U.S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
5200 Auth Road
Camp Springs, MD 20746-4304

Office Note 430

METEOROLOGICAL OBSERVATIONAL DATA COMPRESSION; AN ALTERNATIVE TO CONVENTIONAL "SUPER-OBBING"

R. James Purser*

David F. Parrish†

Michiko Masutani‡
Environmental Modeling Center

May 2000

THIS IS AN UNREVIEWED MANUSCRIPT, PRIMARILY INTENDED FOR INFORMAL EXCHANGE OF INFORMATION AMONG THE NCEP STAFF MEMBERS

* General Sciences Corporation, Beltsville, Maryland; e-mail: jpurser@ncep.noaa.gov
† National Centers for Environmental Prediction, Camp Springs, Maryland; e-mail: dparrish@ncep.noaa.gov
‡ Research and Data Systems Corporation, Greenbelt, Maryland; e-mail: mmasutani@ncep.noaa.gov

1. Introduction

It is increasingly typical for a large proportion of the data assimilated into numerical forecasting models to derive from various remote sensing instruments, such as radar, satellite passive and active sounders and, in the not too distant future, satellite or ground-based Doppler lidar. Although they are rich sources of data, a characteristic they share is the presence among the reports transmitted of a signifant degree of information-redundancy, either by way having linearly dependent weighting functions in the case of passive satellite sounders, or simply by way of their spatial and temporal densities being far in excess of the required resolution of the assimilating system, which is typical (to an extreme degree) of radar data.

Redundant data impose a burden on an operational assimilation system since each datum is laboriously processed, usually involving repetitive interpolations from the analysis grid to its location, and adjoint interpolations back again, performed each iteration of the large-scale linear solver (typically a conjugate-gradient or quasi-Newton method). This effort is carried out for each datum regardless of the information that can be attributed to it in the overall assimilation. The time and storage expended on mutually redundant data could be better spent on improving other aspects of an assimilation. Therefore, it is desirable, ahead of time if possible, to effect whatever data compression the ensemble of fresh observations allow, subject to the resulting degradation of the otherwise "optimal" (in some formal sense only) analysis being negligible. The term for a surrogate datum which replaces several partially redundant actual data is a "super observation" (sometimes abbreviated to "super-ob").

A systematic construction of super observations obeying certain desirable criteria related to the known statistical properties of the background error has been described by Lorenc (1981). This method seeks to obtain surrogate data explicitly uncorrelated with the background (a strong constraint), using a modification of the formalism of optimal interpolation to produce the desired super observation's value. This approach was demonstrated by Purser (1990) to possess a simple generalization to multi-component super observations, with applications to the assimilation of "retrieved" satellite soundings derived from passive multi-channel radiometer measurements. However, this approach is oriented specifically to one particular assimilation model's background field characteristics, which is a disadvantage, not only because it means the resulting super observations lack independence from the selected assimilation system, but also because a prerequisite of such a method is a statistical description of part of one particular forecast system which is not always conveniently available when and where the occasion for data compression is appropriate. For radar data, it would be desirable to compress the data at the measurement site, not at the numerical forecasting center. Likewise, for satellite lidar, the communication bottle-neck between satellite and ground would be relieved if the construction of the super observation could be organized on the satellite itself prior to transmission.

Here we propose some alternative general methods of constructing surrogate observations which require no detailed information about the statistics of the background field with which the data are destined to be combined. Also, we show how the concepts of information theory, together with their adaptation suggested by Huang and Purser (1996) to allow the local attribution of "information density" contained in meteorological observations of various kinds, provide us with incisive tools to determine the extent to which putative methods of constructing the super observations will degrade the information available. In this way we may be guided in formulating practical and information-efficient prescriptions for the systematic generation of the super observations that many new and future data sources require.

The following section describes two general conditions in which the construction of super observations is appropriate. The linear algebra of the construction itself is given in section 3. The remaining sections deal with the assessment of the information content of the data in unprocessed and super observation form.

2. Conditions enabling data compression in a statistical assimilation

The optimal linear analysis of a meteorological variable $\mathbf{x}^a$ whose prior "background" estimate is $\mathbf{x}^b$, given independent measurements $\mathbf{y^o}$, is obtained from the well-known equations of "optimum interpolation" (Gandin, 1963; Daley 1991). Using the notation of Ide et al. (1997) and assuming linearity:

$$\mathbf{x}^a - \mathbf{x}^b = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{d}, \tag{2.1}$$

with

$$\mathbf{d} = \mathbf{y^o} - \mathbf{H}\mathbf{x}^b \tag{2.2}$$

or, in equivalent implicit form,

$$\mathbf{x}^a - \mathbf{x}^b = \mathbf{B}\mathbf{H}^T\mathbf{R^{-1}}(\mathbf{y^o} - \mathbf{H}\mathbf{x}^a). \tag{2.3}$$

Here, $\mathbf{B}$ is the covariance $\langle\boldsymbol{\beta}\boldsymbol{\beta}^T\rangle$ of the background error $\boldsymbol{\beta} \equiv \mathbf{x}^b - \mathbf{x}^t$, where $\mathbf{x}^t$ is the true state. $\mathbf{R}$ is the effective covariance of measurement error, taken to comprise the following two terms: the actual covariance $\mathbf{E} \equiv \langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\rangle$ of measurement error $\boldsymbol{\epsilon} = \mathbf{y^o} - \mathbf{H}\mathbf{x}^t$; and the "representativeness error" $\mathbf{F}$ (Lorenc 1986) introduced to account for the existence of detail in the measured true state unresolved by the assimilating model. Thus $\mathbf{R} = \mathbf{E} + \mathbf{F}$. $\mathbf{H}$ is the linearized observation operator. A derivation of these equations for the linear case is provided in Appendix A.

There are various ways in which the set of measurements allow a significant compression of information without a significant degradation in the quality of the resulting analysis. Trivially, some observations may have such large uncertainties (poor quality) that their "precision" weights $\mathbf{R^{-1}}$ render them virtually ineffective; but we shall assume that at least these cases have been weeded out prior to presentation of the data to the assimilation software. It is evident from (2.3) that, when the rows of $\mathbf{H}$ are effectively linearly dependent, a smaller vector $\hat{\mathbf{y}}$ of surrogate data can replace $\mathbf{y^o}$, new weights $\hat{\mathbf{R}}^{-1}$ can replace $\mathbf{R^{-1}}$ and a new measurement operator $\hat{\mathbf{H}}$ of correspondingly fewer rows substituted for $\mathbf{H}$ without significantly changing the analysis, provided the following two conditions are satisfied:

$$\begin{aligned} \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{y}} &= \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y^o}, & (2.4)\\ \hat{\mathbf{H}}^T\hat{\mathbf{R}}^{-1}\hat{\mathbf{H}} &= \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}. & (2.5) \end{aligned}$$

We shall refer to this case of data compression as the creation of a "type-1" super observation. The circumstances conducive to this opportunity for data compression are typical of satellite passive sounding data, whose vertical weighting functions are in large measure mutually redundant. In section 3(a) we present some efficient numerical methods for the construction of such super observations.

Another case which is becoming ever more typical with the progressive advances in remote sensing instrumentation occurs under the following conditions: an analysis and a background error covariance (relative to the background value at one arbitrary fixed point, say z_*) which are both smooth functions devoid of detailed structure within a distance scale which is still large enough to contain several independent measurements. We shall see that this case also enables a reduction in the effective quantity of data without a signifcant loss of information. In this case, which corresponds to the traditional justification for creating super observations, we exploit a known quality (smoothness) of the analysis instead of the inherent internal redundancy implied by the observation operator. Hence, we distinguish this case from that of type-1 super observations, although we shall find that the formalism for construction of super observations in this case, "type-2" super observations, is largely the same. We discuss their creation in section 3(b).

3. Construction of super observations

An optimal analysis has the cumulative property that allows data to be incorporated sequentially (Parrish and Cohn 1985, Cohn and Parrish 1991). Thus, if the optimal analysis using all except some cluster of data is denoted $\mathbf{x}^*$, we can incorporate this remaining cluster optimally merely by applying the formula (2.1) but with $\mathbf{x}^b$ replaced by $\mathbf{x}^*$ and with the original matrix $\mathbf{B}$ now replaced by the covariance of error of $\mathbf{x}^*$. In other words, we incur no loss of generality by assuming here that $\mathbf{y}^{\mathbf{o}}$ denotes only the data in the cluster of interest, with the locations of the individual measuements at z_α.

For reasons to do with the numerical robustness of the process of creating super observations, it is desirable to rewrite the analysis equations rescaled in nondimensional form. Let us suppose diagonal matrix $\mathbf{D}_x$ has elements roughly comparable to the variance components of the background error and that, similarly, diagonal $\mathbf{D}_y$ has components roughly comparable to the variances of the data. Then we may rescale the background, analysis and observations:

$$\tilde{\mathbf{x}}^b = \mathbf{D}_x^{-1/2}\mathbf{x}^b, \tag{3.1}$$

$$\tilde{\mathbf{x}}^a = \mathbf{D}_x^{-1/2}\mathbf{x}^a, \tag{3.2}$$

$$\tilde{\mathbf{y}} = \mathbf{D}_y^{-1/2}\mathbf{y}^{\mathbf{o}}, \tag{3.3}$$

$$\tag{3.4}$$

form the nondimensional background covariance and measurement precision,

$$\tilde{\mathbf{B}} = \mathbf{D}_x^{-1/2}\mathbf{B}\mathbf{D}_x^{-1/2}, \tag{3.5}$$

$$\tilde{\mathbf{R}}^{\mathbf{-1}} = \mathbf{D}_y^{1/2}\mathbf{R}^{\mathbf{-1}}\mathbf{D}_y^{1/2}, \tag{3.6}$$

together with the nondimensional observation operator,

$$\tilde{\mathbf{H}} = \mathbf{D}_y^{-1/2}\mathbf{H}\mathbf{D}_x^{1/2}, \tag{3.7}$$

and express (2.3) in equivalent nondimensional terms:

$$\tilde{\mathbf{x}}^a - \tilde{\mathbf{x}}^b = \tilde{\mathbf{B}}\tilde{\mathbf{H}}^T\tilde{\mathbf{R}}^{\mathbf{-1}}(\tilde{\mathbf{y}} - \tilde{\mathbf{H}}\tilde{\mathbf{x}}^a). \tag{3.8}$$

We shall assume that the nondimensionalization has been done, and drop the tildes for the rest of this section. We describe separately the construction of type-1 and type-2 super observations below.

(*a*) *Type-1 super observations*

We form from (nondimensionally rescaled) **H** a Gram-Schmidt decomposition, with pivoting, based on a definition of inner-product that involves a weighting by the (rescaled) precision, that is:

$$\mathbf{H} = \mathbf{G}\mathbf{U}, \tag{3.9}$$

where the columns of **G** are mutually orthogonal (but not normalized) in the sense:

$$\mathbf{G}^T\mathbf{R}^{-\mathbf{1}}\mathbf{G} = \mathbf{X}, \tag{3.10}$$

for some positive diagonal **X**, and, following a permutation of its columns to "undo" the pivoting, the rows of **U** (which will be fewer than its columns in those cases in which a data compression is successful) belong to the upper triangular part of the matrix with unit elements on the main diagonal. The Gram-Schmidt decomposition is permitted to terminate as soon as the norms of all the remaining "unprocessed" columns of the partially constructed **G** become smaller than some small positive criterion; these negligible columns, and the corresponding rows of **U**, are deleted. The prior rescaling of the variables facilitates the discrimination between significant and negligible columns in the Gram-Schmidt process; then a size criterion for the nondimensional column vectors of **G** of about .01 is usually adequate. When data compression is achievable, the completed **G** has fewer columns than does the original rescaled observation operator, **H**, from which it is processed. A schematic description of an algorithm to perform such a Gram-Schmidt factorization is given in Appendix B. It is then possible to replace the vector of observations $\mathbf{y^o}$ by the shorter vector, $\hat{\mathbf{y}}$, **H** by $\hat{\mathbf{H}}$ and $\mathbf{R}^{-\mathbf{1}}$ by $\hat{\mathbf{R}}^{-\mathbf{1}}$ according to:

$$\hat{\mathbf{y}} = \mathbf{X}^{-1}\mathbf{G}^T\mathbf{R}^{-\mathbf{1}}\mathbf{y^o}, \tag{3.11}$$

$$\hat{\mathbf{H}} = \mathbf{U}, \tag{3.12}$$

$$\hat{\mathbf{R}}^{-\mathbf{1}} = \mathbf{X}. \tag{3.13}$$

The prescription (3.11) –(3.13) describes the simplest of our examples of data compression and would apply to a sounding of satellite radiance measurements in a single vertical column of the atmosphere. However, if the space of vectors $\mathbf{x}^a$ and $\mathbf{x}^b$ is the more inclusive domain of both vertical *and horizontal* dimensions, then the practical procedure is not strictly as it is written above. Instead we should factor the observation operator, **H**, into, first: a preliminary restriction operator, **N**, which, on multiplying $\mathbf{x}^a$, selects from it only those components residing at (or horizontally interpolated to) the vertical column of the sounding; second, a restriction, $(\mathbf{H}\boldsymbol{\tau})$, of the observation operator to the range comprising a single vertical column only. Thus, **H** is replaced by the equivalent, $(\mathbf{H}\boldsymbol{\tau})\mathbf{N}$, in the rescaled counterpart of (2.3), but the Gram-Schmidt process is applied now to the matrix $(\mathbf{H}\boldsymbol{\tau})$ whose rows span only the vertical part, not the whole, of the space of analysis variables. Replacing (3.9), we now have,

$$\mathbf{H}\boldsymbol{\tau} = \mathbf{G}\mathbf{U}. \tag{3.14}$$

Eq. (3.10) applies as before and the definitions (3.11) and (3.13) for $\hat{\mathbf{y}}$ and $\hat{\mathbf{R}}^{-1}$ remain unaltered, but the new observation operator $\hat{\mathbf{H}}$ inherits the factorization into a horizontal restriction followed by a vertical-column operator:

$$\hat{\mathbf{H}} = \mathbf{UN}. \tag{3.15}$$

(*b*) *Type-2 super observations*

The justification for the construction of type-2 super observations relies on the assumption of smoothness in the analysis variables and, on the face of it, would seem likely to lead to a completely different formalism than the one we used for super observations of type-1. However, we shall find that, in fact, the formalisms are practically identical, although the interpretation of some of the terms becomes more general in the case of type-2 super observations.

When we speak of some meteorological field ψ as being "smooth" in the vicinity of a point $z = 0$ of a local coordinate system, we shall use the term to signify that the field in question can be well approximated there by a finite Taylor series. Thus, let $\mathbf{N}$ now denote the linear operator acting on a field in the space of analysis variables which returns the value and spatial derivatives, up to some specified degree p, at $z = 0$, about which are clustered all the discrete observations that we wish to replace by a super observation. Note that the operator $\mathbf{N}$ of the previous subsection was, in the horizontal sense, just the special case of the present $\mathbf{N}$ in which $p = 0$ (signifying the selection of the *value* at $\mathbf{z} = 0$, but none of the proper derivatives). Corresponding to this operator and, in a sense, a generalized inverse of it, we denote by $\boldsymbol{\tau}(z)$ the "moment operator", each column of which is the field (in at least the local portion of analysis space) comprising the powers of the components of the local coordinates $\mathbf{z}$ complementary to those of $\mathbf{N}$ in the finite Taylor series. Some thought will convince one that, again, the special case of this operator in the horizontal with degree $p = 0$ is consistent with the spatial restriction operator $\boldsymbol{\tau}$ of the previous subsection. We shall continue to assume the dependent variables in the analysis are rescaled, but we also assume that the choice of scaling in the local coordinates $\mathbf{z}$ conforms approximately to the characteristic scale of spatial variability, in each direction, inherent in the covariance $\mathbf{B}$. Take as a simple example, the one dimensional case of degree $p = 2$. Then,

$$\mathbf{N}^T \equiv (1, d/dz, d^2/dz^2)|_{z=0}, \tag{3.16}$$

and the three columns of the complementary operator $\boldsymbol{\tau}$ are as given by the right-hand side of:

$$\boldsymbol{\tau}(z) \equiv (1, z, z^2/2!). \tag{3.17}$$

Note that $\mathbf{N}$ and $\boldsymbol{\tau}$ are mutual pseudo-inverses at $z = 0$ in the sense:

$$\mathbf{N}\boldsymbol{\tau} = \mathbf{I}. \tag{3.18}$$

Now we formalize the attribute of "smoothness" of ψ to mean that

$$\psi(z) \approx \boldsymbol{\tau}(z)\mathbf{N}\psi, \tag{3.19}$$

which, for the one-dimensional example of degree $p = 2$, means:

$$\psi(z) \approx \psi(0) + z\frac{d\psi(0)}{dz} + \frac{z^2}{2!}\frac{d^2\psi(0)}{dz^2}, \tag{3.20}$$

in the vicinity of $z = 0$. More specifically, since we shall require this approximation to hold only over the extent of the cluster of raw observations that we aim to replace, we shall regard the field there as "smooth" if and only if the (error-free) *measurements* of its finite Taylor series faithfully reproduce the corresponding measurements of the actual field:

$$\mathbf{H}\psi(z) \approx \mathbf{H}\boldsymbol{\tau}(z)\mathbf{N}\psi. \tag{3.21}$$

The combination, $(\mathbf{H}\boldsymbol{\tau})$, is a matrix of the "moments" of the observation operator for the cluster.

The one dimensional example above can be generalized to more dimensions by extending the vector operator $\mathbf{N}$ and moment vector $\boldsymbol{\tau}$ consistently. Thus, up to second degree in two dimensions, $\mathbf{z} \equiv (z_1, z_2)$, we would have for the operator, $\mathbf{N}$,

$$\mathbf{N}^T = \left(1; \frac{\partial}{\partial z_1}, \frac{\partial}{\partial z_2}; \frac{\partial^2}{\partial z_1^2}, \frac{\partial^2}{\partial z_1 \partial z_2}, \frac{\partial^2}{\partial z_2^2} \right) \Bigg|_{\mathbf{z}=0} \tag{3.22}$$

and the corresponding moment operator, $\boldsymbol{\tau}$, would be:

$$\boldsymbol{\tau} = (1; z_1, z_2; z_1^2/2, z_1 z_2, z_2^2/2), \tag{3.23}$$

but (3.18) still applies.

When the background error covariance $\mathbf{B}$ and the analysis $\mathbf{x}^a$ are smooth in the technical sense we have defined above, then the right-hand side of the analysis equation (3.8) for the rescaled variables must satisfy the approximation,

$$\mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}^{\mathbf{o}} - \mathbf{H}\mathbf{x}^a) \approx \mathbf{B}\mathbf{N}^T\boldsymbol{\tau}^T\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}^{\mathbf{o}} - \mathbf{H}\boldsymbol{\tau}\mathbf{N}\mathbf{x}^a). \tag{3.24}$$

As before, we decompose the combination $(\mathbf{H}\boldsymbol{\tau})$ exactly as in (3.14) using the Gram-Schmidt procedure with orthogonality as in (3.10). Again, we find (3.11), (3.13) and (3.15) to define the multi-component super observation $\hat{\mathbf{y}}$ and its companion operators $\hat{\mathbf{R}}^{-1}$ and $\hat{\mathbf{H}}$. However, this new super observation may now contain information, not only about the value of $\mathbf{x}$, but also about its gradient, and even trend components of higher degree.

Despite the apparent complexity of the formulae above, we note that, in the case of type-2 super observations of zeroth-degree, the algebra implies that the weight $\hat{\mathbf{R}}^{-1}$ is simply the sum of the weights of the original observations and the super observation $\hat{\mathbf{y}}$ itself is just the corresponding weighted sum of the original observations, $\mathbf{y}^{\mathbf{o}}$, as one would intuitively expect. The value of the complete formalism is that it systematically provides for the construction of more general super observations in which gradient information is properly preserved and weighted or, if required, even higher degrees of the Taylor-series characterization of the field analyzed are treated. There is a sense in which the construction of the type-2 super observations described above is equivalent to the method of weighted linear ("least-squares") regression (Menke, 1984) in trend-fitting; the Gram-Schmidt solution to the problem of rank-deficiency we have adopted here applies equally to such problems of linear regression.

It is easily verified that the $\hat{\mathbf{R}}$ defined above is indeed the covariance of error of the associated super observation $\hat{\mathbf{y}}$. When the original data $\mathbf{y}^{\mathbf{o}}$ of the cluster are very numerous compared to the few components of $\hat{\mathbf{y}}$, then we have achieved a significant and valuable compression of data and a numerical simplification of the computational task of data assimilation. By construction, $\hat{\mathbf{R}}^{-\mathbf{1}}$ is nonsingular and the rows of $\hat{\mathbf{H}}$ are measuring independent and informative attributes of $\mathbf{x}$, so, at least within the present cluster, there remains no further redundancy.

For both type-1 and type-2 super observations, we could alternatively employ a singular-value decomposition (SVD) in place of the Gram-Schmidt process in order to handle the problem of rank-deficiency. In this case, the form of the decomposition replacing (3.14) would be

$$\mathbf{R}^{-1/2}\mathbf{H}\boldsymbol{\mathcal{T}} \equiv \mathbf{G}\boldsymbol{\Lambda}\mathbf{F}^T, \tag{3.25}$$

with $\mathbf{G}^T\mathbf{G} = \mathbf{I}$, $\mathbf{F}^T\mathbf{F} = \mathbf{I}$, and positive diagonal $\boldsymbol{\Lambda}$. However, while less elegant formally, the Gram-Schmidt method accomplishes the task adequately and, in practice, is significantly less costly to apply than SVD.

We have not discussed the choice of coordinate origin for the Taylor series used in the above construction. For statistically independent ($\mathbf{R}^{-\mathbf{1}}$ diagonal) observations of homogeneous type (for example, all temperatures or all winds) the natural choice is the precision-weighted centroid, $\bar{\mathbf{z}}$ defined by:

$$\bar{\mathbf{z}} = \frac{\sum_\alpha \mathbf{R}^{-\mathbf{1}}{}_{\alpha,\alpha}\mathbf{z}_\alpha}{(\sum_\alpha \mathbf{R}^{-\mathbf{1}}{}_{\alpha,\alpha})} \tag{3.26}$$

(*c*) *Discussion*

Formally, the process of replacing the multitude of raw data by a smaller set of approximately equivalent quantities must generally entail a finite loss of information. The justification and motivation for carrying out such a procedure is that it reduces the computational load in the subsequent analysis step to more manageable proportions. Indirectly, a judicious data reduction can lead to definite improvements in the analysis. This occurs when the reduction in the data burden so streamlines the execution of each iteration of the analysis scheme's linear solver that *more* iterations than otherwise can be accommodated in the time allotted, thereby improving the final iteration's correspondence with the theoretically optimal state. However, in the case where the iterations are repeated to perfection, the analysis with super observations will generally be of slightly inferior quality to the analysis using the data in their raw form. It is therefore desirable to determine, under various strategies of assigning clusters of data to the same super observation, and under different assumptions of the degree, p, of the Taylor approximation assumed in each super observation's construction, how much information is formally lost by this process. It turns out that this expected information loss, for any particular analysis scenario in which the measurement operators are purely linear and the underlying statistics Gaussian, may be quantified precisely by the formal methods of information theory, which we discuss in the following section.

4. Quantification of information

If $\mathbf{A}_0$ denotes the error-covariance of the analysis derived by the application of (2.1) with measurement errors assumed uncorrelated with background errors, then it is straight-forward to verify that,

$$\mathbf{A} = \mathbf{B} - \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{B}. \tag{4.1}$$

Similarly, we find that the error covariance of the sub-optimal analysis using the super observations is,

$$\hat{\mathbf{A}} = \mathbf{B} - \mathbf{B}\hat{\mathbf{H}}^T(\hat{\mathbf{H}}\mathbf{B}\hat{\mathbf{H}}^T + \hat{\mathbf{R}})^{-1}\hat{\mathbf{H}}\mathbf{B}. \tag{4.2}$$

According to the quantitative theory of information developed by Shannon (1949) (see also Khinchin 1957), when the acquisition of new knowledge about the location of a point in an abstract space enables the "volume" of uncertainty of this location to be reduced in size by a factor of 2^Z for every probability contour, then the gain in the amount of "information", in its technical sense, can be consistently reckoned to be precisely Z "bits" of information. In this technical sense, the information from independent sources combines additively. While the classical theory of information was originally developed with applications to communication in mind, it is also applicable to estimation theory, of which variational meteorological data analysis is a particular example. It is especially useful in the case of remotely sensed data, where the *effective* impact of the numerous, but partially redundant, radiance measurements is usually much harder to quantify by other methods than conventional discrete data. Information theory has been previously applied to remotely sensed satellite observations of the atmosphere by Peckham (1974) and by Eyre (1990) to obtain an estimate of the total information supplied by the data [see also the study of Mateer (1957)]. In a recent paper, Huang and Purser (1996) show that it is possible also to extend the concept of information to a spatial density in a manner consistent with the criterion of total (integrated) information when the underlying statistics are assumed to be Gaussian. In the present context, we may apply these ideas to quantify not only the information, or information density, gained as a result of using the available data in the optimal way, but also we may compare this ideal usage with the sub-optimal analysis that results from the use of our super observations.

For Gaussian statistics, the state-space volumes of the background and analysis are proportional to the square-roots of det $\mathbf{B}$ and det $\mathbf{A}$ respectively. Thus, the total information gained to achieve an analysis with error covariance $\mathbf{A}$ is,

$$Z = -\frac{1}{2}\log_2 \det(\mathbf{A}\mathbf{B}^{-1}), \tag{4.3}$$

which Huang and Purser (1996) show to be equivalent to:

$$Z = \mathrm{Trace}(\mathbf{S}), \tag{4.4}$$

where

$$\begin{aligned} \mathbf{S} &= -\frac{1}{2}\log_2(\mathbf{A}\mathbf{B}^{-1}), & (4.5)\\ &= \frac{1}{2}\log_2(\mathbf{B}\mathbf{A}^{-1}). & (4.6) \end{aligned}$$

In (4.4) we are extending the domain of the *log* function to include square matrices. When the analysis is optimal and $\mathbf{A}$ is given by (4.1), we may construct $\mathbf{S}$:

$$\mathbf{S} = \frac{1}{2}\boldsymbol{\psi}\log_2(\mathbf{I} + \boldsymbol{\Omega})\boldsymbol{\psi}^- \tag{4.7}$$

where $\boldsymbol{\psi}$, $\boldsymbol{\psi}^{-}$, are right and left eigenvectors, $\boldsymbol{\Omega}$ the corresponding diagonal matrix of non-vanishing eigenvalues, of the square matrix, $\mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$, as given by the conditions:

$$\begin{aligned}(\mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\boldsymbol{\psi} &= \boldsymbol{\psi}\boldsymbol{\Omega}, & (4.8)\\ \boldsymbol{\psi}^{-}(\mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}) &= \boldsymbol{\Omega}\boldsymbol{\psi}^{-}, & (4.9)\\ \boldsymbol{\psi}^{-}\boldsymbol{\psi} &= \mathbf{I}. & (4.10)\end{aligned}$$

In practice, it is always easier to obtain the eigen-decomposition based on a symmetric matrix. We can achieve this, together with a valuable reduction in the dimensionality, through a generalized similarity transformation which reduces the eigen-decomposition first to:

$$(\mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{R}^{-1/2})\mathbf{V} = \mathbf{V}\boldsymbol{\Omega}. \qquad (4.11)$$

This then leads to the reconstruction of $\boldsymbol{\psi}$ and $\boldsymbol{\psi}^{-}$:

$$\begin{aligned}\boldsymbol{\psi} &= \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1/2}\mathbf{V}, & (4.12)\\ \boldsymbol{\psi}^{-} &= \mathbf{V}^{-1}\mathbf{R}^{1/2}(\mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1}\mathbf{H}, & (4.13)\end{aligned}$$

which one can readily verify to satisfy the conditions specified in (4.8)—(4.10). If only the total information Z is needed, the manipulations to get $\mathbf{S}$ can be avoided and Z obtained directly from,

$$Z = \frac{1}{2}\mathrm{Trace}\,\log_2(\mathbf{I} + \boldsymbol{\Omega}). \qquad (4.14)$$

However, as discussed by Huang and Purser (1996), it is possible to extract from the diagonal elements of $\mathbf{S}$ knowledge about the spatial distribution of the quantitative *information* resulting from the assimilation of the measurements.

We see from (4.2) that the error covariance of the sub-optimal analysis obtained from the assimilation of super observations is described by a formula of similar form to the one we have treated for the optimal case. Therefore, the same sort of eigen-decompositions can be applied to obtain the information distribution corresponding to the use of the super observations.

5. Remarks

In this note we have described methods for generalizing the concept of a "super observation" to include measures not only of the locally averaged observed value, but also of the gradient, curvature, etc. We have also discussed the application of information theory as an objective means of quantifying the information lost by the conversion of many raw data into few super observations. We suggest that the methods proposed here be considered as a working framework for the systematic reduction of the quantity of observational data when this becomes necessary in 3-D or 4-D variational assimilation by virtue of the otherwise overwhelming quantity of individual observations provided by the various remote sensing instruments. Constructing super observations from dense raw data requires that decisions be made about the size of clusters of the raw data that are combined and, with these new methods, about the degree of higher spatial derivatives that one would choose to accommodate in each super observation. It should be possible to rationalize such decision by using the methods of information theory described in section 4.

A useful property of many observation systems is that the observation errors of individual measurements are mutually independent. We note that this desirable property is automatically preserved by the process we have proposed for the generation of super observations. In this regard, we retain the simplicity of a diagonal covariance operator $\hat{\mathbf{R}}$ for the aggregated data by sweeping any complexities in the definition of these super observations into the effective observation operator, $\hat{\mathbf{H}}$. This latter operator should, in any case, be permitted a greater degree of generality in future remote sensing systems where measurements represent extended spatial integrals of the quantity under observation.

Acknowledgments

This work was partially supported by the NSF/NOAA Joint Grants Program of the US Weather Research Program. This research is also in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

Appendix A

Derivation of standard forms of the equations of 3DVAR

From the basic variational principle of linear 3DVAR, we seek state $\mathbf{x} = \mathbf{x}^a$ that minimizes the cost function:

$$\mathcal{L}(\mathbf{x}) = \frac{1}{2}\left((\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b) + (\mathbf{y^o}-\mathbf{Hx})^T\mathbf{R}^{-1}(\mathbf{y^o}-\mathbf{Hx})\right). \tag{A.1}$$

Setting $\partial\mathcal{L}/\partial\mathbf{x} = 0$ at $\mathbf{x} = \mathbf{x}^a$ leads to,

$$\mathbf{B}^{-1}(\mathbf{x}^a-\mathbf{x}^b) = \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y^o}-\mathbf{Hx}^a), \tag{A.2}$$

from which (2.3) immediately follows. However, by introducing a vector, $\mathbf{f}$, of forcing terms satisfying

$$\mathbf{Rf} = \mathbf{y^o}-\mathbf{Hx}^a, \tag{A.3}$$

(2.3) may be expressed equivalently:

$$(\mathbf{x}^a-\mathbf{x}^b) = \mathbf{BH}^T\mathbf{f}. \tag{A.4}$$

Combining (A.3) and (A.4) and solving for $\mathbf{f}$:

$$\mathbf{f} = (\mathbf{R}+\mathbf{HBH}^T)^{-1}(\mathbf{y^o}-\mathbf{Hx}^b). \tag{A.5}$$

Finally, substituting for $\mathbf{f}$ in (A.4) gives us,

$$(\mathbf{x}^a-\mathbf{x}^b) = \mathbf{BH}^T(\mathbf{R}+\mathbf{HBH}^T)^{-1}(\mathbf{y^o}-\mathbf{Hx}^b) \tag{A.6}$$

which is essentially (2.1).

Appendix B

Modified Gram-Schmidt algorithm

Suppose we wish to approximately factorize an $m_1 \times m_2$ matrix $\mathbf{H}$:

$$\mathbf{H} \approx \mathbf{GUP}, \tag{B.1}$$

to within a certain tolerance, ϵ, where the factors to right obey the following properties. Matrix $\mathbf{G}$ has $m_3 \leq m_2$ mutually orthogonal columns, g_j. $\mathbf{U}$ is a unit upper triangular matrix of m_2 columns u_j of m_3 individual elements U_{kj}. $\mathbf{P}$ is the permutation matrix operating on the m_2 columns of the matrix to its left, and which we may express as a factorization into simple transpositions, encoded by the "pivot-index" list $\{\pi_j\}$,

$$\mathbf{P} \equiv \mathbf{T}_{m_2} \cdots \mathbf{T}_2 \mathbf{T}_1, \tag{B.2}$$

where each transposition $\mathbf{T}_k$ exchanges columns k and π_k of the matrix to its left, where the total permutation $\mathbf{P}$ is chosen to make,

$$X_j \equiv g_j \cdot g_j \geq X_k \tag{B.3}$$

when $j < k$. This factorization is what we refer to as a modified pivoted Gram-Schmidt process. The task is accomplished by the algorithm given schematically below.

```
$\mathbf{G} \leftarrow \mathbf{H}$
$X_j \leftarrow g_j \cdot g_j, \quad \forall j \in [1, m_2]$
do $j = 1, m_2$
    find $\pi_j \equiv l : \forall k \geq j,\ X_l \geq X_k$
    if($l \neq j$)then
        $g_l \leftrightarrow g_j$
        $u_l \leftrightarrow u_j$
        $X_l \leftrightarrow X_j$
    endif
    if ($X_l \leq \epsilon^2$)then
        $m_3 = j - 1$
        return
    endif
    $X_j \leftarrow g_j \cdot g_j$ [*insure against roundoff!*]
    $U_{jk} \leftarrow 0,\ \forall k < j$
    $U_{jj} \leftarrow 1$
    do $k = j + 1, m_2$
        $U_{jk} \leftarrow (g_j \cdot g_k)/X_j$
        $g_k \leftarrow g_k - g_j U_{jk}$
        $X_k \leftarrow X_k - X_j U_{jk}^2$
    enddo
enddo
$m_3 \leftarrow m_2$
```

end

References

Barnes, S. L. 1964 A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.*, **3,** 396–409.

Cohn, S. E., and D. F. Parrish 1991 The behavior of forecast error covariances for a Kalman filter in two dimensions. *Mon. Wea. Rev.*, **119,** 1757–1785.

Daley, R. A. 1991 *Atmospheric Data Assimilation.* Cambridge University Press, 471 pp.

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

Gandin, L. S. 1963 *Objective Analysis of Meteorological Fields*, Leningrad, Gidromet; (Jerusalem, Isreal Program for Scientific Translations; 1965, 242pp).

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

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc 1997 Unified notation for data assimilation: operational, sequential and variational. *J. Meteor. Soc. Japan*, **75,** 181–189.

Khinchin, A. I. 1957 *Mathematical foundation of Information Theory.* New York. Dover, 123pp.

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

Lorenc, A. C. 1986 Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **112,** 1177–1194.

Mateer, C. L. 1957 On the information content of Umkehr observations. *J. Atmos. Sci.*, **22,** 370–381.

Menke, W. 1984 *Geophysical Data Analysis: Discrete Inverse Theory.* New York: Academic Press, 160pp.

Parrish, D. F., and S. E. Cohn 1985 A Kalman filter for a two-dimensional shallow-water model: Formulations and preliminary experiments, Office Note 304, National Meteorological Center, Washington, DC 20233, 64pp.

Peckham, G. 1974 The information content of remote measurements of the atmospheric temperature by satellite IR radiometry and optimum radiometer configurations. *Quart. J. Roy. Meteor. Soc.*, **100,** 406–419.

Purser, R. J. 1990 Vertical aspects of the assimilation of sounding data. *Preprints, WMO Int. Symp. on "Assimilation of Observations in Meteorology and Oceanography"*, Clerment-Ferrand, France; 9–13 July 1990, WMO Report, 501–505.

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