Dynamical Recurrent Neural Networks and Pattern Recognition Methods for Time Series Prediction: Application to Seeing and Temperature Forecasting in the Context of ESO's VLT Astronomical Weather Station

Alex Aussem †‡, Fionn Murtagh ∓, Marc Sarazin †

†Very Large Telescope Division,

European Southern Observatory, Karl-Schwarzschild-Str. 2,

D-85748 Garching bei München - Germany

‡Université René Descartes, UFR de Mathématiques et Informatique,

45, rue des Saints-Pères, 75006 Paris - France

∓ Space Telescope - European Coordinating Facility,

(Affiliated to Astrophysics Div., Space Science Dept., European Space Agency)

European Southern Observatory, Karl-Schwarzschild-Str. 2,

D-85748 Garching bei München - Germany

Running Title: Neural Networks for Seeing and Temperature Forecasting.

Dynamical Recurrent Neural Networks and Pattern Recognition Methods for Time Series Prediction: Application to Seeing and Temperature Forecasting in the Context of ESO's VLT Astronomical Weather Station

Alex Aussem †‡, Fionn Murtagh ∓, Marc Sarazin †

†Very Large Telescope Division,

European Southern Observatory, Karl-Schwarzschild-Str. 2,

D-85748 Garching bei München – Germany

‡Université René Descartes, UFR de Mathématiques et Informatique,

45, rue des Saints-Pères, 75006 Paris – France

∓ Space Telescope - European Coordinating Facility,

(Affiliated to Astrophysics Div., Space Science Dept., European Space Agency)

European Southern Observatory, Karl-Schwarzschild-Str. 2,

D-85748 Garching bei München – Germany

Abstract

The European Southern Observatory's planned Astronomical Weather Station for the Very Large Telescope which is currently under construction at Cerro Paranal in Chile includes (i) advance temperature prediction, which would permit air conditioning in the telescope enclosure to be preset as a function of the next night's expected temperature; and (ii) prediction of seeing, a few hours in advance, to allow flexible scheduling of the most appropriate instrumentation. Extensive data, collected since 1985, are being used to appraise various methodologies. A recurrent neural network is described, which uses arbitrary time-delayed connections to capture the dynamic of time series. This endows the model with a memory of its previous states. The resulting network is time- and space-recurrent, and generalizes most recurrent architectures. The performance of this network is discussed. The results are compared with the k-nearest neighbors method.

1 Introduction

In this article we discuss work carried out in the area of temperature and seeing prediction at La Silla and Paranal, sites of the European Southern Observatory. The studies currently being undertaken are aimed to be incorporated later into the future VLT Weather Station. Since 1985, a monitoring station has been tracking a range of meteorological and atmospheric variables. The continuous expansion of the automatic data acquisition process has effectively taken place since 1989 and affords an opportunity to assess prediction methods. The practical experience and theoretical knowledge of the meteorologists have limitations, which points to the inherent difficulty of deriving rules for local temperature and seeing forecasting. Therefore, prediction is sought without reference to any underlying physical model. The meteorological prior observations are assumed to carry within them the potential for predicting the future. The use of nonlinear models is strongly motivated by the work carried out by Murtagh et al. (1992) which shows that a 1-nearest neighbor method outperforms the fitting of linear autoregressive models. In this paper, we examine the effectiveness of a dynamic and recurrent connectionist network, to infer the rules that govern the temperature and seeing evolution from the regularities in the past meteorological measurements. With a view to highlighting their differences, the connectionist model is compared with a k-nearest neighbor method.

It is difficult to present the dynamic and recurrent connectionist network, especially when the audience consists of a mixture of astronomers and meteorologists, some of whom will know a great deal about connectionist approaches to prediction, while the knowledge of others is limited in scope. However, we have tried to highlight its specificities in a brief review in which neural networks are analyzed as nonlinear extensions of traditional linear models. Neural networks do not build an explicit description of the multidimensional surface but much of their promise comes from the hope that they can learn to emulate unanticipated regularities. In addition, they cope with seasonality or other periodic behavior while most of the standard models deal with stationary data. On the other hand, connectionist models will not help to interpret the underlying dynamic behavior. We have deliberately left aside a few problems, some of which have practical consequences.

Due to the current amount of data and processing capabilities, it is now becoming feasible to venture into the domain of weak models with many parameters and few assumptions about the underlying physical model. But we have been faced with troublesome issues in order to arrive at data sets on which to carry out assessments, in part from the number of missing or spurious measurements, and in part from the relative utility of the different variables. Extensive exploration was required to remove invalid records on the grounds of missing values.

This paper is organized as follows. Section 2 reviews linear autoregressive moving average models as a point of departure for the use of dynamic and recurrent neural networks. In particular, neural networks are analyzed as nonlinear approximation functions to the theoretical optimal predictor of the time series, given a limited number of past observations. Section 3 examines the issues we faced in building a consistent training set. In addition, a brief statistical analysis of the data is carried out and illustrates the limited dependencies between ground-based meteorological variables. Intermediate results are presented in section 4 for the prediction of temperature and seeing 24 hours ahead at sunset. Due to the encouraging results obtained for temperature prediction at La Silla, we present further experiments in which temperature is predicted at each hour of the day, 24 hours in advance. A variety of analysis tools, including the nearest neighbor approach, are applied in order to make meaningful comparisons among the approaches. In line with the nowcasting study carried out in Murtagh and Sarazin (1993), a fuzzy coding approach is discussed in section 5 for the coding of seeing. In this respect, seeing predictability is found to be successful when meteorological variables are available at the same time point. Finally the effectiveness of the predictive approaches are discussed in section 6 with a view to ascertaining the prediction problems for which each method is most suited.

2 Nonlinear ARMA Models and Optimal Predictors

The statistical approach to time series prediction involves the construction of a stochastic model to forecast the observable using the previous observations. Linear autoregressive moving average models (ARMA) have been proposed for the past decades as the "standard" class (Box and Jenkins, 1976; Priestley, 1981) to model a wide range of different types of time series structures. Here we briefly review the cornerstone of linear ARMA models in order to highlight the role of neural networks in approximating the optimal predictors of nonlinear dynamical systems.

2.1 Optimal Predictors

The theory of prediction focuses on optimum prediction of dynamical systems in the minimum mean square sense. One of the earliest class of models which appeared to offer plausible description of a wide range of different types of time series is the linear ARMA model. If x_t denotes the value of the series recorded at time point t, we may write the ARMA(p,q) process in the form,

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{i=1}^{q} b_i e_{t-i} + e_t$$

where it is assumed that $\mathrm{E}[e_t \mid x_{t-1}, x_{t-2}, \ldots] = 0$. This condition is satisfied for instance when e_t are zero mean, independent and identically distributed and are independent of the past x_t's. Without loss of generality, it is assumed that e_t has zero mean and finite variance σ^2. It may be shown that the minimum mean squared error predictor is given by

$$\hat{x_t} = \mathrm{E}[x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1]$$

when the conditional mean exists. x_1 refers to the first available observation. It may be shown that for the linear ARMA model, the optimal predictor is approximatively given by the recursive formula

$$\hat{x}_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{i=1}^{q} b_i \hat{e}_{t-i}$$

where $\hat{e}_{t-i} = x_{t-i} - \hat{x}_{t-i}$. Unfortunately, many types of series do not conform to this model. A natural generalization of the linear ARMA model to the nonlinear case may be given by

$$x_t = f(x_{t-1}, \ldots, x_{t-p}, e_{t-1}, \ldots, e_{t-q}) + e_t$$

where $f()$ is an unknown function. Such a model is termed a NARMA(p,q) process in the literature (Connor et al., 1994). The conditional mean predictor is given by

$$\hat{x_t} = \mathrm{E}[f(x_{t-1}, \ldots, x_{t-p}, e_{t-1}, \ldots, e_{t-q}) \mid x_{t-1}, x_{t-2}, \ldots, x_1] \tag{1}$$

Suppose that the NARMA is invertible in the sense that there exists a function $g()$ such that

$$x_t = g(x_{t-1}, x_{t-2}, \ldots) + e_t$$

then given the infinite past observations, one can in principle use the above equation to compute the e_{t-j} in (1) as a function the past observations $x_{t-1}, x_{t-2}, \ldots$ such that (1) becomes

$$\hat{x_t} = f(x_{t-1}, \ldots, x_{t-p}, e_{t-1}, \ldots, e_{t-q}) \tag{2}$$

where the e_{t-j} are specified by $x_t - g(x_{t-1}, x_{t-2}, \ldots)$. The predictor has a mean-squared error σ^2. Since in practice one has only access to a finite observation record, e_{t-j} cannot be computed exactly; it seems reasonable, as for the linear ARMA process, to approximate the conditional mean predictor (2) by the following formula

$$\hat{x_t} = f(x_t, \ldots, x_{t-p}, \hat{e}_{t-1}, \ldots, \hat{e}_{t-q}) \tag{3}$$

Such a general NARMA scheme provides a valuable framework for modeling a wide range of different types of nonlinear time series structures.

2.2 State Space Paradigm

Chaotic theory also advocates the use of time-lagged vectors to infer the future evolution of time series. Recent studies (Ott, 1993; Sauer, 1993), to cite just a few, indicate that chaotic dynamics are quite common, and that in many cases random behavior is due to low dimensional chaos rather than complicated dynamics involving many irreducible degrees of freedom. When the dynamic of a deterministic chaotic process collapses onto a low dimensional attractor, a relatively small number of parameters can theoretically uniquely specify a distinguishable state of the system which in turn entirely determines the future evolution of the trajectory. The promise of the delay-coordinate embedding technique (Sauer, 1993) is that information about the future can be inferred by studying the evolution of the observable. Given a delay time τ and a dimension d, the *delay coordinate vector* is defined in the Euclidean space as

$$\mathbf{v}_t = (x_t, x_{t-\tau}, \ldots, x_{t-(d-1)*\tau})$$

The central result proven by Takens (Ott, 1993) asserts under reasonable conditions that the correspondence relating $\mathbf{v}_t$ to the state of the system is a smooth local invertible change of coordinates (i.e. it is a diffeomorphism) as long as d is twice as much as the dimension of the attractor. In the light of this, we can in principle reduce the problem of prediction for a complex dynamical system whose motion lies in a low-dimensional attractor to a much simpler forecasting problem and speculate that there exists a nonlinear autoregression (NAR) of the form

$$x_{t+1} = f(x_t, x_{t-1}, \ldots, x_{t-k})$$

provided $k > (d-1) * \tau$. The information concerning the explicit form of the function that accomplishes the desired extrapolation is however not provided by Takens' theorem. That is where the neural network comes into the picture. Here again, we see that chaotic time series may be well fitted by a NAR model, under acceptable conditions, provided the model is fed with a sufficient number of past values.

2.3 Neural Networks towards Predictor Approximation

The identification of the process is thus the identification of the parameters of the optimal predictor $f()$, based on the past observations. Neural networks have recently emerged as a successful tool to approximate smooth static functions. This is due to the versatility of a multi-layer feedforward neural network in approximating an arbitrary static nonlinearity (Cybenko, 1989). It is worth mentioning that the somewhat mysterious ability of connectionist models to infer algorithms to perform generalization is nothing more than real valued function interpolation. For example, Lapedes and Farber (1987) applied the classic multi-layer perceptron architecture as an autoregressive model in which a delayed version of the observable was presented to the input layer. Such architecture has proven to be successful for the prediction of chaotic and noisy real-world time series such as sunspots (Lapedes and Farber, 1987; Weigend et al., 1990).

The predictive ability of such an architecture can be improved by feeding back the prior output errors into the current input as suggested by Eq. (3). Such models, called NARMA, have an advantage over NAR models in much the same way that linear ARMA models have advantages over autoregressive models for some types of series. Connor et al. (1993) have clearly demonstrated that a NARMA network in which the output is fed back to the input, is a special case of the somewhat more general class of fully interconnected networks. For such networks, there is no need to present the $\hat{e}_{t-j}$ to the input since specific non trainable recurrent links from a hidden layer towards itself results in the same functional input-output mapping. Because there is some potential for recurrent networks to provide a better approximation to $f(x_t, \ldots, x_{t-p}, \hat{e}_{t-1}, \ldots, \hat{e}_{t-q})$ than feedforward networks, there has been much appeal to recurrent architectures over the last years.

However, temporal association tasks require a neural network that will store and update context information spanning *a priori* unknown temporal intervals. The time-lagged approach either as a NAR or NARMA model only deals with dynamical problems with fixed, or at least bounded, temporal periods. The length of the delay line of shift register must be chosen in advance to accommodate the longest possible sequence. To overcome this limitation, not only do we use a recurrent configuration, but we also endow the model with an internal dynamic in order to broaden the scope of signals it is possible to model with it.

2.4 The Dynamic and Recurrent Neural Network

This section describes the connectionist model we used to assess the predictability of seeing. A network that is to achieve sequence recognition tasks such as prediction must know about the past. In this perspective, we concentrate on a fully recurrent structure in which the synapses are expressed as autoregressive filters to endow the model with an internal dynamic. By virtue of this capability, the network can identify dynamical systems without having to be explicitly fed with past inputs and outputs and discover contingencies spanning unknown temporal intervals. Note that this architecture encompasses most of the earlier networks introduced for temporal processing tasks (Almeida, 1987; Aussem, 1994; Pineda, 1987; Williams and Zipser, 1989).

The DRNN node equations are:

$$\begin{aligned} o_i(t) &= f(net_i(t)) \\ net_i(t) &= bias_i + \sum_{k=0}^{M} \sum_{j=1}^{N} w_{ji}^k o_j(t-k) \\ f(h) &= \frac{1}{1+\exp(-2\beta h)} \end{aligned} \tag{4}$$

where $o_i(t)$ is the output of node i at time t, w_{ji}^k is a weight from k-th delay of node j to node i, N is number of nodes and M is the order of the autoregressive linear filters. The existence of a fixed-point solution is discussed in Aussem (1994).

The sigmoid performs a smooth mapping from $]-\infty, +\infty[\;\rightarrow\;]0,1[$. Note that the steepness β can be absorbed into weights and biases without loss of generality. Bear in mind that there is a set of terminal units whose only role is to feed input patterns into the rest of the network. There are no connections leading to the input units, since these units are clamped to the values specified in the input pattern. Similarly one unit is taken as the output from which the result of the computation is read off. Note that the desired output is considered as a scalar, without loss of generality, since this study deals with univariate time series. Units having neither input nor output connection to the outside world, are called *hidden* units. In addition, a bias input is achieved by fixing a particular unit activation to 1.0. For notational convenience, we did not distinguish between input and non-input units.

The motivation for using neurons as internal state coordinates is that certain features of the time series can be encoded and help us characterize some of the essential properties that lie behind the observation sequence.

Formally define $\mathbf{x}_t = (x_t, \ldots, x_{t-M})$ where x_t denotes the input value at time t. Let $\mathbf{o}_t^j = (o_t^j, \ldots, o_{t-M}^j)$ for $j = 1, \ldots, N'$ where N' is the number of hidden units. By virtue of the internal dynamic behavior, the DRNN offers the hope to approximate any discrete-time system defined by the following system of deterministic equations:

$$\mathbf{o}_{t+1}^j = g(\mathbf{x}_t, \mathbf{o}_t^1, \ldots, \mathbf{o}_t^{N'}) \quad \text{for} \quad j = 1, \ldots, N' \tag{5}$$

and for which the observable x_{t+1} is given by the combination of the internal parameters $\mathbf{o}_t^j$ according to:

$$x_{t+1} = h(\mathbf{x}_t, \mathbf{o}_t^1, \ldots, \mathbf{o}_t^{N'}) \tag{6}$$

The promise of the DRNN model is to capture the behavior of such dynamical systems, insofar as g and h are "well behaved" functions, accurately approximated by Eq. (4).

The error propagation algorithm employed is a generalization of the back-propagation-through-time algorithm (Hertz et al., 1991; Rumelhart et al., 1986) in which one stores the activations of the units while going forward in time. The backward phase is also backward in time and recursively uses these activations to compute the required gradients. Besides, it benefits from the desirable simplicity of the recurrent-backprop algorithm (Almeida, 1987; Hertz et al., 1991; Pineda, 1987) by relaxing the transposed error network in order to compute the weight changes. From a computational standpoint, the procedure exhibits very attractive features. Order M^2N^2 operations are needed and storage requirements scale as MN. A thorough analysis of the training algorithm is carried out in Aussem (1994) and is briefly sketched out in Aussem et al. (1994).

3 Data Preprocessing for 24-Hour Ahead Predictions

In this section we mainly examine the problems raised by the irregularities of the data records, during the construction of consistent training sets for temperature and seeing prediction at 24 hour intervals. A major lesson from time series analysis is that the *representation problem* plays an important role. A relevant representation can make useful information explicit and strip away obscuring clutter. Different representations can be equivalent in terms of expressive power but differ dramatically in the efficiency to solve problems.

3.1 Temperature Prediction at Sunset

Weather conditions relate to about 165 000 observations in La Silla, and 140 000 observations in Paranal. They are averaged at 20 minute intervals since 1985 in the normal course of action. However, measurements were disrupted due in part from the site construction work, and in part from system breakdowns. In this assessment a succession of atmospheric values 24-hours ahead at sunset were used to predict the temperature at the next such interval. We retained only the data at sunset (i.e. when the sine of the sun elevation is 0 with downward trend) when the telescope enclosure is opened for observation. Seeking to carry out predictions at 24-hour intervals allows diurnal effects to be ignored.

The following set of tuples was first derived from this data for the prediction of the next variation of temperature $\Delta t_{n+1} = t_{n+1} - t_n$ at sunset n:

$$t_n, \Delta p_n, s_n, \sigma_{sn}, d_n \tag{7}$$

where time points $n+1$ refer to the next sunset (i.e. 24 hours later) and

- t_n is the temperature at 2m above ground to be predicted.
- Δp_n is the pressure variation between *n-1* and n.
- s_n and d_n are the wind speed and direction at 10m above ground.
- σ_{sn} represents the standard deviation of s_n during sunset n.

Variation of pressure is used for its information on quickly changing weather conditions, such as incoming weather fronts. All values were linearly range-normalized onto the interval [0,1]. Note that hyperbolic tangent functions were also investigated but did not prove to be relevant for our neural-based prediction. Essentially, such a normalization favors the training of low variations exposed to measurement errors, and significant temperature jumps related to incoming weather fronts are truncated. We believe instead that such incoming weather fronts reveal more intrinsic information of the underlying temperature dynamic than smooth variations. Non-uniform distributions result from the linear normalization and reveal the natural intrinsic drift of the frequencies in which they occur (Figs. 2, 4).

Routines were implemented for culling the data of missing values and assembling the series of 24-hour interval temperature values at sunset. Observations with at least one missing value were dropped from the analysis. The series of course required the succession of values not to be missing which in turn reduced drastically the amount of relevant patterns. Since a few measurement errors had taken place during years of intensive data collection, we retained only those patterns satisfying $\Delta t_n < 9^o C$ according to a preliminary data analysis undertaken on La Silla and Paranal observation sites. The amount of patterns selected from the database is shown in Fig. 1 as a function of time. The strong irregularities exhibited during the period reveal that many deletions occurred during the beginning of the meteorological campaign. A sequence of 1240 observations was retained at La Silla with 124 sequence breaks. We set aside the last 250 patterns corresponding to the year 1993 in order to have a prediction set on which to assess the performance of our forecasts. Unfortunately, only 620 patterns were retained at Paranal with 66 sequence breaks and the last 110 records were left aside for the prediction set.

3.2 Statistical Analysis

Stationarity is a major property of time series and was the first thing we checked. Inspection of the raw temperature plot (Figs. 3, 5) reveals clear nonstationarity of the input data, especially when the wind speed is high. The first-order difference transformation made the output slightly more stationary. Two functions of fundamental importance in the context of time series analysis are plotted, namely, the autocorrelation function (ACF) of the desired output and the intercorrelation of the target with all the input variables (Fig. 6). The rate of decay of the autocorrelation function is roughly a measure of "memory" of the series. The ACF takes the form of a highly damped oscillating function, quickly decaying to zero as the displacement is increased, revealing the absence of relevant information in the past observations. However, the negative correlation ranging from lag 1 to lag 4 indicates that temperature is more likely to vary in the opposite direction in the coming days to the direction observed today. The ACF form tends to show that no periodicities are likely to be present in the data (nonperiodic and stationary time series tend to have an exponential decay). The damping of the ACF advocates the use of small windows of data to be used for modeling purposes, although there is no principled way of choosing the optimal window size from the ACF without knowing the correct functional form of the model.

Besides, the intercorrelation functions between the desired temperature variation and the 5 input parameters were plotted but no clear indication of mutual dependence was found (Fig. 6). The temperature variation shows a major Gaussian-like distribution (Figs. 2; 4, plot *a*). Residual peaks spuriously arise, due to the relatively low amount of data and the high sampling rate of the histogram. The differences between histograms are manifest and give clear indications of the peculiarities of each physical parameter. Note for instance the clear dichotomy of the wind direction (plot d) which reveals north-south predominant orientations.

3.3 Night Average Seeing Prediction

In this section we report on seeing prediction at 24-hour intervals at Paranal from 1989 to 1994. Given concerns about the somewhat chaotic behavior of seeing (Jorgenson et al., 1991), the lack of long sequence measurements, and the absence of a clear relation between seeing and ground-based meteorological conditions, we mainly focused on predictions of night averages, rather than expect the network to predict the exact seeing magnitude at a given time point. In this regard, our objective was to forecast the average seeing value during the next night, according to most of the available meteorological parameters. The average seeing histogram is plotted in Fig. 7.

The following set of tuples was derived for the prediction of the average seeing value during the night, starting at the end of the day $n+1$:

$$\max(t_n),\ \max(p_n),\ \max(s_n),\ \max(\sigma_{sn}),\ \max(rh_n),$$
$$\min(t_n),\ \min(p_n),\ \min(s_n),\ \min(\sigma_{sn}),\ \min(rh_n),$$
$$\mathrm{aver}(d_n)$$

where *max*, *min* and *aver*, refer to the maximum, the minimum and the average values recorded between the sunrise and the sunset during the day n respectively. The average seeing value was computed only when more then 4 values were measured during the night. Such a prediction aims at giving a rough indication of the next night's quality. 860 patterns were collected from the original database and 160 were left out for prediction purposes.

4 24-hour Ahead Predictions

4.1 Intermediate Results

In this section, we report on details concerning performance measures, selection of network parameters, and result analysis. The DRNN used for temperature and seeing forecasts can be regarded as a four-layer network with $5\times5\times5\times1$ nodes and order 2:0:0 (order 0 means no delay at all). In addition, feedback connections of order 0 were introduced between the third and the second layer. The resulting network had 141 links including bias. Selection of these dimensions was mostly by trail and error along with various heuristics. Since the first hidden layer acts as a bank of linear filters, selection of the filter order was motivated by the autocorrelation decay rate which didn't indicate substantial correlation after lag 4. In performing these preliminary experiments, we stress that we have made no effort to obtain an optimal architecture for each problem. Instead, we focused on this configuration and tried to obtain the lowest mean squared error. Note that, in general, selection of dimensions for neural networks remains a difficult problem in need of further research.

The network used to learn the series has a total of 141 degrees of freedom and was fit with between 600 and 1000 training input-output patterns potentially posing the problem of overfitting. A topological reduction approach (Hertz et al., 1991; Weigend et al, 1990), consisting of pruning appropriately non-useful connections as training proceeds, prevented the network from overfitting the data. In addition, the network was stopped when the minimum of the mean squared error in the cross validation set was reached.

Due to the neural network internal memory, breaks in the sequence were troublesome. This issue is particularly important when the network has to discover contingencies spanning long temporal intervals. Hopefully, this was not the case with the meteorological variables so that we could suspend the training process for a few time steps only, waiting for the network to re-build its memory. In these experiments, the training suspension time interval was taken to be proportional to the maximum connection order (see Aussem, 1994).

A measure of fit is given by the *normalized mean squared error* (Gershenfeld and Weigend, 1993):

$$\text{NMSE} = \frac{1}{\sigma^2 N}\sum_{k=1}^{N}(x(k) - \hat{x}(k))^2 \tag{8}$$

where $x(k)$ is the true value of the sequence, $\hat{x}(k)$ is the prediction, and σ^2 is the variance of the true sequence over the prediction duration N. A value of the NMSE = 1 thus corresponds to predicting the unconditional

		NMSE	
data	*site*	training	prediction set
Δt_n	La Silla	0.75	0.70
Δt_n	Paranal	0.79	1.20
seeing	Paranal	0.81	0.50

Table 1: Neural network performance for prediction at sunset.

Experiment	2 degrees	1 degree	0.5 degree
Neural net	72.5%	42.5%	20.0%
nearest neighbor	59.8%	37.8%	18.3 %
Carbon copy $t_n = t_{n-1}$	56.7%	31.3%	18.8%

Table 2: Results of 24-hour ahead prediction of temperature at sunset La Silla.

mean. Table 1 summarizes various NMSE values for single-step predictions on both the training set and the prediction set.

Carbon-copy prediction, i.e. the use of the last value as estimate of the next target, was used in an attempt to give a basic index and allow comparisons (Fig. 8). Note that linear autoregressive models were appraised in a recent study and did not improve the performance of carbon-copy predictions (Murtagh et al., 1992).

The learning rate was nominally set to 0.01 and then varied during the course of training. Initial weight values were chosen randomly. A training run typically took over a night on a Sun SPARC 10 and required several thousand passes through the training set.

Very curiously, better performance was achieved on the prediction set than on the training set at La Silla and for seeing prediction at Paranal (Table 1). We believe this is mainly due to the fact that the training set precedes the prediction set and that erroneous measurements were more likely the further into the past we go since early technical defects were progressively corrected. We assume that more regularities were present in the prediction set at La Silla in the sequence of temperature, and at Paranal for the seeing.

Finally, the nearest neighbor approach, described below, performed very poorly with a NMSE of 1.69 as against 0.70 for the neural net. We conclude that relevant information was present in the near past observations, to anticipate the future temperature fluctuations.

As can be seen in Fig. 8, the prediction ability was hardly noticeable except for the temperature at La Silla where the direction of the temperature variations was correctly forecast most of time. Interestingly, plot *e* clearly shows two cases for which good seeing (i.e. low values) were anticipated by the model. Note that the NMSE value for seeing prediction is lower than the NMSE value for the temperature forecasts at La Silla but does not reflect better forecasting abilities. The reason is that high fluctuations were characteristic of the temperature variations whereas seeing averages were more compact. Ultimately, the network was not able to learn anything else than the noise in the temperature variation at Paranal.

4.2 24-hour Ahead Temperature Forecasts at Each Hour of the Day

In the light of previous analysis, it appears that forecasts at sunset are somewhat opaque and need further research with a view towards obtaining predictions with reasonable assuredness. It is not clear so far if the amount of data extracted from the database was enough for the connectionist model to extract useful information and achieve accurate prediction. Nor is it clear if the use of wind speed and direction is more relevant in the context of temperature prediction than prior temperature and pressure values. Therefore, we concentrate on 24 hour predictions at any hour of the day in order to increase the amount of available data and use successive temperature and pressure values instead of wind characteristics. The data used in this assessment relate to over 18000 observations carried out in austral summer 1989-1990. The following set of tuples was derived (Murtagh et al., 1992) for the temperature t_n at time-point n:

$$t_{n-1}, t_{n-2}, t_{n-3}, p_{n-1}, p_{n-2}, \tag{9}$$

where time-point $n-1$, $n-2$ and $n-3$ refer to the previous 24, 48 and 72 hours respectively. All the parameters were range normalized. A static recurrent architecture was used since the memory was made external to the network itself. A set of 5260 observation tuples, with no missing values, was used. The first 5000 tuples were used for training purposes, and predictions were carried out on the last 260 tuples.

Besides the neural network, prediction was performed using a 1-nearest neighbor approach. That is to say, a best matching chronologically prior tuple was sought using the ordinary Euclidean distance, to provide a predicted value of t_n. The matching was carried out using the 5-tuples (9). To reflect a practical implementation, forecasts were carried out in one of the two following ways. Either a chronologically prior tuple could be used as the basis for a prediction, and in addition all predictions were carried out among the most recent 1260 observations (E1); or the prediction was based on tuples within a stock of 4000 prototype tuples only, and again carried out among the 1260 most recent observations (E2). The latter experiment, albeit less natural, can more easily be compared with the connectionist approach since the stock of samples can be regarded as a training set, using the neural network terminology. A third experiment (E3), similar to E1, was conducted except that pressure variables were not used.

Table (3) presents results on the accuracy of prediction, and the base-line result provided by carbon-copy prediction, i.e use t_{n-1} as estimate of t_n. A multidimensional autoregressive model (ARX) was used; it involves the fitting of the linear model:

$$t_n = a_1 t_{n-1} + a_2 t_{n-2} + a_3 t_{n-3} + a_4 p_{n-1} + a_5 p_{n-2} + \epsilon$$

where the noise, ϵ, is Gaussian and of zero mean, and where the coefficients a_j for $j = 1, \ldots, 5$ are estimated.

Two static neural networks were used in this experiment: a four-layer network with $5 \times 14 \times 14 \times 1$ nodes without any delayed connection (E4), and a recurrent $5 \times 10 \times 10 \times 1$ network where backward links connect the third layer to the second (E5). Both neural networks have approximatively 300 adjustable weights. They were trained for around 1000 epochs with a constantly decreasing learning step from 0.01 to 0.0001. The first 5000 tuples were retained to form a training set, while the last 260 were left aside for prediction purposes.

In this experiment in which the past observations are presented spatially to the network, the nearest neighbor method is shown to outperform the other approaches (Table 3). One sees that E1 outperforms E2, since more historical information is being used in the former case for arriving at a prediction. There is every reason to prefer the experimental arrangement associated with E1 instead of E2 in a practical implementation.

The feedforward network appears to be more appropriate than the static recurrent architecture for this particular problem but is far less good compared to the first nearest neighbor method (E1). It would be unwise to draw any conclusion from these preliminary results. It seems however that the nearest neighbor approach is well suited when the memory is presented via a delay vector and when all the chronologically prior tuples are used to provide the predicted value. It is also worth mentioning that the absence of dynamic in the network clearly limits the predictive capability of the connectionist model.

Finally, we note that the pressure variables are useful (cf. E1 versus E3), and that carbon copy performs better at times than the ARX process.

Experiment	2 degrees	1 degree	0.5 degree
E1 - nearest neighbor 1	83.6%	72.7%	62.3%
E2 - nearest neighbor 2	66.7%	48.6%	32.3%
E3 - nearest neighbor 3	68.1%	50.2%	37.1%
E4 - Neural net 1	71.1%	45.1%	22.0%
E5 - Neural net 2	69.7%	41.8%	21.2%
E6 - Carbon copy $t_n = t_{n-1}$	63.2%	36.3%	18.8%
E7 - ARX model	63.7%	35.8%	16.3%

Table 3: Results of 24-hour ahead prediction of temperature at any hour of the day. See text for further details.

5 Fuzzy Nowcasts of Seeing

5.1 Nowcasting versus Forecasting

Our intent in forecasting seeing 24 hours in advance was not to clearly demonstrate the predictability of seeing, but rather to show that good seeing was sometimes and somehow related to prior meteorological conditions. Recent studies (Jorgenson et al., 1991) have given indications, to some extent, that seeing *might* be more accurately modeled by a chaotic process than by a random process. Confirmation of this fact would introduce the possibility of predicting the short-term fluctuations but preclude long-term predictability. This would in turn lend validity to our earlier statement that seeing is hardly predictable one day in advance. Consequently, the framework pursued in this section attempts to relate seeing measurements with meteorological measurements *at the same time point.* This is not quite prediction, but rather *nowcasting* (Murtagh and Sarazin, 1993) since we are using the meteorological variables here without trying to forecast them. In the terminology of statistics, this is multiple regression of seeing on a set of meteorological variables. We assume that good estimates can be provided by an *ad hoc* meteorological model, and we attempt to capture the time-local determinants of the seeing dynamic. The basic objective is the exploration of ways to improve our seeing prediction abilities by splitting the problem into two parts: (i) the prediction of meteorological conditions one day in advance, (ii) the nowcasting of seeing according to these predicted parameters.

This experiment is devoted to nowcasting of seeing given:

$$t_n, p_n, s_n, \sigma_{sn}, rh_n \tag{10}$$

Coding of the data plays a pivotal role. An intuitively plausible and interpretation-friendly form of coding was adopted.

5.2 Seasonally-Referenced Fuzzy Coding

Each variable was categorized into three categories: low-valued (=0), a linearly varying moderate-value, and a high value (=1). The boundaries between these values were set at the 33rd and 67th percentiles. Such coding is termed "fuzzy coding" or "piecewise linear coding". Based on this coding, exploratory data analysis was carried out in Murtagh et al. (1994). Correspondence analysis (a method with various similarities to principal components analysis) was used to seek strong relationships between meteorological variables including seeing. Each case (i.e. a set of temperature, pressure, etc. values) refers to a median of night-time recorded values, and is therefore associated with an observing night. In a practical scenario, we believe it would be relatively easy to obtain high quality estimates of such variables all the more so as we require only minimal information, i.e. high, low, or moderate. The coding of seeing was referenced to the season. A season is a set of about 90 nights which are batched together in time. In fact, around 50 nights remained when missing values were removed. The seasonal referencing which required a relevant set of 33rd and 67th percentiles was provided by the immediately preceding "season" of 49 cases. The data for all seasons between spring 1989 and autumn 1991 were from Cerro Paranal. Construction then started at this site and the data for the four final seasons, summer 1992 to spring 1992, were collected from the nearby Cerro NTT. These data were collected and gathered to form a 605-tuples dataset. While the records all correspond to the same physical parameter, namely seeing, we noticed that the measurement quality and the data regularity were not uniform over the dataset. Thus, two prediction sets of 105 tuples were used to appraise the neural network performance and to make comparisons with the nearest neighbor method as consistent as possible. The first is related to Cerro NTT during winter-spring 1992, while the other is related to Cerro Paranal during spring-summer 1989.

5.3 A Cost-Adaptive Connectionist Approach

In this section, we report some preliminary experimental results obtained with a connectionist network. We note that an interest in this paper, beyond the quest for reliable forecasts, is in optimizing the scheduling observation time assignment procedure at ESO. A preliminary cost analysis has been carried out given that the queue policy amounts to selecting the observation that best fits the estimated seeing value: see Table 4. While the fuzzy nearest neighbor method provides unequivocal estimates of seeing, the neural net output interpretation is dependent on this particular performance objective.

In the following, we concentrate on a network without delays to capture the static dependencies between time-local meteorological conditions and the fuzzily coded seeing. The static network used for the seeing experiment is a four-layer feedforward network with $5 \times 14 \times 14 \times 1$ nodes without any delayed connection. A recurrent network did not help improve the results. The network had 309 links including bias. Overfitting was prevented by a weight elimination method (Aussem, 1994; Weigend et al., 1990). In performing these experiments, we have made no effort to find an optimal architecture. Further investigations are being conducted at present by means of a dynamic network to benefit from the median seeing short-term inertia and will be reported elsewhere. Indeed, short-term correlations are present in the seeing time series (in view of the ACF, Fig. 11), despite the breaks in the sequence.

Since the network output $o(t)$, at any time t, fell mostly inside the interval [0,1] , two thresholds (s_1, s_2) were required to interpret the output value. The prediction was assumed to be low whenever $o(t) < s_1$, a moderate value when $s_1 \leq o(t) \leq s_2$ and a high value when $o(t) > s_2$. These thresholds were optimally estimated from the performance achieved on the training set and then applied to the prediction set to appraise the predictive performance of the model. The thresholds could be adjusted on-line in a practical implementation as time progresses, among the chronologically most recent tuples. Consequently, the results presented are a lower bound of the expected performance. An upper bound of the expected prediction performance is also provided by optimally adjusting the thresholds on the prediction set in suitable way, according to a given cost function (see Tables 10 and 14). We emphasize that the choice of the cost function can be tailored in a convenient way to meet the changing requirements of the engineers and astronomers of the observatory.

With a view to presenting the performance of the connectionist model with the above objective in mind, four pairs of thresholds were used, related to the following performance measures of interest:

- P1 – maximum hit rate, when actual and estimated seeing are in the same category.
- P2 – maximum conditional probability to have good seeing when good seeing was predicted.
- P3 – minimum of the cost function specified by Table 4.
- P4 – minimum incorrect estimate rate, when actual and estimated seeing are in opposite categories (good-bad or bad-good).

estimated \ actual seeing	good	moderate	bad
good	-1	+1	+1
moderate	+1/2	0	+1/4
bad	+1/2	+1/4	-1/4

Table 4: Costs associated with each possibility. Positive and negative values express failure and success respectively. Observing time is lost when an observation requiring good seeing is scheduled during a bad or moderate actual seeing night. Conversely, observing time is inefficiently used when an observation with no specific seeing requirements is scheduled during a good seeing night.

The results obtained for both prediction sets are cross-tabulated in Tables 6 to 14. Various measures of interest can be derived from these contingency tables. They are described below in a convenient way if one wishes to diminish the number of "false alarms", i.e. bad seeing when good seeing is expected, or to minimize a given cost function, etc.

5.4 Neural Network Nowcasts

5.4.1 Winter 92 to Spring 92 at Cerro NTT

The NMSE on the training and prediction sets were 0.73 and 0.56 respectively. Better results were obtained on the prediction set which yields more accurate nowcasts on this set: Fig. 9. Cumulative prediction histograms are plotted in Fig. 12 and neural network outputs in Fig. 10.

- For policies P1 and P3, the conditional probability of good and bad seeing, given that good seeing was predicted, is 80.0% and 6.7% respectively. The possibility that good seeing is nowcast, that is, the number of good nowcasts / total number of nowcasts, is 28.6%.

- For policy P2, the conditional probability of good and bad seeing, given that good seeing was predicted, is 87.0% and 4.3% respectively. The possibility that good seeing is nowcast is 21.9%.

- For policy P4, we found a minimum incorrect estimate rate of 0%. The conditional probability of good seeing and bad seeing, given that good seeing was predicted, is 90.0% and 0% respectively. The possibility that good seeing is nowcast is 9.52%.

- When the thresholds are determined from the prediction set to minimize the cost, the conditional probability of good seeing and bad seeing, given that good seeing was predicted, is 85.7% and 3.6% respectively. The possibility that good seeing is nowcast is 26.7%.

As an indication, the three parameter configuration is related to the costs -15, -11.75 and -14.75 respectively. The fact that result 1 has a lower cost value than result 3 is not paradoxical in that the cost was minimized on the training set and not on the prediction set.

5.4.2 Spring 89 to Summer 89 at Cerro Paranal

The NMSE on the training and prediction sets were 0.70 and 0.84 respectively. Logically, better results were obtained on the training set.

- For policies P1 and P3, the conditional probability of good seeing and bad seeing, given that good seeing was predicted, is 66.7% and 20.0% respectively. The possibility that good seeing is nowcast is 28.6%.
- For policy P2, the conditional probability of good seeing and bad seeing, given that good seeing was predicted, is 72.7% and 18.2% respectively. The possibility that good seeing is nowcast is 10.5%.
- For policy P4, we found a minimum incorrect estimate rate of 9.52%. The conditional probability of good seeing and bad seeing, given that good seeing was predicted, is 66.7% and 5.7% respectively. The possibility that good seeing is nowcast is 16.7%.
- When the thresholds are determined from the prediction set to minimize the cost, the conditional probability of good seeing and bad seeing, given that good seeing was predicted, is 70.4% and 14.8% respectively. The possibility that good seeing is nowcast is 25.7%.

As an indication, the three parameter configuration is related to the costs -0.75, $+13.5$ and -0.75 respectively.

5.5 Nearest Neighbor Nowcasts

The nearest neighbor approach is extended to deal with the fuzzy coding. Given a set of meteorological variables, we seek among all those stored in our database, the fuzzy seeing value associated with the nearest neighbor vector as the estimate of the desired value. Multidimensional closeness has been measured in the temperature experiments, using the Euclidean distance between range normalized variables. For seeing, in order to be consistent with the coding used, a fuzzy Jaccard measure (Miyamoto, 1990; Murtagh et al., 1994) is used to find the best match. The Jaccard measure is well-known in information retrieval, and in some tests we carried out, it outperformed a fuzzy Euclidean distance. The definition, for vectors $\{A_i\}$ and $\{B_i\}$ is:

$$\bigcap_i (A_i, B_i) / \bigcup_i (A_i, B_i) = \sum_i \min(A_i, B_i) / \sum_i \max(A_i, B_i)$$

When equivalued similarities were found, we averaged the seeing values in an *ad-hoc* way, using the ordinary definition of mean. The experimental "leave one out" strategy consists in searching for the most similar set of meteorological variables in the entire database which also contains data collected at time beyond the case currently investigated. The seeing associated with the most similar set of meteorological variables provided the estimate. The bigger our database, the more chance we have of capturing every meteorological combination of interest.

The results obtained are presented in Table 5 and may be summarized as follows:

- Conditional probabilities of good and bad seeing, given that good seeing was estimated, are 48.2% and 22.0% respectively.
- Actual and estimated seeing in the same category: 40.2% of cases.
- Actual and estimated seeing in opposite categories (good–bad or bad–good): 9.6% of cases.

Here, one may remark that an uninformed policy based on the definition of good seeing, would lead to 33% of cases. Bear in mind in the sequel that the cost associated with the nearest neighbor nowcasts is +16.8 (this cost was normalized to be consistent with the 105 predictions carried out on the prediction set of the neural networks).

6 Discussion

Throughout the experiments presented in this paper, we found both methodologies, i.e. the connectionist model and the nearest neighbor method, to perform well in certain circumstances. Though briefly discussed here, it would be interesting to ascertain the problems for which each model is most suited. A question of importance relates to when and how *high-accuracy* predictions can be made.

It was found that the quality of the results obtained with the k-nearest neighbor method rested on a sufficiently large number of near-identical past prototypes of the series. This appears, empirically, to be a precondition for this approach to work well. Indeed, Farmer and Sidorowich (1988) pointed out that local approximation produces better fits, particularly for large data sets, compared to global approximation. Thus the consideration which motivates the use of local approximation schemes relates to the density of occupancy in the multidimensional space. How well does this local approximation work? Not only does it depend on the number of data points N but also on the space dimension D, the prediction time and the signal to ratio noise S. Suppose we want to approximate an analytic function f with the nearest neighbor approach, that is a first order local approximation, then the error depends on the first derivative $f'(x)$. This implies that the errors are proportional to ϵ, the typical spacing between data points. If $\epsilon \simeq N^{1/D} < S^{-1}$ then the forecast is limited by noise. If, on the other hand, $\epsilon > S^{-1}$ then the accuracy of forecasts is limited by the number of data points. In this case, provided f is invertible, the normalized mean squared error E satisfies

$$E \sim N^{-1/D}$$

As an example, one deduces from the above formula that the expected error would be decreased by about 30% when using 5000 data instead of 1000 in the 24-hours ahead temperature experiments. In fact, such a local approximation will fail if the system is high-dimensional like the average seeing prediction input-output mapping, or is nonstationary, because in these cases, there is no longer a simple surface to build. Furthermore, we are never relieved from the uncertainty of having a one-to-one correspondence between the input pattern and the output. This is rarely the case in practice, resulting in spurious estimates with the nearest neighbor approach. One way to potentially get around this problem is to use a delayed version of the meteorological observations following the state-space reconstruction technique (Sauer, 1993), at the expense of the space dimension. This yielded significant results in the 1-day ahead temperature forecats at any hour of the day in section 5.

Sparsity of the data motivates the use of connectionist models. Other considerations relate to the high level of noise present in the ground-based measurements essentially due to random meteorological effects, the relatively low density of occupancy of the multidimensional space in which the matching is carried out, and the nonstationarity of the series. Fortunately, nonlinear and global adaptive surface approximation around points of interest are allowed for by neural networks which thereby cope with season effects. Nonetheless, most global representation schemes such as the neuronal method reach a point of diminishing returns when adding more parameters, e.g. weights, or more data only give a marginal improvement in accuracy. Indeed, adding more parameters entails the problem of fitting the noise, termed overfitting in the terminology adopted in the connectionist field.

We see that neither approach is uniquely preferable. The domains of applicability are not the same, and the choice of which to use depends mostly on the sparsity of the data in the multidimensional space, the stationarity, and the smoothness of the surface to be fitted. Further substantiation through more experiments and analysis will be proposed in future papers.

7 Conclusion

We have described a connectionist approach for predicting temperature and astronomical seeing at two observatory sites, La Silla and Paranal. A k-nearest neighbors method was also applied in order to appraise both methodologies and make meaningful comparisons among them in the context of real-world time series forecasting. Attempts to achieve reliable forecasts at sunset, one day in advance, did not prove to be illuminating. We found the dynamical and recurrent neural network to perform poorly in the temperature and seeing experiment at Paranal, but acceptably well in the temperature experiment at La Silla. An additional experiment has been carried out in which temperature values were to be estimated 24 hours in advance at each hour of the day. In this particular case in which the number of data is significantly augmented, better results were obtained with the nearest neighbor method. Insufficient data and numerous disruptions in the record sequence were the major stumbling blocks for the nearest neighbor and the connectionist approach respectively. Fortunately, reliable fuzzy seeing estimates were obtained by a static neural network given the present environmental condition parameters. The flexible neuronal methodology discussed for fuzzy seeing nowcasts enables the observatory to select appropriately the performance objective function most suited to the present observation requirements. The adequacy of the nowcasting approach for seeing prediction is confirmed and future work should aim to obtain encouraging results for the prediction of fuzzily coded seeing one hour in advance. Finally, the effectiveness of these predictive methods to achieve reliable forecasts were discussed with a view to ascertaining the prediction problems for which each method is most suited.

The assessment figures for nowcasting of seeing which are quoted in this article, and the connectionist model which has been prototyped, are a contribution to a future telescope decision support system.

estimated \ actual seeing	good	moderate	bad
good	68	42	31
moderate	109	117	104
bad	27	49	58

Table 5: Contingency table exemplifying estimated versus actual seeing values with the k-nearest neighbor method in the overall data set – Same category: 40.2%, Partially wrong: 50.2% and Incorrect: 9.6%

estimated \ actual seeing	good	moderate	bad
good	24	4	2
moderate	7	9	4
bad	3	20	32

Table 6: 1992 – Contingency table exemplifying estimated versus actual seeing values in prediction set – Output thresholds $(0.34; 0.68)$ have been optimally determined from the training set to maximize the hit rate (P1) – Same category: 61.9%, Partially wrong: 33.3% and Incorrect: 4.8%

estimated \ actual seeing	good	moderate	bad
good	20	2	1
moderate	11	11	5
bad	3	20	32

Table 7: 1992 – Contingency table exemplifying estimated versus actual seeing values in prediction set – Output thresholds $(0.27; 0.77)$ have been optimally determined to maximize the good-good conditional probability (P2) – Same category: 60.0%, Partially wrong: 36.2% and Incorrect: 3.8%

estimated \ actual seeing	good	moderate	bad
good	24	4	2
moderate	4	2	1
bad	6	27	35

Table 8: 1992 – Contingency table exemplifying estimated versus actual seeing values in prediction set – Output thresholds $(0.55; 0.68)$ have been optimally determined to minimize the cost function (P3) – Same category: 58.1%, Partially wrong: 34.3% and Incorrect: 7.6%

estimated \ actual seeing	good	moderate	bad
good	9	1	0
moderate	25	31	37
bad	0	1	1

Table 9: 1992 – Contingency table exemplifying estimated versus actual seeing values in prediction set – Output thresholds $(0.12; 0.88)$ have been optimally determined to minimize the incorrect estimate rate (P4) – Same category: 39.0%, Partially wrong: 61.% and Incorrect: 0%

estimated \ actual seeing	good	moderate	bad
good	24	3	1
moderate	7	8	3
bad	3	22	34

Table 10: 1992 – Contingency table exemplifying estimated versus actual seeing values in prediction set – Output thresholds $(0.4; 0.74)$ have been optimally determined from this set to give an upper performance bound – Same category: 62.9%, Partially wrong: 33.3% and Incorrect: 3.8%

estimated \ actual seeing	good	moderate	bad
good	20	4	6
moderate	6	13	9
bad	9	18	20

Table 11: 1989 – Contingency table exemplifying estimated versus actual seeing values – Output thresholds $(0.39; 0.71)$ have been optimally determined to maximize the hit rate (P1), they also minimize the cost function (P3) – Same category: 50.5%, Partially wrong: 35.2% and Incorrect: 14.3%

estimated \ actual seeing	good	moderate	bad
good	8	1	2
moderate	21	25	22
bad	6	9	11

Table 12: 1989 – Contingency table exemplifying estimated versus actual seeing values – Output thresholds $(0.3; 0.9)$ have been optimally determined to maximize the good-good conditional probability (P2) – Same category: 41.9%, Partially wrong: 50.5% and Incorrect: 7.6%

estimated \ actual seeing	good	moderate	bad
good	4	1	1
moderate	22	16	14
bad	9	18	20

Table 13: 1989 – Contingency table exemplifying estimated versus actual seeing values – Output thresholds $(0.39; 0.91)$ have been optimally determined to minimize the incorrect estimate rate (P4) – Same category: 38.1%, Partially wrong: 52.4% and Incorrect: 9.5%

estimated \ actual seeing	good	moderate	bad
good	19	4	4
moderate	3	2	2
bad	13	29	29

Table 14: 1989 – Contingency table exemplifying estimated versus actual seeing values – Output thresholds $(0.66; 0.75)$ have been optimally determined from this set to give an upper performance bound – Same category: 47.6%, Partially wrong: 36.2% and Incorrect: 16.2%

References

[1] Almeida, L.B. (1987) Backpropagation in perceptrons with feedback, In *Neural Computers*, Ed. R. Eckmiller and Ch. von der Malsburg, Springer-Verlag, Berlin, pp. 199-208.

[2] Aussem, A. (1994) Dynamical and recurrent neural networks towards time series prediction, in preparation.

[3] Aussem, A., Murtagh, F. and Sarazin, M. (1994) Fuzzy astronomical seeing nowcasts with a dynamical and recurrent connectionist network, *Neurocomputing*, submitted.

[4] Box, G.E. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

[5] Connor, J.T., Martin, R.D. and Atlas, L.E. (1994) Recurrent neural networks and robust time series prediction, IEEE *Trans. on Neural Networks* **5**, 240-254.

[6] Cybenko, G. (1989) Continuous value neural networks with two hidden layers are sufficient, *Math. Control Signals and Systems* **2**, 303-314.

[7] Farmer, J.D. and Sidorowich, J.J. (1988) Exploiting chaos to predict the future and reduce noise, In *Evolution, Learning, and Cognition*, Ed. Lee, Y.C., World Scientific Press, Singapore, p. 277.

[8] Gershenfeld, N.A. and Weigend, A.S. (1993) The future of time series: learning and understanding, In *Time Series Prediction: Forecasting the Future and Understanding the Past*, Ed. Weigend, A.S. and Gershenfeld, N.A., Addison-Wesley, New Jersey, pp. 1-70.

[9] Hertz, J., Krogh, A. and Palmer, R. (1991) *An Introduction to the Theory of Neural Computation*, Addison-Wesley, New Jersey.

[10] Jorgenson, M.B., Aitken, G.J.M. and Hege, E.K. (1991) Evidence of a chaotic attractor in star-wander data, *Optics Letters* **16**, 64-66.

[11] Lapedes, A. and Farber, R. (1987) Non linear signal prediction using neural networks: prediction and system modelling, Preprint, Los Alamos National Laboratory LA-UR-87-2662.

[12] Miyamoto, S. (1990) *Fuzzy Sets in Information Retrieval and Cluster Analysis*, Kluwer, Dordrecht.

[13] Murtagh, F., Sarazin, M. and Adorf, H.M. (1992) Statistical prediction of astronomical seeing and of telescope thermal environment, In *ESO Conference on Progress in Telescope and Instrumentation Technologies*, Ed. M.H. Ulrich, ESO, Garching, pp. 377-380.

[14] Murtagh, F., Aussem, A. and Sarazin, M. (1994) Nowcasting astronomical seeing: towards an operational approach, *PASP*, submitted.

[15] Murtagh, F. and Sarazin, M. (1993) Nowcasting astronomical seeing: a study of ESO La Silla and Paranal, *PASP* **105**, 932-939.

[16] Ott, E. (1993) *Chaos in Dynamical Systems*, University Press, Cambridge.

[17] Pineda, F.J. (1987) Generalization of back-propagation to recurrent neural networks, *Physical Review Letters* **59**, 2229-2232.

[18] Priestley, M.B. (1981) *Spectral Analysis and Time Series*, Vol. I, II, Academic Press, London.

[19] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1988) *Numerical Recipes in C*, Cambridge University Press.

[20] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning internal representations by error propagation, In *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, Ed. Rumelhart, D.E. and McClelland, J.L., Vol. 1, MIT Press/Bradfords Books, Cambridge, MA, pp. 318-362.

[21] Sauer, T. (1993) Time series prediction by using delay coordinate embedding, In *Time Series Prediction: Forecasting the Future and Understanding the Past*, Ed. Weigend, A.S. and Gershenfeld, N.A., Addison-Wesley, New Jersey, pp. 175-193.

[22] Weigend, A.S., Rumelhart, D.E. and Huberman, B.A. (1990) Predicting the future: a connectionist approach, *International Journal of Neural Systems* **1**, 193-220.

[23] Williams, R.J. and Zipser, D. (1989) Experimental analysis of the real-time recurrent learning algorithm, *Connection Science* **1**, 87-111.

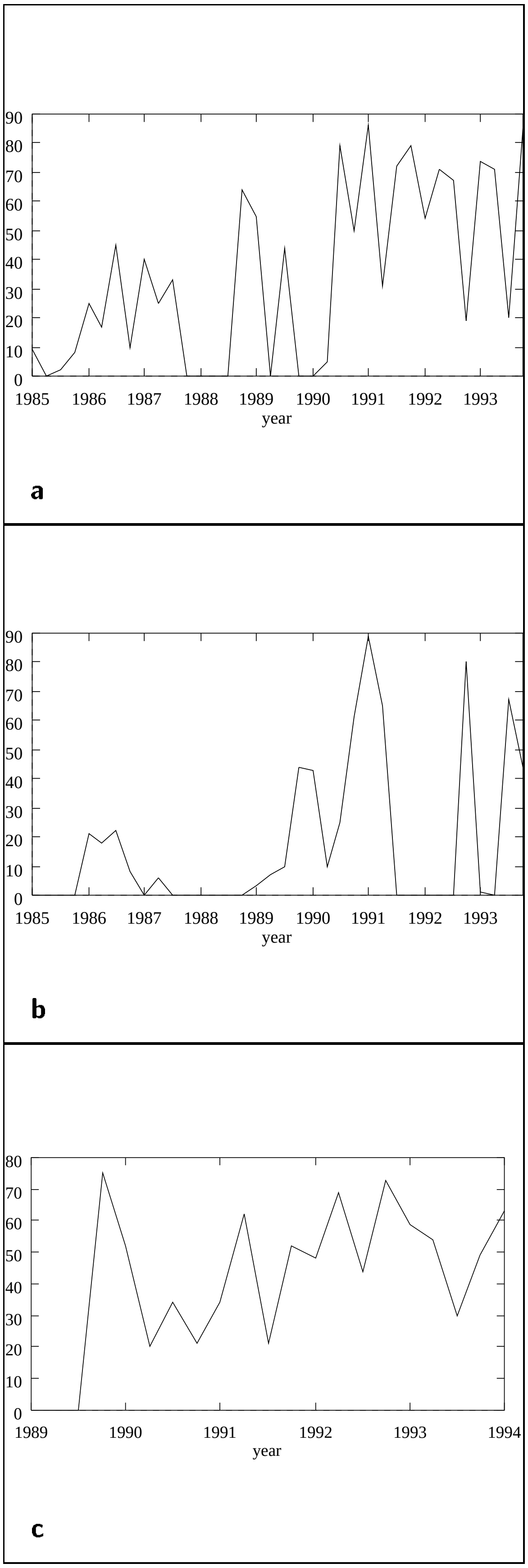

Figure 1: Number of data extracted from database plotted as a function of time, in La Silla (upper plot) and in Paranal for prediction experiments 1 day ahead at sunset (temperature in middle plot and average seeing in plot below).

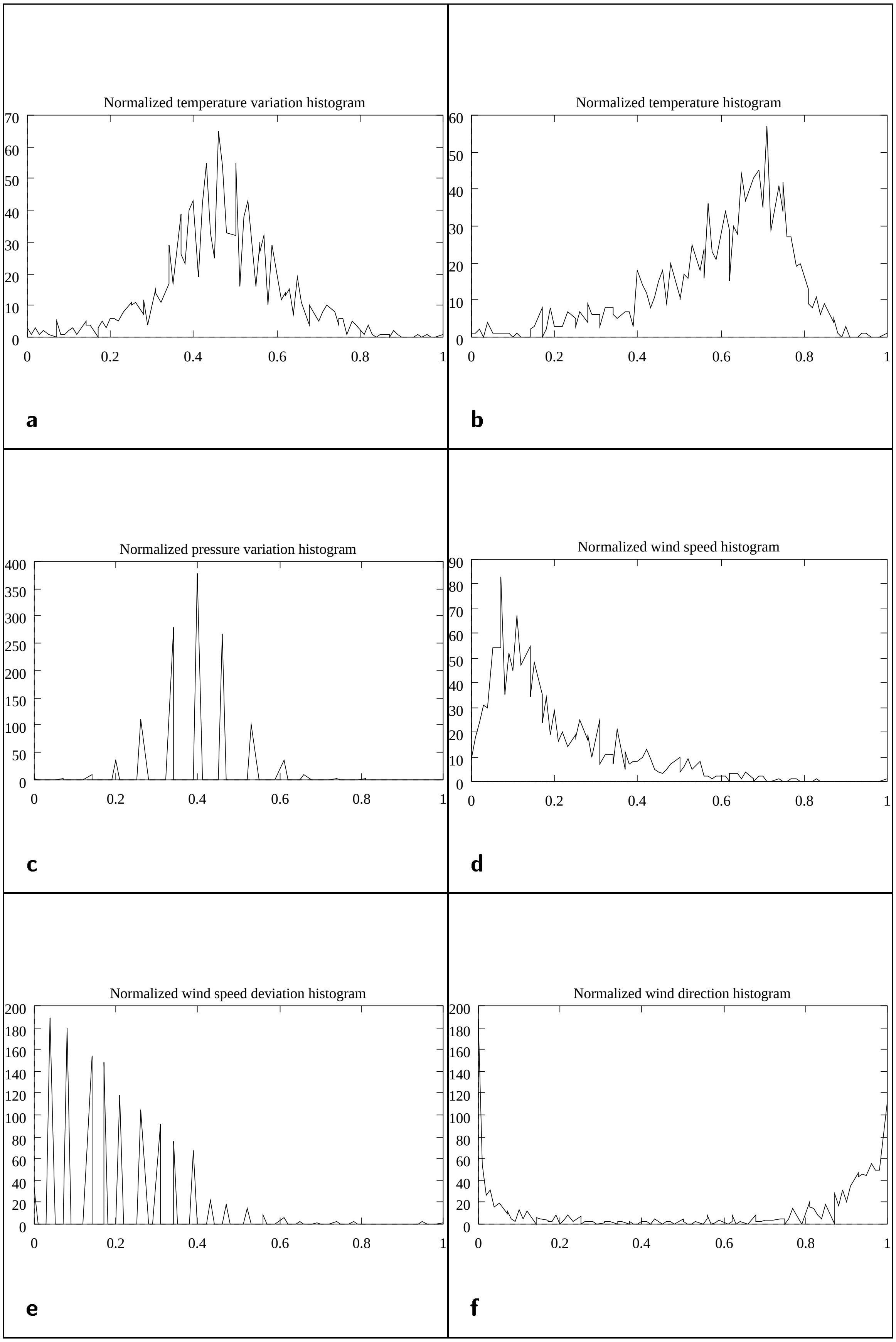

Figure 2: Cumulative histograms of meteorological data at sunset in La Silla: (a) temperature variations, (b) temperature, (c) pressure variation, (d) wind speed, (e) wind speed deviation, (f) wind direction.

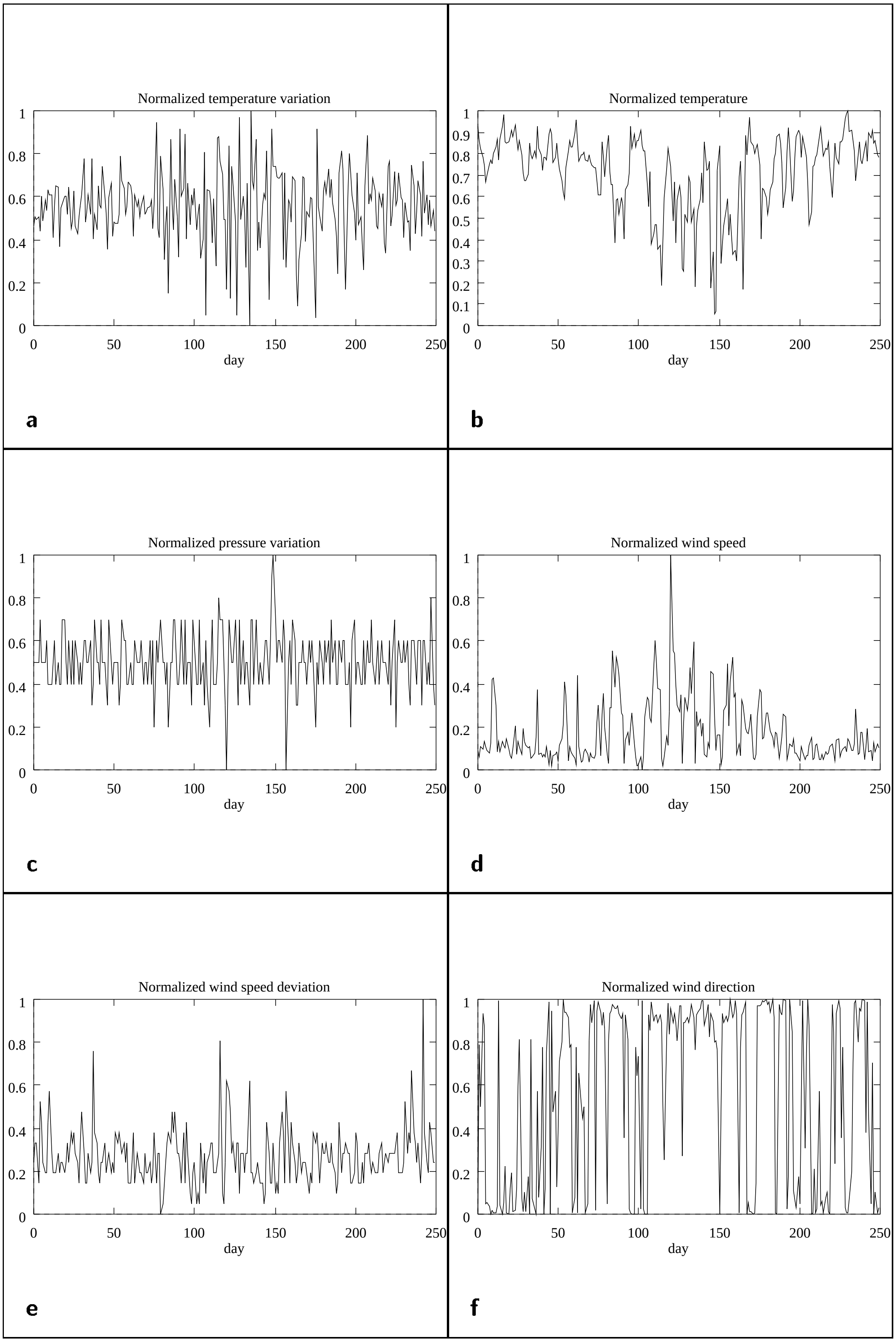

Figure 3: Linearly normalized meteorological data at sunset during year 1993 in La Silla: (a) temperature variations, (b) temperature, (c) pressure variation, (d) wind speed, (e) wind speed deviation, (f) wind direction.

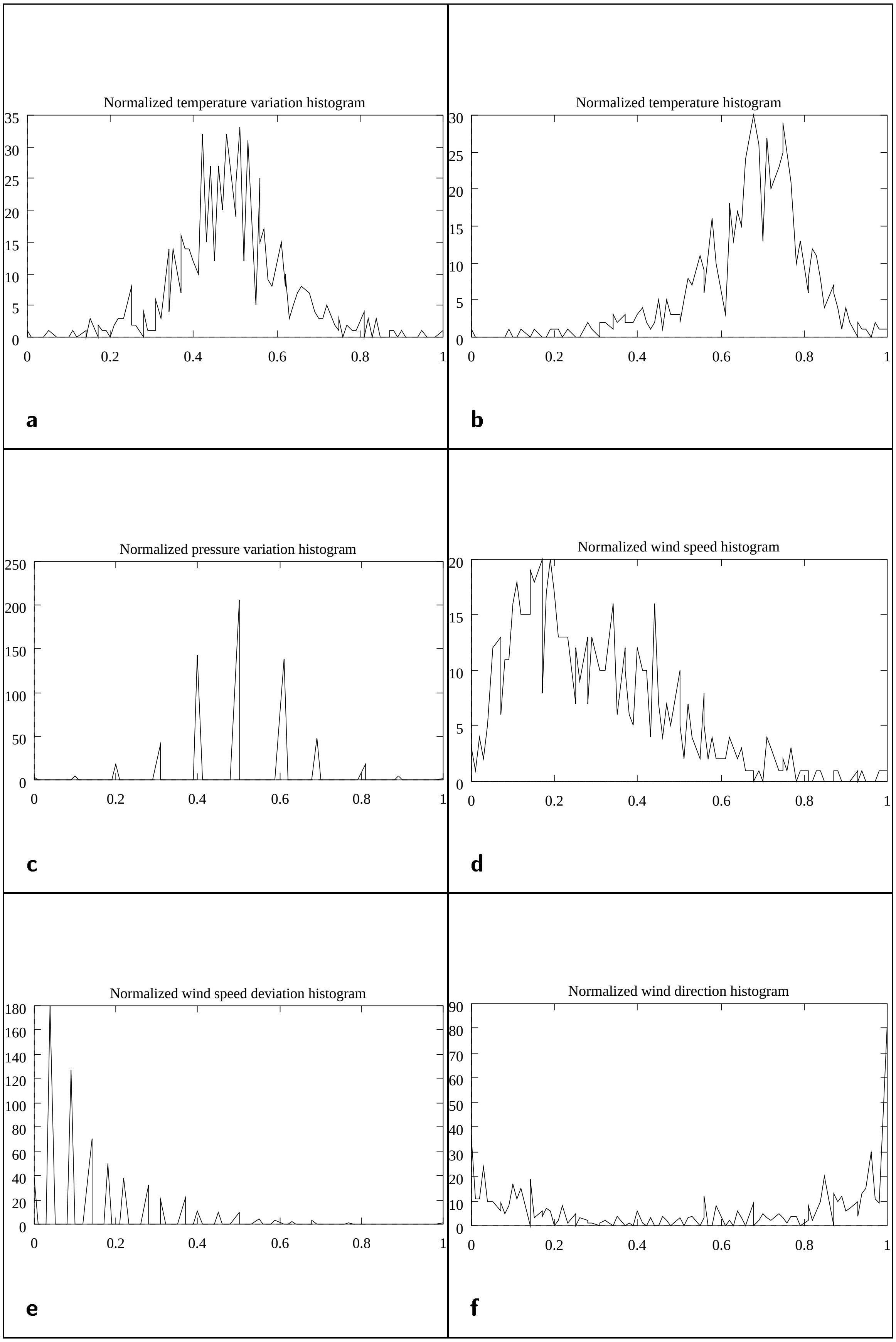

Figure 4: Cumulative histograms of meteorological data at sunset in Paranal: (a) temperature variations, (b) temperature, (c) pressure variation, (d) wind speed, (e) wind speed deviation, (f) wind direction.

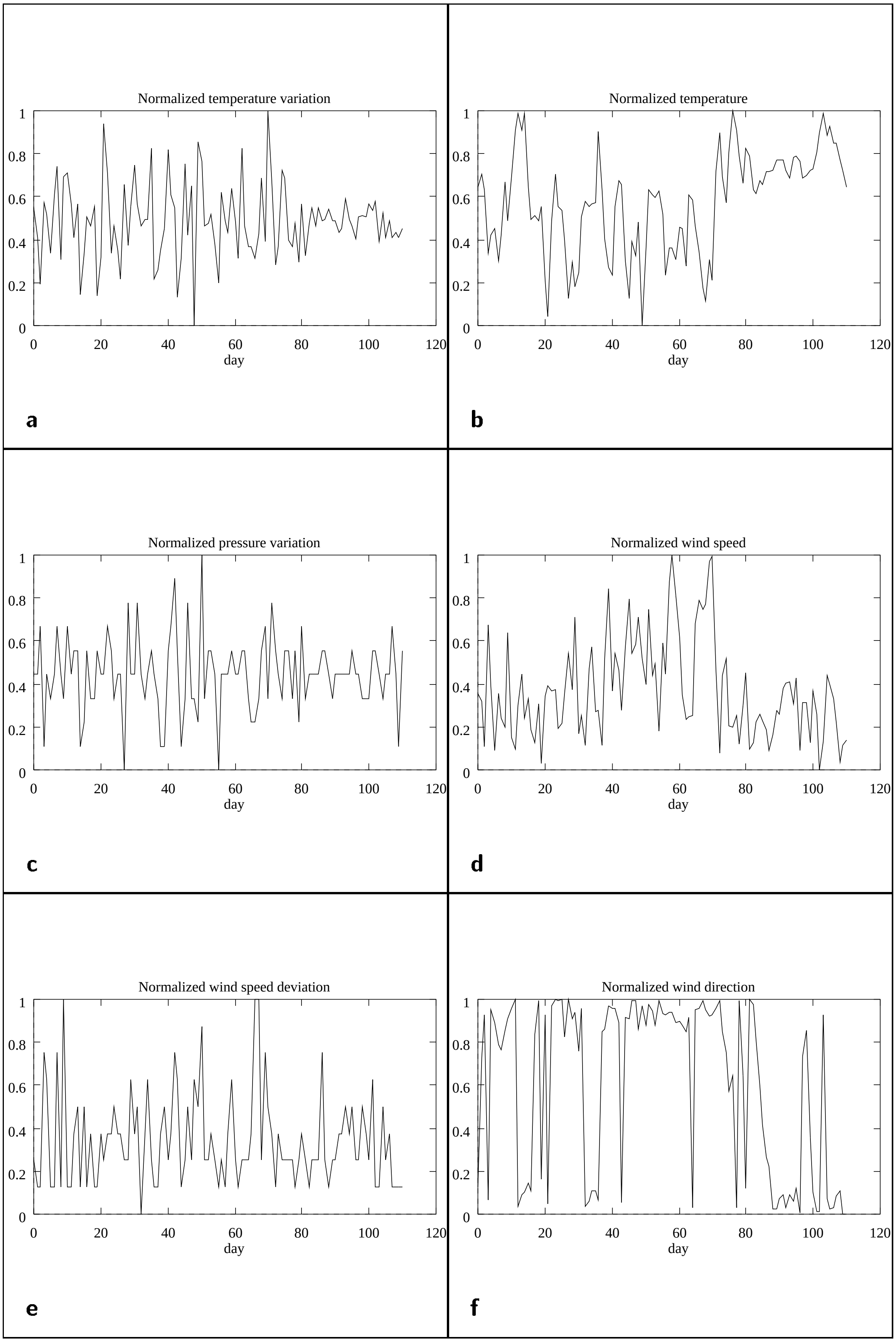

Figure 5: Linearly normalized meteorological data at sunset during year 1993 in Paranal: (a) temperature variations, (b) temperature, (c) pressure variation, (d) wind speed, (e) wind speed deviation, (f) wind direction.

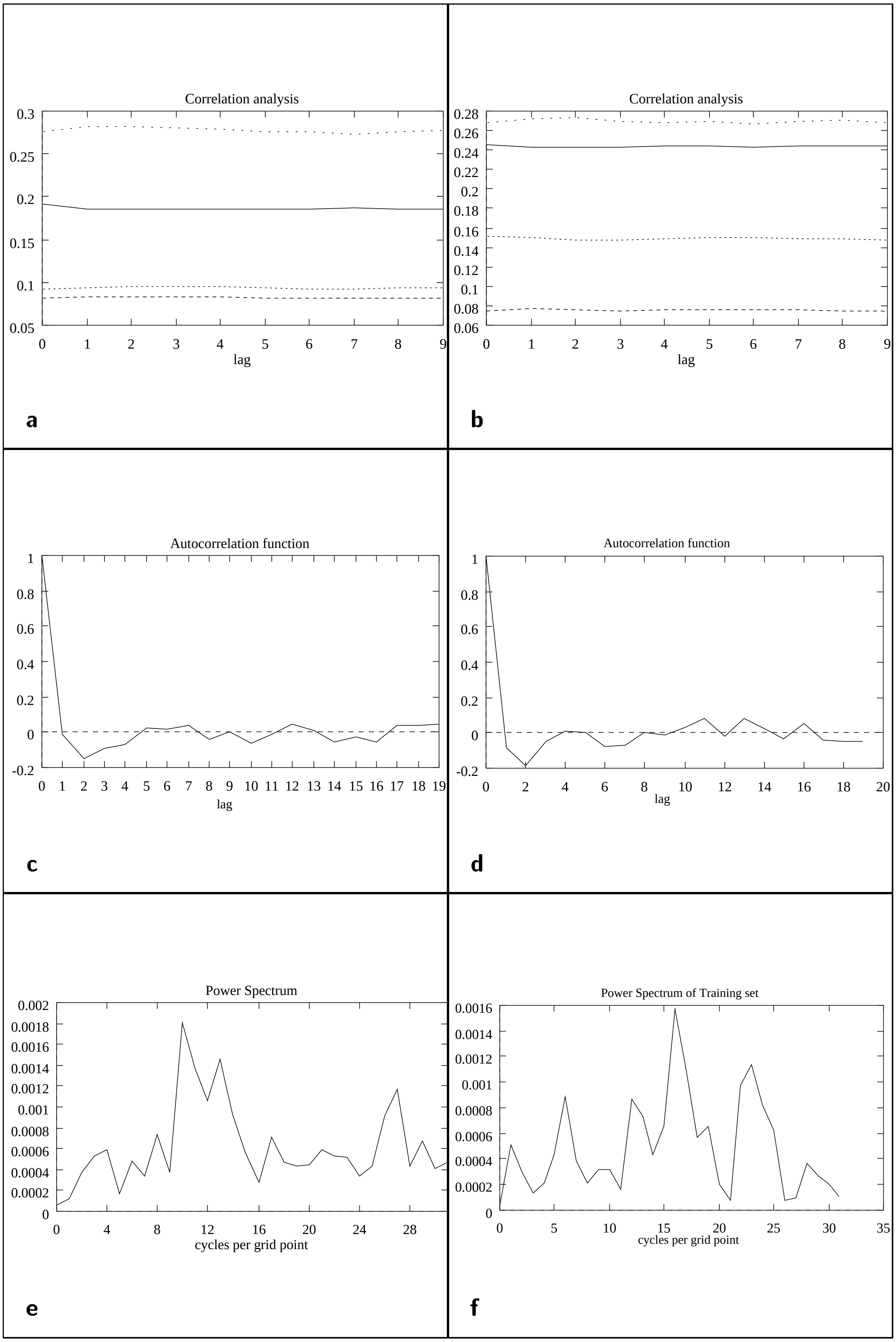

Figure 6: Left side: La Silla – right side: Paranal.(a), (b) Intercorrelation function between Δt_n and, starting from the upper curve, d_n, Δp_n, s_n and σ_{s_n}. (c), (d) Autocorrelation of Δt_n. (e),(f) target mean square amplitude at frequency $j * f_c/32$ ($f_c = 1/48\ h^{-1}$ denotes the Nyquist critical frequency)

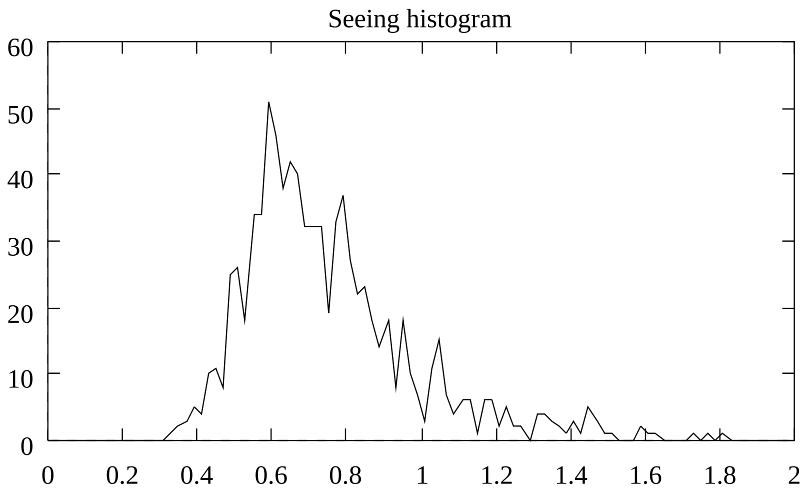

Figure 7: Average seeing histogram at night in Paranal between 1989 and 1994.

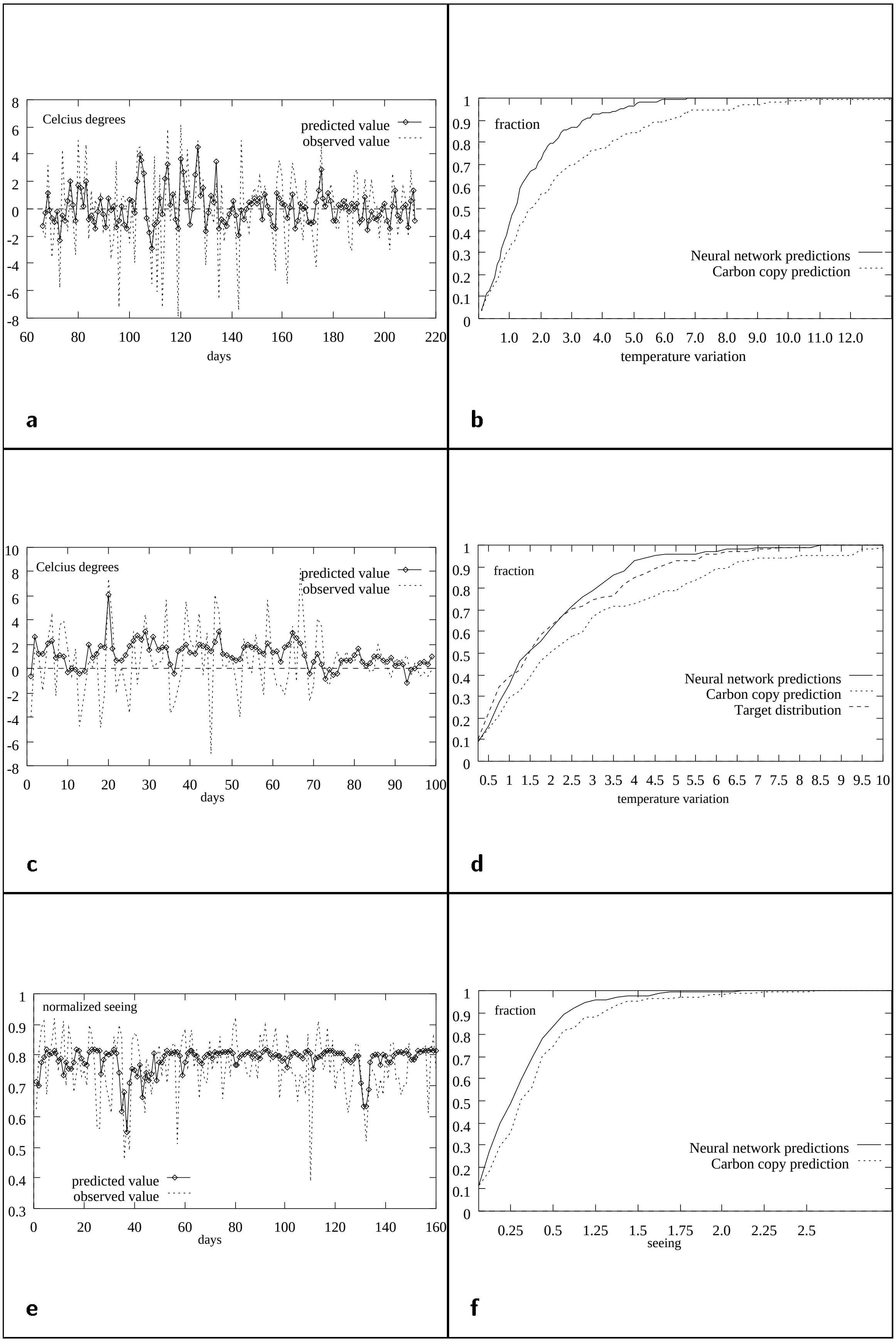

Figure 8: Neural network forecasts and cumulative histograms: (a) and (b) temperature variation predictions at La Silla, (c) and (d) temperature variation predictions at Paranal, (e) and (f) average seeing prediction at Paranal.

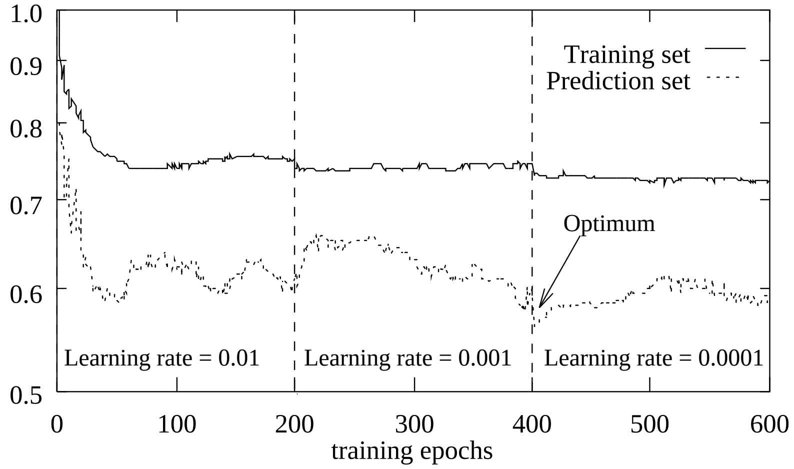

Figure 9: NMSE in both the training set and the prediction set as a function of the training time in epochs. Better performance was obtained on the prediction set which consists of data recorded during winter and spring 1992 at Cerro NTT. The reason is that more regularities were present in the prediction set than in the training set.

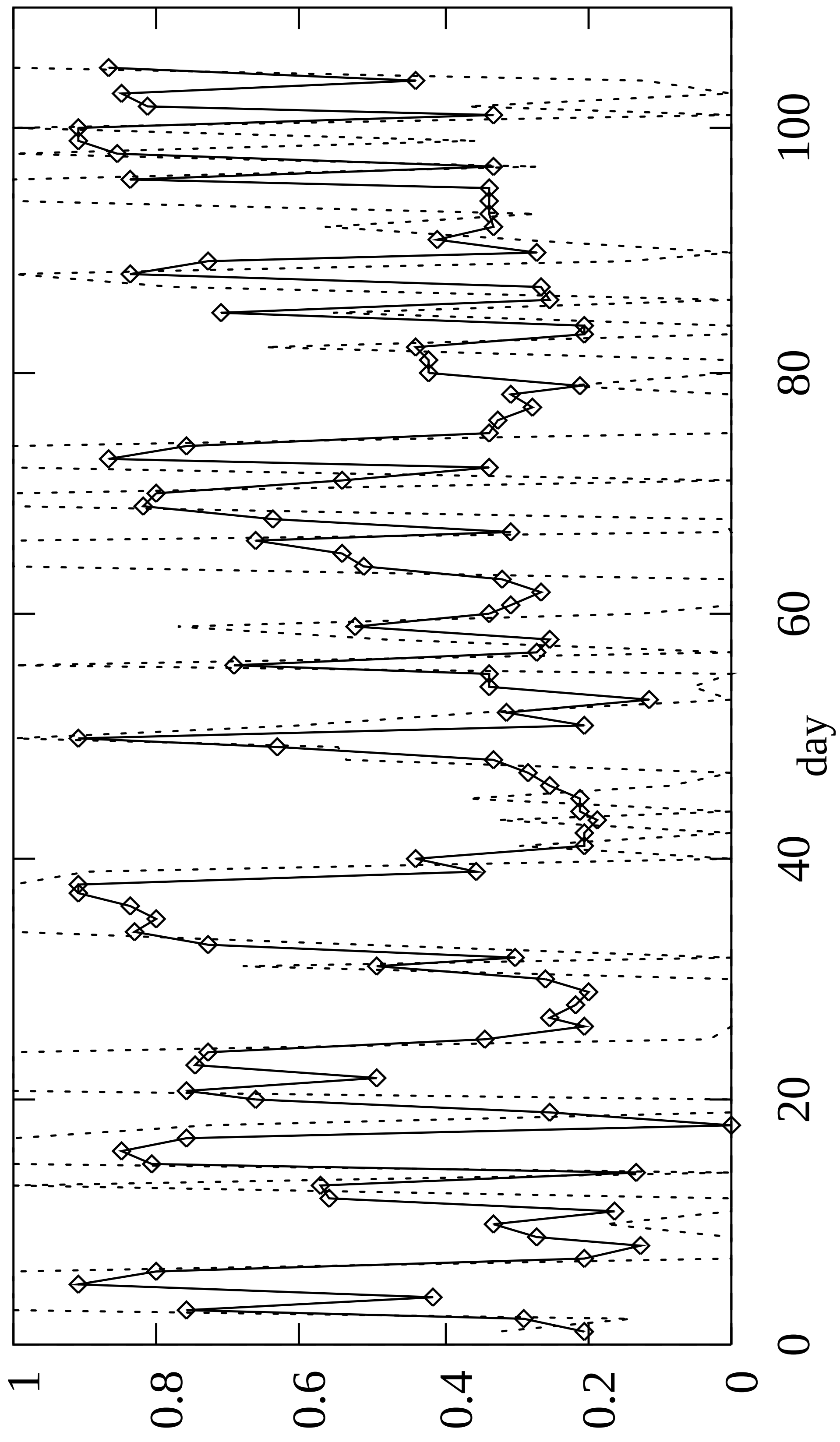

Figure 10: Fuzzy coded seeing (dotted line) during winter and spring 1992 and single-step forecasts (plain line) in the prediction set.

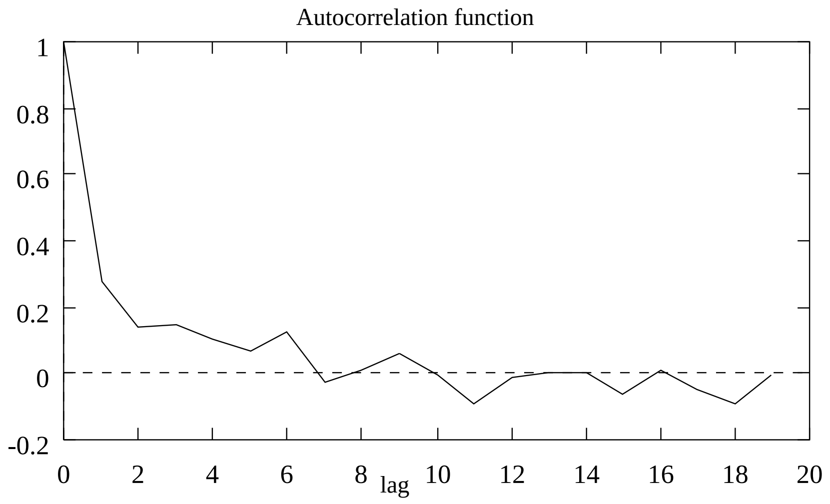

Figure 11: Autocorrelation function of fuzzy seeing sequence.

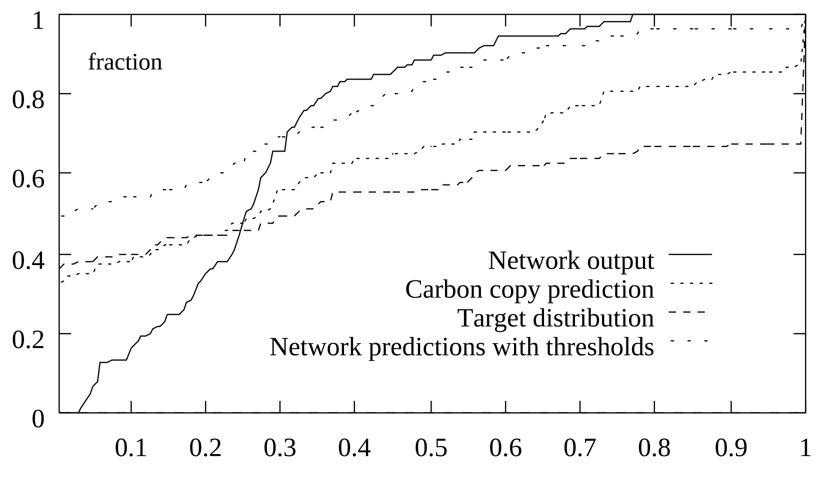

Figure 12: Fuzzily coded seeing cumulative histograms during winter and spring 1992. The plain line corresponds to the raw output of a $1 \times 14 \times 14 \times 1$ feedforward static network. The dotted line corresponds to the output decoded by means of two thresholds optimally determined to maximize the hit rate (P1).