## APPENDIX B The Technique of Multivariate Least Squares Regression

### B1 Technical Background

As discussed in Appendix A previously, the commonly used method of least squares, linear regression assumes that one data set is deterministic, ie the reference data $\{x\}$, and the other, ie the predictor data $\{y\}$, is dependent on it. This is equivalent to saying that all the 'error' resides in one data set and that the other can be measured without error. In the context of correlating wind speeds from two spatially separated sites it is clear that the 'error' cannot be considered to reside exclusively in either data set, and therefore it cannot be presumed that the estimators derived using the least squares, linear regression method are unbiased.

Where both data sets are subject to error it is necessary to adopt a multivariate approach to obtain unbiased estimators. Physically, this corresponds to the assumption that both data sets, $\{x\}$ and $\{y\}$, are dependent on a third variable, in this case, atmospheric conditions at higher altitudes, $\{w\}$. Hence:

$$x_i = m_x w_i + c_x + \varepsilon_{x_i}$$
$$y_i = m_y w_i + c_y + \varepsilon_{y_i}$$

where the uncertainties are taken to be normally distributed:

$$\varepsilon_{x_i} \sim N(0,\sigma_x) \quad ; \quad \varepsilon_{y_i} \sim N(0,\sigma_y) \qquad \text{B.2}$$

Eliminating $w_i$ gives:

$$y_i = mx_i + c + (\varepsilon_{y_i} - m\varepsilon_{x_i}) \qquad \text{B.3}$$

where:

$$m = \frac{m_y}{m_x} \quad ; \quad c = c_y - mc_x \qquad \text{B.4}$$

This is not equivalent to the uncertainty model in Appendix A - see equation A.2 - and the assumptions of that model are violated. Instead it is suggested that there we may be able to find some angle, $\alpha$, at which the residuals become normally distributed - see section B2 for a description of the procedure used to determine the this angle.

The residuals $d_i$, when defined in the direction set by $\alpha$, take the form:

$$d_i = \frac{y_i - \hat{m}x_i - \hat{c}}{\hat{m}\sin\alpha + \cos\alpha} \qquad \text{B.5}$$

So that the sum of squares of the residuals becomes:

$$S = \frac{\sum(y_i - \hat{m}x_i - \hat{c})^2}{(\hat{m}\sin\alpha + \cos\alpha)^2} \qquad \text{B.6}$$

The partial derivatives of S set the conditions on our estimators of m and c:

$$\frac{\partial S}{\partial \hat{m}} \propto \sin\alpha \sum (y_i - \hat{m}x_i - \hat{c})^2 + (\hat{m}\sin\alpha + \cos\alpha)\sum x_i (y_i - \hat{m}x_i - \hat{c}) = 0 \qquad \text{B.7}$$

$$\frac{\partial S}{\partial \hat{c}} \propto \sum (y_i - \hat{m}x_i - \hat{c}) = 0$$

So that $\hat{m}$ and $\hat{c}$ can be determined from:

$$\hat{m} = \frac{\sin\alpha \sum (y_i - \bar{y})^2 + \cos\alpha \sum y_i (x_i - \bar{x})}{\sin\alpha \sum y_i (x_i - \bar{x}) + \cos\alpha \sum (x_i - \bar{x})^2} \; ; \; \hat{c} = \bar{y} - m\bar{x} \qquad \text{B.8}$$

This reduces to the usual least squares estimate of the gradient if we set $\alpha = 0°$ - see equation A.6. The regression coefficient is given by:

$$r = \frac{\sum (x_i y_i - \overline{xy}) - \sum x_i y_i}{\sqrt{\left(\sum (x_i - \bar{x})^2 - \left(\sum (x_i - \bar{x})\right)^2\right).\left(\sum (y_i - \bar{y})^2 - \left(\sum (y_i - \bar{y})\right)^2\right)}} \qquad \text{B.9}$$

Given $\hat{m}$ and $\hat{c}$, the variance of the estimator, $\hat{y}$, is not so easy to calculate. However, it is assumed that, if the co-ordinates are transformed to frame$_1$, so that the $y_1$ axis lies in the direction in which the residuals are determined, we can proceed in this frame as before. This corresponds to a rotation of the co-ordinates through an angle of $\alpha$. Our estimate of the variance of $\hat{y}$ can then be achieved by rotating back to the original co-ordinate system.

So, in the new frame:

$$var(\hat{y}_1) = var(\hat{c}_1) + x_1^2 var(\hat{m}_1) + 2 x_1 covar(\hat{m}_1, \hat{c}_1) \qquad \text{B.10}$$

The variance of the error in the $y_1$ direction is given by:

$$\hat{\sigma}_1^2 = \frac{S}{(N-2)} \qquad \text{B.11}$$

And hence the variance of the estimator of $\hat{y}_1$, for a given value of $x_1$ is:

$$\text{var}\left(\hat{y}_1\right) = \frac{\hat{\sigma}_1^2}{\Sigma\left(x_{1,i} - \bar{x}_1\right)^2}\left(x_1 - \bar{x}_1\right)^2 \qquad \text{B.12}$$

which can easily be expressed in terms of the original data sets through the co-ordinate transforms given by:

$$x_1 = x\cos\alpha + y\sin\alpha \qquad \text{B.13}$$

$$y_1 = y\cos\alpha - x\sin\alpha$$

Using these transformations we obtain estimators for x and finally y in the original frame:

$$\hat{x} = \frac{x_1 - \hat{c}\sin\alpha}{\hat{m}\sin\alpha + \cos\alpha} \qquad \text{B.14}$$

So that:

$$\hat{y} = \frac{\hat{m}x_1 + \hat{c}\cos\alpha}{\hat{m}\sin\alpha + \cos\alpha} \qquad \text{B.15}$$

Given these, and var($\hat{y}$), confidence intervals can be defined at the desired levels, as follows:

$$\hat{x} = \frac{x_1 - \hat{c}\sin\alpha}{\hat{m}\sin\alpha + \cos\alpha} \pm t_{\beta/2,N-2}\sin\alpha\sqrt{var(\hat{y}_1)} \qquad \text{B.16}$$

$$\hat{y} = \frac{\hat{m}x_1 + \hat{c}\cos\alpha}{\hat{m}\sin\alpha + \cos\alpha} \pm t_{\beta/2,N-2}\cos\alpha\sqrt{var(\hat{y}_1)} \qquad \text{B.17}$$

where $t_{\beta/2,N-2}$ is the value of the Student's t-distribution for a confidence level of $\beta/2$ and N-2 degrees of freedom. To estimate y, for a given value of x, equation B.12 is substituted into equation B.16 then rearranged to give a quadratic in $x_1$, ie:

$$x_1^2\left(1 - \frac{t^2\hat{\sigma}^2\sin^2\alpha}{\sum(x_{1,i} - \bar{x}_1)^2}\right) + 2x_1\left(\frac{\bar{x}_1 t^2\hat{\sigma}^2\sin^2\alpha}{\sum(x_{1,i} - \bar{x}_1)^2} - c\sin\alpha - \hat{x}(m\sin\alpha + \cos\alpha)\right) + \left([\hat{x}(m\sin\alpha + \cos\alpha) + c\sin\alpha]^2 - \frac{\bar{x}_1^2 t^2\hat{\sigma}^2\sin^2\alpha}{\sum(x_{1,i} - \bar{x}_1)^2}\right) = 0 \qquad \text{B.18}$$

where we write '$t$' as a shorthand for $t_{\beta/2,N-2}$. This is then solved for the value of $x_1$ corresponding to x. Note that the two roots arise as any information about the sign of t is lost in this process. The lower root corresponds to the upper confidence level, ie positive t, and the upper one to the lower confidence level, ie negative t. The value of $x_1$ is then substituted into equation B.17 to give the final expression for the estimator of y.

## B2 Determination of Alpha

The key to the success of this technique is the correct determination of $\alpha$. As it proved impossible to determine an analytical expression for $\alpha$, The form for equation B.19 was determined empirically from analysis of a large volume of wind speed data. The relation used to identify an appropriate value in the current context, ie relating wind speeds from two, spatially separated sites, is:

$$\alpha = \tan^{-1}\left(m\frac{\sigma_x^2}{\sigma_y^2}\right) \qquad \text{B.19}$$

where $\sigma_x$ and $\sigma_y$ represent the variances associated with the two data sets, and m the 'best fit' gradient, an approximate value for which is obtained from least squares, linear regression. The values of $\sigma_x$ and $\sigma_y$ are obtained from the error signals after 3 hour moving average values have been subtracted from the raw $\{x\}$ and $\{y\}$ time series data.

A more rigorous justification for the form of equation B.19 can be obtained through inspection of the Chi Squared merit function, which is a measure of how well a statistical model describes a given data set. For the usual linear model with uncertainty in one data set, eg {x}, this can be shown to be:

$$\chi^2(m,c) = \sum\left(\frac{y_i - mx_i - c}{\sigma_i}\right)^2 \qquad \textbf{B.20}$$

Minimising this, with respect to m and c, gives 'best fit' model parameters, and it is easy to verify that, in the case of least squares, linear regression, where $\sigma_i$ is constant and independent of $x_i$, that these are identical to those of equation A.6.

With uncertainty in both data sets, the variance in the denominator of the above expression must be replaced with:

$$\begin{aligned} var(y_i - mx_i - c) &= var(y_i) + m^2\, var(x_i) \\ &= \sigma_y^2 + m^2\sigma_x^2 \end{aligned} \qquad \textbf{B.21}$$

So that equation B.20 becomes:

$$\chi^2(m,c) = \sum\frac{(y_i - mx_i - c)^2}{\sigma_y^2 + m^2\sigma_x^2} \qquad \textbf{B.22}$$

Substituting equation B.19 into this expression gives:

$$\chi^2(m,c) = \frac{\sum(y_i - mx_i - c)^2}{m\sin\alpha + \cos\alpha} \cdot \frac{\cos\alpha}{\sigma_y^2} \qquad \textbf{B.23}$$

Minimising this with respect to m and c gives exact forms for their estimators, and it is straightforward to verify that the results are identical to those of equations B.8. This is strong evidence in support of the form chosen for $\alpha$ in equation B.19.