

Evaluation of statistical protocols for quality control of ecosystem carbon dioxide fluxes

Jorge F. Perez-Quezada,

Universidad de Chile, Santiago, Chile

Nicanor Z. Saliendra,

US Department of Agriculture Forest Service, Rhinelander, USA

William E. Emmerich

US Department of Agriculture Agricultural Research Service, Tucson, USA

and Emilio A. Laca

University of California at Davis, USA

[Received February 2006. Revised July 2006]

Summary. The process of quality control of micrometeorological and carbon dioxide (CO₂) flux data can be subjective and may lack repeatability, which would undermine the results of many studies. Multivariate statistical methods and time series analysis were used together and independently to detect and replace outliers in CO₂ flux data derived from a Bowen ratio energy balance system. The results were compared with those produced by five experts who applied the current and potentially subjective protocol. All protocols were tested on the same set of three 5-day periods, when measurements were conducted in an abandoned agricultural field. The concordance of the protocols was evaluated by using the experts' opinion (mean \pm 1.96 standard deviations) as a reference interval (the Bland–Altman method). Analysing the 15 days together, the statistical protocol that combined multivariate distance, multiple linear regression and time series analysis showed a concordance of 93% on a 20-min flux basis and 87% on a daily basis (only 2 days fell outside the reference interval), and the overall flux differed only by 1.7% (3.2 g CO₂ m⁻²). An automated version of this or a similar statistical protocol could be used as a standard way of filling gaps and processing data from Bowen ratio energy balance and other techniques (e.g. eddy covariance). This would enforce objectivity in comparisons of CO₂ flux data that are generated by different research groups and streamline the protocols for quality control.

Keywords: Bowen ratio energy balance; Multiple linear regression; Multivariate distance; Net ecosystem carbon dioxide flux; Quality control; Time series analysis

1. Introduction

Independent testing of results in different laboratories is a fundamental requirement of science. This is particularly important for studies that have far-reaching socio-economic implications, such as the quantification of carbon fluxes and inventories in relation to global climate change. Increasing efforts are under way to expand the network of carbon flux measurements by different research groups (Aubinet *et al.*, 2000; Svejcar *et al.*, 1997; Baldocchi *et al.*, 2000). Modellers

Address for correspondence: Jorge F. Perez-Quezada, Departamento de Ciencias Ambientales y Recursos Naturales Renovables, Facultad de Ciencias Agronómicas, Universidad de Chile, Casilla 1004, Santiago, Chile. E-mail: jorgepq@uchile.cl

use such data to calibrate and validate models to extrapolate and predict carbon budgets over space and time (e.g. Schimel *et al.* (2001), Ito and Oikawa (2002) and Melesse and Hanley (2005)). Thus, it is essential to assess the potential inconsistencies and differences between data sets that could be attributed to quality control (QC) by different individuals. QC is defined as

‘a system of routine technical activities to measure and control the quality of the inventory as it is being developed’

(Intergovernmental Panel on Climate Change, 2000).

Current knowledge about carbon fluxes is largely based on micrometeorological measurements with Bowen ratio energy balance (BREB) and eddy covariance systems. Individual experts in different laboratories subject data from these measurements to QC, and it is not known how consistent this process is across experts. Carbon flux measurements with BREB and eddy covariance systems involve several signals that are sampled at rates as high as 10 Hz. Raw data are typically aggregated to one set of values every 10–30 min. Aggregated data are inspected by an expert who eliminates observations that are identified as ‘incorrect’ and fills the gaps that are created by violations to micrometeorological assumptions, equipment malfunction and poor weather (Falge *et al.*, 2001). Typically, outliers in BREB and eddy covariance data are replaced with linearly interpolated values based on records surrounding the erroneous measurements. This method is appropriate only when the gaps are much shorter than the shortest meaningful pattern in the data, and it ignores the potentially rich information that is contained in the auto-correlations and cross-correlations of the signals. This QC process depends both on clear rules and on subjective decisions made by each expert.

The subjectivity of the QC process may affect both the identification of observations as erroneous and the values that are used to fill gaps. It is therefore very important to develop a protocol that can be used as a standard method among scientists. When applied to the same raw data, the protocol should always yield identical results, regardless of who performs the QC.

Scientists working on carbon flux measurements have not used the body of theory and techniques that is available for the analysis of signals (e.g. Chatfield (2004)), although the main issue in QC is a smoothing problem. A ‘smoother’ estimates the state of a system (i.e. values of all signals from the flux system) at time t on the basis of measurements before and after t . Estimation can be purely empirical or it can be based on a mechanistic model of the system. The benefits of the latter are multiple, but it requires detailed knowledge of the ecosystem characteristics and climate conditions.

The objective of this work is to test the use of multivariate statistics, time series analysis and a combination of both to define QC protocols. We evaluate the concordance of these statistical protocols with the current, potentially subjective, method based on expert opinions. We assess the degree of agreement between methods because it is not possible to know which one is correct. We use the statistical protocols for both outlier detection and prediction of carbon dioxide (CO₂) flux data, as measured by a BREB system. After testing several protocols, we report here on the four protocols that showed better agreement with the average opinion of an expert panel, who independently applied the current QC protocol to the same three data sets.

2. Methods

2.1. Site and data description

The study area is in northern Kazakhstan, near Shortandy (51° 37' N–71° 05' E), 60 km north of Astana. The climate is cold semiarid with average annual precipitation 323 mm, 60% of which occurs between May and September. The average temperature for this same period is 15.8 °C

whereas the annual average is 1.6 °C (Gilmanov *et al.*, 2004). Considering the June–October period, the 2002 growing season was very close to that of the average year, with a total precipitation of 180 mm and average temperature of 13.0 °C. These values are almost identical to the averages for 1936–2000 (188.4 mm and 13.9 °C respectively).

This area is in the Kazakh steppe ecoregion (Olson *et al.*, 2001) and has not been cultivated by wheat farmers during the last decade because of the effects of economic policy changes on input–output prices and subsidies. Abandoned lands are currently dominated by weeds (genera *Sonchus* and *Cirsium*) and shrubs (genera *Artemisia* and *Matricaria*) and are undergoing secondary succession (Laca *et al.*, 2003).

Flux measurements were done continuously with a BREB tower from May 18th until November 22nd (days of the year (DOYs) 138–326) of 2002, except for occasional failures of the equipment. Three sets of BREB data consisting of five consecutive days each were selected from different periods of the growing season, when canopy heights were 30, 90 and 80 cm. Each period consisted of a time series of 360 data, which is the result of averaging the BREB data every 20 min during 5 days. All three periods included complete sets of micrometeorological variables and CO₂ and water fluxes (FCO₂ and FH₂O), some of which are reported in Table 1. In this study, negative values of FCO₂ indicate ecosystem net CO₂ uptake, as photosynthesis exceeds ecosystem respiration during the day, whereas positive values denote net CO₂ efflux from the ecosystem into the atmosphere. Even though the first period (PE-1) and the second period (PE-2) had different micrometeorological conditions, they had similar average daily FCO₂ (approximately $-17.5 \text{ g m}^{-2} \text{ s}^{-1}$) (Table 1). The third period (PE-3) had a much lower average flux ($0.61 \text{ g m}^{-2} \text{ s}^{-1}$), which represents respiration because it occurred during vegetation senescence.

2.2. Bowen ratio energy balance technique

A BREB system (model 023/CO₂, Bowen ratio system, Campbell Scientific Inc., Logan, Utah, USA) was used for continuous monitoring of FCO₂ and FH₂O between the atmosphere and the abandoned field. The Bowen ratio β is defined as the ratio between sensible heat H and latent heat flux LE (Bowen, 1929). FCO₂ was indirectly estimated by the BREB system as described in detail by Campbell Scientific (1998). The equations that are used in this study are contained in Appendix A, and the complete list of variables measured and calculated is given in Table 2. For details about instrumentation see Gilmanov *et al.* (2004).

The FCO₂ data were corrected for changes in CO₂ density that were caused by the simultaneous flux of water vapour (Webb *et al.*, 1980), as described by Campbell Scientific (1998). The correction for the effect of H -flux was not performed because it has been found that the air streams entering the infra-red gas analyser sample and reference cells have the same temperature (Angell *et al.*, 2001).

2.3. Quality control process

The analysis of QC protocols was divided in two sections. The first evaluates the current protocol that is used by several scientists who are specialists in the field of biometeorology. The second part contains details of four alternative protocols that are based on statistical methods, which showed higher agreement with the current protocol. Agreement is used in the sense of Bland and Altman (1999), who suggested the use of the standard method as a ‘gold standard’. This does not mean that it is free of error, but instead that it is our ‘best guess’ of the true flux value. In the case of this study, we use the average and variability of the experts’ opinion (mean \pm 1.96 standard deviations (SDs)) as the limits of agreement. Agreement of each protocol is quantified as the percentage of data points that fall within this reference interval.

Table 1. Micrometeorological characteristics† of the three sampling periods and results of current and statistical QC protocols (daily average CO₂ flux)

Period	DOY	Micrometeorological and original flux					QC by experts					Statistical protocols				
		T _a (°C)	RH (%)	U (m s ⁻¹)	F (g m ⁻² day ⁻¹)	\bar{F} (g m ⁻² day ⁻¹)	SD (g m ⁻² day ⁻¹)	NOM (%)	PR-A	PR-B	PR-C	PR-D	PR-A	PR-B	PR-C	PR-D
1	162	15.2	73	5.4	-14.00	-19.51	0.44	9	-19.30	-19.94	-19.54	-19.19	-19.30	-19.94	-19.54	-19.19
1	163	13.1	68	3.7	-24.28	-18.54	0.20	11	-20.27‡	-19.34‡	-18.03‡	-18.10‡	-20.27‡	-19.34‡	-18.03‡	-18.10‡
1	164	13.1	80	3.6	-31.47	-16.93	1.43	18	-18.33	-18.50	-10.81‡	-18.29	-18.33	-18.50	-10.81‡	-18.29
1	165	12.8	74	3.8	-19.24	-17.23	1.23	15	-18.42	-18.00	-17.55	-18.53	-18.42	-18.00	-17.55	-18.53
1	166	15.1	57	2.7	4.59	-14.97	0.91	9	-13.35	-17.19‡	-14.18	-13.44	-13.35	-17.19‡	-14.18	-13.44
2	202	12.6	76	1.8	7.54	-17.23	4.72	14	-18.89	-18.74	-19.15	-19.15	-18.89	-18.74	-19.15	-19.15
2	203	14.0	69	1.4	8.08	-20.38	3.29	16	-26.24	-25.30	-23.95	-20.88	-26.24	-25.30	-23.95	-20.88
2	204	17.2	72	1.6	75.98	-11.93	12.18	27	-12.65	-19.56	-2.20	-3.52	-12.65	-19.56	-2.20	-3.52
2	205	18.2	74	1.2	5.79	-19.13	3.58	20	-5.68‡	-12.80	-9.79‡	-9.45‡	-5.68‡	-12.80	-9.79‡	-9.45‡
2	206	18.3	65	2.2	8.08	-19.17	0.87	12	-13.28‡	-20.12	-7.19‡	-7.23‡	-13.28‡	-20.12	-7.19‡	-7.23‡
3	257	3.3	62	3.4	0.56	0.40	0.15	5	0.31	0.28	0.42	0.40	0.31	0.28	0.42	0.40
3	258	2.0	56	1.7	3.31	1.54	0.56	15	2.85‡	1.88	1.82	1.36	2.85‡	1.88	1.82	1.36
3	259	5.1	47	2.0	4.59	2.61	0.47	17	3.25	2.87	2.82	2.82	3.25	2.87	2.82	2.82
3	260	8.4	39	2.8	-1.08	-1.06	0.10	9	-1.07	-1.08	-1.14	-1.21	-1.07	-1.08	-1.14	-1.21
3	261	14.3	21	2.4	0.52	-0.46	0.45	11	-0.10	-0.45	-0.13	-0.10	-0.10	-0.45	-0.13	-0.10

†T_a, air temperature; RH, relative humidity; U, wind speed; F, original (uncorrected) average flux; \bar{F} , average flux corrected by the expert panel; SD, standard deviation calculated from five average values generated by the expert panel (see Section 2.3.1); NOM, average percentage of records modified by the expert panel.
‡Daily average fell outside the reference interval ($\bar{F} \pm 1.96$ SDs).

Table 2. Variables that were used in the multivariate distance method and the subset that was used for multiple linear regression

<i>Variable</i>	<i>Description</i>	<i>Unit</i>	<i>Multiple linear regression</i>
<i>Measured variables</i>			
∂CO_2	CO ₂ concentration gradient between arms	$\mu\text{mol mol}^{-1}$	
$\partial\text{H}_2\text{O}$	Water vapour concentration gradient between arms	mmol mol^{-1}	
∂T	Temperature gradient between arms	$^{\circ}\text{C}$	
CO ₂ reference	Ambient CO ₂ concentration	$\mu\text{mol mol}^{-1}$	✓
T _{air}	Air temperature	$^{\circ}\text{C}$	✓
e _a	Water vapour pressure	mbar	✓
R _n	Net radiation	W m^{-2}	✓
G ₁	Soil heat flux from sensor 1	W m^{-2}	
G ₂	Soil heat flux from sensor 2	W m^{-2}	
T _{soil}	Soil temperature	$^{\circ}\text{C}$	✓
RH	Relative humidity	%	✓
BV	Battery voltage	V	
<i>U</i>	Wind speed	m s^{-1}	✓
PCPN	Total precipitation	mm	
CS615	Volumetric soil water content	$\text{m}^3 \text{m}^{-3}$	
PAR	Photosynthetically active radiation	$\mu\text{mol m}^{-2} \text{s}^{-1}$	✓
Tirga	Infra-red gas analyser temperature	$^{\circ}\text{C}$	
<i>Calculated variables</i>			
SoilH ₂ O	Volumetric soil water content (corrected with gravimetric measures)	$\text{m}^3 \text{m}^{-3}$	✓
<i>C_p</i>	Soil heat capacity	$\text{J m}^{-3} \text{K}$	
<i>G</i>	Heat flux at the soil surface	W m^{-2}	✓
λ	Latent heat of vaporization	J g^{-1}	
β	Bowen ratio	—	
<i>H</i>	Sensible heat flux	W m^{-2}	
ρ_a	Air density	g m^{-3}	
$K_h (K_c)$	Eddy diffusivity for heat; assumed the same for CO ₂	$\text{m}^2 \text{s}^{-1}$	
$d\rho_c$	CO ₂ density differential between arms	g m^{-3}	
ρ_c/ρ_a	CO ₂ density–air density ratio	—	
σ	Webb <i>et al.</i> (1980) correction component	—	
FH ₂ O	Water vapour flux	$\text{g m}^{-2} \text{s}^{-1}$	
K_h	K_h -value used to calculate the CO ₂ flux	$\text{m}^2 \text{s}^{-1}$	
FCO ₂	CO ₂ flux (Webb <i>et al.</i> (1980) corrected)	$\text{mg m}^{-2} \text{s}^{-1}$	
du/dz	Derivative of horizontal <i>U</i> with respect to vertical <i>U</i>	—	
Ri	Richardson number	—	
$\phi(m)$	ϕ -parameter in calculation of K_h	—	
K_H	Eddy diffusivity calculated from aerodynamic method	$\text{m}^2 \text{s}^{-1}$	

2.3.1. Current protocol

To evaluate the current QC method, the three selected data sets were sent to five scientists who are currently using the BREB technique. The specialists independently applied the QC protocol to the data sets in the way that they normally do with their own data. The results were compared to estimate the variability of the results between scientists. These scientists look for BREB problems and spikes in the data and make linear interpolations over the bad data. Their criteria are based on their experience. All experts have equivalent expertise working

with the BREB technique on range land ecosystems; thus the comparison was considered valid.

There are two known problems with the BREB technique when estimating fluxes, which are used in the current QC protocol and are known as Ohmura's criteria. The first case occurs when β approaches -1.0 , because the calculation becomes unstable and yields abnormally high or low FCO_2 values (Ohmura, 1982). The second problem is that the eddy diffusivity for CO_2 (K_c) may not be valid when the direction of H (or LE) has opposite sign from the temperature (or water vapour) gradient (Ohmura, 1982). These two cases have been included in proposed guidelines for QC of flux data using the BREB technique (Payero *et al.*, 2003; Heiser and Sellers, 1995). An alternative method for the opposite flux case is known as the aerodynamic method (Dugas *et al.*, 1999) and it calculates K_c by using wind speed, atmospheric stability and canopy height.

Four of these experts (E1, E2, E3 and E5) use a spreadsheet that was originally created by William Dugas (Blackland Research Center, Temple, Texas, USA) for the US Department of Agriculture 'Rangeland carbon dioxide flux project' (Svejcar *et al.*, 1997). This spreadsheet contains all the formulae for flux calculations plus tests for the Ohmura criteria and graphs to help the user to perform the QC process more interactively. These four experts also share some common criteria when reviewing the data and deciding which observations are outliers.

- (a) Anomalous observations are examined for all pertinent meteorological conditions—such as photosynthetically active radiation, air temperature and precipitation—that exist at the time that the flux spikes occur. If there is no reason for a spike to occur, then that datum or data are linearly interpolated.
- (b) Graphs are examined for general trends and obvious outliers (spikes) are interpolated first; then fine-tuning is done over the rest of the data.
- (c) Ohmura's criteria for opposite flux sign and potentially problematic β -values (typically $-1.3 < \beta < -0.7$) are considered as 'flags' for outlying observations.
- (d) Other flags for outliers are those which define 'fast changing' and 'large fluxes', which in this case were set to 0.2 and $0.75 \text{ mg CO}_2 \text{ m}^{-2} \text{ s}^{-1}$ respectively.
- (e) Because there is a sequential order of the equations for calculating FCO_2 (see Appendix A), the order of interpolation of variables follows the order H , FH_2O and FCO_2 .

Expert E4 uses a computer program for the QC process which is based on setting monthly thresholds for eliminating FH_2O - and FCO_2 -values that are too low, changing too rapidly compared with the previous row or too high. A linear smoother is built with the 'good' records. This linear model smooths all the good values and interpolates the 'bad' values unless there are more than four consecutive bad fluxes, in which case the observation is marked as 'unrepairable' and left out of the daily flux calculation. We considered this case as representative of the current protocol because its result still depends on the expert judgment to define the thresholds.

2.3.2. Statistical methods for developing quality control protocols

We tested the use of multivariate distance (MD), multiple linear regression (MLR) and time series analysis (TSA) for creating a QC protocol that would always yield the same result. A record was defined as an outlier when its standardized residual from a statistical model fell outside the ± 3 SD range. All outliers were removed and gaps were filled as described below.

The Mahalanobis distance was used for detecting multivariate outliers. This MD method was chosen because, when variables are highly correlated in a multivariate sense, it is possible for a point to be unremarkable when seen along one axis but still to be an outlier when it represents an unusual combination of scores on two or more variables (Tabachnick and Fidell, 1996).

Multiple linear regression models were built to obtain predictions of FCO₂ for those records which were detected as outliers by the MD or TSA methods (which are described below). The data that were used to build the MLR models were the observations of all five days remaining after the deletion of outliers. Stepwise MLR was used to select from 10 possible independent variables (Table 2) and to calculate the parameter estimates. We used Cook's distance to determine whether outliers were influential in the MLR analysis (Neter *et al.*, 1996). We also used TSA to detect and replace outlying observations in the time series of FCO₂. Autoregressive integrated moving average models are used in TSA basically for predicting values at time t based on past values (Shumway and Stoffer, 2000).

Several statistical protocols were generated by combining the use of MD, MLR and TSA. In the next section we describe and analyse the four protocols that showed better agreement with the opinion of the expert panel. For the statistical analyses we used the packages JMP IN (Sall *et al.*, 2001) and the freely available R (R Development Core Team, 2004).

3. Results and discussion

3.1. Current quality control protocol

Corrected daily fluxes (average) and the variability of opinion (SD) were fairly stable within periods (Table 1 and Fig. 1). The uncorrected flux F was the only sampled variable that showed a significant relation with SD ($SD = 1.83 + 0.109F$, $R^2 = 0.70$; $P = 0.001$). The other two variables that showed a relationship with SD are the magnitude of the average corrected flux (proportional) and wind speed (inversely proportional) (Table 1). Apparently a lower wind speed biases towards more positive values of F . The highest variability of opinion ($SD = 12.18 \text{ g m}^{-2} \text{ day}^{-1}$) was observed on DOY 204 and coincided with an original (uncorrected) flux value of $76 \text{ g m}^{-2} \text{ day}^{-1}$. Given the combination of factors, period PE-3 had the lowest variability of opinion between experts. The expected proportional effect of precipitation on SD was not clearly observed (Fig. 1). Although period PE-1 had the highest amount of rain in 5 days (29.9 mm), it did not show the highest variability between experts (Fig. 1).

The other variable that we observed to have an effect on SD is the number of 'medium-size' spikes in the series. These are the spikes that are left after the extremely high spikes have been interpolated (usually there is no disagreement in those). If there are many of these medium-size spikes in a series, the QC process will become more variable among experts in terms of which spikes are interpolated. The suspected outliers (spikes) were concentrated during sunrise, sunset and night (Figs 2, 3 and 4), which are the periods when this technique is known to yield erroneous measures (according to Ohmura's criteria).

These results show that, even though the general method to correct this type of data set was similar among experts, the final result that was obtained by different experts varied. This means that when comparing fluxes that are measured by different research groups there is an error (which is expressed here as the variability of opinion when working on identical data sets) that is not currently being estimated. Part of these differences is probably masked when annual budgets are calculated, but this does not make the current protocol better.

3.2. Statistical quality control

In this section we describe four protocols that were defined after several trials and combinations of statistical methods. Briefly, protocol PR-A uses MD to detect outliers on FCO₂ data and MLR to replace those values; protocol PR-B works on the result of PR-A, adding the use of

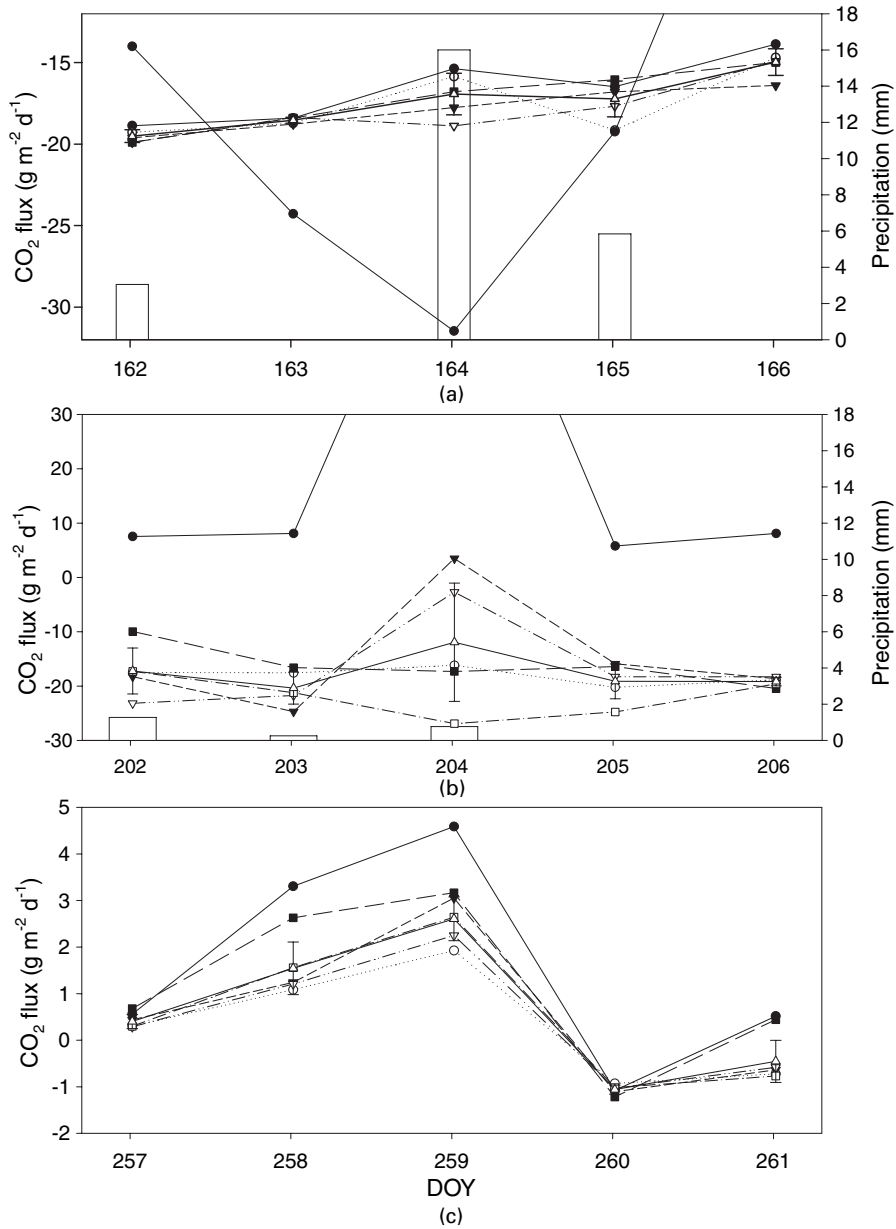


Fig. 1. Original and quality-controlled daily CO₂ fluxes for periods (a) PE-1, (b) PE-2 and (c) PE-3 (values for the original (uncorrected) series on DOY 166 and 204 (4.6 and 76.0 g m⁻² day⁻¹) are not shown to preserve the scale of the plot; error bars are SDs); □, daily precipitation; ●, original; ○, E1; ▼, E2; ▽, E3; ■, E4; □, E5; △, average

TSA to detect a second set of outliers and also uses MLR to replace the deleted values; protocol PR-C uses only TSA on a larger number of variables that influence the calculation of FCO₂, the key step being the definition of potentially faulty β -values; protocol PR-D is similar to protocol PR-C but uses a wider range for defining faulty β -values, which should increase the number of outliers detected.

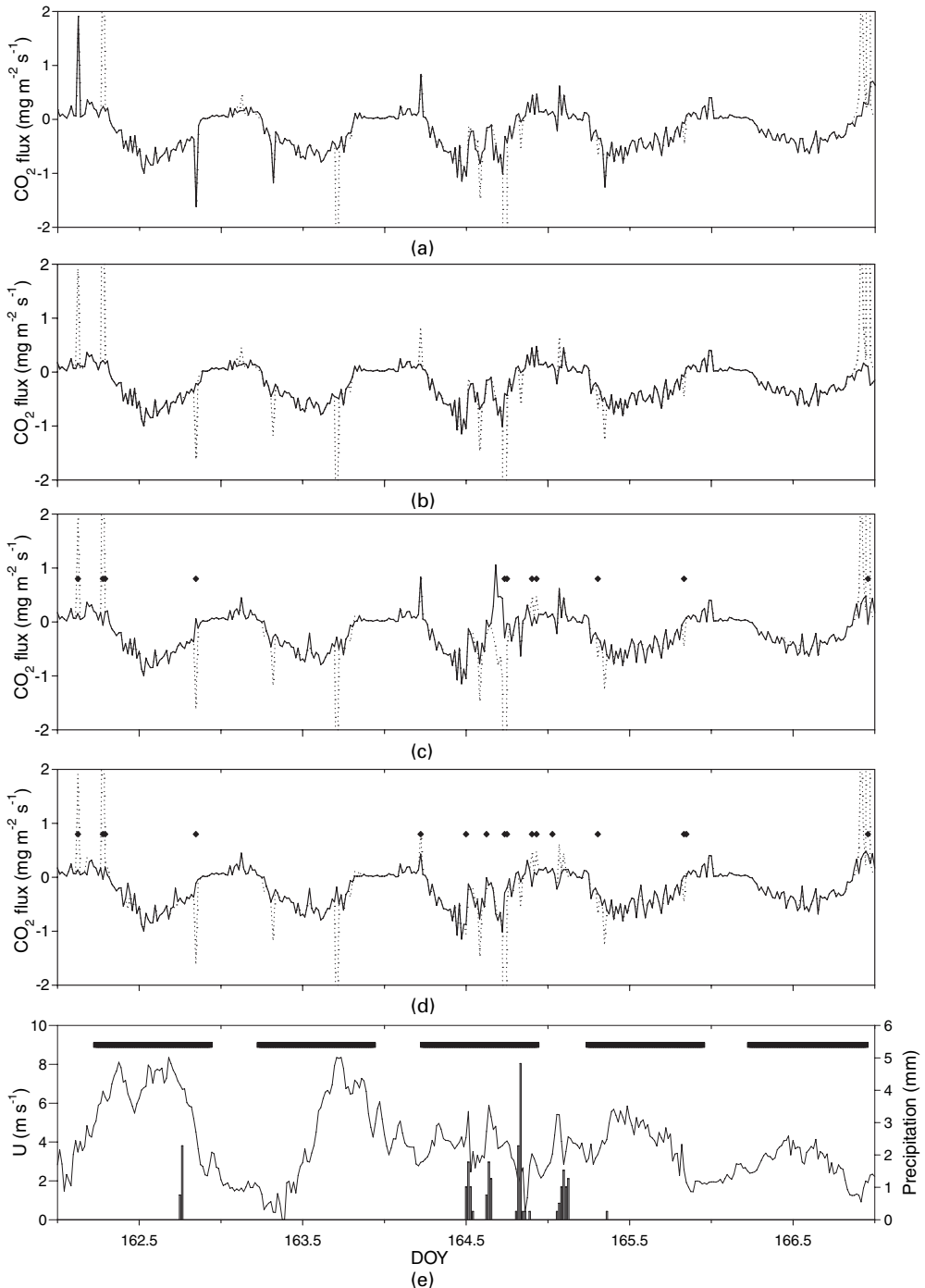


Fig. 2. Original (·····) and quality-controlled (—) FCO₂ 20-min average series for period PE-1 by using protocols (a) PR-A, (b) PR-B, (c) PR-C (◆, $-1.25 < \beta < -0.75$) and (d) PR-D (◆, $-1.4 < \beta < -0.6$), and (e) wind speed (—) and precipitation (■) (—, daytime period): when the original series is not visible it means that the simulation did not differ from it; out-of-scale values (from -11.1 to 12.9 mg m⁻² s⁻¹) are not shown to highlight the differences between protocols

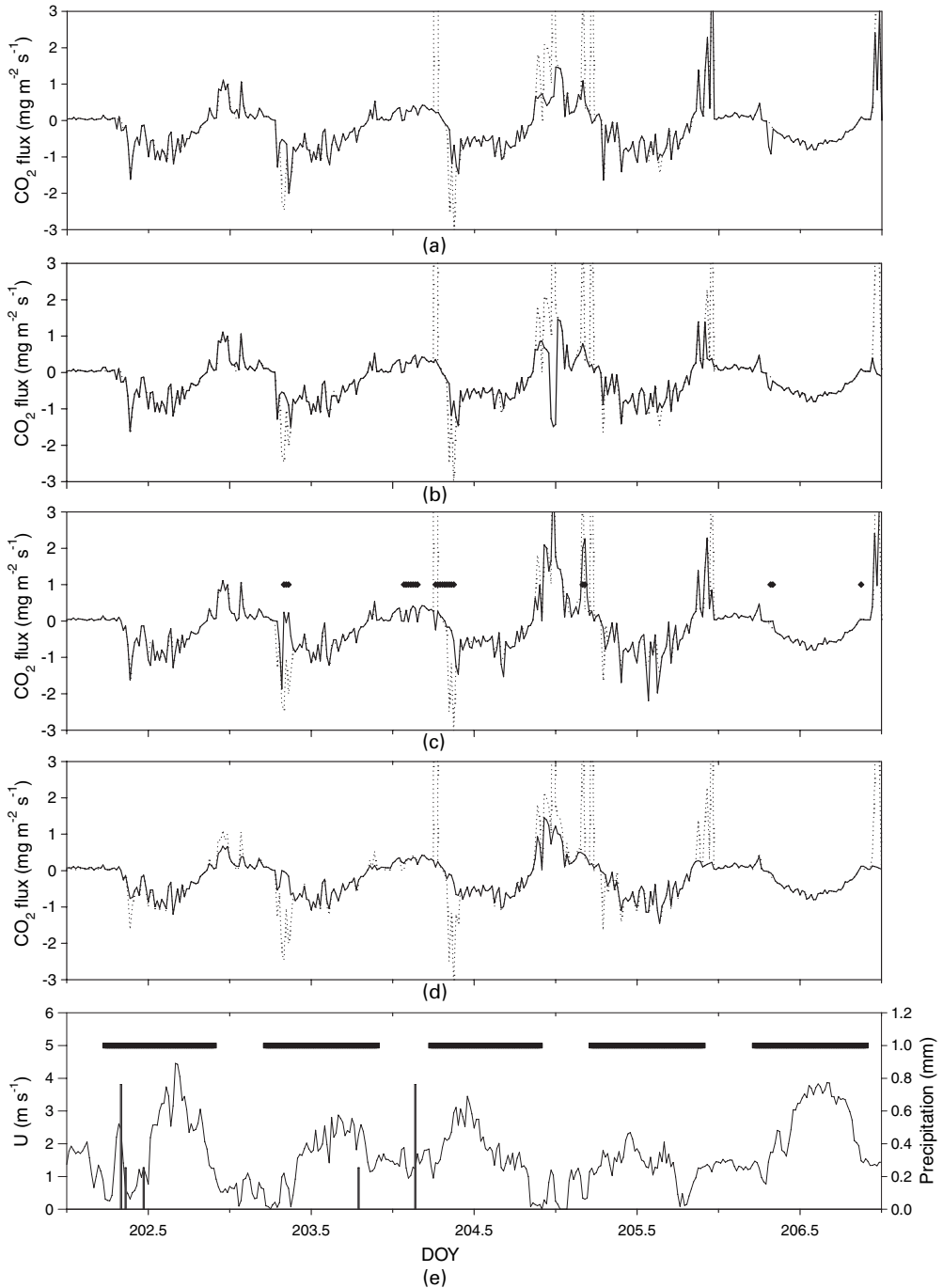


Fig. 3. Original (.....) and quality-controlled (—) FCO₂ 20-min average series for period PE-2 by using protocols (a) PR-A, (b) PR-B and (c) PR-D (◆, $-1.4 < \beta < -0.6$), (d) the average opinion of the expert panel and (e) wind speed (—) and precipitation (■) (—, daytime period): when the original series is not visible it means that the simulation did not differ from it; note the different flux scale compared with Fig. 2; out-of-scale values (from -3.1 to 116 mg m⁻² s⁻¹) are not shown to highlight the differences between protocols

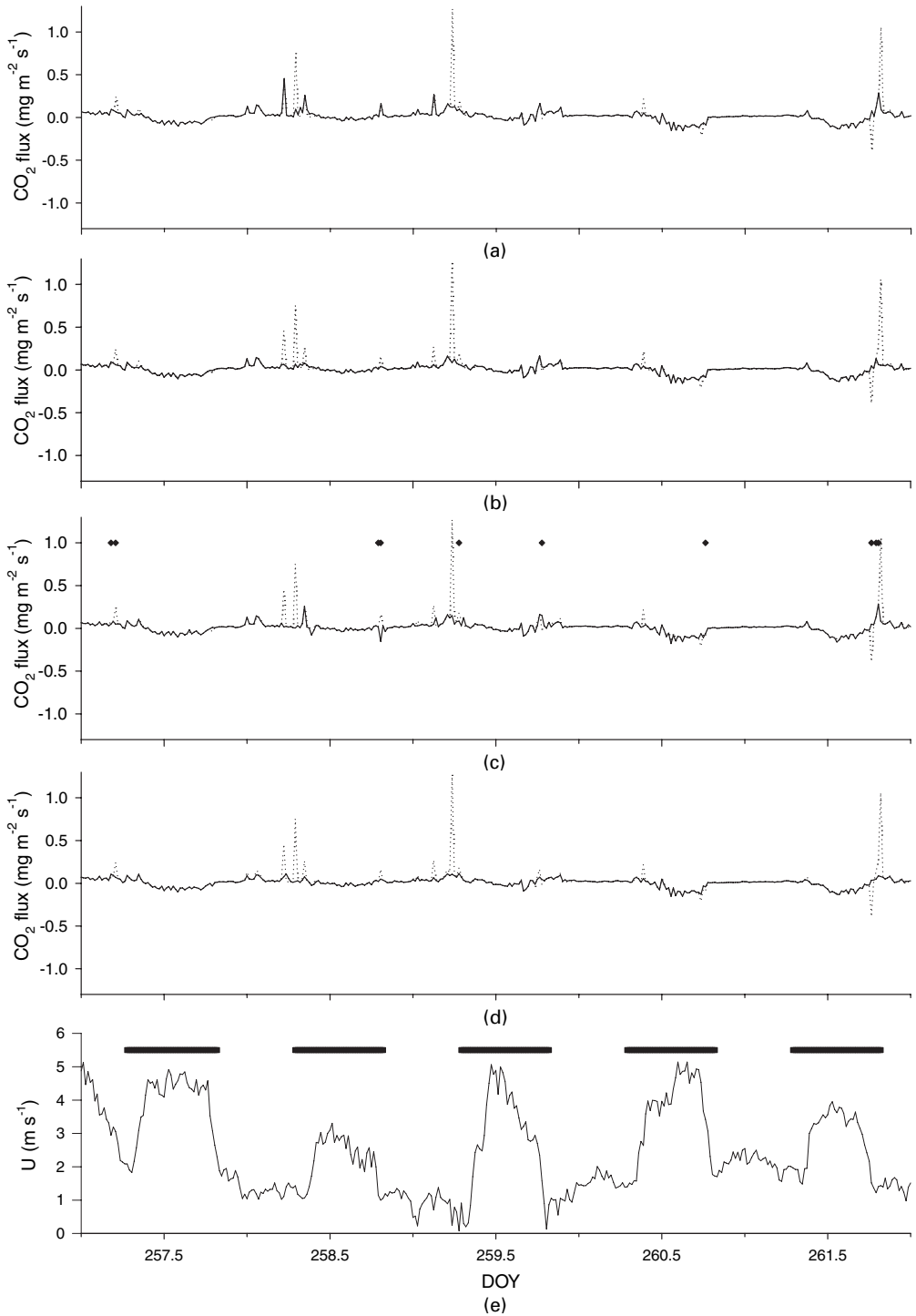


Fig. 4. Original (· · · · ·) and quality-controlled (—) FCO₂ 20-min average series for period PE-3 by using protocols (a) PR-A, (b) PR-B and (c) PR-D (◆, $-1.4 < \beta < -0.6$), (d) the average opinion of the expert panel and (e) wind speed (—) (—, daytime period): note the different flux scale from those in Figs 2 and 3

Protocol PR-A is a combination of MD and MLR methods. The first step was to use MD to detect outliers, for which we used a large number of recorded and calculated variables (Table 2). Using the Mahalanobis distance method, the outliers detected represented 8%, 10% and 7% of the total records on each period, which is lower than the number of modifications that were done by the experts (Table 1). For the remaining records, an MLR model was fitted to obtain a prediction for the 'deleted' records and to replace the corresponding FCO₂-values. The stepwise (forward) method selected the significant variables from an original pool that included the 10 variables that are marked in Table 2, as well as their square and cross-products (65 variables in total). The variables were selected as predictors among the complete set because they were directly measured (except volumetric soil water content and G), not directly involved in the calculation of FCO₂ but are known to be correlates of FCO₂ (the response variable).

The number of variables that were selected was fairly stable between periods, with 19, 23 and 22 variables (plus the intercept) selected for each period. Given that the outlying observations were eliminated before building the models, the adjusted R^2 -values were fairly high (0.72, 0.61 and 0.63 for periods PE-1, PE-2 and PE-3 respectively). When the same variables selected by the stepwise method were used to fit a model to the uncorrected series (360 records), the adjusted R^2 -values were reduced to 0.27, -0.0075 and 0.29. However, using Cook's D-method it was found that no record was extremely influential and therefore no correction was needed. Figs 2(a), 3(a) and 4(a) show the resulting FCO₂-series from applying protocol PR-A, plotted together with the uncorrected series. For period PE-1, although the extremely high and low 20-min flux values were replaced, some spikes are still left. Different is the case in period PE-2, where at the end of DOY 205 and 206 some big spikes are undetected as outliers (Fig. 3(a)).

Protocol PR-B was based on the use of TSA to detect a second set of outliers in the FCO₂ series resulting from protocol PR-A. The purpose was to make use of the information that is given by multiple related variables (protocol PR-A) and the predictive value of past observations (TSA). A second-order autoregressive model (AR(2)) was found to fit better this type of time series. This AR(2) model was fitted to the FCO₂ series and the outlying records were replaced with the prediction from a new MLR model. The result from protocol PR-B (Figs 2(b), 3(b) and 4(b)) was a more smooth series than from protocol PR-A. The outliers detected represented 11%, 12% and 9% of the total for each period, which is still lower than the number of modifications that were done by the experts (Table 1).

Protocol PR-C was developed as an imitation of the sequence of steps that the expert panel followed to do the QC process, as described above. An AR(2) model was fitted to the time series for each variable that is involved in the calculation of FCO₂, and the outliers were replaced with the values that were predicted by the model. The resulting time series at each step were pasted back to the EXCEL spreadsheet so that the calculated variables (e.g. FCO₂) were recalculated. The five-step protocol PR-C consisted in replacing the outliers in key variables of the calculation of FCO₂ in the following order:

- step 1*—directly measured gradients of CO₂, water vapour and temperature (∂CO_2 , $\partial\text{H}_2\text{O}$ and ∂T) and ground heat flux at the surface (G);
- step 2*—sensible heat (H); for all those records when $-1.25 < \beta < -0.75$, H was linearly interpolated and then an AR(2) model was fitted; those records that were interpolated and those detected as outliers were replaced with the model prediction;
- step 3*—water vapour flux (FH₂O); this step does not affect the calculation of K_c but does influence the value of FCO₂;
- step 4*—eddy diffusivity of CO₂ (K_c);
- step 5*—CO₂ flux (FCO₂).

Steps 2–5 were applied to the series that resulted from the correction of the variable in the previous step. Fig. 2(c) shows the resulting FCO₂ series from applying protocol PR-C, plotted together with the uncorrected series for period PE-1. It can be observed that most of the extremely high and low 20-min flux values were replaced but some spikes are still left, especially on day 164. This is probably because the range that was used to interpolate *H* in step 2 ($-1.25 < \beta < -0.75$) is considered to be conservative. The outliers detected represented 30%, 17% and 12% of the total for each period, which means that for period PE-1 it modified a much higher number of records than the experts, but for the other two periods it was very similar (Table 1).

Protocol PR-D is almost identical to protocol PR-C, with the only exception that the range that was used for interpolating *H* in step 2 was increased to $-1.4 < \beta < -0.6$. By comparing Fig. 2(d) with Fig. 2(c) it is clear that protocol PR-D creates a smoother series that should be closer to the expert panel’s opinion, because it tends to eliminate almost all the spikes that are visible in the time series graph. For period PE-2, protocol PR-D did not perform well because of some spikes that were not detected as outliers on DOY 204 and 205 (Fig. 3(c)). In period PE-3, protocol PR-D resembled very well the series product of the expert panel average (Figs 4(c) and 4(d)). The outliers detected increased by only 2% in each period compared with the number from protocol PR-C.

3.3. Comparison of statistical protocols

The four protocols differed not only in the final daily fluxes but also in the number of points that each edited from the original data sets. The number of points edited increased from protocol PR-A to protocol PR-D. This was fully expected according to the way that these protocols were defined. When individual 20-min data points that resulted from the statistical protocols were compared with their corresponding reference interval, all four protocols showed concordance around 90% (Table 3).

The results of daily fluxes of the four protocols for each period are plotted in Fig. 5, in relation to the reference interval defined from the experts’ opinion. The individual values of daily fluxes and their agreement with the reference interval are also summarized in Table 1. The concordance on a daily basis was lower than for 20-min values because the total number of days was low (15). Note for example that the protocol with highest concordance (87%, protocol PR-B), had only 2 days that fell outside the limits of agreement. Protocol PR-D showed higher concordance than

Table 3. Concordance and difference of statistical protocols from the instantaneous, daily and overall average fluxes as determined by the expert panel

Protocol	Concordance (%)		Overall difference†	
	20-min	Day	Flux (g CO ₂ m ⁻²)	Percentage (%)
PR-A	91	73	28.1	14.8
PR-B	91	87	3.2‡	1.7
PR-C	90	73	50.6	26.7
PR-D	89	80	44.7	23.6

†Overall difference calculated from the average overall flux of the five experts, for the 15 days that were used in the study (-189.2 g CO₂ m⁻²).

‡Fell within the reference interval (mean ± 1.96 SDs).

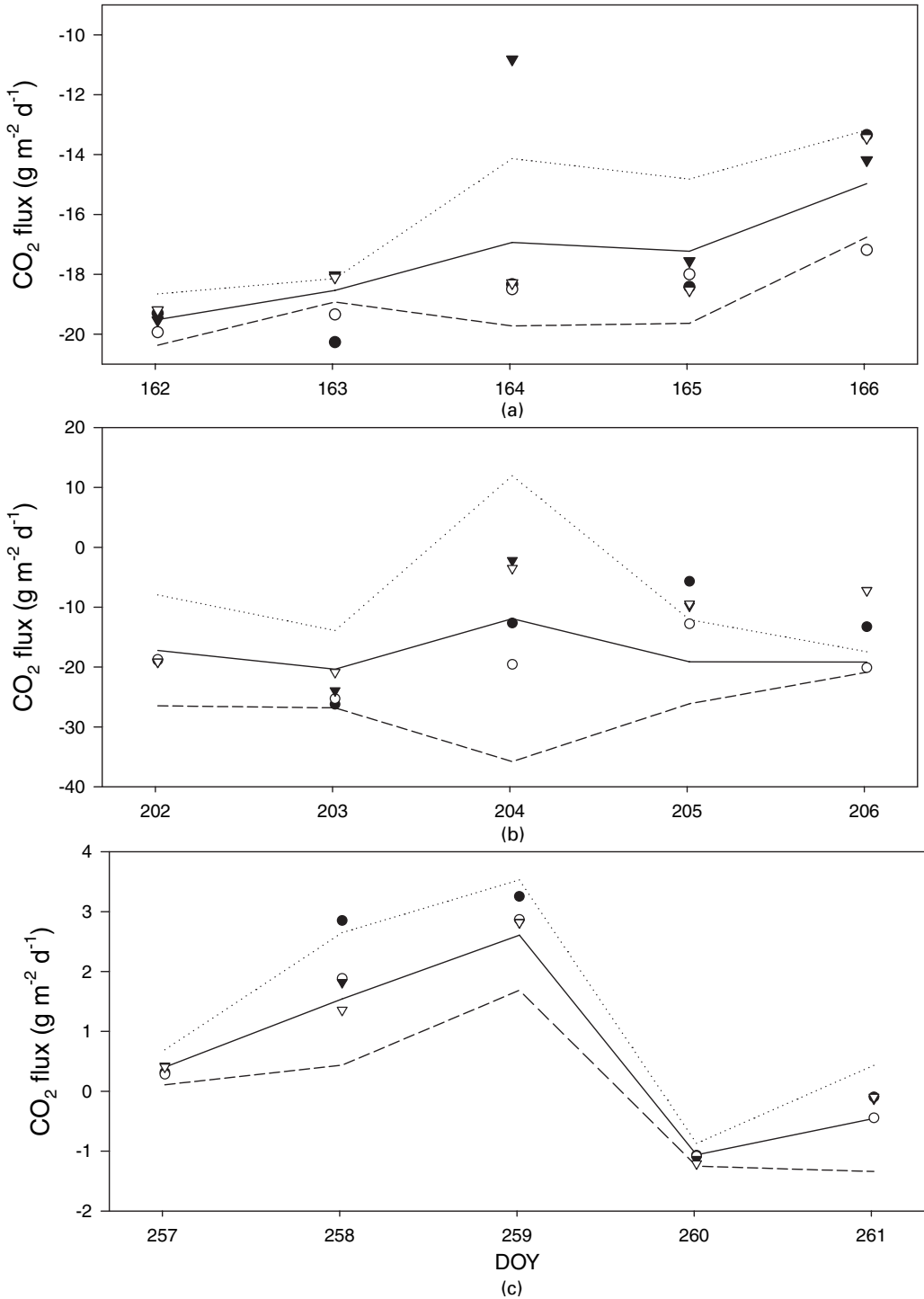


Fig. 5. Average daily CO₂ flux from the expert panel (—) and from protocols PR-A (●), PR-B (○), PR-C (▼) and PR-D (▽) for periods (a) PE-1, (b) PE-2 and (c) PE-3 (·····, upper limit of agreement (mean + 1.96 SDs); - - -, lower limit of agreement (mean - 1.96 SDs))

protocol PR-C, which suggests that the broader range that was used for β in protocol PR-D was more appropriate.

In period PE-1, protocol PR-C failed to represent well DOY 164 (Fig. 5(a)) and this was due apparently to one faulty β that was not captured by the range used. This made the AR(2) model predict positive flux values for a short period of time (Fig. 2(c)). Period PE-2 showed more variability between protocols on a daily basis (Fig. 5(b)), which coincides with the greater variation between experts (Fig. 1). Fig. 3 shows that the difference between protocols PR-A, PR-B and PR-D and the average series was given by how the protocols handled two groups of positive spikes on DOY 205 and one group on DOY 206. For period PE-2, protocol PR-B resulted in daily fluxes that agreed with the experts' opinion (Table 1 and Fig. 5(b)). Protocol PR-B resembled the experts' opinion better because it basically cut out all the spikes. Only at the end of DOY 204 did protocol PR-B wrongly predict opposite (negative) flux values, compared with the average opinion series (Figs 3(b) and 3(d)). Protocol PR-D could not imitate the 'average' pattern because only two faulty β -values were associated with these spikes (Fig. 3(c)). Furthermore, most of these spikes were well described by past values, which explains why they were not pointed out as outliers by the sequence of AR(2) models. Period PE-3 was the easiest for reproducing the general trend of the QC process that was done by the expert panel (Fig. 4). Only one day (DOY 258) of protocol PR-A fell outside the agreement interval (Table 1 and Fig. 5(c)).

Taking the 15 sampled days as a whole, the overall difference in CO₂ flux ranged between 3.2 and 50.6 g m⁻² (Table 3). The former value corresponds to the result of protocol PR-B and represents only 1.7% difference. This was the only value that fell within the reference interval.

4. Conclusions

The variability in average daily flux quality controlled by experts is related primarily to the magnitude of the uncorrected flux (determined in part by wind speed) and to the number of medium-size spikes in the series. The explanation for the latter to be important is because the decision to eliminate this type of spike depends more on the opinion of each expert.

The four statistical protocols that were described here were successful in eliminating the large outliers in every period. This was reflected in the very similar average fluxes on a daily basis compared with the original flux series for each particular period. On a 20-min basis, the agreement with the experts' method was around 90%.

Using TSA in combination with the multivariate methods (multiple distance regression and MLR) in protocol PR-B represented in general an improvement from using only the latter (protocol PR-A). Protocols PR-C and PR-D used only the relationship of present observations with past values but a larger number of variables were quality controlled. The broader range of faulty β -values that were used in protocol PR-D represented an improvement compared with protocol PR-C, especially in period PE-1.

Protocol PR-D did not resemble the average of the experts' opinion in period PE-2 because it did not recognize some spikes as outlying observations. There are at least two possible explanations for this. One is that these spikes are erroneous measures and that the AR(2) model fails to detect them when they happen consecutively. In favour of this explanation is the fact that there are no precipitation events on DOY 205 and DOY 206 that could explain these large FCO₂-values, which is probably what made the experts think that these were outlying observations. The alternative explanation is that the spikes that were not detected as outliers are plausible values and that protocol PR-D may be yielding an accurate estimation. Because there was no independent estimation of the flux during this period, we cannot rule out either of these two explanations.

The results that are presented here show that statistically based methods can be used to control the quality of series of FCO₂ data. Protocols PR-B and PR-D appear to be the most promising because they resembled successfully the average opinion of the expert panel. Additionally, protocol PR-D also quality controlled other relevant variables that are used in the calculation of FCO₂ (i.e. ∂T , ∂CO_2 , ∂e , G , H , LE , K_c and FH_2O). Protocol PR-B was the protocol that matched more closely the average opinion of the expert panel, considering the 15 days together. It differed only by 3.2 g m^{-2} , which represents only 1.7% difference.

If any of these protocols or a similar protocol were used as a standard method for a QC process, it would eliminate the variability that is generated by the response of different experts. Moreover, the statistical protocols can be easily automated, which would make the process considerably faster. Such a procedure may be the base for generating a standard protocol for QC of data from similar micrometeorological techniques, such as the eddy covariance method. Further research is needed to test these protocols under different climate conditions and types of vegetation so that its application can lead to the benefits that have just been described.

Acknowledgements and disclaimer

We thank the five experts who are associated with the US Department of Agriculture ‘Rangeland carbon dioxide flux project’, who did the manual QC of our data sets. This work resulted from joint efforts between the University of California at Davis and the Global Livestock Collaborative Research Support Program, funded in part by US Agency for International Development grant PCE-G-00-98-00036-00, and by scholarships from the University of California at Davis to J. Perez-Quezada. We also thank two reviewers for their comments that helped to improve our paper.

The opinions that are expressed herein are those of the authors and do not necessarily reflect the views of the US Agency for International Development.

Appendix A: Basic equations used for the analyses of Bowen ratio energy balance data

The Bowen ratio β is defined as the ratio between sensible heat H and latent heat LE (Bowen, 1929):

$$\beta = H/\text{LE}, \quad (1)$$

$$H = \rho_a C_p K_H \frac{\partial T}{\partial z}, \quad (2)$$

$$\text{LE} = \frac{\lambda \rho \varepsilon K_v}{P} \frac{\partial e}{\partial z}, \quad (3)$$

where ρ_a (kg m^{-3}) is the air density, C_p ($\text{J kg}^{-1} \text{K}^{-1}$) is the specific heat of air at constant pressure, K_H ($\text{m}^2 \text{s}^{-1}$) is the eddy diffusivity for sensible heat, K_v ($\text{m}^2 \text{s}^{-1}$) is the eddy diffusivity for water vapour, T ($^\circ\text{C}$) is the air temperature, z (m) is the height of the instrument arms, λ (J kg^{-1}) is the latent heat of vaporization, ε is the ratio of the molecular weight of water to that of dry air (0.622), e (kPa) is the vapour pressure, P (kPa) is atmospheric pressure and ∂ represents the difference between the two BREB arms.

Since the vertical gradients of temperature (∂T) and vapour pressure (∂e) can be estimated from measurements at two heights above the canopy (in this study they were at 1.0 and 2.0 m), β can be calculated by substituting equations (2) and (3) into equation (1) and assuming that $K_H = K_v$ (similarity theory):

$$\beta = \gamma \frac{\partial T}{\partial e}, \quad (4)$$

where $\gamma = C_p P / \varepsilon \lambda$ ($\text{kPa } ^\circ\text{C}^{-1}$) is the psychrometric constant.

The one-dimensional (vertical) surface energy balance is defined as

$$R_n - G = H + LE, \quad (5)$$

where R_n is the net radiation, G is the soil heat flux and all the components of the balance are expressed in watts per metre squared. Both R_n and G can be directly measured, allowing the estimation of LE as

$$LE = \frac{R_n - G}{1 + \beta}. \quad (6)$$

Water vapour flux F_{H_2O} is equal to LE/λ . The value of LE that is obtained from equation (6) is used to calculate K_v with equation (3). As mentioned above, K_v is assumed to be equal to K_H and also to be equal to the eddy diffusivity for CO_2 (K_c). Therefore, the CO_2 flux FCO_2 is estimated as

$$FCO_2 = \rho_a K_c \frac{\partial c}{\partial z}, \quad (7)$$

where ∂c is the CO_2 concentration gradient estimated from measurements at the same two heights from which the gradients of temperature (∂T) and vapour pressure (∂e) are estimated above the canopy.

References

- Angell, R. F., Svejcar, T., Bates, J., Saliendra, N. Z. and Johnson, D. A. (2001) Bowen ratio and closed chamber carbon dioxide flux measurements over sagebrush steppe vegetation. *Agric. For. Meteorol.*, **108**, 153–161.
- Aubinet, M., Grelle, A., Ibrom, A., Rannik, U., Moncrieff, J., Foken, T., Kowalski, A. S., Martin, P. H., Berbigier, P., Bernhofer, C., Clement, R., Elbers, J., Granier, A., Grunwald, T., Morgenstern, K., Pilegaard, K., Rebmann, C., Snijders, W., Valentini, R. and Vesala, T. (2000) Estimates of the annual net carbon and water exchange of forests: the EUROFLUX methodology. *Adv. Ecol. Res.*, **30**, 113–175.
- Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw, U. K. T., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K. and Wofsy, S. (2000) FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bull. Am. Meteorol. Soc.*, **82**, 2415–2434.
- Bland, J. M. and Altman, D. G. (1999) Measuring agreement in method comparison studies. *Statist. Meth. Med. Res.*, **8**, 135–160.
- Bowen, I. S. (1929) The ratio of heat losses by conduction and by evaporation from any water surface. *Phys. Rev.*, **27**, 779–787.
- Campbell Scientific (1998) *023/CO₂ Bowen Ratio System with CO₂ Flux: Instruction Manual*, revision 4/98. Logan: Campbell Scientific.
- Chatfield, C. (2004) *The Analysis of Time Series: an Introduction*, 5th edn. Boca Raton: Chapman and Hall–CRC.
- Dugas, W. A., Heuer, M. L. and Mayeux, H. S. (1999) Carbon dioxide fluxes over bermudagrass, native prairie, and sorghum. *Agric. For. Meteorol.*, **93**, 121–139.
- Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, G., Clement, R., Dolman, H., Granier, A., Gross, P., Grunwald, T., Hollinger, D., Jensen, N. O., Katul, G., Keronen, P., Kowalski, A., Lai, C. T., Law, B. E., Meyers, T., Moncrieff, J., Moors, E., Munger, J. W., Pilegaard, K., Rannik, U., Rebmann, C., Suyker, A., Tenhunen, J., Tu, K., Verma, S., Vesala, T., Wilson, K. and Wofsy, S. (2001) Gap filling strategies for long term energy flux data sets. *Agric. For. Meteorol.*, **107**, 71–77.
- Gilmanov, T. G., Johnson, D. A., Saliendra, N. Z., Akshalov, K. and Wylie, B. K. (2004) Gross primary productivity of the true steppe in Central Asia in relation to NDVI: scaling up CO_2 fluxes. *J. Environ. Mangmt*, **33**, S492–S508.
- Heiser, M. D. and Sellers, P. J. (1995) Production of a filtered and standardized surface flux data set for FIFE 1987. *J. Geophys. Res. Atmos.*, **100**, 25631–25643.
- Intergovernmental Panel on Climate Change (2000) Quality assurance and quality control. In *Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories* (eds J. Penman, D. Kruger, I. Galbally, T. Hiraishi, B. Nyenzi, S. Emmanul, L. Buendia, R. Hoppaus, T. Martinsen, J. Meijer, K. Miwa and K. Tanabe), pp. 8.1–8.17. Kanagawa: Institute for Global Environmental Strategies.
- Ito, A. and Oikawa, T. (2002) A simulation model of the carbon cycle in land ecosystems (Sim-CYCLE): a description based on dry-matter production theory and plot-scale validation. *Ecol. Modllng*, **151**, 143–176.
- Laca, E. A., Yurchenko, V., Parsaev, E. and Pittroff, W. (2003) Secondary succession in former wheat fields of Kazakstan steppe. In *Proc. 7th Int. Rangeland Congr., Durban* (eds N. Allsopp, A. R. Palmer, S. J. Milton, K. P. Kirkman, G. I. H. Kerley, C. R. Hurt and C. J. Brown), pp. 366–369. Durban: Grassland Society of Southern Africa.
- Mellesse, A. M. and Hanley, R. S. (2005) Artificial neural network application for multi-ecosystem carbon flux simulation. *Ecol. Modllng*, **189**, 305–314.

- Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996) *Applied Linear Regression Models*, 4th edn. New York: McGraw-Hill.
- Ohmura, A. (1982) Objective criteria for rejecting data for Bowen-ratio flux calculations. *J. Appl. Meteorol.*, **21**, 595–598.
- Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P. and Kassem, K. R. (2001) Terrestrial ecoregions of the world: a new map of life on Earth. *Bioscience*, **51**, 933–938.
- Payero, J. O., Neale, C. M. U., Wright, J. L. and Allen, R. G. (2003) Guidelines for validating Bowen ratio data. *Trans. Am. Soc. Agric. Engrs*, **46**, 1051–1060.
- R Development Core Team (2004) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Sall, J., Lehman, A. and Creighton, L. (2001) *JMP Start Statistics*. Pacific Grove: Duxbury.
- Schimel, D. S., House, J. I., Hibbard, K. A., Bousquet, P., Ciais, P., Peylin, P., Braswell, B. H., Apps, M. J., Baker, D., Bondeau, A., Canadell, J., Churkina, G., Cramer, W., Denning, A. S., Field, C. B., Friedlingstein, P., Goodale, C., Heimann, M., Houghton, R. A., Melillo, J. M., Moore, B., Murdiyarso, D., Noble, I., Pacala, S. W., Prentice, I. C., Raupach, M. R., Rayner, P. J., Scholes, R. J., Steffen, W. L. and Wirth, C. (2001) Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature*, **414**, 169–172.
- Shumway, R. H. and Stoffer, D. S. (2000) *Time Series Analysis and Its Applications*. New York: Springer.
- Svejcar, T., Mayeaux, H. and Angell, R. F. (1997) The rangeland carbon dioxide flux project. *Rangelands*, **19**, 16–18.
- Tabachnick, B. G. and Fidell, L. S. (1996) *Using Multivariate Statistics*, 3rd edn. New York: HarperCollins.
- Webb, E. K., Pearman, G. I. and Leuning, R. (1980) Correction of flux measurements for density effects due to heat and water-vapor transfer. *Q. J. R. Meteorol. Soc.*, **106**, 85–100.