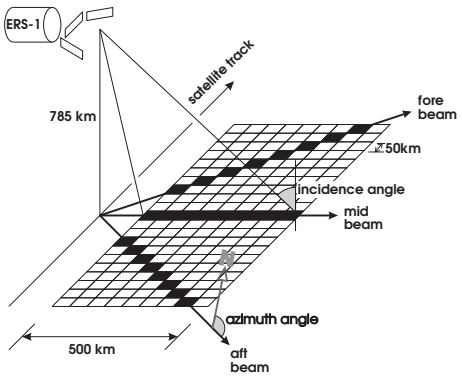


Figure 1: Schematic illustration of the geometry of the ERS-1 satellite showing the footprints of the three radar scatterometers.

The ERS-1 satellite was launched in July 1991 by the European Space Agency [Offiler, 1987]. The on-board microwave radar operates at 5.3 GHz and measures the amount of backscatter generated by small ripples on the ocean surface of around 5 cm wavelength, although this depends on the incidence angle of the radar beam. Measured backscatter from the ocean surface is given as the Normalised Radar Cross Section, and generally denoted by σ^o, which has units of decibels. A 500 km wide swathe is swept by the satellite along the track of its polar orbit, with nineteen cells sampled across the swathe, each cell having dimensions of roughly 50 by 50 km (see Figure 1). Thus there is some overlap between cells. Also, each cell is sampled from three different directions by the fore, mid and aft beams respectively giving a triplet of observations, $(\sigma^o_f, \sigma^o_m, \sigma^o_a)$. This $\boldsymbol{\sigma^o}$ triplet, together with the incidence angle of the mid-beam (which varies across the swathe) can be used to determine the average wind vector within the cell [Offiler, 1994].

Many methods to compute wind vectors from scatterometer data exist. Most have considered model based techniques [Offiler, 1994; Wentz, 1991; Stoffelen and Anderson, 1992; Stoffelen and Anderson, 1997] where a physically based mapping from wind vectors to $\boldsymbol{\sigma^o}$ is formulated. In [Thiria *et al.*, 1993] the mapping from $\boldsymbol{\sigma^o}$ to wind vectors was modelled using simulated data and a neural network based classifier, which gave probabilities of the wind direction being in each of thirty-six intervals. Simulated data was used since real $\boldsymbol{\sigma^o}$ measurements were not available at the time the work was undertaken.

While the outputs of the networks in [Thiria *et al.*, 1993] were interpreted as probabilities they are not strictly such since they can take negative values and are not required to sum to one. The wind direction network has 30 inputs, 2 hidden layers each of 25 units and 36 output units giving a total of 2361 weights. With a training set of size approximately 6000 observations there is considerable danger of over-fitting [Bishop, 1995]. Despite this the networks in [Thiria *et al.*, 1993] appear

to have performed very well on the simulated data. In this study we use a neural network to estimate the full local conditional probability density of the wind direction given $\boldsymbol{\sigma^o}$ in a simple and well-founded manner [Bishop and Nabney, 1996].

2 The Geophysical Model

Much effort has been put into understanding the theoretical relationship between σ^o and wind direction [Wentz, 1991; Stoffelen and Anderson, 1997]. This has been based on studies of the physical processes that govern backscattering from water surfaces [Ebuchi *et al.*, 1993] together with a statistical analysis of the relation between wind vectors (both buoy observed and NWP derived) and scatterometer measurements [Offiler, 1994]. From these studies empirical forward models between single σ^o's and relative wind direction (ϑ) have been established of the general form

$$\sigma^o \sim b_0 + b_1 cos(\vartheta) + b_2 cos(2\vartheta) \tag{1}$$

where the coefficients are complicated functions of the scatterometer incidence angle (θ) and the wind speed ($\|u\|$). The most widely used, and currently operational, forward model is known as CMOD4 [Offiler, 1994; Stoffelen and Anderson, 1997]. There are three σ^o measurements for each cell and these together define a self-intersecting cone-like manifold in 3 dimensional space, which has been shown to approximate a Lissajous curve [Thiria *et al.*, 1993]. For most $\boldsymbol{\sigma^o}$ triples, which are observed with noise, there is ambiguity over the optimal direction to select (see Figure 2).

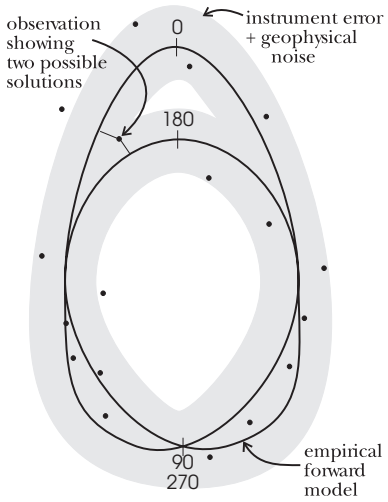
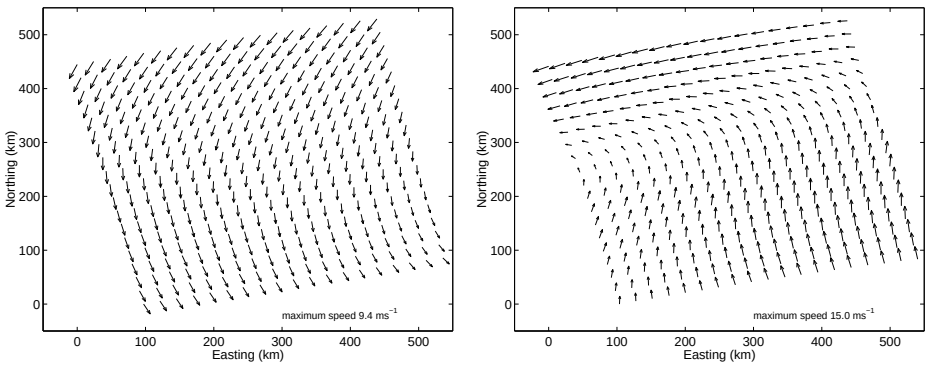


Figure 2: Sketch of a cross section (at constant wind speed) of the 2D manifold (embedded in a 3D space) of the mapping from $(\sigma_1^o, \sigma_2^o, \sigma_3^o)$ to direction (θ). The solid line gives the empirical forward model $\theta \rightarrow \boldsymbol{\sigma^o}$ (e.g. CMOD4) while the grey area gives an estimate of the uncertainty due to instrument error and geophysical noise. Example observations are plotted as black dots, with one labelled to show that there is generally more than one possible wind direction for a given σ^o triplet.

This is typical of many inverse problems in the geophysical sciences, where the forward model output (i.e. $\boldsymbol{\sigma^o}$ as a function of wind direction) is uni-valued for a given set of inputs but the inverse model (i.e. wind direction as a function of $\boldsymbol{\sigma^o}$) is multi-valued. It is known that the relation between wind speed and σ^o is uni-valued [Thiria *et al.*, 1993]. Since the wind speed is largely uncorrelated with the wind direction *relative* to the satellite azimuth[1] angle, the problem

[1]The azimuth angle gives the clockwise angle from North of the scatterometer beam incident on the cell.

of modelling wind vectors can be split into modelling the speed and direction separately.



(a) (b)

Figure 3: Two scenes showing the target wind vectors for (a) a gradient in wind speed and direction and (b) a cyclonic circulation.

Operationally, the problem of obtaining local wind directions from scatterometer data is resolved using the CMOD4 forward model and minimising some cost function (which is typically a mean square error) between the observed $\boldsymbol{\sigma^o}$ triplets and the manifold defined by CMOD4. In general up to four valid solutions are obtained (although there are often two dominant modes with approximately 180 degree ambiguity — the true and alias solutions). Disambiguation methods [Chelton *et al.*, 1989; Schultz, 1990; Shaffer *et al.*, 1991] (such as smoothing filters) are then applied globally to decide which local direction is to be selected, often based on the spatial correlation present in wind fields (see examples in Figure 3). Ambiguity removal is not discussed in this paper but will be addressed in future work[2]. An advantage of our probabilistic models is that this allows Bayesian methods to be applied to the disambiguation problem. Here we consider only the local prediction of wind speed and direction given the local $\boldsymbol{\sigma^o}$ observation.

3 Neural Networks for Modelling Scatterometer Data

Many applications of neural networks can be formulated in terms of a multivariate non-linear mapping from an input vector $\boldsymbol{x}$ to a target vector $\boldsymbol{t}$. A conventional neural network approach, based on a least squares error function, for example, leads to a network mapping which approximates the regression (i.e. the conditional average) of $\boldsymbol{t}$ given $\boldsymbol{x}$. However, for mappings which are multi-valued, such as wind direction in this application, this approach breaks down, since the average of two solutions is not necessarily a valid solution.

This problem can be resolved by recognising that the conditional mean is just one aspect of a more complete description of the relationship between input and target, obtained by estimating the full conditional probability density of $\boldsymbol{t}$ conditioned on $\boldsymbol{x}$, written as $p(\boldsymbol{t}|\boldsymbol{x})$. The least squares approach then corresponds to maximum likelihood for the special case in which $p(\boldsymbol{t}|\boldsymbol{x})$ is modelled by a Gaussian distribution which is spherically symmetric in $\boldsymbol{t}$-space and which has an $\boldsymbol{x}$-dependent mean and constant variance. The mapping from $\boldsymbol{\sigma^o}$ to direction will typically be multi-valued and thus our model of $p(\vartheta|\boldsymbol{\sigma^o})$ cannot be modelled by a uni-modal distribution; instead we show how

[2] More details of the project this work forms part of can be found at `http://www.ncrg.aston.ac.uk/Projects/NEUROSAT/`.

mixtures of uni-modal distributions can be used. The wind speed mapping, on the other hand, is uni-valued and can be sensibly modelled using a regression approach, as outlined below. Thus we address prediction in $(\|u\|, \vartheta)$ space rather than using the Cartesian vector components.

3.1 Neural Networks for Modelling Wind Speed

Our model for predicting wind speed is a fully connected two layer multi-layer perceptron with sigmoidal activation functions in the hidden layer and exponential units in the output layer, to ensure that only positive speeds are generated:

$$z_j = g\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i\right), \tag{2}$$

with the output layer:

$$\|u_p\| = exp\left(\sum_{j=0}^{H} w_j^{(2)} z_j\right), \tag{3}$$

where $\|u_p\|$ is the predicted wind speed, H is the number of hidden (sigmoidal) units z_j, d is the number of inputs x_i (for our networks this was four - the $\boldsymbol{\sigma^o}$ triplet and the mid-beam incidence angle). The w's represent the weights of the network with $w_{ji}^{(1)}$ being the first layer weights from the ith input to the jth hidden unit and $w_j^{(2)}$ being the second layer weight from the jth hidden unit to the output (wind speed). $w_{j0}^{(1)}$ and $w_0^{(2)}$ are the bias parameters for the hidden and output units respectively. The function g is the sigmoid function:

$$g(a) = \frac{1}{1 + \exp(-a)}. \tag{4}$$

The network was trained using a sum of squares error function:

$$E_{\|u_p\|} = \frac{1}{2}\sum_{k=1}^{n} \left(\|u_p\|_k - \|u\|_k\right)^2 \tag{5}$$

where n is the number of observations in the training set, $\|u_p\|_k$ is the output of the network for the kth example for the training set and $\|u\|_k$ is the kth target value; that is the observed speed for the kth example for the training set. Back-propagation (to determine the gradient of the error function with respect to the network weights) together with a conjugate gradient optimisation algorithm was applied to determine the optimal weights in the networks. Only 500 iterations of this algorithm were required for convergence for all numbers of hidden units investigated. Early stopping using independent training and validation sets [Bishop, 1995, Section 9.2.4] reduced the possibility of over-fitting and different numbers of hidden units were investigated.

In order to assess the degree of non-linearity in the wind speed retrieval problem, linear and quadratic regression models of the form

$$\|u_p\| = \exp\left(w_0 + \sum_{j=1}^{P}\sum_{i=1}^{d} w_{ij} x_i^j\right), \tag{6}$$

where P is the order of the polynomial (1 or 2) and d is as before, were also tested on the same datasets. The parameters were computed using standard least squares estimation on the training set [Press *et al.*, 1992].

3.2 Neural Networks for Modelling Wind Direction

When modelling wind direction some care must be taken since not only is the target variable multi-valued, it is also periodic. This section describes one method of estimating the conditional (probability) density of periodic variables: see also [Bishop and Nabney, 1996] for computational details.

3.2.1 Density Estimation for Periodic Variables

A commonly used technique for *unconditional* density estimation is based on mixture models of the form

$$p(\boldsymbol{t}) = \sum_{i=1}^{l} \alpha_i \phi_i(\boldsymbol{t}), \tag{7}$$

where α_i are the mixing coefficients, and the l component functions, or kernels, $\phi_i(\boldsymbol{t})$, are typically chosen to be Gaussians [Titterington *et al.*, 1985; McLachlan and Basford, 1988]. In order to turn this into a model for *conditional* density estimation, the mixing coefficients, as well as any adaptive parameters in the component densities, are set to be functions of the input vector $\boldsymbol{x}$:

$$p(\boldsymbol{t}|\boldsymbol{x}) = \sum_{i=1}^{l} \alpha_i(\boldsymbol{x}) \phi_i(\boldsymbol{t}|\boldsymbol{x}). \tag{8}$$

These functions are likely to be non-linear, so we set the mixing coefficients and kernel parameters from the outputs of a neural network which takes $\boldsymbol{x}$ as input. This underlies the 'mixture of experts' model [Jacobs *et al.*, 1991] and has also been considered by a number of other authors [Bishop, 1994; Liu, 1994].

In this section two methods for modelling the conditional density $p(\vartheta|\boldsymbol{x})$ of a periodic variable ϑ conditioned on an input vector $\boldsymbol{x}$ are reviewed. Both methods use the same kernel functions (and output error function) but allow different sets of parameters to be varied.

3.2.2 Circular Normal Densities

By using a mixture of kernel functions in (8) which are periodic themselves the overall conditional density function will be periodic. The kernel function of the wind direction ϑ is given by:

$$\phi_i(\vartheta) = \frac{1}{2\pi I_0(m_i)} \exp\{m_i \cos(\vartheta - \psi_i)\}, \tag{9}$$

which is known as a *circular normal* or *von Mises* distribution [Mardia, 1972]. The normalisation coefficient is expressed in terms of the zeroth order modified Bessel function of the first kind, $I_0(m_i)$, and the parameter m_i is analogous to the inverse variance parameter in a conventional normal distribution. The parameter ψ_i is the mean of the density function.

A multi-layer perceptron, with a single hidden layer of sigmoidal units (2) and linear output units, is used to set the parameters in the mixture model (8) and (9). The linear outputs are given by:

$$z_k^o = \sum_{j=0}^{H} w_{kj}^{(2)} z_j, \tag{10}$$

where H is the number of hidden units and $w_{kj}^{(2)}$ weight from the jth hidden unit (z_j) to the kth linear output. We divide the network outputs into three classes, corresponding to mixing coefficients, means and variances and apply a suitable transform to each class of output.

In order to ensure that the mixture model in (8) is a probability density function, it is sufficient that the mixing coefficients $\alpha_i(\boldsymbol{x})$ satisfy the constraints:

$$\sum_{i=1}^{l} \alpha_i(\boldsymbol{x}) = 1, \qquad 0 \leq \alpha_i(\boldsymbol{x}) \leq 1, \tag{11}$$

for all $\boldsymbol{x}$. This can be achieved by choosing the α_i to be related to the corresponding network outputs by a normalised exponential, or *softmax* function [Jacobs *et al.*, 1991]:

$$\alpha_i = \frac{\exp\left(z_i^{\alpha}\right)}{\sum_{j=1}^{l} \exp\left(z_j^{\alpha}\right)}, \tag{12}$$

where z_j^{α} represent the corresponding network outputs (i.e. the component of z_k^o which represents the mixing coefficients). The centres $\psi_i(\boldsymbol{x})$ of the kernel functions are represented directly by the network outputs since these may take any value over the reals. This is also motivated by the corresponding choice of an uninformative Bayesian prior, assuming that the relevant network outputs have uniform probability distributions [Jacobs *et al.*, 1991; Berger, 1985]. The inverse variance parameters $m_i(\boldsymbol{x})$ of the kernel functions are *scale* parameters and so it is convenient to represent them in terms of the exponentials of the corresponding network outputs. This ensures that $m_i(\boldsymbol{x}) > 0$ and discourages $m_i(\boldsymbol{x})$ from tending to 0, which corresponds to a pathological solution. Again, it can be motivated by the concept of an uninformative prior in the Bayesian framework.

3.2.3 Expansion in Fixed Kernels

The other technique used in this paper involves a conditional density model as in (8) consisting of a fixed set of periodic kernels, again given by circular normal functions as in (9). In this case the mixing coefficients alone are determined by the outputs of a neural network (through a *softmax* activation function (12)) and the centres ψ_i and scale parameters m_i are fixed. We selected a uniform distribution of centres, and set $m_i = m$ for each kernel, where the value for m was chosen to give moderate overlap between the kernel functions. Fixed kernels are only appropriate for targets from low dimensional spaces since the number needed grows exponentially with the dimension of the target space.

3.2.4 Computational Details

The $\boldsymbol{\sigma^o}$ triple together with the (mid-beam) incidence angle and wind speed predicted by (3) were used as inputs to the mixture density network (i.e. $\boldsymbol{x} = (\boldsymbol{\sigma^o}, \theta, \|u_p\|)$). The use of incidence angle as an additional input to our neural network makes our model more flexible than those in [Thiria *et al.*, 1993] where a separate network was trained for each of the 10 incidence angles (they considered only every other cell). Wind speed was included as an input since [Thiria *et al.*, 1993] suggested that the relationship between direction and $\boldsymbol{\sigma^o}$ is dependent on the wind speed.

For both models the adaptive parameters of the model (the weights and biases in the network) are optimised by maximising the likelihood of the data given the model. In practice it is convenient to minimise an error function E given by the negative logarithm of the likelihood function. Derivatives of E with respect to the network weights can be computed using the rules of calculus [Bishop and

Nabney, 1996], and these derivatives can then be used with standard optimisation procedures to find a minimum of the error function. In such non-linear problems care must be taken to ensure good initialisation of the model parameters to avoid bad local minima. In this case we initialise the network weights so that the centres are approximately evenly spaced, the scale parameters are large enough to ensure overlapping of the kernel functions and the mixing coefficients are approximately equal. These are, however, set with a small random component so that we can investigate the effect of different initialisations.

A conjugate gradient algorithm was used to minimise the mixture density error function. Early stopping [Bishop, 1995] was used to ensure reasonable generalisation performance. The network that was used on the test set was the one with the lowest validation set error E_v. Generally after around 1200 iterations the early stopping rule had selected the best fully adaptive Circular Normal (CN) network, and this was always attained within 2000 iterations. When training the Fixed Kernel (FK) networks convergence was much quicker, usually occuring within 250 iterations. Several network architectures were investigated by varying the number of hidden units and the number of circular normal functions used. When using fixed kernels the scale parameter m was also varied.

Once the networks were trained, a committee of networks [Bishop, 1995] was constructed which combined the predictive conditional densities $p(\vartheta|\boldsymbol{x})$ from several models. The networks forming the committee were weighted equally since all had similar accuracies.

4 Evaluation Function

In order to compare different models a *Figure of Merit* (FoM) evaluation function has been proposed by David Offiler of the UK Meteorological Office [Stoffelen and Anderson, 1997]. The FoM reflects the extent to which the transfer model meets the design specifications of $\pm 2\ ms^{-1}$ for wind speed and $\pm 20^o$ for wind direction. The FoM is computed over the $4 - 24\ ms^{-1}$ wind speed range. A FoM of greater than one indicates the transfer function is performing to within these specifications, although the exact form is rather ad-hoc. The FoM details can be found in Appendix I. The FoM is considered in both weighted and unweighted forms, which take into account the performace of the algorithm at different wind speeds. If we wish to perform well on the sum of the weighted and unweighted evaluation functions then we must either explicitly adapt the cost function in the network training (by weighting the error function by a factor depending on the wind speed class) or carefully select the training set. In this study we chose the latter option.

The wind direction used when computing the FoM was chosen using a very simple disambiguation algorithm. Initially the most likely direction is selected. If this is more than $\pm 90^o$ from the observed direction the second most likely direction is chosen. If this is still more than $\pm 90^o$ from the observed direction then the third most likely direction is chosen and so on until the first four modes have been considered. This must be done to allow a sensible value of the direction errors to be calculated in the absence of a more sophisticated ambiguity removal algorithm. The algorithm used here is not applicable in practice since it requires knowledge of the target values, and is simply used to compare the performance of 'local models'.

5 Data

The data used in this study was compiled by the European Space Agency in collaboration with the UK Meteorological Office. The database consisted of 115 scenes, each of which contains 19 by

19 cells, or observation locations, corresponding to a square of area roughly 500 by 500 km. Two examples can be seen in Figure 3. The scenes were classified into 6 cases:

- low wind speeds (10 scenes)
- homogeneous cases (15 scenes)
- gradients in speed or direction (59 scenes)
- cyclonic circulations (10 scenes)
- anti-cyclonic circulations (11 scenes)
- fronts and other 'difficult' cases (10 scenes)

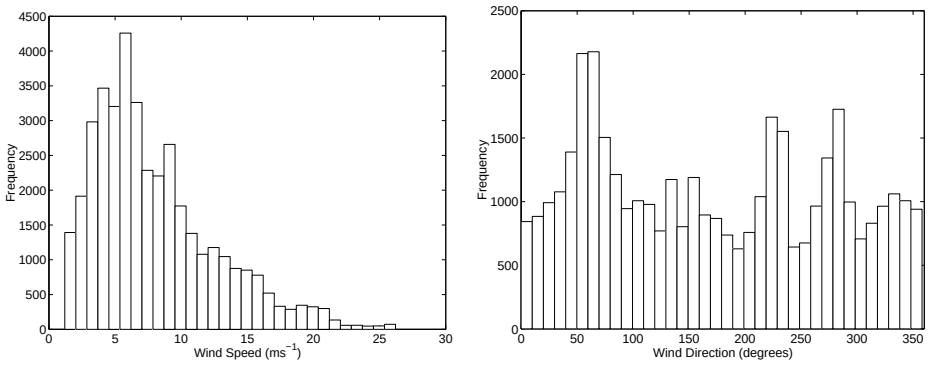
This classification was made by meteorologists. Each of the 39,611 observations in this dataset contained information on its location, the $\boldsymbol{\sigma^o}$ triple, incidence and azimuth angles, wind speed and wind direction. In this work all wind directions were computed relative to the azimuth angle, so the wind direction in the training data does not give the absolute direction, but rather the relative direction. The wind speed and direction were obtained from the UK Meteorological Office Unified Model[3], which has a horizontal resolution of approximately 150 km [Milton and Wilson, 1996]. Thus the targets (i.e. wind speed and direction) are interpolated from a model designed to represent the large scale dynamics of the atmosphere [Haltiner and Williams, 1980]. This means there is likely to be considerable smoothing of the target values.

We use NWP model wind fields since these are the best estimates of the true winds available over the ocean surface that are collocated with the scatterometer observations. Care must be taken since there is a danger in using one model to define the targets for another model. However the NWP model derived estimates of wind speed and direction are based on assimilated data which combines those indepedent observations available from bouys and ships with a good forecast from the previous state of the atmosphere, and represents the best available estimate of the true winds. Due to the sparsity of observations over the oceans the forecast state dominates the analysis and the location and intensity of features in the wind fields may be in error, thus the quality of the targets was improved by matching the position of low pressure centres in the forecast winds with those observed by the scatterometer by linearly translating the model wind fields.

Despite having nearly 40,000 observations, there are in reality far fewer truly independent observations, since the within-scene spatial correlations between both wind speed and wind direction will be very high. This implies that care must be taken when selecting data to train, validate and test the neural network models. The local models we are training consider only the information from within the relevant cell to infer wind speed and direction. Thus to retrieve all speeds and directions well we require a training set with all possible combinations of wind speed, direction and beam incidence angle represented. In order to retain some totally independent data for testing 13 scenes were selected from the whole dataset and removed from further analysis. The remaining 102 scenes were used for parameter estimation.

Figure 4 shows the distribution of the target values in these 102 scenes. Wind speed has a distribution strongly skewed towards lower speeds, which reflects the distribution of surface wind speeds in the atmosphere. We wish to train our networks to learn the transfer function at all speed ranges to minimise the FoM evaluation function. Thus when selecting a subset of 3,000 observations to train the networks, 1,500 observations were selected to give as uniform a distribution of wind speed as possible. A further 1,500 randomly chosen observations were also selected. This implies that equal weight is given to the weighted and unweighted FoM. The validation set, which was used in the

[3]See *http://www.meto.govt.uk/sec5/NWP/NWP.html.*



(a) (b)

Figure 4: Histograms showing the distribution of (a) wind speed and (b) direction.

early stopping procedure, was selected in a similar manner. Finally a test set of 3,000 observations was chosen randomly from the test scenes. All variables (except wind speed and direction) were linearly transformed to have zero mean and unit variance, using the mean and standard deviation derived from the direction training dataset. All results in this paper refer to the test set.

6 Results

6.1 Wind Speed

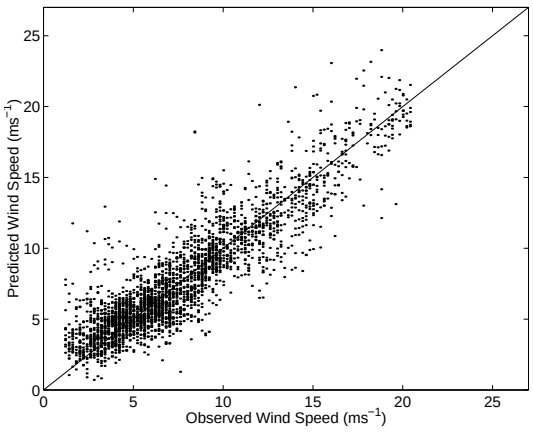


Figure 5: Scatter-plot of predicted versus observed wind speeds using the 4 hidden unit multilayer perceptron.

The relationship between wind speed and σ^o is the less challenging problem and can be approached using standard regression techniques. Figure 5 shows a scatter-plot of the results for the 4 hidden

unit network, demonstrating that a reasonable approximation is made at all wind speeds, although there are some large residuals. The wind speed results are also computed for different wind speed bins, as suggested in the section on evaluation functions.

Table 1: Results for the 4 hidden unit multi-layer perceptron, binned by wind speed and a linear regression model. The regression model included terms up to quadratic in the variables but no interaction terms. All figures given in ms^{-1}.

Speed Range	Bias	SD[a]	RMSE[b]	N[c]
< 4	1.37	1.71	2.19	512
$4 - 8$	0.18	1.43	1.44	1324
$8 - 12$	0.05	1.85	1.86	703
$12 - 16$	-0.74	2.13	2.25	324
$16 - 20$	-0.30	2.17	2.19	126
> 20	-0.94	1.23	1.55	11
whole test set	0.23	1.80	1.81	3000
optimal regression model	-0.14	1.85	1.85	3000

[a]Standard Deviation
[b]Root Mean Square Error
[c]Number of observations

From Table 1 it is clear that the network is having some difficulty learning the transfer function for speed at higher and lower wind speeds (shown especially in the large biases). This is a feature common to all reported transfer models [Offiler, 1994; Wismann, 1992] and the regression model. The root mean square error of 1.81 ms^{-1} for the full test set is within the design specification of the instrument of 2 ms^{-1}.

6.2 Wind Direction

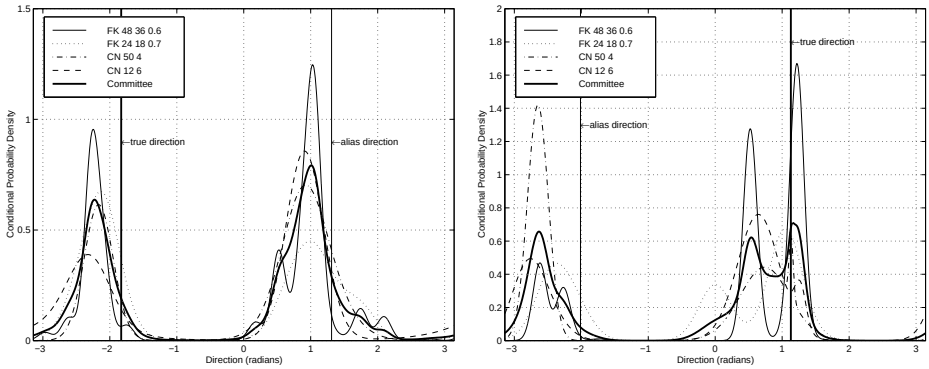
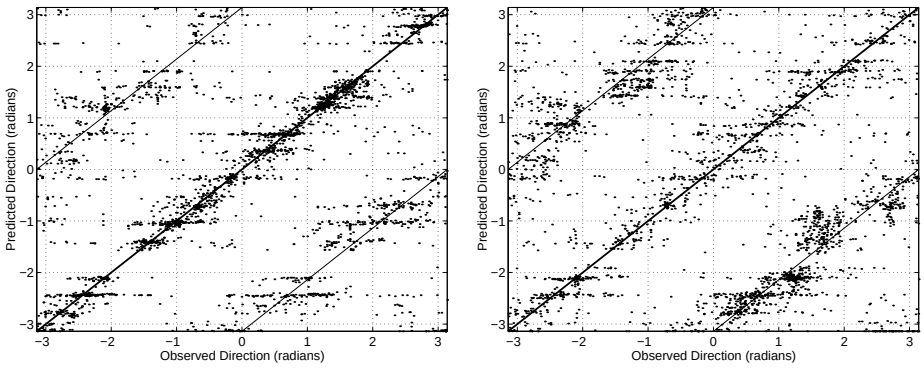


Figure 6: Conditional density functions for 2 cells showing the results of 2 circular normal, 2 fixed kernel and the committee of these 4 models. Both the true and aliased (i.e. incorrect by 180 degrees) targets are shown.

Figure 6 illustrates that the different techniques of using adaptive circular normals and expansion in fixed kernels produces similar conditional densities for the wind direction given the scatterometer inputs. There is however some variability in the results, suggesting that a committee of networks might improve performance. In both cases there is good agreement between the observed direction



(a) (b)

Figure 7: Scatter-plots of observed versus predicted direction for (a) the most likely and (b) the second most likely solutions. The lines show perfect fit and 180^o alias solutions.

and the predicted conditional densities. Figure 7 shows the results of the committee of networks, choosing the most likely and second most likely vectors (since generally there are two dominant solutions; the true one and its 180 degree alias). Also marked on Figure 7 is the line of correct solutions (observed = predicted) in the centre, and the two lines (observed = predicted ± 180 degrees) that are the alias solutions. The data clusters around these lines in both figures, although there is considerable scatter.

Table 2: Results of the network techniques applied to the direction data. Columns give the percentage of observations within $\pm 20^o$ of the target solution considering the most likely, first and second most likely and the first four most likely directions. For each example the best fitting solution is chosen.

Network Configuration / Technique	1 solution	2 solutions	4 solutions
FK 24 18 0.7[a]	53.0	72.0	79.0
FK 48 36 0.6	51.7	70.9	86.5
CN 12 6[b]	47.2	70.8	77.5
CN 50 4	45.1	72.9	78.7
4 net committee	53.8	74.2	84.0

[a]FK = fixed kernel, number of hidden units, number of fixed kernels, scale of kernels relative to inter-kernel spacing.

[b]CN = circular normals, number of hidden units, number of adaptive kernels.

Table 2 shows selected results for two of the better fixed kernel and circular normal approaches as well as the committee results. By considering both first and second solutions and picking the better one, using the circular normal technique we obtain the correct solution within 20 degrees more than 70% of the time. The committee of networks outperform all their members when considering only one or two solutions, however when considering the four most likely solutions some of the fixed kernel results are marginally better.

Table 3: Comparison of results using different algorithms (results computed in speed range the 4–24 ms^{-1}). Note that the final CMOD4 result uses a completely different dataset, and that the other studies used a different ambiguity removal procedure than was used in this study. All units ms^{-1} (speed) and ^{o} (direction).

Method Used	Speed Bias	Speed SD[a]	Dir. Bias	Dir. SD
Neural networks - this study	-0.01	1.73	0.73	23.05
CMOD4 - same data	0.3	1.8	–	–
Subset of data [26]	-0.44	∼2	-1.37	∼20
CMOD4 [8]	0.1	1.9	-1.6	17.0
CMOD4 [11]	0.06	1.65	0.76	16.69

[a]Standard Deviation

7 Discussion and Conclusions

The equivalent results for the operational CMOD4 algorithm are only available for an (unknown) subset of the dataset we used, excluding those cases with wind speeds < 4 ms^{-1}, and are shown in Table 3. These may be comparable with our results since we have used a representative sample of the full (40,000 observation) dataset in testing.

Table 3 shows the limited number of comparable results published; however CMOD4, being the operational algorithm, can be used as a benchmark. The neural networks used in this study produced comparable results to CMOD4 on speed, however they did not perform as well on direction. Given that the estimated noise on the wind speed targets is of the order of 2 ms^{-1} the speed results may well represent the best that can be achieved given the data available, and fall within the specification of the instrument, which required less than 2 ms^{-1} root mean square error. It must be noted that the neural networks trained in this study were trained on a combination of uniformly and randomly (i.e. as the data) distributed wind speed data, and thus are unlikely to produce an optimal solution over the observed distribution of wind speeds. This was done so that the networks would minimise the *Figure of Merit* evaluation function. However, the almost uniform distribution of relative wind direction means there should be no such problems with respect to wind direction.

Results for the neural network approach to wind direction retrieval proposed in [Thiria *et al.*, 1993], are not directly comparable since they used simulated data as well as a spatial input context. They obtained 72% of the first solution within 20 degrees and 98.4% of the first two solutions within 20 degrees. Our committee of networks obtained the correct solution to within 20 degrees roughly 75% of the time considering the two most likely directions only. In further work we have performed using a different training set based on co-located scatterometer $\boldsymbol{\sigma^o}$ triples and wind vectors from the ECMWF numerical weather prediction model the same committee produced results of 77.4% within 20^o of the true direction, taking the first two most likely solutions. A more representative data set, with far more patterns in the training set is likely to improve results. The more simple CMOD4 forward model was trained using almost 40,000 observations.

Our techniques produced *Figure of Merit* scores of 1.08 (weighted average over the speed bins) and 1.16 (unweighted), which compare with the CMOD4 scores of 1.13 (weighted) and 1.16 (unweighted). These scores cannot be directly compared since they were computed from different data sets, however there is some potential benefit in the neural network approach. It would be considerably less computer intensive since once the networks are trained, new wind vectors can be computed using forward propagation in the model, whereas the current operational models require the minimisation of a complex function or the use of large look-up tables.

It is clear that the use of the circular normal distribution with all parameters (or just mixing coefficients) determined by neural networks represents a feasible method for the solution of inverse problems (i.e. multi-valued mappings) where the target is periodic. The ability to approximate the conditional probability of wind direction allows us to use sophisticated disambiguation algorithms. There is some concern that the neural networks were having difficulty in learning the appropriate mapping because they were being trained on very noisy data which did not cover the full range of input values. The satellite scatterometer measurements are prone to noise from larger ocean waves (swell), wave breaking and rainfall [Badran *et al.*, 1991]. Also, the target data from numerical weather prediction models is over-smooth, particularly in the vicinity of fronts and strong gradients in wind speed or direction. Many of these problems have yet to be quantified in terms of their impact on wind retrieval.

Future work will obtain better target data (increased numbers of observations from which to select a data set which covers a broader input and target range and increased data quality through manual removal of extreme outliers in $\boldsymbol{\sigma}^o$ space) and investigate modelling the conditional distribution in terms of the Cartesian wind components, where the error model is better understood.

Acknowledgements

We are grateful to the European Space Agency and the UK Meteorological Office for making available the ERS-1 data. The authors wish to thank David Offiler of the UK Meteorological Office for his input. This work is funded by the European Union under contract ENV4–CT96–0314, and forms part of the NEUROSAT project. MATLAB software to perform the computations outlined in this project is available from `http://www.ncrg.aston.ac.uk/ netlab/index.html`.

References

Badran, F., S. Thiria, and M. Crepon 1991. Wind Ambiguity Removal by the Use of Neural Network Techniques. *Journal of Geophysical Research* **96**, 20521–20529.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis* (Second ed.). New York: Springer.

Bishop, C. M. 1994. Mixture density networks. Technical Report NCRG/4288, Neural Computing Research Group, Aston University, U.K.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bishop, C. M. and I. T. Nabney 1996. Modelling Conditional Probability Distributions for Periodic Variables. *Neural Computation* **8**, 1123–1133.

Chelton, D. B., M. H. Freilich, and J. R. Johnson 1989. Evaluation of Unambiguous Vector Winds from the SeaSat Scatterometer. *Journal of Atmospheric and Oceanic Technology* **6**, 1024–1039.

Daley, R. 1991. *Atmospheric Data Analysis*. Cambridge: Cambridge University Press.

Dickinson, S. and R. A. Brown 1996. A Study of Near-Surface Winds in Marine Cyclones Using Multiple Satellite Sensors. *Journal of Applied Meteorology* **35**, 769–781.

Ebuchi, N., H. Kawamura, and Y. Toba 1993. Physical Processes of Microwave Backscattering from Laboratory Wind Wave Measurements. *Journal of Geophysical Research* **98**, 14669–14681.

Haltiner, G. J. and R. T. Williams 1980. *Numerical Prediction and Dynamic Meteorology*. Chichester: John Wiley.

Harlan, J. and J. J. O'Brien 1986. Assimilation of Scatterometer Winds into Surface Pressure Fields Using a Variational Method. *Journal of Geophysical Research* **91**, 7816–7836.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton 1991. Adaptive mixtures of local experts. *Neural Computation* **3**, 79–87.

Levy, G. 1994. Southern-Hemisphere Low-Level Wind Circulation Statistics from the SeaSat Scatterometer. *Annales Geophysicae - Atmospheres, Hydroshperes and Space Sciences* **12**, 65–79.

Liu, Y. 1994. Robust neural network parameter estimation and model selection for regression. In *Advances in Neural Information Processing Systems*, Volume 6, pp. 192–199. Morgan Kaufmann.

Lorenc, A. C., R. S. Bell, S. J. Foreman, C. D. Hall, D. L. Harrison, M. W. Holt, D. Offiler, and S. G. Smith 1993. The Use of ERS-1 Products in Operational Meteorology. *Advances in Space Research* **13**, 19–27.

Mardia, K. V. 1972. *Statistics of Directional Data*. London: Academic Press.

McLachlan, G. J. and K. E. Basford 1988. *Mixture models: Inference and Applications to Clustering*. New York: Marcel Dekker.

Milton, S. F. and C. A. Wilson 1996. The impact of parameterized subgrid-scale orographic forcing on systematic errors in a global NWP model. *Monthly Weather Review* **124**, 2023–2045.

Offiler, D. 1987. Wind Measurements from the Earth Remote-sensing Satellite (ERS-1). *Meteorological Magazine* **116**, 279–285.

Offiler, D. 1994. The Calibration of ERS-1 Satellite Scatterometer Winds. *Journal of Atmospheric and Oceanic Technology* **11**, 1002–1017.

Press, W. H., S. A. Teukolsky, W. T. Vettering, and B. P. Flannery 1992. *Numerical Recipes in C* (2nd Edition ed.). Cambridge, UK: Cambridge University Press.

Schultz, H. 1990. A Circular Median Filter Approach for Resolving Directional Ambiguities in Wind Fields Retrieved from Spaceborne Scatterometer Data. *Journal of Geophysical Research* **95**, 5291–5303.

Shaffer, S. J., R. S. Dunbar, S. V. Hsiao, and D. G. Long 1991. A Median-Filter-Based Ambiguity Removal Algorithm for NSCAT. *IEEE Transactions on Geoscience and Remote Sensing* **29**, 167–174.

Stoffelen, A. and D. Anderson 1997. Scatterometer Data Interpretation: Estimation and Validation of the Transfer Function CMOD4. *Journal of Geophysical Research* **102**, 5767–5780.

Stoffelen, A. and D. L. T. Anderson 1992. ERS-1 Scatterometer Data Characteristics and Wind Retrieval Skill. In B. Kaldeich (Ed.), *Proceedings of the First ERS-1 Symposium - Space at the Service of the Environment*, pp. 41–47. Cannes, France, ESA.

Thiria, S., C. Mejia, F. Badran, and M. Crepon 1993. A Neural Network Approach for Modeling Nonlinear Transfer Functions: Application for Wind Retrieval from Spaceborne Scatterometer Data. *Journal of Geophysical Research* **98**, 22827–22841.

Titterington, D. M., A. F. M. Smith, and U. E. Makov 1985. *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley.

Wentz, F. J. 1991. A Simplified Vector Algorithm for Satellite Scatterometers. *Journal of Atmospheric and Oceanic Technology* **8**, 697–704.

Wismann, V. 1992. A C-Band Wind Scatterometer Model Derived from the Data Obtained During the ERS-1 Calibration / Validation Campaign. In B. Kaldeich (Ed.), *Proceedings of the First ERS-1 Symposium - Space at the Service of the Environment*, pp. 55–59. Cannes, France, ESA.

Appendix I

In order to compare different models a *Figure of Merit* 'evaluation function' has been proposed by David Offiler of the UK Meteorological Office (personal communication). This evaluation function takes the form:

$$FoM = \frac{(F1 + F2 + F3)}{3} \tag{13}$$

where $F1 = 40/(U_{bias} + 10U_{sd} + D_{bias} + D_{sd})$, $F2 = (2/U_{rms} + 20/D_{rms})/2$, $F3 = 4/V_{rms}$, U gives the wind speed, D the wind direction and V the wind vector. This reflects the extent to which the model meets the instrument specifications of $\pm 2 ms^{-1}$ and $\pm 20^{o}$. A FoM of greater than one indicates the transfer function is performing to within these specifications, although this is a rather ad-hoc measure. The bias is given by:

$$U_{bias} = \frac{1}{n} \sum_{i=1}^{n} U_{res(i)} \tag{14}$$

where the residual wind speed $U_{res} = U_{pred} - U_{obs}$, the predicted speed ($U_{pred}$) coming from the neural network, the observed speed (U_{obs}) from the numerical weather prediction model and n is the number of observations. The standard deviation of the residuals is given by:

$$U_{sd} = \sqrt{\left(\frac{1}{n} \sum_{i=1}^{n} (U_{res(i)})^2 \right) - (U_{bias})^2} \tag{15}$$

Similarly the root mean square error U_{rms} is given by:

$$U_{rms} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (U_{res(i)})^2} \tag{16}$$

The vector residual V_{res} is given by:

$$V_{res} = \sqrt{U_{obs}^2 + U_{pred}^2 - 2U_{obs}U_{pred}\cos(D_{res})} \tag{17}$$

The wind speed is known to affect the ability of the on-board instruments to resolve wind direction. Thus the FoM is computed by binning the cases in 5 wind speed classes; $4-8\ ms^{-1}$, $8-12\ ms^{-1}$, $12-16\ ms^{-1}$, $16-20\ ms^{-1}$ and $> 20\ ms^{-1}$. The bias, standard deviation and root mean square error (of the residuals) are computed for each bin and a final FoM is computed from both weighted (by the number of observations in each bin) and unweighted means of the binned statistics. The weighted mean takes into account the distribution of wind speed values in the atmosphere. The unweighted mean gives much larger importance to performance at higher wind speeds, which are arguably the cases of greatest interest to atmospheric scientists.