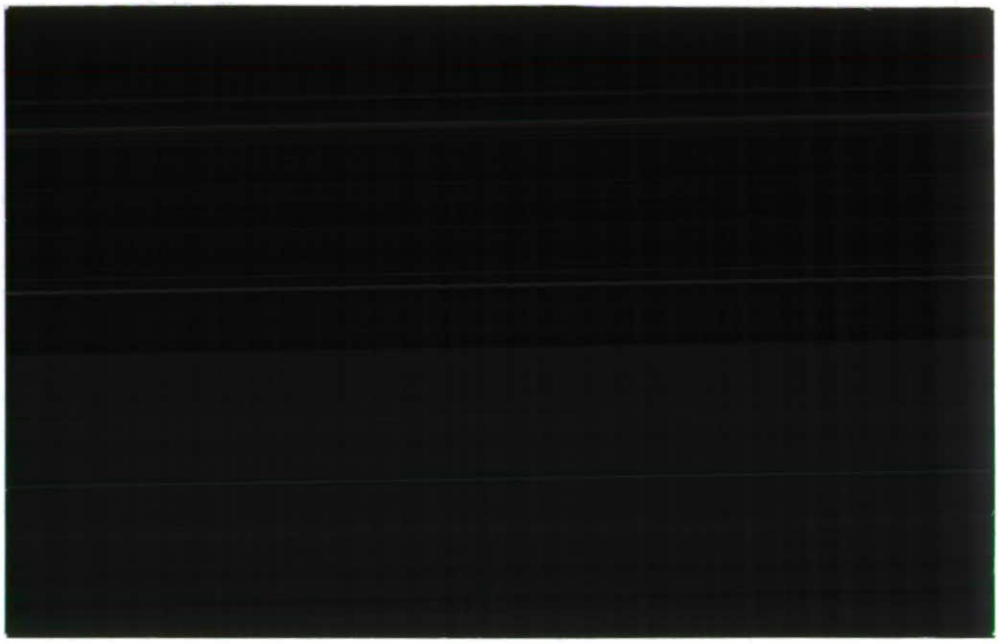
Institute of Hydrology



CALIBRATION OF THE SEASONAL GROWTH MODEL

FINAL REPORT

by

C. Huntingford and R. J. Harding

"The coupling of the vegetation growth model through the scheme under development in the MITRE Project, to the land surface scheme in the Hadley Centre, General Circulation model (GCM) and the calibration of any parameterisation changes which are necessary between The Meteorological Office and The Natural Environment Research Council, Institute of Hydrology on Agreement No. CB/Met 2a/0437."

Institute of Hydrology
Crowmarsh Gifford
Wallingford
Oxfordshire
OX10 8BB
UK

Tel: 01491 838800
Fax: 01491 692424
Telex: 849365 Hydrol G

IH Project No. T06050KI

June 1995

Contents

Chapter 1

Summary

This report gives an account of the joint research which has taken place between the Institute of Hydrology and the Hadley Centre in developing a seasonal growth model for use in the Hadley Centre GCM.

The development of a description of vegetation within a GCM which interacts with its environment is an important improvement to climate models and one which is essential if the models are to make accurate assessments of the possible effects of future man-made changes on the climate. Interactive vegetation will be described using a carbon budget, with the major term being the carbon dioxide exchanges at the surface. Current land surface schemes used in GCMs describe the heat and water exchanges at the surface. However, the fluxes of carbon dioxide are inextricably linked to those of water (both are regulated through stomatal opening) and it is possible to extend the evaporation calculations to include carbon and indeed a number of combined water and carbon dioxide schemes already exist. As yet, very few of these schemes have been tested against field data or are simple enough to be included within GCMs.

Development within this project has been to use meteorological, climate and carbon flux data from field experiments to calibrate and verify parameterisations which have been developed using laboratory studies. A variety of statistical and optimisation tools have been used to verify the appropriateness of applying such laboratory studies to field data and where successful, inferences made about any unknown empirical functional forms and parameters.

There are three main strands to this work:

(1) Understanding the instantaneous carbon dioxide fluxes.

Relationships between the photosynthetic rate and stomatal opening have been checked, and an algorithm developed that can easily relate the evaporative and carbon dioxide fluxes through the common variable, bulk stomatal conductance.

This work has used the data collected from a Kansas prairie (FIFE experiment). Key results (presented in detail in Chapter 2) are:

- a simple algebraic relationship between bulk stomatal behaviour, humidity and net photosynthetic rate performs well. This relationship has been developed from laboratory experiments but work undertaken within the framework of this project provides one of the first tests against actual field data.
- calibration of the photosynthesis model can only produce a sensible and realistic response when an additional dependence on soil water status is included.

(2) Investigation of the dependence of the bulk stomatal conductance on environmental variables.

This has been studied elsewhere through simple curve fitting procedures to field or laboratory data. Here, a more formal statistical analysis has been performed extending to a novel use of Neural Networks as a means of providing an accurate description of stomatal behaviour. The key new result is that the commonly used Stewart-Jarvis formulation is an adequate description of the dependence of the conductance on the environmental variables which can be either measured or predicted by the GCM. There is, however, evidence of "missing" physics which most probably lies in inadequacies in the representation of seasonal and interannual variations of leaf area index. These results are summarised in more detail in the results and conclusions section of Chapter 3.

(3) Modelling of seasonal growth.

The calibrated models described above should be taken and a time dependence introduced such that seasonal growth may be explicitly modelled. It is hoped this third aspect of the work will continue through the productive collaboration between the Institute of Hydrology and the Hadley Centre. Additional data sets of combined water and carbon dioxide fluxes will become available in the next few years from tropical rainforest, boreal forest and tundra experiments. These data sets should be used to extend the testing and calibration of the flux and growth models for these important biomes.

The contents of this report will also appear as two papers broadly based on Chapters 2 and 3. These are in draft form and currently being reviewed internally. The intention is to submit both to scientific journals by August 1995.

Chapter 2

A discussion of canopy level relationships between water and carbon dioxide fluxes and bulk stomatal conductance with verification against data from a Kansas prairie

2.1 Nomenclature and Units (this Section)

A	Available energy (W m^{-2})
c_a	CO_2 concentration at weather station level (μ mol m^{-3})
c_c	CO_2 concentration at leaf level (μ mol m^{-3})
c_i	Internal leaf CO_2 level (μ mol m^{-3})
F_s	Soil respiration (μ mol m^{-2} s^{-1})
g_S	Bulk stomatal conductance to water vapour (m s^{-1})
g_{a_v}	Aerodynamic conductance to water vapour (m s^{-1})
H	Sensible heat flux (W m^{-2})
h_s	Relative humidity
k_T	Temperature dependent slope of photosynthetic CO_2 response (mol m^{-2} s^{-1})
LAI	Leaf area index

N	Measured net downward CO_2 flux above canopy (μ mol m^{-2} s^{-1})
P	Net photosynthetic rate (μ mol m^{-2} s^{-1})
P_G	Gross photosynthetic rate (μ mol m^{-2} s^{-1})
Q_p	Quantum flux density (μ mol m^{-2} s^{-1})
R	Gas constant (J K^{-1} mol^{-1})
R_D	Plant respiration (μ mol m^{-2} s^{-1})
R_{PAR}	Photosynthetically active radiation (W m^{-2})
R_{solar}	Total incoming solar radiation (W m^{-2})
T	Temperature at leaf level (°C)
T_a	Temperature at weather station level (°C)
u_a	Mean horizontal windspeed at weather station level (m s^{-1})
u_*	Friction velocity (m s^{-1})
V_T	Temperature dependent rate of rubisco capacity (μ mol m^{-2} s^{-1})
α	Quantum efficiency (mol m^{-1})
δq	Humidity deficit (g kg^{-1})
θ	Volumetric soil moisture content
λE	Latent heat flux (W m^{-2})
τ	Momentum flux (kg m^{-1} s^{-2})
Ψ_l	Leaf water potential (MPa)
Ψ_s	Soil water potential (MPa)

2.2 Introduction

This paper is concerned with developing a simple model that relates the evaporation flux, λE, the net photosynthetic rate, P, and the bulk stomatal conductance, g_S, and any dependencies these quantities may have on the local environment. Many of the intervariable relationships have been proposed as a consequence of studying data at leaf level (both in the field and within laboratories) and at canopy level (in the field). The aim here is to use the dataset from the FIFE experiment (Sellers *et al.*, 1992) which include measurements of both λE and net CO_2 flux, N (from which P is inferred), thereby enabling an intercomparison of proposed models of mass and carbon transfer at canopy level. Upon selection of the "best" models, a scheme is proposed that would be suitable for implementation within Global Circulation Models, and it is hoped this lower boundary condition will enhance current simulations of the carbon and water cycles.

The processes studied are explained through four equations (see below). There are four unknown variables, $\lambda E, g_S, P$ and the internal carbon dioxide concentration, c_i - with the number of unknowns equalling the number of equations, the system is therefore closed. Three equations are well established, these being the Penman-Monteith equation, a photosynthesis model

$P = P(c_i, R_{PAR}, T)$ and

$$c_a - c_i = P\left[\frac{1.6}{g_S} + \frac{1.4}{g_{a_h}}\right] \tag{2.1}$$

where the factors of 1.4 and 1.6 account for the different diffusivities of carbon dioxide and water in air and the stomata. Equation (2.1) assumes between the measurement level and vegetation canopy, the carbon dioxide flux is given by P, and thus not affected by the soil flux F_s. This is the equivalent to using the Shuttleworth and Wallace (1985) model with the meeting point of modelled conductances at measurement height. Calculating c_i but using $c_c - c_i = 1.6P/g_S, c_a - c_c = 1.4N/g_{a_h}$ produced only slightly different values of c_i: this can only be interpreted as a consequence of $g_S \ll g_{a_h}$.

A fourth equation is needed, which may be regarded as a model for g_S. One possibility is to prescribe a form for g_S depending directly on $R_{solar}, T, \delta q$ and θ, following the work of Jarvis (1976). Such calculations for the FIFE datasets are given in Stewart and Verma (1992). Seen adversely, these forms are little more than curve-fitting exercises and thus inherently empirical. For a scheme developed for use in climate models, it is advantageous to rely on equations with a physical basis, for calibrations of empirical forms with unknown parameters will always be data intensive and overly site specific. Alternative closures have been developed by Ball *et al.* (1987), and extended by Leuning (1995) whereby the stomatal conductance may be expressed explicitly as a function of the other unknown variables (most notably P) along with a dependency on the local environment but with relatively few tunable constants. It is these latter closures that are tested here.

2.3 Data

2.3.1 The FIFE datasets

The 1987 FIFE data was collected from a study area in a Kansas prairie consisting of 82 % ungrassed C_4 species of grass (Verma *et al.*, 1992) between May and October, site reference 16 (4439-ECV). Measurements of the fluxes $H, \lambda E, N$ and τ were made using eddy correlation techniques, averaged over 30 minutes. Half hourly measurements of radiation variables, R_{net}, R_{PAR} and meteorological variables T_a, q_a, u_a were collected; also weekly measurements of soil moisture content θ (expressed as a percentage and over the depth of 0.1-1.4m) and leaf area index, LAI. The values of the data used in this study are selected in an identical fashion to Stewart and Verma (1992), that is daylight hours only and after 1000 local time to prevent evaporation from dew distorting results. To exclude evaporation of intercepted rain, days when rainfall occurred and the day after were not included in the dataset for model testing. After this reduction of the dataset, 174 points remain.

2.3.2 Data Processing

The data allows direct inversion of the Penman-Monteith equation to yield inferred measurements of bulk stomatal conductance, g_S. Less simple is obtaining the net photosynthetic rate, P. To eliminate negative contributions to the net photosynthetic rate, N, the upward fluxes of CO_2 from the soil must be subtracted. That is, the relationship between the fluxes is given by

$$P = N + F_s.$$

Following Norman *et al.* (1992), the soil respiration at the FIFE site can be described by

$$F_s = p_1 \left(\frac{\theta - p_2}{40.0 - p_2} \right) e^{p_3(T-25)} \tag{2.2}$$

where the value of T is assumed to be an approximation to that measured at the 10 cm depth. Equation (2.2) is optimised against nighttime measurements of N (when it is assumed $P \equiv 0$), yielding estimates $p_1 = 17.8$ μ mol CO_2 m^{-2} s^{-1}, $p_3 = 0.20$ and $p_3 = 0.062$ ° C $^{-1}$. However, this best fit only explained 50.7 % of the variance in F_s. Respiration from the vegetation other than through the stomata (that is stem and root respiration) is neglected here under the assumption that its contribution to the CO_2 fluxes is negligible. This is verified in Norman *et al.*, 1992.

A relationship between the soil moisture content, θ, and the soil water potential, Ψ_s, is required. Observed values in Stewart and Verma (1992) give $\Psi_s = -1.5$ MPa at $\theta = 0.205$ and $\Psi_s = -0.033$ MPa at $\theta = 0.387$. Fitting a function of the form $\Psi_s = a\theta^{-b}$ for constants a and b gives $\Psi_s = -1.11 \times 10^{-4}\theta^{-6}$.

Finally, in the absence of measurements, the value for the ambient CO_2 concentration at weather station height, c_a, is assumed fixed at 0.5 g kg^{-1}. This is converted into units of μ mol CO_2 m^{-3} which fluctuates due to atmospheric density changes.

2.4 Theory

2.4.1 Proposed models for bulk stomatal conductance, g_s

The "Jarvis" Model

The model proposed by Jarvis (1976) is that the stomata respond to the individual climatic variables through normalised functions, f_i, and such that their combined effect on overall stomatal opening is multiplicative. That is

$$g_S = g_{S_{max}} \prod_i f_i(X_i). \tag{2.3}$$

Individual forms for f_i represent a family of functions, characterised by one or two unknown parameters which may be estimated through least squares fitting

of data. Such parameters are regarded as plant specific, but may also include effects of scaling up to canopy level. Stewart and Verma (1992) fit a model of the form (2.3) to the same FIFE dataset.

The "Ball-Woodrow-Berry" Model

Frequent observations of a relationship between photosynthetic rate, P, and stomatal conductance, g_S, led to Ball *et al.* (1987) proposing the model

$$g_S = g_{S_{min}} + \frac{aPh_s}{c_c} \tag{2.4}$$

where a and $g_{s_{min}}$ are unknown fitting parameters. Dependencies of g_S on the water and carbon dioxide status of the local atmosphere is accounted for through the variables h_s and c_a.

The "Leuning" Model

Leuning (1995) suggests a revised equation based on the Ball-Woodrow-Berry model: here the dependence on relative humidity is replaced by the humidity deficit, δq. Leuning (1995) propose the formulation

$$g_S = g_{S_{min}} + \frac{aP}{(c_a - c_*)\left(1 + \frac{\delta q}{\delta q_c}\right)} \tag{2.5}$$

which now has four unknown constants, $g_{S_{min}}, a, c_*$ and δq_c. However, for C_4 plants, the constant c_* should be set to zero. This formulation has the advantage that the atmospheric humidity is expressed in units commonly used in meteorological models.

2.4.2 The photosynthesis model

The photosynthesis model for C_4 plants follows that given by Collatz *et al.* (1992). Appendix B. The gross photosynthesis, P_g, is either limited by the internal CO_2 concentration, or by some combination of leaf level temperature and levels of photosynthetically active radiation. The first equation finds the temperature and light limited value for P_g, that is $P_{g_{max}}$ which satisfies

$$\beta_1 P_{g_{max}}^2 - P_{g_{max}}(V_T(T) + \alpha Q_p) + V_T(T)\alpha Q_p = 0$$

where

$$\begin{aligned} Q_p &= 4.57 R_{PAR} \\ \beta_1 &= 0.83 \\ V_T &= \frac{V_{max} Q_{10}^{0.1(T-25)}}{(1 + e^{0.3(13-T)})(1 + e^{0.3(T-36)})} \\ Q_{10} &= 2.0 \end{aligned}$$

The β_1 value allows for co-limitation between the temperature and light dependencies and the smallest root to the quadratic for $P_{g_{max}}$ is selected. Similarly, a function for co-limitation is produced between a light-temperature dependence, and a dependence on the carbon dioxide concentration, c_i. Now

$$\beta_2 P_g^2 - P_g \left(P_{g_{max}} + k_T(T)\frac{c_i^*}{p} \right) + \frac{P_{g_{max}} k_T(T) c_i^*}{p} = 0 \qquad (2.6)$$

where c_i^* is the internal leaf CO_2 level in Pascals (Pa) to keep with the notation of Collatz, 1992, p is the atmospheric pressure fixed at 1.013×10^5 Pa and $\beta_2 = 0.90$. Hence,

$$c_i^* = R(T + 273.15)c_i \times 10^{-6}$$

where $R = 8.3144$ J K^{-1} mol^{-1}. Again, the smallest root is selected to the above quadratic returning a value for P_g. Finally the net photosynthetic rate satisfies

$$P = P_g - R_d(T).$$

The rate "constant" k_T and leaf respiration term R_d are assumed to be linearly related to V_T, that is they have identical temperature dependencies (Collatz *et al.*, 1992; p531). Hence, writing

$$\begin{aligned} V_T &= V_{max} f(T) \\ R_T &= R_D f(T) \\ k_T &= k f(T) \end{aligned}$$

the value of R_D is set as $R_D = 0.025 V_{max}$, which is an adjustment given by recent work of —, whilst the value for k is unknown.

Plots of leaf measurements at the FIFE site for the three dominant C_4 grass types at varying values of c_i, R_{PAR} and T are given in Figure A, Polley *et al.* (1992), in part verifying the functional forms above.

2.5 Performance of models against FIFE data

2.5.1 Methodology

The four equations analysed are the Penman-Monteith equation and equations (2.1), (2.5 and (2.6), that is the preferred bulk stomatal conductance model will be taken as the Leuning model due to it's simplicity (although the other descriptions are tested as well). There are unknown fitting parameters in (2.5) and (2.6) which can only be found by comparison of model predictions against data. The datasets contain two variables against which intercomparisons may be made, namely both fluxes λE and inferred P. Having two measurements is a distinct advantage when calibrating models as individual components of the four overall equations may be checked in isolation. Here the natural choice of equations to examine separately are (2.5) and (2.6), that is the bulk stomatal conductance model and the photosynthesis model.

2.5.2 The fitting of models for g_S

Seven fits were made in total for the FIFE dataset, and all are presented in Table 2.1 where SV represents the Stewart and Verma (1992) model (application of equation (2.3)). For models where g_S has an explicit dependence on P, the inferred value from the data is used. Similarly, these values of P are used implicitly to evaluate c_c.

The Ball-Woodrow-Berry achieves a fit of 88.1% variance explained in g_S, with selected parameters given by $g_{S_{min}} = 1.84 \times 10^{-3}$ m s^{-1} and $a = 5.72$. Initial optimised runs of the Leuning model yielded small values for $g_{S_{min}}$ (8.97×10^{-4} m s^{-1}), and so the simplified form

$$g_s = \frac{aP}{(c_a - c_*)\left(1 + \frac{\delta q}{\delta q_c}\right)} \tag{2.7}$$

is used. This gives values $a = 5.41$ and $\delta q_c = 53.6$ g kg^{-1} with a fit of 88.2% variance explained. The simple fit of direct proportionality returns a value of $a = 5.34$ with corresponding performance statistic of 84.0%.

The full Stewart-Verma approach explains 78.6% of the variance, which is a figure less than that quoted in their paper. However, the dataset used for this paper includes the last intensive field campaign. Stewart and Verma (1992) ignored this on the basis that the vegetation was senescent and the low heat flux measurements not so reliable.

These results are especially reassuring given that the simplified Leuning model contains only two tunable parameters. The result of simple proportionality between g_S and P (with just one parameter, p_4) also performs very well, and underpins the basic assumption of a direct relationship between g_S and P. The results in Table 2.1 for a reduced Stewart-Verma model suggest that a large proportion of the variance of g_S may be explained by the dependence on θ alone. As a good fit is achieved by assuming g_S is almost proportional to P, it may be expected that the net photosynthetic rate also has a dependency on the soil moisture content. This hypothesis is considered below in the optimisation of the model for photosynthesis.

2.5.3 The fitting of the model for P

There are three unknown parameters to be found in model (2.6), namely, α, V_{max} and k. Values given in Collatz *et al.*, 1992 for C_4 plants are $\alpha = 0.04$ mol m^{-2} s^{-1}, $V_{max} = 39$ μ mol m^{-2} s^{-1} and $k = 18 \times 10^3 V_{max}$. Optimisation of the model against the inferred values of P to find the unknown fitting parameters (where the internal CO_2 concentration is found through (2.1) but by substituting the calculated value of g_S and the measured values of P) yields $\alpha = 0.038$ mol m^{-2} s^{-1}, $V_{max} = 441$ μ mol m^{-2} s^{-1} and $k = 360 V_{max}$. However, these values are physically unrealistic, generating a $P - c_i$ curve where the photosynthetic rate is

permanently carbon dioxide limited (see Figure A). The percentage of variance explained in P is 75.9.

This phenomenon may be explained as follows. Recall there is a near linear relationship between P and g_S, where the g_S has a strong dependency on θ. In regions where the photosynthesis rate is limited through c_i (that is $P = P(c_i)$) where $c_i = c_a - P/g(\theta)$, then P depends on θ but through the stomatal conductance. In regions where P is limited by light and temperature there can be no dependence of P on θ and thus with realistic parameters for the photosynthesis model, a bad fit will ensure. To adjust therefore, the least squares optimisation will select a value of the fitting parameters such that P is always dependent on c_i in order to capture the θ dependence in P. This observation, whereby a calibration exercise suggests a good fit but with a model using obviously incorrect parameterisations, should be regarded as a cautionary tale when using optimisation procedures.

To obtain a realistic parameterisation of the model for the net photosynthetic rate, it is necessary to introduce a dependence of P on the soil status. This is done through the soil water potential, Ψ_s: the proposed new model for photosynthesis, P, satisfies

$$P = P'(c_i, R_{PAR}, T)\left\{1 - \frac{\Psi_s}{\Psi_{sc}}\right\} \tag{2.8}$$

where $P'(c_i, R_{PAR}, T)$ is the original model proposed in equation (2.6) and Ψ_{lc} (critical leaf water potential) is set as -1.5 MPa. Fitting this model yields a fit of 89.4% variance explained and a realistic $P - c_i$ curve, as in Figure A. The dependence $P = P(\Psi_s)$ has not been extensively reported in the literature, although there is some evidence in Figure A, Polley *et al.* (1992). It has also been discussed in a theoretical context in Friend (1991). A further optimisation using Ψ_l with a root resistance prescribed as an additional tunable parameter did not significantly improve the model. However, whether the apparent dependence on the soil status has physical basis within the leaves themselves (and hence the correct variable is Ψ_l), or through a hormonal signal from the soil (and so requiring Ψ_s) is not clear.

2.5.4 Creation of a simple algorithm

Combining (2.1) with (2.7) yields

$$1 - \frac{c_i}{c_a} = \frac{1.6}{a}\left(1 + \frac{\delta q}{\delta q_c}\right) \tag{2.9}$$

and so specification of δq enables immediate calculation of c_i. This, combined with knowledge (or prediction) of R_{PAR}, T, Ψ_{soil} returns a value of P. Returning to equation (2.7), g_S can be predicted and then through the Penman-Monteith equation, a value of λE. In summary, variables are found as follows

$$R_{PAR}, T, \Psi_{soil}, \delta q \Longrightarrow c_i \Longrightarrow P \Longrightarrow g_S \Longrightarrow \lambda E.$$

This is a change from traditional models where the emphasis is placed on the calculating the hydrological aspect first.

2.6 Conclusions

The inherent relationships between stomatal conductance, carbon and vapour fluxes and their dependence on the local environment have been studied. The aim has been to develop a simple algorithm such that given local meteorological and soil conditions, then P and λE may be evaluated through the intermediatory variable g_S. Such a scheme has been devised through the development of relationships between these variables proposed in various earlier papers; in this paper selection and verification of the most appropriate governing equations has been decided on their ability to replicate fluxes at canopy level for C_4 grasses.

The most notable adjustment to previous work has been the necessity to introduce a *direct* dependence of the net photosynthetic rate, P on Ψ_l in order to reproduce the carbon fluxes. Such an effect may be seen in the data, although there appears to be little discussion in the theoretical literature of photosynthesis of such a process.

The finalised, calibrated forms for the models perform well, reproducing above eighty percent of variance explained in both fluxes. Code has subsequently been written to incorporate these descriptions of the water and carbon fluxes, and it now awaits trials within the Hadley Centre GCM. However, before such an exercise, it is essential to check the above conclusions for a range of vegetation types. This should be possible using the carbon dioxide fluxes measured in recent international experiments.

2.7 References

Ball, J.T., Woodrow, I.E. and Berry, J.A.: 1987, 'A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions', in: 'Progress in photosynthesis research (ed. I. Biggins)', *Martinun Nijhoff Publishers, Netherlands*, 221-224.

Collatz, G.J., Ribas-Carbo, M. and Berry, J.A.: 1992, 'Coupled photosynthesis-stomatal conductance model for leaves of C_4 plants', *Aust. J. Plant Physiol.* **19**, 519-538.

Friend, A.D.: 1991, 'Use of a model of photosynthesis and leaf microenvironment to predict optimal stomatal conductance and leaf nitrogen partitioning', *Plant Cell Environ.* **14**, 895-905.

Leuning, R.: 1995, 'A critical appraisal of a combined stomatal-photosynthesis

model for C_3 plants', *Plant, Cell and Environment* **18**, 357-364.

Monteith, J.L.: 1965, 'Evaporation and the environment', *Symp. Soc. Expl. Biol.* **19**, 205-234.

Norman, J.M., Garcia, R. and Verma, S.B.: 1992, 'Soil surface CO_2 fluxes and the carbon budget of a grassland', *J. Geophysic. Res.* **97**, D17, 18,845-18,853.

Polley, H.W., Norman, J.M., Arkebauer, T.J., Walter-Shea, E.A., Greegor, D.H., Jr. and Bramer, B.: 1992, 'Leaf gas exchange of *Andropogon gerardii* Vitman, *Panicum virgatum* L., and *Sorghastrum nutans* (L.) Nash in a tallgrass prairie', *J. Geophysic. Res.* **97**, D17, 18,837-18,844.

Sellers, P., Hall, F.G., Asrar, G., Strebel, D.E. and Murphy, R.E.: 1992, 'An overview of the First International Satellite Land Surface Climatology Project (ISLSCP) Field Experiment (FIFE)', *J. Geophysic. Res.* **97**, D17, 18,345-18,373.

Stewart, J.B. and Verma, S.B.: 1992, 'Comparison of surface fluxes and conductances at two contrasting sites within the FIFE area', *J. Geophysic. Res.* **97**, D17, 18,623-18,638.

Verma, S.B., Kim, J. and Clement, R.J.: 1992, 'Momentum, water vapour, and carbon dioxide exchange at a centrally located prairie site during FIFE', *J. Geophysic. Res.* **97**, D17, 18,629-18,639.

Table 2.1: Intercomparison of models for g_S: FIFE data

Model	Percentage of variance explained
SV-θ dependence only	65.4
SV-D dependence only	52.9
SV-R_{solar} depndence only	3.5
SV - full model	78.6
Ball-Woodrow-Berry	86.7
Leuning (Simpified)	87.0
$g_S = p_4 P$	84.5

2.8 Figure A

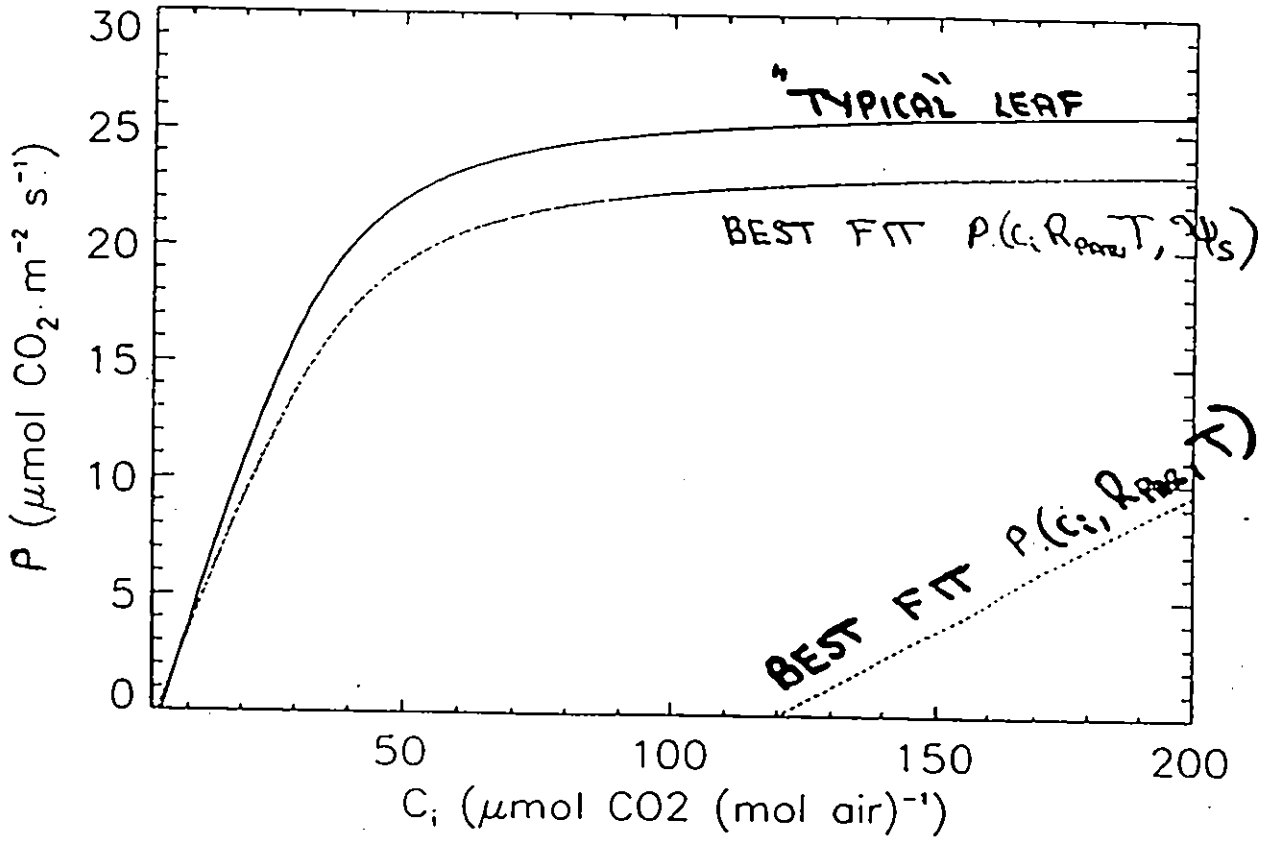


Figure A: P vs c_i curves with and without a dependence on Ψ_s.

Chapter 3

Use of statistical and neural network techniques to detect how stomatal conductance responds to changes in the local environment

3.1 Introduction

The radiative energy received at the Earth's surface is partitioned into sensible and latent heat fluxes as a function of the properties of the surface. The magnitude of such fluxes can have an important effect on the climate and hydrological cycle at a variety of temporal and spatial scales (see for instance Amazonian numerical simulations in Lean and Rowntree, 1993): an accurate description of these fluxes is required as a lower boundary conditions to Global Circulation Models (GCMs). The partitioning of energy into sensible and latent heat is controlled by the opening and closing of stomata where the more open the stomata, the easier it is for vapour to diffuse from inside the leaves and into the local atmosphere. Collins and Avissar (1992) demonstrate of all land-surface characteristics described in atmospheric models, results are most sensitive to the parameterisation of stomatal behaviour.

It is therefore essential that the opening and closing of stomata is properly parameterised, and this is achieved in its simplest form for a homogeneous

canopy through the bulk stomatal conductance, g_S (m s^{-1}). Simultaneous measurements of both meteorological variables and energy fluxes enable calculation of g_S through inversion of the Penman-Monteith Big Leaf Model (Monteith, 1965). Inferences can then be made regarding the dependence of g_S on local climatological variables (solar radiation, leaf temperature and humidity deficit at leaf level, the soil water potential in the root zone, the leaf area index and possibly the time of day). This is clearly a regression problem for g_S in these six local variables ($X_j, j = 1,6$). However, there are a variety of difficulties which are always associated with such problems. How are the dominant dependencies of g_S determined? Having identified an important environmental variable within the stomatal conductance description, what is the true functional dependency? How may collinearity amongst the driving variables weaken conclusions about g_S? Are some of the apparent fits to data covering for other missing dependencies not included in the list X_j? How can the addition of a new environmental variable be tested for significance?

This paper attempts to address these difficulties. The theory behind regression analysis is very well established, particularly for linear models. This is used to make initial studies, but it is soon apparent that the response of g_S to the local climate is highly nonlinear. Whilst this makes it more difficult to understand the effect of each regressor variable, it is conjectured the nonlinearities help counteract any effects of collinearity (that is problems associated with variables moving together). The functions $f_j(X_j)$ as proposed by Jarvis (1976) and implemented by Stewart (1988) for pine forest clearly do well at predicting g_S for a given climatology in dry conditions. To see if these can be improved further, the technique of neural networks is used. In essence this is a large unconstrained optimization, whereby any nonlinear functional dependencies a physical process may have are approximated to using a series of transfer functions. In adopting any final forms for g_S to predict energy fluxes, the simple Big Leaf Model must be used. It is shown (for instance in Huntingford, 1995) that the final calibrated forms for g_S are dependent on the aerodynamic description.

Consideration should be given to the relationship between direct canopy observations of g_S and laboratory measurements of g_{ST}, the latter leading to the initial forms for f_j. This is partly understood through a simple relationship dependent on LAI. Avissar (1992) observed large variability at leaf level of both stomatal conductance and local climate, and concluded the integrated effect, giving a canopy g_S, can only be calculated from leaf measurements of stomatal conductance g_{ST} using complex multi-layer canopy models. However, there is little doubt even simple models such as the Penman-Monteith Big Leaf model, using the forms f_j based on laboratory experiments, can do well at predicting g_S. It can only be concluded the spatial variation in leaf-scale processes are implicitly included in any carefully optimised and thus calibrated canopy value of g_S.

Accepting the concerns raised in the last two paragraphs, this paper attempts to carry out a rigorous analysis of the bulk stomatal response, g_S, to the local

climate for a Scots pine forest. The structure of this paper is as follows. A site description is provided in Section 3.2, followed in Section 3.3 by a review of the analysis by Stewart (1988). Section 3.4 discusses the application of linear analysis to the problem. In particular, Section 3.4.2 describes relevant theory for the linear regression analysis of g_S and Section 3.4.3 applies such theory to a straight dependence on the X_j. Section 3.4.4 demonstrates the improved model emulations when using the functions $f_j(X_j)$ and the importance of nonlinearities in the dependencies of g_S. It is therefore natural to see if the response curves f_j found by Jarvis, 1976, may be improved upon. The nonlinear analogue to linear regression is the use of neural network techniques, providing a completely unconstrained optimisation including the functional forms themselves for g_S (Section 3.5). It becomes apparent there is an upper limit on the possible fit of g_S to X_j, but the remaining errors cannot be exclusively attributed to experimental error (Section 3.6).

It is acknowledged that this paper is primarily concerned with modelling tools. Little new physical insight into the behaviour of g_S is presented. However, what is achieved is the description of rigorous analysis enabling both the validity of previous modelling attempts to be assessed, and formally demonstrate areas where "missing physics" is present. It is hoped that the techniques are presented in a sufficiently general fashion as to allow easy transfer to any similar problems in Soil Vegetation Atmosphere Transfer scheme (SVAT) calibration.

3.2 Data Considerations

3.2.1 Scots Pine Forest data

The data used for this study was collected from a Pine Forest site near Thetford, Norfolk. Meteorological and surface flux data were collected, the latter using Bowen ratio measurements. Data were restricted to dry canopy conditions in daylight hours, and during the years 1974, 1975 and 1976. The soil moisture deficit (mm) over the top 1.0 m was found by using a neutron probe at a series of depths once every few days and intergrating to a value $\delta\theta$, whilst interpolating to obtain values for days when measurements were not made. The estimate of the LAI profile is assumed identical for all three years. A full description of both the site and experimental equipment is given in Stewart (1988); in all, there were 584 hourly measurements. Scatter plots of the dataset against each X_j are given in Figure 1.

3.2.2 The Penman-Monteith equation

Values of the bulk stomatal conductance g_S are found through inversion of the Penman-Monteith big leaf model (Monteith, 1965):

$$g_S = \frac{\gamma g_{a_h}}{\left[\frac{A\Delta}{\lambda E} + \frac{\rho c_p D g_{a_h}}{\lambda E} - \Delta - \gamma\right]} \tag{3.1}$$

and it is assumed

$$g_S = f_S(LAI)\, g_{ST}. \tag{3.2}$$

Variables used in (3.1) and (3.2) are latent heat flux, λE ($\mathrm{Wm^{-2}}$); density of air, ρ ($\mathrm{kg\ m^{-3}}$); psychometric "constant", γ ($\mathrm{kPa\ ^\circ K^{-1}}$); specific heat of air c_p ($\mathrm{J\ kg^{-1}\ K^{-1}}$); available energy, A ($\mathrm{Wm^{-2}}$); vapour pressure deficit at weather station height, D (kPa); aerodynamic conductance of sensible and latent heat, g_{a_h} ($\mathrm{m\ s^{-1}}$) and gradient of saturated vapour pressure with temperature, $\Delta = \mathrm{d}e_s/\mathrm{d}T$ ($\mathrm{kPa\ ^\circ K^{-1}}$). Often f_S is taken as the identity function $f_S(LAI) \equiv LAI$ although this expression represents a crude scaling from leaf to canopy level. Later on an alternative form for f_S generated by the neural network analysis is proposed.

By calculating the bulk stomatal conductance for each timestep through substitution of driving variables into (3.1), a data set is created listing both g_S and the variables g_S is believed to depend on, namely R_{solar}, δq, T, $\delta\theta$, LAI, t_{day}, where R_{solar} is the total incoming solar radiation ($\mathrm{W\ m^{-2}}$), δq is the humidity deficit ($\mathrm{g\ kg^{-1}}$) and t_{day} is the number of hours past midnight (hours). The motivation for a time of day function is that it may capture a short timescale soil dependence due to drying of soil throughout the day in the immediate vicinity of the roots, this being an effect not distinguishable in the measurements of $\delta\theta$. The humidity deficit is chosen instead of the vapour pressure deficit to keep in with notation used by Stewart (1988) and a good approximation is given by $D \approx \delta q/0.622$. As in Stewart (1988), the approximation $g_{a_h} = 0.167\ \mathrm{m\ s^{-1}}$ is made: the aerodynamic conductances are an order of magnitude greater than the stomatal conductances, and as shown in Huntingford (1995) the evaporation rate as predicted by the Penman-Monteith equation are insensitive to g_{a_h} for this site. A further approximation is made that the environmental variables, δq and T, measured at the weather station height may be taken as variables values affecting the stomata at leaf height. Again, this is likely to be valid for the Thetford Forest, as the aerodynamic resistance is relatively small.

3.2.3 Training and test data

The data is split into training and test data. The training data accounts for 75 % of available data, and is used to calibrate and model. Such models are then run to assess their predictive ability against the remaining 25 % of data, called the test data. The division of data into training and test is determined using a random number generator.

3.3 The Stewart-Jarvis approach

3.3.1 The nonlinear functions f_j

These will be referred to as the Stewart-Jarvis functions. Performing laboratory experiments, Jarvis (1976) looked for the response of g_{ST} to each variable by individually allowing it to vary with other conditions kept constant. Normalised functions were selected to closely follow the assimilated experimental data. The particular forms used by Stewart (1988) to model the pine forest are repeated here for clarity:

$$\begin{aligned} f_1(R_{solar} : a_1) &\equiv \frac{R_{solar}}{a_1 + R_{solar}} \left(1 + \frac{a_1}{R_{\text{solar}_{max}}}\right), \\ f_2(\delta q : a_2, a_3) &\equiv \begin{cases} 1 - a_2 \delta q & 0 < \delta q < a_3 \\ 1 - a_3 a_2 & \delta q \geq a_3 \end{cases} \\ f_3(T : a_4) &\equiv \left(\frac{T - T_L}{a_4 - T_L}\right) \left(\frac{T_U - T}{T_U - a_4}\right)^{\frac{T_U - a_4}{a_4 - T_L}}, \\ f_4(\delta\theta : a_5) &\equiv 1 - 0.340 e^{a_5(\delta\theta - 69.7)}, \end{aligned}$$

where $R_{\text{solar}_{max}} = 1000$, $T_L = 0$, $T_U = 40$ and $f_5(LAI)$ is the identity function. The parameters $a_1, \ldots, a_5$ are initially unknown and estimated through model calibration.

3.3.2 Modified Stewart-Jarvis approach to include time (Model One)

The calculations performed in Stewart (1988) form the basis for the first model of g_S in this study (Model One), but with two important differences. First a dependence on time of day is introduced where:

$$f_6(t_{day}) \equiv (1 - a_6(t_{day} - 8)).$$

Second the model is calibrated against 438 training data points (to find $a_1, \ldots, a_6$) and the subsequent functional forms are used predictively to estimate g_S for the remaining 146 test data points. Optimising through least squares the function

$$g_S = g_{ST_{max}} \prod_{i=1}^{6} f_i(X_i) \tag{3.3}$$

yields the following statistics: percentage of variance explained (see next section for statistic description) in the training data; 75.7 %, whilst for the test data (using the model optimised on the training data); 72.2 %. These results are repeated in Table 3.1, row 1. The selected optimised parameters satisfy $g_{ST_{max}}$ = 7.906 x 10^{-3} , a_1 = 82.2, a_2 = 0.0682, a_3 = 10.1, a_4 = 16.3, a_5 = 0.0842 and a_6 = 0.0246.

3.4 Linear Regression

3.4.1 Introduction

A linear model for the stomatal conductance g_S is of the form

$$g_S = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \ldots \beta_J x_J + \epsilon \tag{3.4}$$

where $x_1, \ldots, x_J$ are the regressor variables and the β_j are the regression coefficients. In this study, the X_j are either the local environmental variables X_j or the Stewart-Jarvis functions $f_j(X_j)$. The ϵ term describes the model error.

Values of β_j are estimated through least squares optimisation. For a well-posed problem, the β_j allow physical insight into the dependency of g_S on x_j, and further enable predictions to be made of g_S when values of only x_j are known. A common problem is the presence of collinearity, or "moving together" amongst the regressor variables. This has three undesirable consequences. First the ability to accurately determine the β_j is weakened, and so the physical interpretation of selected values must be treated with caution. Second, should the model be used predictively for a new dataset of x_j where the collinearity between variables is different or nonexistent, then incorrect model emulations of g_S are likely to occur. This second difficulty should be seen as occurring when using a model predictively, but in regions of variables space where it has not been calibrated or verified due to collinearity in the original data for which g_S is known. The third problem is if two variables are collinear, where one is a real dependency of g_S, whilst the second is questionable. It is therefore impossible to determine through optimisation which is the correct variable to retain in a model.

This section briefly reviews the theory of detection of collinearity and its implications in determining the degrees of freedom of the data and their subsequent use in a model for g_S. Further a standard statistical test for assessing the hypothesis that g_S may not depend on a particular variable is described.

3.4.2 Mathematical description of optimization of model for g_S

Least squares optimization

Much of this subsection is covered in standard texts such as Myers, 1989, but the relevant parts appropriate for the determination of a linear model for g_S are reproduced here for continuity. There are $J = 6$ environmental variables X_j, all measured $I = 438$ times (training data only), and for each time a value of g_S exists. Assumptions are that the linear model is the correct model, and that the regressor variables are measured with zero error. Hence, the model errors come about as a result of measurements errors in g_S. As at each time interval i, g_S is found by inversion of the Penman-Monteith equation, the implication here is

that errors are allowed in the flux measurements, but not the other variables X_j. Given by the random variable, ϵ, these errors are assumed to be independently normally distributed N(0,σ^2). These are very strong and limiting assumptions; the approach taken is to assume them to hold until it is clear violations exist, at which point a more generalised model is required.

The regressor data $x_{i,j}$ is given by matrix $\mathbf{X}$, values of g_{S_i} by the vector $\mathbf{g}_S$ and β_j and ϵ_i by the vectors β and ϵ. In this paper, $x_{i,j}$ will either equal $X_{i,j}$ directly ($X_{i,j}$ is the i^{th} observation of environmental variable X_j), or some function of $X_{i,j}$, that is $x_{i,j} = f_j(X_{i,j})$.

The true model is

$$\mathbf{g}_S = \mathbf{X}\beta + \epsilon$$

where

$$\mathbf{g}_S = \begin{pmatrix} g_{S_1} \\ \vdots \\ g_{S_{438}} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,6} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{438,1} & \cdots & x_{438,6} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_6 \end{pmatrix}.$$

The value $x_{i,j}$ is the i^{th} observation of variable j and g_{S_i} the i^{th} observation of g_{S_i}. The estimate of β is called b and is a vector random variable due to the variability in ϵ. Least squares optimization is such that b minimises the sum of squares of the residuals, $SSR = (\mathbf{g}_S - \mathbf{X}\mathbf{b})^T(\mathbf{g}_S - \mathbf{X}\mathbf{b})$ and in such circumstances b has the properties

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{g}_S \qquad \mathrm{E}(\mathbf{b}) = \beta \qquad \mathrm{E}(\mathbf{b}-\beta)(\mathbf{b}-\beta)^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \quad (3.5)$$

where the diagonal terms of the last matrix give the variance of the b_j, whilst σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{SSR}{(I-(J+1))} = \frac{SSR}{431}. \quad (3.6)$$

The percentage of variance explained statistic, PVE, provides a simple assessment of model fit to g_S, given by

$$PVE = 100\left(1 - \frac{SSR}{\sum_{i=1}^{438}(g_{S_i} - \bar{g}_S)^2}\right) \qquad \text{where} \qquad \bar{g}_S = \frac{\sum_{i=1}^{438} g_{S_i}}{438}$$

Used predictively, for given $x_1, \ldots, x_6$, (that is, $J = 6$) then

$$\hat{g}_S = b_0 + \sum_{j=1}^{6} b_j x_j$$

is an unbiased estimator for g_S.

The importance of individual variables may be tested. Studying one variable in the presence of others, the following hypothesis may be tested at the required

level of significance:

$$\begin{aligned} H_0 : b_j &= 0 \\ H_1 : b_j &\neq 0. \end{aligned}$$

The test is a student's t-test (having $I-(J+1)=431$ degrees of freedom) with test statistic given by

$$t_{431} = \frac{b_j}{\hat{\sigma}\sqrt{e_{jj}}} \tag{3.7}$$

Here, e_{jj} are the diagonal terms of the 6 x 6 matrix, $(\mathbf{X}^T\mathbf{X})^{-1}$.

Centering and scaling; collinearity

A first step in the understanding of collinearity is to present the data in a "normalised" form. Variables are centred and scaled creating a data matrix $\mathbf{X}^*$ with components $x^*_{i,j}$ where

$$x^*_{i,j} = \frac{x_{i,j}-\mu_j}{S_j} \quad \text{where} \quad \mu_j = \frac{\sum_{i=1}^{438} x_{i,j}}{438}, \qquad S_j = \sqrt{\sum_{i=1}^{438}(x_{i,j}-\mu_j)^2}.$$

The 6 x 6 matrix given by $\mathbf{X}^{*T}\mathbf{X}^*$ is called the correlation matrix. Diagonal terms are unity, whilst collinearity between individual regressor variables are indicated by off-diagonal terms near unity. The correlation matrix is symmetric with real eigenvalues $\lambda_1,\ldots,\lambda_6$ corresponding to normalised eigenvectors $\mathbf{v}_1,\ldots,\mathbf{v}_6$. The number of eigenvalues near zero represents the number of overall collinearities and subsequent reduction (from six) in degrees of freedom within the data. The statistic $\lambda_{max}/\lambda_{min}$ is sometimes used to indicate the total effect of collinearity on a system of data, with very high values indicating problems.

In the centred and scaled co-ordinates, the model is written as $\hat{g}_S = \beta^*_0 + \sum_{j=1}^{6}\beta^*_j x^*_j$ where least squares estimation generates estimates b^*_j of β^*_j. The effect of collinearity in weakening the estimation of regressor coefficients is expressed through variation inflation factors

$$VIF_j = \frac{\mathrm{Var}(b^*_j)}{\sigma^2} \qquad j=1,6,$$

where the values may be found from the last expression in (3.5). In an ideal situation, with little collinearity between variables, the correlation matrix and its inverse would be near the identity matrix. Hence the VIF_j would all be approximately one. From correlation matrix theory,

$$\mathrm{E}(\mathbf{b}^*-\beta^*)^T(\mathbf{b}^*-\beta^*) = \sigma^2\mathrm{tr}(\mathbf{X}^{*T}\mathbf{X}^*)^{-1} = \sigma^2\sum_{j=1}^{6}\frac{1}{\lambda_j}, \qquad \sum_{j=1}^{6}\lambda_j = 6,$$

and so it is apparent that for ill-conditioned data due to collinearity, at least one eigenvalue will be near zero, and hence the VIF_j must be large for some j. The equivalent t-statistic to (3.7) but for centred and scaled variables should return identical values. It is very important to note that collinearity weakens the power of the test given in (3.7). That is, collinearity makes it more difficult to reject the null hypothesis for any given regressor variable.

Principal Component Analysis

The centred and scaled regressor variables $x^*_{i,j}$ are transformed into principal components given by $\mathbf{Z} = \mathbf{X}^*\mathbf{V}$, where the columns of $\mathbf{V}$ are the eigenvalues $\mathbf{v}_j$. The 6 x 6 matrix $(\mathbf{Z}^{*^T}\mathbf{Z})_{ij} = \lambda_j$ if $i = j$, zero otherwise (that is off diagonal terms are zero), demonstrating the rotation of variables from x^*_i to z_i is such that the new variables are orthogonal. Trivially $(\mathbf{Z}^{*^T}\mathbf{Z})^{-1}_{ij} = \lambda_j$ if $i = j$, zero otherwise, and consequently $VIF_j = \mathrm{Var}(\alpha_j)/\sigma^2 = 1/\lambda_j$ where the regression model is expressed as $\hat{g}_S = \beta^*_0 + \sum_{j=1}^{6} \delta_j z_j$; the correlation matrix for these new variables would be the identity matrix. The variance in the system has not been eliminated, but eigenvalues λ_j may now be directly associated with particular variables z_j. The larger eigenvalues will correspond to the dominant linear trends within the data whilst the smaller eigenvalues correspond to the components that are orthogonal to such principal variations.

The t-test given in (3.7) becomes

$$t_{431} = \frac{d_j\sqrt{\lambda_j}}{\sigma}$$

where d_j estimates δ_j. Due to the orthogonality of the principal components, elimination of any variable z_j leaves the remaining d_j unaltered in any reduced least squares re-optimization. This makes (3.4.2) a valid order of elimination statistic, for remaining values are unaltered after component elimination. Terms are dismissed for showing no evidence at a particular statistical level that they are important.

Statistic (3.4.2) provides the important link between the behaviour and trends of the data (through λ_j), and the demands of g_S itself (through d_j). For problems that are ill-posed, the order of elimination may not be in increasing λ_j. A small λ_j may correspond to a large d_j, meaning g_S depends on some of the weaker "signals" in the data. Similarly, a large λ_j may correspond to a small d_j, meaning g_S does not respond to the strongest movements in the driving data.

Mallows' C_p statistic

The t-statistic given above allows elimination of unnecessary principal components and the subsequent reduction in the degrees of freedom of the problem. However a test does exist whereby rotation to principal components is not

required. The Mallows' C_p statistic provides the balance between an overfitted model with unnecessary degrees of freedom, and one which is underfitted due to the elimination of relevant physics. An overfitted model will have large variances introduced into regressor coefficients (particularly with collinearity present) which as remarked above, renders the model inaccurate when used predictively. Similarly, an underfitted model will have an inherent bias. There is therefore an optimum number of dependencies to be included in any regression of g_S, and this number is given by the C_p statistic. The statistic is

$$C_p = p + \frac{(s^2 - \sigma^2)(I - p)}{\sigma^2}, \qquad p = J' + 1,$$

where J' is the number of remaining regressor variables, decreasing from six during elimination. s^2 is the estimated variance of the residuals for a particular optimization, given by the sum of squares of residuals for the training data divided by $(I - p)$. If $C_p > p$, this indicates overfitting, $C_p < p$ represents underfitting; the ideal number of principal components is where $C_p \approx p$. Unfortunately, for g_S, the error variance σ^2 is unknown. Traditionally in such circumstances, the estimate of σ^2 is given by $\sigma^2 = \hat{\sigma}^2 = s^2$ for $p = 7$, that is using all the components (J'=6) as given in (3.6), and so $C_7 = 7$.

The models for g_S are optimised with components gradually eliminated to derive values of the C_p statistic. It must be stressed that the C_p statistic takes no account of any distinction between real and questionable dependencies in the event of collinearity. The reduced model will be accurate predictively provided the structure of the dataset is invariant. However, should the model be used in new regimes of variable space, it would be very misleading. This is not a fault of any statistical techniques presented; the only way a model is of use throughout all possible regions of the driving variables is to have data and measurements of g_S everywhere, against which the model is initially calibrated. The process of model reduction through principal components is more rigorous and especially has the advantage that setting any rejected components equal to zero will give the regions of variable space in which the model may be used with confidence.

3.4.3 Application to stomatal conductance - regression in the environmental variables (Model Two)

Least squares optimization (Model Two)

The linear regressions are performed using the training data of 438 points. The regression variables are given by $R_{solar}, \delta q, T, \delta\theta, LAI$ and t_{day}, where this order is used throughout the paper. To avoid excessive use of exponents, all stomatal conductances are now expressed in units of (mm s^{-1}). Predictive ability is assessed using the regression to estimate the stomatal conductance at the

remaining 146 test data points. The least squares optimization of

$$g_S = \beta_0 + \sum_{j=1}^{6} \beta_j X_j + \epsilon$$

gives

$$\hat{g}_S = 5.41 + 0.00361 R_{solar} - 0.729\delta q + 0.271 T - 0.0960\delta\theta + 2.96 LAI - 0.198 t_{day} \tag{3.8}$$

with percentage of variance explained in the training data of 60.2 %. The equivalent statistic for the test data is 50.6 % (see also Table 3.1). The estimate of the variance of model errors (using (3.6)) is

$$\hat{\sigma}^2 = 3.93.$$

The t-statistics on individual variables return the values

$$5.56, \quad -11.7, \quad 6.04, \quad -10.1, \quad 7.09, \quad -4.70,$$

and so at the 99 % level, the null hypothesis that a particular variable is unimportant is rejected for each variable. This is a strong result, particularly as any effects of collinearity have not yet been addressed. Please note further that this result includes the variable t_{day}.

Centering and scaling; collinearity (Model Two)

Scaling and centring data gives the correlation matrix

$$\mathbf{X}^{*T}\mathbf{X}^{*} = \begin{pmatrix} 1 & 0.389 & 0.343 & 0.246 & 0.055 & -0.283 \\ . & 1 & 0.838 & 0.508 & 0.208 & 0.443 \\ . & . & 1 & 0.739 & 0.477 & 0.324 \\ . & . & . & 1 & 0.675 & 0.065 \\ . & . & . & . & 1 & 0.045 \\ . & . & . & . & . & 1 \end{pmatrix}. \tag{3.9}$$

The top four correlations (all with values greater than 0.5) are the pairs $(\delta q, T)$, $(T, \delta\theta)$, $(\delta\theta, LAI)$ and $(\delta\theta, \delta q)$. The second and fourth correlation are a surprise; the temperature and humidity deficit have a diurnal fluctuation whereas measured soil moisture deficit values vary on a longer timescale. Similarities between (3.9) and the equivalent matrix for Bramley's apple trees may be seen in Jones and Higgs (1989). The means and standard deviations of the training and test data (including g_S) are given in Table 3.2 and Table 3.3.

The eigenvalues, λ_j, of the correlation matrix, $\mathbf{X}^{*T}\mathbf{X}^{*}$, are given by

$$2.974, \quad 1.313, \quad 1.084, \quad 0.336, \quad 0.197, \quad 0.0966.$$

There are no eigenvalues very near zero, and $\lambda_{max}/\lambda_{min} = 30.79$, suggesting in fact collinearity is not overly severe. It is unlikely the data will be reduced to just one or two degrees of freedom.

The optimization in the scaled and centred variables is given by

$$\hat{g}_S = 6.21 + 14.4R^*_{solar} - 52.5\delta q^* + 30.2T^* - 36.7\delta\theta^* + 20.1LAI^* - 12.8t^*_{day} \quad (3.10)$$

whilst the VIF_j for each regressor coefficient β^*_j has values

$$1.71, \quad 5.10, \quad 6.36, \quad 3.34, \quad 2.04, \quad 1.89.$$

The statistics for percentage of variance explained and estimate of error variance, $\hat{\sigma}^2$ remain unaltered under coordinate transformation to scaled and centred variables (or indeed principal components) when all components are retained.

Principal component analysis (Model Two)

The dataset is transformed to principal components, giving the regression

$$\hat{g}_S = 6.21 - 18.7z_1 - 22.5z_2 - 13.4z_3 + 26.5z_4 - 19.6z_5 + 60.6z_6$$

and VIF_j factors given by the inverse of the eigenvalues, that is

$$0.336, \quad 0.762, \quad 0.923, \quad 2.976, \quad 5.076, \quad 10.4$$

and so the redistribution of variance is readily observed.

Values of $t_{431} = (d_j\sqrt{\lambda_j})/\hat{\sigma}$ are given by

$$-16.3 \quad -13.0, \quad -7.05, \quad 7.74, \quad -4.40 \quad 9.50$$

and every value is significant at the 99 % level, that is H_0 is rejected for all principal components. However, although this indicates all components are of importance, the order statistics do not have values whose magnitudes are in descending order. The order of elimination of the principal components is 5, 3, 4, 6, 2, 1 (the ideal is 6, 5, 4, 3, 2, 1), suggesting the problem to be partially ill-posed. The first and second strongest variations, or "frequencies" within regression variables Z_j are those found within g_S. This is not true for the third "signal" onwards. Three possible physical explanations are proposed: either the environmental variables g_S is believed to depend on are chosen incorrectly, a linear regression model itself is insufficient thereby causing incorrect inferences about variable dependence, or simply that the answer is correct. The effect of eliminating components on PVE is given in Table 3.4. The last principal component is given by

$$z_6 = 0.0734R^*_{solar} - 0.619\delta q^* + 0.743T^* - 0.209\delta\theta^* - 0.101LAI^* + 0.0812t^*_{day}$$

and although collinearity is not so strong as to eliminate it, the solutions $z_6 = 0$ does show the plane where most driving data is near. Notably, the largest coefficents show the predicted relationship between δq and T.

Mallows' C_p test (Model Two)

Values of the C_6 statistic, that is eliminating single variables, are

$$36.0 \quad 142, \quad 41.4, \quad 108, \quad 50.6, \quad 26.7.$$

As no values are near six, then this suggests the regression requires seven degrees of freedom, that is all the regression variables. C_p statistics for elimination of principal components are given in Table 3.4. This agrees with the t-tests above that no components may be eliminated.

3.4.4 Regression in the previously optimized functions f_j (Model Three)

Least Squares Optimisation (Model Three)

This experiment in regression is simply as follows. The six previously optimized functions f_j as given in Stewart (1988) and Section 3.3.2 ($a_6 = 0.02455$) are taken as regression variables in a further linear regression,

$$g_S = \beta_0 + \sum_{j=1}^{6} \beta_j f_j(X_j) + \epsilon. \tag{3.11}$$

The rationale for studying this particular form is twofold. First it demonstrates if the success of the Stewart-Jarvis approach is due to either the assumption of multiplicity (equation 3.3), or the nonlinearity of the functions f_j themselves. Second, by partly reducing the nonlinearity of the system to a linear regression in the f_j, the full weight of linear regression theory as outlined in Section 3.4.2 may be employed to understand the optimisation. This includes the ability to add new terms to (3.11) and test their importance at a given level of significance through t-tests. This makes the search for possible new environmental dependencies g_S may have easier to evaluate.

To compare the optimization (3.11) directly with the analysis in Section 3.4.3 is probably unfair. The least squares solution to (3.11) may be regarded as having more than seven parameters to optimize, but where the extra fitting has come from the original work by Stewart (1988). However, if all the unknown parameters were studied simultaneously, there would undoubtedly be excess overfitting due to strong collinearity between the (a_j, β_j) pairs.

Results of the optimization give

$$\hat{g}_S = -31.1 + 6.00f_1 + 8.48f_2 + 6.59f_3 + 9.59f_4 + 3.40f_5 + 7.15f_6$$

which explains 71.9 % of variance in the training data and 70.4% in the test data. The estimate of the error variance is given by

$$\hat{\sigma}^2 = 2.78.$$

All three statistics suggest the new model is better than the original simple regression (Model Two). The decrease in the estimate of $\hat{\sigma}^2$ suggest this model to be less biassed. The reduction in difference of PVE between the train and test data suggests the new model to be robust, and thus better posed. This will be partially verified by noting the reduced effects of collinearity below.

The t_{431}-statistics for the individual f_i are given by

$$6.12, \quad 11.5, \quad 8.36, \quad 13.6, \quad 9.55, \quad 4.38$$

which are all significant at the 99 % level.

Centering and scaling; collinearity (Model Three)

The transformation of variables by scaling and centring does not cause such a large magnitude change in their values as for Model Two; the new regression variables f_i (except $f_5(LAI)$) are already normalised to lie on the region [0,1]. The correlation matrix is

$$\mathbf{X}^{*T}\mathbf{X}^{*} = \begin{pmatrix} 1 & -0.297 & 0.129 & -0.178 & 0.085 & 0.320 \\ . & 1 & -0.309 & 0.533 & -0.333 & 0.511 \\ . & . & 1 & -0.362 & 0.547 & 0.075 \\ . & . & . & 1 & -0.638 & 0.024 \\ . & . & . & . & 1 & 0.045 \\ . & . & . & . & . & 1 \end{pmatrix} \tag{3.12}$$

and so it is immediately apparent the magnitude of correlations between variables has reduced. Further, the new variables have rearranged the order of correlations with the strongest magnitude. The eigenvalues of the correlation matrix above are

$$2.493 \quad 1.385 \quad 1.003 \quad 0.651 \quad 0.316 \quad 0.153$$

and so $\lambda_{max}/\lambda_{min} = 16.33$. The optimisation itself in centred and scaled variables is

$$\hat{g}_S = 6.21 + 13.4f_1^* + 34.5f_2^* + 16.9f_3^* + 35.3f_4^* + 23.2f_5^* + 11.3f_6^*$$

with VIF_j for each coefficient given by

$$1.728 \quad 3.228 \quad 1.474 \quad 2.410 \quad 2.116 \quad 2.413.$$

All these statistics point to a reduction in collinearity (and subsequent better-posedness) over Model Two.

The possible reduction in collinearity through the use of nonlinear functions can be seen through the following, albeit extreme, "thought experiment". Suppose two regressor variables x_1 and x_2 are related to y, and an initial model is

given simply as $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Suppose further the data satisfies $x_{i,1} \approx x_{i,2}$ for most data points. Although the approximation for y may be quite good, the estimates of β_j carry no physical significance having large variances. Now let a new nonlinear model be proposed. Let

$$y = \beta_1 \mathrm{H}(x_1 - 1) + \beta_2 \mathrm{H}(x_2 - 2) + \epsilon$$

where $\mathrm{H}(x)$ is the Heavyside "step" function equal to zero for $x < 0$ and unity for $x \geq 0$. If this new model is correct, then the jumps at $x_1 = 1$ and $x_2 = 2$ should be reflected in the g_S, and so the estimates of β_1 and β_2 do now have physical significance and may be accurately determined. Of course, if the second model is correct, the linear model in x_1 and x_2 must be incorrect, and thereby invalidates many of the assumptions described in Section 3.4.3. It is proposed the Stewart-Jarvis functions allow the effects of collinearity to be weakened in a similar fashion.

Principal component analysis (Model Three)

Transformed to principal components, the regression is given by

$$\hat{g}_S = 6.21 - 15.3z_1 + 25.6z_2 + 23.3z_3 + 32.4z_4 - 32.7z_5 - 3.88z_6$$

where the VIF_j are

$$0.401, \quad 0.722, \quad 0.997, \quad 1.536, \quad 3.16, \quad 6.54.$$

Values of $t_{431}(d_j\sqrt{\lambda_j})/\hat{\sigma}$ are given by

$$-14.46 \quad 18.08, \quad 14.00, \quad 15.71, \quad -11.02 \quad -0.91$$

and this time all are significant at the 99% lever except the last component. This is evidence the optimization in the f_i has five degrees of freedom. The order of elimination of the principal components is 6,5,3,1,4,2 which, when again compared with the ideal of 6,5,4,3,2,1, demonstrates the strongest dependencies in g_S are not those seen in the data. Indeed, this time the optimization selects the second principal component as accounting for the maximum variance in g_S. Rejecting the last component means the optimised model for g_S is only valid when

$$z_6 = -0.339f_1^* - 0.649f_2^* - 0.0668f_3^* + 0.416f_4^* + 0.162f_5^* + 0.510f_6^* \approx 0.$$

Interestingly, the nonlinearity has spread the collinearity between variables more evenly throughout the $f_j(X_j)$; notably between the functions of $R_{solar}, \delta q, \delta\theta$ and t_{day}. The previously strong relationship between T and δq is lost and this can also be seen in the correlation matrix above.

Mallows' C_p statistic (Model Three)

Values of the C_6 statistic are

$$42.0, \quad 138, \quad 74.4, \quad 190, \quad 95.6, \quad 23.6.$$

Clearly no single value of f_j may be eliminated to provide an equally satisfactory model. However, the principal component analysis in Section 3.4.4 shows that the data, $f_j(X_j)$, does have one degree of collinearity which may be eliminated through ignoring a single principal component. This is verified through the C_p statistics associated with the eliminated z_j presented in Table 3.5. The spread of collinearity between variables is therefore such that it cannot be absorbed into a single elimination of a f_j regressor variable.

The immediate conclusion so far of this section is that nonlinearity is essential for a good fit, and the problem becomes better-posed through its introduction. Collineanity in the f_j, as opposed to the simple X_j, is reduced overall, but what does remain can be channelled into a single principal component. That is, there are five degrees of freedom in the $f_j(X_j)$ dataset. However this reduction cannot be introduced through the simple elimination of a single f_j.

3.4.5 Intercomparison of full model with original Stewart-Jarvis approach

The ability of models to produce similar estimates for g_S is assessed. Two models may have nearly identical statistics for the percentage of variance explained, but this is not proof they are producing identical model emulations; the models may be explaining completely different parts of variance in g_S. The following statistic is calculated, and presented in Table 3.8: momentarily, assume model p is correct. Then calculate the percentage of variance explained in output from model p by model q. The key result for this section is a comparison of Model One (adjusted Stewart-Jarvis model) to Model Three (summing functions f_i) where the intercomparisons for training and test data give values of 94.3 % and 96.0 % respectively. By noting statistics for comparing Model Two (linear regression in environmental variables) with Model Three and One, which show very different model predictions, this is confirmation that the nonlinearity in the functions f_i is of far more importance than whether they are multiplied or summed.

The necessity of using nonlinear functions has been demonstrated. To check the functions used are appropriate for this optimization, the theory of neural networks is now appealed to.

3.5 Neural network method (Model 4)

3.5.1 Introduction

In this section an attempt is made to find a general nonlinear mapping between the environmental variables, X_j and the stomatal conductance, g_S, using a "Backpropagation" neural network (????). Neural networks of this type are capable of reproducing any (reasonable) nonlinear function given sufficient accurate "training" data (Funahushi, 1989, Hornik *et al.*, 1989). Here the network approach allows a derivation of the best fit model for g_S without requiring the prescription of separate functional forms for each environmental dependency (as is required for the Stewart-Jarvis model). Instead the neural network produces a form for g_S which cannot be separated into distinct dependencies on each input. As such, the network derived function can be viewed as an "unconstrained" nonlinear fit, which provides a benchmark against which more physically based models of g_S can be judged. In addition the neural network enables the inclusion of dependencies for which physical models are not available (for instance, time of day).

3.5.2 Theory

A schematic representation of a typical network architecture is shown in Figure 2. The relationship between the inputs, X_j (the environmental variables) and the output, $\hat{Y}$ (the modelled bulk stomatal conductance, $\overline{g}_S$ is described using a network of nodes or "neurons". Each neuron produces an output, Z_k, which is a strongly nonlinear function of its weighted input:

$$Z_k = \Psi\left(\sum_j W_{k,j} X_j + W_{k,0}\right)$$

where $W_{k,j}$ is the weight linking the j^{th} input variable to the k^{th} neuron, $W_{k,0}$ is an offset for the k^{th} neuron and Ψ is known as the "transfer" function. Here Ψ is usually taken to be a sigmoidal function:

$$\Psi(u) = \frac{1}{1+e^{-u}} - \frac{1}{2}.$$

The output variable is derived as the weighted sum of the outputs from each neuron:

$$\hat{g}_S = \hat{Y} = \sum_k V_k Z_k + V_0$$

where V_k is the weight linking the k^{th} neuron to the output variable and V_0 is a constant offset.

For the network to be useful, it is necessary to determine the weights $W_{k,j}$ and $V_{j,k}$ which minimise the difference between the known outputs, Y_i, and the modelled outputs $\hat{Y}_i$ where i labels the i^{th} datapoint. Again, this is achieved through selecting values that give the smallest sum of squares of the residuals, that is minimising

$$SSR = \sum_i (Y_i - \hat{Y}_i)^2$$

where i labels the i^{th} datapoint. In neural networks SSR is normally minimised by an iterative procedure in which each of the training datapoints, $X_{i,j}, Yi$, are presented to the network in turn (recall $X_{i,j}$ is the i^{th} measurement of variable X_j), many times (or "epochs"), and on each occasion the weights are updated according to a learning algorithm.

In the backpropogation network the learning algorithm amounts to a steepest descent on a surface in the weight space, where the height of the surface corresponds to the error, SSR. The reader is referred to Rumelhart *et al.* (1986), for a full derivation; here only the equations for calculating the weight changes $\Delta_i W_{kj}$ and $\Delta_i V_k$ which result from the presentation of the i^{th} datapoints are listed:

$$\Delta_i V_k = \eta Z_k (Y_i - \hat{Y}_i) + \mu \Delta_{i-1} V_k$$

$$\Delta_i V_0 = \eta (Y_i - \hat{Y}_i) + \mu \Delta_{i-1} V_0$$

$$\Delta_i W_{k,j} = \eta X_{i,j} \Psi' \left(\sum_j W_{k,j} X_{i,j} + W_{k,0} \right) (Y_i - \hat{Y}_i) V_k + \mu \Delta_{i-1} W_{k,j}$$

$$\Delta_i W_{k,0} = \eta \Psi' \left(\sum_j W_{k,j} X_{i,j} + W_{k0} \right) (Y_i - \hat{Y}_i) V_k + \mu \Delta_{i-1} W_{k,0}$$

where $\Delta_{i-1} W_{k,j}$ and $\Delta_{i-1} V_k$ are the weight changes produced by the previous datapoint, $\Psi' = \Psi(u)/du$, and η and μ are the "learning rate" and the coefficient of the "momentum term" respectively. Each of the latter two may be chosen to optimise the performance of the network.

3.5.3 Results

Implementation

In order to diagnose possible overfitting, the practice of separating the data into 438 training points and 146 test points is retained. The network weights are updated as described above using the training data only. At each epoch the network is also used in predictive mode in order to calculate the percentage of

the variance explained in both the training and test datasets. An improving fit to the training data which is associated with a worsening fit to the test data is symptomatic of overfitting and a degrading predictive capability. In all cases the network input (X_j) and output variables (Y) are centred and scaled using means and variances calculated from the entire dataset (as in Section 3.4.2).

Fitting the Penman-Monteith Equation

Before proceeding to model the bulk stomatal conductance of the Thetford Forest, the ability of the network to fit a known nonlinear function, namely the Penman-Monteith equation, is first assessed. In this case the actual ("measured") bulk stomatal conductance is amongst the inputs, whilst the output variable is the predicted latent heat flux, λE. The remaining inputs are the driving variables in the Penman-Monteith equation (see Section 3.2.2), thus for this Section *only*:

$$X_j = (\ A^*,\ \ \delta q^*,\ \ T^*,\ \ g_S^*,\ \ \Delta^*\ \ ,\rho\ )\ \ ;\ \ Y = \lambda E^*\ \ ;\ \ \hat{Y} = \lambda \hat{E}^*$$

where, as before, the * represent centred and scaled variables, and $\hat{}$ represents a modelled quantity. Note that the aerodynamic conductance, g_{a_h} is not amongst the input variables because it is again assumed to be constant (0.167 m s^{-1}).

Figure 3 shows the evolution of the percentage of variance explained in the training data (a) and test data (b). Here the "Epoch Number" is the number of times in which the entire training dataset has been presented to the network. Results are shown for networks consisting of 2, 4 and 6 neurons and in each case $\mu = 0.5$ but the learning rate, η, varys with the number of neurons, K : $\eta = 0.005 \times K$.

Each of the networks fits the Penman-Monteith equation well and there is no evidence of overfitting. The 6 neuron network explains in excess of 99.9 % of the variance in the training data and more than 99.8 % of the variance in the test data after 500 epochs. The backpropogation network is obviously capable of reproducing a nonlinear function (in this case the Penman-Monteith equation) to high accuracy given good training data (in this case it is perfect noiseless data).

Fitting the Bulk Stomatal Conductance

The dependence of the bulk stomatal conductance of the Thetford forest are now fitted. Thus once again the output variable is g_S and the inputs are as described in Section 3.2.2 (but centred and scaled):

$$X_j = (\ R^*_{solar},\ \ \delta q^*,\ \ T^*,\ \ \delta\theta^*,\ \ LAI^*,\ \ t^*_{day}\ )\ \ ;\ \ Y = g_S^*\ \ ;\ \ \hat{Y} = \hat{g_S^*}$$

Figure 4 shows the evolution of the variance explained for 2, 4 and 6 neurons and Table 3.1 lists the final figures for the 4 neuron network after 2000 epochs .

The learning rates and momentum coefficients are as laid out above. The dotted line in Figure 4 represents the fits obtained using the modified Stewart-Jarvis model (Section 3.3.2). The networks with 4 and 6 neurons fit the training data significantly better than 2 neurons (81.0% and 85.0% respectively as opposed to 75.7%) and do slightly better on the test data (74.0% and 77% respectively as opposed to 72.2%). However, the greater disparity between the test and train errors suggests that the network fit does not generalise as well, and there is some evidence of overfitting in the 6 neuron version. This is not entirely unexpected since the number of adjustable weights, P, increases as:

$$P = K(J+1) + K + 1$$

where J is again the number of input variables (6 in this case) and K is the number of neurons. Thus the six neuron network has a total of 49 adjustable weights. Even if the inputs to each neuron are assumed to be the "true" independent variables (by analogy to principal components) there are still 7 adjustable output weights (V_k). With this in mind, all further analysis is based on the 4 neuron network which appears to have a dimensionality more in keeping with the conclusions of Section 3.4.4 and which shows no evidence of overfitting.

Comparison with the Stewart-Jarvis Model

Figure 5 shows the response of the network modelled $\hat{g}_S$ to the 6 input variables. The dotted lines show the corresponding response curves from the modified Stewart-Jarvis model. It should be noted here that each of the response functions were produced by varying a single input between its minimum and maximum values whilst all other inputs were held fixed at their means (as defined from the entire Thetford dataset). In fact the neural network produces highly convoluted dependencies on the environmental inputs which are not separable into distinct responses to each variable, so this figure must be interpreted with some caution.

Nevertheless some interesting conclusions may be drawn from Figure 5. It is clear that the gross dependencies in the two models are similar: g_S decreases with δq, $\delta\theta$, t_{day}, and at high T; g_S increases with R_{solar} and LAI. However, the detailed response shapes differ significantly. Thus, the radiation response is concave in the network fit but convex in the Stewart-Jarvis model and the network does not produce a regime in which g_S is independent of the humidity deficit. In addition the neural network response shows an increase in g_S at low temperatures rather than the more physiologically realistic decrease. This is most likely a consequence of supplying unrealistic driving variables (for instance, mean R_{solar} but low T) which imply a regime for which the network has not been trained. Under these circumstances the purely empirical model has been incorrectly extrapolated out of its region of calibration and unrealistic results are likely to follow.

Figure 6(a) shows a scatter plot of the Model Four (neural network) derived g_S versus Model One (modified Stewart-Jarvis) derived g_S and Figure 6(b) is likewise for the former versus the "measured" g_S values found by inversion of the Penman-Monteith equation. Qualitatively the models may appear to perform similarly despite their differing response functions, with both underestimating the high conductance values in the data. However, a comparison of the models' prediction of values for g_S suggests this to be misleading, with the network only explaining 88.1% of the train variance and 81.7% of the test variance as simulated by Model One (modified Stewart-Jarvis); see also Figure 6a. This is a crucial result. If the behaviour of all possible models for $g_S = g_S(X_j)$ were viewed in terms of the returned values of PVE, no single clear dominant minimum may be observed. The only possible explanation for this is that there are additional, currently missing, dependencies of g_S, and were these to be included, the optimisation would become better-posed, and a definite minimum found through the techniques presented above.

3.6 Some possible causes of model error

3.6.1 Introduction

The variance explained by Model One and Model Four are within 6 % of each other (Table 3.1). However, as remarked above and presented in Table 3.8 and Figure 6(a), the models are explaining different parts of the variance. Errors in g_S through ϵ are unlikely to cause such an effect, and it can only be concluded that missing physics is causing difficulties for the optimisations. The two models are explaining very well different parts of the dataset, but no single model does well overall; in this section, possible causes of modelling deficiencies are given.

3.6.2 Interannual variability

Throughout this paper, a seasonal pattern or profile of LAI is assumed to be true for all years. However, such an assumption should be questioned; the LAI profile may depend on the recent history of environmental conditions such as the rainfall (Stewart, personal communication), and indeed 1976 was an exceedingly dry year. Further, the LAI profile could have a "memory" between years where a dry year affects the bud size for the next year. The root structure and subsequent response to soil deficit could vary from year to year due to growth, or damage sustained in continued extreme conditions.

The student's t-test provides a statistical method to test for differences in means between years (Moroney, 1951). The neural network (considered here as the best model) is run throughout the entire training data set, and then against the test data. This corresponds to Model 4, Table I. Residual errors are then calculated (modelled g_S minus measured g_S), and subdivided into

year. As the years 1974, 1975, and 1976 are, in general, progressively drier, it is hypothesised that the "running mean" of errors moves from negative to positive. This corresponds to the stomatal conductance (and hence evaporative flux) being overestimated during the drier periods.

The mean of the errors, μ_p, $p = 1, 2, 3$ corresponding to 1974, 1975 and 1976 are tested against each other, where the one-tailed hypothesis tested is

$$\left.\begin{array}{ll} \mathrm{H}_0: & \mu_p = \mu_q \\ \mathrm{H}_1: & \mu_p < \mu_q \end{array}\right\} p < q. \tag{3.13}$$

The t-statistic is given by

$$t_{n_p+n_q-2} = \frac{\mu_p - \mu_q}{\sqrt{\frac{n_p S_p^2 + n_q S_q^2}{n_p+n_q-2}}\sqrt{\frac{1}{n_p}+\frac{1}{n_q}}}$$

where n_p, n_q are the number of values, and S_p^2, S_q^2 are the variances of the errors. The statistics required for this test are presented in Table 3.6, whilst the results are given in Table 3.7.

There is evidence at the 95 % level that there is an interannual variability that is as yet unaccounted for in models for g_S.

3.6.3 The effect of an understorey

Throughout this paper, the contribution of the understorey to the evaporation flux is implicitly included in the above canopy measurements. However, the LAI profile used in modelling g_S is that assumed for the trees, thereby ignoring any differences in the equivalent profile for the bracken. Further, the assumption of a single stomatal function describing the entire ecosystem may be at fault, and that in changing environmental conditions with different responses of both the trees and bracken, the Penman-Monteith Big Leaf Model will not perform adequately. A discussion of the LAI and local stomatal conductance for the bracken is described in Roberts *et al.* (1980) and Roberts *et al.* (1984).

Within the framework of the modelling techniques above, the understorey could be included with relative ease. A two source analogue to the Big Leaf Model that explicitly models a stomatal conductance of both the canopy and understorey is provided in Shuttleworth and Wallace (1985). However, precise calibration of such a model against the dataset used above will probably be difficult due to the collinearity between the two values of g_S. Extra information on the understorey could avoid this; an equivalent example but for Sahelian fallow savannah is given in Huntingford *et al.*, 1995, although the additional complexity of a two-source model for this vegetation type represented little improvement over the Big Leaf Model when predicting total evaporative fluxes.

The Neural Network circumnavigates making the lengthy algebraic manipulations necessary within a Two-Source model should the model output be the

total evaporative flux, λE. Presenting the network with the additional input of LAI for the understorey, then in theory the derived values for the adjustable weights should implicitly contain such multi-layer modelling. However, if such an approach is both practicable and useful remains an interesting question. Further to questions of multi-layer modelling, the form suggested by the network for the dependency of g_S on LAI *may* be an artifact of integrating up from leaf level to canopy level.

3.7 Summary of results and discussion

The aim of this paper is to find the best possible fit of a model for g_S to the given environmental variables, X_j and once found, this provides a benchmark for existing stomatal conductance models. In the search for such a solution, statistical techniques were applied to the Thetford data. It was found that for this particular data, collinearity was not a major problem and hence the model could not be dramatically reduced in degrees of freedom. However, to obtain a good model, nonlinear response functions for g_S were required and compared with a naive linear regression, the problem became far better posed. This observation led to performing a neural network analysis over the data, where such a technique may be viewed as an unconstrained nonlinear optimisation.

The neural network was found to fit better over the training data, but did not generalise to the test data so well. That is, there was a significant drop in the variance explained, which implies the Stewart-Jarvis model is very robust. This is because prior knowledge of likely plant physiological knowledge was used to obtain the functional forms. However, the neural network is unconstrained and subsequently suffers to a greater extent from overfitting. Like all purely empirical models, it cannot be justifiably used to extrapolate into regions where the data does not exist. Indeed, it could be argued this is the most important outcome of any principal component analysis; the components show where the data does not exist.

Although both the Stewart-Jarvis and neural network approach do well in reiterating the importance of nonlinearity in the stomatal response, both models explain different elements of the variability in g_S. This suggests that each is capturing different parts of the response, and such an effect cannot be attributed to experimental noise alone. Furthermore, given perfect training data, we expect the backpropergation neural network to replicate any reasonable nonlinear function to arbitrary accuracy. This strongly suggests there is missing driving variable(s) in the original Thetford measurements. It is believed these are associated with the longer term variabilities in the plant response, as is the evidence in Section 3.6.2. There are a number of possible causes of this modelling deficiency. Most likely is the seasonality assumption of unchanging LAI between years, particularly given the exceptionally dry period experienced in 1976. In addition, there are likely to be corresponding plant structural changes such as

root resistance and distribution.

The work reported in this paper points strongly to a need to include longer term changes in plant structure, even when modelling the variability in instantaneous flux measurements. This requires either detailed measurements of the dynamics of leaf and root growth and senescence, or else the development of models which include these processes.

3.8 References

Avissar, R.:1992, 'Conceptual aspects of a statistical-dynamical approach to represent landscape subgrid-scale heterogeneities in atmospheric models', *J. Geophysical Res.* **97** 2729-2742.

Collins, D.C. and Avissar, R.:1994, 'An evaluation with the Fourier Amplitude Sensitivity Test (FAST) of which land-surface parameters are of greatest importance in atmospheric modeling', *J. Climate* 7, 681-703.

Funahashi, K.: 1989, 'On the approximate realization of continuous mappings by neural networks', *Neural Networks* **2**, 183.

Hornik, K., Stinchcombe, M. and White, H.: 1989, 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**, 359-366.

Huntingford, C.:1995, 'Non-dimensionalisation of the Penman-Monteith model', *J. Hydrology*, In press.

Huntingford, C., Allen, S.J. and Harding, R.J.:1995, 'An intercomparison of single and dual-source vegetation-atmosphere transfer models applied to transpiration from Sahelian savannah', *Bound-Layer Meteorol*, In press.

Jarvis, P.G.: 1976, 'The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field', *Phil. Trans. Roy. Soc. Lond.*, **B 273**, 593-610.Lean, J. and Rowntree, P.R.:1993, 'A GCM simulation of the impact of Amazonian deforestation on climate using an improved canopy representation', *Q. J. R. Meteorol. Soc.* **119**, 509-530.

Jones, H.G. and Higgs, K.H.: 1989, 'Empirical models of the conductance of leaves in apple orchards', *Plant, Cell and Environment* **12**, 301-308.

Monteith, J.L.: 1965, 'Evaporation and the environment', *Symp. Soc. Expl. Biol.*, **19**, 205-234.

Moroney, M.J.: 1951, 'Facts from figures', Penguin, Harmondsworth, 472pp.

Myers, R.H.:1989, 'Classical and modern regression with applications', PWS-KENT, Boston, 488pp.

Roberts, J., Pymar, C.F., Wallace, J.S. and Pitman, R.M.: 1980, 'Seasonal changes in leaf area, stomatal and canopy conductances and transpiration from bracken below a forest canopy', *J. Appl. Ecol.* **17**, 409-422.

Roberts, J., Wallace, J.S. and Pitman, R.M.: 1984, 'Factors affecting stomatal conductance of bracken below a forest canopy', *J. Appl. Ecol.* **21**, 643-655.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J: 1986, in "Parallel Distributed Processing", Vol. 1, chapter 8, MIT Press.

Shuttleworth, W.J. and Wallace, J.S.:1985, 'Evaporation from sparse crops - an energy combination theory', *Quart. J. R. Met. Soc.* **111**, 839-855.

Stewart, J.B.: 1988, 'Modelling surface conductance of pine forest', *Agric. and Forest Met.*, **43**, 19-35.

Table 3.1: The effect of fitting different model descriptions of stomatal conductance, g_S, to pine forest data

Model no.	Model description	% var. exp'd training data	% var. exp'd test data
One	$g_S = g_{ST_{max}} LAI \left(\prod_{j=1}^{4} f_j(Y_j)\right)(1 - a_6(\text{hour} - 8))$	75.7	72.2
Two	$g_S = a_0 + \sum_{j=1}^{6} a_j X_j$	60.2	50.6
Three	$g_S = a_0 + \sum_{j=1}^{6} a_j f_j$	71.9	70.4
Four	Neural network with 4 neurons	81.0	74.0

Table 3.2: Summary statistics for training data (438 points)

Variable	R_{solar}	δq	T	$\delta\theta$	LAI	t_{day}	g_S (mm s^{-1})
Mean	419	6.41	17.0	54.2	2.27	11.8	6.21
Standard deviation	185	3.90	5.92	19.8	0.331	2.90	3.12

Table 3.3: Summary statistics for test data (146 points)

Variable	R_{solar}	δq	T	$\delta\theta$	LAI	t_{day}	g_S (mm s^{-1})
Mean	412	5.93	16.5	52.3	2.22	11.9	6.08
Standard deviation	192	3.46	5.58	18.6	0.339	3.08	3.22

Table 3.4: The effect of eliminating principal components on the linear regression in natural variables, X_j

Remaining components	1-2-3-4-5-6	1-2-3-4-*-6	1-2-*-4-*-6	1-2-*-*-*-6	1-2-*-*-*-*	1-*-*-*-
Training data Per. Var. Explained	60.2	58.4	53.9	48.3	40.0	24.4
C_p statistic	7	24.3	72.0	130.0	218.2	385.3
Test data Per. Var. Explained	50.6	49.9	48.3	38.2	27.0	18.5

Table 3.5: The effect of eliminating principal components on the linear regression in $f_j(X_j)$

Remaining components	1-2-3-4-5-6	1-2-3-4-5-*	1-2-3-4-*-*	1-2-*-4-*-*	*-2-*-4-*-*	*-2-*-*-
Training data Per. Var. Explained	71.9	71.8	63.9	51.0	37.4	21.3
C_p statistic	7	5.82	125	319	526	771
Test data Per. Var. Explained	70.4	70.3	63.5	43.4	31.0	4.6

Table 3.6: Statistics for error in prediction of g_S (mm s^{-1}) by year, using the Neural Network model

Year	1974	1975	1976
TRAINING DATA			
Number of points (I)	66	97	275
Mean error (μ)	-0.258	-0.131	0.316
Error variance (S^2)	2.759	1.889	1.599
TEST DATA			
Number of points (I)	23	28	95
Mean error (μ)	-0.550	-0.398	0.139
Error variance (S^2)	3.625	1.730	2.761

Table 3.7: t-test statistics for the intercomparison of error in prediction of g_S by year, using the Neural Network model

Years	t-statistic	Degrees of of freedom	Significant at 95 % level?
TRAINING DATA			
1974-1975	-0.529	161	No
1975-1976	-2.915	370	Yes
1974-1976	-3.091	339	Yes
TEST DATA			
1974-1975	-0.329	49	No
1975-1976	-1.559	121	No
1974-1976	-1.717	116	Yes

Table 3.8: An intercomparison *between* models of predicted values for g_S. Model k is assumed "correct", whilst model l is the test model.

Model k	Model l	Training data	Test data
1	2	74.6	71.6
1	3	94.3	95.9
1	4	88.1	81.7
2	3	78.8	74.1
2	4	63.3	51.5
3	4	85.0	80.2

3.9 Figures 1-6

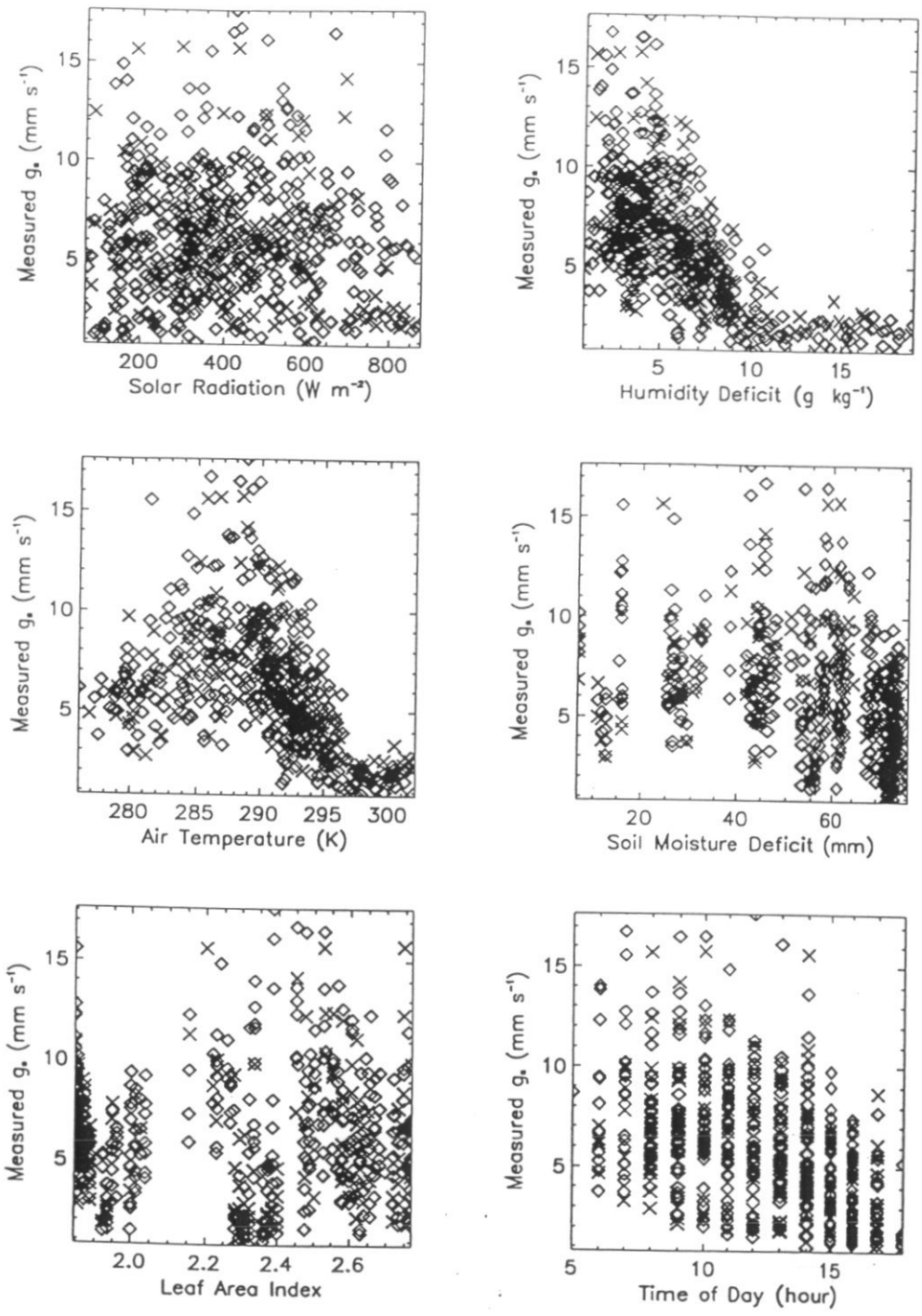


Figure 1: Thetford Forest bulk stomatal conductance versus the 6 driving variables.

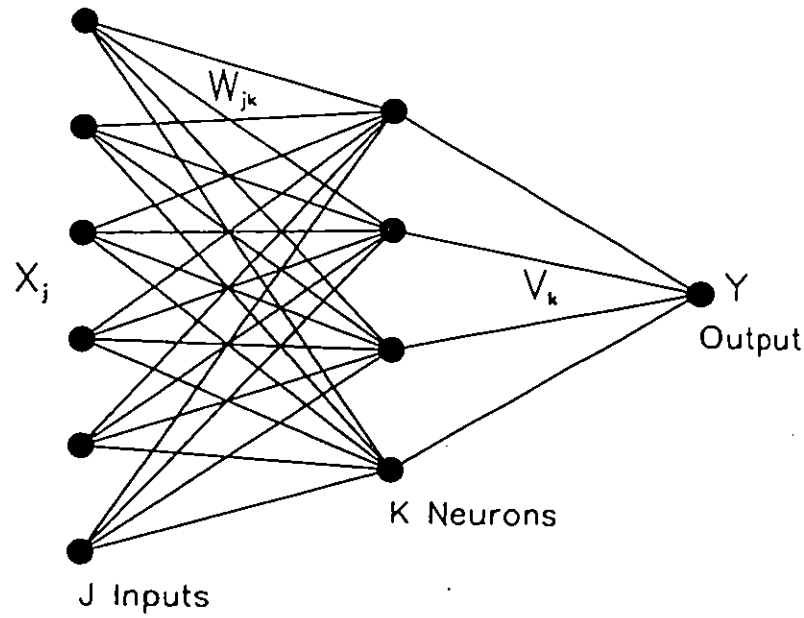


Figure 2: Schematic representation of the backpropogation neural network. For the modelling of the bulk stomatal conductance the inputs are as described in section 2.2, and the output is $\hat{g}_s$. The input weights, W_{jk}, and output weights, V_k, are updated by the learning algorithm.

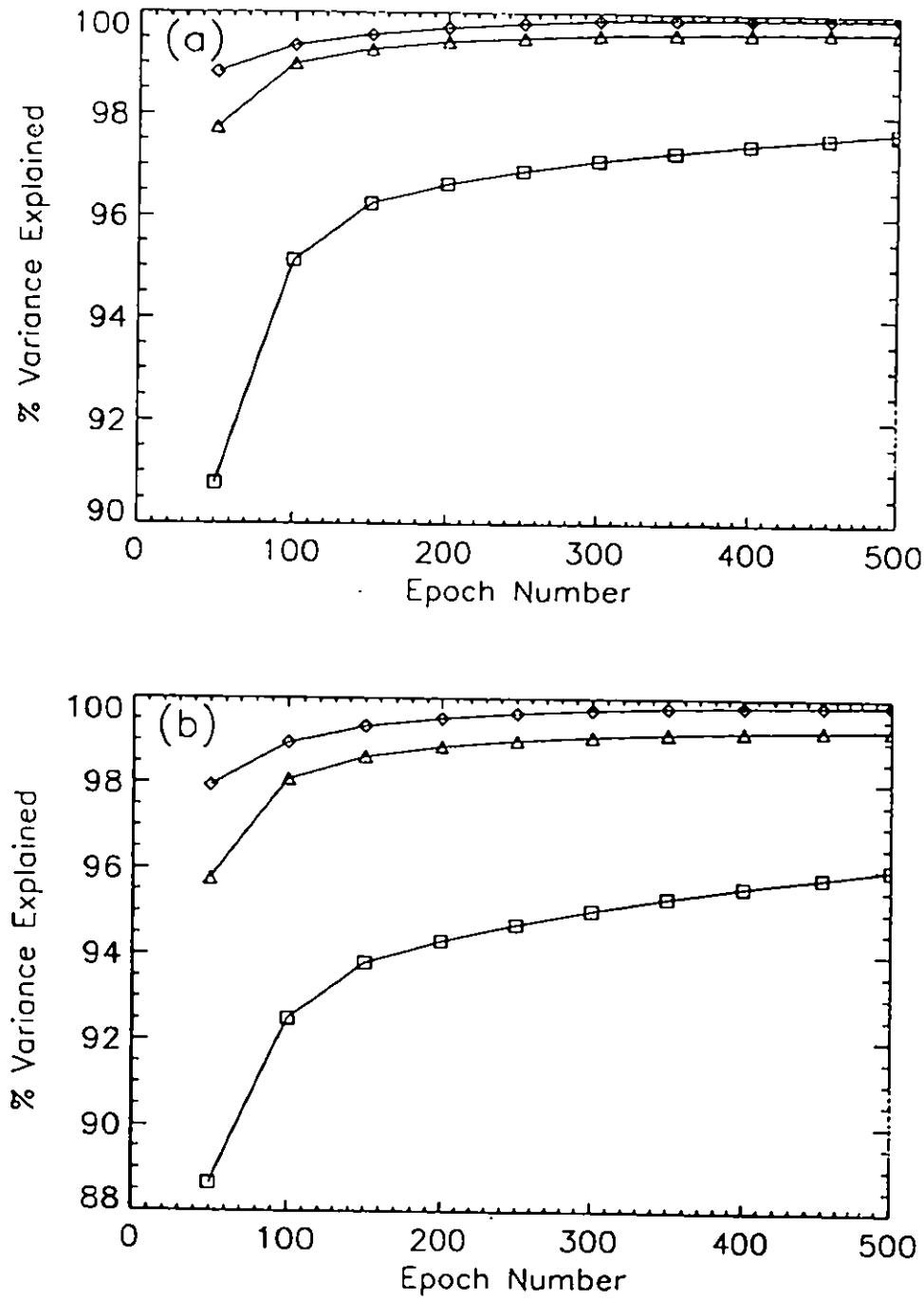


Figure 3: Fit of the backpropogation network to the Penman-Monteith equation versus training epoch, for 2 ($\triangle$), 4 ($\diamond$) and 6 (+) neurons. (a) training data ; (b) test data.

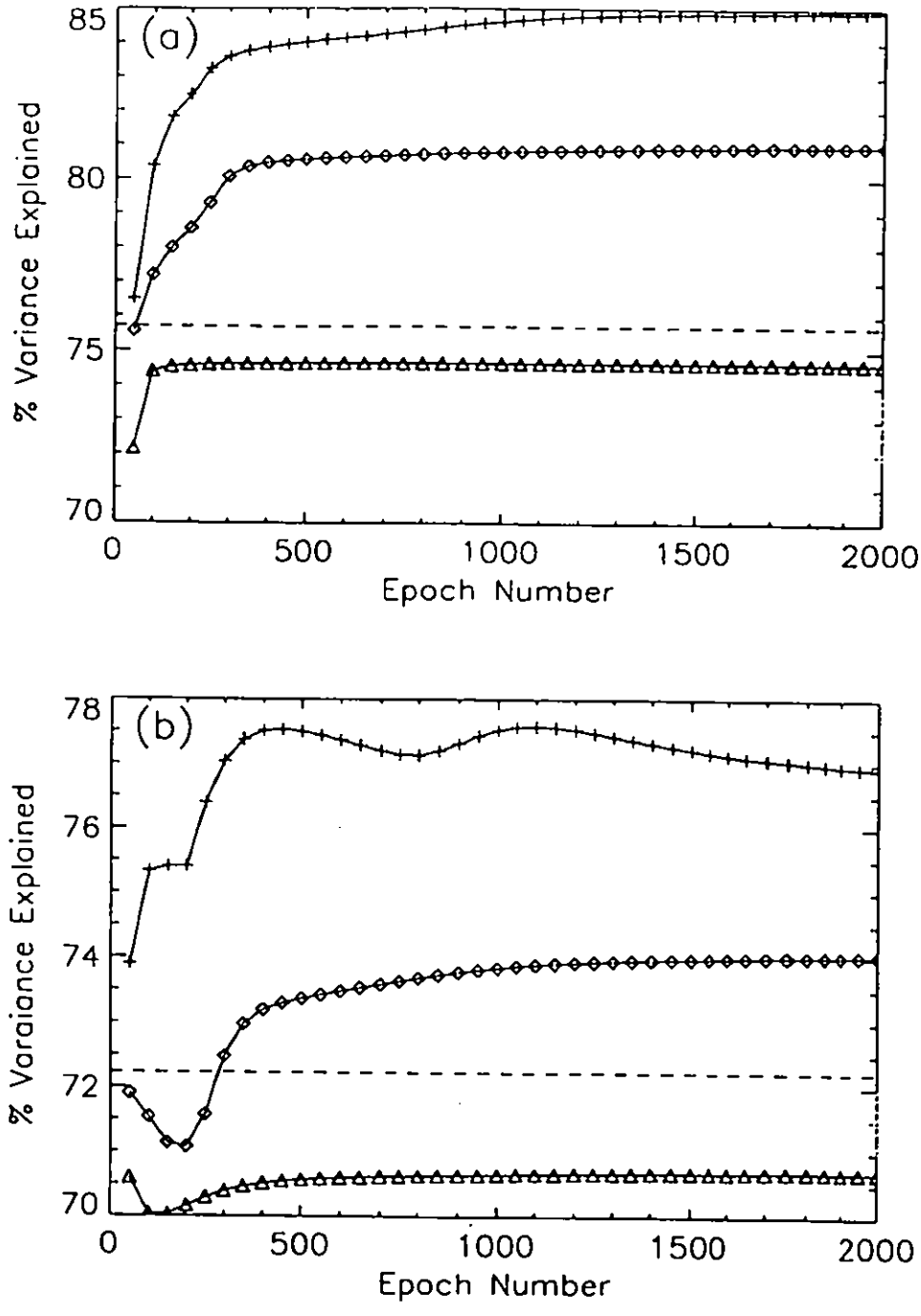


Figure 4: Fit of the backpropogation network to the Thetford bulk stomatal conductance versus training epoch, for 2 ($\triangle$), 4 ($\diamond$) and 6 (+) neurons. The corresponding variances explained by the modified Stewart-Jarvis model are given by the dotted lines. (a) training data ; (b) test data.

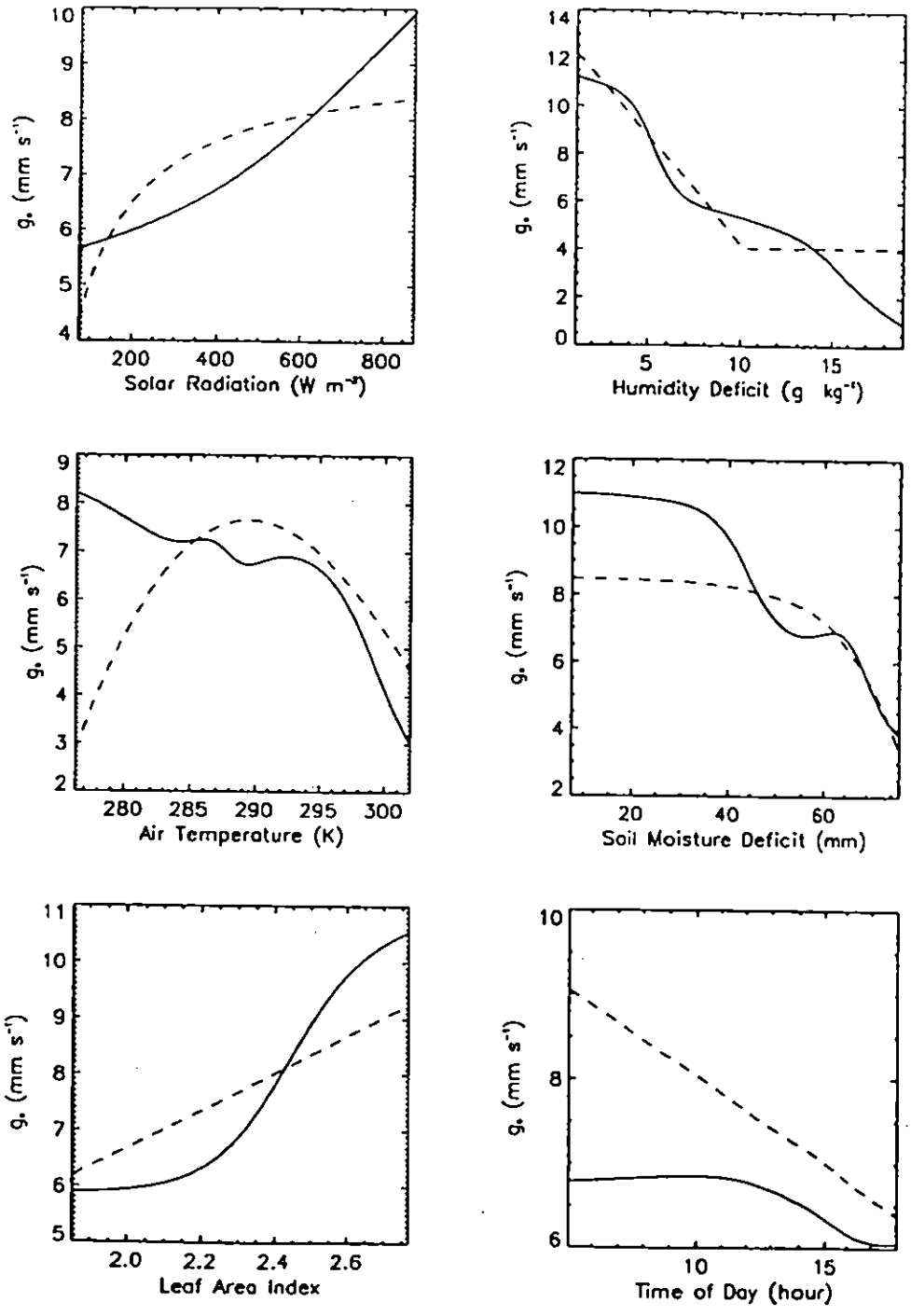


Figure 5: Response functions for the 4 neuron network (continuous line) and the modified Stewart-Jarvis model (dotted line). Each of these curves is derived by varying a single input between its minimum and maximum values whilst holding all other inputs at their mean values.

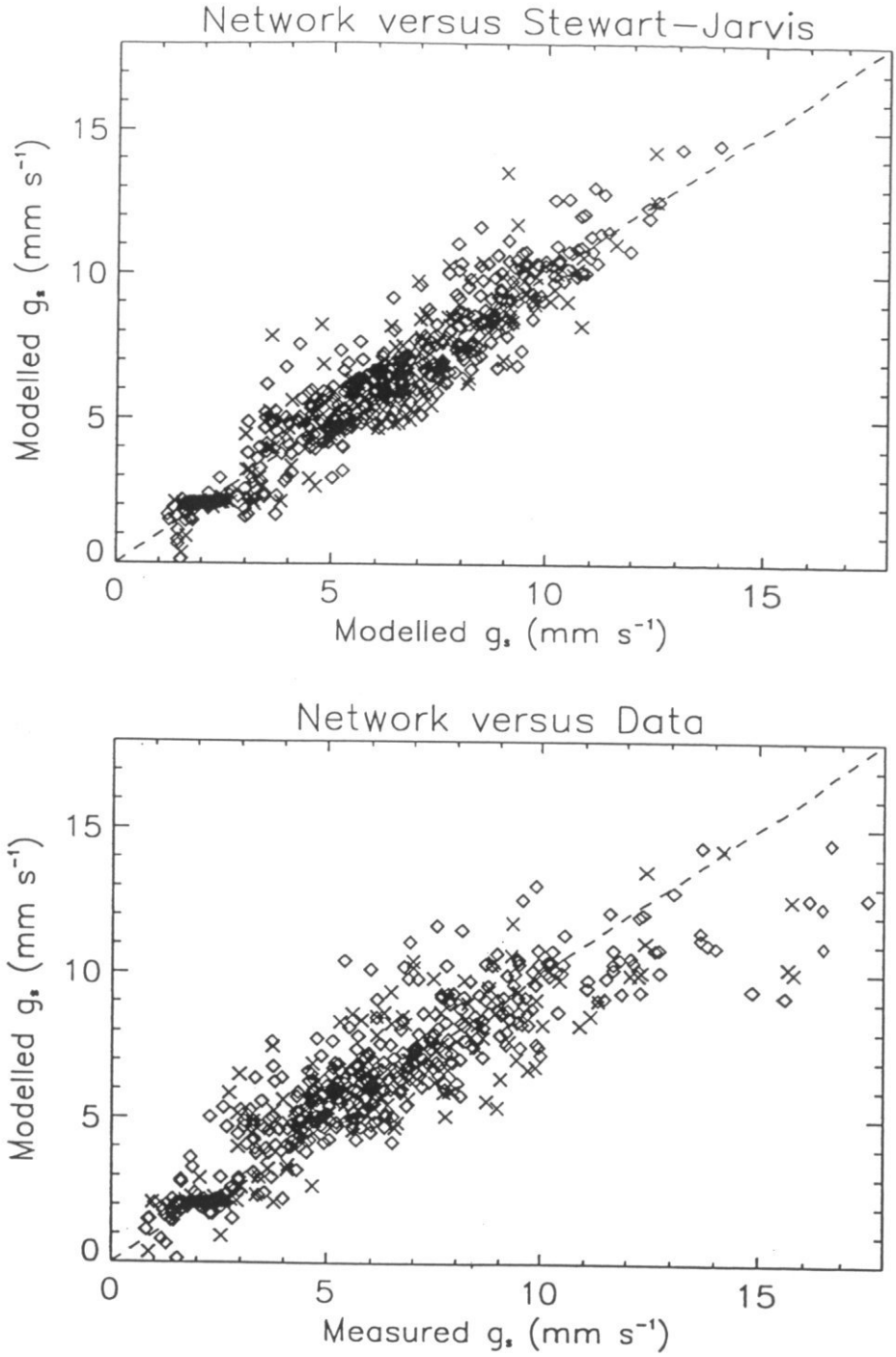


Figure 6: Scatter plot of the networked modelled g_S versus (a) the Stewart-Jarvis modelled g_S, (b) the "measured" g_S values.

Chapter 4

Overall conclusions and future work

The work under this contract has made important progress towards the implementation of interactive vegetation within the GCM of the Hadley Centre. In Chapter 2 a carbon assimilation routine suitable for use within a climate model is developed and tested. Considerable use is made of one of the few data sets currently available which contains both water vapour and carbon dioxide fluxes at the field scale. This work has confirmed the validity of a simple relationship between stomatal behaviour, humidity, soil water status and photosynthetic rate. In Chapter 3 a detailed study of the dependence of stomatal behaviour on environmental variables is presented, this work confirms in a rigorous way that the current formulations are adequate and suggests directions in which they can be improved. Again in this study there is a focus on the comparison with model performance against measurements made at the field scale.

The two studies described above will provide the basis for algorithms to calculate instantaneous carbon fluxes within a climate model. The next step will be to introduce a time dependence such that seasonal growth may be explicitly modelled. It is to be hoped that the same careful and systematic methodology will be followed in this final step, that is to investigate dependencies and sensitivity of the algorithms to environmental parameters and to test thoroughly against field data.

It is suggested that future work should have three components:

(1) Continued testing of the instantaneous flux algorithms with field data. In the next two years a number of important field data sets containing combined water vapour and carbon dioxide fluxes will become available. These data sets which were collected during recent international experiments in which the Institute

of Hydrology has participated cover a number of the important global biomes (boreal forest, tropical rain forest, semi-arid savannah and tundra surfaces) and will thus provide validation over a large range of vegetation types and climates.

(2) Development of the seasonal growth model from the instantaneous flux algorithms. The components for this model already exist but it remains to incorporate these in a form suitable for inclusion within a climate model. This should involve simplification and rationalisation aided by sensitivity analysis.

(3) Testing of the seasonal growth model against field and satellite data. This should be an essential and ongoing activity. A number of data sets already exist to undertake this activity.

Chapter 5

Acknowledgements

The work described in this report represents a truly collaborative effort between the Institute of Hydrology and the Hadley Centre. We would particularly like to acknowledge the input of Peter Cox of the Hadley Centre who has had a very significant input into the work described in Chapters 2 and 3. It is expected that Chapters 2 and 3 will appear in some form as papers in the open literature under the joint authorship of staff from the Institute and the Hadley Centre.


Natural
Environment
Research
Council