Some statistical considerations associated with the data assimilation of precipitation observations

By RONALD M. ERRICO1*, LUC FILLION2, DOUGLAS NYCHKA1 and ZHAN-QIAN LU3
1National Center for Atmospheric Research, Boulder, USA
2Atmospheric Environment Service, Canada
3The Hong Kong University of Science and Technology, Hong Kong

(Received 15 June 1998; revised 4 May 1999)

SUMMARY

Bayes’s theorem is applied to the problem of analysing temperature and moisture in a volume of air given a single observation of precipitation amount, utilizing a model of non-convective precipitation and prior estimates of the fields. Results using different statistics and shapes of probability distributions are examined. These include normal, truncated normal, and log normal distributions with special treatment of the value zero. The uncertainty of the model’s formulation is considered in addition to uncertainty of observations. The posterior distribution is multi-modal due to the model’s formulation using a conditional expression. The dominant mode may be predicted as a non-precipitating state by the model, although the observation indicates precipitation is present. Means and modes of posterior distributions depend sensitively both on the assumed statistics and the shapes of the underlying distributions. The results suggest that the usual minimization of a cost-function should not be used cavalierly to assimilate precipitation observations.

KEYWORDS: Bayes’s theorem Humidity analysis Temperature analysis

1. INTRODUCTION

Intuitively, it may appear that observations of precipitation amounts or rates should be useful for better defining atmospheric states of temperature, moisture, and even wind. Qualitatively, some states, like dry ones, can certainly be ruled out when precipitation is present, but with an adequate model that relates precipitation rates to states there should also be useful quantitative information for use in a data assimilation procedure. Of course, precipitation data in terms of rates or short-period (i.e., a day or less) accumulations are almost always of poor quality, either due to sparseness of the observing networks or to the indirectness of retrievals from remote-sensing instruments. Current precipitation models, especially of convection, also have less than the desired accuracy. In regions where more conventional data is lacking, however, even this somewhat inaccurate information may be valuable. In particular, it can ameliorate spin-up problems, which often cause models to be slow to develop precipitation (Kasahara et al. 1996).

One of the early attempts to assimilate precipitation observations using a numerical weather prediction model was by Wang and Warner (1988) who used a simple nudging technique. It was basically unsuccessful, due to several ad hoc assumptions that were necessary and also due to the imbalances created by the nudging process (Bao and Errico 1997). More recently, Krishnamurti et al. (1993) have developed a more elaborate nudging and adjustment technique and, independently, Kasahara et al. (1996) have experimented with a combined balancing and adjustment technique. Precipitation amounts have also been assimilated using slightly more elaborate, but much more computationally demanding, four-dimensional variational techniques (commonly called 4D-Var) in studies by Zupanski and Mesinger (1995), Zou and Kuo (1996), and Tsuyuki (1996). Their main contribution to the assimilation problem has been to suggest that the incorporation of precipitation data may indeed be useful and feasible. Most details of the techniques implemented in these recent studies, however, should not be considered

* Corresponding author: National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307-3000, USA. email: rmerrico@ucar.edu
appropriate for an optimal operational data assimilation system, partly for reasons to be made apparent in this paper.

None of the studies referenced above makes explicit use of the statistical nature of the assimilation problem; e.g. the consideration that all the observations, models, and prior estimates available actually have unknown errors that can only be described statistically. This consideration is especially critical where precipitation is concerned, since its associated errors are often quite large. The issue is not simply that the assimilation technique should be variational, with weights determined by the estimated variances of the associated errors; other properties of the probability distributions describing the errors should also be considered, particularly because precipitation rate is a non-negative quantity and therefore its associated errors are not precisely normally distributed. By implication, the result of the assimilation itself also has possible error, and the character of its probability distribution should also be considered if some single ‘best’ analysis is to be made and its accuracy properly interpreted. These important considerations will be discussed more formally in the next section.

There are in fact many issues regarding the assimilation of precipitation that have not been confronted to a significant degree in the assimilation of more conventional data. These include the relatively poor quality both of observations and of models of precipitation, the significant nonlinearity of the models, and the importance of dynamic imbalances. Two of these issues have been partly addressed in Fillion and Errico (1997) and Fillion and Mahfouf (1999). Our goal in this paper is to continue the systematic examination of these issues by beginning with the simplest relevant situation, namely the statistical assimilation of a single observation of precipitation within a single grid box. While such simplicity is too extreme to offer sufficient guidance in developing a truly optimal assimilation system, it will be illustrative of some issues that must be considered. Our focus will be on the effects of model nonlinearity and non-normal probability distributions for the observation error and model error on the shapes of resulting posterior probability distributions.

The statistical assimilation problem for our simple context is presented in section 2 using a general Bayesian approach. The simple precipitation model is introduced in section 3. Specific probability distributions are considered in the following two sections, beginning with Gaussian functions. Conclusions are offered in section 6.

2. ASSIMILATION OF PRECIPITATION DATA FROM A STATISTICAL VIEWPOINT

‘Data assimilation’, as it is known in meteorology, ‘inverse problem theory’ as it is known in other geophysical disciplines, or ‘estimation theory’ as it is known in engineering, has a broad literature base. In particular, Tarantola (1987), Daley (1991), Lorenz (1986) and a collection of papers edited by Ghil et al. (1997) are excellent general references for meteorological applications. They each describe assimilation as a fundamentally statistical problem rather than principally as a simple interpolation or nudging problem. While the formulation of a particular, statistically based scheme may resemble nudging or interpolation, the actual result of the scheme will critically depend on details of its implementation. Those details are likely to be very different were different statistical considerations and assumptions applied (Lorenc 1986) or no explicit statistical considerations made at all.

The statistical approach to data assimilation presented by Tarantola (1987) will be followed here, with a change of notation to make the variable names more familiar and explicit in the context to follow. Also the notation recommended in Ide et al. (1997) will be used where applicable. All the kind of information to be considered and determined
will be expressed in terms of their conditional probability distributions. Here, these distributions will actually be described in terms of their corresponding conditional probability-density functions (PDFs), denoted by $\rho$.

The values to be analysed will be single values of both specific humidity, $q$, and temperature, $T$, considered as mass-weighted mean values within a model grid box. A single value, $p_o$, of pressure will be assumed to be given without any error, consistent with the fact that in the lower troposphere variations of $T$ and $q$ within the size of their uncertainties typically yield much greater variations of model-determined precipitation results than do variations of pressure within the size of its uncertainty. All the relevant PDFs for this problem can therefore be plotted as 1- or 2-dimensional figures, which is the primary motivation for the imposed simplicity.

The precipitation will be expressed here as the ratio, $r$, of the mass of water precipitated out of a volume of air to the original mass of moist air within that volume. This $r$ is related to $T$ and $q$ (given $p_o$ and other ‘constants’) by an imperfect model of precipitation (more generally called the ‘observation operator’ or ‘forward model’):

$$ r = r_m(T, q) + \epsilon_m, \quad (1) $$

where, when $T$ and $q$ are given without error, $\epsilon_m$ is the ‘modelization’ error due to an imperfect model formulation. More generally, the relationship between $T$, $q$, and $r$ can be expressed statistically by a conditional PDF of $r$ given $T$ and $q$: $\rho(r|T, q)$. Here, however, we will restrict considerations to those $\rho(r|T, q)$ that can be expressed in terms of $\rho(r|r_m)$; i.e. distributions that are identical for all values of $T$ and $q$ that yield identical values of $r_m$. For model-error descriptions that are more general than (1), see section 1.3 in Tarantola (1987).

Errors such as $\epsilon_m$ are often called errors of representativeness in older atmospheric data assimilation literature. This terminology is likely to stem from the fact that, until recently, the only errors due to modelling explicitly considered in atmospheric data assimilation were those due to coarse representation of spatially continuous fields by discretized grids or a finite number of specified functions (Daley 1991). This nomenclature pre-dates attempts to consider explicit physical models, such as for precipitation and radiative transfer. These errors are now sometimes called ‘forward model error’ (e.g. Eyre 1989; Anderson et al. 1994; McNally and Vesperini 1996), or simply ‘modelization error’ (Tarantola 1987). Note that some papers that use the term ‘representativeness error’ (e.g. Lorenc 1986; Derber and Wu 1998) make clear that it includes forward model error, but Zou and Kuo (1996) and Tsuyuki (1996), for example, do not consider the rather substantial error in a forward precipitation model.

A single observation of precipitation will be considered, with the value denoted by $r_o$. This observation is necessarily imperfect due to inaccuracies of the observing instrument or system. The relation between $r_o$ and $r$ is most generally described by a conditional probability distribution with PDF $\rho(r_o|r)$.

Imperfect estimates of $T$ and $q$ prior to consideration of the observation will be denoted by subscript $p$. Without additional constraints, this prior information is necessary to construct a problem with a unique solution when a single state is to be selected as a ‘best estimate’, requiring that the number of values to be estimated (2 here) is not greater than the number of values given (3 here). In any case, fairly accurate prior information produced by a short-term forecast is often available and should not be casually neglected. The previous information is also most generally described as a probability distribution, here expressed in terms of a conditional PDF $\rho(T, q|T_p, q_p)$ to explicitly render its dependence on $T_p$ and $q_p$. 
The information given in terms of \( \rho(r_0|r) \), \( \rho(r|T, q) \), and \( \rho(T, q|T_p, q_p) \) can be combined using Tarantola's theory of information or Bayes's theorem (Tarantola 1987: sections 1.5.2 and box 1.4, respectively) to yield the joint probability distribution for \( T \) and \( q \) conditioned on \( T_p, q_p, \) and \( r_0 \):

\[
\rho(T, q|T_p, q_p, r_0) = \text{const} \times \rho(T, q|T_p, q_p) \int_{-\infty}^{\infty} \rho(r_0|r)\rho(r|T, q) \, dr. \tag{2}
\]

The integrand and integral in (2) are respectively proportional to \( \rho(r, r_0|T, q) \) and \( \rho(r_0|T, q) \). If all the conditional PDFs on the right-hand side of (2) are specified and the multiplications and integrations performed, then the analysis problem is solved in a general sense. This solution (2) is called the a posteriori or analysis PDF and will be simply denoted as \( \rho_a \). The constant in (2) is such that the integral of \( \rho_a \) over all possible values of \( T \) and \( q \) is unity.

Although \( \rho_a \) contains all the known information about the state being analysed (Tarantola 1987: section 1.5), often it is necessary to choose a single state as the 'optimal' or 'best' analysis. The two most popular choices are the estimates associated with either the mode or mean of the distribution. The former is the location where \( \rho_a \) is a maximum, indicating which neighbourhood of \( T, q \) has the greatest posterior probability. The latter is the state that yields the minimum expected variance of the analysis errors. This relationship between the mean of the distribution and the minimum variance estimate is independent of both the shape of the distribution and the norm used to measure the variance (e.g. see Cohn 1997). There is another, related estimate, called the 'best linear unbiased estimate' (or BLUE) that will be discussed in section 4. Actually, it is useful to know the shape of the distribution if the appropriateness of using the mean, mode, or some other estimate of the state is to be considered (e.g. see chapter 5 in Tarantola 1987).

For some uses of the analysis, the consequences of choosing different states as the best estimate should also be considered. For example, if it is critical that the possible occurrence of extreme events should not be missed, then the selection of an optimal analysis should be partially weighted towards ones that lead to such events. Tarantola (1987) briefly discusses a similar point and the common difficulty of including such considerations. One example where a single analysis need not be chosen is in ensemble forecasting (e.g. Leith 1974).

### 3. The Model

The model to be considered here is one of stratiform precipitation (see Houze 1997 for a discussion of precipitation types). The mass of precipitation generated per mass of air within some volume that has mean temperature \( T \), specific humidity \( q \), and pressure \( p_0 \), will be simply modelled as

\[
 r_m(T, q) = \begin{cases} 
 a(T, p_0)\{q - q_s(T, p_0)\} & \text{if } q > q_s(T, p_0), \\
 0 & \text{otherwise}, 
\end{cases}
\tag{3}
\]

where \( q_s \) is the saturation specific humidity and \( a \) is a factor that depends on \( \partial q_s/\partial T \). Both \( a \) and \( q_s \) are determined as functions of \( T \) and \( p_0 \) from thermodynamic and other constraints as described in appendix A. These functions are nonlinear in \( T \), but over a range of a few degrees Kelvin they can be well approximated by linearized functions, although the nonlinear formulations are used here.
The function $r_m(T, q)$ is presented in Fig. 1. Note that the contours are almost equally spaced straight lines, indicating that for the ranges of $T$ and $q$ presented, $q_s$ is almost a linear function of $T$ and $a$ does not vary much. The slopes of the contours are approximately $\delta q_s/\delta T$. The principal nonlinearity in (3) therefore comes from the conditional there, and $r_m$ can be considered as an approximately piece-wise linear function of $T$ and $q$.

In the examples presented in sections 4 and 5, two values of $r_0$ will be considered. These are 0.0004 and 0.00125, which correspond to moderate and strong non-convective precipitation rates of 1.2 and 3.8 mm h$^{-1}$, respectively, using relationship (A.4) and the parameters in appendix A. For all the experiments, $p_o = 80$ kPa, $T_p = 280$ K, and $q_p = 7.4$ g kg$^{-1}$. The relative humidity for the state’s prior estimate is $q_p/q_s(T_p, p_o) = 0.95$.

4. Gaussian Distributions

We will begin by making some assumptions commonly invoked in the context of data assimilation for numerical weather prediction. Although these assumptions are actually inappropriate for this model and data, the PDFs take simple forms, are easily manipulated, and have convenient properties. Also, the forms are equivalent to some invoked implicitly using different assumptions elsewhere for solving the precipitation assimilation problem (e.g. Zou and Kuo 1996). More appropriate assumptions will be introduced in the next section, and their consequences contrasted with those here.
Specifically, in this section we will assume

$$\rho(r_0|r) \sim \mathcal{N}(r, \sigma_0^2),$$  \hskip 1cm (4)

$$\rho(r|T, q) \sim \mathcal{N}(r_m(T, q), \sigma_m^2),$$  \hskip 1cm (5)

$$\rho(T, q|T_p, q_p) \sim \mathcal{N}_2(T_p, q_p, \mathbf{B}),$$  \hskip 1cm (6)

where the tilde ($\sim$) should be read as 'is expressed by' and $\mathcal{N}$ denotes a normal (i.e. the $\ell_2$ Gaussian) distribution whose two arguments are the distribution's mean and variance, respectively. In the case of (6), $\mathcal{N}_2$ denotes a bivariate distribution with vector mean $T_p$, $q_p$ and $2 \times 2$ covariance matrix $\mathbf{B}$.

For a conditional probability of a variable $x$ that is normally distributed having $y$ as its mean, its PDFs depend on $x$ and $y$ only in the form of their difference $x - y$. For the distributions here, where one of $x$ or $y$ is the assumed but unknown true physical value, that difference may be interpreted as an error; e.g. an observational error, a model error, or an error in the prior estimate. Since the true state is unknown, these errors are also unknown, but may be considered random with some probability distribution. In fact, assuming (4)–(6) is equivalent to assuming that the corresponding errors are normally distributed with zero mean; i.e. that the errors are unbiased. Additionally, assuming that the variance of the distribution is independent of $y$ is equivalent to assuming that the errors are statistically independent of the state. This equivalence of assumptions can be invoked in deciding whether the assumed distributions are indeed reasonable, as discussed further in section 5.

When (4) and (5) are integrated together as in (3), the result is

$$\int_{-\infty}^{\infty} \rho(r_0|r) \rho(r|T, q) \, dr = \text{const} \times \exp[-0.5 \sigma_d^{-2}(r_m(T, q) - r_0)^2],$$  \hskip 1cm (7)

with

$$\sigma_d^2 = \sigma_0^2 + \sigma_m^2.$$  \hskip 1cm (8)

Note that $\sigma_d$ depends similarly on the observation error and modelization error variances.

Combining (6) and (7) yields

$$\rho_a = \text{const} \times \exp\{-J(T, q)\},$$  \hskip 1cm (9)

with

$$J(T, q) = \frac{1}{2} \left[ (T - T_p)^T \mathbf{B}^{-1} (T - T_p) + \sigma_d^{-2}(r_m(T, q) - r_0)^2 \right],$$  \hskip 1cm (10)

where the dependence on the prior state values is written in (10) as a 2-component vector, superscript $T$ indicates a transpose, and the constant in (9) is such that the integral of $\rho_a$ over all possible values of $T$ and $q$ equals 1. Although (9) and (10) resemble a PDF for a multivariate normal distribution, it is only if $r_m(T, q)$ is a linear function of its arguments.

The function $J$ is sometimes called the ‘misfit function’, or ‘cost function’ (Ide et al. 1997). The $T$ and $q$ that minimize $J(T, q)$ determine the maximum of (9) and hence the mode of the distribution. If the model were linear this minimizer would also be the mean of the distribution, since the mean and mode are identical for a normal distribution (also conditions for the BLUE apply as discussed below).
Often, meteorological data assimilation is expressed as a variational problem without explicit assumptions about underlying distributions such as (4)–(6). The solution sought is the minimizer of a $J$ expressed as in (10). The case when the underlying distributions are normal has just been discussed. If the model is linear, and the means of the distributions of $T - T_p$, $q - q_p$, and $r_m - r_o$ are zero with corresponding variances expressed by $B$ and $\sigma_d$, then the minimizer of $J$ is the best linear unbiased estimator or BLUE (see, e.g. Talagrand 1997). The BLUE is the mean of the distribution corresponding to $\rho_a$, irrespective of whether or not the underlying distributions are normal. If they are not normal however, then (9) does not apply, $\rho_a$ may be non-normal, and the BLUE may be an inappropriate estimate (e.g. see problem 1.9 and chapter 5 in Tarantola 1987). Also, if the model is nonlinear (as it is here) or if the means of the distributions of the differences in (10) are non-zero, the conditions for the BLUE are not satisfied. The minimizer of $J$ is also termed the ‘least-squares estimate’, regardless of whether $B$ and $\sigma_d$ are the variances of the underlying distributions or whether the means of the distributions of $T - T_p$, $q - q_p$, and $r_m - r_o$ are zero. In this case however, little can be said regarding the PDF for the analysis or the appropriateness of any estimate, except that the weights applied to the least-squares measure do not necessarily reflect the uncertainty inherent in the various components of $J$. If both the underlying distributions are non-normal and the model is nonlinear, there is indeed little that can be claimed about a minimizer of (10) as an estimator, since (9) no longer applies.

(a) Results for normal distributions

Four examples of distributions $\rho_a$ are presented in this section. These are generated using 4 sets of prior statistics. The uncertainty of the prior estimate is assumed to have standard deviations $\sigma_T = 1.6$ K, $\sigma_q = 0.0007$ for the indicated variables with no correlation, yielding a diagonal matrix for $B$ with elements $\sigma_T^2$ and $\sigma_q^2$. These uncertainties were used by Fillion and Errico (1997) as derived from a data analysis system employed for the Canadian forecasting system.

The PDF of the prior estimate appears in Fig. 2. Contours appear almost as circles because the ranges of values of $T$ and $q$ presented are both approximately 6 standard deviations and $T - T_p$ and $q - q_p$ have been considered to be uncorrelated. The particular contours displayed for each PDF are selected such that the probability associated with the fields of $T$, $q$ values enclosed by successive contours are successive multiples of 10%. The contours are labelled with these percentages. All PDFs that are functions of both $T$ and $q$ will be presented using the same contour selection technique.

For all examples in this section, $\sigma_0$ and $\sigma_m$ will be specified as the same value, designated $\sigma_{om}$. For each value of $r_o$ considered, two values of $\sigma_{om}$ will be considered, specified as either 20% or 50% of the value used for $r_o$, respectively indicating relatively certain or uncertain model and observations. The 50% factor was chosen to illustrate a particular point to be discussed below. Although this factor yields large values of $\sigma_{om}$ (which are actually unrealistic for the normal distributions (4)–(5)), it does not grossly over-estimate the magnitude of error statistics for the simple model (3) and for many precipitation observing systems.

A pair, $T$, $q$, will be designated by $[\cdot, \cdot]$ with respective units of K and g kg$^{-1}$. The mean and primary (dominant) mode of any plotted distribution will be indicated by $M$ and $H$ respectively.

One example of the integrand appearing in (2) is shown in Fig. 3 as a function of $T$ and $q$ for a particular value of $r_o$. It is determined for the case of an observed moderate precipitation rate with relatively uncertain model and observations. Since this integrand
is only a function of \( r_0 \), \( \sigma_{om} \), and \( r_m \), and the first two are independent of \( T \) and \( q \), the contours are parallel to those of \( r_m \) appearing in Fig. 1. Values are a maximum along the curve where \( r_0 = r_m \). The resulting probability density \( \rho_a \) is proportional to the product of the function in Fig. 3 with that in Fig. 2.

In Fig. 4(a), \( \rho_a \) is presented as a function of \( T \) and \( q \) for the moderate precipitation case with relatively certain observation and model. The mean is \([278.61, 7.90]\) and the mode is nearly identical. The PDF is approximately normal, because in the range where both the integrand and \( \rho(T, q|T_p, q_p) \) have significant magnitude the model is approximately linear.

A secondary mode of \( \rho_a \) for the parameters used for Fig. 4(a) also exists, although unnoticeable in the way that the PDF is presented. In the region where \( q < q_s \), which includes the location of the prior, \( r_m = 0 \) and hence \( r_0 - r_m \) and therefore (7) are constants. The only variation of \( \rho_a \) in that region is therefore due to the variation of \( \rho(T, q|T_p, q_p) \), which has a maximum at \([T_p, q_p]\). A local maximum of \( \rho_a \) therefore exists at this same point.

The \( \rho_a \) in Fig. 4(b) is determined in the same way as for Fig. 4(a), but using the 50% factor to specify \( \sigma_{om} \). The primary mode is \([278.89, 7.79]\). The mean \([279.69, 7.51]\) is between this mode and the secondary mode \([T_p, q_p]\) which is now noticeable; in fact, it is closer to the latter. The discontinuity along the contours occurs at the points where \( q = q_s(T) \); i.e. where the conditional in (3) changes sense. Compared to Fig. 4(a), the primary mode and mean for \( T \) have changed by 0.3 K and 0.9 K, respectively, due to
the increased weight of the prior relative to that of $r_m - r_0$. A change in the choice of $T$
by this amount can have profound effects on a numerical weather forecast.

The $\rho_a$ corresponding to those in Figs. 4(a) and (b), except for the larger observed precipitation rate, appear in Figs. 4(c) and (d), respectively. In Fig. 4(c), the mode and mean are shifted to larger values of supersaturation compared with Fig. 4 to yield more precipitation, and the breadth of the PDF is increased because $\sigma_{om}$ has been specified proportionally to the larger specified value of $r_0$. The mode in Fig. 4(d) is $[T_p, q_p]$, which is a non-precipitating state according to (3). If it is desirable to determine state estimates that are precipitating according to the model and as revealed by the observations, the secondary mode at $[279.18, 7.71]$ may be a better estimate than the primary mode. If $\rho_a$ in Fig. 4(d) were instead produced using a 45% factor to specify $\sigma_{om}$, the primary mode would yield precipitation according to the model. A small change in $\sigma_{om}$ can therefore have a profound effect on determination of the primary mode.

Note that if the distribution for the prior were uniform over the range of values presented, $\rho_a$ would be proportional to the function in Fig. 3. There would be a continuum of values that were either the maximum likelihood estimate or the mean of the distribution. Also note that any single pair $T, q$ selected for either estimate would be very different from the estimates determined with the normally distributed prior. The information assumed in the prior therefore has a profound impact on the resulting analysis.
Figure 4. The probability density function $p(q, T | q_p, q_p, r_0)$ for observations $r_0 = 0.0004$ in (a) and (b), and $r_0 = 0.00125$ in (c) and (d) with variance $\sigma_{\text{off}}$ as (a) $8 \times 10^{-5}$, (b) $2 \times 10^{-4}$, (c) $2.5 \times 10^{-4}$, (d) $6.25 \times 10^{-4}$, produced assuming normal prior, observation, and modelling distributions. The contours are presented as those in Fig. 2. The mode and mean of the distribution are indicated by the letters H and M, respectively. See text for further details.
Figure 4. Continued.
5. **NON-GAUSSIAN DISTRIBUTIONS**

The assumption of Gaussian distributions in section 4 is inappropriate, foremost because \( r, r_0, \) and \( r_m \) are always non-negative. Here, other distributions for which the concerned variables are strictly non-negative are considered.

If we consider the variables \( r, r_0, \) and \( r_m \) as continuously valued, rather than as falling into discrete bins as observations would commonly be reported, then the probability of any specific positive value occurring would be zero. If the values were binned, the probability of a value falling within any particular bin would tend to be approximately proportional to the size of the bin as the range of values defining the bin were decreased. This is not true for the value zero, however, or for a bin containing the value zero. There is indeed a finite probability that \( r_0 = 0 \) given \( r \neq 0 \) (i.e. that precipitation occurs but is unobserved) or that \( r = 0 \) given that \( r_m = 0 \) (i.e. that the model often accurately predicts when no precipitation is occurring). The value zero must therefore be treated in some special way.

As a conditional distribution of an \( x \) given \( y \) that treats zero more appropriately, we consider here:

\[
P(x = 0|y) = P_0(y), \tag{11a}
\]
\[
P(0 < x \leq x_2|y) = (1 - P_0(y)) \mathcal{F}(x_2, y), \tag{11b}
\]

where \( P_0 \) is a point probability for \( x = 0 \) whose value depends on \( y \) and other parameters, and \( \mathcal{F} \) is a distribution defined only for \( x > 0 \) that also depends on \( y \) and other parameters. Some properties of this composite distribution are presented in appendix B.

Since we do not have a dataset to estimate a functional form of \( P_0 \), we will simply assume that as \( y \) increases, the conditional probability that \( x = 0 \) given \( y \) decreases monotonically. A simple function with this property is the exponential

\[
P_0(y) = \alpha \exp(-y/\beta) \tag{12}
\]

with parameters \( \alpha, \beta \).

In the case of (11) and (12) applying to \( \rho(r_0|r) \), we will restrict our presentations to results with \( \alpha = 1 \) and \( \beta = 0.0004 \). The former implies that when there is truly no precipitation there is no chance of precipitation being observed; i.e. the conditional probability of \( r_0 = 0 \), given \( r = 0 \), is 1. When \( r = 0.0004 \) or 0.00125 (which are the values considered for \( r_0 \) in the examples) this \( P_0 \) has values 0.37 and 0.04, respectively. In the case of (11) applying to \( \rho(r|r_m) \) we use \( \alpha = 0.8 \) and \( \beta = 0.0004 \), which indicates that when \( r_m = 0 \), there is only an 80% probability that the model 'prediction' is correct.

For these non-normal distributions, the integral appearing in (2) has been computed by numerical integration, using 10,000 values of \( r \) equally or exponentially spaced. These integrals were checked to be satisfactory by examining how they converged as the number of values of \( r \) increased.

(a) *The truncated normal distribution*

Two choices for \( \mathcal{F} \) will be considered. The first is the truncated normal with corresponding PDF

\[
\phi_1(x|y) = \text{const} \times \begin{cases} 
\exp\left(-\frac{1}{2}(x - y)^2/\sigma^2\right) & \text{if } x > 0, \\
0 & \text{otherwise},
\end{cases} \tag{13}
\]

where the constant is such that the integral of \( \phi_1 \) over all \( x \) is 1. This distribution is chosen because it has similar shape to the distributions in section 4, except in the
region where $x \leq 0$ where it vanishes. We can, therefore, begin by examining effects of considering the values $x = 0$ and $x < 0$ more properly, without affecting some other aspects of the distribution. We do not claim this is an otherwise appropriate distribution for the specific distributions to which (11) is applied.

The results shown in Fig. 5(a) and (b) are produced using the same parameters to define $\phi_1$ as for $\rho$ used to produce Fig. 4(b) and (d), respectively, although due to the truncation and treatment of zero, these parameters are not the mean and variance of the distribution (11). The means of the distribution shown in Fig. 5(a) and (b) are [279.19, 7.69] and [279.16, 7.70], respectively. The respective primary modes are [278.68, 7.89] and [278.77, 7.84]. The mean values of $T$ differ from the non-truncated Gaussian results by 0.5 K and 0.55 K, respectively, which can be significant differences. The modes shown in Figs. 5(b) and 4(d) are shifted by 1.25 K. Results for the parameters that produced Fig. 4(a) and (c) are very similar to those obtained using the non-truncated normal distribution with identical parameters (not shown) because, when the standard deviation is small compared to the mean of the distribution, very little of the normal distribution is truncated.

(b) The log-normal distribution

The second choice for $\mathcal{F}$ examined is another common distribution, the log-normal one. This will be denoted $\mathcal{L}(x_0, s)$ for a variable $x$ whose natural logarithm is distributed normally with mean $\ln x_0$ and variance $s^2$. The corresponding probability distribution in terms of $x$ rather than $\ln x$ is described in appendix B. The $\mathcal{L}$ distribution is often used to approximate distributions of observed precipitation rates (e.g. Wilheit et al. 1991), but this does not imply it is also appropriate for describing distributions of observations or model errors. We use it as much for its familiarity as for any other reason.

For specifying $\rho(r_0|r)$, the parameter $x_0$ used to define $\mathcal{L}$ will be taken as the maximum of $r$ and a parameter $c_0$. Similarly, for specifying $\rho(r|r_m)$, $x_0$ will be the maximum of $r_m$ and $c_0$. This $c_0$ is used to ensure that corresponding $\ln x_0$ remain defined when $r$ or $r_m$ vanish. We use the value $c_0 = 1 \times 10^{-5}$ which is fairly small compared to most values of $r$ and $r_m$ that need to be considered.

The same moderate and heavy rates for $r_0$ will be used here as for the normal and truncated normal distributions. Two values of $s$ will be considered: $s = 1.3$ and $s = 2$. The former indicates relatively certain observations and model, and the latter uncertain. We shall make comparisons with our earlier results based on this characterization, although no strict correspondence exist because not only the PDFs but also the parameters are very different.

An example of the integrand appearing in (2) analogous to that shown in Fig. 3, but for the distribution (11) with $\mathcal{F} = \mathcal{L}$, is shown in Fig. 6. Like Fig. 3, it is for the observed moderate precipitation rate with relatively uncertain model and observations. It appears similar to Fig. 3, except that the maximum values of the integrand are displaced towards greater $q$ for a given $T$, and the spacing between contours is more irregular (particularly less symmetric about the line of maximum values). This is due to the asymmetric character of the log-normal distribution.

The $\rho_0$ produced using $\mathcal{L}$ applied to (11) with $r_0 = 0.0004$ and $s = 1.3$ appears in Fig. 7(a). It appears somewhat similar to the approximate ellipse in corresponding Fig. 4(a), except it is slightly broader along the ellipse's minor axis and more flattened on the side of lower $r_m$. The mode of the distribution is not significantly affected, but the mean value of $T$ is reduced by 0.2 K compared with that in Fig. 4(a).
Figure 5. The probability density function $\rho_0(T, q|T_0, q_0, r_0)$ produced as for (a) Fig. 4(b), and (b) Fig. 4(d) but with the distribution of Eqs. (11) using a truncated normal distribution for $F$ for the observation and model information. See text for further details.
Figure 6. The integrand appearing in Eq. (2) as a function of $T$ and $q$ assuming the distributions of Eqs. (11) with $F = L$ for the observation and model information, using $r_0 = 0.0004$ and $s = 2$ to determine the means and variances. These and other parameters that specify the distribution are defined in the text. The contour interval is 500.

Figure 7(c) is produced as for Fig. 7(a), except with $r_0 = 0.00125$; it is also a distorted ellipse. Its mean and mode are respectively reduced by 0.4 K and 0.17 K compared with the corresponding Fig. 4(c).

For the $s = 2$ examples in Fig. 7(b) and (d) the shapes of the PDFs appear similar to those for $s = 1.3$, but broader. There is no suggestion of bimodality in Fig. 7(d) in contrast to the corresponding Fig. 4(d), but a secondary mode does exist at $[T_p, q_p]$ for the same reasons as explained in section 4. The mean of the distributions in Fig. 7(b) and (d) differ by more than 1.5 K and 0.5 g kg$^{-1}$ compared with those in the corresponding Fig. 4(b) and (d). The modes also have shifted considerably.

Other values for $\alpha$, $\beta$, and $c_0$ were investigated. Values $1 \leq \alpha \leq 0.5$ produced very similar means and modes of the probability distribution for the analysis. Decreasing $\beta$ by a factor of 4 or increasing $c_0$ by a factor of 4 had little impact on the means or modes.

6. Conclusions

Bayes's theorem has been applied to the problem of assimilating observations of precipitation using a simple, but nonlinear, prediction model. The example here highlights some important properties of the problem.

The first important point, discussed in section 4, is that the weight assigned to the observational fit term in the usual cost function defined in variational analysis should include consideration of both model and observational error. If the model and observational errors are independent of each other and both Gaussian, then the weight for the square of the misfit of the observation to model value should be the inverse
Figure 7. The probability density function $p_{q}(T, q|T_{p}, q_{p}, r_{o})$ for observations $r_{o} = 0.0004$ in (a) and (b), and $r_{o} = 0.00125$ in (c) and (d) with $s = 1.3$ in (a) and (c), and $s = 2$ in (b) and (d), produced assuming a Gaussian distribution for the prior, but the distribution of Eqs. (11) with a log-normal distribution in place of $F$ for consideration of the observation and model. The presentation is as for Fig. 4. See text for further details.
Figure 7. Continued.
of the sum of the observation and model error variances (8). In the case when highly inaccurate models of precipitation are considered, the contribution by the model error variance can, therefore, become very significant, although not necessarily additive since the Gaussian assumption also becomes inappropriate. This is not a new result, but is frequently ignored in meteorological data assimilation literature.

The probability distribution for the analysis can be bimodal due to model non-linearity that, in this case, is primarily a consequence of the discontinuous formulation of the precipitation model. In fact, if precipitation is observed but none is predicted using the usual model evaluated from the prior information, and if Gaussian errors are assumed for the observation, model, and prior information, then bimodality is always present in the examples here. One mode may be extremely subdominant, however, as discussed in connection with Fig. 4(a). In the examples here, the primary and secondary modes differ by as much as a few degrees in temperature and tens of percent in relative humidity. Schemes that determine a mode simply by finding where the gradient of the misfit function vanishes may, therefore, find very different modes depending on the specific algorithm and its starting point (e.g. see Dharssi et al. 1992).

The ‘best’ estimate for the analysis is not necessarily the primary mode of the a posteriori distribution. In particular, when the observation or model are relatively uncertain, the primary mode may be a non-precipitating one defined by the prior information, even though precipitation is observed and certainly present. For some applications, choosing a precipitating secondary mode may be better. The problem of determining an ‘optimal’ estimate given bimodality of the distribution for the analysis is, therefore, not necessarily simply ensuring that the primary, rather than secondary, mode is found. This is equivalent to stating that an optimal analysis is not necessarily that which minimizes a cost function.

Since actual, model, and observational values of precipitation cannot be negative but may be zero or close to it, an assumption of normally distributed errors is unrealistic. Also, the value of precisely zero is special since it has a significant probability of occurring; e.g. no precipitation may fall into a network of buckets sparsely distributed within some area although precipitation is occurring within that area. For these reasons, non-normal distributions that addressed these issues were investigated. One incorporated a truncated normal distribution and the other a log-normal distribution. When the observation and model information were considered to be relatively uncertain, these non-normal distributions yielded significantly different means for distributions of the analysis compared with those produced assuming normal distributions.

In the examples presented in sections 4 and 5, altering the variances of the assumed distributions for observation and model information had significant impacts (up to 2 K and 1.0 g kg⁻¹) on the means of the distributions for the analyses. The primary modes were affected less, except when a change of variance caused a secondary mode to become the primary one.

It is premature to base a new analysis scheme for the assimilation of precipitation observations on these results. First, too many aspects of the problem require more careful attention; some of these aspects were mentioned in the introduction. Second, until descriptions of realistic observation and model error distributions are available, it is difficult to evaluate simplified error distributions or model physics since such comparisons depend so strongly on details, particularly the prescribed statistics. Such evaluations also may change as more realistic data assimilation systems are considered. For example, in 4D-Var procedures, mechanisms may be present to render some simplifications better approximations (e.g. because of additional smoothing when spatially correlated prior information is considered) or as worse ones (e.g. because more degrees of freedom allow
greater multimodality). Our intention is to take a sequential approach to examining all the important aspects of this problem in order to better understand more complex results that we expect to eventually obtain.

**APPENDIX A**

*Details of the precipitation model*

The function $q_s$ is the saturation specific humidity

$$q_s(T, p_o) = 0.622 \frac{v_p(T)}{p_o - v_p(T)}, \quad (A.1)$$

where $v_p$ is the saturation vapour pressure. It is approximated using a form of Teten's formula:

$$v_p(T) = 611.2 \exp \left(17.67 \frac{T - 273.15}{T - 29.65}\right), \quad (A.2)$$

(units of Pa with $T$ expressed in K).

As supersaturated water vapour is condensed it changes $q$ by $\Delta q$ and simultaneously heats the air to change the temperature by

$$\Delta T = \frac{L}{c_p} \Delta q, \quad (A.3)$$

where $L = 2.5104 \times 10^6$ J Kg$^{-1}$ is the latent heat of condensation (approximated by a constant here) and $c_p = 1005.7$ J Kg$^{-1}$K$^{-1}$ is the specific heat of the air at constant pressure (also approximated by a constant here). The net changes conserve both enthalpy $E = m(c_p T + L q)$ and water mass $m(r + \Delta q) = 0$, where $m$ is the mass of air in the volume. Additionally, it is assumed that $q_{\text{new}} = q_s(T_{\text{new}}, p_o)$ at the end of the process. The simultaneous satisfaction of these conditions is modelled only approximately by applying

$$a(T, p) = \left(1 + \frac{L}{c_p \partial q_s \partial T}\right)^{-1} \quad (A.4)$$

in (3). This factor of $a$ would yield the exact equilibrated result, properly accounting for the accompanying heating, if $q_s$ were a linear function of $T$.

The value $r_m$ can be related to a precipitation rate by assuming that the same rate of generation occurs within several, vertically stacked volumes, each volume a size appropriate for defining (3), and that the precipitation out from the bottom volume is simply the sum of that generated within each volume. For a net vertical pressure thickness of the stack denoted as $\Delta p$, the stack's total mass of air per unit horizontal area may be expressed hydrostatically as $\Delta p/g$, $g$ being the acceleration of gravity. If all the precipitation falls through the box in time $\Delta t$, the precipitation rate in metres per unit horizontal area per unit time through the bottom of the stack is

$$R = \frac{r_m \Delta p}{c_w g \Delta t}, \quad (A.5)$$

where $c_w$ is the density of liquid water. In sections 4 and 5, values of $\Delta p = 10$ kPa, $g = 9.8 \text{ m s}^{-2}$, $c_w = 1000 \text{ kg m}^{-3}$, and $\Delta t = 20 \text{ min}$ (the latter is the time step in a
typical global forecast model) are used to relate \( r_m \) to \( R \). Relation (A.5) is rather crude but adequate for scaling considerations.

APPENDIX B

The non-normal distributions

The PDF for the log-normal distribution \( \mathcal{L} \) with parameters \( x_0 \) and \( s^2 \) is

\[
\rho(\ln x) = \frac{1}{(2\pi)^{\frac{1}{2}} s} \exp\{-0.5s^{-2}(\ln x - \ln x_0)^2\}. \tag{B.1}
\]

The mean and variance of \( \ln(x) \) for this distribution are \( \ln(x_0) \) and \( s^2 \), respectively. In terms of \( x \) rather than \( \ln(x) \), however, this PDF is expressed as

\[
\rho(x) = \frac{1}{(2\pi)^{\frac{1}{2}} sx} \exp\{-0.5s^{-2}[\ln(x/x_0)]^2\}. \tag{B.2}
\]

It has mean and variance

\[
\bar{x} = x_0 \exp(s^2/2), \tag{B.3}
\]

\[
(x - \bar{x})^2 = x_0 \exp(s^2/2)[\exp(s^2) - 1]^{\frac{1}{2}}. \tag{B.4}
\]

The mean and variance of the distribution in (11) can be easily determined from the value of \( P_0 \) and the mean \( \mu_F \) and variance \( \sigma_F^2 \) of \( F \):

\[
\bar{x} = (1 - P_0)\mu_F, \tag{B.5}
\]

\[
(x - \bar{x})^2 = (1 - P_0)(\sigma_F^2 + P_0\mu_F^2). \tag{B.6}
\]

ACKNOWLEDGEMENTS

The authors wish to thank Kevin Raeder and three anonymous reviewers for their comments on an earlier draft of the manuscript. This work has been partially supported at NCAR by U.S. Navy contract N00173-96-MP-00152 and NASA contract W-18,077, Mod. 6.

The National Center for Atmospheric Research is sponsored by the US National Science Foundation.

REFERENCES


Bao, J.-W. and Errico, R. M.

Cohn, S. E.

Daley, R.

Derber, J. C. and Wu, W.-S.

Dharssi, I., Lorenc, A. C. and Ingleby, N. B.

Eyre, J. R.


1991 *Atmospheric data analysis*, Cambridge University Press, UK


