

A sampling method for quantifying the information content of IASI channels

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Fowler, A. M. ORCID: <https://orcid.org/0000-0003-3650-3948>
(2017) A sampling method for quantifying the information
content of IASI channels. Monthly Weather Review, 145 (2).
pp. 709-725. ISSN 0027-0644 doi: 10.1175/MWR-D-16-0069.1
Available at <https://centaur.reading.ac.uk/68961/>

It is advisable to refer to the publisher's version if you intend to cite from the
work. See [Guidance on citing](#).

Published version at: <https://doi.org/10.1175/MWR-D-16-0069.1>

To link to this article DOI: <http://dx.doi.org/10.1175/MWR-D-16-0069.1>

Publisher: American Meteorological Society

All outputs in CentAUR are protected by Intellectual Property Rights law,
including copyright law. Copyright and IPR is retained by the creators or other
copyright holders. Terms and conditions for use of this material are defined in
the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

A Sampling Method for Quantifying the Information Content of IASI Channels

ALISON MARGARET FOWLER

National Centre for Earth Observation and Data Assimilation Research Centre, University of Reading, Reading, United Kingdom

(Manuscript received 23 February 2016, in final form 21 October 2016)

ABSTRACT

There is a vast amount of information about the atmosphere available from instruments on board satellites. One example is the Infrared Atmospheric Sounding Interferometer (IASI) instrument, which measures radiances emitted from Earth's atmosphere and surface in 8461 channels. 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 that 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 has been based on the linear theory that is at the heart of operational variational data assimilation. However, the retrieval of information about the atmosphere from the satellite radiances can be highly nonlinear.

Here, a sampling method for calculating the mutual information that is free from assumptions about the linearity of the relationship between the observed radiances and the state variables is examined. It is found that large linearization errors can indeed lead to large discrepancies in the value of mutual information. 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. It is anticipated that accounting for the nonlinearity in the channel selection will be beneficial when using nonlinear data assimilation methods currently in development.

1. Introduction

Satellites provide 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 that contain the most useful information (Collard 2007; Rabier et al. 2002). Within this study we will concentrate on the Infrared Atmospheric Sounding Interferometer (IASI) instrument, an infrared Fourier transform spectrometer, on board the MetOp series of satellites in a polar orbit of Earth. IASI measures

radiances emitted from 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, 27–39) 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 has been based on the linear theory that is at the heart of operational variational data assimilation. However, the retrieval of information about the atmosphere from the satellite radiances can be highly nonlinear. To understand the importance and potential impact of the nonlinear 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 numerical weather prediction (NWP) model. The result,

 Denotes content that is immediately available upon publication as open access.

Corresponding author e-mail: Alison Margaret Fowler, a.m.fowler@reading.ac.uk



This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

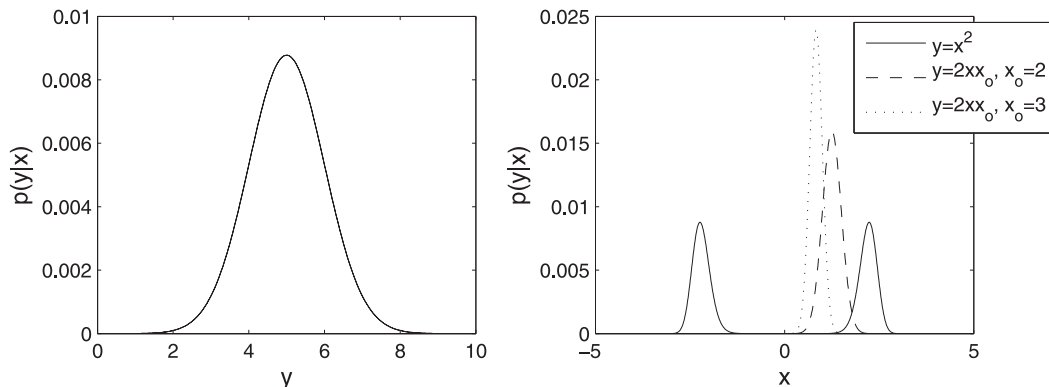


FIG. 1. Illustration of the effect of a nonlinear observation operator on the likelihood distribution in state space: (left) $p(y|x) = N(5, 1)$ plotted as a function of y and (right) $p(y|x)$, this time plotted as a function of $x = \sqrt{y}$ (solid line) and as a function of the linearized estimate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

known as the analysis, can be used to give initial conditions for the next forecast.

Many data assimilation schemes are derivable from Bayes's theorem, which states

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

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 as it is independent of $\mathbf{x}$.

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, was 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 proven to be useful (see, e.g., Eyre 1989).

A simple illustration of the effect of a nonlinear observation operator is given in Fig. 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; that is, $y = x^2$. 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 the information content of the observations when the likelihood was in fact non-Gaussian but the observation operator was linear. In the case of approximating a nonlinear 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 Fig. 1, where we see that the linearized estimate of the likelihood is very poor and strongly depends on the choice of the linearization state (dashed and dotted lines in Fig. 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.

a. The observation operator

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

$$\mathbf{y} = H(\mathbf{x}) + \varepsilon_o. \quad (2)$$

There may be uncertainty in $H(\mathbf{x})$, for example, because of missing processes or if the observations $\mathbf{y}$ are sampling scales smaller than can be represented by the state variables $\mathbf{x}$. The latter is often referred to as representation error. However, for simplicity we shall assume that the error in $H(\mathbf{x})$ is negligible.

In this study the observations $\mathbf{y}$ are top-of-the-atmosphere (TOA) brightness temperatures T_B , which can be directly related to TOA radiances L^{TOA} using

Planck's law (e.g., [Salby 1996](#), p. 209). The state $\mathbf{x}$ is a vector of temperature and specific humidity on 51 model levels. The TOA radiances may be modeled as a function of the frequency ν and the angle of incidence θ as follows:

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

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 ([Hocking et al. 2011](#)). Recall

$$B(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{\exp\left(\frac{h\nu}{kT}\right) - 1}, \quad (4)$$

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

In this work RTTOV, which is a fast radiative transfer model developed within EUMETSAT's Satellite Application Facility for Numerical Weather Prediction [NWP SAF; see [Hocking et al. \(2011\)](#)] is used to evaluate (3) for each of the considered IASI channels. The transmittances are computed using a linear regression approach in optical depth based on the input vector variables (in this case, temperature, humidity, and trace gases). The accuracy of the observation operator is fundamental in data assimilation, and channels that are known to be poorly modeled are neglected in the assimilation, as are observations made in poorly modeled atmospheric conditions, for example, regions of cloud ([Chevallier et al. 2004](#); [Pavelin et al. 2008](#)).

b. Measuring information content

A measure of information content should quantify the impact of the observations on the analysis. Mutual information (MI) 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

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

([Cover and Thomas 1991](#), 13–42). An observation with a large impact is therefore one that results in a large change in the posterior distribution compared to the prior.

In this case, MI can be interpreted as the relative entropy weighted with the probability of all possible

realizations of the observations, where the relative entropy (RE) is defined as

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

Because of 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 (i.e., the way its measurement relates to the state variable and its error statistics) rather than the value observed. However, as will be seen, this makes it much more costly to compute in the case of a nonlinear observation operator.

The focus of this work is on understanding how the linearization 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 that 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 nonlinear estimate of MI, it is demonstrated that this method may suffer detrimentally from the problem of undersampling. This issue will be addressed in [section 4](#) and in [section 5](#) we will see how this allows us to apply the algorithm to a more realistically sized problem. A summary of the key conclusions is then finally presented in [section 6](#).

2. Estimating mutual information

When a nonlinear 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 2a](#). 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 the marginal distribution, $p(\mathbf{y})$, so that an accurate estimate of MI may be given. This method for evaluating MI is described in [section 2b](#).

a. A linearized estimate

If we assume that the observation operator can be accurately linearized, then the posterior and additionally

the marginal distributions become Gaussian (under the assumption that both the prior and likelihood are 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:

$$\text{MI}^G = \frac{1}{2} \ln |\mathbf{B} \mathbf{P}_a^{-1}| \quad (7)$$

(Rodgers 2000, 27–39). The superscript G refers to the Gaussian approximation.

Within this linear framework, the posterior error variance is given by $\mathbf{P}_a = (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1}$, where $\mathbf{H}$ 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 linearization 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.

b. A nonlinear estimate

Here, we propose a method for calculating 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 $p(\mathbf{x}|\mathbf{y})$. Because of the nonlinear mapping between the state and observation space, it is not possible in general to give an analytical expression for $p(\mathbf{y})$ and $p(\mathbf{x}|\mathbf{y})$. Instead, we propose a sampling method for approximating 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 fact in section 3, when an iterative selection of the channels with the highest MI is performed, the prior distribution is not assumed to be Gaussian after the first iteration.

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 $p(\mathbf{x})$:

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

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

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

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:

$$p(\mathbf{x}|\mathbf{y}_j) = \sum_{i=1}^N w_{ij} \delta(\mathbf{x} - \mathbf{x}_i), \quad (10)$$

where these weights are given by

$$w_{ij} = \frac{p(\mathbf{y}_j|\mathbf{x}_i)}{N p(\mathbf{y}_j)}. \quad (11)$$

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

$$p(\mathbf{y}_j) = \int p(\mathbf{x}, \mathbf{y}_j) d\mathbf{x}. \quad (12)$$

Using Bayes's theorem and (9), we see that this can be evaluated as

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

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

Given (10) 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 expressions into (6), the relative entropy for this sample from the likelihood is given by

$$\text{RE}_j = \sum_{i=1}^N w_{ij} \ln(N w_{ij}). \quad (14)$$

It is possible to express RE in this form because of the collocation 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, for example, 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.

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, leading to 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.

c. Mutual information of IASI channels

Before comparing the two different estimates of MI, we begin by looking at the convergence rate of the

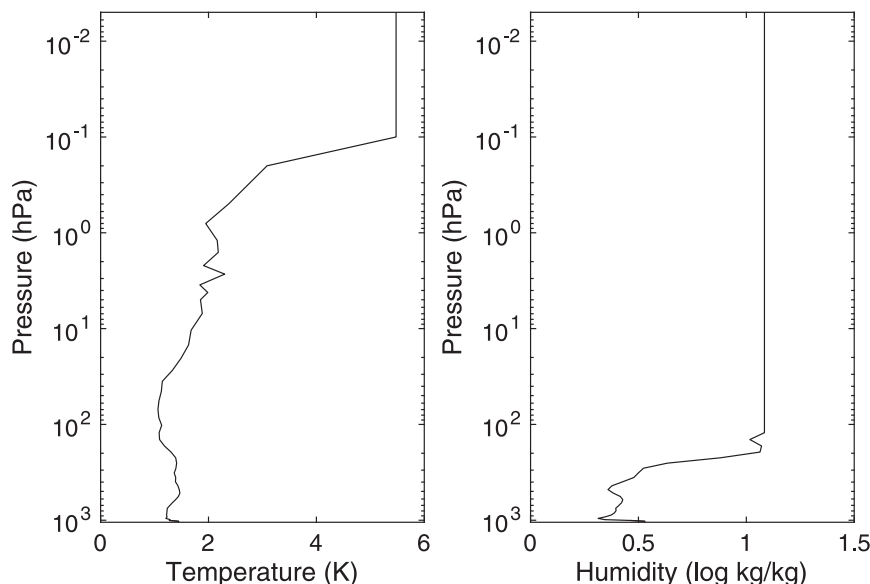


FIG. 2. Background error standard deviations.

sampling estimate of MI described in [section 2b](#). From experiments (not shown) it is known that the sample estimate of MI 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 **B** and **R** error covariance matrices, necessary for generating the initial samples, have been provided by the NWP SAF 1D-Var package. The background error standard deviations for temperature and humidity are given in [Fig. 2](#) and the correlation structure is given in [Fig. 3](#). To represent **R**, a diagonal matrix is used. The square root of the diagonal elements of **R** is given as a function of wavenumber in [Fig. 4](#). Note that although the error in the measured radiance value is assumed

invariant under scene temperature (average brightness temperature), the corresponding error in the brightness temperature is not. The “true” atmospheric profile, from which the samples are generated, represents mid-latitude cloud-free conditions.

[Figure 5](#) shows the convergence of MI with increasing sample size N for 10 different channels of IASI (stars). For each choice of N , MI has been estimated 10 times with different realizations of the random error in the observations and the prior estimate. Most channels seem to have begun to converge by $N = 2000$. It has therefore been decided from these experiments to initially use $N = 2000$ to compute MI for all channels to compare to the linear estimate, and then in [section 3](#) to use $N = 3200$ when performing the channel selection, for which

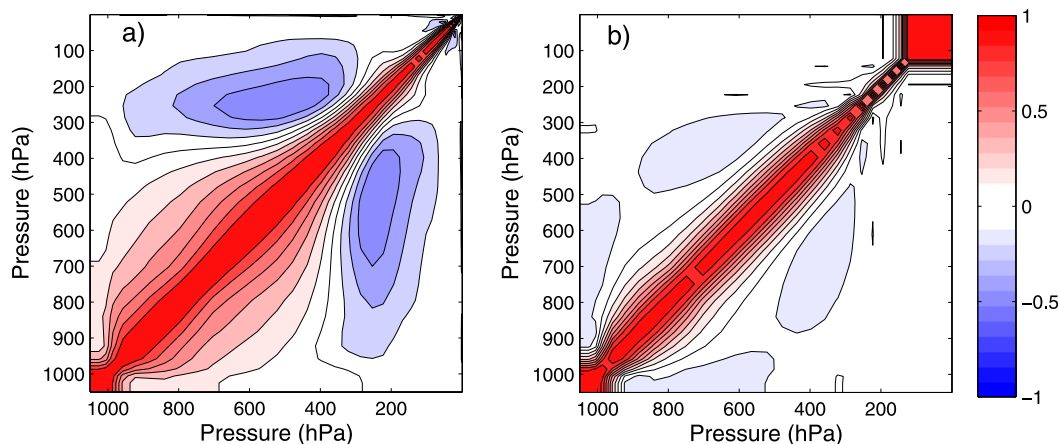


FIG. 3. Background error correlations for (a) temperature and (b) humidity. There are no multivariate correlations.

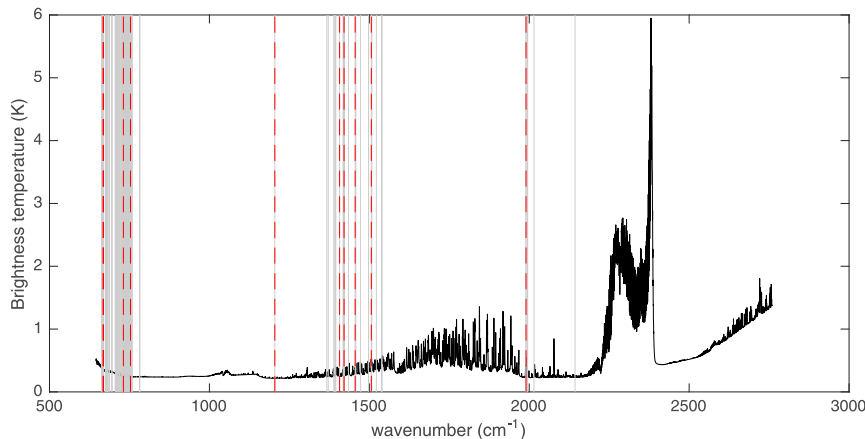


FIG. 4. The observation error standard deviations as a function of wavenumber for all 8461 IASI channels. Red dashed lines mark the 10 channels used to illustrate the proposed method in section 3. The gray lines mark the 100 channels used in section 5.

sampling error may have a greater impact as we are interested not only in the absolute value of MI but the value relative to the other channels.

Details of the 10 IASI channels given in Fig. 5 can be found in Table 1. The first column in Table 1 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 fourth columns, details of the wavelength and wavenumber 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” CO₂ channels. A total of 65 channels were selected by Collard (2007) in this initial selection. “Main” refers to the channel selection when water vapor channels were reintroduced. These channels will be used in section 3 to develop the method for optimal channel selection based on the nonlinear estimate of MI.

For these 10 channels we now compare the sample estimate of MI to the linear estimates, from section 2a. In Fig. 5 the linear estimates to MI calculated as $0.5 \ln |\mathbf{I}_n + \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}|$ [see (7)] are given by the lines. The three lines represent estimates when the observation operator has been linearized about (i) $\mathbf{x}_{\text{truth}}$ (solid line), (ii) $\mathbf{x}_{\text{truth}} - \sigma_b$ (dashed line), and (iii) $\mathbf{x}_{\text{truth}} + \sigma_b$ (dotted 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, which is assumed to be much closer to the truth than $\mathbf{x}_{\text{truth}} \pm \sigma_b$.

The accuracy of the linear estimate of MI (as compared to the sample estimate) and its sensitivity to the

linearization 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, while 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 linearization error of the observation operator for each of the channels. The linearization error can be quantified as

$$\varepsilon_{\text{lin}} = H(\mathbf{x} + \delta\mathbf{x}) - H(\mathbf{x}) - \mathbf{H}\delta\mathbf{x}. \quad (15)$$

This can be deemed adequately small if ε_{lin} is much smaller than the observation error. In incremental variational data assimilation, the perturbation $\delta\mathbf{x}$ can be expected to decrease with iteration as the nonlinear cost function is minimized (Courtier et al. 1994).

In Fig. 6, the linearization error normalized by the standard deviation of the observation error is plotted as a function of perturbation size $\delta\mathbf{x}$. In the experiment shown, $\delta\mathbf{x}$ has been chosen to be a fraction of the standard deviation of the background error. This is an arbitrary choice to illustrate that a large error in the linear estimate to mutual information (see Fig. 5) corresponds to a large linearization error. In reality this does not give profiles consistent with the assumed background errors as it does not take into account the vertical correlation seen in Fig. 3 and as such may overestimate the error due to the linearization that will be seen in practice.

In Fig. 7 MI has been computed for *all* of the IASI channels using the two approximations. Channels with

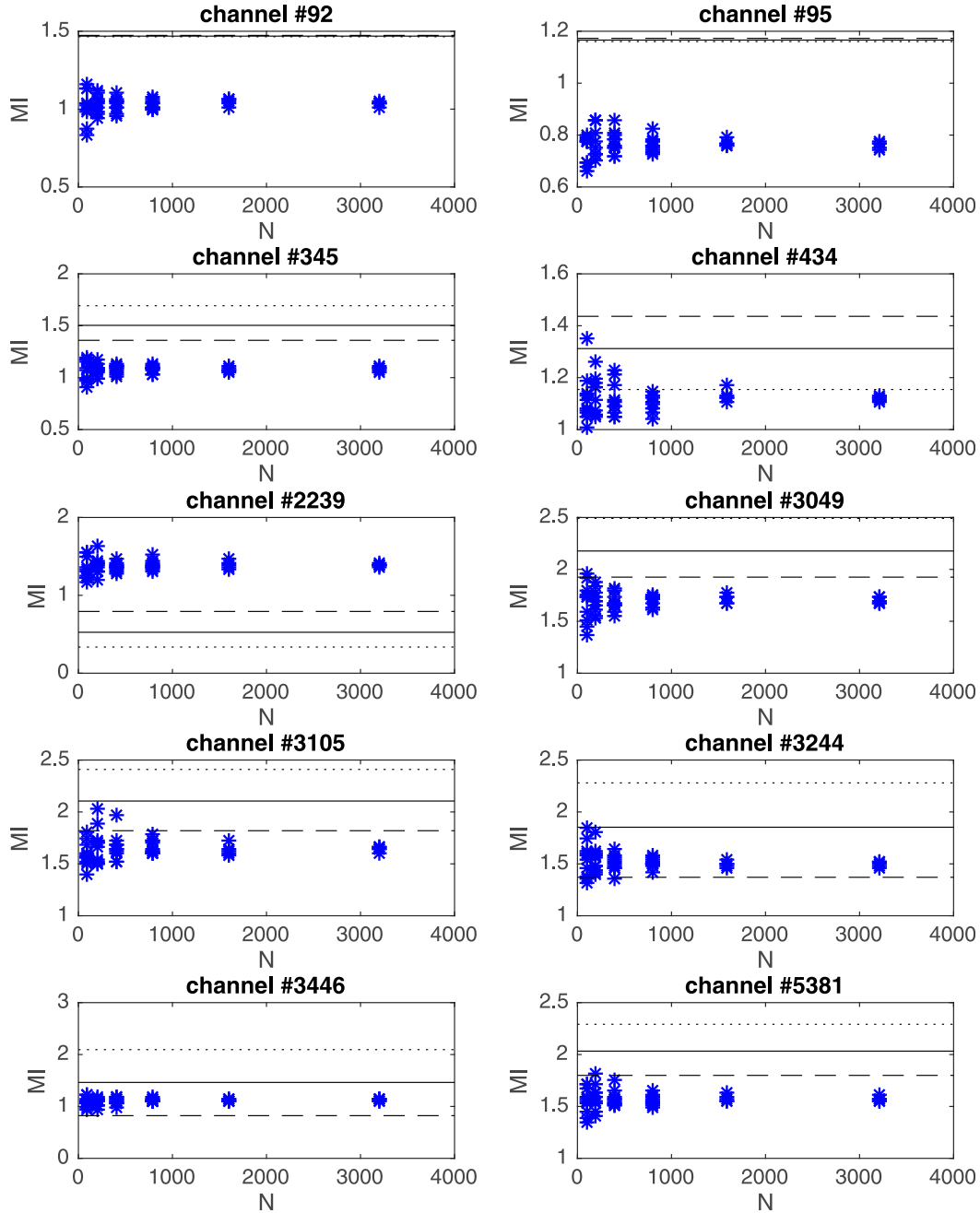


FIG. 5. MI is approximated using the sampling method (blue stars, discussed in [section 2b](#)) for different random realizations of the prior and likelihood when $M = 100$ and N is allowed to vary. The lines show the linear estimate of MI [see (7)] when the observation operator has been linearized about the truth: $\mathbf{x}_{\text{truth}}$ (solid line), $\mathbf{x}_{\text{truth}} - \sigma_b$ (dashed line), and $\mathbf{x}_{\text{truth}} + \sigma_b$ (dotted line).

large sensitivities to the surface, water vapor, and ozone are marked at the top of [Fig. 7](#). Some of these channels are removed in the preselection runs detailed by [Collard \(2007\)](#). Also indicated at the bottom of the figure are the blacklist channels due to large sensitivities to trace gases (CH_4 , CO , and N_2O), solar irradiance, and channels in

the CO_2 band that are affected by non-local thermodynamic equilibrium (LTE) effects. The difference between the two estimates indicates regions where nonlinearity for this problem is larger (e.g., water-sensitive channels, which could be expected as the dependence of the temperature Jacobian on humidity is

TABLE 1. Channels used within the selection in section 3. The rightmost column refers to the order in which the channels were selected by Collard (2007).

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

not accounted for in the linear approximation). When the linear estimate of MI is larger than the sample estimate of MI, this could be indicative of a situation similar to that illustrated in Fig. 1. In this case $p(\mathbf{x} | \mathbf{y})$ has a much larger variance than the linear estimate, so we can expect that, despite the non-Gaussian structure, the information in the observations is overestimated by the linearization of the observation operator. However, when the linear estimate of MI is smaller than the sample estimate of MI, this could be indicative of a situation when the linear estimate of the variance is similar to the true variance but lacks the structure of the true posterior density function. Note that only temperature and humidity are part of the state vector.

3. Channel selection for IASI instrument

In the last section it was shown that there are indeed instances when the linear and nonlinear 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 and how this affects the amount of independent information measured in the different channels.

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

- 1) Initially channels that are known to be poorly modeled by RTTOV are removed from the channels available for selection (e.g., those dominated by trace species); see Fig. 7.
- 2) Then 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 process are repeated until the required number of channels has been selected.

This is a time-consuming procedure that is performed offline. To deal with the nonlinearity, Collard (2007) and Rabier et al. (2002), while using a linear estimate of MI, repeated this channel selection procedure for a number of different atmospheric states and averaged the results. In the comparison that follows, between our sampling method and the linear method, we only use one true state to highlight the effects of the nonlinearity.

It is also important to note that we are unable to take into account interchannel error correlations in either the sampling or the linear method because of the sequential nature of the two methods. The effect of systematic errors in the radiative transfer model on the linear channel selection method was studied by Ventress and Dudhia (2014) but how to include these in the sampling method is left for future work; one possibility is to transform the observations using $\mathbf{R}^{-1/2}$ so that in the transformed space the observations remain uncorrelated.

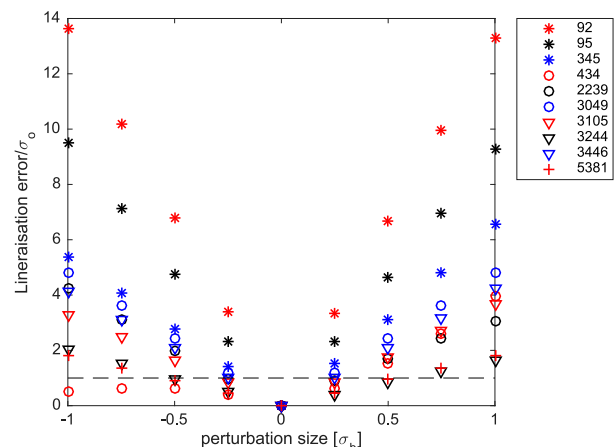


FIG. 6. Linearization error ϵ_{lin} normalized by the observation error standard deviation σ_o as a function of perturbation size (a fraction of the background error standard deviation σ_b). The dashed line shows $\epsilon_{lin}/\sigma_o = 1$.

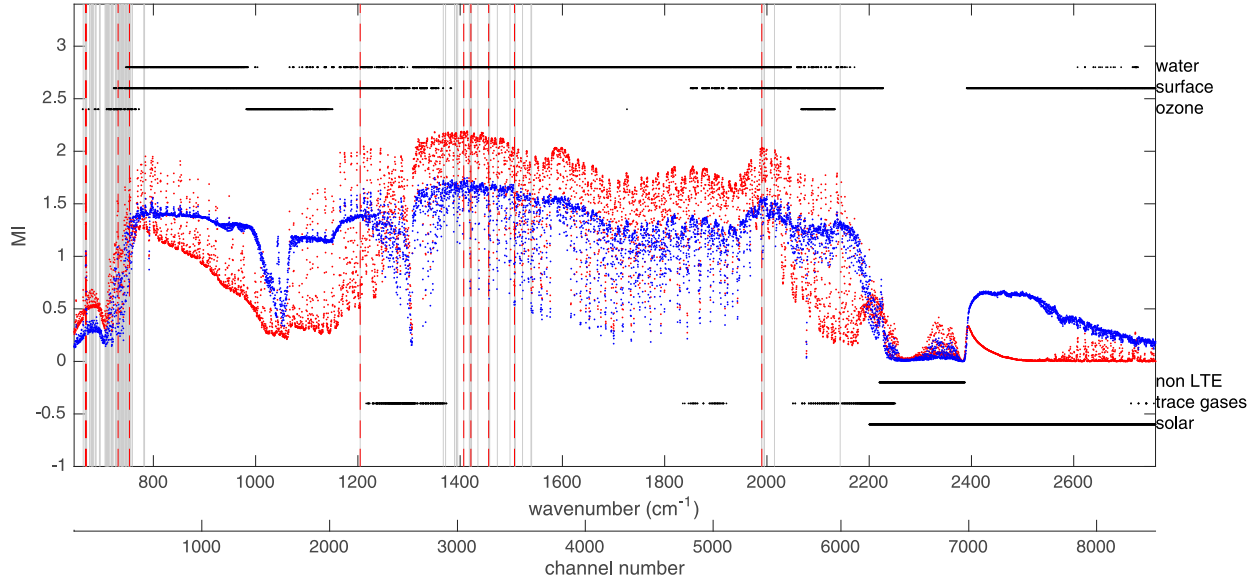


FIG. 7. MI approximated using the sampling method (blue dots; discussed in section 2b) when $M = 100$ and $N = 2000$ and using the linear method [red dots; see (7)] when the observation operator has been linearized about the truth. Red dashed lines mark the 10 channels used to illustrate the proposed method in section 3. The gray lines indicate the 100 channels to be used for selection in section 5.

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 they were considered to have a large information content by Collard (2007). The weighting functions of the 10 channels used are given in Fig. 8. It can be seen that these channels are sensitive to temperature and humidity throughout the troposphere as well as providing information about temperature for the stratosphere.

In Fig. 9, MI and the subsequent channel selection are shown for when (i) the sample estimate of MI is used and (ii) when the linear estimate of MI is used with $\mathbf{H}$ linearized about the truth. As suggested by Fig. 5, the values of MI for the initial selection (first column) can differ significantly between the two different estimates. In addition to this the amount of independent information in the channels differs between the two estimates, so that although the first channel to be selected is the same in each case the remaining channels are selected in a different order. For example channel number 10 is deemed to be the third most important channel using the linear method but only the fifth most important using the sampling method.

In Fig. 10, the effective sample size ess of the sample estimate is shown for the first realization from the likelihood. This is defined as

$$ess_1 = \frac{1}{\sum_{i=1}^N (w_{i,1})^2} \quad (16)$$

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., $w_{i,j} = 1/N \forall i, j$), then the effective sample size is N . As the variance of the weights used to describe the posterior distribution in (11) increases, ess decreases.

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 center of the distribution, where the probability is high, are given a large weight while 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 indicates that it is no longer useful for subsequent channel selection, as shown in Fig. 5. This problem will increase as the number of channels to be selected increases and the amount of

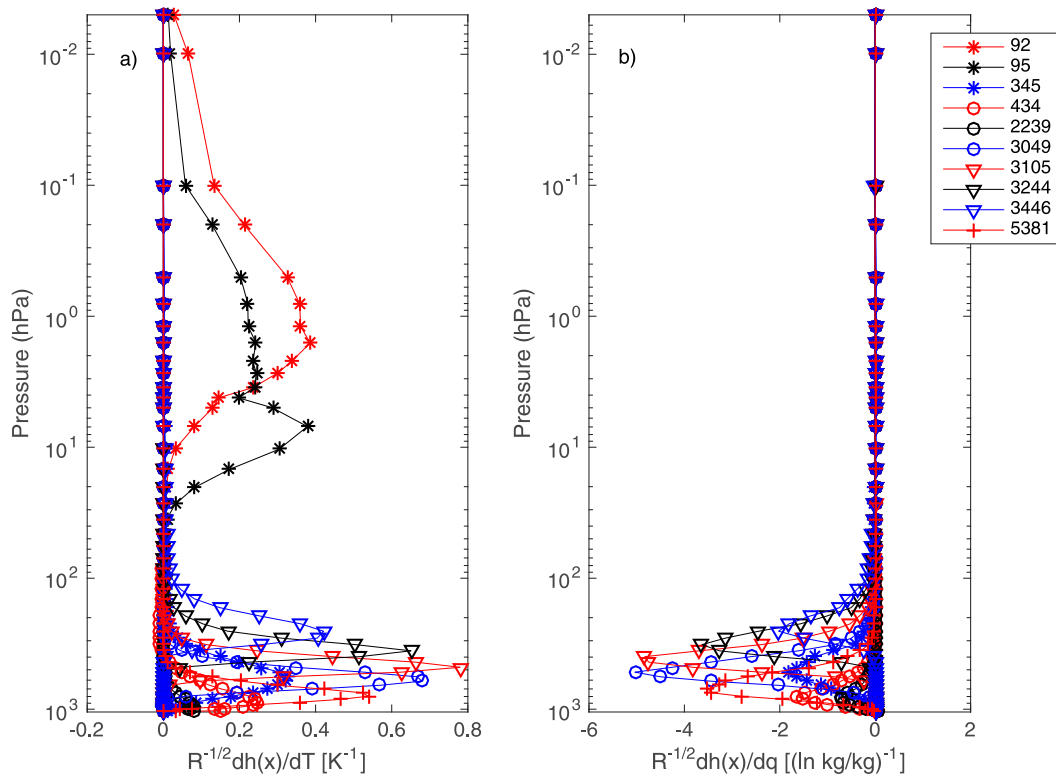


FIG. 8. Weighting functions normalized by the observation error standard deviation, for the 10 channels used in the channel selection. Shown are sensitivities to changes in (a) temperature and (b) humidity.

information available in consecutive observations is increased.

4. Improving the effective sample size

As seen in Fig. 5, the nonlinear 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 close to constant throughout the channel selection procedure. For this reason increasing the sample size is not the solution; first an unnecessarily (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 second the accuracy of the MI estimate would change throughout the channel selection process as the effective sample size decreases.

An alternative to assimilating one channel at a time and resampling from the posterior after each channel is selected is to assimilate an increasingly large number of channels. In practice this means that on each round an extra channel is assimilated in addition to those already selected, but the number to consider is still reduced by one. This would have several advantages, namely there would be no need to resample from the posterior

distribution as the prior would be unchanged for each channel selection, and it would also allow for the interchannel error correlations to be accounted for. However, the issues with diminishing sample size are still prevalent, as illustrated in Figs. 11–13. It is clear that as the number of channels assimilated in one go increases, the number of samples needed to estimate MI also increases, as indicated by the effective sample size. For example when a sample size of 8000 is used, the effective sample size after assimilation ranges from about 2000 to 2500 when 1 channel is assimilated (Fig. 11), 10 to 130 when 5 channels are assimilated (Fig. 12), and 1 to 9 when 10 channels are assimilated (Fig. 13). It is therefore unfeasible to estimate MI for all channels at one time.

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 Gordon et al. 1993). This idea has been used extensively in the particle filter (e.g., Kim et al. 2003; Lui

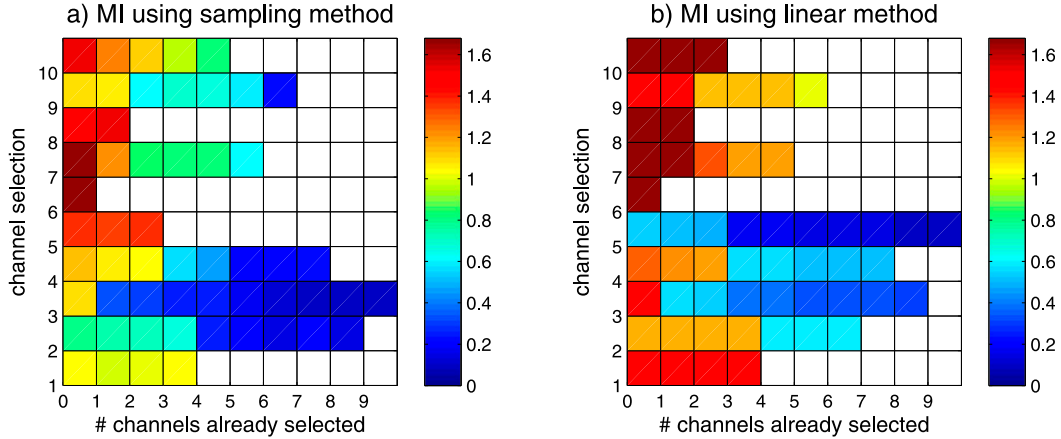


FIG. 9. Channel selection for a subset of 10 channels given in Table 1. Channel selections using (a) a sample estimate of MI and (b) a linear estimate of MI. The colors represent the value of MI estimated in each case.

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 function (e.g., Van Leeuwen 2010, 2011); the idea being that we sample from a proposal density function that is similar to the posterior distribution that we wish to represent. This generally makes use of the observations to “draw” the sample toward the region of high likelihood. This is complicated in the case of channel selection because (i) we do not know a priori which channel will be selected and (ii) we need to average over the observation space. Therefore, this technique would involve a prohibitively 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 is made. Because of the nonlinear observation operator, we expect the posterior to be non-Gaussian, and we 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 the Akaike information criterion (AIC) is at a minimum (Burnam and Anderson 2002, 60–64) but each component is represented by a large enough sample to ensure a good estimate of the covariance matrix for each of the Gaussian components.

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

$$p(\mathbf{x}) = \sum_{k=1}^G \alpha_k N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (17)$$

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), and $\boldsymbol{\Sigma}_k$ (the covariances of the Gaussian components). These parameters may be found using the expectation–maximization (EM) method [see Bishop (2006, 424–435) for an introduction].

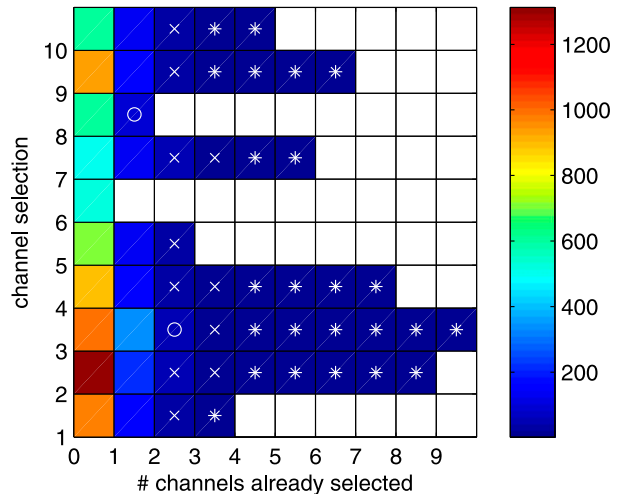


FIG. 10. Effective sample size of the posterior distribution: circles, $\text{ess} < 100$; crosses, $\text{ess} < 50$; and stars, $\text{ess} < 10$.

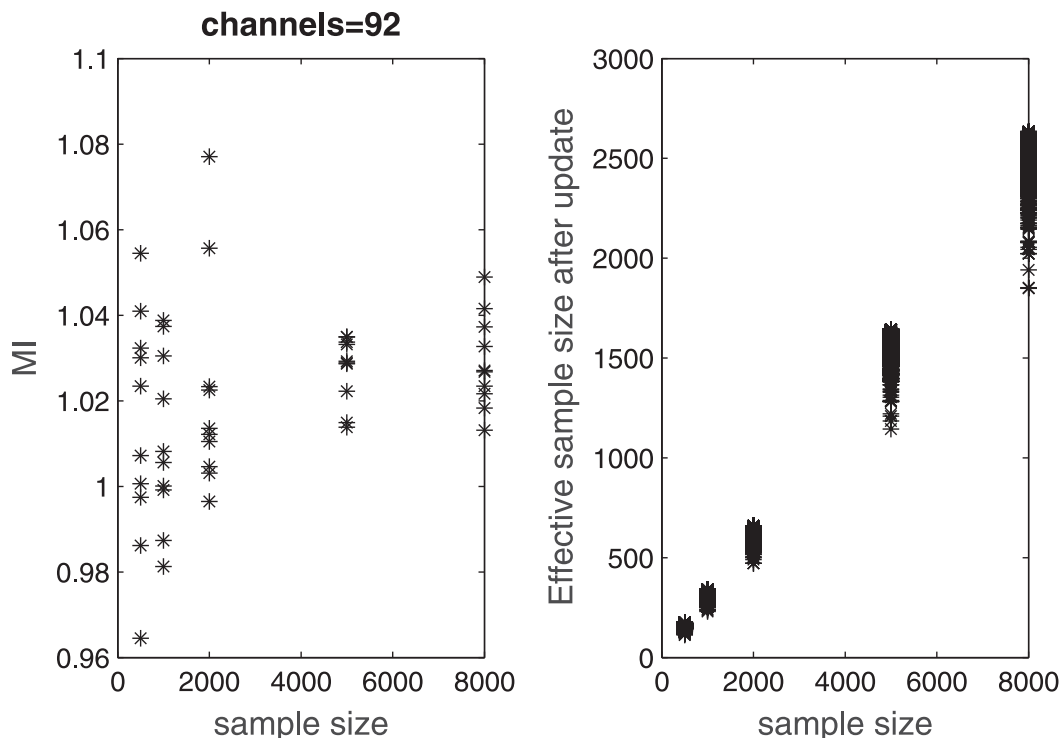


FIG. 11. (left) The convergence of MI of one channel as sample size increases. (right) The effective sample size of the updated sample.

Once the Gaussian mixture has been fitted to the sample, it is straightforward 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 .

a. Fitting a Gaussian mixture in practice

In practice the fitting of a Gaussian mixture is not straightforward and may be performed in many different ways. Here, the method used to provide the following results, which makes use of the MATLAB statistics toolbox function “fitgmdist,” is outlined. In many cases the methods are chosen for their pragmatism and the author accepts that different approaches could be equally valid and perhaps better.

To begin the process of fitting a distribution to the weighted sample, a new sample is generated that is equally weighted. This can be done in many different ways but here probabilistic resampling is used (Gorden et al. 1993). As discussed above, this is simply a method for replicating samples with a high weight and deleting those with a small weight, which only leads to an artificial increase in the effective sample size.

Once the probabilistic resampling is complete, an initial estimate of the parameters is needed as a first guess for the iterative EM method. Here, we use a

randomly selected sample point to represent the means, the weights are uniform (each equal to $1/G$), and covariance matrices are initially diagonal with variances equal to the variance of the sample. Alternatively, a k -clustering algorithm can be used, which assigns each of the samples to different groups [again see Bishop (2006, 424–435), for an introduction]. The parameter estimates are then refined by iterating the EM method 100 times.

To decide how many Gaussian components are necessary to describe the sample, the Gaussian mixture model is estimated for an increasing number of components until one group has too few members (i.e., $N_k < N_{\min}$). The model with the smallest AIC is then selected. Within this work each component is represented by at least 204 samples ($N_{\min}$). 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), while still being small enough to allow for a good deal of structure in the fitted distribution. In addition to this we also add a small regularization term of 10^{-5} to the diagonal of the covariance matrices to ensure that the estimated covariances are always positive definite.

An illustration of the process on an artificially generated sample is shown in Figs. 14 and 15. This

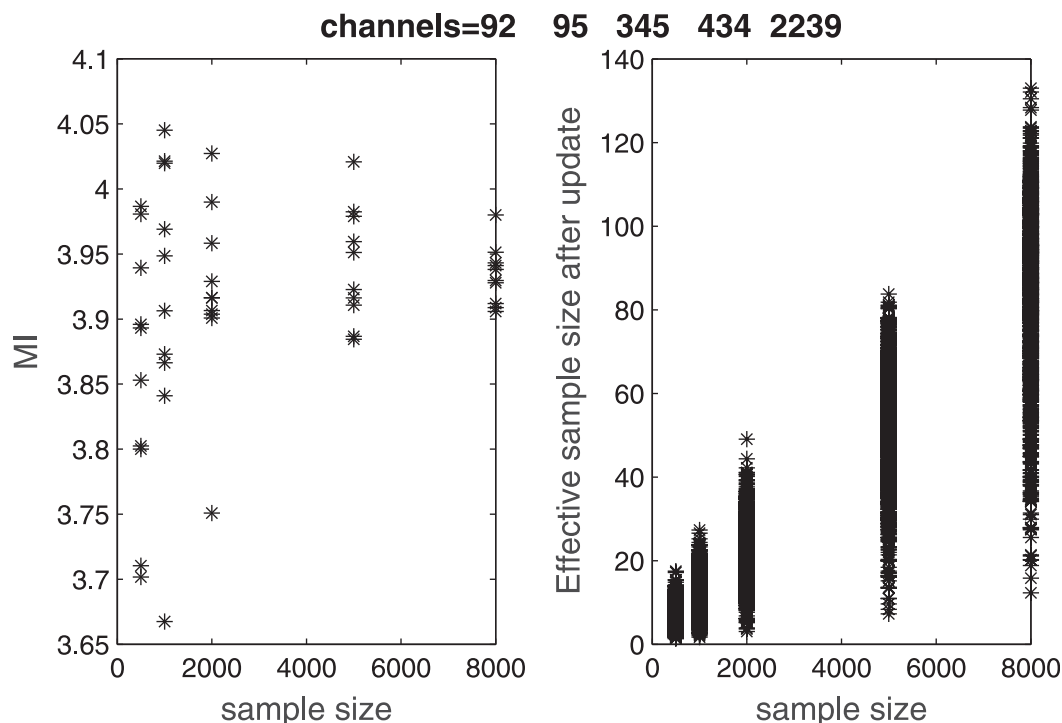


FIG. 12. (left) The convergence of MI of five channels as the sample size increases. (right) The effective sample size of the updated sample.

two-dimensional sample has a size of 400 but has been weighted according to its distance from the point (0, 1) marked by the red star. This results in an effective sample size of 197. In the right-hand panel of Fig. 14 a new sample is generated using probabilistic resampling. The density of the sample is difficult to see from the scatter-plot, as many of the sample points are duplicates and therefore cannot be distinguished from each other. In Fig. 15 we see that the GM resampling step has decided that the distribution can be accurately described by three components, illustrated by the different colors. This new sample is seen to have much better coverage of the high-probability regions and the marginal frequency histograms for the two variables match closely to the original sample. In Table 2 the first four moments of the original sample and the sample from the GM fit are given.

b. Applying the sampling method with Gaussian mixture resampling to the channel selection problem

In Fig. 16 the channel selection is repeated for the 10 channels; this time resampling from a Gaussian mixture distribution after each channel is selected. After the first channel is selected there are some differences in the value of the sample estimate to MI (cf. to Fig. 9); in particular, we see that although the first six and last four

channels to be selected are the same in each case, the order that they have been selected in has changed. The substantial increase in the effective sample size after each channel selection is made allows for 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. Application to a larger-sized problem

In the previous section we developed a sampling method to allow for the objective selection of channels based on their mutual information. Within this section the adapted algorithm will be applied to a larger-sized problem. Applying the method to the full IASI channel set is beyond the scope of this paper, in which the aim is the demonstration of a new method. However, it is important to note that the method can be run on parallel processors, so applying it to the full-sized problem is feasible and will be addressed in future work. In this section the channel selection algorithm will be applied to a channel set of size 100, given by the first 100 channels selected by Collard (2007) (in both the preselection and the main run).

The 100 channels used within this experiment are shown in Fig. 7 by the gray vertical lines. We see that the majority of the channels are in the $670\text{--}710\text{ cm}^{-1}$ region

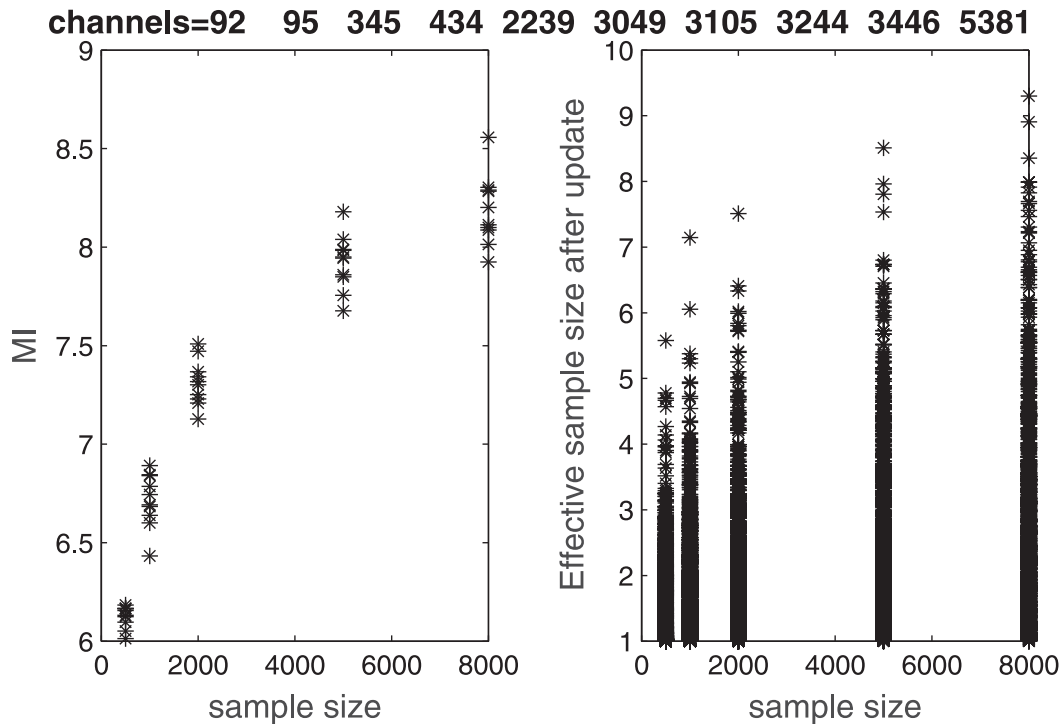


FIG. 13. (left) The convergence of MI of 10 channels as sample size increases. (right) The effective sample size of the updated sample.

because of their sensitivity to the upper troposphere and lower stratosphere. This is explained in Collard (2007) to be related to the relatively high a priori temperature errors in this region compared to the troposphere.

In Fig. 17, 50 of the 100 channels available are selected using both the sampling and linear approximations. It can be seen that the sampling method appears to have chosen a larger spread of the wavenumbers available

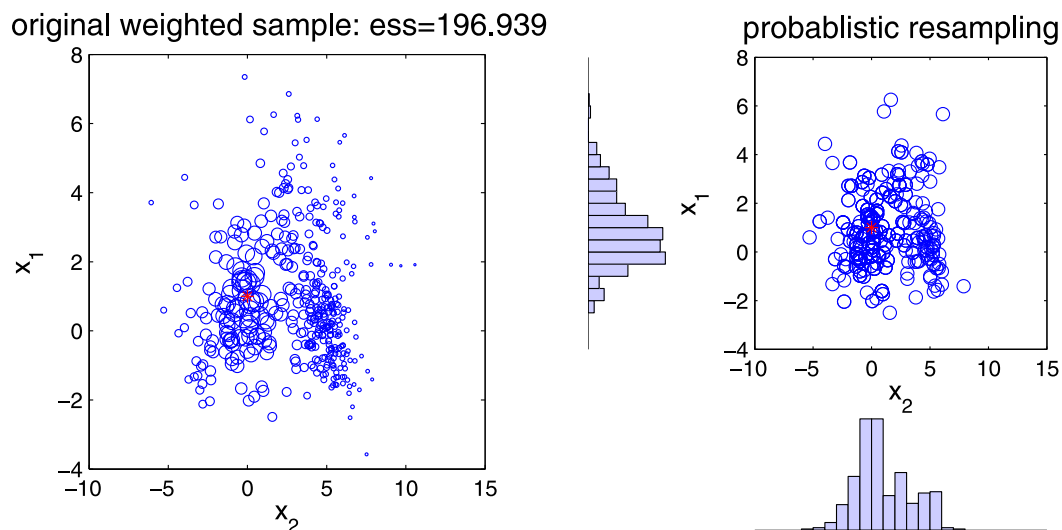


FIG. 14. Illustration of the probabilistic resampling step. (left) The original weighted sample is shown (the size of the marker is proportional to its weight). (right) A new sample has been generated using probabilistic resampling so that the sample is equally weighted. As this involves deleting samples with a small weight and duplicating sample members with a high weight, the increase in effective sample size is clearly artificial; the histograms show the frequency of each of the variables.

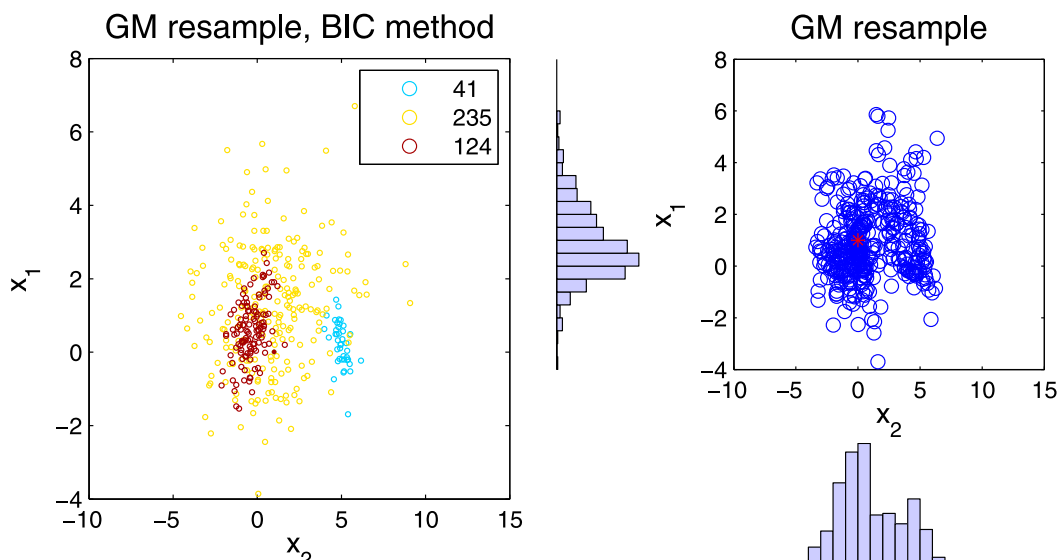


FIG. 15. Illustration of how a GM distribution is fitted to the equally weighted sample generated in Fig. 14 and a new sample is generated. (left) Samples from three different Gaussian distributions with different proportions are shown. (right) The same sample, but without distinction between the different groups, is given; the histograms show this sample to be a good match to the original sample but the effective sample size has increased.

than the linear method. For example, 26 of the channels selected by the linear method are in the 600–800 cm^{-1} band whereas only 20 are selected in this band by the sampling method.

The mutual information of the selected channels is given in Fig. 18 for each of the two selection methods. It is clear that the sampling method quantifies there to be less information in the channels than the linear methods for the first few channels selected, but as the channel selection progresses, both estimates diagnose a similar amount of additional information in the remainder of the channels selected.

6. Discussion

Satellite observations are a nonlinear 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 not just the first two moments. To estimate mutual information, a sampling technique that 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.

To obtain a good estimate of the 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

TABLE 2. Comparison between the first four moments calculated from the original sample and the GM resample for the illustrations in Figs. 14 and 15.

Variable	Original sample				GM resample			
	μ	σ	Skewness	Kurtosis	μ	σ	Skewness	Kurtosis
x_1	0.971	2.36	0.447	2.55	0.982	2.40	0.411	2.15
x_2	0.864	1.44	0.560	3.39	0.881	1.42	0.600	3.88

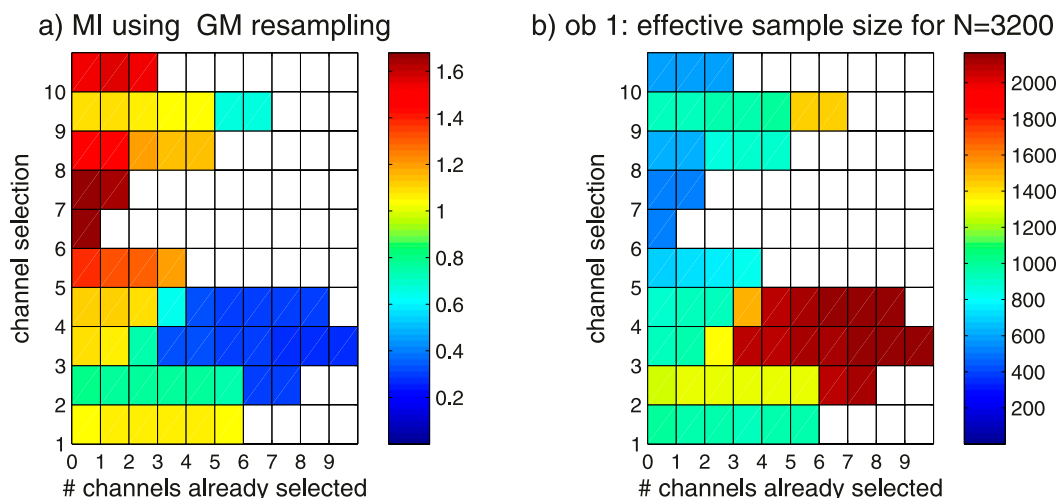


FIG. 16. (a) Channel selection using sample estimate of MI with GM resampling (the colors represent the value of MI estimated). (b) Effective sample size of the posterior distribution before GM resampling is applied.

each channel has been selected. Resampling 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 “offline,” 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 that could be applied to all atmospheric conditions. This helps to reduce some of the effect of the nonlinearity. An advantage of the proposed sampling method is that, by accounting explicitly for the nonlinearity, it is possible to give an optimal channel list for a specific prior distribution. However, it may be possible to identify different classes of prior distribution that lead to a similar channel selection allowing for the computations to be performed offline.

It is important to note that taking into account the nonlinearity of the observation operator in the channel selection is only beneficial if this is consistent with the way the observations are to be assimilated, that is, 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 is relaxed. The author therefore anticipates the need to reassess 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. I would also like to thank Christina Prates for her help with the setup of RTTOV and Andrew Collard for providing me with the

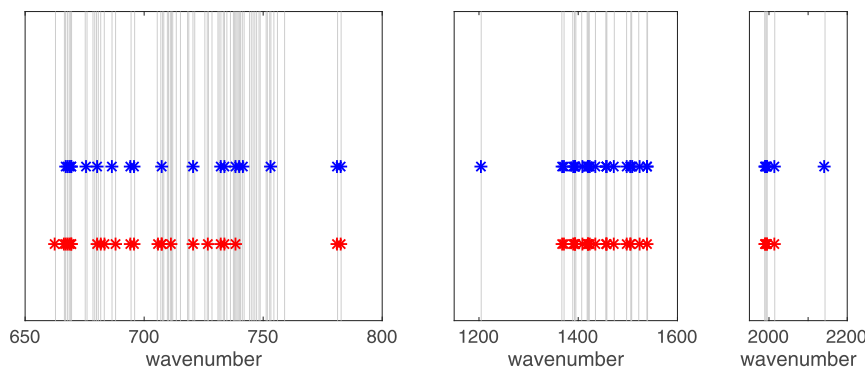


FIG. 17. Channels selected by the two different methods: sampling (blue) and linear approximation (red). The available channels for selection are represented by the gray lines.

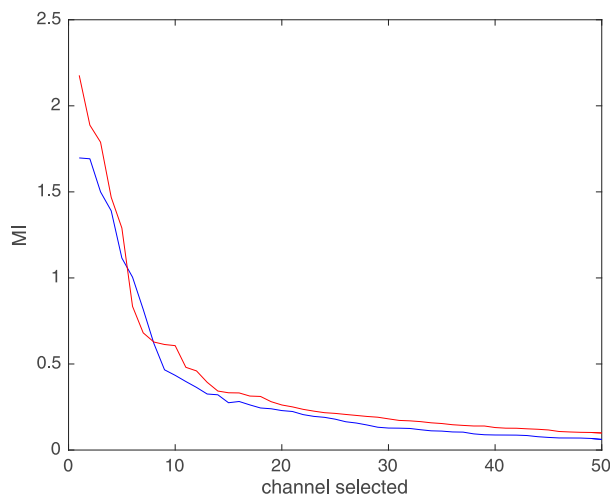


FIG. 18. Change in MI as the channel selection progresses: sampling method (blue) and linear approximation (red).

channel list used in his original paper. Thanks also go to the anonymous reviewers of a previous version of this manuscript. This work has been funded under the “ESA Advanced Data Assimilation Methods” project, contract ESRIN 4000105001/11/I-LG, and as part of NERC’s support for the National Centre of Earth Observation (NCEO).

REFERENCES

- Bishop, C., 2006: *Pattern Recognition and Machine Learning*. Springer, 738 pp.
- Burnam, K. P., and D. R. Anderson, 2002: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. Springer, 488 pp.
- Chevallier, F., P. Lopez, A. M. Tompkins, M. Janisková, and E. Moreau, 2004: The capability of 4D-Var systems to assimilate cloud-affected satellite infrared radiances. *Quart. J. Roy. Meteor. Soc.*, **130**, 917–932, doi:10.1256/qj.03.113.
- Collard, A. D., 2007: Selection of IASI channels for use in numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **133**, 1977–1991, doi:10.1002/qj.178.
- Courtier, P., J.-N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J. Roy. Meteor. Soc.*, **120**, 1367–1387, doi:10.1002/qj.49712051912.
- Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory*. Wiley Series in Telecommunications, John Wiley and Sons, 542 pp.
- Eyre, J. R., 1989: Inversion of cloudy satellite sounding radiances by nonlinear optimal estimation. I: Theory and simulation for TOVS. *Quart. J. Roy. Meteor. Soc.*, **115**, 1001–1026, doi:10.1002/qj.49711548902.
- Fowler, A. M., and P. J. Van Leeuwen, 2013: Measures of observation impact in data assimilation: The effect of a non-Gaussian measurement error. *Tellus*, **65**, 20035, doi:10.3402/tellusa.v65i0.20035.
- Gorden, N. J., D. J. Salmond, and A. F. M. Smith, 1993: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proc.*, **144**, 107–113, doi:10.1049/ip-f-2.1993.0015.
- Hocking, J., P. Rayer, R. Saunders, M. Marticardi, A. Geer, and P. Brunel, 2011: RTTOV v10 users guide. NWP SAF Tech. Rep. NWPSAF-MO-UD-023, 92 pp. [Available online at http://nwpsaf.eu/oldsite/deliverables/rtm/docs_rtto10/users_guide_10_v1.5.pdf.]
- Hoteit, I., X. Luo, and D.-T. Pham, 2012: Particle Kalman filtering: A nonlinear Bayesian framework for ensemble Kalman filters. *Mon. Wea. Rev.*, **140**, 528–542, doi:10.1175/2011MWR3640.1.
- Kim, S., G. L. Eyink, J. M. Restrepo, F. J. Alexander, and G. Johnson, 2003: Ensemble filtering for nonlinear dynamics. *Mon. Wea. Rev.*, **131**, 2586–2594, doi:10.1175/1520-0493(2003)131<2586:EFND>2.0.CO;2.
- Lui, J. S., and R. Chen, 1998: Sequential Monte Carlo methods for dynamical systems. *J. Amer. Stat. Assoc.*, **90**, 567–576, doi:10.2307/2669847.
- Musso, C., N. Oudjane, and F. Le Gland, 2001: Improving regularized particle filters. *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. de Freitas, and N. Gordon, Eds., Springer, 247–271.
- Pavelin, E. G., S. J. English, and J. R. Eyre, 2008: The assimilation of cloud-affected infrared satellite radiances for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **134**, 737–749, doi:10.1002/qj.243.
- Rabier, F., N. Fourrié, D. Chafäi, and P. Prunet, 2002: Channel selection methods for Infrared Atmospheric Sounding Interferometer radiances. *Quart. J. Roy. Meteor. Soc.*, **128**, 1011–1027, doi:10.1256/0035900021643638.
- Rodgers, C. D., 1996: Information content and optimisation of high spectral resolution measurements. *Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research II*, P. B. Hays and J. Wang, Eds., International Society for Optical Engineering (SPIE Proceedings, Vol. 2830), 136–147, doi:10.1117/12.256110.
- , 2000: *Inverse Methods for Atmospheric Sounding*. Series on Atmospheric, Oceanic and Planetary Physics, Vol. 2, World Scientific Publishing, 256 pp.
- Salby, M. L., 1996: *Fundamentals of Atmospheric Physics*. Academic Press, 627 pp.
- Smith, K. W., 2007: Cluster ensemble Kalman filter. *Tellus*, **59A**, 749–757, doi:10.1111/j.1600-0870.2007.00246.x.
- Tamminen, J., and E. Kyrölä, 2001: Bayesian solution for nonlinear and non-Gaussian inverse problems by Markov chain Monte Carlo method. *J. Geophys. Res.*, **106**, 14 377–14 390, doi:10.1029/2001JD900007.
- Van Leeuwen, P. J., 2003: A variance-minimizing filter for large-scale applications. *Mon. Wea. Rev.*, **131**, 2071–2084, doi:10.1175/1520-0493(2003)131<2071:AVFFLA>2.0.CO;2.
- , 2009: Particle filtering in geophysical systems. *Mon. Wea. Rev.*, **137**, 4089–4114, doi:10.1175/2009MWR2835.1.
- , 2010: Nonlinear data assimilation in geosciences: An extremely efficient particle filter. *Quart. J. Roy. Meteor. Soc.*, **136**, 1991–1996, doi:10.1002/qj.699.
- , 2011: Efficient non-linear data assimilation in geophysical fluid dynamics. *Comput. Fluids*, **46**, 52–58, doi:10.1016/j.compfluid.2010.11.011.
- Ventress, L., and A. Dudhia, 2014: Improving the selection of IASI channels for use in numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **140**, 2111–2118, doi:10.1002/qj.2280.