Tellus (2006), 59A, 198–209
Printed in Singapore. All rights reserved

TELLUS

Measuring information content from observations for data assimilation: relative entropy versus shannon entropy difference

By QIN XU*, *NOAA/National Severe Storms Laboratory, Oklahoma, USA*

(Manuscript received 16 June 2006; in final form 21 November 2006)

ABSTRACT

The relative entropy is compared with the previously used Shannon entropy difference as a measure of the amount of information extracted from observations by an optimal analysis in terms of the changes in the probability density function (pdf) produced by the analysis with respect to the background pdf. It is shown that the relative entropy measures both the signal and dispersion parts of the information content from observations, while the Shannon entropy difference measures only the dispersion part. When the pdfs are Gaussian or transformed to Gaussian, the signal part of the information content is given by a weighted inner-product of the analysis increment vector and the dispersion part is given by a non-negative definite function of the analysis and background covariance matrices. When the observation space is transformed based on the singular value decomposition of the scaled observation operator, the information content becomes separable between components associated with different singular values. Densely distributed observations can be then compressed with minimum information loss by truncating the components associate with the smallest singular values. The differences between the relative entropy and Shannon entropy difference in measuring information content and information loss are analysed in details and illustrated by examples.

1. Introduction

Analyses of irregularly distributed observations, especially large amounts of remotely sensed observations are common in geophysical research and play increasingly important roles in environmental monitoring and operational weather predictions (Daley, 1991; Bennett, 1992). Remotely sensed observations such as those from satellites and ground-based radars are characterized by their huge amounts and dense spatial and/or temporal distributions. Over their covered areas, remotely sensed observations are often much denser than the analyses grids. For example, the operational (WSR-88D) Doppler radar wind observations have typically a resolution of 250 m in the direction along the radar beam and a resolution of 1° in the azimuthal direction of the radar scan (Doviak and Zrnic, 1993), while the analysis grids used for current operational tests at the National Centers for Environmental Predictions (NCEP) have horizontal resolution in the range from 5 to 10 km (Liu et al., 2005b). A development plan to transmit level-II data real-time from all 158 WSR-88D radars and assimilate them into the operational numerical weather prediction (NWP) models has been implemented at the NCEP. Since each radar scans continuously every 5–10 min per volume and each volume scan can contain 10^5 observations, the amount of data received from all the radars over each 6-hr analysis time window is huge (up to $150 \times 36 \times 10^5$). Even though such a huge amount of data can be processed real time with quality controls (Gong et al., 2003; Zhang et al., 2005; Liu et al., 2005a), it is not feasible to assimilate all the radar data operationally. Since the spatial and temporal densities of radar data are far in excess of the resolution of the analysis systems, there can be a significant degree of information redundancy. It is not only unfeasible but also unnecessary to assimilate all the radar data with the current analysis systems. Similar situations are seen for satellite observations, although the information redundancy is caused mainly by linearly dependent weighting functions in case of passive satellite sounders (Peckham, 1974; Eyre, 1990; Huang and Purser, 1996).

Redundant observations not only impose unnecessary computational burdens on a data analysis system but can also cause the cost function used in the analysis ill-conditioned (especially in the presence of nearly collocated and/or correlated observations). To reduce or eliminate information redundancy from observations for data assimilation, it is important to know (i) how to measure the information content extracted from observations (or compressed observations) by an optimal (or presumably optimal) analysis and (ii) how to quantify the degree of information

*Correspondence
e-mail: Qin.Xu@noaa.gov
DOI: 10.1111/j.1600-0870.2006.00222.x

redundancy from observations. It is also necessary to address practical issues concerning (iii) how to compress observations (such as densely distributed radar data) to eliminate or reduce information redundancy with minimum possible loss of information and thus to minimize possible degradation of the otherwise 'optimal' analysis. These issues are not new, and similar issues were raised and examined in early studies of satellite observations of the atmosphere (Peckham, 1974; Eyre, 1990; Huang and Purser, 1996). These previous studies advocated the use of the Shannon entropy (Shannon, 1949) of probability density function (pdf) as a measure of information content and, in particular, they used the Shannon entropy difference between the analysis pdf and background pdf to measure the information content extracted from observations by the analysis. Purser et al. (2000) further refined the use of the Shannon entropy difference in quantifying the information content from observations, and proposed some general methods for constructing surrogate observations or, say, super-observations. As an extension these previous studies, the current study is intended to revisit the above issues by considering the relative entropy versus the Shannon entropy difference in terms of measuring information content from observations.

The Shannon entropy has some unique features in quantifying the uncertainty of a pdf (see section 15.1 of Papoulis, 1991; and section 6.2 of Majda and Wang, 2006). The original expression of Shannon entropy is not invariant with respect to a variable transformation and thus does not provide a consistent measure of the information content. The Shannon entropy difference is invariant with respect to a linear variable transformation but not to a nonlinear transformation. The relative entropy, also known as the Kullback–Liebler (non-symmetric) distance, is non-negative definite and invariant respect to any smooth invertible transformation of variables and thus provides a consistent measure of the information content of a pdf with respect to another pdf. The Shannon entropy difference is additive for successive inclusions of observations into the analyses as pointed out by James Purser (personal communication 1 Aug. 2006), but it measures only the dispersion part of the information content. The relative entropy is not additive and thus is less convenient than the Shannon entropy difference in computing cumulative information content from successive group of observations, but it measures both the signal and dispersion parts. Such a signal-dispersion combined measure has received considerable attention in the statistics literature (see Bernardo and Smith, 1994 for an overview). Recently, Kleeman (2002) and Majda et al. (2002) advocated the use of the relative entropy of the prediction and climatological pdfs as a measure of the utility of a particular statistical prediction. Advantages of using the relative entropy for predictability studies are further demonstrated in subsequent studies (Kleeman and Majda, 2005; Haven et al., 2005). Inspired by these previous studies, this study examines the use of the relative entropy, in comparison with the Shannon entropy difference, as a measure of the amount of information extracted from observations (or compressed super-observations) by an optimal analysis with a prior background pdf.

The paper is organized as follows. The Shannon entropy and relative entropy are reviewed briefly in the next section. The relative entropy is compared with the Shannon entropy difference in measuring the information content from observations in Section 3. The relative entropy is used to quantify possible information redundancy from observations and to measure information loss caused by super-observations in Section 4. Examples are given in Section 5 to illustrate the differences between the relative entropy and Shannon entropy difference in measuring information content and information loss. Conclusions follow in Section 6.

2. Shannon entropy and relative entropy

Consider a discrete pdf $p(x) = \sum p_i \delta(x - x_i)$ with $p_i \geq 0$ ($i = 1, 2, \ldots n$) and $\sum p_i = 1$, where $\delta(x - x_i)$ is the delta function at point x_i. The Shannon entropy of this pdf is expressed by

$$S(p) = -\langle \ln p \rangle = -\sum p_i \ln p_i, \tag{2.1}$$

where $\langle\rangle$ denotes the statistical average (expectation). According to Shannon's intuition from the theory of communication (Shannon, 1949), this entropy can be thought as a measure of how un-informative the pdf is about the state of $x \in \{x_i \mid i = 1, 2, \ldots I\}$. This intuition can be illustrated by a simple example as reviewed in section 6.2 of Majda et al. (2006). Consider the set of all binary data with digit length n. Clearly, this set has $N = 2^n$ elements, and $\log_2 N = n$ measures the amount of information needed to completely determine an element in the set, that is, a n-digit binary datum. The nature of this measure does not change when it is scaled by an constant factor such as ln 2. This implies that $\ln N = (\ln 2)\log_2 N$ can be used to quantify the total information needed to determine an element in a set A that contains N elements. If we know that an element of A belongs to the ith disjoint subset A_i that contains N_i elements of A, then $\ln N_i$ is the amount of information needed to determine this element. Thus, $\ln N - \ln N_i = -\ln p_i$ measures the lack of information relative to the total information $\ln N$. Note that $p_i = N_i/N$ is the probability of an element belonging to A_i, so the average lack of information is the Shannon entropy in (2.1).

When $p(x)$ is a continuous pdf of $x \in R$, the Shannon entropy is still expressed by $-\langle \ln p \rangle$ but takes the following integral form:

$$S(p) = -\langle \ln p \rangle = -\int dx\, p(x) \ln p(x). \tag{2.2}$$

Apart from an linear transformation with an arbitrary constant factor (such as ln 2), the logarithmic rule used in (2.2) is the unique impartial, symmetric, proper scoring rule (see sections 2.54–2.58 of O'Hagan, 1994). With this rule, $-\int dx p(x) \ln q(x)$ can measure the expected loss in using an arbitrary pdf $q(x)$ as an approximation of $p(x)$, and this measure is minimized at $q(x) = p(x)$ with the minimum given by the Shannon entropy of $p(x)$ (see section 15-4 of Papoulis, 1991). The Shannon entropy is

essentially unique as it increases monotonically with increasing uncertainty and satisfies the composition law (see section 6.2 of Majda et al., 2006, or section 15-1 of Papoulis, 1991). It is also known that the Shannon entropy is not invariant with respect to a smooth invertible transformation from $x \epsilon R$ to $x' \epsilon R$, because $p(x') = p(x)|dx/dx'|$ and this implies that $dxp(x)$ is invariant but $\ln p(x)$ is not in (2.2).

The relative entropy, also known as the Kullback–Liebler distance, is defined by

$$R(p, q) = \sum p_i \ln (p_i/q_i)$$

for discrete pdfs $p(x) = \sum p_i \delta(x - x_i)$ and $q = \sum q_i \delta(x - x_i)$ with $q_i \geq 0$ $(i = 1, 2, \ldots n)$ and $\sum q_i = 1$ similar to p_i in (2.1). In this case, if $q_i = 1/n$ is used as an approximation of $p = \sum p_i \delta(x - x_i)$ with $p_i \neq 1/n$, then the average lack of information is increased from $-\sum p_i \ln p_i$ to the maximum $-\sum q_i \ln q_i = \ln(n)$. This implies that q with $q_i = 1/n$ is a non-informative pdf and the 'absolute' information content of p can be defined by its relative entropy with respect to this non-informative pdf, that is, $R(p, q) = \sum p_i \ln(p_i/q_i) = \ln(n) - S(p)$ with $S(p)$ given by (2.1). However, since $S(q) = \ln(n) \to \infty$ as $n \to \infty$, the concept of non-informative pdf and associated 'absolute' information measure cannot be extended systematically to the infinite case ($n \to \infty$ or $x \epsilon R$). The amount of information is thus a relative concept in general and can be always properly measured by the relative entropy.

When $p(x)$ and $q(x)$ are continuous pdfs of $x \epsilon R$, the relative entropy takes the following integral form:

$$R(p, q) = \int dx\, p(x) \ln[p(x)/q(x)]. \tag{2.3}$$

This entropy provides a natural and consistent measure of the information content of p with respect to another pdf q which is considered as an approximation of p. Note that both $dxp(x)$ and $\ln[p(x)/q(x)]$ in (2.3) are invariant with respect to any smooth invertible variable transformation, and so is their composed integral – the relative entropy defined in (2.3). Note also that $R(p, q)$ is defined in (2.3) as the difference between the aforementioned expected loss, $-\int dxp(x) \ln q(x)$, and its minimum given by the Shannon entropy, so the relative entropy is non-negative definite and becomes zero if and only if $p = q$. Clearly, both $R(p, q)$ and $R(q, p)$ are positive but not equal to each other unless $p = q$, so the relative entropy is a non-symmetric measure of the distance between two pdfs in the function space. As a distance from q to p, $R(p, q)$ quantifies how informative q is about p. When q is considered as an approximation of p, this distance measures the information content of p with respect to q.

In the next section, the relative entropy will be used to measure the information content of the pdf produced by an optimal analysis of observations (or compressed super-observations) with respect to a prior background pdf used by the analysis. When observations are given with a pdf, the optimal analysis produces a posterior conditional pdf which is an improvement upon a prior background pdf. Thus, the background pdf can be always considered as an approximation of the analysis pdf. The information content extracted from the observations by an optimal analysis can be then measured by the relative entropy of the analysis pdf with respect to the background pdf. In this sense, the relative entropy measures indirectly the information content provided by the observations in terms of the changes produced in the pdf by the analysis. Clearly, with this measure, the information content from observations is not independent of the background pdf used by the analysis.

3. Information content from observations

When observations are assimilated into a NWP model by an optimal analysis scheme (Daley, 1991, Jazwinski, 1970), such as the statistical interpolation (often called the optimal interpolation or, simply OI), three-dimensional variational method (3dVar), or Kalman Filter (KF), the background mean field, denoted by vector $\mathbf{b} \epsilon R^n$, is provided by the prediction of the NWP model and the background pdf is assumed to be Gaussian with an pre-estimated (in OI or 3dVar) or predicted (in KF) covariance matrix, denoted by $\mathbf{B}$. With the predicted mean $\mathbf{b}$ and estimated covariance matrix $\mathbf{B}$, the background Gaussian pdf has the following general form:

$$q(x) = [(2\pi)^n \mathrm{Det}(\mathbf{B})]^{-1/2} \exp[-(\mathbf{x} - \mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x} - \mathbf{b})/2]. \tag{3.1a}$$

Since the observation pdf is also assumed to be Gaussian, the pdf of the analysed field (obtained with a linear or linearized observation operator) is thus also Gaussian and can be expressed by

$$p(x) = [(2\pi)^n \mathrm{Det}(\mathbf{A})]^{-1/2} \exp[-(\mathbf{x} - \mathbf{a})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{x} - \mathbf{a})/2], \tag{3.1b}$$

where $\mathbf{a}$ and $\mathbf{A}$ denote the mean and covariance of the analysis, respectively. Substituting (3.1) into the vector form of (2.3) gives (see appendix A)

$$R(p, q) = (\mathbf{a} - \mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{a} - \mathbf{b})/2 + [\ln \mathrm{Det}(\mathbf{B}^{1/2} \mathbf{A}^{-1} \mathbf{B}^{1/2}) + \mathrm{Tr}(\mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{-1/2}) - n]/2, \tag{3.2}$$

where Det() and Tr() denote the determinant and trace of (), respectively. The right-hand side of (3.2) consists of two parts. The first part is caused purely by the analysis increment, $\mathbf{a} - \mathbf{b}$, that updates the mean (from $\mathbf{b}$ to $\mathbf{a}$), and it is called the signal part of $R(p, q)$. The second part is caused purely by the change of the covariance (from $\mathbf{B}$ to $\mathbf{A}$) and is called the dispersion part of $R(p, q)$. Such a signal-dispersion partition can be generalized for non-Gaussian pdfs (see section 3 of Majda et al., 2002).

As mentioned in the introduction, the Shannon entropy difference between the analysis pdf and background pdf was used to measure the information content from observations (Peckham, 1974; Eyre, 1990 Huang and Purser, 1996; Purser et al., 2000). The Shannon entropy of $p(\mathbf{x})$ can be derived by

substituting (3.1b) into the vector form of (2.2) and the result is

$$S(p) = [-\ln \mathrm{Det}(\mathbf{A}) + n + n\ln(2\pi)]/2. \tag{3.3}$$

As explained in the previous section, the Shannon entropy is not invariant with respect to a smooth invertible transformation. As we can see from (3.3), even with respect to a linear transformation, say, from $\mathbf{x} \in R^{\mathrm{n}}$ to $\mathbf{x}' = \mathbf{L}\mathbf{x} \in R^{\mathrm{n}}$, lnDet($\mathbf{A}$) is changed to $\mathrm{lnDet}(\mathbf{L}\mathbf{A}\mathbf{L}^{\mathrm{T}}) = \mathrm{lnDet}(\mathbf{A}) + \mathrm{lnDet}(\mathbf{L}\mathbf{L}^{\mathrm{T}})$ according to (A.11) and hence $S(p)$ is altered by the amount of $[\mathrm{lnDet}(\mathbf{L}\mathbf{L}^{\mathrm{T}})]/2$. By using (3.3) and (A.11), the Shannon entropy difference between p and q can be expressed by

$$S(q) - S(p) = [\ln \mathrm{Det}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2})]/2. \tag{3.4}$$

This entropy difference is invariant with respect to a linear transformation but not to a nonlinear transformation [that transforms $q(\mathbf{x})$ and $p(\mathbf{x})$ from Gaussian to non-Gaussian or vice versa]. The Shannon entropy difference in (3.4) is the same as the first term in the dispersion part of the relative entropy in (3.2) and, clearly it does not have the signal part. On the other hand, the relative entropy is strictly invariant and measures both the signal and dispersion parts. A comprehensive comparison between the relative entropy and Shannon entropy difference was given in section 2.4 of Majda et al. (2002), and it shows that the relative entropy is superior to the Shannon entropy difference in quantifying predictive information content. In the case of predictability, there is a very clear interpretation of the signal term and in fact the terminology 'signal' derives from the interpretation. For the applications that concern this study, we need to re-examine the meaning and significance of the signal term in the present context and explore the advantages and disadvantages of the relative entropy versus the Shannon entropy difference.

The signal part of the relative entropy in (3.2) is a quadratic form of the analysis increment vector weighted by $\mathbf{B}^{-1}$. When an optimal analysis scheme (such as OI, 3dVar or KF) is used, this increment vector is given by

$$\mathbf{a} - \mathbf{b} = \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\mathbf{d}, \tag{3.5}$$

where $\mathbf{d} = \mathbf{y} - H(\mathbf{b})$ is the innovation vector, $\mathbf{y}$ is the observation vector (composed of m observations used by the analysis) in the observation space R^{m}, $H()$ denotes the observation operator that transforms the state vector from the background space R^{n} to the observation space R^{m}, $\mathbf{H}$ is the linearized $H()$ at $\mathbf{x} = \mathbf{b}$, and $\mathbf{R}$ is the covariance matrix of the observation pdf. Substituting (3.5) into the signal part of the relative entropy in (3.2) gives

$$\begin{aligned}
&(\mathbf{a}-\mathbf{b})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{a}-\mathbf{b})/2 \\
&\quad = \mathbf{d}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{T} + \mathbf{R})^{-1}\mathbf{d}/2 \\
&\quad = \mathbf{d}^{\mathrm{T}}\mathbf{R}^{-1/2}\left(\mathbf{M}\mathbf{M}^{\mathrm{T}} + \mathbf{I}_m\right)^{-1}\mathbf{M}\mathbf{M}^{\mathrm{T}}\left(\mathbf{M}\mathbf{M}^{\mathrm{T}} + \mathbf{I}_m\right)^{-1}\mathbf{R}^{-1/2}\mathbf{d}/2 \\
&\quad = \mathbf{d}'^{\mathrm{T}}\left(\mathbf{\Lambda}^2 + \mathbf{I}_m\right)^{-1}\mathbf{\Lambda}^2\left(\mathbf{\Lambda}^2 + \mathbf{I}_m\right)^{-1}\mathbf{d}'/2 \\
&\quad = \sum d'^2_i\lambda_i^2\left(1+\lambda_i^2\right)^{-2}/2.
\end{aligned} \tag{3.6}$$

Here, $\mathbf{I}_{\mathrm{m}}$ is the $m\times m$ identity matrix; $\mathbf{M} \equiv \mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}^{1/2}$ is a $m\times n$ matrix for the scaled observation operator (with the range and domain of $\mathbf{H}$ scaled by $\mathbf{R}^{1/2}$ and $\mathbf{B}^{-1/2}$, respectively); $\mathbf{\Lambda} \equiv \mathrm{diag}\{\lambda_1, ..., \lambda_\mu\} = \mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{V}$ is a diagonal matrix composed of the singular values of $\mathbf{M}$ with $\boldsymbol{\lambda}_1 \geq \lambda_2 \geq \ldots \geq \lambda_\mu \geq 0$ and $\mu \equiv \min(m, n)$; $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrix composed of the left and right singular vectors of $\mathbf{M}$, respectively (see theorem 2.3-1 of Golub and Van Loan, 1983); $\mathbf{d}' = \mathbf{U}^{\mathrm{T}}\mathbf{R}^{-1/2}\mathbf{d}$; d'_{i} denotes the ith element of $\mathbf{d}'$; $\sum$ denotes the summation over i from 1 to μ.

For the optimal analysis in (3.5), the covariance matrix of the analysis pdf is given by the following two equivalent forms (see chapter 7 of Jazwinski, 1970):

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}, \tag{3.7a}$$

$$\mathbf{A} = \mathbf{B} - \mathbf{B}\mathbf{H}^{T}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\mathbf{H}\mathbf{B}. \tag{3.7b}$$

Either of the two forms can be used to analyse the dispersion part of the relative entropy in (3.2), and they yield the same result. The first form in (3.7a) was used by Peckham (1974) and Huang and Purser (1996) and the second form in (3.7b) was used by Purser et al. (2000) to analyse the Shannon entropy difference in (3.4). Here, it is convenient to use (3.7a) and (3.7b) to analyse the two matrix terms $\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}$ and $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, respectively, in the dispersion part of the relative entropy in (3.2). Substituting (3.7a) into $\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}$ gives

$$\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2} = \mathbf{I}_n + \mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{I}_n + \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{\mathrm{T}}, \tag{3.8a}$$

where $\mathbf{I}_{\mathrm{n}}$ the $n\times n$ identity matrix. Substituting (3.7b) into $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ gives

$$\begin{aligned}
\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} &= \mathbf{I}_n - \mathbf{M}^{\mathrm{T}}\left(\mathbf{M}\mathbf{M}^{\mathrm{T}} + \mathbf{I}_m\right)^{-1}\mathbf{M} \\
&= \mathbf{I}_n - \mathbf{V}\mathbf{\Lambda}\left(\mathbf{\Lambda}^2 + \mathbf{I}_m\right)^{-1}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}.
\end{aligned} \tag{3.8b}$$

Substituting (3.8a) into (3.4) gives

$$S(q) - S(p) = \ln \mathrm{Det}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2})]/2 = \sum \ln(1 + \lambda_{\mathrm{i}}^2)/2. \tag{3.9}$$

Substituting (3.8a) and (3.8b) into the dispersion part of the relative entropy in (3.2) gives

$$\begin{aligned}
&[\ln \mathrm{Det}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}) + \mathrm{Tr}(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}) - n]/2 \\
&\quad = \sum\left[\ln\left(1 + \lambda_i^2\right) - \lambda_i^2\left(1 + \lambda_i^2\right)^{-1}\right]/2.
\end{aligned} \tag{3.10}$$

The singular-value form of the Shannon entropy difference in (3.9) is essentially the same as the eigenvalue forms presented in the literature [see section 3 of Peckham, 1974, (2.12) of Huang and Purser, 1996, or (4.14) of Purser et al., 2000]. Note that both $\ln(1 + \lambda_i^2)$ and $\ln(1 + \lambda_i^2) - \lambda_i^2(1 + \lambda_i^2)^{-1}$ become zero at $\lambda_{\mathrm{i}} = 0$ and increase monotonically with λ_{i}, so either (3.9) or (3.10) can be used as a measure of the dispersion part of the information content from observations. The derivative of (3.9) with respect to λ_i is one at $\lambda_i = 0$. The derivative of (3.10) with respect to λ_i is zero at $\lambda_i = 0$, so the ith term in (3.10) has a smooth zero

minimum at $\lambda_i = 0$. Thus, as λ_i approaches to zero, the ith term in (3.10) decreases faster than that in (3.9). This difference is also illustrated by the examples in Section 5 (see Dsi and SDi curves in Figs. 2 and 4).

As shown by (3.9) [also see (3.7a) and (A.11)], the Shannon entropy difference is additive for successive inclusions of observations into the analyses. The relative entropy is not additive and thus is not as convenient as the Shannon entropy difference in computing cumulative information content from successive groups of observations. The major difference between the two measures, however, is the signal term in (3.6). As shown by (3.2) and (3.6), this term measures the (signal) part of the information content from observations that improves the mean through the analysis. When the innovation vector happens to be zero [that is, $\mathbf{d} = \mathbf{y} - H(\mathbf{b}) = 0$], the analysis has the same mean as the background (that is, $\mathbf{a} = \mathbf{b}$), so the mean is not improved and the signal part becomes zero. The error covariance of the analysis, however, is always improved by observations (unless $\mathbf{M} = \mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}^{1/2} = 0$ which means that the observations are infinitely inaccurate relative to the background and thus become useless). The dispersion term in (3.10) or the Shannon entropy difference in (3.9) measures the (dispersion) part of the information content from observations that improves the covariance through the analysis. This part depends on the observation operator but is independent of the observation vector and associated innovation vector. The significance and utilities of these terms in measuring information loss caused by super-observations are examined in the next two sections.

4. Information redundancy and information loss

The sum of (3.6) and (3.10) gives the following singular-value form of the relative entropy:

$$R(p, q) = \sum_r \left[d_i'^2\lambda_i^2\left(1+\lambda_i^2\right)^{-2} + \ln\left(1+\lambda_i^2\right) - \lambda_i^2\left(1+\lambda_i^2\right)^{-1}\right]/2, \tag{4.1}$$

where $\sum_r$ denotes the summation over i from 1 to r, and $r = \text{rank}(\mathbf{\Lambda}) = \text{rank}(\mathbf{M}) \le \mu \equiv \min(m, n)$. Since $\lambda_i = 0$ for $i > r$, only the first r components of $\mathbf{d}'$ can contribute to the signal part in (3.6), and the original equation (3.5) can be rewritten into

$$\begin{aligned}\mathbf{a} - \mathbf{b} &= \mathbf{B}^{1/2}\mathbf{M}^{\mathrm{T}}\left(\mathbf{M}\mathbf{M}^{\mathrm{T}} + \mathbf{I}_m\right)^{-1}\mathbf{R}^{-1/2}\mathbf{d} \\ &= \mathbf{B}^{1/2}\mathbf{V}\mathbf{\Lambda}(\mathbf{\Lambda}^2 + \mathbf{I}_m)^{-1}\mathbf{d}' \\ &= \mathbf{B}^{1/2}\mathbf{V}\Lambda\left(\mathbf{\Lambda}^2 + \mathbf{I}_j\right)^{-1}\mathbf{d}'_j \quad \text{for } m \ge j = r,\end{aligned} \tag{4.2}$$

where $\mathbf{I_j}$ is the $j\times j$ identity matrix, $\mathbf{d}'_j$ is a truncated vector composed of the first j components of $\mathbf{d}'$, and j is the truncation number. There will be no truncation if $j = m$. Applying $\mathbf{I}_j$ (as a projection onto the truncated subspace R^s) to $\mathbf{d}'$ gives $\mathbf{d}'_j = \mathbf{I}_j\mathbf{d}' = \mathbf{I}_j\mathbf{U}^{\mathbf{T}}\mathbf{R}^{-1/2}\mathbf{d} = \mathbf{U}_j^{\mathrm{T}}\mathbf{R}^{-1/2}\mathbf{d}$, where $\mathbf{U}\mathbf{I}_j = \mathbf{U}_j \equiv (\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_j)$ and $\mathbf{u}_i$ denotes the ith column vector of $\mathbf{U}$. Hence, $\mathbf{d}'_j$ can be viewed as a truncated linear transformation of the scaled innovation vector $\mathbf{R}^{-1/2}\mathbf{d}$ and the transformation is implemented by $\mathbf{I}_j\mathbf{U}^{\mathrm{T}}$. We may call $\mathbf{d}'_j$ the super-innovation vector, $\mathbf{y}'_j \equiv \mathbf{U}_j^{\mathrm{T}}\mathbf{R}^{-1/2}\mathbf{y}$ the super-observation vector, $H'_j() \equiv \mathbf{U}_j^{\mathrm{T}}\mathbf{R}^{-1/2}H()$ the super-observation operator, and $\mathbf{H}'_j \equiv \mathbf{U}_j^{\mathrm{T}}\mathbf{R}^{-1/2}\mathbf{H}$ the linearized super-observation operator. Here, the truncated linear transformation $\mathbf{I}_j\mathbf{U}^{\mathrm{T}}$ is applied consistently to all the vectors and operators in the observation space scaled by $\mathbf{R}^{1/2}$, so $\mathbf{d}'_j = \mathbf{y}'_j - H'_j(\mathbf{b})$. It is easy to see that the scaled observation covariance matrix is given by $\mathbf{R}^{-1/2}\mathbf{R}\mathbf{R}^{-1/2} = \mathbf{I}_{\mathrm{m}}$ and the super-observation covariance matrix is given by $\mathbf{R}'_j = \mathbf{I}_j\mathbf{U}^{\mathrm{T}}\mathbf{I}_{\mathrm{m}}\mathbf{U}\mathbf{I}_j = \mathbf{I}_j$.

By using the above super-observation notations, (3.5) or (4.2) can be further rewritten into

$$\mathbf{a}_j = \mathbf{b} + \mathbf{B}\mathbf{H}'^{\mathrm{T}}_j\left(\mathbf{H}'_j\mathbf{B}\mathbf{H}'^{\mathrm{T}}_j + \mathbf{R}'_j\right)^{-1}\mathbf{d}'_j. \tag{4.3}$$

This equation has the same form as the original (3.5) except that the original-observation notations $\mathbf{y}$, $\mathbf{R}$, $H()$ and $\mathbf{H}$ are replaced by their respective super-observation counterparts and thus $\mathbf{a}$ is replaced by $\mathbf{a}_{\mathrm{j}}$. Similarly, (3.7a) and (3.7b) can be rewritten, respectively, into

$$\mathbf{A}_j^{-1} = \mathbf{B}^{-1} + \mathbf{H}'^{\mathrm{T}}_j\mathbf{R}'^{-1}_j\mathbf{H}'_j \tag{4.4a}$$

and

$$\mathbf{A}_j = \mathbf{B} - \mathbf{B}\mathbf{H}'^{\mathrm{T}}_j\left(\mathbf{H}'_j\mathbf{B}\mathbf{H}'^{\mathrm{T}}_j + \mathbf{R}'_j\right)^{-1}\mathbf{B}\mathbf{H}'_j. \tag{4.4b}$$

These two equations have the same forms as those in (3.7) except that the original-observation notations are replaced by their respective super-observation counterparts and thus $\mathbf{A}$ is replaced by $\mathbf{A}_j$. When $j \ge r$, $\mathbf{a}_j$ in (4.3) is exactly the same as $\mathbf{a}$ in (3.5) and $\mathbf{A}_j$ in (4.4) is exactly the same as $\mathbf{A}$ in (3.7). When $j < r$, (4.3) and (4.4) still give the optimally analysed mean and covariance, although they are no longer the same as their (non-truncated) counterparts in (3.5) and (3.7).

The results in (4.3) and (4.4) indicate that the truncated linear transformation $\mathbf{I_j}\mathbf{U}^{\mathbf{T}}$ will cause no information loss as long as $j \ge r$. Thus, the degree of information redundancy can be quantified by $(m - r)/m$ in terms of reducible percentage, and $m-r$ is the reducible number of observations. Clearly, if m is larger than n, then the degree of information redundancy will be at least as large as $(m - n)/m$. Note that rank($\mathbf{M}$) is the dimension of the range of $\mathbf{M}$ and the range of $\mathbf{M}$ is the complement of the null space of $\mathbf{M}^{\mathrm{T}}$ in R_{m} (see section 1.2 of Golub and Van Loan, 1983), so the dimension of the null space of $\mathbf{M}^{\mathrm{T}}$ is $m - \text{rank}(\mathbf{M}) = m - r$ and hence gives the reducible number of observations. If the observations are sufficiently dense (relative to the background resolution) and the background covariance is local or virtually local (becomes zero or or virtually zero beyond a certain range of spatial separation), then the null space of $\mathbf{M}^{\mathrm{T}}$ can be non-empty and the observations can be redundant (even if $m \le n$). In this case ($m \le n$), $n - r$ is the dimension of the null space of $\mathbf{M}$. If the observations are locally distributed and the background

covariance is local or virtually local, then the dimension of the null space of $\mathbf{M}$, that is, $n - r$ can be very large or even close to n. If the observations are also dense in this case, then r can be smaller than m. This is another scenario that can cause redundant observations (even if $m <$ or $\ll n$).

If the truncation number j becomes smaller than r, then the above truncated linear transformation $\mathbf{I}_j\mathbf{U}^{\mathrm{T}}$ will cause an information loss according to (4.1)–(4.4). The information loss (IL) can be quantified by

$$\mathrm{IL}_j = \mathrm{SIL}_j + \mathrm{DIL}_j \equiv \sum_{j,r} \left[\mathrm{d}'^2_i \lambda_i^2 \left(1+\lambda_i^2\right)^{-2} \right] /2 + \sum_{j,r} \left[\ln\left(1+\lambda_i^2\right) - \lambda_i^2 \left(1+\lambda_i^2\right)^{-1} \right] /2, \tag{4.5}$$

where $\sum_{j,r}$ denotes the summation over i from $j+1$ to r. The first summation is the signal part of IL_j, called signal information loss and denoted by SIL_j, while the second summation is the dispersion part of IL_j, called dispersion information loss and denoted by DIL_j. The signal information loss SIL_j depends not only on the truncated non-zero singular values but also on the truncated components of $\mathbf{d}'_{j,r} \equiv (d'_{j+1}, d'_{j+2}, \ldots d'_{\mathrm{r}})^{\mathrm{T}}$. Here, $\mathbf{d}'_{j,r} = \mathbf{U}^{\mathrm{T}}_{j,r}\mathbf{R}^{-1/2}\mathbf{d}$ is the projection of the scaled innovation vector $\mathbf{R}^{-1/2}\mathbf{d}$ onto the truncated subspace spanned by $\mathbf{U}_{j,r} \equiv (\mathbf{u}_{j+1}, \mathbf{u}_{j+2}, \ldots \mathbf{u}_{\mathrm{r}})$. If $\mathbf{R}^{-1/2}\mathbf{d}$ happens to be orthogonal to $\mathbf{U}_{j,r}$, then $\mathbf{d}'_{j,r} = 0$ and thus $\mathrm{SIL}_j = 0$ according to (4.5). On the other hand, if $\mathbf{R}^{-1/2}\mathbf{d}$ happens to be completely in $\mathbf{U}_{j,r}$, then $\mathbf{d}' = \mathbf{d}'_{j,r}$ and SIL_j is maximized to the total amount in (3.6). As we will see in the next section, SIL_j measures the information loss that degrades the analysis mean.

The dispersion information loss DIL_j depends only on the truncated non-zero singular values and it measures the information loss that degrades the analysis covariance. Since the truncation is made to the smallest non-zero singular values in the transformed observation space, it causes the minimum information loss in the dispersion part for a given truncation number j ($<r$). This information loss can be used as a benchmark to evaluate possible additional information loss in the dispersion part caused by any other truncated linear transformations (versus $\mathbf{I}_j\mathbf{U}^{\mathrm{T}}$) used in producing super-observations. From (3.9), it is easy to see that the information loss measured by the Shannon entropy difference is

$$\mathrm{SDIL}_j \equiv \sum\nolimits_{j,r} \left[\ln\left(1+\lambda_i^2\right) \right] /2. \tag{4.6}$$

Since SDIL_j also measures the dispersion information loss that degrades the analysis covariance, SDIL_j is similar to DIL_j but different from SIL_j. The difference between SDIL_j and SIL_j is thus the main difference between SDIL_j and IL_j. This main difference will be illustrated by examples in the next section.

5. Illustrative examples

5.1. Direct measures of analysis mean degradation and covariance degradation

In this section, radar observed velocities and model produced background velocity fields are used to illustrate the differences between the relative entropy and Shannon entropy difference in measuring information content and information loss. To evaluate how closely SIL_j measures the information loss that degrades the analysis mean caused by the truncation in (4.3), we quantify the analysis mean degradation (MD) by

$$\mathrm{MD}_j \equiv |\Delta \boldsymbol{a}_j|, \tag{5.1}$$

where $\Delta\mathbf{a}_j = \mathbf{a}_j - \mathbf{a}$, $\mathbf{a}_j$ is given in (4.3), and $|\Delta\mathbf{a}_j|$ denotes the absolute value of vector $\Delta\mathbf{a}_j$ (which is also the l_2-norm of $\Delta\mathbf{a}_j$). To evaluate how closely DIL_j (or SDIL_j) measures the information loss that degrades the analysis covariance caused by the truncation in (4.4), we quantify the covariance degradation (CD) by

$$\mathrm{CD}_j \equiv ||\Delta\mathbf{A}_j||_{\mathrm{F}}, \tag{5.2}$$

where $\Delta\mathbf{A}_j = \mathbf{A}_j - \mathbf{A}$, $\mathbf{A}_j$ is given in (4.4), and $||\Delta\mathbf{A}_j||_{\mathrm{F}}$ denotes the Frobenius norm of $\Delta\mathbf{A}_j$ defined by the square root of the sum of the squared absolute values of all the elements in $\Delta\mathbf{A}_j$ [see (2.2-4) of Golub and Van Loan, 1983]. As functions of the truncation number j, SIL_j and MD_j are expected to have similar variations with j, because they both measure the degradation of the mean. Similarly, DIL_j (or SDIL_j) and CD_j are expect to have similar variations with j, because they both measure the degradation of the covariance. These similarities will be illustrated by the examples in Sections 5.3 and 5.4.

5.2. Descriptions of the data

The observational data are selected from the radial-component velocities scanned by the NSSL phased array radar from 2100 to 2200 UTC when a four-quadrant electronic-scan strategy was tested on 2 June 2004. During this period, a squall line moved southeastward through the central Oklahoma area in the radial range (140 km) of the phased array radar scans (see Fig. 1 of Xu et al., 2005). The original radar data have a spatial resolution of 240 m in the radial direction and 1.6° in the azimuthal direction. The data are processed through quality control (as in Xu et al., 2005) and then thinned to 3 km resolution along the radar beam (see Fig. 1). The thinned observations are not correlated and the estimated observation error variance is $\sigma_{\mathrm{o}}^{b2} = 6.4\ \mathrm{m}^2\ \mathrm{s}^{-2}$ according to Xu et al. (2005). For the illustrative purpose in this section, a single beam of radial-velocity observations is used. This beam was scanned at 2108 UTC along 0.75° elevation angle and 97.8° azimuthal angle (positive clockwise from the north), and it contain 40 thinned observations. The observation error

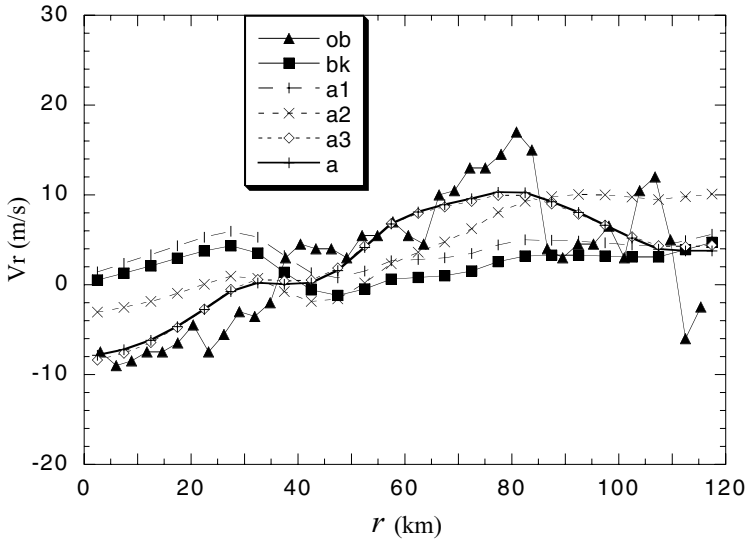

Fig. 1. Thinned radial-velocity observations (with $m = 40$ and 3 km resolution) along a single beam from the radar (shown by the thin solid curve with solid triangles as denoted by 'ob' in the legend); background radial-velocity field on an one-dimensional grid of 5 km resolution along the radar beam with $n = 24$ grid points (shown by the thin solid curve with solid squares as denoted by 'bk' in the legend); optimally analysed field by using the total 40 observations, that is, **a** in (3.5) (shown by the thick solid curve with '+' signs as denoted by 'a' in the legend); optimally analysed fields by using super-observations, that is, $\mathbf{a}_j$ in (4.3) with $j = 1$, 2 and 3 (shown by the thin dashed curves as denoted by 'a1', 'a2' and 'a3', respectively, in the legend). The horizontal coordinate r is the radial distance from the radar.

covariance matrix is thus given by $\mathbf{R} = \sigma_o^{b2}\mathbf{I}_m$ with $m = 40$ (the number of observations).

The background field is produced by the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS, Hodur 1997). The model is configured with three nested domains centred over the state of Oklahoma with resolutions of 54, 18 and 6 km for the coarse, medium and fine grids, respectively, and 30 levels in the vertical. The predicted wind fields on the 6 km grid are projected onto the aforementioned radar beam to obtain the background field on an one-dimensional grid of 5 km resolution with $n = 24$ grid points (see Fig. 1). Since the rotational and divergent parts of the estimated background vector velocity error variance are roughly the same (see section 4 of Xu et al., 2005), the background radial-velocity error covariance in the above one-dimensional grid space along the radar beam can be modeled approximately by the Gaussian function, that is, $\sigma^2\exp[-(\Delta r)^2/(2L^2)]$ according to (2.7) and (3.2) of Xu and Gong (2003), where Δr is the distance between two correlated points (along the radar beam), σ^2 and L denote the background radial-velocity error variance and de-correlation length scale, respectively. The estimated variance is $\sigma^2 = 70\ \mathrm{m^2\ s^{-2}}$ and the estimated de-correlation length is $L = 40$ km (see section 4 of Xu et al., 2005). The above data and parameter values will be used by the first example-1 in the next section. These data and parameter values will be also used by the second example in Section 5.4 except that the de-correlation length is reduced from $L = 40$ to 15 km.

5.3. Example-1

By using the data and parameter values described in the previous section, $\mathbf{M} \equiv \mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}^{1/2}$ is constructed and decomposed into $\mathbf{U}\Lambda\mathbf{V}^{\mathrm{T}}$. The largest singular value of $\mathbf{M}$ is $\lambda_1 = 7.45$. The remaining singular values are scaled by λ_1 and plotted as a function of i in Fig. 2, where

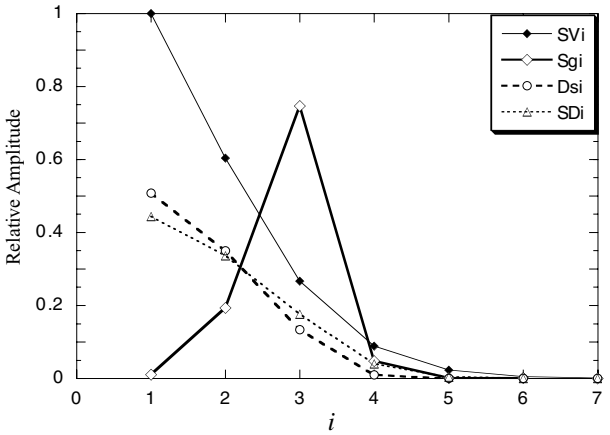

Fig. 2. Scaled singular value SVi in (5.3), signal term Sgi in (5.4), dispersion term Dsi in (5.5) and Shannon entropy difference term SDi in (5.6) plotted as functions of i for example-1, where i is the sequential number associated with the ith the singular value. The curves are denoted by their respective symbols in the legend.

$$\mathrm{SV}_i \equiv \lambda_i/\lambda_1 \tag{5.3}$$

denotes the ith scaled singular value. Individual signal terms in (3.6), dispersion terms in (3.9) and Shannon entropy difference terms in (3.10) are scaled by their respective sums and plotted as functions of i in Fig. 2, where

$$\mathrm{Sg}_i \equiv d'^2_i\lambda_i^2\left(1+\lambda_i^2\right)^{-2}/(2Sg),\ \ Sg \equiv \sum[d'^2_i\lambda_i^2\left(1+\lambda_i^2\right)^{-2}]^{-1}/2; \tag{5.4}$$

$$\mathrm{Ds}_i \equiv \left[\ln\left(1+\lambda_i^2\right) - \lambda_i^2\left(1+\lambda_i^2\right)^{-1}\right]/(2\mathrm{Ds}),$$
$$\mathrm{Ds} \equiv \sum\left[\ln(1+\lambda_i^2) - \lambda_i^2\left(1+\lambda_i^2\right)^{-1}\right]/2; \tag{5.5}$$

$$\mathrm{SD}_i \equiv \left[\ln\left(1+\lambda_i^2\right)\right]/(2\mathrm{SD}), \quad \mathrm{SD} \equiv \sum\left[\ln\left(1+\lambda_i^2\right)\right]/2. \tag{5.6}$$

As shown in Fig. 2, SV_i decreases monotonically to nearly zero ($<10^{-3}$) as i increases from 1 to 7. This indicates that the rank of $\mathbf{M}$ can be virtually as small as 6. With $r = 6$, the reducible number of observations is $m - r = 34$ and the degree of information redundancy is $(m - r)/m = 85\%$, so the observations are highly redundant even though they have been thinned from 512 to 40. Fig. 2 also shows that as i increases from 1 to 7, SD_i decreases monotonically to virtually zero (10^{-4} at $i = 6$ and 5×10^{-6} at $i = 7$), and Dsi decreases even more rapidly (to 10^{-7} at $i = 6$ and 5×10^{-9} at $i = 7$). The signal term Sg_i, however, increases rapidly and reaches the peak value of 0.75 as i increases from 1 to 3, and then drops very rapidly to virtually zero (4×10^{-4} at $i = 6$ and 5×10^{-7} at $i = 7$) as i increases further to 7. For the dispersion part measured by Ds_i (or SD_i), the first two terms ($i = 1$ and 2) contain most (>80%) of the information content. For the signal part, the third term contains most (75%) of the information content. The computed values for the sums in (5.4)–(5.6) are Sg = 14.86, Ds = 3.01 and SD = 4.54. The total information content measured by the relative entropy in (4.1) is thus given by $R(p, q) = \mathrm{Sg} + \mathrm{Ds} = 17.87$. As $\mathrm{Sg} = 14.86 \gg \mathrm{Ds} = 3.01$, the signal part is much larger than the dispersion part and thus contributes dominantly to the total.

When the truncation number j is zero, all the terms in (4.5) and (4.6) are truncated, so SIL_j and DIL_j in (4.5) and SDIL_j in (4.6) reach their maxima, that is, $\mathrm{SIL}_0 = \mathrm{Sg}$, $\mathrm{DIL}_0 = \mathrm{Ds}$ and $\mathrm{SDIL}_0 = \mathrm{SD}$, respectively, according to (5.4)–(5.6). When $j = 0$, the analysis mean degradation MD_j in (5.1) and covariance degradation CD_j in (5.2) also reach their maxima $\mathrm{MD}_0 \equiv |\Delta\mathbf{a}_0| = |\mathbf{b} - \mathbf{a}|$ and $\mathrm{CD}_0 \equiv ||\Delta\mathbf{A}_0||_\mathrm{F} = ||\mathbf{B} - \mathbf{A}||_\mathrm{F}$, respectively, where $\mathbf{a}_0 = \mathbf{b}$ and $\mathbf{A}_0 = \mathbf{B}$ are used according to (4.3) and (4.4). In Fig. 3, SIL_j, DIL_j and SDIL_j are scaled by their respective maxima and plotted as functions of the truncation number j in comparison with the scaled analysis mean degradation $\mathrm{MD}_j/\mathrm{MD}_0$ and covariance degradation $\mathrm{CD}_j/\mathrm{CD}_0$. As shown in Fig. 3, the scaled signal information loss $\mathrm{SIL}_j/\mathrm{Sg}$ varies with j in the same way as the scaled analysis mean degradation $\mathrm{MD}_j/\mathrm{MD}_0$. Thus, as expected, SIL_j does indeed measure the information loss that degrades the analysis mean. As j increases, the scaled dispersion information loss $\mathrm{DIL}_j/\mathrm{Ds}$ decreases smoothly following the scaled covariance degradation $\mathrm{CD}_j/\mathrm{CD}_0$. Thus, as expected, DIL_j does measure the information loss that degrades the analysis covariance. The scaled Shannon entropy difference information loss, $\mathrm{SDIL}_j/\mathrm{SD}$, also decreases smoothly following $\mathrm{CD}_j/\mathrm{CD}_0$ but not as closely as $\mathrm{DIL}_j/\mathrm{Ds}$. Clearly, $\mathrm{SDIL}_j/\mathrm{SD}$ is larger than $\mathrm{DIL}_j/\mathrm{Ds}$ while the latter is larger than $\mathrm{CD}_j/\mathrm{CD}_0$.

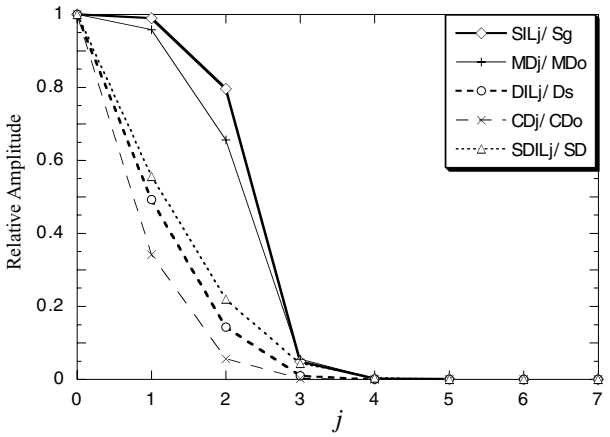

Fig. 3. Scaled signal information loss $\mathrm{SIL}_j/\mathrm{Sg}$, dispersion information loss $\mathrm{SIL}_j/\mathrm{Ds}$ and Shannon entropy difference information loss $\mathrm{SDIL}_j/\mathrm{SD}$ plotted as functions of the truncation number j in comparison with the scaled analysis mean degradation $\mathrm{MD}_j/\mathrm{MD0}$ and the scaled covariance degradation $\mathrm{CD}_j/\mathrm{CD}_0$ for example-1. The curves are denoted by their respective symbols in the legend, and the symbols are defined in (4.5)–(4.6) and (5.1)–(5.2).

Note again that $\mathrm{Sg} = 14.86 \gg \mathrm{SD} = 4.54 > \mathrm{Ds} = 3.01$, so the total information loss IL_j measured by the relative entropy in (4.5) is dominated by its signal part SIL_j. When $j \geq 3$, $\mathrm{SDIL}_j/\mathrm{SD}$ is very close to $\mathrm{IL}_j/(\mathrm{Sg} + \mathrm{Ds}) \approx \mathrm{SIL}_j/\mathrm{Sg}$, so the Shannon entropy difference and relative entropy both indicate that the information loss is small as long as $j > 2$. However, when $j = 2$, $\mathrm{SIL}_j/\mathrm{Sg} = 0.80$, $\mathrm{IL}_j/(\mathrm{Sg} + \mathrm{Ds}) = 0.69$ and $\mathrm{SDIL}_j/\mathrm{SD} = 0.22$. In this case, the relative entropy indicates that the information loss is large (in the signal part), but the Shannon entropy difference indicates that the information loss is still small (because it does not measure the signal part). With $j = 2$, the scaled analysis mean degradation is $\mathrm{MD}_j/\mathrm{MD}_0 = 0.66$ and the scaled covariance degradation is $\mathrm{CD}_j/\mathrm{CD}_0 = 0.06$. The analysis mean degradation caused by the truncation to $j = 2$ (or 1) is also clearly illustrated by the deviation of the truncated super-observation analysis $\mathbf{a}_2$ (or $\mathbf{a}_1$) from the total-observation analysis $\mathbf{a}$ in Fig. 1.

5.4. Example-2

This second example is the same as the above first example except that the background error de-correlation length is reduced to $L = 15$ km. In response to the reduction of L, the largest singular value of $\mathbf{M}$ is reduced from $\lambda_1 = 7.45$ to 5.24 and the virtual rank of $\mathbf{M}$ is increased from 6 to 14 (since $\mathrm{SV}_i \leq 10^{-3}$ as $i \geq 14$ in this case). The reducible number of observations is thus $m - r = 40 - 14 = 26$, and the degree of information redundancy is $(m - r)/m = 65\%$ which is still high but not as high as in example-1. As shown in Fig. 4, when i increases from 1 to 15, SD_i decreases to virtually zero (1.7×10^{-6} at $i = 15$), and Ds_i decreases more rapidly (to 6.8×10^{-7} at $i = 12$). The signal term, Sg_i, reaches the first peak value of 0.34 as i increases from 1 to

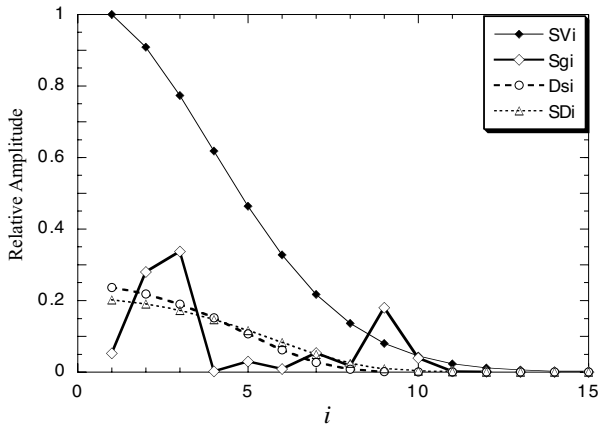

Fig. 4. As in Fig. 2 but for example-2.

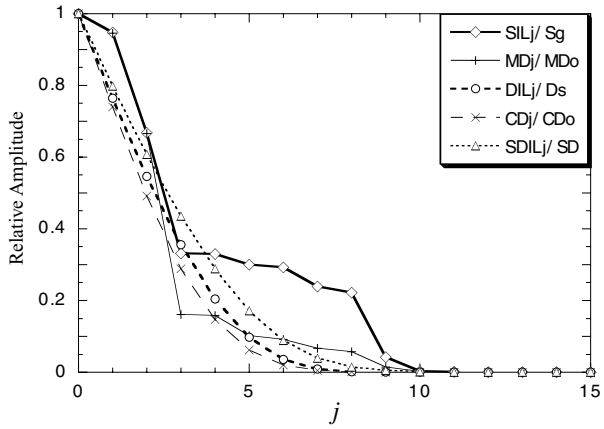

Fig. 5. As in Fig. 3 but for example-2.

3, then drops to nearly zero (0.002) at $i = 4$ and stays below 0.06 until i increases to 9 and Sg_i reaches the second main peak value (0.18 at $i = 9$). The computed values for the sums in (5.4)–(5.6) are Sg = 9.93, Ds = 5.04, and SD = 8.29. The total information content measured by the relative entropy is thus given by $R(p, q) = \mathrm{Sg} + \mathrm{Ds} = 14.97$. The signal part is still larger than the dispersion part but not as much as in example-1.

Figure 5 shows that the scaled signal information loss $\mathrm{SIL}_j/\mathrm{Sg}$ varies with j basically in the same way as the scaled analysis mean degradation $\mathrm{MD}_j/\mathrm{MD}_0$, while the scaled dispersion information loss $\mathrm{DIL}_j/\mathrm{Ds}$ decreases as j increases in the same way as the scaled analysis covariance degradation $\mathrm{CD}_j/\mathrm{CD}_0$. Thus, again as expected, SIL_j measures the information loss that degrades the analysis mean, while DIL_j measures the information loss that degrades the analysis covariance. The scaled Shannon entropy difference information loss $\mathrm{SDIL}_j/\mathrm{SD}$ also decreases following $\mathrm{CD}_j/\mathrm{CD}_0$ but not as closely as $\mathrm{DIL}_j/\mathrm{Ds}$. Since $\mathrm{Sg} = 9.93 > \mathrm{SD} = 8.29 > \mathrm{Ds} = 5.04$, the total information loss measured by IL_j in (4.5) is still dominated by its signal part SIL_j.

In Fig. 5, the Shannon entropy difference and relative entropy both indicate that the information loss is nearly zero (<0.003) for $j \geq 10$. The two measures, however, start to become different when j decreases to 9, and in this case $\mathrm{IL}_j/(\mathrm{Sg} + \mathrm{Ds}) \approx \mathrm{SIL}_j/\mathrm{Sg} = 0.04$ but $\mathrm{SDIL}_j/\mathrm{SD} = 0.004$. When j decreases further to 8 and then to 5, $\mathrm{SIL}_j/\mathrm{Sg}$ becomes significantly larger than $\mathrm{SDIL}_j/\mathrm{SD}$. In this case, although $\mathrm{SDIL}_j/\mathrm{SD}$ is closer to $\mathrm{MD}_j/\mathrm{MD}_0$ than $\mathrm{SIL}_j/\mathrm{Sg}$, its variation (with j) does not follow the variation of $\mathrm{MD}_j/\mathrm{MD}_0$, and the reason is because SDIL_j does not measure the information loss that degrades the analysis mean.

The main features of SIL_j, DIL_j and SDIL_j and their differences illustrated by the examples in this and previous subsections are also seen from other examples not presented in this paper. In these additional examples, radial-velocity observations are sampled along different beams from not only the NSSL phased array radar but also the Oklahoma City operational KTLX radar during the stormy weather on 2 June 2004, while the background fields are provided by COAMPS predictions with the observation and background error covariances estimated as in (Xu et al., 2005, 2007). The examples so far examined are one-dimensional. When radar radial-velocity observations are analysed for practical applications, analyses often need to be performed either on each two-dimensional conical surface of the radar scans (Xu et al., 2006) or in a full three-dimensional volume of the radar scans. In this case, the information content from observations can still be measured differently by the relative entropy (with the signal-dispersion partition) and the Shannon entropy difference, but the detailed differences between the two measures and their related features are expected to be more complex than illustrated by the one-dimensional examples in this paper.

6. Conclusions

In this paper, the relative entropy is compared with the Shannon entropy difference as a measure of the amount of information extracted from observations by an optimal analysis. The main differences between the two measures can summarized as follows: (i) The relative entropy measures both the signal and dispersion parts of the information content from observations, but the Shannon entropy difference measures only the dispersion part; (ii) The relative entropy provides a consistent measure as it is strictly invariant with respect to any smooth invertible transformation of variables, but the Shannon entropy difference is invariant only to a linear transformation and (iii) the Shannon entropy difference is additive for successive inclusions of observations into the analyses while the relative entropy is not additive and thus is not as convenient as the Shannon entropy difference in computing cumulative information content from successive groups of observations. The first two differences are among those summarized in section 2.4 of Majda et al. (2002) that favor the relative entropy over the Shannon entropy difference for quantifying predictive information content. The third difference was pointed out by James Purser (personal communication). The first difference appears to be the most

significant one in terms of measuring information content from observations. The significance of this difference is illustrated by examples in which radar radial-velocity observations are analysed in one-dimensional numerical model produced background velocities.

The information content can be defined only in a relative sense in general (except for a finite discrete case as discussed in the last paragraph of Section 2). The information content extracted from observations by an optimal analysis with a prior background pdf is thus indirectly measured in terms of the changes in the pdf produced by the analysis with respect to the background pdf. Hence, the information content from observations depends not only on the observation pdf but also on the background pdf.

When the observation and background pdfs are Gaussian (or transformed to Gaussian if doable), the integral form of the relative entropy yields an explicit formulation in which the signal part is given by the inner-product of the analysis increment vector weighted by the inverse of the background covariance matrix [see (3.6)] and the dispersion part is a non-negative definite function of the analysis covariance matrix multiplied by the inverse of the background covariance matrix [see (3.10)].

The above formulation can be further simplified via a linear transformation (in the observation space scaled by $\mathbf{R}^{1/2}$) by using the left orthogonal matrix obtained from the singular value decomposition (SVD) of the scaled observation operator [that is, $\mathbf{M} \equiv \mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}^{1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$, see (3.6)]. With this transformation, the information content becomes separable between different components associated with different singular values [see (4.1)], so the observations can be compressed without information loss by discarding the components associated with zero singular values, that is, the components in the null space of the scaled observation operator.

A further compression can be made to observations, but not without information loss, by discarding the components associate with the smallest non-zero singular values. The signal part of the information loss, also called signal information loss, depends on the truncated non-zero singular values and truncated components of the scaled innovation vector in the transformed observation space. The dispersion part of the information loss, also called dispersion information loss, depends only on the truncated non-zero singular values. As the truncation is made to the smallest non-zero singular values in the transformed observation space, its causes the minimum information loss in the dispersion part of the information content for a given truncation number.

As illustrated by the examples in Section 5, the signal information loss (SIL) can be closely related to the analysis mean degradation (MD) (see $\mathrm{SIL}_j/\mathrm{Sg}$ and $\mathrm{MD}_j/\mathrm{MD}_0$ curves in Figs. 3 and 5) while the dispersion information loss (DIL) can be closely related to the analysis covariance degradation (CD) (see $\mathrm{DIL}_j/\mathrm{Ds}$ and $\mathrm{CD}_j/\mathrm{CD}_0$ curves in Figs. 3 and 5). The Shannon entropy difference information loss (SDIL) can be also related to CD but not as closely as DIL (see $\mathrm{SDIL}_j/\mathrm{SD}$ curves in Figs. 3 and 5). As the truncation number increases from zero (that is, the number of singular-value terms retained by the truncation), the SIL often decreases slowly in the first few steps and then drops at one or two irregular large steps. The stepwise irregularity depends on the innovation vector. Contrary to SIL, DIL and SDIL are independent of the innovation vector, so they decrease rapidly and approach zero smoothly as the truncation number increases. Since DIL and SDIL do not measure the signal part, they tend to underestimate the information loss. When DIL (or SDIL) is used alone to determine the super-observation truncation, the actual information loss can be significantly larger than DIL (or SDIL). It is thus safer to use SIL and DIL in combination to measure the information loss and determine the super-observation truncation. A simple and natural combination is given by the sum of SIL and DIL, that is, the total information loss measured by the relative entropy.

There can be a significant degree of information redundancy if the observations are dense (relative to the background resolution) and the background covariance is local (becomes zero or virtually zero beyond a certain range of spatial separation). Densely distributed observations (such as those remotely sensed from satellites and ground-based radars) can be compressed in principle with minimum information loss by truncating the components associate with zero and smallest non-zero singular values in the transformed observation space based on the aforementioned SVD. The SVD and related matrix computations, however, are efficient only if the observation space or background space is not too large. Hence, the SVD-based compression is practical only if observations are assimilated serially in small batches. In this case, the background covariance must be also updated serially as in some of the ensemble Kalman filter techniques (Houtekamer and Mitchell, 2001; Whitaker and Hamill, 2002). For most operationally used data assimilation techniques, such as the Grid-point statistical interpolation which is currently being tested for radar data assimilation at the NCEP (Wu et al., 2002; Purser et al., 2003; Liu et al., 2005b), the background covariance is pre-estimated and not updated with the analysis, so all the observations collected during each data assimilation cycle are analysed together (rather than serially). In this case, the observation space and background space are both very large and the above SVD-based compression becomes impractical unless it is implemented locally in each properly divided observation subspaces. Such a localization can greatly improve the computational efficiency but will cause some additional information loss.

Purser et al. (2000) proposed some general methods for localized observational data compression based on a Gram-Schmidt decomposition of the observation operator. Localized observational data compression techniques may be also designed by using other types of truncated linear transformations without matrix decomposition to gain further computational efficiency. The super-obbing strategy proposed by Purser et al. (2000) is based on the Shannon entropy difference, so it allows the super-obbing

weights to be formulated beforehand based on the expected observation locations (observation operator) and observation and background covariances before the observations are taken. A super-obbing strategy based on the relative entropy, however, will depend not only on the observation locations but also on the observation values, so the super-obbing weights cannot be formulated before the observations are actually taken. Because of this, the relative entropy is not suitable for a super-obbing strategy in which the super-obbing weights must be formulated before-hand (to increase the computational efficiency). For radar observations, the exact observation locations and coverage are often unknown until the observations are actually taken, so the super-obbing weights cannot be formulated before-hand. In this case, a super-obbing strategy can be designed based on the Shannon entropy difference if the goal is to minimize the information loss that degrades the analysis covariance. However, if the goal is to minimize the total information loss that degrades both the analysis mean and covariance, then the super-obbing strategy should be designed based on the relative entropy.

The minimum information loss measured by the relative entropy or Shannon entropy difference in the singular-value form [see (4.5) or (4.6)] can be used as a benchmark to evaluate the additional information loss caused by a localized observational data compression technique that is designed to be sufficiently efficient for operational uses. The relative entropy and especially its singular-value form may be also used to measure the optimality of a remote sensing strategy (such as radar scanning strategy) in terms of maximizing the information content from observations for a given data assimilation system. This and other related applications of the relative entropy deserve further studies.

7. Acknowledgments

The author is thankful to Dr. James Purser and Dr. Richard Kleeman for their insightful comments and suggestions that improved the presentation of the results and to Li Wei and Kang Nai for their help in preparing the examples presented in Section 5. The research was supported by the ONR Grant N000140410312 to the University of Oklahoma and by FAA contract IA# DTFA03-01-X-9007 to NSSL.

8. Appendix A

Derivation of Eq. (3.2)

When the scalar variable x is extended to a vector variable $\mathbf{x}$, the relative entropy defined in (2.3) has the following form:

$$R(p,q) = \int d\mathbf{x} p(\mathbf{x}) \ln[p(\mathbf{x})/q(\mathbf{x})] = \langle \ln p(\mathbf{x}) \rangle_p - \langle \ln q(\mathbf{x}) \rangle_p, \quad (A.1)$$

where $\langle () \rangle_{\mathrm{p}} \equiv \int d\mathbf{x} p(\mathbf{x})()$ denotes the expectation of () based on $p(\mathbf{x})$. From (3.1), we have

$$\ln p(\mathbf{x}) = -[n \ln(2\pi) + \ln \mathrm{Det}(\mathbf{A}) + (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{x}-\mathbf{a})]/2, \quad (A.2a)$$

$$-\ln q(\mathbf{x}) = [n \ln(2\pi) + \ln \mathrm{Det}(\mathbf{B}) + (\mathbf{x}-\mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}-\mathbf{b})]/2. \quad (A.2b)$$

Note that $\mathbf{A}$ is constant with respect to $\mathbf{x}$ and $\int d\mathbf{x} p(\mathbf{x}) = 1$, so

$$\langle n \ln(2\pi) + \ln \mathrm{Det}(\mathbf{A}) \rangle_{\mathrm{p}} = n \ln(2\pi) + \ln \mathrm{Det}(\mathbf{A}). \quad (A.3)$$

Applying $\langle () \rangle_p$ to the last term in (A.2a) gives

$$\begin{aligned} \langle (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{x}-\mathbf{a}) \rangle_p &= \int d\mathbf{x} p(\mathbf{x}) (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{x}-\mathbf{a}) \\ &= (2\pi)^{-\mathrm{n}/2} \int d\mathbf{x} [\mathrm{Det}(\mathbf{A})]^{-1/2} (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{x}-\mathbf{a}) \\ &\quad \times \exp[-(\mathbf{x}-\mathbf{a})^{T} \mathbf{A}^{-1} (\mathbf{x}-\mathbf{a})/2] \\ &= (2\pi)^{-\mathrm{n}/2} \int d\mathbf{x}' |\mathbf{x}'|^2 \exp(-|\mathbf{x}'|^2/2) = n, \end{aligned} \quad (A.4)$$

where $\mathbf{x}' = \mathbf{A}^{-1/2}(\mathbf{x}-\mathbf{a})$ and $d\mathbf{x}' = d\mathbf{x}\mathrm{Det}(d\mathbf{x}'/d\mathbf{x}) = d\mathbf{x}\mathrm{Det}(\mathbf{A}^{-1/2}) = d\mathbf{x}[\mathrm{Det}(\mathbf{A})]^{-1/2}$ are used. Substituting (A.3) and (A.4) into $\langle (\mathrm{A.2a}) \rangle_p$ gives

$$\langle \ln p(\mathbf{x}) \rangle_p = -[n + n \ln(2\pi) + \ln \mathrm{Det}(\mathbf{A})]/2. \quad (A.5)$$

Applying $\langle () \rangle_p$ to the first two terms in (A.2b) gives

$$\langle n \ln(2\pi) + \ln \mathrm{Det}(\mathbf{B}) \rangle_p = n \ln(2\pi) + \ln \mathrm{Det}(\mathbf{B}). \quad (A.6)$$

The last term in (A.2b) can be expanded into $(\mathbf{x}-\mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}-\mathbf{b}) = (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}-\mathbf{a}) + (\mathbf{a}-\mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{a}-\mathbf{b}) + (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{a}-\mathbf{b}) + (\mathbf{a}-\mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}-\mathbf{a})$. Applying $\langle () \rangle_p$ to each expanded term gives

$$\langle (\mathbf{a}-\mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{a}-\mathbf{b}) \rangle_p = (\mathbf{a}-\mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{a}-\mathbf{b}), \quad (A.7)$$

$$\langle (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{a}-\mathbf{b}) \rangle_p = \langle (\mathbf{a}-\mathbf{b})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}-\mathbf{a}) \rangle_p = 0, \quad (A.8)$$

$$\begin{aligned} \langle (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}-\mathbf{a}) \rangle_p &= \int d\mathbf{x} p(\mathbf{x}) (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}-\mathbf{a}) \\ &= (2\pi)^{-n/2} \int d\mathbf{x} [\mathrm{Det}(\mathbf{A})]^{-1/2} (\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}-\mathbf{a}) \\ &\quad \times \exp[-(\mathbf{x}-\mathbf{a})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{x}-\mathbf{a})/2] \\ &= (2\pi)^{-n/2} \int d\mathbf{x}' (\mathbf{x}'^{T} \mathbf{A}^{1/2} \mathbf{B}^{-1} \mathbf{A}^{1/2} \mathbf{x}') \\ &\quad \times \exp(-|\mathbf{x}'|^2/2) = \mathrm{Tr}(\mathbf{A}^{1/2} \mathbf{B}^{-1} \mathbf{A}^{1/2}), \end{aligned} \quad (A.9)$$

where $\mathbf{x}' = \mathbf{A}^{-1/2}(\mathbf{x}-\mathbf{a})$ is used as in (A.2). Substituting (A.6)–(A.10) into $\langle (\mathrm{A.2b}) \rangle \mathrm{p}$ gives

$$\begin{aligned} -\langle \ln q(\mathbf{x}) \rangle_p &= -\int d\mathbf{x} p(\mathbf{x}) \ln q(\mathbf{x}) \\ &= [(\mathbf{a}-\mathbf{b})^{T} \mathbf{B}^{-1} (\mathbf{a}-\mathbf{b}) + \mathrm{Tr}(\mathbf{A}^{1/2} \mathbf{B}^{-1} \mathbf{A}^{1/2}) \\ &\quad + \ln \mathrm{Det}(\mathbf{B}) + n \ln(2\pi)]/2. \end{aligned} \quad (A.10)$$

The multiplicative property of determinant gives

$$\begin{aligned}\mathrm{Det}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}) &= \mathrm{Det}(\mathbf{B}^{1/2})\mathrm{Det}(\mathbf{A}^{-1})\mathrm{Det}(\mathbf{B}^{1/2}) \\ &= \mathrm{Det}(\mathbf{B})/\mathrm{Det}(\mathbf{A}), \end{aligned} \tag{A.11}$$

so $\ln\mathrm{Det}(\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}) = \ln\mathrm{Det}(\mathbf{B}) - \ln\mathrm{Det}(\mathbf{A})$. Since $\mathbf{A}$ and $\mathbf{B}$ are real symmetric matrices, we have the following decompositions (see theorem 8.1-1 of Golub and Van Loan, 1983): $\mathbf{A} = \mathbf{U}_a\mathbf{\Lambda}_a^2U_a^{\mathrm{T}}$ and $\mathbf{\Lambda}_a^{-1}\mathbf{U}_a^{\mathrm{T}}\mathbf{B}\mathbf{U}_a\mathbf{\Lambda}_a^{-1} = \mathbf{C} = \mathbf{U}_c\mathbf{\Lambda}_c^2\mathbf{U}_c^{\mathrm{T}}$, where $\mathbf{\Lambda}_a^2$ (or $\mathbf{\Lambda}_c^2$) is the diagonal matrix composed of the eigenvalue of $\mathbf{A}$ (or $\mathbf{C}$) and $\mathbf{U}_a$ (or $\mathbf{U}_c$) is the orthogonal matrix composed of the eigenvectors of $\mathbf{A}$ (or $\mathbf{C}$). With these decompositions, we have $\mathbf{A}^{1/2} = \mathbf{U}_a\mathbf{\Lambda}_a\mathbf{U}_a^{\mathrm{T}}$ and $\mathbf{B}^{-1} = \mathbf{U}_a\mathbf{\Lambda}_a^{-1}\mathbf{U}_c\mathbf{\Lambda}_c^{-2}\mathbf{U}_c^{\mathrm{T}}\mathbf{\Lambda}_a^{-1}\mathbf{U}_a^{\mathrm{T}}$. By using these expressions, one can verify that

$$\begin{aligned}\mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2}) &= \mathrm{Tr}\big(\mathbf{\Lambda}_c^{-2}\big), \\ \mathrm{Tr}(\mathbf{A}\mathbf{B}^{-1}) &= \mathrm{Tr}\big(\mathbf{U}_a\mathbf{\Lambda}_a^2\mathbf{U}_a^{\mathrm{T}}\mathbf{U}_a\mathbf{\Lambda}_a^{-1}\mathbf{U}_c\mathbf{\Lambda}_c^{-2}\mathbf{U}_c^{\mathrm{T}}\mathbf{\Lambda}_a^{-1}\mathbf{U}_a^{\mathrm{T}}\big) \\ &= \mathrm{Tr}\big(\mathbf{U}_a\mathbf{\Lambda}_a\mathbf{U}_c\mathbf{\Lambda}_c^{-2}\mathbf{U}_c^{T}\mathbf{\Lambda}_a^{-1}\mathbf{U}_a^{T}\big) = \mathrm{Tr}\big(\mathbf{\Lambda}_c^{-2}\big), \\ \mathrm{Tr}(\mathbf{B}^{-1}A) &= \mathrm{Tr}\big(\mathbf{U}_a\mathbf{\Lambda}_a^{-1}\mathbf{U}_c\mathbf{\Lambda}_c^{-2}\mathbf{U}_c^{\mathrm{T}}\mathbf{\Lambda}_a^{-1}\mathbf{U}_a^{\mathrm{T}}\mathbf{U}_a\mathbf{\Lambda}_a^2\mathbf{U}^{\mathrm{T}}\big) \\ &= \mathrm{Tr}\big(\mathbf{U}_a\mathbf{\Lambda}_a^{-1}\mathbf{U}_c\mathbf{\Lambda}_c^{-2}\mathbf{U}_c^{T}\mathbf{\Lambda}_a\mathbf{U}_a^{T}\big) = \mathrm{Tr}\big(\mathbf{\Lambda}_c^{-2}\big), \end{aligned}$$

where $\mathbf{U}_a^{\mathrm{T}}\mathrm{Ua} = \mathbf{I}$ is used. This shows that $\mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2}) = \mathrm{Tr}(\mathbf{A}\mathbf{B}^{-1}) = \mathrm{Tr}(\mathbf{B}^{-1}\mathbf{A})$. Similarly, one can show that $\mathrm{Tr}(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}) = \mathrm{Tr}(\mathbf{B}^{-1}\mathbf{A})$, and thus

$$\mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2}) = \mathrm{Tr}(\mathbf{B}^{-1}\mathbf{A}) = \mathrm{Tr}(\mathbf{A}\mathbf{B}^{-1}) = \mathrm{Tr}(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}). \tag{A.12}$$

Substituting (A.5) and (A.10)–(A12) into (A.1) give (3.2).

References

Bennett, A. F. 1992. *Inverse Method in Physical Oceanography*. Cambridge University Press, New York, 346 pp.

Bernardo, J. M. and Smith, A. F. M. 1994. *Bayesian Theory*. John Wiley and Sons, New York, 586 pp.

Daley, R. 1991. *Atmospheric Data Analysis*. Cambridge University Press, New York, 457 pp.

Doviak, J. D. and Zrnic, D. S. 1993. *Doppler Radar and Weather Observations.* 2nd Edition. Academic Press, New York, 562 pp.

Eyre, J. R. 1990. The information content of data from satellite sounding systems. A simulation study. *Q. J. R. Meteor. Soc.* **116**, 401–434.

Golub, G. H. and Van Loan, C. F. 1983. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 476 pp.

Gong, J., Wang, L. and Xu, Q. 2003. A three-step dealiasing method for Doppler velocity data quality control. *J. Atmos. Ocean. Technol.* **20**, 1738–1748.

Haven, K., Majda, A. J. and Abramov, R. 2005. Quantifying predictability through information theory. small sample estimation in a non-Gaussian framework. *J. Comp. Phys.* **206**, 334–362.

Hodur, R. M. 1997. The Naval Research Laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). *Mon. Wea. Rev.* **125**, 1414–1430.

Houtekamer, P. L. and Mitchell, H. L. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.* **129**, 123–137.

Huang, H.-L. and Purser, R. J. 1996. Objective measures of the information density of satellite data. *Meteorol. Atmos. Phys.* **60**, 105–117.

Jazwinski, A. H. 1970. *Stochastic Processes and Filtering Theory*. Academic Press, San Diego, 376 pp.

Kleeman, R. 2002. Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.* **59**, 2057–2072.

Kleeman, R. and Majda, A. J. 2005. Predictability in a model for geophysical turbulence. *J. Atmos. Sci.* **62**, 2864–2879.

Liu, S., Xu, Q. and Zhang, P. 2005a. Quality control of Doppler velocities contaminated by migrating birds. Part II: Bayes identification and probability tests. *J. Atmos. Oceanic Technol.* **22**, 1114–1121.

Liu, S., Xue, M., Gao, J. and Parrish, D. F. 2005b. Analysis and impact of super-obbed Doppler radial velocity in the NCEP grid-point statistical interpolation (GSI) analysis system. Extended abstract. *17th Conf. Num. Wea. Pred.* Washington DC, Amer. Meteor. Soc. 13A.4.

Majda, A. J., Kleeman, R. and Cai, D. 2002. A mathematical framework for quantifying predictability through relative entropy. *Methods Appl. Anal.* **9**, 425–444.

Majda, A. J. and Wang, X. 2006. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*. Cambridge University Press, New York, 565 pp.

O'Hagan, A. 1994. *Kendall's Advanced Theory of Statistics. Volume 2B. Bayesian Inference*. Oxford University Press. New York, 332 pp.

Peckham, G. 1974. The information content of remote measurements of atmospheric temperature by satellite infra-red radiometry and optimum radiometer configurations. *Q. J. R. Meteor. Soc.* **100**, 406–419.

Papoulis, A. 1991. *Probability, Random Variables, and Stochastic Processes*. 3rd Edition, McGraw-Hill, New York, 666 pp.

Purser, R. J., Parrish, D. F. and Masutani, M. 2000. Meteorological observational data compression; An alternative to conventional "super-Obbing". *Office Note 430*, National Centers for Environmental Prediction, Camp Springs, MD, 12 pp. [available on line http://www.emc.ncep.noaa.gov/officenotes/FullTOC.html#2000]

Purser, R. J., Wu, W.-S., Parrish, D. F. and Roberts, N. M. 2003. Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances. *Mon. Wea. Rev.* **131**, 1524–1535.

Shannon, C. E. 1949. Communication in the presence of noise. *Proc. I.C.E.* **37**, 10–21.

Whitaker, J. S. and Hamill, T. M. 2002. Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.* **130**, 1913–1924.

Wu, W.-S., Purser, R. J. and Parrish, D. F. 2002. Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.* **130**, 2905–2916.

Xu, Q., Nai, K., Wei, L., Zhang, P., Wang, L. and co-authors. 2005. Progress in Doppler radar data assimilation. *32nd Conference on Radar Meteorology*, 24–29 October 2005, Albuquerque, New Mexico, Amer. Meteor. Soc., CD-ROM, JP1J7.

Xu, Q., Nai, K. and Wei, L. 2007. An innovation method for estimating radar radial-velocity observation error and background wind error covariances. *Q. J. R. Meteor. Soc.*, in press.

Xu, Q., Liu, S. and Xue, M. 2006. Background error covariance functions for vector wind analyses using Doppler radar radial-velocity observations. *Q. J. R. Meteorol. Soc.* in press.

Zhang, P., Liu, S. and Xu, Q. 2005. Quality control of Doppler velocities contaminated by migrating birds. Part I: Feature extraction and quality control parameters. *J. Atmos. Oceanic Technol.* **22**, 1105–1113.