A Strategy for Calibrating Atmospheric Transport Models

I. G. Enting

DIVISION OF ATMOSPHERIC RESEARCH TECHNICAL PAPER No. 9

COMMONWEALTH SCIENTIFIC AND INDUSTRIAL RESEARCH ORGANIZATION, AUSTRALIA 1985

A Strategy for Calibrating Atmospheric Transport Models

By I. G. Enting

Division of Atmospheric Research Technical Paper No. 9

Commonwealth Scientific and Industrial
Research Organization, Australia
1985

CSIRO Aust. Div. Atmos. Res. Techn. Pap. No.9, 1-25(1985)

A STRATEGY FOR CALIBRATING ATMOSPHERIC TRANSPORT MODELS

I.G. Enting

CSIRO Division of Atmospheric Research,
Private Bag No.1, Mordialloc, Australia, 3195

Abstract

The distribution of long-lived atmospheric constituents is strongly influenced by the extent to which they are transported by large-scale atmospheric circulations. At present our knowledge of the effects of these circulations is incomplete and can in principle be improved by incorporating information on tracer distributions. Since each tracer cycle has its own inherent uncertainties as well as the uncertainties in the atmospheric transports, the overall model calibration problem becomes one of considerable complexity once a number of different tracers are considered. The application of transformations developed for analysing siesmic data can separate the calibration problem into a number of essentially independent parts, reducing the complexity to a more manageable level and allowing for systematic inclusion of additional tracer information as it becomes available.

1. Modelling Atmospheric Transports

This report investigates some of the problems involved in modelling the transports of passive tracers in the atmosphere. The 'passive' tracers are those that do not directly change the dynamics of the atmosphere and so the modelling can be performed using a kinematic description of the transports.

The main aims of studies involving the transport of trace atmospheric constituents usually fall into one of the following classes:

(i) Deducing the distributions of sources and sinks from the observed distribution of a tracer.

(ii) Using the observed distribution of a tracer to deduce large-scale features of the atmospheric circulation such as transport rates between the two hemispheres.

(iii) Predicting the future concentrations of various atmospheric constituents, given scenarios for changes in the sources.

(iv) Extending such studies from 'normal' conditions to the 'abnormal' conditions ranging from the relatively small anomalies in transports in particular years right through to the major changes that might be associated with a nuclear war.

The various studies of atmospheric constituents will use various types of transport models with varying degrees of spatial resolution. The primary feature determining the appropriate spatial resolution is the time-scales involved in the cycles of a particular tracer. Figure 1.1 is a schematic representation of the time scales associated with atmospheric mixing on various length scales. One important aspect is the differences between the length vs time curves in the three different directions.

These scale differences mean that if the time scales involved are too short to justify a zero-dimensional model, ignoring atmospheric transports altogether, then a one-dimensional model with only vertical resolution may be appropriate. Alternatively a one-dimensional model with only horizontal mixing may be appropriate in other cases - the mixing times from north to south and from troposphere to stratosphere are similar. The most appropriate direction to resolve is the one in which there is most variability.

For situations with shorter time scales or with significant variability in both horizontal and vertical directions, a two-dimensional (zonally-averaged) model becomes appropriate, so long as the time scales are long enough to ensure zonal mixing. Finally if very short time-scales are of interest or if there is variability in all three spatial directions, a full three-dimensional model is appropriate.

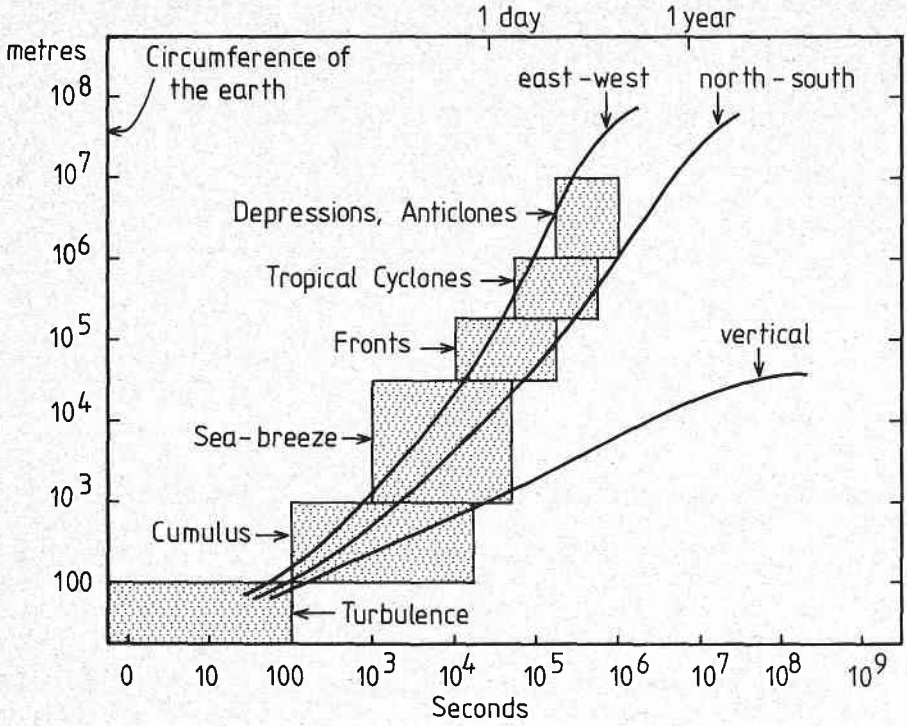

Figure 1.1: The length and time scales of various atmospheric processes showing how on particular time scales different length scales will be involved in the three directions.

When using the lower-dimensionality models, it is important to note the different possible ways in which they can be interpreted. Bauer (1980) has listed three different ways in which one-dimensional (height only) models can be interpreted, namely

(i) as representing global average height distributions;

(ii) as representing hemispheric averages;

(iii) as representing typical mid-latitude (30°N) distributions.

Similarly, one-dimensional models in which only latitude is resolved can be interpreted as representing

(i) surface values (averaged at that latitude);

(ii) tropospheric averages (at that latitude);

(iii) atmospheric averages (at that latitude).

Some insight into the ways in which these different types of one-dimensional model should be treated can be obtained by investigating the way in which such models relate to a fuller two-dimensional model of the atmospheric transports.

Zero-dimensional (single reservoir models) can represent either the troposphere or the atmosphere as a whole. Two-dimensional models are almost always taken as representing zonally averaged distributions.

This report is concerned with the question of calibrating atmospheric transport models and while much of the analysis is quite general, it is two-dimensional models that are of most concern here. Two-dimensional models have much more complexity than one-dimensional models because they can include both advective and diffusive parameterization of the transports. Two-dimensional models also have complications that are missing from three-dimensional modelling studies because the process of zonal averaging transforms some three-dimensional advective process (in particular those involving eddies) into two-dimensional 'diffusive' processes in a poorly determined manner. Thus two-dimensional models may be more complicated to calibrate than three-dimensional models because they represent some atmospheric processes in an indirect manner.

The remainder of this report is as follows. Section 2 reviews some of the different tracers that are of possible use in calibrating two-dimensional models and gives a discussion of the uncertainties in the source/sink terms. Section 3 describes the calibration problem and the way in which the use of a large number of tracers, each with its own uncertainties, leads to a calibration problem that appears to have an unmanageable degree of complexity. Section 4 shows how techniques developed by Pavlis and Booker (1980) for analysing earthquake data can separate the overall calibration problem into a set of managable sub-problems. Section 5 extends this approach and analyses an iterative calibration scheme that is needed because of non-linear aspects of the problem and because of the approximate nature of the initial separation into subproblems. Section 6 discusses the computational requirements for implementing this calibration scheme.

2. Data for Calibrating two-dimensional models

There are four main classes of quantity that can be used in the calibration of atmospheric transport models. These are the 'meteorological variables', other natural trace constituents, products of nuclear testing, an other anthropogenic trace constituents.

(a) Meteorological variables

The quantities involved are:

(i) Mass: The mass distribution is often used as the basis of two-dimensional modelling with the vertical distribution being expressed relative to pressure (i.e. mass) coordinates. If advective flows are expressed in terms of stream functions then mass conservation follows automatically.

(ii) Entropy: Entropy (heat) is of limited value in calibrating transport models directly, partly because it is not a passive tracer but mainly because it is not conserved by the transport processes. As described by Hantel and Haase (1983), one of the main uncertainties in the atmospheric heat budget is the vertical transport by transient eddies - a process that will have a non-linear dependence on the heat distribution. One way in which the heat distribution constrains the transport coefficients is that the

major axis of the diffusion tensor is expected to be along the isentropes. (Plumb pers. comm.).

(iii) Water: In the troposphere water is of little use as a tracer since it has an active dynamical role and in particular it is not conserved by the transport processes. It has however been used by Hidalgo and Crutzen (1977) for determining eddy diffusion coefficients in the stratosphere.

(iv) Momentum/Angular momentum: Oort (1983, Appendix B) shows that over much of the earth, particularly in the southern hemisphere, the mean meridional circulation cannot be determined directly from the available observations but must be deduced indirectly from angular momentum balance. Thus if these mean circulations are used then angular momentum balance is already implicit in the transport data. However Plumb (1979) has suggested that effective circulations related to the Lagrangian mean may give better descriptions of zonally averaged transports.

(b) Natural trace constituents

(i) Carbon dioxide: The natural seasonal cycle of CO_2 gives a useful tracer, particularly for studies of interhemispheric transport. The problems are that the source and sink strengths are very poorly known, involving both the biosphere and the oceans.

(ii) Carbon-13: The studies of seasonal variability of carbon isotope ratios could help to resolve the problems of the relative strengths of biospheric and oceanic contributions to the cycle but the signal is so small that the measurements are very difficult.

(c) Products of Nuclear Testing

The use of any of these tracers is limited by the uncertainties in the release rates. However many aspects of the source function are the same for all of these tracers - the main exception being the distinction between fission and fusion. In addition the location and timing of the major releases is fairly well known - it is the amounts released that are uncertain. The most important species are:

(i) Carbon-14: This has been widely observed and because of the large difference between the stratospheric and tropospheric concentrations, it should provide useful information in the seasonal variability of transport between the stratosphere and troposphere.

(ii) Strontium-90: This is a fission product and is subject to relatively rapid rainout from the troposphere once formed. It has been widely measured because of concern over the dangers to health arising from its chemical similarity to calcium (Zirconium-97 is a similar fission product).

(iii) Tritium: A product from fusion only. Tritium distributions in the oceans have been subject to extensive study but atmospheric studies have been much more restricted.

(d) Other Anthropogenic Constituents

(i) Chloroflourocarbons: Compounds such as CCl_3F (Freon-11) and CCl_2F_2 (Freon-12) have been released at rates that increased steadily at least till the mid-seventies. The main sources were in the northern hemisphere. This has led to a gradient between the northern and southern hemispheres and between the stratosphere and troposphere and these gradients can help to calibrate the transport rates between these regions. There is however a problem with CCl_3F. Fraser et al. (1983) have indicated that the observed rates of change of concentration appear to be inconsistent with the release rates that are calculated from production data. They calculated an alternative release function that gives agreement with the observations. That calculation must be regarded as preliminary because the model itself was partly calibrated using distributions of CCl_3F. Such an approach can quickly lead to circularity in the calibration procedure. A consistent approach is to estimate corrections to the release rate and the transport parameters simultaneously, and then estimate the combined uncertainties. This calculation can be simplified by using a separation procedure similar to that described in Section 4, below.

(ii) Krypton-85: This nuclide is produced by nuclear reactors and has a half-life of 10.76 years. Weiss et al. (1983) have observed interhemispheric gradients of Krypton-85 at sea-level in the Atlantic.

(iii) Carbon Dioxide: The steady increase in fossil carbon release has led to an interhemispheric gradient in concentration but the analysis is complicated by contributions from seasonal covariance of atmospheric transport and ocean uptake. The vertical distribution may be of interest in determining rates of transport into the stratosphere, especially in view of the problems associated with CCl_3F.

3. The calibration problem

The discussion in the previous section reveals gaps in our knowledge of the sources and sinks of each of the potential tracers. Any model calibration technique that ignores this problem runs the risk of being biased (if the uncertainties in the sources and sinks are ignored) or circular (if an iterative procedure is used to determine the sources and sinks and the atmospheric transports from a single data set). The preferred approach is to estimate the transport parameters and the sources and sinks simultaneously with as much accuracy as is possible given the data available. In practice the estimation procedure will almost certainly be under-determined and some form of constrained inversion (Twomey, 1977) will be required.

Any direct attempt to estimate both the source/sink strengths and the transport parameters simultaneously is likely to be unacceptably complicated. As an example, the distribution of CCl_3F is very difficult to interpret even when no other tracers are considered. Hyson et al (1980) used

the horizontal and vertical distributions of CCl_3F to 'tune' the diffusive transports in their two-dimensional model, assuming that the direct estimates of releases were correct. Later studies by Fraser et al. (1983) used time series for CCl_3F concentrations to infer that there were inconsistencies in the release data, and modified release rates were calculated by fitting the observed time series to the results of the model. This type of procedure, using similar data to tune both the transports and the release rates, involves a grave risk of circularity even though in this particular case subsequent calculations (Enting, unpublished) seem to confirm that the problem with the CCl_3F release data in not an artifact of the calibration procedure. Obviously a more detailed analysis of the use of the CCl_3F data is needed in order to determine which data can be used in which aspect of the calibration. Apart from the uncertainties mentioned above, the atmospheric lifetime of CCl_3F is not entirely certain and one of the main aims of the observational and modelling programs is to determine this lifetime. Again, a detailed analysis of the use of the data is required to determine whether the lifetime can be estimated, given all the other uncertainties. Formally we could attempt to obtain simultaneous estimates of an atmospheric lifetime and corrections to horizontal and vertical transports and corrections to the direct estimates of release rates. Since however it seems that these four factors can be estimated from four different sets of CCl_3F data, (horizontal and vertical gradients for the respective transport coefficients, atmospheric inventories for the lifetime and the curvature in the time series for the release rates) a mathematical procedure that exploits this approximate division into subproblems is desirable.

Even when different tracers are considered, the calibration problems are not independent because in each case the atmospheric transports are the same so long as the tracers are conserved. The calibration involves determining the uncertainties in a number of source characteristics that are connected because of the uncertainties in the transports.

The computational procedure should reflect this structure of inter-related subproblems so far as is possible for the following reasons:

(i) to reduce the size of the individual calibration problems;

(ii) to improve the numerical stability of the procedure;

(iii) to give a more comprehensible description of the operation of the calibration procedure;

(iv) to allow for the inclusion of additional tracers, to the extent that the relative significance of new data can be assessed in the context of one 'subproblem' without having to completely recalibrate the model.

The comprehensibility of the calibration procedure is perhaps the most important of these considerations. For many tracers, the process of simultaneous improvement of our knowledge of the tracer cycle and the atmospheric transports will be highly assymmetric. For some tracers, the uncertainties in the sources and sinks will be so large that our (incomplete) knowledge of atmospheric transports enables the net release rates to be determined more accurately while for other tracers, at least some aspects of

their cycle will be known sufficiently well for it to be possible to make significant improvements in the determination of atmospheric transport parameters. The calibration procedure should be set up so as to indicate which data are significant for which aspects of the calibration. This is a study of what Jackson (1972) has called 'the marginal utility of data'.

4. The linear approximation and the analogy with seismic analysis

The calibration problem described in the previous section was one of deducing a single set of transport coefficients from a set of independent responses to independent sources each of which has its own independent uncertainties. The solution described in this section is taken by direct analogy from the problem of deducing velocity profiles from earthquake data. This problem can be described in almost identical words to the atmospheric problem. The requirement is to deduce the 'transport' coefficients (the velocity profile for compressive (or 'p') waves) given a set of independent responses to independent sources (i.e. individual earthquakes) each of which has its own uncertainties (i.e. the positions of the hypocentres and the times of the original shock are unknown). Pavlis and Booker (1980) showed that, in a linear approximation, the earthquake problem can be transformed so that it breaks into separate subproblems. They were most interested in separating the 'velocity-profile' subproblem from the 'source' subproblem so that different inversion techniques can be applied to each subproblem. They did however point out that the separation produced an independent subproblem for each separate source (i.e. each distinct earthquake). The same transformations apply equally well to the atmospheric transport calibration problem because in mathematical terms the linearised forms of the problem are identical. (The use of these solutions to the linearized problem to give an iterative solution of the non-linear problem is discussed in the following section).

The linear problem described by Pavlis and Booker (1980) can be regarded as a linearisation about some initial solution described by P transport parameters $x_p^{(o)}$ and K source parameters $s_k^{(o)}$. In the seismic problem the x_p represent a parameterisation of the velocity profile while in the atmospheric transport problem they would be a parameterisation of the atmospheric transport processes. The s_k describe the times and positions of the hypocentres in the seismic problem or the various unknown source strengths in the atmospheric transport problem.

$$y_i = y_i^{(o)}(\underline{x}^{(o)}, \underline{s}^{(o)}) + \sum_{k=1}^{K} A_{ik} s_k + \sum_{p=1}^{P} G_{ip} x_p, \qquad i=1,I \tag{4.1}$$

where $y_i^{(o)}$ is the initial model prediction of the ith of I observations m_i and the s_k and x_p represent deviations from the initial parameters. Fitting the observations m_i reduces to solving

$$\mu_i = \sum_{k=1}^{K} A_{ik} s_k + \sum_{p=1}^{P} G_{ip} x_p, \qquad i=1,I \tag{4.2}$$

where

$$\mu_i = m_i - y_i^{(o)} , \quad i=1,I . \tag{4.3}$$

The procedure described by Pavlis and Booker is to evaluate

$$\hat{s}_k = \sum_i (A^{\dagger})_{ki} \mu_i \tag{4.4}$$

where $\underline{A}^{\dagger}$ is the Moore-Penrose pseudo-inverse (see Deutsch, 1965) of $\underline{A}$. This matrix is expressed in terms of the singular-value decomposition of $\underline{A}$, which can be written as

$$A_{ik} = \sum_{i'=1}^{I} \sum_{k'=1}^{K} U_{ii'} \Lambda_{i'k'} V_{k'k} \tag{4.5}$$

where $\underline{U}$ and $\underline{V}$ are orthogonal matrices and $\Lambda_{i'k'}$ is zero except for the first r elements on the leading diagonal which are equal to λ_i, the r non-zero singular values of $\underline{A}$. The pseudo-inverse is

$$(A^{\dagger})_{ki} = \sum_{k'} \sum_{i'} V_{kk'} (\underline{\Lambda}^{-1})_{k'i'} U_{i'i} \tag{4.6}$$

where $\underline{\Lambda}^{-1}$ is zero apart from the first r elements of the leading diagonal which are given by λ_i^{-1}.

The reasons for choosing the solution (4.4) are (Deutsch, 1965):

(i) It is a least-squares solution that minimises the sum of squares of residuals $\Sigma\rho_i^2$

where $$\rho_i = \mu_i - \Sigma A_{ik} s_k . \tag{4.7}$$

and (ii) Where the least-squares solution is not unique (i.e. if r, the rank of matrix $\underline{A}$ is less than K the number of source parameters) then solution (4.4) selects the least squares solution that minimises the deviations s_k (i.e. Σs_k^2 is minimised subject to the constraint of $\Sigma\rho_i^2$ being a minimum).

The other advantage of the use of the pseudo-inverse in equation (4.4) is that it allows for the separation of the 'transport problem' (i.e. estimating the x_p) from the 'source problem'.

Initially there are I measurements (d in the notation of Pavlis and Booker) from which to estimate K source parameters (p in the notation of Pavlis and Booker) plus a number of transport parameters.

The estimate $\hat{\underline{s}}$ defined by (4.4) lies within an r dimensional subspace of possible $\underline{s}$ vectors. It is assumed that K < I and of necessity, r $\leq$ K so

$$r \leq K < I \tag{4.8}$$

The effective use of r independent equations to obtain $\hat{\underline{s}}$ leaves K-r independent equations from which to estimate $\underline{x}$. Pavlis and Booker show how these equations can be constructed explicitly using the orthogonal matrices $\underline{U}$ and $\underline{V}$ that generate the singular value decomposition (4.6). The equations correspond to fitting the (transformed) linearised model to a transformed data set that they call the annulled data set.

In the iterative solution to the non-linear calibration problem (see Section 5 below) this independence disappears and so the generalised formalism does not attempt to transform the data set explicitly. In any case the procedure described below uses a solution matrix that is only approximately equal to the pseudo-inverse and so the precise separation would not occur. The quantities that are of interest in the general case are the sizes of the subspaces involved since these indicate how much of the data is actually used in each of the subproblems.

5. Iterative solution of the non-linear calibration problem

In general, models of atmospheric transport can not be represented linearly and so linear formalisms such as that of Section 4 can only be used iteratively. This Section describes a calibration procedure based on iterative use of a linear scheme closely related to that described in Section 4.

In order to describe the general calibration procedure it is first necessary to define the notation:

α,β are used as indices denoting particular tracers. Sums over α,β are sums over all tracers and the limits are not given explicitly in the equations below;

i,j are indices for observations of tracer concentrations and are used in combinations α,i; β,i etc. The ranges of i,j will depend on the tracer. Sums over α,i etc. are sums over all observations of all tracers;

k,k' are indices for the parameters describing the tracer source/sink strengths and are used in combination such as α,k; β,k' etc. The ranges of k,k' depend on the number of parameters used to characterise the sources and sinks of a particular tracer. Sums over α,k etc. are over all the source/sink parameters of all the tracers;

p,p' are indices for the transport parameters. The range of these indices depends on the resolution that is chosen for the description of the transports;

$s_{\alpha k}$ is the kth parameter describing the source/sink strengths for the αth tracer. Statistical estimates of this parameter are denoted $\hat{s}_{\alpha k}$.

$r_{\alpha k}$ is a prior estimate of the value of $s_{\alpha k}$. This estimate should be obtained independently of the tracer concentration data which will be used to produce the refined estimates $\hat{s}_{\alpha k}$. The ideas behind the statistical description of constrained inversion calibrations are discussed by Rodgers (1977) and discussions of the application of these ideas in a geophysical context are given by Enting (1983) and Enting and Pearman (1983). It is assumed that the units of all of the $r_{\alpha k}$ and the corresponding $s_{\alpha k}$ are chosen so that the variances of the prior estimates $r_{\alpha k}$ are equal to 1 and that the parameterization is such that covariances between distinct $r_{\alpha k}$'s are all zero. This assumption is made for the purposes of simplifying the description given in this section. It does not represent a fundamental restriction of the method because any parameterization that is in a more general form can be transformed into the special form used here. (Enting and Pearman, (1983) describe and apply a similar formalism using the more general description in which variances are not restricted to unity but covariances are required to be zero. This degree of generality is probably the most appropriate form for actual calculations).

x_p is the pth transport parameter. Again the common statistical notation x_p is used to denote estimates of the 'true' value that should be assigned to the variable x_p.

q_p is a prior estimate of the value of x_p. This estimate should be obtained independently of the observations $m_{\alpha i}$. Again it is assumed that the units of x_p and q_p are such that the variances of the q_p are all equal to 1 and the parameterization is such that all the covariances are zero.

$m_{\alpha i}$ is the ith observed value of the concentration of the αth tracer. It is assumed that the units of all the $m_{\alpha i}$ (and the corresponding $y_{\alpha i}$) are such that the variances of the $m_{\alpha i}$ are all equal to 1, and that the $m_{\alpha i}$ represent independent observations so that the covariances are zero. Again this restriction is for the purposes of simplifying the present description and Enting (1983) describes a corresponding formalism with no restriction of the variances of the $m_{\alpha i}$.

$y_{\alpha i}$ is the model prediction for $m_{\alpha i}$. This prediction will depend on the values of the transport parameters x_p and on the source strengths $s_{\alpha k}$. The calibration procedure consists of adjusting the parameters x_p and $s_{\alpha k}$ so that the $y_{\alpha i}$ agree with the $m_{\alpha i}$.

The calibration procedure that is proposed here is based on minimising the sum of squares θ, given by

$$\theta = \sum_{\alpha i}(y_{\alpha i}-m_{\alpha i})^2 + \gamma \sum_{p}(x_p-q_p)^2 + \sum_{\alpha k} \eta_\alpha (s_{\alpha k}-r_{\alpha k})^2 \quad . \qquad (5.1)$$

The sum of squares θ is to be minimised with respect to the x_p and the $s_{\alpha k}$. The weighting factors should ideally be set to one - Enting (1983) describes circumstances in which other values may be appropriate and discusses the interpretation of the results in these cases. In addition, the computational experience involved in obtaining the results described by Enting and Pearman (1983) suggests that the use of weighting factors greater than 1 in the early steps of an iterative procedure can improve the numerical stability of such procedures. (This use of the weighting factors is essentially the Marquardt method for function minimisation - see Bard, 1974, Section 5.8).

Enting (1983) has pointed out that there are a number of different approaches that suggest minimising θ (for $\gamma = \eta_\alpha = 1$).

(i) The minimisation of θ can simply be regarded as a least squares fit to an extended data set consisting of 3 classes of data: the $m_{\alpha i}$, the q_p and the $r_{\alpha k}$.

(ii) The analysis can be interpreted as a Bayesian estimation procedure (see Box and Tiao, 1973). The data $m_{\alpha i}$ are used to obtain refined (posterior) estimates of the parameters $s_{\alpha k}$ and x_p for which prior estimates ($r_{\alpha k}$ and q_p) are available. This interpretation is most direct when the multivariate distributions are normal.

(iii) The analysis can be regarded as a constrained inversion (see for example Jackson, 1972; Twomey, 1977) of the type common in geophysical analysis. The information $r_{\alpha k}$, q_p acts as constraints when the information $m_{\alpha i}$ would lead to an ill-conditioned estimation procedure if used on its own. Rodgers (1977) has emphasised that the appropriate statistical interpretation of constrained inversion techniques is obtained by reducing the procedure to case (i) above.

The calibration procedure embodied in the minimisation of θ is formally the 'all-at-once' procedure described in Section 3. Since this is, formally, optimal (in the sense of minimum variance of the estimates) it is this problem that we really want to solve. What is presented in this Section is a technique for solving the combined problem as a sequence of subproblems in a way that reproduces some of the desirable features of the analysis given in Section 4.

ALGORITHM

A. Initialise $s_{i\alpha}$ to $r_{i\alpha}$, x_p to q_p and γ, η_α to appropriate values;

B. For each α, estimate the $s_{i\alpha}$ by 'minimising' θ with fixed values of x_p and $s_{i\beta}$ $(\beta \neq \alpha)$.

C. Estimate x_p by 'minimising' θ with the $s_{i\alpha}$ fixed at the values obtained in B.

D. Repeat B and C to obtain iterative improvements to the x_p and $s_{i\alpha}$, adjusting the weighting factors - the final estimates should use weighting factors of 1.

Remarks

(i) The minimisation step in B and C need not be an exact minimisation at each stage. An approximate solution of the subproblems is adequate so long as the overall procedure converges. In particular linearised approximations to the subproblems can be solved at each stage.

(ii) If the overall calibration problem were linear then the technique would be equivalent to a Gauss-Seidel solution of the linear equations applied to blocks of variables. Convergence will be ensured if the 'diagonal' elements are sufficiently large, and in particular if the weighting factors γ and η_α are sufficiently large.

(iii) The 'Newton' method of minimising θ would be to use the first and second derivatives of θ with respect to all of the unknown parameters, (i.e. the x_p and the $s_{\alpha k}$). This technique would have quadratic convergence (see Bard, 1974, Section 5.6). By breaking the calibration into subproblems we are precluding any use of the 'mixed' second derivatives $\frac{\partial^2\theta}{\partial x_p \partial s_{\alpha k}}$, $\frac{\partial^2\theta}{\partial s_{\alpha k} \partial s_{\beta k'}}$

The most convenient way of minimising θ seems to be to apply a further restriction and only consider those contributions to the second derivative of θ that involve products of first derivatives of the factors whose squares are summed. (Bard, 1974, Section 5.9 refers to this as the Gauss method).

It is readily seen that any solutions obtained by this iterative procedure will be solutions of the original problem of minimising θ with respect to all the x_p and $s_{\alpha k}$. This is because convergence of the iteration occurs when the changes produced by each step go to zero, i.e.

$$\frac{\partial\theta}{\partial x_p} = 0 \qquad \text{for all } p \qquad (5.2a)$$

and

$$\frac{\partial\theta}{\partial s_{\alpha k}} = 0 \qquad \text{for all } \alpha, k \qquad (5.2b)$$

The combined equations (5.2a,b) constitute the so-called 'normal equations' defining the minimum of θ.

At this point the analogy with the seismic problem described in Section 4 becomes useful as a basis for describing the behaviour of the calibration procedure.

We define a matrix $\underline{A}^{(\alpha)}$ such that

$$A_{ik}^{(\alpha)} = \frac{\partial y_i}{\partial s_{\alpha k}} \tag{5.3}$$

so that in a linearised approximation, the minimisation subproblem for tracer α uses

$$\theta_\alpha \cong \sum_i (y_{\alpha i}^{(o)} + \sum_k A_{ik}^{(\alpha)} s_{\alpha k} - m_{\alpha i})^2 + \eta_\alpha \sum_k (s_{\alpha k} - r_{\alpha k})^2 \quad . \tag{5.4}$$

Dropping the tracer index α and minimising with respect to the s_k gives

$$\sum_i \left(y_i^{(o)} + \sum_{k'} A_{ik'} s_{k'} - m_i \right) A_{ik} + \eta (s_k - r_k) = 0$$

$$\sum_{k'} \left(\sum_i A_{ik} A_{ik'} + \eta I_{kk'} \right) s_{k'} = \eta r_k + \sum_i \left(m_i - y_i^{(o)} \right) A_{ik} \tag{5.5}$$

The corresponding equation with $\eta = 0$ can be written in matrix form as

$$\underline{A}^T \underline{A} \underline{s} = \underline{A}^T \underline{\mu} \tag{5.6}$$

The pseudo-inverse gives the solution of equation 5.6 as

$$\underline{s} = \underline{A}^\dagger \underline{\mu} \tag{5.7}$$

since

$$(\underline{A}^T \underline{A}) \ \underline{A}^\dagger \underline{\mu}$$

$$= \underline{A}^T (\underline{A} \underline{A}^\dagger) \underline{\mu} \ , \quad \text{(associativity)} \tag{5.8a}$$

$$= \left((\underline{A} \underline{A}^\dagger)^T \underline{A} \right)^T \underline{\mu} \ , \quad \left((\underline{X} \underline{Y})^T = \underline{Y}^T \underline{X}^T \right) \tag{5.8b}$$

$$= (\underline{A} \underline{A}^\dagger \underline{A})^T \underline{\mu} \ , \text{(since } \underline{A} \underline{A}^\dagger \text{ is symmetric)} \tag{5.8c}$$

$$= (\underline{A})^T \underline{\mu} \ , \qquad (\underline{A} \underline{A}^\dagger \underline{A} = \underline{A} \text{ defines } \underline{A}^\dagger) \tag{5.8d}$$

as required.

(The various matrix properties of the pseudoinverse are given by Deutsch (1965)). Thus even when the equations

$$\underline{A}\,\underline{s} = \underline{\mu} \tag{5.9}$$

have no solution the related equations $\underline{A}^T\,\underline{A}\,\underline{s} = \underline{A}^T\,\underline{\mu}$ always do have at least one solution which is a least squares solution of equations (5.9).

When we have already reduced the problem to the least squares form (5.6) it is more convenient to use the pseudo inverse of $\underline{A}^T\underline{A}$ which is given by

$$(\underline{A}^T\underline{A})^\dagger = \underline{A}^\dagger(\underline{A}^T)^\dagger \quad . \tag{5.10}$$

This can be proved by substituting (5.10) into the defining equation

$$(\underline{A}^T\,\underline{A})(\underline{A}^T\,\underline{A})^\dagger\,(\underline{A}^T\,\underline{A}) = \underline{A}^T\,\underline{A} \tag{5.11}$$

The solution to (5.6) is written as

$$\underline{s} = (\underline{A}^T\underline{A})^\dagger\,\underline{A}^T\,\underline{\mu}$$

$$= \underline{A}^\dagger\,\underline{A}^{T\dagger}\,\underline{A}^T\,\underline{\mu} \qquad \text{(from 5.10)} \tag{5.12a}$$

$$= \underline{A}^\dagger\,\underline{A}^{\dagger T}\,\underline{A}^T\,\underline{\mu} \tag{5.12b}$$

$$= \underline{A}^\dagger\,(\underline{A}\,\underline{A}^\dagger)^T\,\underline{\mu} \qquad \text{(as in 5.8b)} \tag{5.12c}$$

$$= \underline{A}^\dagger\,\underline{A}\,\underline{A}^\dagger\,\underline{\mu} \qquad \text{(as in 5.8c)} \tag{5.12d}$$

$$= \underline{A}^\dagger\,\underline{\mu} \qquad \text{(from basic property of } A^\dagger\text{)} \tag{5.12e}$$

as before

Thus the same solution is obtained by solving $\underline{A}\,\underline{s} = \underline{m}$ using $\underline{A}^\dagger$ or by solving (5.6) using $(\underline{A}^T\,\underline{A})^\dagger$. In terms of the singular value decomposition described in Section 4 we have

$$A_{ik} = \sum_{i'} \sum_{k'} U_{ii'} \Lambda_{i'k'} V_{kk'}$$

$$(\underline{A}^{\dagger})_{ki} = \sum_{i'} \sum_{k'} V_{kk'} (\Lambda^{-1})_{k'i'} U_{i'i}$$

and $$(\underline{A}^T \underline{A})^{\dagger}_{kk'} = \sum_{k''} \sum_{k'''} V_{kk''} (\underline{\Lambda}^{-2})_{k''k'''} V_{k'''k'} \qquad (5.13)$$

where $\underline{\Lambda}^{-2}$ is zero apart from the leading diagonal whose first r elements are λ_i^{-2} where the λ_i are the non-zero singular values of A and the non-zero eigenvalues of $\underline{A}^T \underline{A}$ are λ_i^2.

The two pseudo inverses $(\underline{A}^T \underline{A})^{\dagger}$ and $\underline{A}^{\dagger}$ have been used as solution matrices for the sets of equations (5.6) and (5.9) respectively.

The equations (5.5) with non-zero η would correspond to

$$(\underline{A}^T \underline{A} + \eta \underline{I})\, \underline{s} = \underline{m}^*$$

The matrix $\underline{A}^T \underline{A} + \eta \underline{I}$ will always have an inverse which can be written in terms of the singular value decomposition as

$$\left((\underline{A}^T \underline{A} + \eta \underline{I})^{-1}\right)_{kk'} = \sum_{k''} \sum_{k'''} V_{kk''} \phi_{k''k'''} V_{k'''k'} \qquad (5.14)$$

where $\phi_{k''k'''}$ is a KxK diagonal matrix with diagonal elements given by $1/(\lambda_i^2 + \eta)$ where only the first r λ_i are non-zero. For $\lambda_i >> \eta$ the contributions to the solution matrix (5.14) approximate those of the pseudo-inverse of $\underline{A}^T \underline{A}$ while for $\lambda_i \approx 0$ the contributions are $1/\eta$ rather than the 0 values used in 5.13.

Applying the solution matrix (5.14) to the right hand side of (5.5) it will be seen that in directions corresponding to eigenvectors with $\lambda_i >> \eta$ the right-hand side will be dominated by the $\underline{A}^T \underline{m}$ term and the solution matrix will approximate the pseudo-inverse. In these directions the solution will approximate the solution of (5.7). When $\lambda_i \approx 0$ the solution matrix differs from the pseudo inverse which leads to differences but since the right hand side will be dominated by the ηr_k term, the equations are approximately $\eta \underline{I}\, \underline{s} = \eta\, \underline{r}$ and so the solution in these directions is only weakly dependent on the $\underline{m}$. Thus as in the pseudo-inverse approach used in the seismic problem, the estimates of the $\underline{s}$ are based on the combinations of data corresponding to the large singular values of $\underline{A}$. The uniform treatment of all directions avoids consideration of distinct cases depending on the relative sizes of the subspaces (as is necessary in inversion formalisms such as that used by Bolin et al. 1983).

The solution matrix (5.14) is seen to be an approximation to the pseudo-inverse of $\underline{A}^T \underline{A}$. The same solution can also be obtained by adding additional rows to $\underline{A}$ in (5.9) so that the equations become (in block form)

$$\begin{pmatrix} \underline{A} \\ \cdots \\ \sqrt{\eta}\underline{I} \end{pmatrix} [\underline{s}] = \begin{pmatrix} \tilde{\underline{m}} \\ \cdots \\ \sqrt{\eta}\underline{r} \end{pmatrix} \tag{5.15}$$

$$= \underline{A}^*(\gamma)\ \underline{s}$$

Since $$(\underline{A}^*(\gamma))\ \ \underline{A}^*(\gamma)^T = \underline{A}\ \underline{A} + \gamma \underline{I}^T\ , \tag{5.16}$$

the singular values of $\underline{A}^*(\gamma)$ must be the square roots of the eigenvalues of the right hand side (i.e. $(\lambda_i^2 + \eta)^{\frac{1}{2}}$). Lawson and Hanson (1974, Section 4 of chapter 25) show this explicitly by applying Givens rotations to equation (5.15).

The correspondence between the solution matrix (5.14) that occurs in the calibration procedure outlined above and the pseudo-inverse used by Pavlis and Booker (1980) shows that we can legitimately use the distribution of eigenvalues as a measure of the extent to which a given set of data is giving information about the source parameters $s_{\alpha k}$.

The present calibration procedure does not attempt to construct an 'annulled data set' of the type used by Pavlis and Booker. The reasons for omitting this step are:

(i) Model predictions $y_{\alpha i}$ for any tracer will be obtained by a numerical integration of the full transport model. The amount of computation will be proportional to the number of tracers that will be essentially independent of the number of data points and so an annulled data set does not give any computational saving. The exception is the case in which a tracer can be completely ignored for the purposes of calibrating the transports because the source function is so poorly known that all data is used in attempting to determine the release rate.

(ii) The independence properties associated with the use of the annulled data set are desirable but will inevitably be lost in any iterative calibration of a non-linear system. There is no point in having a 'local' independence in the substeps when it does not lead to any 'global' independence property of the overall calibration.

(iii) Since the estimates of the $s_{\alpha k}$ and x_p are being 'biased' by the prior estimates $r_{\alpha k}$ and q_p it is reasonable to use any one particular observation (or combination of observations) to adjust both sets, $\{s_{\alpha k}\}$ and $\{x_p\}$ so as to obtain the best compromise between fitting the prior estimates and the observations.

In the substep in which the transport parameters x_p are estimated the distribution of eigenvalues can again be used to indicate the amount of information in the residual data. The most appropriate way of assessing the relative contributions of the different tracers appears to be repeating this step with each successive tracer removed. In special cases it may be appropriate to remove smaller data sets to determine the importance of a particular type of observation. It may also be of interest to determine the usefulness of proposed observational programs by generating possible data sets and calibrating the system using this hypothetical data. This approach has been used by Enting and Pearman (1984) in carbon cycle studies.

Apart from the parameter estimates ($\hat{x}_p$ and $\hat{s}_{\alpha k}$) it is also desirable to give indications of the uncertainties in these estimates and of the uncertainties in any quantities that may be calculated using the model.

Enting (1983) has described a computationally convenient technique for calculating the variance of any quantity Z that depends on the parameters of a model. The procedure is to minimise

$$\phi = \theta - 2\varepsilon Z \tag{5.17}$$

with respect to all the parameters. This gives a set of perturbed parameters, the amount of perturbation being proportional to ε, and Z is evaluated using these values to give a quantity $Z(\varepsilon)$. Enting (1983) shows that, to a linear approximation,

$$\text{Var}\ (Z) = (Z(\varepsilon) - Z(0)/\varepsilon \tag{5.18}$$

6. Implementation of the formalism

From the general formalism described in Section 5, three distinct classes of calculation can be identified:

(i) using a reference data set, calibrate the system, estimating $\{s_{\alpha k}\}$, $\{x_p\}$;

(ii) investigate perturbations about the reference calibration for purposes such as calculating variances using equation (5.18);

(iii) recalibrate the system with a slightly different data set and compare the results to the reference calibration. These calculations can be either for studies of the marginal utility of various data sets or for extending the reference data set as new information becomes available.

The implementation of these calculations can best be described by starting with the data structures that are used. Since the calculations above represent a set of inter-related computations that are likely to be performed at a variety of times, it is clear that the system must be built around a set of permanent files that preserve the information between runs. A schematic example is shown in Figure 6.1.

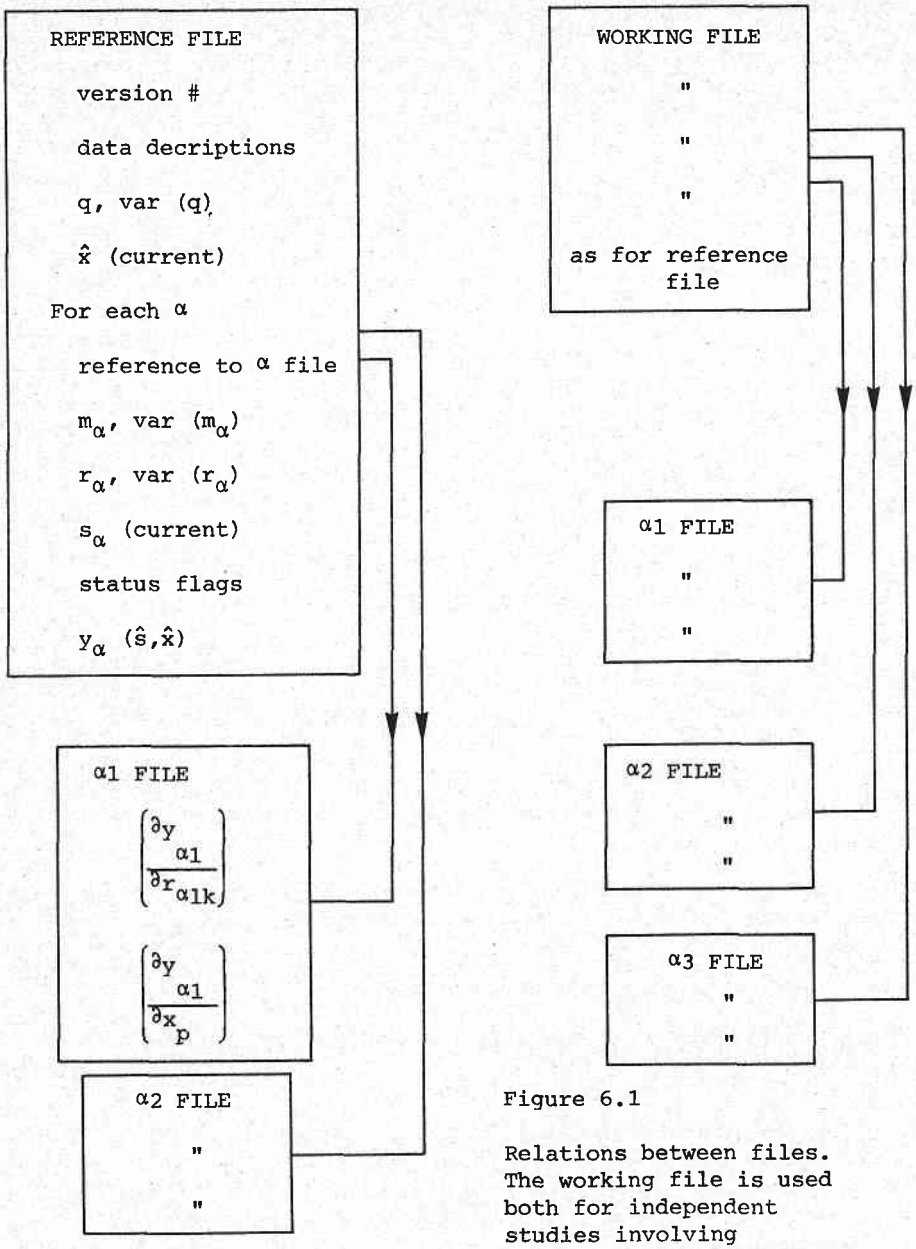

Figure 6.1

Relations between files. The working file is used both for independent studies involving perturbations about the reference and as a step in extending the data set (or parameter set) used as a reference.

The main features of the file set are:

(i) A reference file: This holds the inputs ($\underline{r}$, $\underline{q}$ and their variances) and outputs (the current $\underline{x}$, the current s and the current $\underline{y}(\underline{x})$). It also holds basic descriptors of the various quantities and pointers to other files (one for each tracer) containing the derivatives of $\underline{y}$.

The reference file holds status flags for each of these files indicating which of the rows of the derivative matrix have been calculated using the current $\underline{x}$ rather than a previous $\underline{x}$.

The file also includes a version number that is incremented each time $\underline{x}$ is changed and which serves to distinguish the reference file from the working file.

(ii) A working file: Used in studies of variations about the reference calibration. The structure is the same as for the reference file although the contents of the file will reflect the changes that are being investigated. For initial iteration steps or simple perturbation studies it may be possible to make use of derivatives calculated at the reference parameters, and the derivative files will be copied from those associated with the reference file.

(iii) Derivative files: There is one file for each tracer, holding $\left(\frac{\partial y_\alpha}{\partial s_\alpha}\right), \left(\frac{\partial y_\alpha}{\partial x}\right)$. When these files are first set up they are only skeletons - the rows (i.e. derivatives of all observations of tracer α, with respect to an x_p or an $s_{\alpha k}$) are written as they are calculated.

The various computations can be defined in terms of the changes that they induce in these files. The following list gives the main processes described in terms of these changes. The interrelations between these processes are shown in Figure 6.2.

(i) INITIALISE: Read inputs $\underline{r}$, $\underline{q}$, $\underline{m}$ and their variances and specifications of the tracers and their weight factors. Create reference file and skeletons of the derivative files.

(ii) MODIFY: Create a working file which is a copy of the reference file, subject to certain changes. Create skeleton derivative files and copy in as much information as possible from the previous derivative files. If the data set for a particular tracer is being extended then all rows in that tracer matrix will be recalculated.

(iii) MATRIX: Work through the status flags and recalculate any matrix rows that are unknown or only known for earlier parameter sets. This process is designed to be restartable. The status flags keep track of its progress. This process starts to be recalculating $\underline{y}_\alpha(\underline{x})$ if necessary.

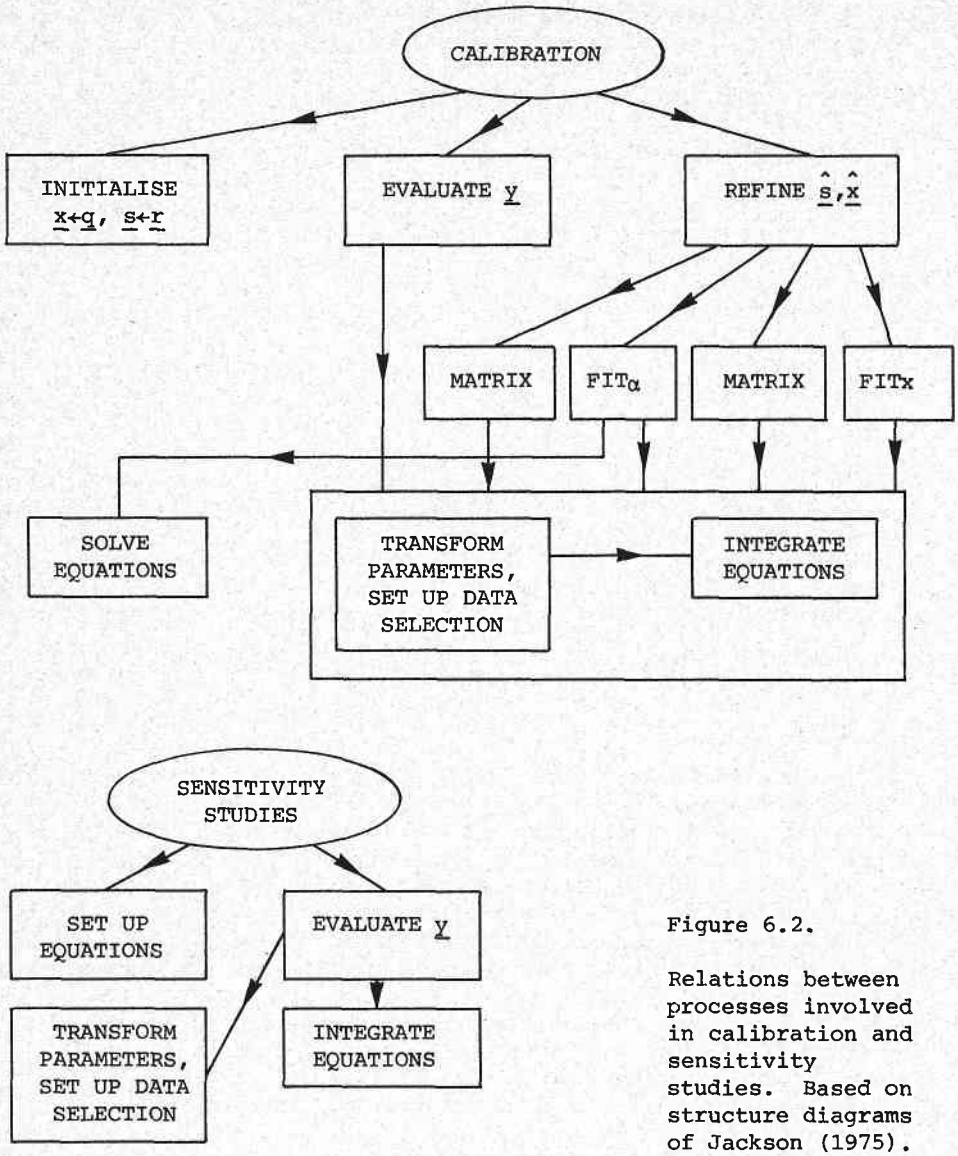

Figure 6.2.

Relations between processes involved in calibration and sensitivity studies. Based on structure diagrams of Jackson (1975).

(iv) FITα: Calculate an improved set of source parameters s_α and revise $\underline{y}_\alpha(\underline{x},\underline{s}_\alpha)$. Reset status flags for tracer α.

(v) FIT x: Calculate an improved set of transport parameters $\underline{x}$ and revise $\underline{y}_\alpha$ for all tracers and reset all status flags. This process should be made restartable if several tracers are involved.

(vi) SENSITIVITY: Calculate a revised set of s_α* and x* that minimises $\theta - \alpha Z$ for a specified Z. Evaluate $y_\alpha(\underline{s}^*_\alpha,\underline{x}^*)$ and tabulate the differences from $\underline{y}_\alpha(\underline{s},\underline{x})$.

For these main processes there are four main classes of subprocess.

(vii) FILE CREATION: Used by INITIALISE and MODIFY.

(viii) EQUATION SOLVING: Used by FITα, FITx and SENSITIVITY.

(ix) PARAMETER HANDLING: To convert the lists of parameters $\underline{s}_\alpha$ and $\underline{x}$ into the form required to integrate the model equations. Set up list to select $\underline{y}_\alpha$.

(x) INTEGRATION: Integrate the model equations for tracer α, storing the particular set of results, corresponding to the observations, as $\underline{y}_\alpha$.

7. Discussion

The procedures described in the preceding sections were designed mainly for calibrating two-dimensional atmospheric transport models. The analogy with the analysis of seismic data (Pavlis and Booker, 1980) provides a demonstration of the feasibility of splitting the calibration procedure into a series of subproblems associated with each tracer's sources and sinks followed by a subproblem involving the atmospheric transport parameters. This type of approach would be likely to have much wider applications in global biogeochemical studies; some possible applications are suggested here.

(i) Most two-dimensional atmospheric transport models use typical circulations for all years. With sufficient data it may be possible to look at interannual variability in tracer transport, applying the analysis above to the 'anomalies' in concentrations and transport rates.

(ii) It may be possible to apply similar techniques to the modelling of reactive atmospheric constituents. For example the distribution of OH radical is a common point of contact for a number of atmospheric chemical cycles that are otherwise loosely coupled and so the techniques described in this report may be applicable to modelling atmospheric chemistry as a set of subsystems that are nearly independent apart from their coupling to the OH radical distribution.

References

Bard, Y. (1974): Non-linear Parameter Estimation. Academic Press (New York), 341 pp.

Bauer, E. (1980): A study of stratosphere-to-troposphere transfer using radioactive tracer data in a one-dimensional parameterization. Institute for Defence Analysis paper P.1456. 84 pp.

Bolin, B., Björkström, A., Holmén, K. and Moore, B. (1983): The simultaneous use of tracers for ocean circulation studies. Tellus 35B: 206-236.

Box, G.E.P. and Tiao, G.C. (1973): Bayesian Inference in Statistical Analysis. Addison-Wesley (Reading, Mass.), 588 pp.

Enting, I.G. (1983) : Error analysis for parameter estimates from constrained inversion. CSIRO Division of Atmospheric Research, Technical Paper No.2, 18 pp.

Enting, I.G. and Pearman, G.I. (1983): Refinements to a one-dimensional carbon cycle model. CSIRO Division of Atmospheric Research, Technical Paper No.3, 35 pp.

Enting, I.G. and Pearman, G.I. (1984): The use of observations in calibrating and validating carbon cycle models. Presented at 6th Oak Ridge National Laboratory Life Sciences Symposium - Knoxville, November 1983. Proceedings: "The Global Carbon Cycle: Analysis of the Natural Cycle and Implications of Anthropogenic Alterations for the Next Century" - in press.

Hantel, M. and Hasse, S., (1983). Mass consistent heat budget of the zonal atmosphere. Bonner Meteorologische Abhandlungen. No. 29. 84 pp.

Hidalgo, H. and Crutzen, P.J. (1977): The tropospheric and stratospheric composition perturbed by NO_x emissions of high-altitude aircraft. J.Geophys.Res. 82: 5839-5866.

Jackson, D.D. (1972): Interpretation of inaccurate, insufficient and inconsistent data. Geophys.J.R.Astr.Soc. 28: 97-109.

Jackson, M.A. (1978): Principles of Program Design. Academic Press (London) 299 pp.

Lawson, C.L. and Hanson, B.J.: Solving Least Squares Problems. Prentice-Hall (Englewood Cliffs, N.J.). 339 pp.

Pavlis, G.L. and Booker, J.R. (1980): The mixed discrete - continuous inverse problem: Application to the simultaneous determination of earthquake hypocentres and velocity structure. J.Geophys.Res. 85 B9 : 4801-4810.

Plumb, R.A. (1979): Eddy fluxes of conserved quantitites by small-amplitude waves. J.Atmos.Sci. 36, 1699-1704.

Rodgers, C.D. (1977): Statistical principles of inversion theory. pp.117-134 of Inversion Methods in Atmospheric Remote Sounding. A. Deepak (ed), Academic Press (New York) 622 pp.

Twomey, S. (1977): Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements. Elsevier (Amsterdam) 243 pp.

Weiss, W., Sittkus, A., Stockburger, H. and Sartorius, H. (1983): Large-scale atmospheric mixing derived from meridional profiles of Krypton 85. J.Geophys.Res. 88C: 8574-8578.