A sampling method for the objective thinning of IASI channels

Alison M. Fowler

Data assimilation research centre, University of Reading, Reading, UK.

November 25, 2013

Abstract

There is a vast amount of information about the current state of the atmosphere from instruments measuring the top of the atmosphere radiances on board satellites. One example is the IASI (Infrared Atmospheric Sounding Interferometer) instrument which measures radiances emitted from the Earth's atmosphere and surface in 8461 channels. In many cases it is difficult to transmit, store and assimilate such a large amount of data. A practical solution to this has been to select a subset of a few hundred channels based on those which contain the most useful information.

Different measures of information content for objective channel selection have been suggested for application to variational data assimilation. These include mutual information and the degrees of freedom for signal. To date, the calculation of these measures of information content have been based on the linear theory which is at the heart of operational variational data assimilation. However the retrieval of information about the atmosphere from the satellite radiances can be highly non-linear. Here we look at a sampling method for calculating the mutual information which is free from assumptions about the linearity of the relationship between the observed radiances and the state variables. How this new estimate of information content can be used in channel selection is addressed, with particular attention given to the efficiency of the new method.

---

1 Introduction

Satellites provides a wealth of information about the current state of the atmosphere by hosting instruments measuring the top of the atmosphere radiances. In general, the amount of data available from satellites is more than can be practically assimilated let alone stored and transmitted [Collard, 2007]. A practical solution to this has been to select a subset of a few hundred channels based on those which contain the most useful information [Collard, 2007, Rabier et al., 2002]. Within this study we will concentrate on the IASI (Infrared Atmospheric Sounding Interferometer) instrument, an infrared Fourier-transform spectrometer, on board the METOP series of satellites circumnavigating the Earth in a polar orbit. IASI measures radiances emitted from the Earth's atmosphere and surface in 8461 channels.

Different measures of information content for objective channel selection have been suggested by Rodgers [1996] and Rodgers [2000] for application to variational data assimilation. These include mutual information and the degrees of freedom for signal. To date, the calculation of these measures of the information content have been based on the linear theory which is at the heart of operational variational data assimilation. However, the retrieval of information about the atmosphere from the satellite radiances can be highly non-linear. To understand the importance and potential impact of the non-linear relationship between satellite data and the atmospheric state we shall first introduce the data assimilation problem.

Data assimilation allows for satellite data and other atmospheric observations to be combined with a NWP (numerical weather prediction) model. The result, known as the analysis, can be used to give initial conditions for the next forecast. The more accurate the analysis's representation of the true initial conditions the more accurate the forecast will be.

Many data assimilation schemes are derivable from Bayes' theorem:

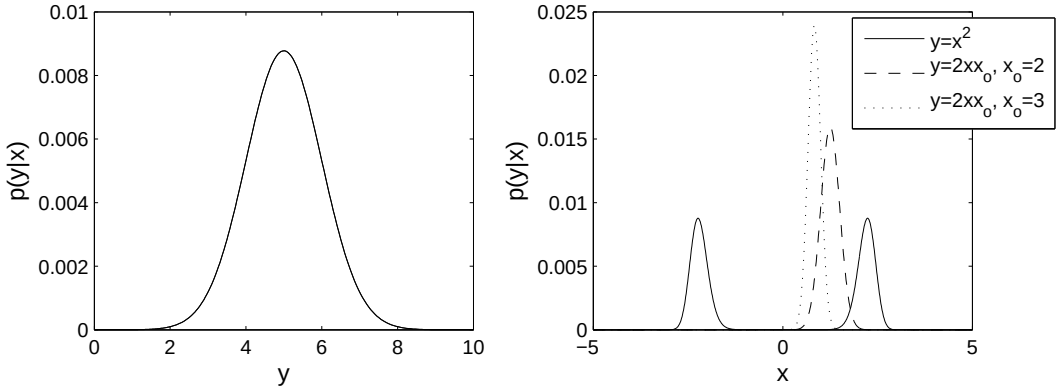

Figure 1: Illustration of the effect of a non-linear observation operator on the likelihood distribution in state space. Left-hand panel: $p(y|x) = N(5, 1)$ plotted as a function of y. Right hand panel: $p(y|x)$ this time plotted as a function of $x = \sqrt{y}$ (solid line) and as a function of the linearized estimate $x = y/2x_o$ when x_o is 2 (dashed line) and x_o is 3 (dotted line).

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \tag{1}$$

The aim is to find the posterior probability of the state given the observation, $p(\mathbf{x}|\mathbf{y})$, when the probability of the observation measuring the state, $p(\mathbf{y}|\mathbf{x})$, and the probability of the state prior to the observations being made, $p(\mathbf{x})$, are known. In (1) the marginal distribution, $p(\mathbf{y})$, is often simply thought of as a normalization factor.

An adequate approximation, in many cases, to the probability distributions $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x})$ is a Gaussian distribution. If it was then assumed that the observation operator, a transform mapping from state to observation space, were also linear then the posterior distribution would also be Gaussian. The analysis state could then be defined as the mode of the posterior distribution, giving both the most likely and minimum error variance estimate of the true state. This is a large simplification in the case of satellite data assimilation, but has proved to be useful (see e.g. Eyre [1989]).

A simple illustration of the effect of a non-linear observation operator is given in figure 1. In the left hand panel a Gaussian likelihood is shown as a function of the observation variable y. In the right hand panel the likelihood is plotted as a function of the state variable x for the case when the observation measures the square of the state variable, i.e. $y = x^2$. The likelihood (solid black line) is clearly no longer Gaussian in the state space, with the two peaks representing the uncertainty in the sign of x. From (1), this means that the posterior distribution will also be non-Gaussian.

In previous work, Fowler and van Leeuwen [2013], it was shown that approximating a non-Gaussian error distribution with a Gaussian (i.e. just allowing for the first two moments) resulted in a small underestimate of mutual information when the likelihood was in fact non-Gaussian but the observation operator was linear. In the case of approximating a non-linear observation operator with its tangent linear, the non-Gaussian structure of the likelihood in state space is again underestimated. However, the approximation is no longer as simple as fitting a smooth Gaussian to the non-Gaussian likelihood. This is illustrated in figure 1, where we see that the linearized estimate of the likelihood is very poor and strongly depends on the choice of linearization state (dashed and dotted lines in figure 1). For this reason, the results derived in Fowler and van Leeuwen [2013], which assumed that the non-Gaussian distribution and its Gaussian approximation share the same first two moments, cannot be applied here.

1.1 The observation operator

The mapping between the observation and state variable is given by the observation operator, H, plus a small measurement error, ϵ_o.

$$\mathbf{y} = H(\mathbf{x}) + \epsilon_o. \tag{2}$$

There may be errors in $H(\mathbf{x})$, for example, due to missing processes or if the observations, $\mathbf{y}$, are sampling smaller scales than can be represented by the state variables, $\mathbf{x}$, which are often given on a fixed model grid. However, we shall assume for simplicity that the error in $H(\mathbf{x})$ is negligible.

In this study the observations, $\mathbf{y}$, are TOA radiances, L^{TOA}. The state, $\mathbf{x}$, is a vector of temperature and specific humidity on 51 model levels. The TOA radiances may be modeled as follows:

$$L^{\text{TOA}}(\nu,\theta) = \tau_s(\nu,\theta)\varepsilon_s(\nu,\theta)B(\nu,T_s) + \int_{\tau_s}^{1} B(\nu,T)d\tau + (1-\varepsilon_s(\nu,\theta))\tau_s^2(\nu,\theta)\int_{\tau_s}^{1}\frac{B(\nu,T)}{\tau^2}d\tau, \tag{3}$$

where τ_s is the surface to space transmittance, ε_s is the surface emissivity and $B(\nu,T)$ is the Planck function for a frequency ν and temperature T. Recall

$$B(\nu,T) = \frac{2h\nu^2}{c^2}\frac{1}{\exp(\frac{h\nu}{kT})-1}, \tag{4}$$

where k is the Boltzmann constant, h is the Planck constant and c is the speed of light. The transmittances depend on humidity and atmospheric constituents of gases such as ozone and water vapor.

In this work RTTOV (a fast radiative transfer model developed within NWP SAF [Hocking et al., 2011]) is used to evaluate (3) for each of IASI's 8461 channels. The transmittances are computed using a linear regression in optical depth based on the input vector variables (in this case, temperature, humidity and ozone). The accuracy of the observation operator is fundamental in data assimilation, as such channels which are known to be poorly modeled are neglected in the assimilation, as are observations made in poorly modeled atmospheric conditions, e.g. regions of cloud [Chevallier et al., 2004, Pavelin et al., 2008].

1.2 Measuring information content

A measure of information content should quantify the impact of the observations on the analysis. Mutual information measures this impact as the change in entropy (uncertainty) when an observation is made. It is given in terms of the prior and posterior distributions as

$$MI = \int p(\mathbf{y}) \int p(\mathbf{x}|\mathbf{y}) \ln \frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})} d\mathbf{x} d\mathbf{y} \tag{5}$$

[Cover and Thomas, 1991]. An observation with a large impact is therefore one which results in large change in the posterior distributions compared to the prior.

MI can be interpreted as the relative entropy weighted with the probability of all possible realizations of the observations, where the relative entropy is defined as:

$$RE = \int p(\mathbf{x}|\mathbf{y}) \ln \frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})} d\mathbf{x}. \tag{6}$$

Due to the extra integral in (5), MI is independent of the realization of the observation random error. This is a beneficial property as it provides a measure of information content based on the instrument characteristics rather than the value observed. However, as will be seen, this makes it much more costly to compute in the case of a non-linear observation operator.

The focus of this work is to understand how the linearisation of the observation operator affects the information content of observations as calculated by mutual information. The impact this has on channel selection for IASI data will also provide insight into how the information content of one observation relative to another can be changed. In section 2 we will first look at how MI may be calculated in practice, introducing a method which does not rely on the assumption that the observation operator is near linear. In section 3 it will be shown how these estimates of MI may be applied to the problem of channel selection. When performing the channel selection using the non-linear estimate of MI it is

demonstrated that this method may suffer detrimentally from the problem of under sampling. This issue will be addressed in section 4. A summary of the key conclusions is then finally presented in section 5.

2 Estimating Mutual information

When a non-linear observation operator is considered it is not possible to give an analytical expression for MI. Assumptions must therefore be made. As already discussed, one assumption that has proved to be useful is that the observation operator can be linearized. The expression for MI that this leads to is given in section 2.1. Alternatively it is possible to avoid the assumption of near-linearity by sampling from the prior, $p(\mathbf{x})$, and likelihood, $p(\mathbf{y}|\mathbf{x})$, distributions and assuming that the sample size is large enough to give an accurate approximation to the posterior distribution, $p(\mathbf{x}|\mathbf{y})$, and marginal distribution, $p(\mathbf{y})$, so that an accurate estimate of MI may be given. This method for evaluating MI is given in section 2.2.

2.1 A linearized estimate

If we assume that the observation operator can be accurately linearized then the posterior and additionally the marginal distributions become Gaussian. In this case it is possible to calculate the mutual information in terms of the prior and posterior error covariances alone, $\mathbf{B}$ and $\mathbf{P}_a$ respectively,

$$MI^G = \frac{1}{2} \ln |\mathbf{B}\mathbf{P}_a^{-1}| \tag{7}$$

[Rodgers, 2000]. The superscript G refers to the Gaussian approximation.

In this linear framework, the posterior error variance is given by $\mathbf{P}_a = (\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}$. $\mathbf{H}$ is the linearized observation operator, usually linearized about the analysis which is assumed to be the mode of the posterior and $\mathbf{R}$ is the observation error covariance matrix. This estimate of MI is therefore sensitive not only to linearisation error in the observation operator (a function of the state) but also to the estimates of the prior and observation error covariances and fundamentally the assumption that these alone are enough to characterize the prior and likelihood.

2.2 A non-linear estimate

Here we propose a method to calculate the mutual information without linearizing the observation operator. To calculate (5), it is necessary to have an estimate of the probability distributions; $p(\mathbf{x})$, $p(\mathbf{y})$ and $p(\mathbf{x}|\mathbf{y})$. Due to the non-linear mapping between the state and observation space it is not possible to give an analytical expression for $p(\mathbf{y})$ and $p(\mathbf{x}|\mathbf{y})$. Instead we propose a sampling method to approximate these distributions.

Let $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x})$ have Gaussian distributions with means $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ and covariances $\mathbf{R}$ and $\mathbf{B}$ respectively. Note that the proposed method is not restricted to these assumptions but in order to generate the initial distributions some assumptions are necessary.

In order to represent $p(\mathbf{y})$ and $p(\mathbf{x}|\mathbf{y})$ we shall first take M samples from $p(\mathbf{y}|\mathbf{x})$ and N samples from $p(\mathbf{x})$:

$$\begin{array}{ll} \mathbf{x}_i \sim N(\boldsymbol{\mu}_x, \mathbf{B}) & \text{for } i = 1, ..., N \\ \mathbf{y}_j \sim N(\boldsymbol{\mu}_y, \mathbf{R}) & \text{for } j = 1, ..., M \end{array} \tag{8}$$

The prior distribution can now be expressed as a sum of delta functions

$$p(\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i). \tag{9}$$

Similarly the likelihood can now be expressed as

$$p(\mathbf{y}|\mathbf{x}) \approx \frac{1}{M} \sum_{j=1}^{M} \delta(\mathbf{y} - \mathbf{y}_j). \tag{10}$$

Substituting (9) into (1) allows for the posterior distribution conditioned on the j^{th} sample from $p(\mathbf{y}|\mathbf{x})$ to be expressed as a weighted sum of delta functions:

$$p(\mathbf{x}|\mathbf{y}_j) = \sum_{i=1}^{N} w_{i,j}\delta(\mathbf{x} - \mathbf{x}_i). \tag{11}$$

where these weights are given by

$$w_{i,j} = \frac{p(\mathbf{y}_j|\mathbf{x}_i)}{p(\mathbf{y}_j)}. \tag{12}$$

$p(\mathbf{y}_j|\mathbf{x}_i)$, is evaluated using the prescribed Gaussian distribution. It is then assumed that the sample from $p(\mathbf{x})$ is large enough to imply

$$p(\mathbf{y}_j) = \sum_{i=1}^{N} p(\mathbf{y}_j|\mathbf{x}_i). \tag{13}$$

This has the effect of normalizing the weights so that $\sum_{i=1}^{N} w_{i,j} = 1$.

Given equations (11) and (9) it is now possible to evaluate the relative entropy given by the j^{th} sample from $p(\mathbf{y}|\mathbf{x})$. Substituting these expression into (6), the relative entropy for this sample from the likelihood is given by

$$RE_j = \sum_{i=1}^{N} w_{i,j} \ln\left(N w_{i,j}\right). \tag{14}$$

It is possible to express RE in this form due to the co-location of the sample representing the prior and posterior. Such an expression would therefore not be possible if a direct sample from the posterior was made, e.g. using a Markov chain Monte Carlo type method as in Tamminen and Kyrölä [2001].

Performing this calculation for each of the M samples from the likelihood allows us to build up the statistics for $p(\mathbf{y})$ to then be able to calculate the mutual information as

$$MI = \sum_{j=1}^{M} \left(\sum_{i=1}^{N} p(\mathbf{y}_j|\mathbf{x}_i) \right) RE_j. \tag{15}$$

This estimate of MI is clearly more computationally expensive than the linear estimate given by (7). However, given a large enough sample this estimate should have a much smaller error giving a better evaluation of the 'true' information content of the satellite channels. In doing so we can then assess how detrimental the linear approximation is.

2.3 Mutual information of IASI channels

Before comparing the two different estimates of MI, we begin by looking at the convergence rate of the sampling estimate of MI described in section 2.2. From experiments (not shown) it is known that the estimate given by (15) is most sensitive to the size of N rather than M. For this reason M will be kept fixed at a value of 100 for the remainder of the experiments and the sensitivity of MI to the value of N alone is now studied. The $\mathbf{B}$ and $\mathbf{R}$ error covariance matrices, necessary for generating the initial samples, have been provided by the NWPSAF 1DVar package. The 'true' atmospheric profile represents cloud free conditions.

Figure 2 shows the convergence of MI with increasing sample size N for 6 different channels of IASI (blue stars). For each choice of N, MI has been estimated 10 times with different realizations of the random error in the observations and prior estimate. It can be seen that the convergence rate of MI depends on the channel number. Most channels seem to have begun to converge by $N = 5000$. The small sample error that is present should not be enough to impact on the channel selection. It has therefore been decided from these experiments to keep N =5000.

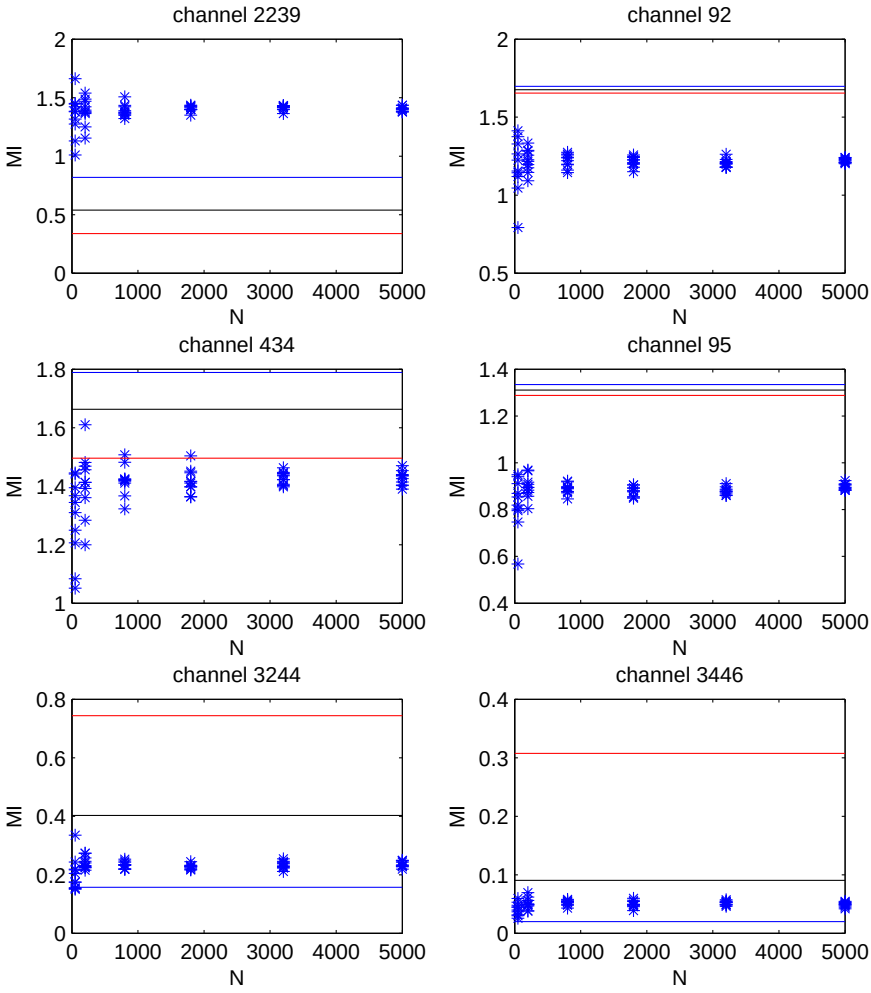

Figure 2: MI approximated using the sampling method (blue stars, discussed in section 2.2) for different random realization of the prior and likelihood when $M = 100$ and N is allowed to vary. The solid lines show the linear estimate of MI (see (7)) when the observation operator has been linearized about the truth, $\mathbf{x}_{\text{truth}}$ (black line), $\mathbf{x}_{\text{truth}} - \sigma_b$ (red line) and $\mathbf{x}_{\text{truth}} + \sigma_b$ (blue line).

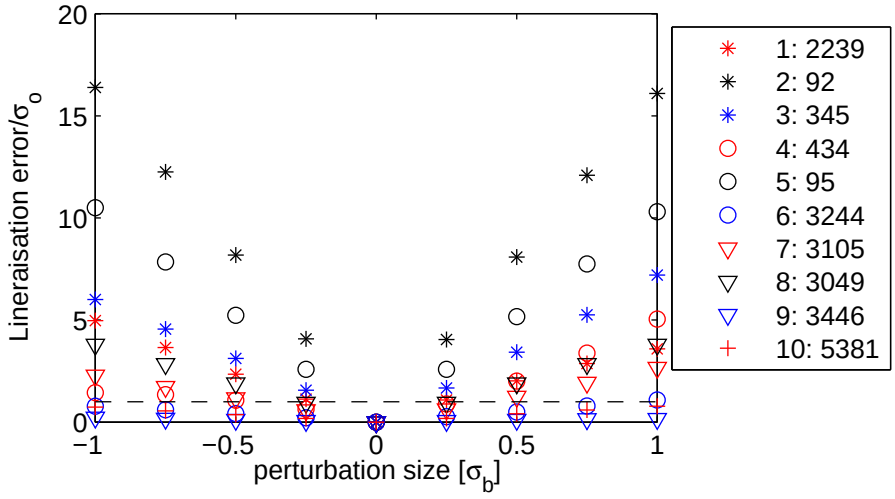

Figure 3: Linearisation error, ε_{lin}, normalized by the observation error standard deviation (σ_o) as a function of perturbation size (a fraction of the background error standard deviation, σ_b). The dashed line shows $\varepsilon_{lin}/\sigma_o = 1$.

For these 6 channels we now compare the sample estimate of MI to the linear estimates, given in 2.1. In figures 2 the linear estimates to MI calculated as $0.5 \ln |\mathbf{I}_n + \mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}|$ (see (7)) are given by the solid lines. The three lines represent estimates when the observation operator has been linearized about i) $\mathbf{x}_{\mathrm{truth}}$ (black line) ii) $\mathbf{x}_{\mathrm{truth}} - \sigma_b$ (red line) and iii) $\mathbf{x}_{\mathrm{truth}} + \sigma_b$ (blue line), where σ_b is the background error standard deviation (square root of the diagonal elements of $\mathbf{B}$). In practice the observation operator is linearized about the analysis, assumed to be much closer to the truth than $\mathbf{x}_{\mathrm{truth}} \pm \sigma_b$.

The accuracy of the linear estimate of MI (as compared to the sample estimate) and its sensitivity to the linearisation state differs for each of the channels. For some channels (e.g. channel 3244) the sample estimate is within the range of the linear estimate indicating that the observation operator may be considered near linear, whilst for others (e.g. channel 95) the sample estimate is well outside the range of the linear estimate.

These results give an indication of the size of the error caused by the linear estimate to the observation operator. This can be corroborated by plotting a measure of the linearisation error of the observation operator for each of the channels. The linearisation error can be quantified as:

$$\epsilon_{lin} = H(\mathbf{x} + \delta\mathbf{x}) - H(\mathbf{x}) - \mathbf{H}\delta\mathbf{x}. \tag{16}$$

This can be deemed adequately small if ϵ_{lin} is much smaller than the observation error.

In figure 3, the linearisation error normalized by the standard deviation of the observation error is plotted as a function of perturbation size, $\delta\mathbf{x}$, (a fraction of the standard deviation of the background error). It is seen that a large error in the linear estimate to mutual information (see figure 2) corresponds to a large linearisation error.

Details of the 10 IASI channels given in figure 3 can be found in table 1. The first column refers to the channel selection experiments performed in section 3. The second column gives the IASI channel number (ranging from 1 to 8461). In the third and forth columns, details of the wavelength and wave number are given. The final column refers to the order in which the channels were selected by Collard [2007]. 'Temp' is the initial channel selection in which channels most sensitive to water vapor or ozone were removed so that the temperature information primarily comes from the 'relatively-linear' CO_2 channels. 65 channels were selected by Collard [2007] in this initial selection. 'Main' refers to the channels selection when water vapor channels were reintroduced.

Channel selection no.	IASI channel no.	wavelength (μm)	wavenumber (cm^{-1})	Collard [2007]
1	2239	8.3	1205	Temp 1
2	92	15.0	668	Temp 2
3	345	13.7	731	Temp 3
4	434	13.2	753	Temp 4
5	95	14.9	669	Temp 5
6	3244	6.9	1456	Main 1
7	3105	7.0	1421	Main 2
8	3049	7.1	1407	Main 3
9	3446	6.6	1506	Main 4
10	5381	5.0	1990	Main 5

Table 1: Channels used within the selection.

3 Channel selection for IASI instrument

In the last section it was shown that there are indeed instances when the linear and non-linear estimates of mutual information can provide very different results. The impact these differences will have on applications such as channel selection will depend on how the relative values of mutual information between the different channels differs for the two different estimates.

A method similar to that of Collard [2007] and Rabier et al. [2002] can be followed for the channel selection:

1. Initially channels which are known to be poorly modeled by RTTOV are removed from the channels available for selection. e.g. those dominated by trace species,
2. MI is calculated for each of the remaining channels.
3. The channel with the greatest MI is selected.
4. The prior is then updated given the information from this channel choice.
5. Steps 2-4 of the channel selection are repeated until the required number of channels have been selected.

This is a timely procedure which is performed off-line. To deal with the non-linearity Collard [2007] and Rabier et al. [2002], whilst using a linear estimate of MI, repeated this channel selection procedure for a number of different atmospheric states and averaged the results.

3.1 Some initial channel section results

An initial attempt at channel selection has been performed for a subset of 10 IASI channels (see table 1). These channels have been chosen as those selected by Collard [2007] as to have the greatest information content[1]. The weighting functions of the ten channels used are given in figure 4. It can be seen that these channels are sensitive to temperature and humidity throughout the lower atmosphere as well as providing information about temperature for the upper levels.

In the top two panels of figure 5, MI and the subsequent channel section is shown for when the sample estimate of MI is used (left) and when the linear estimate of MI is used (right). As suggested by figure 2 the values of MI for the initial selection (first column) can differ significantly between the two different estimates. Sometimes the linear estimate attributes a greater amount of information to a channel than the sampling estimate (e.g. channels 2, 3, 4, 5 and 8) and sometimes the linear estimate attributes a

[1]The channel selection has been restricted to this subset of IASI channels in order to give a clear illustration of the channels selection algorithm, however there is no reason the channel selection may not be performed on the full channel list.

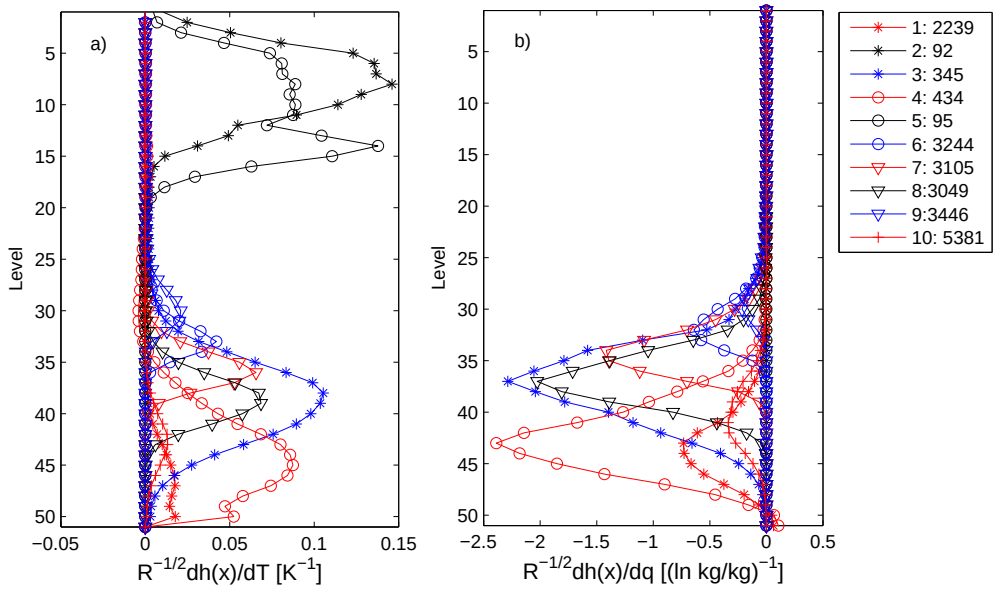

Figure 4: Weighting functions normalized by the observation error standard deviation, for the ten channels used in the channel selection. a) Sensitivity to changes in temperature. b) Sensitivity to changes in humidity.

lesser amount of information to a channel than the sampling estimate (e.g. channel 1). This is seen to lead to the channels being selected in a different order. For example the first five channels selected using the sampling method are 4, 2, 3, 1 and 5, whilst the first five channels selected using the linear method are 3, 2, 4, 5 and 7.

In figure 5c), the effective sample size, ess, of the sample estimate is shown. This is defined as

$$ess_j = 1/\sum_{i=1}^{N}(w_{i,j})^2 \tag{17}$$

and gives an estimate of the number of samples that have any significance in approximating the posterior distribution. If the weights are all equal (i.e. $1/N$) then the effective sample size is N. ess decreases as the variance of the weights used to describe the posterior distribution in (12) increases.

It is seen that the channel selected corresponds to the largest reduction in ess because this channel has had the greatest impact in refining the area of high probability. The samples at the centre of the distribution, where the probability is high, are given a large weight whilst the samples on the periphery of the distribution, where the probability is low, are given a small weight and are effectively discarded.

After the first channel is selected, ess reduces quickly until at the end of the selection process there is only one sample with any significance in representing the posterior. Therefore the error in the estimate of MI used for channel selection becomes progressively worse as each channel is selected. The size of the error in the sampled estimate after the first channel selection, means that it not useful for subsequent channel selection. This problem will increase as the number of channels to be selected increases and the amount of information available in consecutive observations is increased.

4 Improving the effective sample size

As seen in figure 2, the estimate of mutual information is sensitive to the sample size. We would therefore like to have some control over the effective sample size so that it remains constant throughout the channel selection procedure. For this reason increasing the sample size is not the solution; firstly an unnecessary (and unfeasibly) high sample size at the beginning of the channel selection would be needed in order for the effective sample size to be adequate by the end of the channel selection, and secondly the accuracy

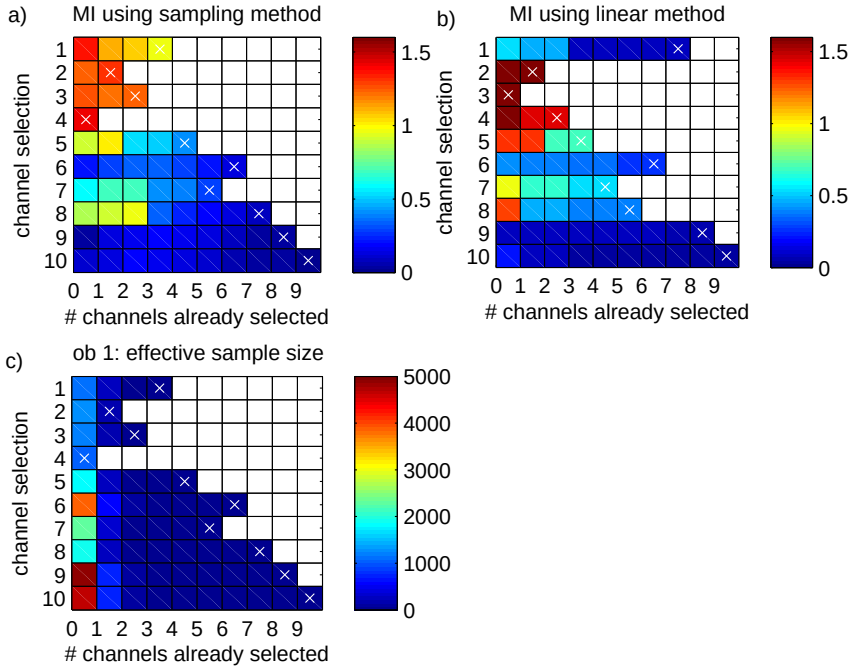

Figure 5: Channel selection for a subset of 10 channels given in table 1. a) Channel selection using sample estimate of MI. b) Channel selection using linear estimate of MI. c) Effective sample size of the sampling estimate.

of the MI estimate would change throughout the channel selection process as the effective sample size decreases.

The problem of a small effective sample size is a common problem in the particle filtering technique. As such there is a large amount of literature discussing possible options for overcoming this problem (see van Leeuwen [2009] for a review of proposed techniques).

One idea would be to resample from the current sample after each channel selection, replicating samples with a large weight and deleting samples with a small weight (see Gorden et al. [1993]). This idea has been used extensively in the particle filter (e.g. Kim et al. [2003], Lui and Chen [1998], van Leeuwen [2003]) but because we do not include a stochastic model (dynamic or otherwise) there is no way for identical samples to differ as the channel selection progresses. As such the accuracy of the estimate of MI would not increase despite the value of ess remaining high.

A more sophisticated approach would be to make use of a proposal density (e.g. van Leeuwen [2010, 2011]). The idea being that we sample from a proposal density which is similar to the posterior distribution which we wish to represent. This generally makes use of the observations to 'draw' the sample towards the region of high likelihood. This is complicated in the case of channel selection because a) we do not know *a priori* which channel will be selected and b) we need to average over observation space. Therefore this technique would involve a prohibitory large number of forward runs of RTTOV.

An alternative approach would be to generate a new sample from the prior updated after each channel selection. This would reset the sample size back to N after each channel is selected. To do this we would need to fit a PDF to our weighted sample representation of the posterior after each channel selection. Due to the non-linear observation operator we expect the posterior to be non-Gaussian, and would like to keep any non-Gaussian structure within our sample. As such we wish to consider moments greater than the first and second order.

We propose fitting a Gaussian mixture to the sample with the number of Gaussian components chosen such that each component is represented by at least 300 samples. This number has been chosen somewhat arbitrarily but should be large enough to ensure a good estimate of the covariance matrix for each of the Gaussian components (our state size is 102), whilst still being small enough to allow for a good deal of structure in the fitted distribution. The idea of using a Gaussian mixture model has been applied to the particle filter by Smith [2007] and Hoteit et al. [2012]. A similar approach, which we do not consider, is

resampling using kernel density estimation (e.g. Musso et al. [2001])

The Gaussian mixture model is given by

$$p(\mathbf{x}) = \sum_{k=1}^{G} \alpha_k N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{18}$$

where G is the number of Gaussian components. We therefore need to find $3G$ parameters: α_k (the weights of each of the Gaussian components), $\boldsymbol{\mu}_k$ (the means of the Gaussian components), $\boldsymbol{\Sigma}_k$ (the covariances of the Gaussian components). These parameters may be found using the expectation-maximization, EM, method (see Bishop [2006] for an introduction). This is an iterative method and so to supply the first guess of the parameters a k-clustering algorithm can be used, which assigns each of the samples to different groups (again see Bishop [2006] for an introduction).

Once the Gaussian mixture has been fitted to the sample it is straight forward to draw a new sample of size N from this distribution. Each of the new samples has equal weight and so the effective sample size is returned to N.

An example of some of the sample estimates of the posterior joint distributions are shown in figure 6. Skewed and multi-modal distributions are evident. The new sample generated from fitting a Gaussian mixture is shown in figure 7. In this case 7 Gaussian components are necessary to describe the structure in the distribution.

In figure 8 the channel selection is repeated this time resampling from a Gaussian mixture distribution after each channel is selected. After the first channel is selected there are some small differences in the value of the sample estimate to MI (compare to figure 5). Although the effect on the channel selection appears to be small, the difference in the effective sample size after each channel selection is substantial giving greater confidence in the statistical estimates. As such this method should be necessary when performing channel selection for the full list of available channels.

5 Discussion

Satellite observations are a non-linear function of the atmospheric state variables of interest. As such a linear estimate of their information content may be erroneous. Within this paper we have illustrated the potential effect of assuming a linear relationship between the observations and state variables by looking at how this can change the choice of channels for data assimilation.

Many different measures of information content have been used for channel selection. We have focused on mutual information as this takes into account the impact of the observations on the full posterior density function. In order to estimate mutual information a sampling technique which is free from assumptions about linearity has been developed. This has shown that for some channels the linear approximation is indeed poor and can lead to a different interpretation of the observation's value.

In order to obtain a good estimate of mutual information the sample size needs to remain high throughout the channel selection process. This was a fundamental flaw with the original scheme proposed as the effective sample size can be seen to decrease as the number of channels selected is increased and the region of high probability is reduced. This problem can be alleviated by fitting a Gaussian mixture to the weighted sample after each channel has been selected. Re-sampling from this given distribution resets the effective sample size back to the chosen value, N.

In the previous studies of Collard [2007] and Rabier et al. [2002] the channel selection was performed 'off line' giving an optimal set of channels over a range of atmospheric conditions. The channel list was then averaged, for example by taking the most frequently selected channels, to give a list which could be applied to all atmospheric conditions. This helps to reduce some of the effect of the non-linearity. An advantage of the proposed sampling method is that, by accounting explicitly for the non-linearity, it is possible to give an optimal channel list for a specific prior distribution.

It is important to note that taking into account the non-linearity of the observation operator in the channel selection is only beneficial if this is consistent with the way the observations are to be assimilated, i.e. if the observation operator is not assumed to be linear in the assimilation method. There is currently much interest in developing data assimilation techniques applicable to the geosciences in which the assumption of linearity and Gaussian error statistics are relaxed. The author therefore anticipates

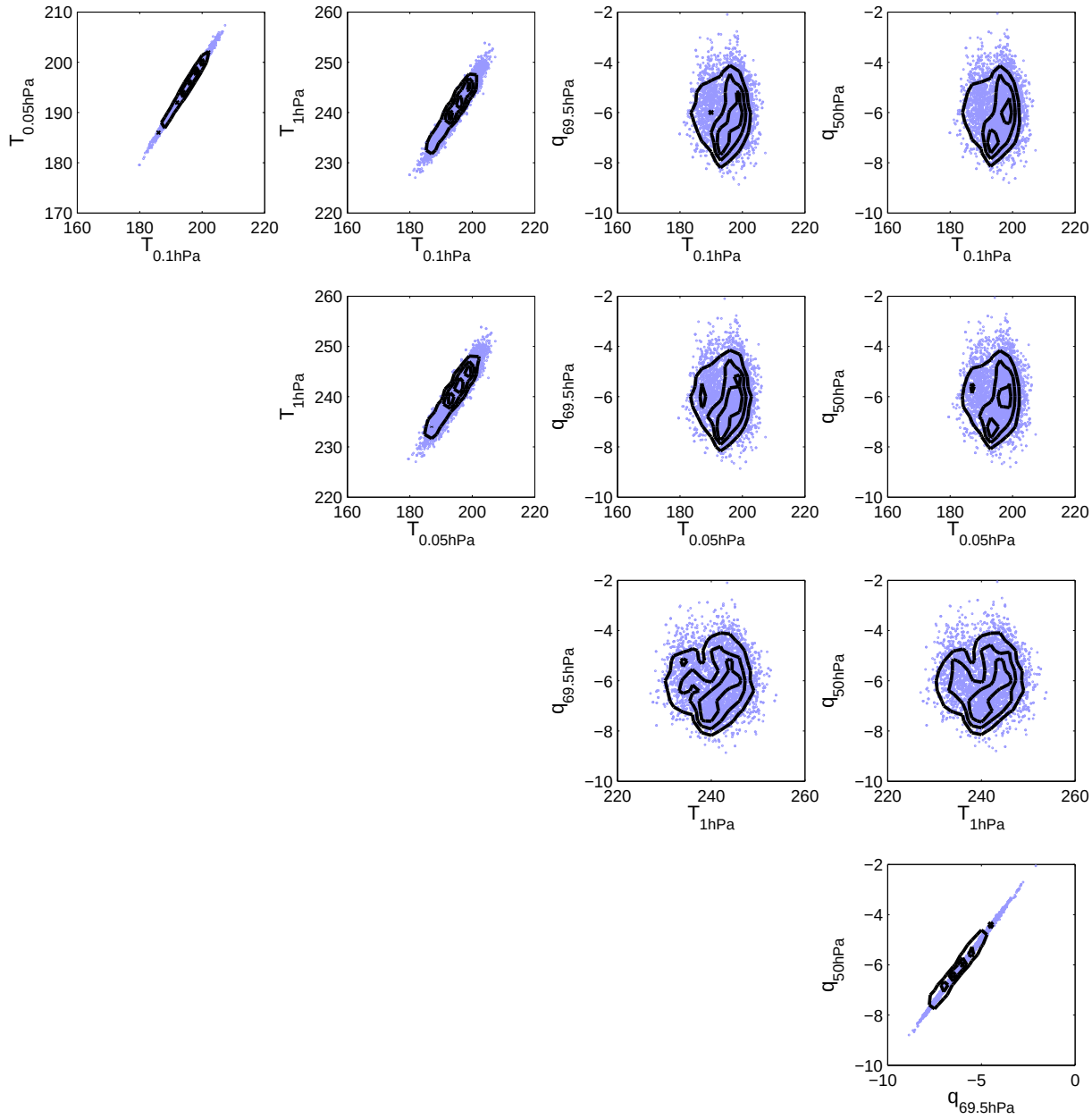

Figure 6: Example of some joint distributions from the posterior distribution after the 10th channel is selected.

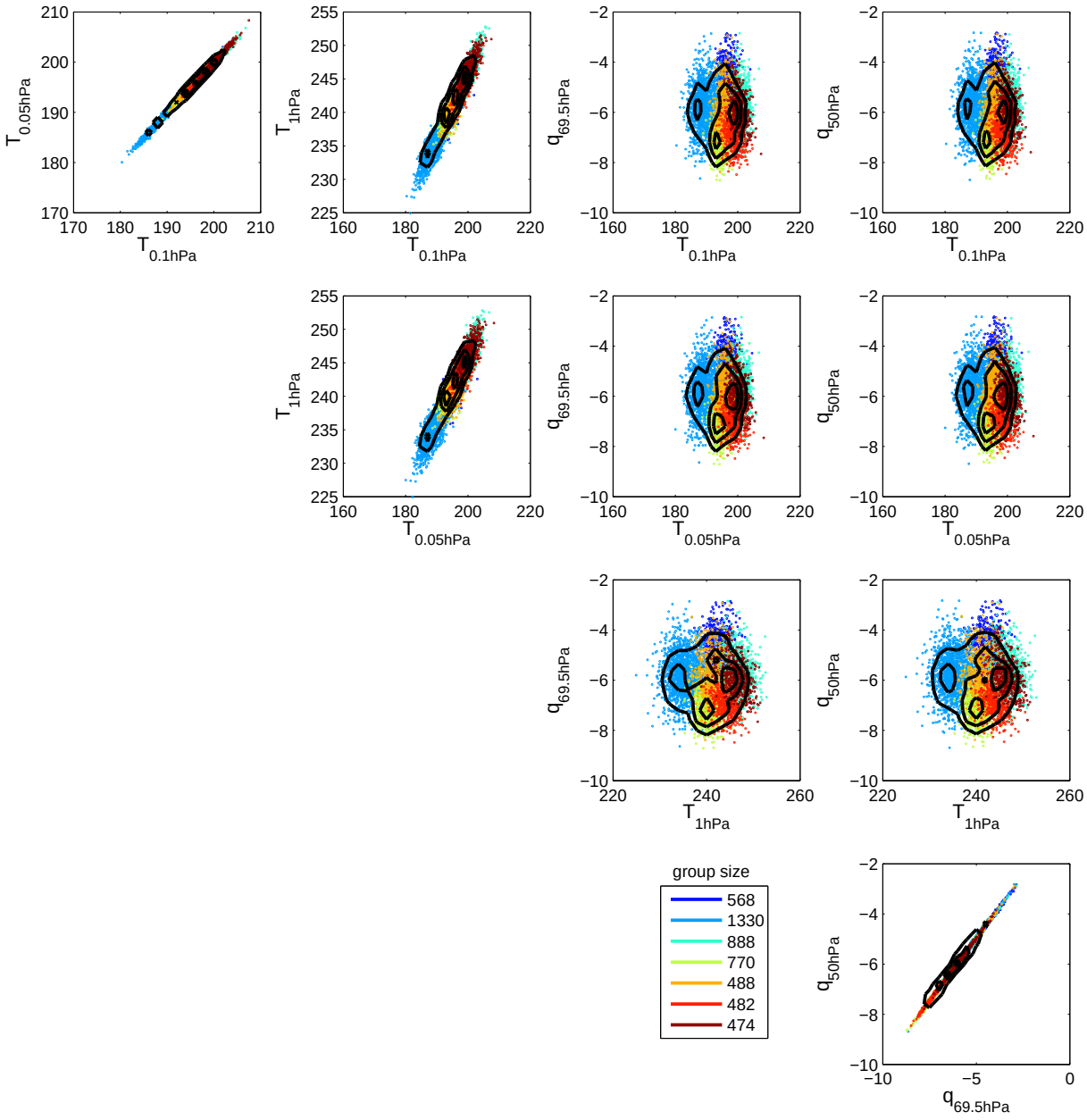

Figure 7: GM resampling of the sample given in figure 6. The different colors represent the different components of the Gaussian mixture, the legend shows the number of samples drawn from the individual Gaussian components.

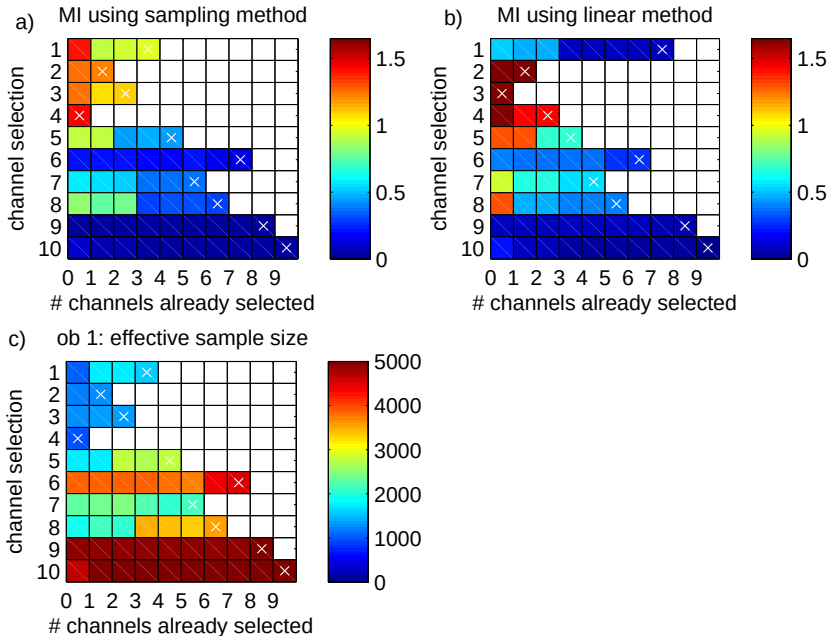

Figure 8: Same as figure 5 but using GM resampling.

the need to re-asses the information content of observations in these advanced data assimilation systems.

Acknowledgments The author would like to thank Peter Jan van Leeuwen and Stefano Migliorini for their valuable feedback on this manuscript. This work has been funded under the project "ESA Advanced Data Assimilation Methods" contract number ESRIN 4000105001/11/I-LG.

References

C. Bishop. *Pattern recognition and machine learning.* Springer Science + Business Media, LLC, New York, 2006.

F. Chevallier, P. Lopez, A. M. Tompkins, M. Janisková, and E. Moreau. The capability of 4D-Var systems to assimilate cloud-affected satellite infrared radiances. *Q. J. R. Met. Soc.*, 130:917–932, 2004. doi: 10.1256/qj.03.113.

A. D. Collard. Selection of IASI channels for use in numerical weather prediction. *Q. J. R. Met. Soc.*, 133:1977–1991, 2007.

T. M. Cover and J. A. Thomas. *Elements of information theory (Wiley series in Telecommunications).* John Wiley and Sons, Inc, New York., 1991.

J. R. Eyre. Inversion of cloudy satellitee sounding radiances by nonlinear optimal estimation. I: Theory and simulation for TOVS. *Q. J. R. Met. Soc.*, 115:1001–1026, 1989.

A. M. Fowler and P. J. van Leeuwen. Measures of observation impact in data assimilation: the effect of a non-Gaussian measurement error. *Tellus*, 65:20035, 2013.

N. J. Gorden, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proc.*, 144:107–113, 1993.

J. Hocking, P. Rayer, R. Saunders, M. Marticardi, A. Geer, and P. Brunel. RTTOV v10 Users Guide. Technical report, NWPSAF-MO-UD-023, 2011.

I. Hoteit, X. Luo, and D.-T. Pham. Particle Kalman filtering. a nonlinear Bayesian framework for ensemble Kalman filters. *Mon. Wea. Rev.*, 140:528–542, 2012.

S. Kim, G. L. Eyink, J. M. Restrepo, F. J. Alexander, and G. Johnson. Ensemble filtering for nonlinear dynamics. *Mon. Wea. rev.*, 131:2586–2594, 2003.

J. S. Lui and R. Chen. Sequential monte-carlo methods for dynamical systems. *J. Amer. Stat. Assoc.*, 90:567–576, 1998.

C. Musso, N. Oudjane, and F. Le Gland. *Sequential Monte Carlo methods in practice*, chapter Improving regularized particle filters, page 247. Springer-Verlag, New York, 2001.

E. G. Pavelin, S. J. English, and J. R. Eyre. The assimilation of cloud-affected infrared satellite radiances for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, 134:737–749, 2008.

F. Rabier, N. Fourrié, D. Chafa i, and P. Prunet. Channel selection methods for infrared atmospheric sounding interferometer radiances. *Q. J. R. Met. Soc.*, 128:1011–1027, 2002.

C. D. Rodgers. *Inverse methods for atmospheric sounding.* World Scientific Publishing, Singapore., 2000.

C. D. Rodgers. Information content and optimisation of high spectral resolution measurements. *Proc. SPIE*, 2830:136–147, 1996.

K. W. Smith. Cluster ensemble kalman filter. *Tellus*, 59A:749–757, 2007.

J. Tamminen and E. Kyrölä. Bayesian solution for nonlinear and non-Gaussian inverse problems by Markov chain Monte Carlo method. *J. Geophys. Res.*, pages 14377–14390, 2001.

P. J. van Leeuwen. A variance-minimizing filter for large-scale applications. *Mon. Wea. Rev*, 131:2071–2084, 2003.

P. J. van Leeuwen. Particle filtering in geophysical systems. *Mon. Wea. Rev.*, 137:4089–4114, 2009.

P. J. van Leeuwen. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Q. J. Meteorol. Soc.*, 136:1991–1996, 2010.

P. J. van Leeuwen. Efficient non-linear data assimilation in geophysical fluid dynamics. *Computers and Fluids*, 2011. doi: 10.1016/j.compfluid.2010.11.011.