LONG TERM MEASUREMENTS OF WEATHER AND FLUXES OF CO_2: TREATMENT OF DISCONTINUOUS DATA

E.J.Moors and A.J. Dolman

Alterra, PO Box 47 AC, Wageningen, the Netherlands

1. INTRODUCTION

Two of the main purposes for performing long term measurements are:

- to create a long time series to analyse for possible trend and
- to create a parameter space as wide as possible so as to enable us to understand the feedback mechanism of an ecosystem for as wide a range of different conditions as possible.

For this there is a need for continuos long data series of driving variables such as measured by automatic weather stations and flux data. However, measurements on a short time step, for example on a (half-)hour basis, of environmental variables with the objective of creating a long time series of data are inclined to have gaps. Causes for gaps can be power failure, instrument breakdown, or rejection of data after quality check etc.

The treatment of gaps depends on the variable and the length of the gap. For instance, hourly data of a variable with a strong diurnal signal can be missing or daily data of a variable with a seasonal cycle. Both sorts of gaps may be treated differently.

The most important objective of gap filling should be to create *consistent* data series. This means that before the gaps are filled, a data quality check should be executed to ensure high quality data. It should also be clear that the experimentalist should try and do his outmost to create a high quality data series of prognostic values as uninterrupted as possible, and that the procedures outlined here are only meant as a last resort and can never replace high quality measurements.

In this study we use measurements of energy fluxes and NEE (net ecosystem exchange) over a mid latitude Scots pine forest in the Netherlands for a five year period (1995-1999).

2. SITE AND METHODS

2.1. Site description

The forest site is an extensive Scots pine forest in the center of the Netherlands on sandy soil, origination from glacial deposits (e.g. Dolman et al., 1998). The forest was planted in the beginning of the previous century. Existing soil organic material was largely removed at that time. There is an understorey of *Deschampsia flexuosa*, a grass that can reach a height of 50 cm. The soil is a sandy soil (humuspodzol) with a 10cm top layer of organic material.

2.2. Measurements

Fluxes of latent and sensible heat and momentum were obtained by the eddy correlation method from scaffolding towers since early 1995. In 1996 the system was extended to measure also the flux of CO_2. The fetch is at least 1.5 kilometres in all directions and consists of similar forests with the same species of similar age and height. For this site Elbers et al. (1996) calculated that most of the flux originates from 500 m around the tower, with the maximum flux contribution at 120 m for neutral atmospheric conditions.

The eddy correlation system consists of a 3-D sonic anemometer (Solent 1012 R2), a Krypton hygrometer (Campbell, Inc., USA) and a LiCOR 6262 infrared gas analyser linked to a notebook computer. The computer calculates on-line variances and co-variances at half-hourly intervals using an moving average filter with a time constant of 200 s. Measurements were taken at 10 Hz. All raw data were saved on a removable hard disk and collected every week.

At five levels (2.5, 5.0, 8.4, 23.5 & 26.0 m) within and above the canopy measurements were made of the CO_2 concentration using a single channel CIRAS infrared gas analyser (PP Systems, UK). These measurements were taken at each level for five minutes. The profiles thus obtained were time-differenced and vertically integrated to estimate the total change in CO_2 storage. The CIRAS data is used to calibrate off- line the LiCOR system.

An automated weather station took measurements of incoming and reflected solar (Kipp and Zonen CM21) and long wave (Kipp en Zonen, CG1) radiation, soil heat flux (TNO-WS 31 and Hukseflux SH1), windspeed (Vector A101ML), wind direction (W200P) and temperature and relative humidity (Vaisala HMP35A). Note that no measurements of net radiation were taken, as the separate components of the energy balance, shortwave and longwave radiation were determined independently. The incoming long-wave instruments was cooled by a fan to minimise temperature differences between the housing and the environment.

Rainfall was measured above the canopy and in the open field with automated tipping bucket rain gauges. Power was supplied by 12 V batteries connected to solar panels and a wind generator. A diesel generator was used as backup and installed downwind at about 100 m of the tower. During the autumn and winter a considerable part of the required power is supplied by the diesel generator. No effect of the exhaust fumes on the CO_2 signal could be found in the tower CO_2 profile measurements.

Corrections for signal loss of the eddy correlation equipment due to sensor separation, path length, finite instrument response time and tube length were calculated following Moore (1986), Moncrieff et al. (1997) and Aubinet et al. (2000). The reference frame of the co-variances was rotated for every half hourly flux

measurement to align the fluxes perpendicular to the mean streamline. For further details on the methodological aspects of the measurements see Aubinet et al. (2000).

2.3. Gap filling

There are several techniques to fill gaps in time series, such as: (non)-linear interpolation, look up tables, a method based on mean diurnal variations of previous periods (MDV), (multiple) (non)-linear regression, artificial neural networks, or (semi-)empirical models.

As most relations in nature are complex, *(non)-linear interpolation* is only reliable for filling very short gaps (e.g. one 30 min. timestep) if other data series without gaps of more or less physically related variables show the same trend (e.g. temperature and radiation).

Falge *et al*. (2001a) give an example of how to use *look up tables*. Missing measurements of fluxes are looked-up based on meteorological conditions. They used PPFD (photosynthetic photon flux density) and vapor pressure deficit as the driving meteorological variables for latent and sensible heat flux. For the net ecosystem exchange they used air temperature and PPFD. The tables were split up in 6 or 4 parts to represent different seasonal conditions. The advantage of lookup tables if compared with semi-empirical methods such as non-linear regression is that lookup tables do not depend on a fixed response function.

Missing data can be filled using the *mean diurnal variation* method. The missing data are replaced by the mean for that time slot based on measurements of previous and subsequent days. Using mean diurnal variations to fill gaps in data series is based on the assumption that the average of the previous and subsequent periods is representative for the period with the missing data. Falge *et al*. (2001a) recommended for (half-)hourly data averaging windows of 7 or 14 days. Larger averaging intervals are in general not recommended as they will introduce bigger deviations of the mean for each time slot. The strong point of MDV is that there are no data needed of other variables, which is especially useful for remote locations when there are no data available at all. This at the same time is also the weak point of this method, because there is no response to different conditions that may influence the variable to be filled (Falge *et al*., 2001b).

A well calibrated *(semi-)empirical model* based on more or less known physical functions can be used to fill gaps in data series of fluxes. An example of such a model is the often used relation between soil temperature and radiation and NEE (e.g. Lloyd and Taylor, 1994). However, the disadvantage of such models is that they are based on pre-defined functions that are not necessarily representative for the measurements. For example the assumption of energy balance closure although theoretically correct will not always be true for the measurements at a specific site. A site may be influenced by advection or the energy storage in the soil or canopy is not accurately measured, fluxes derived from a model using the assumption of energy balance closure will then be offset compared to the measured fluxes.

Neural networks can be considered a non-linear regression tool and as such are capable of reproducing highly non-linear relations that are common in nature (Huntingford and Cox, 1997). A typical neural network consists of a number of input nodes, a number of nodes in a hidden layer and one output node (see Fig. 1). To ensure the greatest flexibility and enable the network to reproduce the non-linear behaviour of natural processes a sigmoid activation function such as:

$$y = -1 + \frac{2}{1 + e^{-x}}$$

can be used for the nodes in the hidden layer.

To prevent different weights at the initialisation all input signals can be scaled to values between –1 and 1.

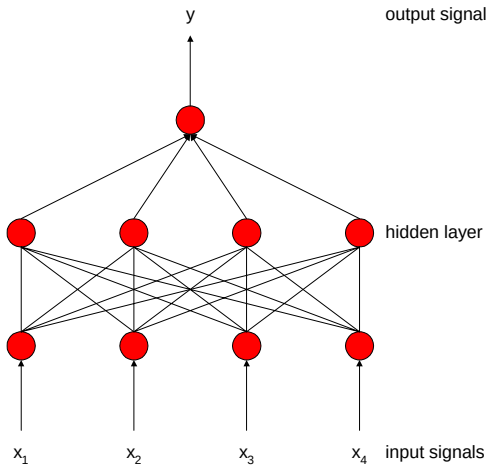


Figure 1: Example of a neural network configuration with 4 input signals, a hidden layer with 4 nodes and 1 output signal.

A distinct advantage of a neural network in gap filling is that it is not necessary to make any assumptions about a physical relation between variables. One can introduce for example time of day or day of year as an input signal.

3. RESULTS

3.1. Automatic Weather Station Data

For weather data there are a number of techniques available which in general are based on comparing data at a specific site with those measured at neighboring sites. Xia *et al*. (1999) tested six different methods to estimate missing values in forest climatology data for Bavaria, Germany. They found that a method based on multiple regression analysis (using the five closest weather stations) with least absolute criteria gave the best estimations. The most accurately estimated variables were maximum and mean temperature, and water vapor followed by minimum temperature. The poorest results were obtained for wind speed and precipitation. A problem arises when the station is located in extended forested area and the neighboring stations are located on

for example grassland. If the forest influence is not taken into account the errors will increase significantly.

An alternative to using neighboring stations is to use other variables measured at the same site and a non-linear regressor such as a neural net. A neural net also gives the possibility of including non-linear relations in the data between different sites. The ability to fit the data depends on the number of degrees of freedom. For a neural network these depend on the number of input variables (i.e. input nodes) and the number of nodes in the hidden layer(s). In Table 1 an example is given of possible network configurations to fill the gaps in the wind speed data at the Loobos site. The neural net was not trained on the wind speed directly, but on the differences in wind speed measured by two sensors.

Variable	Output	Input	Nodes	RMSE
u	$u\text{-}u_{sonic}$	u_{sonic}, u_*, $(z\text{-}d)/L$	2	0.128
u	$u\text{-}u_{level1}$	u_{level1}, u_{level2}, u_{dir}, T	2	0.578
u	$u\text{-}u_{level1}$	u_{level1}, u_{level2}, u_{loc}, R_g	3	0.437
u	$u\text{-}u_{loc}$	u_{loc}, u_{dir}, T R_g	2	0.589

Table 1. Different neural network configurations to fill gaps in wind speed data. In the column Nodes the number of nodes used in the hidden layer is indicated, RMSE is the error presented as the root of the sum of squares divided by the number of residuals. (u_* = friction velocity, (z-d)/L = stability parameter, u_{dir} = wind direction, T = temperature, R_g = global radiation, the subscripts stand for: sonic = sonic anemometer, level1/2 = cup anemometer at different height, loc = anemometer at different location.)

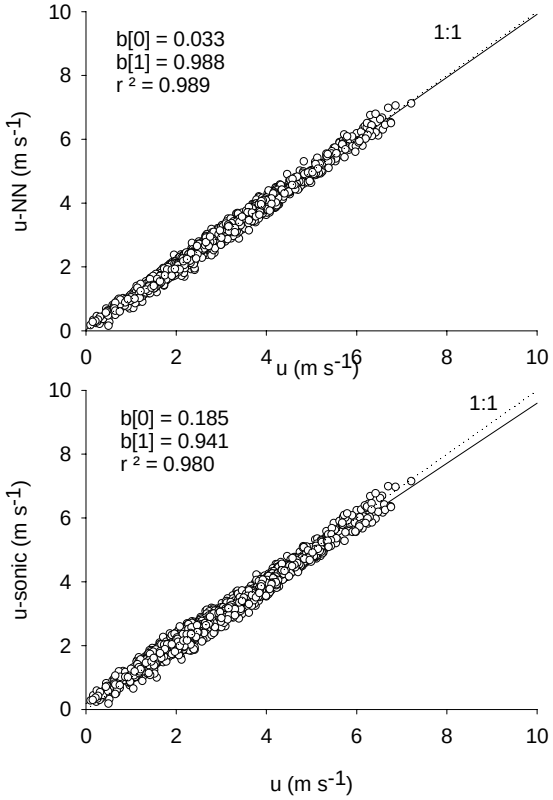
In Fig. 2 the results are plotted of the first neural net (NN) configuration of Table 1. This figure shows that although this specific relation is almost linear the neural net improves the average and the variance explained.

3.2. Flux data

For flux data it is in general not practical to use neighboring stations to fill missing data.

Van Wijk and Bouten (1999) used several combinations of input variables R_g (global radiation), T (temperature), D (vapour pressure deficit), u (wind speed), L_{AI} (leaf area index), T_{ofD} (time of day) and N_{day} (day number) to test the performance of neural networks to determine the half hourly latent heat flux and NEE. They excluded wet day and night time data with wind speeds below 2.5 to 3.0 m s^{-1} to prevent possible side effects due to interception evaporation and CO_2 storage build up during stable nights. The neural net was calibrated for a subset of the remaining data. The validation was done on the data of the same year not used for the calibration. They obtained the worst results

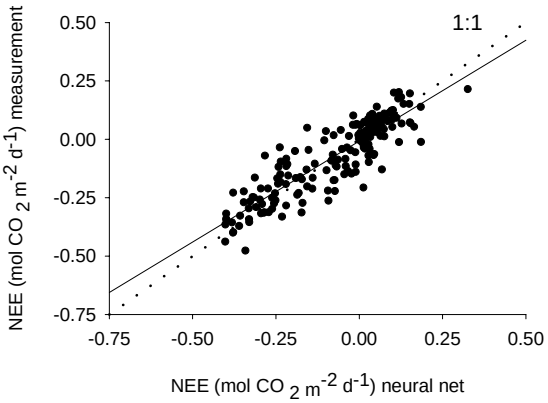
Fig. 2: Wind speed measured by a cup-anemometer (u) versus the derived wind speed using a neural net (NN)



(top graph) and as measured by a sonic anemometer (u-sonic) (bottom graph).

(NRMSE = 0.77 and R^2 = 0.66) using R_g and T as input and 3 nodes in the hidden layer and the best results (NRMSE = 0.69 and R^2 = 0.70) using R_g, T, D, L_{AI} and T_{ofD} as input and also with 3 hidden nodes.

Fig. 3: NEE measured and derived from a neural net.



Because we were not interested in trying to model NEE using a neural net, but only in obtaining a continuous data set, we used a slightly different approach. Instead of half hourly data we used daily data, which reduces the uncertainty in the data. First the missing data for the meteorological and other data were replaced using different neural net configurations as indicated in Table 1. Secondly the missing data in the latent heat flux were replaced. And at last the NEE gaps were filled in using all automatic weather station data, soil moisture data and

fluxes of latent heat. This procedure ensured that we could use nearly all-available information at the site contained in the measurements of the different variables. In Fig. 3 the validation results for 1999 data are plotted for a neural net trained on 1997 and 1998 data. Here the slope of the regression line is 0.863 and the variance explained $r^2 = 0.78$.

4. DISCUSSION AND CONCLUSIONS

From the above it is clear that filling missing data is not an easy and straightforward task. It is clear that the use of semi-empirical models is not recommended because of the risk of creating biased data. For missing precipitation data the use of neighbouring stations is often the only method. For other variables measured at forested sites care should be taken if neighbouring stations are used based on other land use.

For flux data Falge *et al*. (2001a) prefer look-up tables over mean diurnal variation methods as they preserve the response to main meteorological conditions. An other technique that preserves this response are the use of non-linear regressors. Here it was demonstrated that a neural network being a non-linear regressor is a good tool to fill missing data in both meteorological and flux data series. Falge et al. (2001a) showed that different filling techniques may give different results, and thus standardisation of filling techniques is preferable. However, as shown in Table 1. there are a number of different neural net configurations possible. The best configuration depends on the site characteristics as well as the data available. To get the most out off the available data different methods will have to be applied. To ensure correct interpretation of a filled data set, it is recommended to add to each data set information on the filling techniques used as well as an estimation of the accuracy.

Advantages of the use of neural networks are that clear predefined relationships are not needed and it is easy to incorporate daily or seasonal trends by including the time of the day or the day of the year. However, it is also demonstrated here that the results improve if those easy predictable fluctuations are removed prior to the training of the network. The use of neural networks contributes to highly consistent data series. This data series, however is not necessarily accurate, as during the training phase of the network, the possible errors of the object variable are included. Data quality remains the building block of a good data series. A possible advantage of the use of a neural network is also that there is no need to find a consensus model to generate the flux data: consensus on the input data and the transfer function is sufficient.

REFERENCES

Aubinet, N., Grelle, A. Ibrom, M., Rannik, U., Moncrieff, J., Foken, T., Kowalski, A.S., Martin, P. H., Berbigeier, P., Bernhoferr, C., Clement, R., Elbers, J., Granier, A., Grunwald, T., Morgenstern, K., Pilegaard, K., Rebmann, C., Snijders, W., Valentini, R., Vesala, T., 2000. Estimates of the annual net water and carbon exchange of European forests. Adv. Ecol. Res., 30: 113-175.

Dolman,A.J., Moors, E.J., Elbers, J.A., Snijders, W. (1998). Evaporation and surface conductance of three temperate forests in the Netherlands. Ann. Forest., 55: 255-270.

Elbers, J.E., Dolman, A.J., Moors, E.J. en Snijders, W., 1996. Hydrologie en waterhuishouding van bosgebieden in Nederland. Fase II: experimentele opzet en eerste meetresultaten. Rapport 333, Staring Centrum, Wageningen. 76 pp (in Dutch).

Lloyd, J., Taylor, J.A., 1994. On the temperature dependence of soil respiration. Funct. Ecol. 8: 315-323.

Moncrieff J.B., Massheder, J.M., de Bruin, H.A.R., Elbers, J., Friborg, T., Heusinkveld, B., Kabat, P., Scott, S., Soegaard., H. and Verhoef. A., 1996. A system to measure surface fluxes of momentum, sensible heat, water vapour and carbon dioxide. J. Hydrol., 188/189: 589-611.

Moore, C.J., 1986. Frequency response corrections for eddy correlation systems. Bound. Lay. Meteorol., 37: 17-35.

Falge, E., D. Baldocchi, R.J. Olson, P. Anthoni, M. Aubinet, C. Bernhofer, G. Burba, R. Ceulemans, R. Clement, H. Dolman, A. Granier, P. Gross, T. Grünwald, D. Hollinger, N.-O. Jensen, G. Katul, P. Keronen, A. Kowalski, C. Ta Lai, B. E. Law, T. Meyers, J. Moncrieff, E. Moors, J. W. Munger, K. Pilegaard, Ü. Rannik, C. Rebmann, A. Suyker, J. Tenhunen, K. Tu, S. Verma, T. Vesala, K. Wilson, S. Wofsy. 2001a. Gap Filling Strategies for Defensible Annual Sums of Net Ecosystem Exchange. Agr For Meteorol. 107: 43-69

Falge, E., D. Baldocchi, R.J. Olson, P. Anthoni, M. Aubinet, C. Bernhofer, G. Burba, R. Ceulemans, R. Clement, H. Dolman, A. Granier, P. Gross, T. Grünwald, D. Hollinger, N.-O. Jensen, G. Katul, P. Keronen, A. Kowalski, C. Ta Lai, B. E. Law, T. Meyers, J. Moncrieff, E. Moors, J. W. Munger, K. Pilegaard, Ü. Rannik, C. Rebmann, A. Suyker, J. Tenhunen, K. Tu, S. Verma, T. Vesala, K. Wilson, S. Wofsy. 2001b. Gap Filling Strategies for Longterm Energy Flux Data Sets. Agr For Meteorol. 107:71-77

Huntingford, C. and P.M. Cox, 1997. Use of statistical and neural network techniques to detect how stomatal conductance responds to changes in the local environment. Ecol. Model., 80: 217-246.

Van Wijk, M.T. and W. Bouten, 1999. Water and carbon fluxes above European coniferous forests modelled with artificial neural networks. Ecol. Model., 120: 181-197

Xia, Y., P. Fabian, A. Stohl, M. Winterhalter,1999. Forest climatology: estimation of missing values for Bavaria, Germany. Agr. For. Meteorol., 96: 131-144.

* Corresponding author
E.J. Moors
Alterra, PO Box 47, 6700 AC Wageningen, the Netherlands
email: e.j.moors@alterra.wag-ur.nl