Inversion methods for satellite sounding data
April 1991

By **J. R. Eyre**

Table of contents

1. BASIC IDEAS

1.1 The radiative transfer equation for emission to space

The monochromatic radiation intensity at frequency ν emitted along a vertical path at the top of the atmosphere and incident at a satellite-borne instrument is given by:

$$R_\nu = (I_0)_\nu \tau_\nu(z_0) + \int_{z_0}^{\infty} B_\nu\{T(z)\} \frac{\mathrm{d}\tau_\nu(z)}{\mathrm{d}z} \mathrm{d}z \ , \qquad (1)$$

where

$(I_0)_\nu$ is the emission from the earth's surface at height z_0,
$\tau_\nu(z)$ is the vertical transmittance from height z to space,
$T(z)$ is the vertical temperature profile,
and $B_\nu\{T(z)\}$ is the corresponding Planck function profile.

© ECMWF, 2002

Here we have neglected molecular scattering in and out of the beam - a good approximation in the infrared and microwave regions. We have also assumed, for the moment, that no cloud is present and we shall return to the cloud problem later. If the earth's surface reflects radiation significantly, then we acquire a third term in (1) representing radiation emitted downwards by the atmosphere and reflected back in the direction of the satellite. For simplicity we have ignored this term, which is equivalent to assuming that the surface is black (often a good approximation in the infrared). Reflection of solar radiation by the surface may usually be neglected also.

Equation (1) may also be used to represent radiation emitted along a slant (non-vertical) path if the transmittance is computed appropriately. Making the approximation of a plane-parallel atmosphere, for a viewing path through the atmosphere at angle θ to the vertical, then

$$\tau_\nu(z,\theta) = \exp\left\{-\sec\theta\int_z^\infty k_\nu(z')\, c(z')\, \rho(z')\, \mathrm{d}z\right\}, \tag{2a}$$

where $\rho(z)$, $c(z)$, $k_\nu(z)$ are respectively the vertical profiles of atmospheric density, absorbing gas mixing ratio and absorption coefficient.

It is often more convenient to choose pressure as the vertical coordinate. Then, using the hydrostatic approximation ($\mathrm{d}p = -g\rho\,\mathrm{d}z$), we obtain

$$\tau_\nu(p,\theta) = \exp\left\{-\sec\theta\int_0^p k_\nu(p')\, c(p')\, \mathrm{d}p'/g\right\}. \tag{2b}$$

1.3 Integration over frequency

Real satellite instruments sense radiation over a range of frequency rather than monochromatic radiation, and it is usually necessary to perform an integration over frequency to obtain radiances of adequate accuracy as “seen” by the satellite instrument, i.e.

$$R = \frac{\int R_\nu f_\nu \,\mathrm{d}\nu}{\int f_\nu \,\mathrm{d}\nu}, \tag{2}$$

where f_ν is the relative response of the instrument to radiation at frequency ν. This complicates the calculations involved in the interpretation of the data, but does not change the basic nature of the inversion problem. Therefore, for this discussion, we shall ignore it and work only with the monochromatic equations.

1.4 Weighting functions

Equation (1) may be written as

$$R_\nu = (I_0)_\nu \tau_\nu(z_0) + \int_{z_0}^\infty B_\nu\{T(z)\} K_\nu(z)\,\mathrm{d}z\,. \tag{3}$$

$K_\nu(z) = \mathrm{d}\tau_\nu(z)/\mathrm{d}z$ is called a WEIGHTING FUNCTION; it weights the Planck function in the atmospheric component of the emitted radiation. It specifies the layer from which the radiation emitted to space originates, and hence it determines the region of the atmosphere which can be sensed from space at this frequency. Fig. 1 shows the transmittance profiles and corresponding weighting functions at two frequencies for which the atmospheric absorption is different. Since the weighting function is the derivative of the transmittance profile, it

© ECMWF, 2002

will peak higher in the atmosphere for the frequency at which the absorption is stronger. In this way, a carefully selected family of frequencies can be chosen to sense radiation from different layers in the atmosphere.

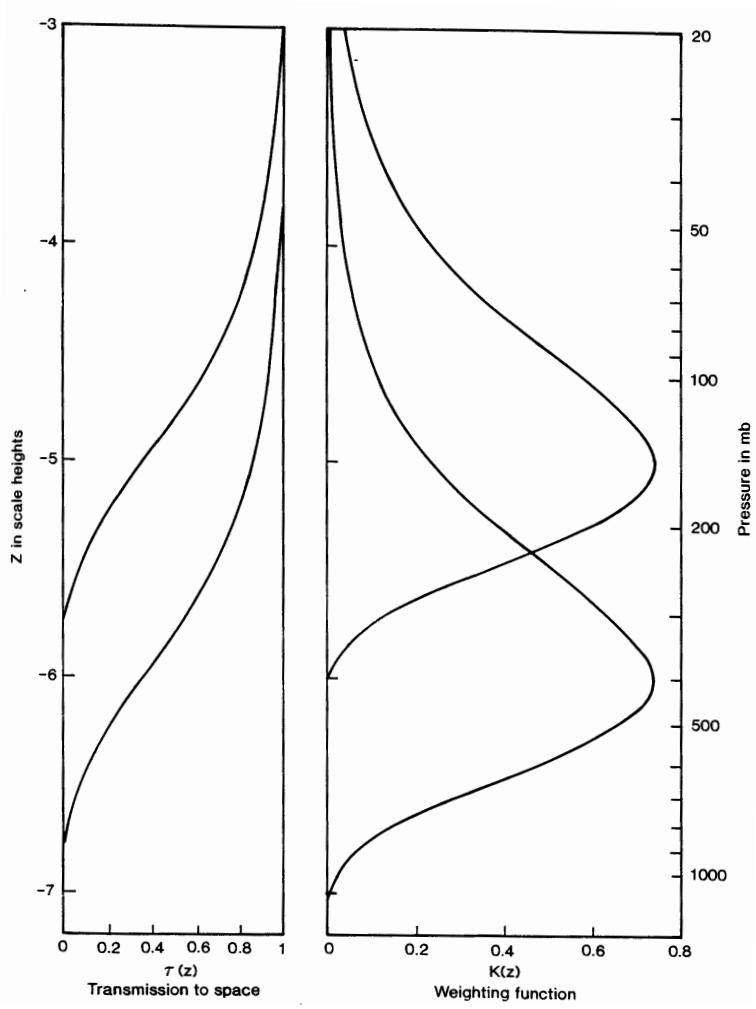

Figure 1. Idealised transmittance profiles and weighting functions at two frequencies with different absorption coefficients. Vertical coordinate: scale height = $-\log_e$(pressure).

To understand qualitatively why the weighting functions take this form, we can consider the emission to space from air parcels of unit volume at different heights in the atmosphere. The radiation emitted to space is determined by three factors:

(a) the temperature of the air parcel, i.e. the variable we hope to measure,

(b) the number of molecules of emitting gas, which is determined by the atmospheric density (and also by the mixing ratio of the absorbing constituent, although for the principal gases used in temperature sounding—carbon dioxide and oxygen—the mixing ratio can be assumed constant and known),

(c) the transmittance of the atmosphere from the air parcel to space.

This is illustrated in Fig. 2 for three air parcels at different heights. For the lowest parcel, the atmospheric density is high and so the amount of radiation emitted is high, but almost all is absorbed by the atmosphere above it and very little reaches space. For the highest parcel, the transmittance to space is high, but comparatively little radiation

© ECMWF, 2002

is emitted because atmospheric density decreases exponentially with height. These two conflicting effects combine in such a way that, at some intermediate height, the contribution of a parcel to the radiation reaching space is a maximum. The variation of the radiance to space as a function of height is shown by the curve on the right of Fig. 2 . Most of the radiation to space originates in a layer around the peak of this function (which is actually the product of the weighting function and the Planck function profile, i.e. the integrand in equation (3)). From knowledge of the atmosphere's composition and spectroscopic parameters we can calculate where in the atmosphere this layer will be. Then the intensity of the radiation can be interpreted in terms of the mean temperature of the layer. Using radiation at different frequencies for which the absorption strength is different, we can build a family of weighting functions, which provide information on the mean temperatures of many such layers, thus leading to the idea that we might be able to RETRIEVE information on the atmospheric temperature profile from a set of multi-frequency measurements.[1]

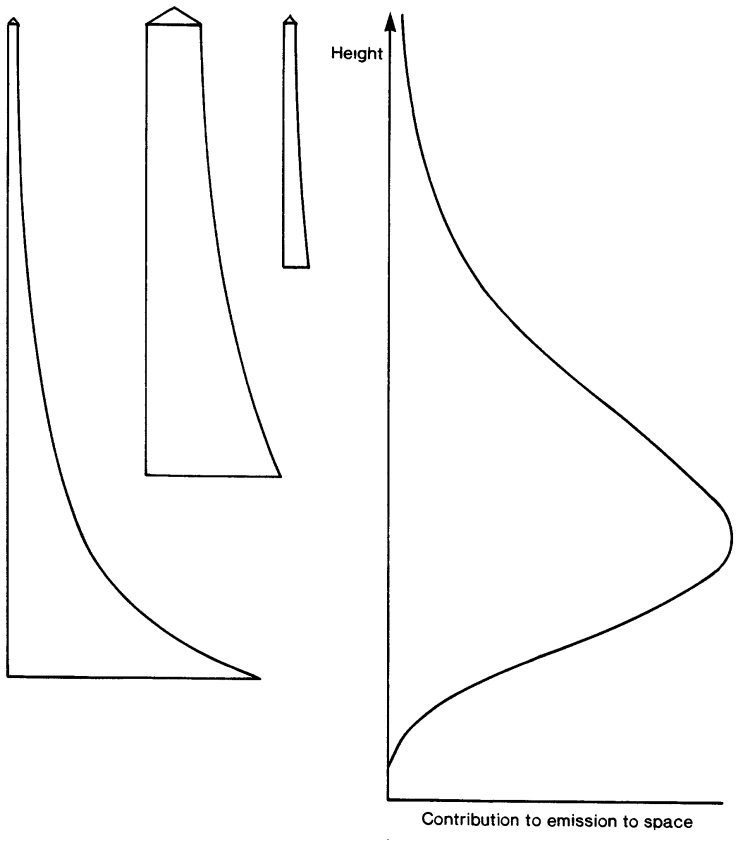

Figure 2. Left—illustrating the attenuation of upwelling radiation emitted from three heights in the atmosphere. Right—the corresponding vertical profile of the contribution to the emission to space.

1.5 Characteristics of weighting functions

At this point, two aspects of the problem are worthy of note. Firstly, the weighting functions are broad (i.e. several

1. We note in passing that ground-based measurements of downwelling atmospheric radiation do not have associated with them weighting functions of the same form. Here, the atmospheric density and transmittance from air parcel to instrument both decrease with height, and so (for an absorber mixing ratio which is constant with height) the largest contribution to the measured radiance is always from close to the instrument, whatever the frequency.

© ECMWF, 2002

kilometres). This means that the satellite instrument can sense the mean properties of broad layers very well, but it is only able to sense the characteristics of single levels or narrow layers insofar as they are correlated with the properties of the broad layers. The width of the weighting functions severely limits the capability of satellite sounders to detect atmospheric structure which has relatively small scale in the vertical. The finite width of the weighting functions is a fundamental feature of passive remote sensing techniques. However, the precise width is determined by technological considerations, as explained below.

Secondly, for most instruments, the family of weighting functions are highly overlapping. One consequence of this is that, although the instrument may make measurements at N separate frequencies, we obtain fewer than N pieces of independent information. The implications of this in the inversion problem are discussed below.

© ECMWF, 2002

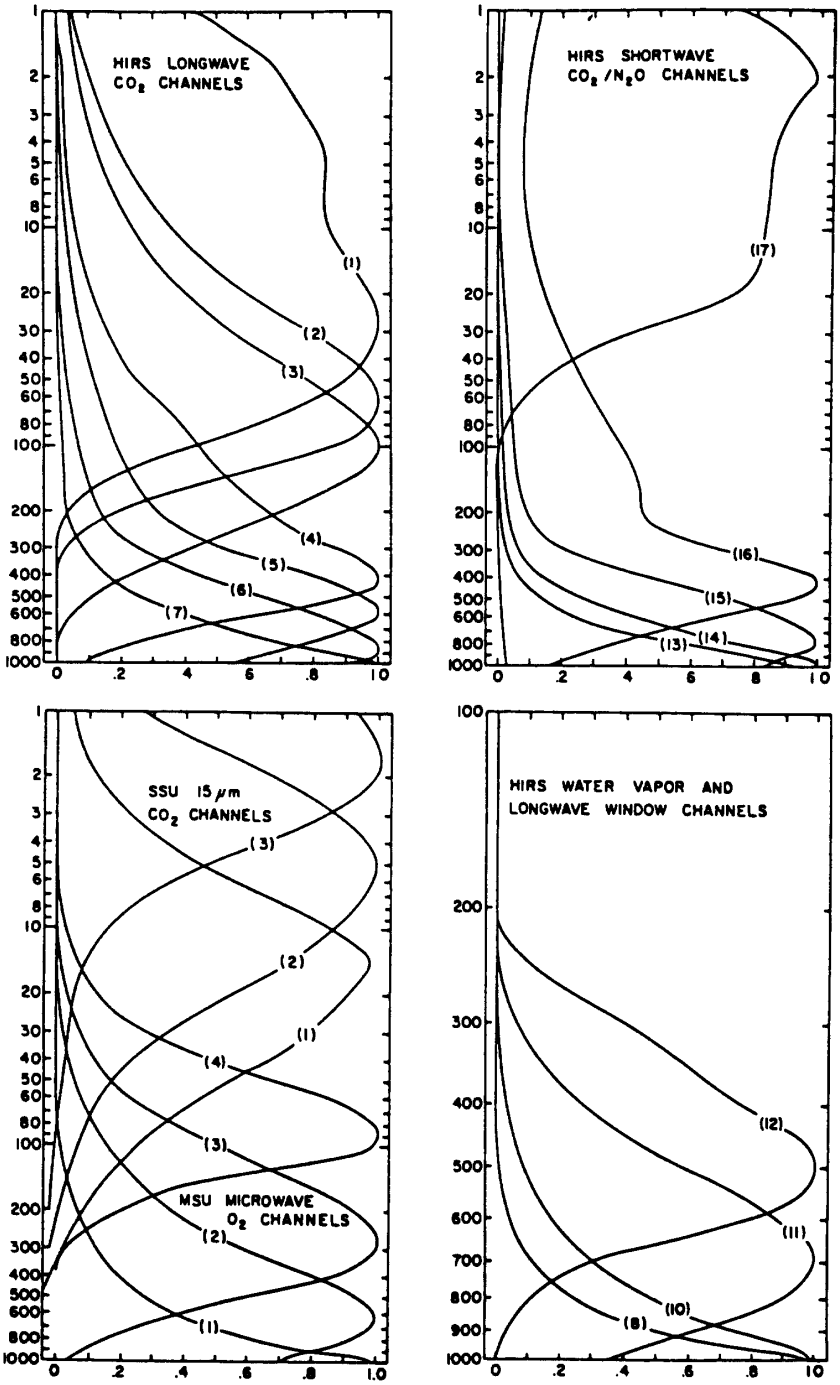

Figure 3. TOVS normalised weighting functions (from Smith *et al.* 1979).

Figure 3 illustrates the weighting functions for the present operational sounding system—the TIROS-N Operational Vertical Sounder (TOVS) which consists of 3 instruments: the High-resolution Infrared Radiation Sounder (HIRS/2), the Microwave Sounding Unit (MSU) and the Stratospheric Sounding Unit (SSU). For further information on TOVS, see Smith *et al.* (1979) or Schwalb (1978).

For the microwave channels, the spectral responses at the individual measurement frequencies (usually called "channels") are much narrower than the widths of the atmospheric absorption lines. Therefore the weighting func-

© *ECMWF, 2002*

tions are close to their monochromatic limit. However, it is possible to improve the vertical resolution of the microwave sounder by adding more channels. This will be done for the Advanced Microwave Sounding Unit (AMSU) which, along with HIRS, will constitute the Advanced TOVS on the next generation of polar orbiting satellites (from about 1994). The weighting functions for AMSU are illustrated in Fig. 4 .

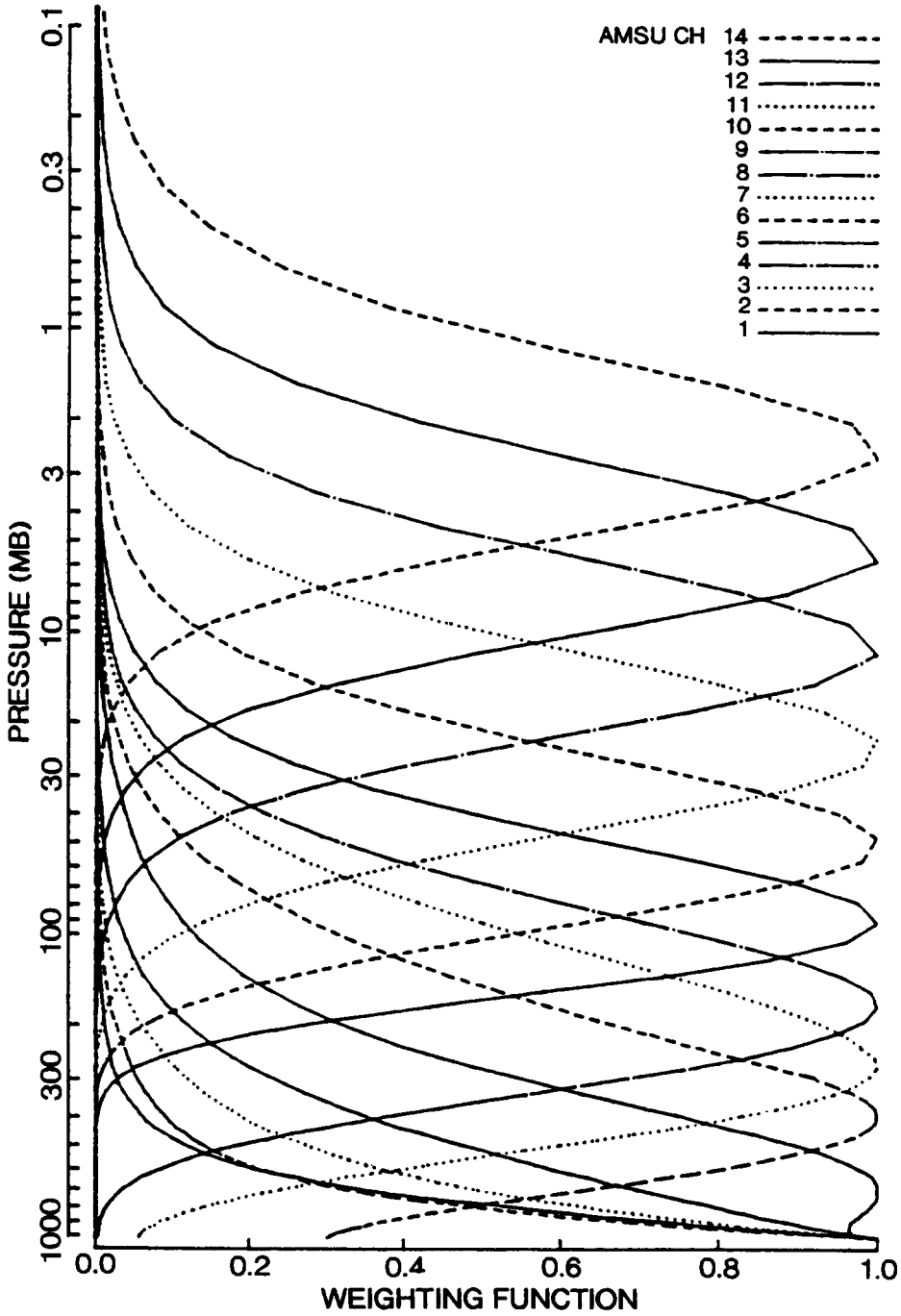

Figure 4. AMSU normalised weighting functions.

For the infrared channels the position is very different: HIRS is a filter radiometer and its channels have spectral widths typically hundreds of times greater than the atmospheric absorption lines. Therefore they average together

© ECMWF, 2002

frequencies for which the absorption strengths are very different. This has the effect of broadening the weighting functions considerably. By using instruments of much higher spectral resolution, such as interferometers or grating spectrometers, it is possible to achieve spectral resolutions closer to the widths of the atmospheric absorption lines. In this way instruments with several thousand channels and much sharper weighting functions can be built. Fig. 5 illustrates the weighting functions from such an instrument. Similar instruments are planned for satellite missions in the late 1990s.

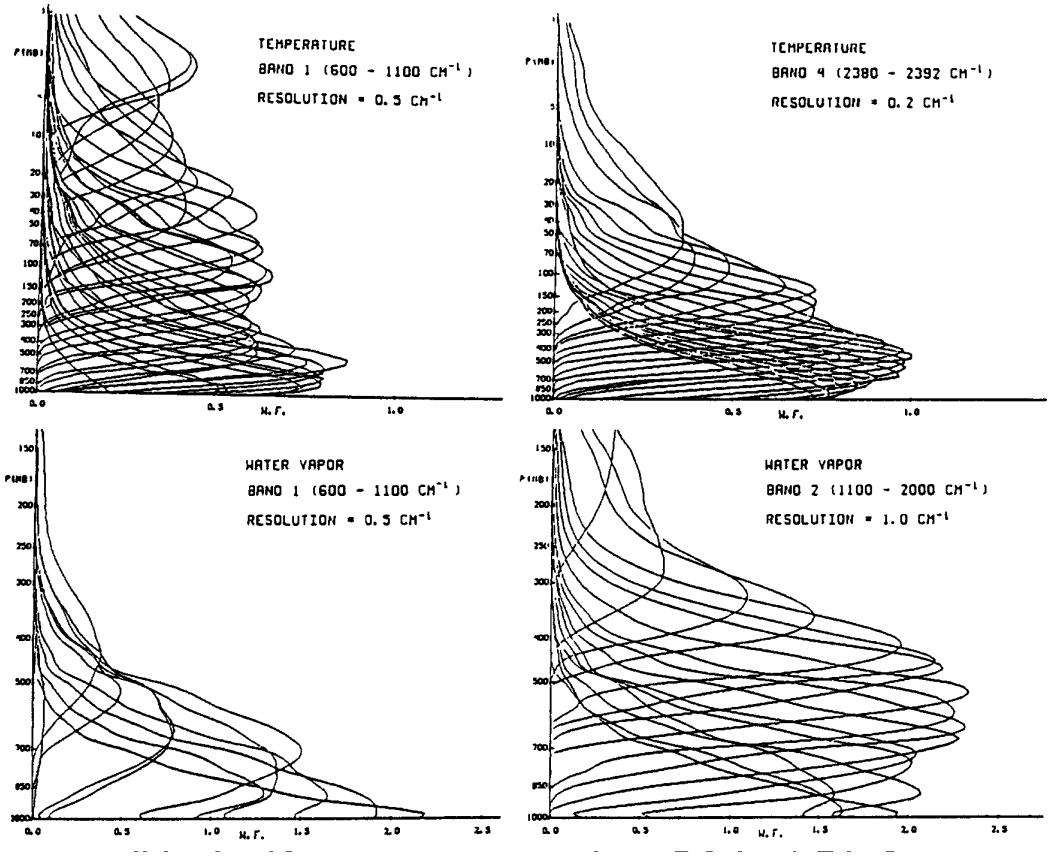

Figure 5. A selection of weighting functions for several bands of the interferometer proposed for the Atmospheric InfraRed Sounder (AIRS) instrument.

1.6 The forward and inverse problems

The instrument makes measurements of radiance in a number of channels ν_i. For each channel, we can write a radiative transfer equation:

$$R_i = (I_0)_i \tau_i(z_0) + \int_{z_0}^{\infty} B_i\{T(z)\} K_i(z)\, \mathrm{d}z \,. \tag{4}$$

This equation expresses the FORWARD PROBLEM for the channel, i.e. given the state of the atmosphere, the solution of this equation tells us the radiance incident at the satellite in this channel. However, when presented with satellite measurements, we are faced with the INVERSE PROBLEM: given the measurements, what is the state of the atmosphere (in terms of its vertical profiles of temperature and constituents). Let us concentrate on the temperature profile inversion problem and return to the constituent problem later.

© ECMWF, 2002

Since we have a limited number of channels ($i = 1 \rightarrow I$), we can see immediately that the inversion of equation (4) is ILL-POSED or UNDER-CONSTRAINED. This is because we are trying to retrieve $T(z)$, a continuous function of height (which, in general, requires an infinite number of parameters to represent it fully), from a finite number of measurements. This means that there exists an infinite number of profiles $T(z)$ which satisfy the measurements. Our problem is to find one which is reasonable and, if possible, to find the profile which is best or most reasonable in some sense.

In addition, the measurements always contain some error or "noise". This further increases the ill-posed nature of the problem, and we must find a method of solution which does not amplify the noise to an unacceptable degree.

1.7 A vector–matrix representation

At this point, it is convenient to change from the notation of continuous profiles and integrals, as in equation (4), to discrete profiles and the notation of vectors and matrices. We consider the atmosphere to be composed of many thin layers (numbered, from the top, $j = 1 \rightarrow (J-1)$) with mean temperature T_j and Planck function $B_{ij} = B_i\{T_j\}$. Let the transmittance from the bottom of layer j to space be denoted τ_{ij}. Then equation (4) becomes

$$R_i = (I_0)_i \tau_i(z_0) + \sum_{j=1}^{J-1} B_{ij}(\tau_{ij-1} - \tau_{ij})$$

or (5)

$$R_i = (I_0)_i \tau_i(z_0) + \sum_{j=1}^{J-1} B_{ij} K_{ij}.$$

To solve this equation for B, it is convenient to find a transformation of B which is independent of i (i.e. of frequency). For channels which are very close together in frequency, we can use the Planck function at a central frequency. However, this is rarely an adequately accurate approximation. One solution is to specify a reference frequency for the Planck function and "adjust" all the measured radiances to it. Alternatively, we can convert radiances to some other quantity, such as equivalent black body temperature, which is independent of frequency. These are technical details on which we need not dwell; it is only necessary to appreciate that it is relatively straightforward to find a channel-independent form of B so that we may write

$$R_i = (I_0)_i \tau_i(z_0) + \sum_{j=1}^{J-1} B_j K_{ij}. \quad (6)$$

We can also "absorb" the surface term as the J th term in the summation by setting $B_J = I_0$ and $K_{ij} = \tau_i(z_0)$. Then

$$R_i = \sum_{j=1}^{J} B_j K_{ij}. \quad (1a)$$

If we now represent the radiance in all channels by a vector R and the Planck function profile by a vector B, equation (6) may be written for all channels simultaneously as

$$\mathbf{R} = \mathbf{K} \cdot \mathbf{B}. \quad (7)$$

© ECMWF, 2002

$\mathbf{K}$ is a matrix containing discrete weighting function elements K_{ij}. Thus, our measurements are a vector $\mathbf{R}$ (elements R_i, $i = 1 \rightarrow I$), our unknowns are a vector $\mathbf{B}$ (elements B_j, $j = 1 \rightarrow J$), and $\mathbf{K}$ is a matrix of size $I \times J$. Our problem is to invert equation (7) to find $\mathbf{B}$. Then the temperature profile is obtained directly as a known function of $\mathbf{B}$.

1.8 Linearity

It is usually important to appreciate the degree of LINEARITY of any given inverse problem. By this we mean the degree to which we can separate out the knowns and unknowns of the problem into a linear equation. For example, in the case of equation (7), it represents a linear problem if $\mathbf{K}$ is independent of $\mathbf{B}$. If $\mathbf{K}$ is a function of $\mathbf{B}$, we have a nonlinear problem. In the temperature retrieval problem, $\mathbf{K}$ consists of differences between transmittances to space from the top and bottom of the layer (see equation (5)). The transmittances, when expressed in pressure co-ordinates (see equation (2b)), are strong functions of the mixing ratio and pressure of the absorbing gas and its spectroscopic parameters, but the latter are only weakly temperature dependent. Therefore, the problem is almost linear. This means that we can calculate a reasonable approximation to $\mathbf{K}$ without accurate prior knowledge of the unknown $\mathbf{B}$. It also means that the weighting functions for a given channel are almost independent of the atmospheric conditions.

The near linear nature of the temperature inversion problem has allowed the development of appropriate inversion methods based on linear theory. Nevertheless, the nonlinearities are significant and must be considered carefully when accurate results are required.

An excellent discussion of linear inversion theory applicable to a wide range of geophysical problems is given by Menke (1984).

2. TEMPERATURE PROFILE INVERSION METHODS

2.1 Some simple solutions and their problems

We noted above that the problem of retrieving a continuous profile from a finite set of measurements (inverting equation (4)) is ill-posed. With the discrete formulation of equation (7), if $J > I$, then the problem is still ill-posed because the number of unknowns exceeds the number of simultaneous equations represented by this single matrix equation. We can make the problem WELL-POSED by reducing the number of layers over which the profile is specified or by expanding the profile in terms of the coefficients of a restricted set of BASIS FUNCTIONS:

$$\mathbf{B} = \Phi \cdot \mathbf{b}, \tag{8}$$

where Φ is a matrix of basis functions (size $J \times K$) and $\mathbf{b}$ is vector of coefficients (length K). Then

$$\mathbf{R} = \mathbf{K} \cdot \Phi \cdot \mathbf{b} = \mathbf{A} \cdot \mathbf{b}, \tag{9}$$

where $\mathbf{A}$ has dimensions $I \times K$. If $K = I$, then $\mathbf{A}$ is square and equation (9) may be inverted directly:

$$\mathbf{b} = \mathbf{A}^{-1} \cdot \mathbf{R}, \tag{10}$$

where $^{-1}$ denotes matrix inverse. Substitution into equation (8) then gives the solution for $\mathbf{B}$. However, this solution is found to be unsatisfactory (see Rodgers, 1976), because the problem is usually ILL-CONDITIONED. By this, we mean that the elements of $\mathbf{A}^{-1}$ tend to have large magnitudes (positive and negative) leading to an

© ECMWF, 2002

amplification of small errors in $\mathbf{R}$ into large errors in $\mathbf{B}$. This arises because the weighting functions are broad and overlapping—we do not have I independent pieces of information. We have found one of the family of profiles which are mathematically consistent with the measurements, but usually one which is far from the true profile.

It is possible to improve on this by restricting the number of basis functions to $K < I$. Now we have an OVER-CONSTRAINED problem—the number of equations exceeds the number of unknowns and we may seek a solution which is a least-squares fit to the measurements, i.e. we minimize

$$\sum_{i=1}^{I} \left(R_i - \sum_{k=1}^{K} A_{ij} b \right)^2 \tag{11}$$

with respect to all the elements of $\mathbf{A}$. The solution is then

$$\mathbf{b} = (\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A})^{-1} \cdot \mathbf{A}^{\mathrm{T}} \cdot \mathbf{R}, \tag{12}$$

where $^{\mathrm{T}}$ denotes matrix transpose.

This least-squares solution tends to be better than the exact solution represented by equation (10) but is still usually unsatisfactory. For a successful method, the basis functions must be carefully chosen and well suited to the types of profiles which are found in the atmosphere. For example, simple functions of height such as polynomials are not suitable. Also, it is often found that the solution over-fits the measurements (see Rodgers, 1976, for more discussion).

2.2 The estimation problem and the role of constraints

From our appreciation of the ill-posed nature of the problem and experience with simple solutions, we are forced to the conclusion that we need information additional to the measurements in order to CONSTRAIN the profile and to choose a reasonable profile from the infinite number of mathematically possible profiles. Fortunately, for problems of interest in atmospheric remote sensing, additional information is available.

In seeking a method of solution, we accept from the start that we cannot find the “true” profile exactly - the ill-posed nature of the problem and the noise in the measurements preclude this. We must look instead for an ESTIMATE of the true profile which is acceptably accurate or the best estimate in some statistical sense.

We may look towards probability theory or statistical techniques to tell us how to combine our radiation measurements with other information in order to select from all the possible profiles the best one. In this case, it is the other (“prior” or “background”) information which provides the constraints on the solution. There are many of these methods in the literature. They are all interrelated, and some are examined below.

Alternatively, we may take an empirical approach and look for an *ad hoc* method which finds a solution to the problem—one of many, but one which is found through experience to be acceptable. There are a number of such methods in the literature and we shall look at one in section 2.3 (e). These methods do not address the estimation problem directly. They use a variety of constraints which are not always obvious. These can take the form of limits on the smoothness of the profile or the requirement that it is composed of a linear combination of some basis functions.

It is interesting to note that all the characteristics of the satellite sounding inversion problem are also characteristics of the data analysis problem for numerical weather prediction (NWP). The two are mathematically equivalent; in general they are both estimation problems which are ill-posed without the use of prior constraints.

© ECMWF, 2002

2.3 Practical solutions

2.3 (a) The maximum probability solution. It is often useful to think of our knowledge of a variable in terms of a probability density function (PDF): let $P(x)$ express the probability that a scalar variable has a value x. If we know the estimates of x have mean value x_0 and errors which, statistically, are normally distributed (Gaussian) with standard derivation σ, then we can say that x has a probability described by a PDF,

$$P(x) \alpha \exp\left[-\frac{1}{2}\frac{(x-x_0)^2}{\sigma^2}\right]. \tag{13}$$

When considering a vector variable $\mathbf{x}$, the equivalent equation is:

$$P(\mathbf{x}) \alpha \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^{\mathrm{T}} \cdot \mathbf{B}^{-1} \cdot (\mathbf{x}-\mathbf{x}_0)\right], \tag{14}$$

where $\mathbf{B}$ is the error covariance about the mean value $\mathbf{x}_0$.

A very powerful tool in probability theory is BAYES THEOREM, which helps us to manipulate the probabilities of two states or events occurring together. Let us consider two states, x and y. The probability that x and y will occur *together* is called their JOINT PROBABILITY and is denoted $P(x, y)$. The probability that x will occur when y occurs is called the CONDITIONAL PROBABILITY of x given y, and it is denoted $P(x|y)$. It is evident that the two are related by

$$P(x, y) = P(x|y)P(y), \tag{15}$$

i.e. that the probability of x and y occurring together is the probability that x occurs when y occurs multiplied by the probability that y occurs.

We can also interchange x and y:

$$P(x, y) = P(y|x)P(x). \tag{16}$$

Combining (15) and (16):

$$P(x|y) = P(y|x)P(x) / P(y). \tag{17}$$

This is Bayes theorem.

How is this theory relevant to our problem? Let $\mathbf{x}$ represent some aspect of the atmospheric state such as the vertical temperature profile (or perhaps some function of it such as the Planck function profile, or a vector containing information on the temperature profile and other atmospheric variables—in this way the theory we develop is more generally applicable than that introduced in Section 1). Let $\mathbf{y}^{\mathrm{m}}$ be a vector of measurements, such as satellite sounding data expressed as radiances, brightness temperatures, etc. Our purpose is to find the most probable value of $\mathbf{x}$ given the measurements $\mathbf{y}^{\mathrm{m}}$, i.e. to maximize the conditional probability of $\mathbf{x}$ given $\mathbf{y}^{\mathrm{m}}$:

$$P(\mathbf{x}|\mathbf{y}^{\mathrm{m}}) = \text{maximum}. \tag{18}$$

We apply Bayes theorem in the form,

© ECMWF, 2002

$$P(\mathbf{x}|\mathbf{y}^m) \propto P(\mathbf{y}^m|\mathbf{x})P(\mathbf{x}) . \tag{19}$$

Here we have taken $P(\mathbf{y}^m)$, the prior probability of making a measurement $\mathbf{y}^m$ to be constant (over the range of values allowed by the instrument).

$P(\mathbf{y}^m|\mathbf{x})$ is the probability that we shall make a measurement $\mathbf{y}^m$ when the atmospheric state is $\mathbf{x}$. Let us represent the forward problem by the general expression $\mathbf{y}\{\mathbf{x}\}$.[2] If the measurements were error free, then $P(\mathbf{y}^m|\mathbf{x})$ would be a delta function peaking at $\mathbf{y}^m = \mathbf{y}\{\mathbf{x}\}$. However, the measurements will contain errors which we shall assume are Gaussian with covariance $\mathbf{Y}$. Then the PDF becomes (cf. equation (14))

$$P(\mathbf{y}^m|\mathbf{x}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}^m - \mathbf{y}\{\mathbf{x}\})^T \cdot \mathbf{Y}^{-1} \cdot (\mathbf{y}^m - \mathbf{y}\{\mathbf{x}\})\right]. \tag{20}$$

$P(\mathbf{x})$ contains our information on $\mathbf{x}$ prior to making any measurement. This information may come from a number of sources. For example, we may use a forecast profile from a numerical weather prediction model (along with some estimate of its probable error characteristics) or we may have climatological information such as the climatological mean profile and its covariance about the mean. We shall call such data BACKGROUND INFORMATION and denote it by a background profile $\mathbf{x}^b$ and its error covariance $\mathbf{B}$. Then, for normally distributed background errors, the prior probability of the profile having a value $\mathbf{x}$ is given by (cf. equation (14)):

$$P(\mathbf{x}) \propto \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \cdot \mathbf{B}^{-1} \cdot (\mathbf{x} - \mathbf{x}^b)\right]. \tag{21}$$

It is more convenient to maximize the logarithm of (19) rather than (19) itself; substituting from (20) and (21) and taking the natural log gives

$$\ln\{P(\mathbf{x}|\mathbf{y}^m)\} = -\frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \cdot \mathbf{B}^{-1} \cdot (\mathbf{x} - \mathbf{x}^b) - \frac{1}{2}(\mathbf{y}^m - \mathbf{y}\{\mathbf{x}\})^T \cdot \mathbf{Y}^{-1} \cdot (\mathbf{y}^m - \mathbf{y}\{\mathbf{x}\}) + \text{constant} . \tag{22}$$

2.3 (b) The inversion problem as a variational problem. It is useful to identify the scalar quantity, $-\ln\{P(\mathbf{x}|\mathbf{y}^m)\}$ + constant, with a COST FUNCTION which has to be minimized:

$$J(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \cdot \mathbf{B}^{-1} \cdot (\mathbf{x} - \mathbf{x}^b) - \frac{1}{2}(\mathbf{y}^m - \mathbf{y}\{\mathbf{x}\})^T \cdot \mathbf{Y}^{-1} \cdot (\mathbf{y}^m - \mathbf{y}\{\mathbf{x}\}) . \tag{2a}$$

Now we see that the mathematical problem is equivalent to that which arises in variational NWP data assimilation (e.g. see Lorenc 1988). The quadratic nature of the cost function arises because we have assumed a Gaussian form for the measurement and background errors.

In the extensive literature on satellite sounding inversions, it is unusual to find the problem addressed in this way, i.e. as a variational problem. However, it is helpful to do so, as it makes clear the link with literature on similar variational problems, including that of NWP data assimilation. Moreover, once the general (nonlinear) inversion problem is posed in this way, most of the solutions which appear in literature, some of which are outlined below, can be seen as various levels of approximation to the optimal solution.

The optimal solution, i.e. the most probable profile, is found by minimizing equation (2a) or by solving its gradient

2. Note that equations (4)–(7) are particular (simplified) ways of expressing the forward problem; here we retain a completely general form.

© ECMWF, 2002

equation: if $\mathbf{J}'(\mathbf{x})$ represents the gradient of $J(\mathbf{x})$ with respect to $\mathbf{x}$, then

$$\mathbf{J}'(\mathbf{x}) = \mathbf{B}^{-1} \cdot (\mathbf{x} - \mathbf{x}^{\mathrm{b}}) - \mathbf{K}(\mathbf{x})^{\mathrm{T}} \cdot \mathbf{Y} \cdot (\mathbf{y}^{\mathrm{m}} - \mathbf{y}\{\mathbf{x}\}) = 0\,, \tag{23}$$

where $\mathbf{K}(\mathbf{x})$ is a matrix containing the derivatives, $\mathrm{d}\mathbf{y}\{\mathbf{x}\}/\mathrm{d}\mathbf{x}$. [3]

In general, this problem is not trivial—there is no general analytic solution to this equation. However, many techniques exist for solving it numerically for problems of interest. Here we shall consider only the linear case, i.e. when

$$\mathbf{K}(\mathbf{x}) = \mathbf{K}(\mathbf{x}^{\mathrm{b}}) = \mathbf{K}, \text{ a constant}\,. \tag{24}$$

Integrating $\mathrm{d}\mathbf{y}\{\mathbf{x}\}/\mathrm{d}\mathbf{x} = \mathbf{K}$ gives

$$\mathbf{y}\{\mathbf{x}\} = \mathbf{y}\{\mathbf{x}^{\mathrm{b}}\} + \mathbf{K} \cdot (\mathbf{x} - \mathbf{x}^{\mathrm{b}})\,. \tag{25}$$

Substituting from (25) and (24) into (23) gives

$$\mathbf{x} = \mathbf{x}^{\mathrm{b}} + (\mathbf{B}^{-1} + \mathbf{K}^{\mathrm{T}} \cdot \mathbf{Y}^{-1} \cdot \mathbf{K})^{-1} \cdot \mathbf{K}^{\mathrm{T}} \cdot \mathbf{Y}^{-1} \cdot (\mathbf{y}^{\mathrm{m}} - \mathbf{y}\{\mathbf{x}^{\mathrm{b}}\})\,. \tag{26}$$

Matrix manipulation yields an equivalent equation which is often computationally more efficient (i.e. it involves the inversion of smaller matrices):

$$\mathbf{x} = \mathbf{x}^{\mathrm{b}} + \mathbf{B} \cdot \mathbf{K}^{\mathrm{T}} \cdot (\mathbf{K} \cdot \mathbf{B} \cdot \mathbf{K}^{\mathrm{T}} + \mathbf{Y})^{-1} \cdot (\mathbf{y}^{\mathrm{m}} - \mathbf{y}\{\mathbf{x}^{\mathrm{b}}\})\,. \tag{27}$$

Our inversion therefore takes the following form: we have a background or "first-guess" profile $\mathbf{x}^{\mathrm{b}}$ from which we calculate the corresponding radiances $\mathbf{y}\{\mathbf{x}^{\mathrm{b}}\}$ using a radiative transfer model. We also calculate its derivatives $\mathbf{K}$ (i.e. the weighting functions). With estimates of the error covariance of measurements, $\mathbf{Y}$, and of the background profile, $\mathbf{B}$, we can solve equation (27) for $\mathbf{x}$ using measurements $\mathbf{y}^{\mathrm{m}}$.

It is useful to note that (27) has the form:

$$\mathbf{x} - \mathbf{x}^{\mathrm{b}} = \mathbf{W} \cdot (\mathbf{y}^{\mathrm{m}} - \mathbf{y}\{\mathbf{x}^{\mathrm{b}}\})\,. \tag{28}$$

In the case described here $W = \mathbf{B} \cdot \mathbf{K}^{\mathrm{T}} \cdot (\mathbf{K} \cdot \mathbf{B} \cdot \mathbf{K}^{\mathrm{T}} + \mathbf{Y})^{-1}$, but the same general form of equation with different expressions for $\mathbf{W}$ appears in most inversion schemes. It expresses how differences between measured radiances and calculated radiances (i.e. calculated from the background profile) "map" into differences between retrieved profile and background profile through the INVERSE MATRIX, $\mathbf{W}$. Note that the optimal form of $\mathbf{W}$ above is identical to that obtained in the optimal interpolation technique for NWP data assimilation (see Lorenc 1981), showing again the close correspondence between the two problems.

2.3 (c) The minimum variance solution. Another statistical approach to the estimation problem is to seek the solution which minimizes the mean-square difference between the retrieved and true profiles when averaged over a large number of cases. Let us assume that the forward problem is linear, i.e. that the true radiances $\mathbf{y}\{\mathbf{x}^{\mathrm{t}}\}$ may

3. We use the notation $\mathbf{K}$ because, for the linear temperature retrieval problem, the derivatives of the radiances with respect to the profile elements are the weighting functions, cf. equation (7).

© *ECMWF, 2002*

be calculated from the true profile $\mathbf{x}^t$ by (cf. equation (25)):

$$\mathbf{y}\{\mathbf{x}^t\} = \mathbf{y}\{\mathbf{x}^b\} + \mathbf{K}\cdot(\mathbf{x}^t - \mathbf{x}^b)\,. \tag{29}$$

The measured radiances are the true radiances plus measurement error ε^m:

$$\mathbf{y}^m = \mathbf{y}\{\mathbf{x}^t\} + \varepsilon^m\,, \tag{30}$$

and so

$$\mathbf{y}^m - \mathbf{y}\{\mathbf{x}^b\} = \mathbf{K}\cdot(\mathbf{x}^t - \mathbf{x}^b) + \varepsilon^m\,. \tag{31}$$

We minimize the mean-square departure of $\mathbf{x}$ from $\mathbf{x}^t$ over a large number of cases N:

$$\left(\frac{1}{N}\right)\sum^{N}(\mathbf{x}-\mathbf{x}^t)^{\mathrm{T}}\cdot(\mathbf{x}-\mathbf{x}^t) = \text{minimum}\,. \tag{32}$$

If we seek a solution of the general linear form of equation (28), then by substituting (31) into (28), and (28) into (32), differentiating with respect to all elements of $\mathbf{W}$, and then solving for $\mathbf{W}$, we find

$$\mathbf{W} = \mathbf{B}\cdot\mathbf{K}^{\mathrm{T}}\cdot(\mathbf{K}\cdot\mathbf{B}\cdot\mathbf{K}^{\mathrm{T}} + \mathbf{Y})^{-1}\,, \tag{33}$$

where

$$B = \left(\frac{1}{N}\right)\sum^{N}(\mathbf{x}^b - \mathbf{x}^t)\cdot(\mathbf{x}^b - \mathbf{x}^t)^{\mathrm{T}}\,, \tag{2a}$$

and

$$\mathbf{Y} = \left(\frac{1}{N}\right)\sum^{N}(\varepsilon^m)\cdot(\varepsilon^m)^{\mathrm{T}}\,, \tag{2b}$$

i.e. $\mathbf{B}$ and $\mathbf{Y}$ are the background and measurement error covariances.

We note that the minimum variance solution is equivalent to the maximum probability solution in the linear case when the error characteristics of both background profile and measurements are Gaussian.

2.3 (d) Linear regression. It is possible to tackle the inverse problem without any knowledge of the radiative transfer physics if we have a large set of N measurement pairs of satellite radiances $\mathbf{y}^m$ with atmospheric profiles $\mathbf{x}^m$ (e.g. from radiosondes) closely matched in time and space. We can then look for the coefficients of a linear combination of the radiances which best predict the atmospheric profiles in a least-squares sense. We write a predictive equation

$$\mathbf{x} - \overline{\mathbf{x}^m} = \mathbf{W}\cdot(\mathbf{y}^m - \overline{\mathbf{y}^m})\,, \tag{34}$$

© ECMWF, 2002

where $\mathbf{x}^m$ and $\mathbf{y}^m$ are the mean vectors of our large sample (usually called the dependent sample). Then we find the value of $\mathbf{W}$ for which

$$\sum^{N}(\mathbf{x}-\mathbf{x}^m)^T(\mathbf{x}-\mathbf{x}^m) \;=\; \text{minimum}\,, \tag{35}$$

$$\sum^{N}[\mathbf{x}-\overline{\mathbf{x}^m}-\mathbf{W}\cdot(\mathbf{y}^m-\overline{\mathbf{y}^m})]^T[\mathbf{x}-\overline{\mathbf{x}^m}-\mathbf{W}\cdot(\mathbf{y}^m-\overline{\mathbf{y}^m})] \;=\; \text{minimum}\,. \tag{36}$$

Differentiating with respect to all elements of $\mathbf{W}$, and solving for $\mathbf{W}$, gives

$$\mathbf{W} \;=\; \left[\sum^{N}(\mathbf{x}^m-\overline{\mathbf{x}^m})\cdot(\mathbf{y}^m-\overline{\mathbf{y}^m})^T\right]\cdot\left[\sum^{N}(\mathbf{y}^m-\overline{\mathbf{y}^m})\cdot(\mathbf{y}^m-\overline{\mathbf{y}^m})^T\right]. \tag{37}$$

This is a purely statistical method. Alternatively, the technique may be used in a "physical-statistical" manner by starting from a set of N representative profiles, calculating from them theoretically the corresponding radiances and regressing the two sets of data as in equation (37) (taking care to allow for the measurement error in the radiances).

If the radiative transfer equation is linear, it may be written as

$$\mathbf{y}^m-y\{\overline{\mathbf{x}^m}\} \;=\; \mathbf{y}^m-\overline{\mathbf{y}^m} \;=\; \mathbf{K}\cdot(\mathbf{x}^m-\overline{\mathbf{x}^m})+\boldsymbol{\varepsilon}^m\,. \tag{38}$$

Substituting (38) into (37) and assuming the measurement errors $\boldsymbol{\varepsilon}^m$ are uncorrelated with the profiles, we obtain

$$\mathbf{W} \;=\; \mathbf{C}\cdot\mathbf{K}^T\cdot(\mathbf{K}\cdot\mathbf{C}\cdot\mathbf{K}^T+\mathbf{E})^{-1}\,. \tag{39}$$

This shows that, in the linear limit, linear regression methods are mathematically equivalent to the minimum variance solution described in section 2.3 (c), where the background profile and its error covariance are the mean and covariance of the dependent set.

2.3 (e) Physical iterative methods. The statistical approaches presented above are mainly linear. Some are not well-suited to handling nonlinear problems and those which are do so at the expense of considerable computation. Physical methods (which use *ad hoc* mathematical rather than statistical constraints) are more flexible in this respect. However, they do not attempt to be "optimal" (in the sense in which the statistical methods are) and, if not used with care, can converge on a solution which, although it fits the measurements, is not meteorologically realistic.

We present here the physical, iterative method due to Smith (1970, 1985) adapted slightly to conform with the notation used above. At the n th step of the iteration, we have an estimate of the profile $\mathbf{x}^n$ from which we calculate radiances $\mathbf{y}(\mathbf{x}^n)$. The profile is then updated according to the difference between the measured and calculated radiances:

© *ECMWF, 2002*

$$x_j^{n+1} = x_j^n + \frac{\sum_{i=1}^{I} W_{ij}(y_i^{\mathrm{m}} - y\{x_n\}_i)}{\sum_{i=1}^{I} W_{ij}}. \qquad (40)$$

W_{ij} are empirical weights which could take a number of forms. Smith uses the weighting functions themselves as weights and finds that this gives a stable convergence. The iteration is started from a first-guess profile, such as a forecast profile.

It is also possible to construct hybrid methods in which a statistical regression inversion is used as a first-guess for a physical iterative method, or in which a physical method provides the linearization point for a linear statistical retrieval. With care, a method which is physical and iterative can also be made statistically optimal - see Rodgers (1976) or Eyre (1989).

3. CONSTITUENT PROFILE INVERSION

Satellite measurements at frequencies in absorption bands of atmospheric constituents with variable concentration (such as water vapour and ozone) can be used to estimate the profiles of these constituents. The principles of the methods are generally the same as those presented above, but certain aspects of the problem make constituent profile inversion more difficult than that of temperature.

Firstly, the problem is more nonlinear for constituents. This is because they enter the radiative transfer equation through the mixing ratio profile in the exponent of equation (2a). It is not possible to separate the radiative transfer equation into the product of a simple constituent function and one which is constituent-independent. The consequence is that methods which make assumptions of linearity are less accurate. We can still define a "temperature weighting function" for a constituent-sounding channel as the derivative of the transmittance profile, but this weighting function will vary considerably; its peak will move up in height as the mixing ratio increases.

The second main problem is that, in certain circumstances, the radiances are insensitive to changes in the mixing ratio. To illustrate this, consider the limit of an isothermal atmosphere at temperature T. Then, any mixing ratio profile will result in the same radiances to space (i.e. $B_\nu\{T\}$). In practice, we find this problem in the retrieval of low-level water vapour; because the temperature of the water vapour is close to that of the surface, infrared radiances are relatively insensitive to changes in low-level water vapour. This is not the case for microwave measurements over the sea, where the low emissivity of the sea surface provides an apparently cold background against which changes in low-level water vapour can be detected.

Simple linear regression methods have been used operationally for water vapour profile inversion, but nonlinear methods are potentially superior. For more details, see Smith (1985). In recent years, increasing use has been made of SIMULTANEOUS methods, in which the atmospheric profile vector **x** contains both temperature and water vapour profiles (and possibly also other variables which affect the radiative transfer such as surface emissivity and cloud parameters). See Smith *et al.* (1985) or Eyre (1989).

4. CLOUDS

Clouds create a major problem for temperature retrieval. Not only do they have considerable effects on infrared radiances, but they make the inversion problem highly nonlinear; they cause the weighting functions to change abruptly at the cloud top and thus make them strong functions of cloud-top pressure and amount. Moreover, opaque

© ECMWF, 2002

clouds preclude the sensing of information from below the cloud.

The first approach to the problem is technological, i.e. to sound the atmosphere at frequencies for which clouds are (almost) transparent, such as in the microwave region. There is a tendency for increasing reliance on microwave radiometers as technology in this region advances. However, the infrared has a number of advantages - narrower weighting functions in the troposphere, less variable surface emissivity/reflectivity and higher spatial resolution. For these reasons, a combination of complementary infrared and microwave instruments is likely to be preferred for the foreseeable future.

The second approach is to screen the data carefully for cloud "contamination" and to use only data from cloud-free areas. Many algorithms have been devised to do this (e.g. see McMillin and Dean, 1982). Unfortunately, most of the interesting weather occurs in cloudy areas and so this is only a partial solution. For partly cloudy areas, so-called "cloud-clearing" algorithms have been devised. These estimate from the "cloudy" radiances the radiances which would have been observed if there had been no cloud. Many of these methods are based on the adjacent field of view or "*N*" method originated by Smith (1968). We make the assumption that two adjacent fields-of-view of an instrument (labelled 1 and 2) contain the same temperature and humidity profile and a single layer of cloud with the same cloud-top height, but with different fractional cloud coverages, N_1 and N_2, respectively. Then we may write the following equations for the radiances emitted to space in each field of view:

$$\begin{aligned} R_1 &= (1-N_1)R^{\mathrm{c}} + N_1 R^{\mathrm{o}}, \\ R_2 &= (1-N_2)R^{\mathrm{c}} + N_2 R^{\mathrm{o}}, \end{aligned} \tag{41}$$

where R^{c} and R^{o} are the radiances for cloud-free and completely overcast conditions, respectively. These equations may be solved simultaneously to give

$$R^{\mathrm{c}} = \frac{R_1 - N^* R_2}{1 - N^*}, \tag{42}$$

where $N^* = N_1/N_2$. Alternatively, we may write

$$N^* = \frac{R^{\mathrm{c}} - R_1}{R^{\mathrm{c}} - R_2}. \tag{43}$$

If an estimate of the clear radiance can be obtained for one channel, then N^* can be found through (43). Then, since N^* is independent of channel, the clear radiance is obtained for all other channels using (42). The method fails if $N_1 = N_2$, which includes the (common) case of completely overcast conditions in both fields-of-view.

The third approach to the problem is to perform an inversion directly from the cloudy radiances, by estimating the parameters describing the cloud conditions either simultaneously or iteratively along with the temperature profile (and other atmospheric parameters). Physical methods which take this approach have been devised (e.g. Huang and Smith, 1986; Susskind *et al.*, 1984) and a nonlinear variational method has been demonstrated by Eyre (1989).

5. SATELLITE SOUNDING DATA IN NUMERICAL WEATHER PREDICTION

In recent years it has become increasingly difficult to show that temperature/humidity profiles retrieved from satellite sounding data have a positive impact within operational NWP data assimilation systems, particularly in areas where other observation types are available. This problem has arisen partly because NWP systems have improved

© ECMWF, 2002

to a point at which great care is required in the treatment of any observation type; appropriate quality control is necessary and the error characteristics of each data type must be taken into account. However, there are additional problems with satellite sounding data.

As discussed in section 1.5, the intrinsic vertical resolution of the satellite sounding system is low, both in relation to other temperature sounding observations (i.e. radiosondes) and to the vertical resolutions of modern NWP models. Because of this the background and constraint information used in the inversion affects the retrieved profile considerably. In practice low-order vertical structures in the retrieval are obtained mainly from the radiance information, but high-order structures come largely from the background information. Consequently, considerable care must be taken to avoid components of the retrieved profile which are not derived from the radiance data, but are artefacts of the inversion method, contaminating an otherwise good NWP analysis. Another symptom of the same problem is that retrieved profiles have systematic error structures of a very subtle and specific nature (see Eyre 1987). It is difficult for many analysis systems to suppress the harmful effects of these error characteristics without simultaneously losing the real information contained in the radiance data.

Recent developments in NWP data assimilation seek to solve these problems by making more direct use within the NWP system of radiance observations themselves, rather than retrieved temperature profiles. The French Direction de la Météorologie (Durand 1986) and the UK Meteorological Office (Eyre and Lorenc 1989) both run operational TOVS processing and assimilation systems based on these ideas. At ECMWF, systems for both one- and three-dimensional variational analysis of TOVS radiances are being developed, based on the theory presented in section 2.3 (b) (see Eyre 1990, Pailleux 1990).

REFERENCES

Durand Y., 1986. The use of satellite data in the French high resolution analysis. Report of the Workshop on “High resolution analysis”; ECMWF, Reading; 24-26 June 1985; ECMWF Report, pp. 89-217.

Eyre J.R., 1987. On systematic errors in satellite sounding products and their climatological mean values. Q. J. R. Meteorol. Soc., 113, 279-292.

Eyre J. R., 1989: Inversion of cloudy TOVS radiances by non-linear optimal estimation. Q. J. R. Meteorol. Soc., 115, 1001-1037.

Eyre J.R., 1990. Progress on direct use of satellite sounding radiances in numerical weather prediction. Preprints WMO International Symposium on “Assimilation of Observations in Meteorology and Oceanography”; Clermont-Ferrand, France; 9-13 July 1990; WMO Report, pp. 117-121.

Eyre J.R. and Lorenc A.C., 1989. Direct use of satellite sounding data in numerical weather prediction. Meteorol. Mag., 118, 13-16.

Huang H.-L. A. and Smith W. L., 1986: An extension of the simultaneous TOVS retrieval algorithm the inclusion of cloud. Tech. Proc. 3rd Int. TOVS Study Conf., Madison, Wisconsin; 13-19 August 1986; Ed.: W. P. Menzel; Report of CIMSS, University of Wisconsin-Madison; pp. 118-131.

Lorenc A.C., 1981. A global three-dimensional multivariate statistical interpolation scheme. Mon. Wea. Rev., 109, 701-721.

Lorenc A.C., 1988. Optimal nonlinear objective analysis. Q.J.R. Meteorol. Soc., 114, 205-240.

McMillin L. M. and Dean C., 1982: Evaluation of a new operational technique for producing clear radiances. J. Appl. Meteor., 12, 1005-1014.

Menke W., 1984: Geophysical data analysis: discrete inverse theory. Academic Press.

Pailleux J., 1990. A global variational assimilation scheme and its application for using TOVS radiances. Preprints

© ECMWF, 2002

WMO Int. Symp. on “Assimilation of Observations in Meteorology and Oceanography”; Clermont-Ferrand, France; 9-13 July 1990; WMO Report, pp. 325-328.

Rodgers C.D., 1976: Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. Rev. Geophys. Space Phys., 14, 609-624.

Schwalb A., 1978. The TIROS-N/NOAA A-G satellite series. NOAA Tech. Mem. NESS 95.

Smith W.L., 1968. An improved method for calculating tropospheric temperature and moisture from satellite radiometer measurements. Mon. Wea. Rev., 96, 387-396.

Smith W. L., 1970: Iterative solution of the radiative transfer equation for the temperature and absorbing gas profile of an atmosphere. Appl. Opt., 9, 1993-1999.

Smith W. L., 1985: Handbook of Applied Meteorology (Ed. D. Houghton). Chapter 10: Satellites.

Smith W.L., Woolf H.M., Hayden C.M. and Schreiner A.J., 1985. The simultaneous retrieval export package. Tech. Proc. 2nd. Int. TOVS Study Conf.; Igls, Austria; 18-22 February 1985; Ed.: W.P. Menzel; Report of CIMSS, University of Wisconsin-Madison; pp. 224-253.

Smith W.L., Woolf H.M., Hayden C.M., Wark D.Q. and McMillin L.M., 1979. The TIROS-N Operational Vertical Sounder. Bull. Am. Meteorol. Soc., 60, 1177-1187.

Susskind J., Rosenfield J., Reuter D. and Chahine M. T., 1984: Remote sensing of weather and climate parameters from HIRS2/MSU on TIROS-N. J. Geophys. Res., 89, 4677-4697.

© ECMWF, 2002