An Infrared Atmospheric Sounding Interferometer – New Generation (IASI-NG) channel selection for numerical weather prediction

F. Vittorioso, V. Guidard, Nadia Fourrié

Received: 23 April 2020 | Revised: 9 July 2021 | Accepted: 9 July 2021 | Published on: 12 August 2021

DOI: 10.1002/qj.4129

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

An Infrared Atmospheric Sounding Interferometer – New Generation (IASI-NG) channel selection for numerical weather prediction

F. Vittorioso | **V. Guidard** | **N. Fourrié**

CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Correspondence
F. Vittorioso, Centre National de Recherches Météorologiques, GMAP, 42 av. G. Coriolis, 31057 Toulouse Cedex 1, France.
Email: francesca.vittorioso@meteo.fr

Funding information
Centre National d'Études Spatiales (CNES) within the framework of the IASI-NG project

Abstract

In the framework of the EUMETSAT Polar System–Second Generation (EPS-SG) preparation, a new generation of the Infrared Atmospheric Sounding Interferometer (IASI) instrument has been designed. The IASI-New Generation (IASI-NG) will measure radiances at a doubled spectral resolution compared to its predecessor and with a signal-to-noise ratio improved by a factor of 2. The large amount of data arising from IASI-NG will present many challenges for data transmission, storage and assimilation. Moreover, the full set of measured radiances will not be exploitable in an operational numerical weather prediction (NWP) context. For these reasons, an appropriate IASI-NG channel selection in needed, aiming to select the most informative channels for NWP models. Therefore, the standard iterative channel selection methodology, based on the optimal linear estimation theory and assuming spectrally correlated errors, has been applied to a set of simulated data of the IASI-NG spectrum. The entire simulated IASI-NG spectrum has been first investigated, while finally focusing the channel selection procedure on the most interesting wavelength ranges for the assimilation. Through this process, a total of 500 channels have been chosen to serve as a basis for future channel selections to be provided to NWP centres – 277 temperature, 23 surface-sensitive and 200 water vapour channels. One-dimensional variational (1D-Var) assimilation experiments show that using this selected set of channels leads to a reduction of the standard deviation of the error in temperature (up to 30%) and water vapour (up to 50%) profiles with respect to the *a priori* information.

KEYWORDS

channel selection, DFS, hierarchical clustering, IASI-NG, numerical methods, NWP

1 | INTRODUCTION

Since 2007, the hyper-spectral Infrared Atmospheric Sounding Interferometer (IASI), a key payload element of the European Meteorological Operational Satellite (MetOp) series, has provided a huge contribution to numerical weather prediction (NWP), pollution monitoring and climate research (Hilton *et al.*, 2012). Together with the Atmospheric Infrared Sounder (AIRS; Aumann *et al.*, 2003) and Cross-track Infrared Sounder (CrIS; Glumb *et al.*, 2003), it paved the way to the 21st century class of advanced infrared (IR) sounders. These instruments have changed the way IR satellite data are assimilated, radically increasing the amount of available meteorological information compared to the previous generation.

The IASI instrument is a Fourier-Transform Spectrometer (FTS). It is able to acquire atmospheric emission spectra in the IR range within 645 and 2760 cm^{-1} (3.6–15.5 μm), with a spectral apodized resolution of 0.5 cm^{-1} and a spectral sampling of 0.25 cm^{-1}. Consequently, it collects information in a total of 8,461 wavelengths (or channels). From these spectra, it is possible to derive temperature and humidity profiles with a vertical resolution of 1 km and a 1 K and 10% accuracy respectively (Siméoni *et al.*, 1997;Chalon *et al.*, 2001; Blumstein *et al.*, 2004; Crevoisier *et al.*, 2014).

The main purpose of this atmospheric sounder was originally to profile atmospheric temperature and humidity for meteorological applications. However, IASI has represented over time a significant technological and scientific step forward for measuring atmospheric composition including chemistry and air quality (e.g., Clerbaux *et al.*, 2009). Over the years, a potentially good impact of O_3 and CO on air quality forecasts has been highlighted. Thanks to the instrument's excellent signal-to-noise ratio, reactive chemical species which were previously considered not to be measurable from space (i.e., ammonia) have been measured. Other major atmospheric events, such as volcanic eruptions, fires, dust or pollution, can also be monitored thanks to the data acquired through this IR sounder (George *et al.*, 2009; Karagulian *et al.*, 2010; Capelle *et al.*, 2014).

Given its high potential demonstrated over the years, a new generation of the IASI instrument has been designed within the framework of the preparation for the next European polar-orbiting program. The IASI-New Generation (IASI-NG), to be launched on board the MetOp-Second Generation (MetOp-SG) series in 2022, will be characterized by improved spectral and radiometric characteristics relative to IASI. It will measure at 16,921 wavelengths (or channels) with each sounding pixel benefiting from a spectral sampling of 0.125 cm^{-1}, a spectral resolution of 0.25 cm^{-1} after apodization and a signal-to-noise ratio improved by a factor of 2 compared to its predecessor (Crevoisier *et al.*, 2014; Bermudo *et al.*, 2014). Measurement precision will also be improved compared to IASI. IASI-NG characteristics will lead to further improvements in detection and retrieval of numerous chemical species and aerosols, and in temperature and water vapour retrievals.

The large amount of data resulting from IASI-NG will present many challenges for data transmission, storage and assimilation. Moreover, the total number of channels will not be exploitable in an operational NWP context and the choice of an optimal data subset is needed. On the other hand, data compression using a principal component (PC) analysis is already used as one way of disseminating IASI data. Nevertheless, no NWP centre currently assimilates PC scores despite many investigations (Matricardi and McNally 2014). This means that spectra would need to be reconstructed from PC scores before assimilation. For all these reasons, an appropriate IASI-NG channel selection must be undertaken, aiming to select the most informative channels to be used in NWP models.

Several channel selection methods have been developed in the last two decades. Some of them are based on statistical analysis and information content theory (Rabier *et al.*, 2002; Fourrié and Thépaut 2003; Collard 2007; Ventress and Dudhia 2014; Noh *et al.*, 2017). Some others rely on evaluating the radiances' response to the perturbation of specific chemical species and atmospheric variables, which allows the detection of spectral sensitivity features (Susskind *et al.*, 2003; Gambacorta and Barnet 2012).

The objective of the present study is both to present a methodology to select channels using modern tools and to propose a first IASI-NG channel selection in an NWP context. Such a selection could be used either for dissemination purposes or as a basis for further channel extraction for assimilation. On the other hand, the methodology developed here could be exploited in other future studies on IASI-NG or other instruments.

Our work has been carried out on an observation database of simulated radiances for IASI and IASI-NG (Andrey-Andrés *et al.*, 2018), using a method based on information content theory (Rodgers 1996). Additionally, one-dimensional variational (1D-Var) assimilation experiments have been carried out in order to evaluate the impact of the channel selection on temperature and humidity retrievals.

The paper is arranged as follows. Section 2 presents the future IASI-NG instrument and Section 3 summarizes the methodology. A case-study is described in Section 4. The need to take into account the error correlations among channels has required a study for the evaluation of a full covariance matrix of the observation errors; this is

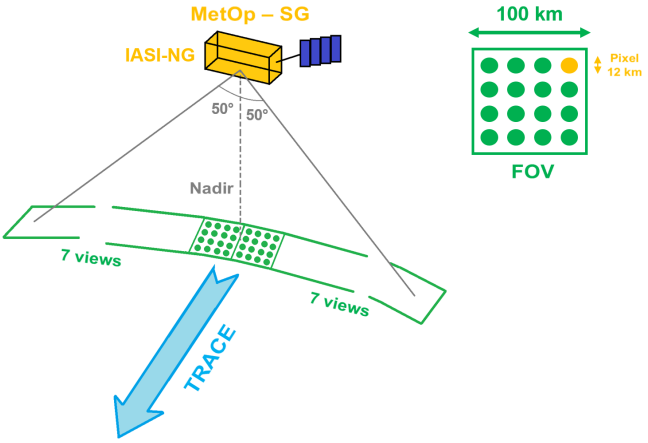

FIGURE 1 IASI-NG swath on MetOp-SG plus a zoom on a nadir Field of View (FOV) [Colour figure can be viewed at wileyonlinelibrary.com]

presented in Section 5. Finally, relying on the data thus obtained, a first selection of channels was performed by assessing the contribution of each individual channel to improving the analysis errors over the background ones. The results obtained are illustrated in Section 6. The relevance of the resulting channel selection is assessed by comparison to a random selection and the results are shown in Section 7. Finally conclusions and perspectives are discussed in Section 8.

2 | IASI-NG CONTEXT

2.1 | Instrument technical specification

IASI-NG is a space-borne FTS, based on the concept of the Mertz interferometer, which measures IR radiation emitted from the Earth.

The first satellite of the MetOp-SG series on which the instrument will be boarded is expected to be launched in 2022 on the same orbit and sensing time as MetOp, which is a late-morning orbit crossing the Equator at 09:30 UTC. The overpass frequency is twice a day at midlatitudes.

In order to achieve a global coverage, the instrument will perform 14 views per scan line across the satellite track (Figure 1). Every swath will cover approximately $2 \times 1{,}000$ km, whereas the surface acquired on each view will cover a square of side length 100 km at nadir. Each instrument single Field of View (FOV) will consist of 4×4 pixels with a 12 km size at nadir. The scanning process will acquire the 14 aforementioned views pointing to calibration targets and then return to the starting position. The total duration of the whole process will be 8 s (Bermudo *et al.*, 2014).

The instrument will cover the same spectral range as IASI. However, the spectral sampling will be improved to a resolution of 0.25 cm^{-1} after apodization (Crevoisier *et al.*, 2014). As a consequence, the spectral samples will be doubled to 16,921.

2.2 | Spectral area pre-selection

Although the spectral coverage will remain unchanged, the IASI-NG spectrum will be split into four bands, compared to three bands for IASI. The limits of the first two bands will be slightly modified, whereas the IASI band 3 will be split into two new bands for the new generation instrument. Limits and characteristics of these four different spectral areas are displayed in Table 1. For IASI-NG, in order to compute the wave number ν having the channel number n, the following formula can be used:

$$\nu = 0.125(n-1) + 645.000. \quad (1)$$

For this study, we focused on bands 1 and 2. Band 1 is mainly affected by the influence of CO_2, surface temperature and O_3. In band 2, among the other gases, water vapour absorption predominates. These features make them the most relevant bands for operational data assimilation in an NWP context.

We selected the first 2,448 contiguous channels of band 1 (from 645.000 to 950.875 cm^{-1}) and the last 3,601 contiguous channels of band 2 (from channel 6,841 up to 10,441, namely from 1,500.000 to 1,950.000 cm^{-1}). The total number of channels considered is 6,049. This choice excludes from the study the ozone-related part of band 1, since the minimisation of this quantity is not yet performed in the operational assimilation at Météo-France. However some NWP centres already have ozone in their control variables, as at ECMWF for instance. On the other hand, the second half of band 2 has been preferred to the

TABLE 1 IASI-NG band characteristics

Band	Wave numbers (cm^{-1})	Channels	Atmospheric information
1	645.000 – 1,150.000	1 – 4,041	Temperature, surface properties, clouds, CO_2 and O_3
2	1,150.125 – 1,950.000	4,042 – 10,441	H_2O, CH_4 and N_2O
3	1,950.125 – 2,300.000	10,442 – 13,241	N_2O, CO and CO_2
4	2,300.125 – 2,760.000	13,242 – 16,961	Surface properties and CO_2

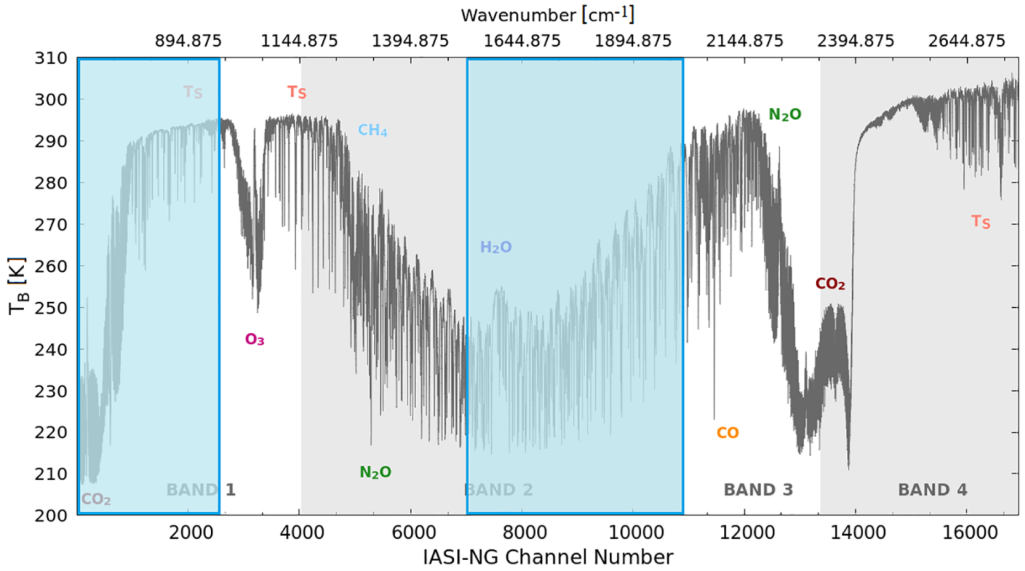

FIGURE 2 Example of a simulated IASI-NG spectrum. The spectrum is divided into four spectral bands, which are here depicted together with the sensitivity of the different spectral regions. The blue boxes highlight the preselected areas to be examined during this study. More specifically, the first 2,448 consecutive channels of band 1 and the last 3,601 consecutive channels of band 2 (from channel 6,841 up to 10,441) have been chosen (for a total of 6,049 channels). The corresponding spectral areas span from 645.000 to 950.875 cm^{-1} and from 1,500.000 to 1,950.000 cm^{-1}, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

first half since it will provide good information about water vapour, without being affected by the influence of trace gases (e.g., N_2O or CH_4). Moreover, this spectral area will be consistent with the one that the Infrared Sounder (IRS), which will fly on board the Meteosat Third Generation (MTG), will provide in the next few years. One can note that, having channels selected on almost the same spectral area that will soon be covered by a geostationary satellite such as MTG-IRS, could have positive side-effects for future studies. The areas just described are highlighted on the spectrum of Figure 2.

3 | METHODOLOGY

3.1 | Data assimilation

The data assimilation process aims at finding an optimal representation of the atmospheric state, by using the information provided by a background model state and by a set of observing systems. In other words, the best compromise between these different data groups is found knowing their respective accuracies (given by error covariance matrices).

The statistically optimal state can be obtained through the minimisation of cost function $J(\mathbf{x})$:

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}\left[\mathbf{y}-H(\mathbf{x})\right]^T\mathbf{R}^{-1}\left[\mathbf{y}-H(\mathbf{x})\right], \quad (2)$$

where $\mathbf{x}$ represents the model state vector, $\mathbf{x}_b$ the *a priori* background state vector estimated before the analysis is carried out, $\mathbf{y}$ the vector of the observations and H the observation operator. The background-error covariance matrix $\mathbf{B}$ and the observation-error covariance matrix $\mathbf{R}$ provide the error characteristics for the two counterparts. As a result, they define the weight to be attributed to each term in the aforementioned process of minimisation. The state that minimizes the cost function is written as $\mathbf{x}_a$ (i.e., the analysis state).

Since the simulated data are vertical profiles, we will use the cost function from Equation 2 in the context of 1D-Var assimilation experiments. The state vector $\mathbf{x}$, on the other hand, will consist of the temperature profile (54 levels over the whole atmosphere), the humidity profile (29 tropospheric levels) and surface temperature.

3.2 | Observation-error estimation

The observation errors consist of instrument noise, error in the observation operator, and representativeness and pre-processing errors. They can be split into systematic and random components. Systematic errors are usually biases that are corrected before the data are assimilated. As a consequence, the observation errors contained in the $\mathbf{R}$ matrix are the random ones.

Most of the channel selection methods employed for the previous generation of the atmospheric sounders primarily used matrices containing only the instrument noise, which was often roughly inflated in order to take the other errors into account. These matrices were diagonal

ones (e.g., Rabier *et al.*, 2002; Collard 2007). Moreover, in the case of IASI, the measurements are apodized. This introduces inter-channel correlation and produces a matrix with non-zero covariance bands. To be able to still use a diagonal matrix while avoiding the effects of apodization, the selected channels were chosen not to be adjacent. Since these approaches can be unrealistic, over the last decade most of the NWP centres started using correlated observation errors in the operational assimilation process (e.g., Bormann *et al.*, 2009).

With this in mind, for the present study we decided to use a full covariance matrix of the observation errors. Furthermore, our aim is rather to perform a selection among all the available channels in our pre-selected spectral area, without excluding *a priori* neighbouring channels.

The diagnostic procedure introduced by Desroziers *et al.* (2005) has been chosen to estimate the structure of a full **R** matrix. This method allows us to obtain variances and covariances of observation errors from *observation-minus-background* $[\mathbf{y} - H(\mathbf{x}_b)]$ and *observation-minus-analysis* $[\mathbf{y} - H(\mathbf{x}_a)]$ statistics. The **B** matrix projected onto the observation space $\mathbf{HBH^T}$ can also be diagnosed, using the difference between the analysis and the background in the observation space. These matrices are given by the following expressions:

$$\mathbf{R} = E\left\{[\mathbf{y} - H(\mathbf{x}_a)][\mathbf{y} - H(\mathbf{x}_b)]^T\right\}, \quad (3)$$

$$\mathbf{HBH^T} = E\left\{[H(\mathbf{x}_a) - H(\mathbf{x}_b)][\mathbf{y} - H(\mathbf{x}_b)]^T\right\}, \quad (4)$$

where E is the statistical expectation operator.

3.3 | Optimal selection

The methodology applied for the channel selection is the one suggested by Rodgers (1996) and has proved to be a good *a priori* method for the determination of an optimal channel set by Rabier *et al.* (2002). This method relies on evaluating the impact of the addition of one channel at a time on a scalar figure of merit reflecting the improvement of the analysis error over the background error. Compared to the original methodology, however, as mentioned in the previous subsection, the **R** matrix to be used will not be diagonal.

The figure of merit we chose to iterate the selection process is the *Degrees of Freedom for Signal* (DFS), which can be defined as:

$$\mathrm{DFS} := \mathrm{Tr}\,(\mathbf{I} - \mathbf{AB}^{-1}) \quad (5)$$

where **I** is the identity matrix and **A** the analysis-error covariance matrix.

It is known that the **A** matrix in the linear case can be expressed as:

$$\mathbf{A} = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}, \quad (6)$$

with **H** the Jacobian matrix, which is the linearized version of the observation operator, and $\mathbf{H}^T$ is its transpose.

This means that the $\mathbf{AB}^{-1}$ term becomes:

$$\begin{aligned}\mathbf{AB}^{-1} &= (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{B}^{-1} \\ &= [\mathbf{B}\,(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\,]^{-1} \\ &= [\mathbf{I} + \mathbf{BH}^T\mathbf{R}^{-1}\mathbf{H}\,]^{-1}. \end{aligned} \quad (7)$$

Consequently, Equation 5 can also be written as

$$\mathrm{DFS} = \mathrm{Tr}[\mathbf{I} - (\mathbf{I} + \mathbf{BH}^T\mathbf{R}^{-1}\mathbf{H})^{-1}]. \quad (8)$$

More to the point, the DFS parameter used in the selection will be determined by the contribution of temperature, humidity and skin temperature. It will be named, from now on, *Total* DFS:

$$\mathrm{DFS_{Tot}} = \mathrm{DFS_{Tem}} + \mathrm{DFS_{Hum}} + \mathrm{DFS_{SkTem}}. \quad (9)$$

Each one of these contributions is computed as the sum of the 54 diagonal elements corresponding to temperature, 29 for humidity (1D-Var performs the minimization for humidity only on the 29 lowest levels) and one for skin temperature.

The first step in selecting channels with this method will consist of computing the $\mathrm{DFS_{Tot}}$ for every single channel among the C pre-selected channels ($C = 6{,}049$ in this study). The one with the highest value of $\mathrm{DFS_{Tot}}$ will represent the first element in our selection. We will then proceed by computing the $\mathrm{DFS_{Tot}}$ for every pair of channels consisting of the first one just selected and each of the remaining $C - 1$ channels not yet taken. The second element in our selection will be the one that, in addition to our first choice, will provide the maximum value of $\mathrm{DFS_{Tot}}$. In a similar way, we will proceed for the subsequent steps. In other words, the channel selected at each iteration will be the one that, together with the others already selected, will produce the maximum value of $\mathrm{DFS_{Tot}}$.

Such an iterative method can be stopped at a certain channel's threshold, to be defined based on the needs of the specific research study, or carried out until all channels have been chosen.

4 | CASE-STUDY

The actual data from IASI-NG will be available only in a few years. Accordingly, in order to be able to perform a

selection among its channels, a carefully simulated dataset is necessary. For this purpose, a database of simulated observations has been built by Andrey-Andrés *et al.* (2018).

The database contains simulated observations using IASI's scan geometry for IASI and two possible versions of IASI-NG (KBr and ZnSe evaluated prism materials). It covers four dates in the middle of each season from 2013 in order to provide a maximum of meteorological variability: 4 February (Northern Hemisphere winter), 6 May (NH spring), 6 August (NH summer) and 4 November (NH autumn). A total of 5,242,047 simulated perfect measurements, in day/night illumination and sea/land conditions, is obtained through the use of the RTTOV (Radiative Transfer for TOVS) radiative transfer model, version 11.3 (Hocking *et al.*, 2015). In our study, we use the simulated IASI-NG data obtained in Andrey-Andrés *et al.* (2018) by adding a random Gaussian noise to the simulated perfect measurements, the noise corresponding to the material KBr which was finally selected (red curve in Figure 6a below).

Atmospheric profiles of temperature, humidity and some chemical compounds used to create the simulated dataset are thus provided. They are used as the 'true' state in our study. Background profiles for our 1D-Var retrievals are created from this true state to which a random Gaussian noise is added. This noise comes from the **B** matrix used in this work, which is the one provided in the NWP SAF package containing the 1D-Var software (Smith 2016). Section 5 gives more details about this matrix.

For the purpose of this project, dealing with more than 5 million simulations would have been too expensive in terms of computational resources. As a result, we selected subsets of data to be used as a case-study dataset.

4.1 | Subset for Desroziers diagnostic

In order to make the computation efficient, the results reliable and the time of execution acceptable, the number of profiles to use for the Desroziers diagnostic procedure should be at least comparable to the 6,049 channels to be treated. Therefore, it has been empirically evaluated that the number of profiles to be retained should be at least 6,000.

With this in mind, to be able to isolate a representative sample of profiles, we required the following criteria to be met:

- Instrument Zenith Angle (IZA) < 2.8° (i.e., at nadir)
- Over sea (water)
- Clear sky
- Day/night illumination conditions.

TABLE 2 Detailed number of profiles to be used for the estimation of the observation error by the Desroziers diagnostic

	February	May	August	November	Total
Day	569	622	1,110	884	3,185
Night	950	1,079	523	530	3,082
Total	1,519	1,701	1,633	1,414	**6,267**

Furthermore, we selected observations with surface temperature above 273.15 K, in order to avoid dealing with areas affected by sea ice.

A total of 30,210 observations was found to match all the aforementioned criteria. However, given the initial preconditions, this number was still too high.

The approach chosen to thin out this subset was to retain a single pixel per FOV on each scan line, or by spacing each observation from another by at least one scan line.

We considered two sets of months with comparable climate features: February/November and May/August. We split the observations between day/night illumination conditions. By doing so, we reached an optimal number of 6,267 observations (Table 2).

Figures 3a,b show the observation distribution by highlighting the month and the illumination conditions. The profile distribution is geographically homogeneous.

4.2 | Subset for channel selection

For the channel selection stage, on the other hand, the subset of 6,267 profiles previously chosen can be further reduced. Performing the selection on a smaller subset of profiles that represent both typical and extreme cases can actually be equally and optimally fulfilling the work targets. For this purpose an *agglomerative hierarchical clustering technique* has been used (e.g., Kaufman and Rousseeuw 2005; Anderberg 2014).

There exist many agglomerative algorithms, which only differ in the way they group similar elements into classes. The parameter we chose to assess the similarity between two profiles i and j is the distance:

$$D(i,j) = \sqrt{\sum_{k=1}^{K} \{D_k(i,j)\}^2}. \tag{10}$$

The quantity $D_k(i,j)$ was defined using the Chevallier *et al.* (2006) approach:

$$D_k(i,j) = \sqrt{\sum_{m=1}^{N} \left(\frac{\theta_{ik}(m) - \theta_{jk}(m)}{\sigma_{\theta k}(m)} \right)^2}, \tag{11}$$

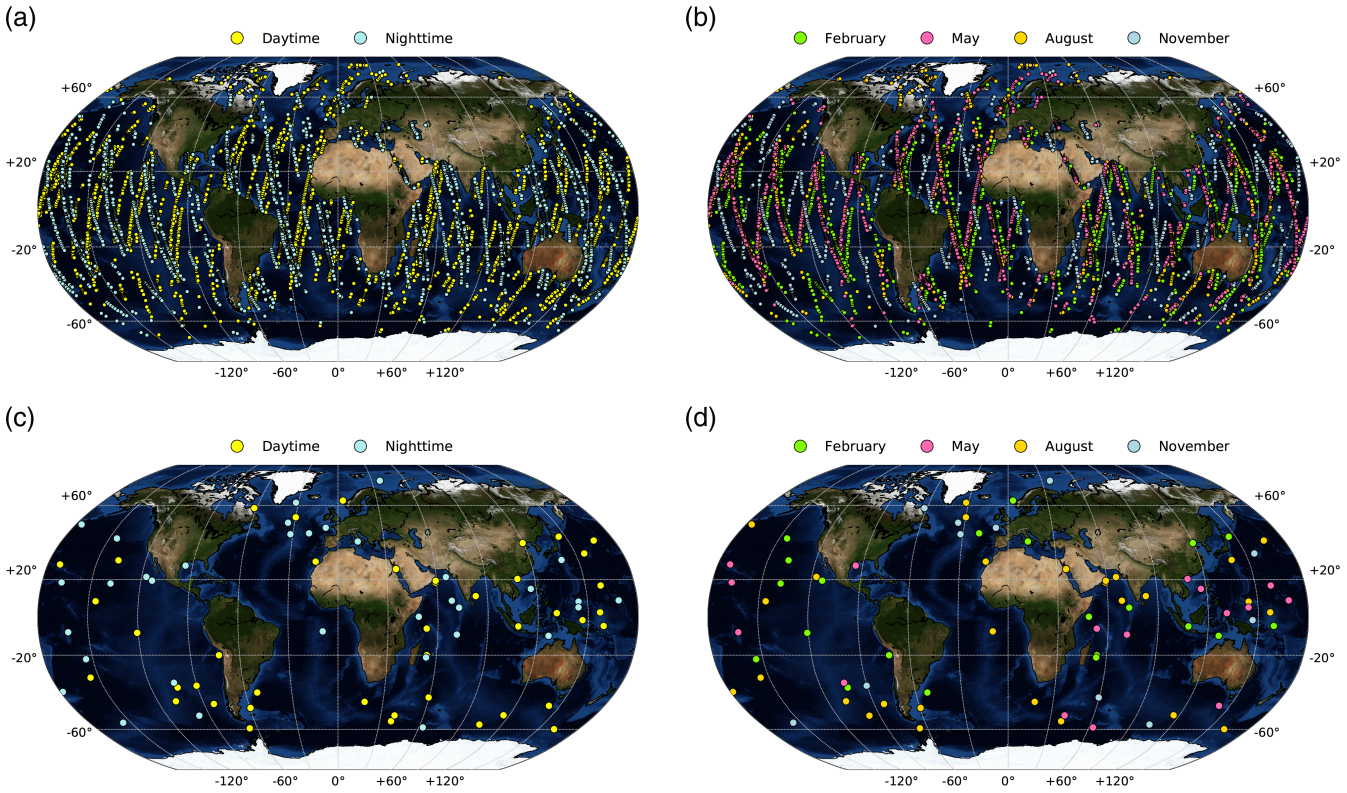

FIGURE 3 Geographical distribution of the profile sets taken as the case-study. (a, b) show the 6,267 profiles preselected to be used for the observation error evaluation. They are displayed in different colours depending on (a) the illumination conditions and (b) the time of the year. (c, d) refer to the 77 profiles picked up, starting from the 6,267 subset, through a hierarchical agglomerative technique in order to carry out the proper selection phase [Colour figure can be viewed at wileyonlinelibrary.com]

where k refers to the variables defining our atmospheric profiles (namely temperature and humidity), $\theta_{ik}(m)$ represents the variable k of the i profile at the m pressure level, N is the number of atmospheric pressure levels ($N = 54$ in our study), and $\sigma_{\theta k}(m)$ is the standard deviation of $\theta_k(m)$.

The difference between each profile and the remaining ones has been computed as in Equation 10. At this point, the profiles deemed to be the most similar based on the values of D, were gathered through an iterative process. The procedure was stopped after 77 groups were assembled. This threshold has been chosen purely for computational efficiency. At this stage, a representative profile of each group has been taken.

In the end, through this process we isolated 77 profiles representative of contrasted atmospheric conditions, as shown in Figure 4 for both temperature and humidity. Table 3 reports the detailed number of chosen profiles per month and time of the day. It shows that the cases are quite well distributed in terms of day/night illumination conditions, with just a slight predominance of daytime scenarios (56% daytime versus 44% night-time). Concerning the different months, November is the one with the lowest number of selected profiles, followed by May, February and August in this order. The geographical distribution of this subset is displayed in Figure 3c,d. This selection ensures the coverage of the three major geographical areas (Tropics, midlatitudes and polar regions). The polar area is the least covered with only three profiles. However, this lack is compensated by several midlatitude cases corresponding to winter scenarios, in both Northern and Southern Hemispheres. This ensures a sufficient quantity of profiles with polar-like characteristics.

5 | RESULTS OF THE DESROZIERS DIAGNOSTICS

In order to implement the Desroziers method for the evaluation of a full **R** matrix, observed, background and retrieved values are required (Subsection 3.2). While the first two can be obtained from the simulated database, to get the retrievals we had to carry out 1D-Var experiments.

To initialise the 1D-Var, **B** and **R** matrices are needed (Subsection 3.1). Again, the **B** matrix exploited for this work is the one provided in the NWP SAF package containing the 1D-Var software (Smith 2016). This **B** matrix consists of one block containing temperature (in K) error

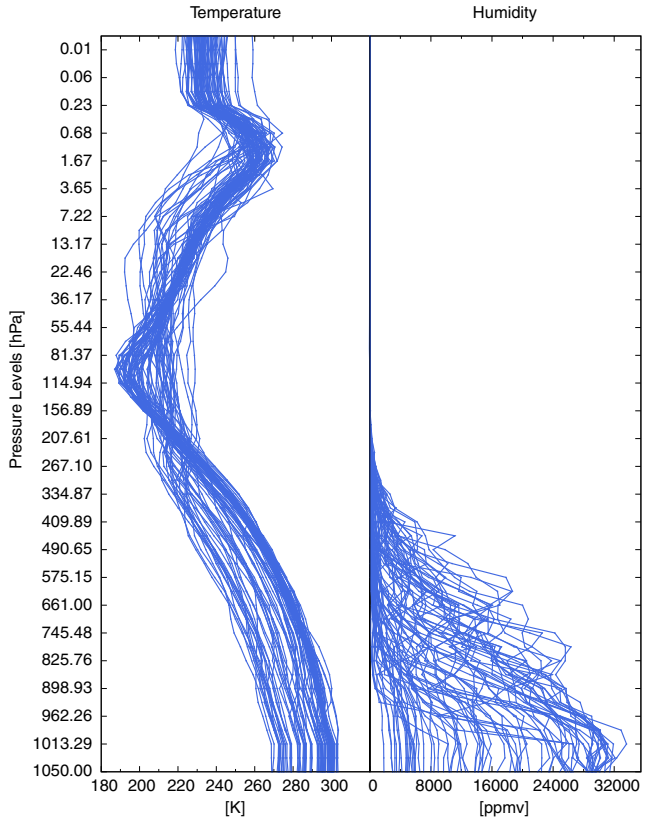

FIGURE 4 Temperature and water vapour content at the 54 used pressure levels of the 77 profiles chosen as a case-study to perform the channel selection [Colour figure can be viewed at wileyonlinelibrary.com]

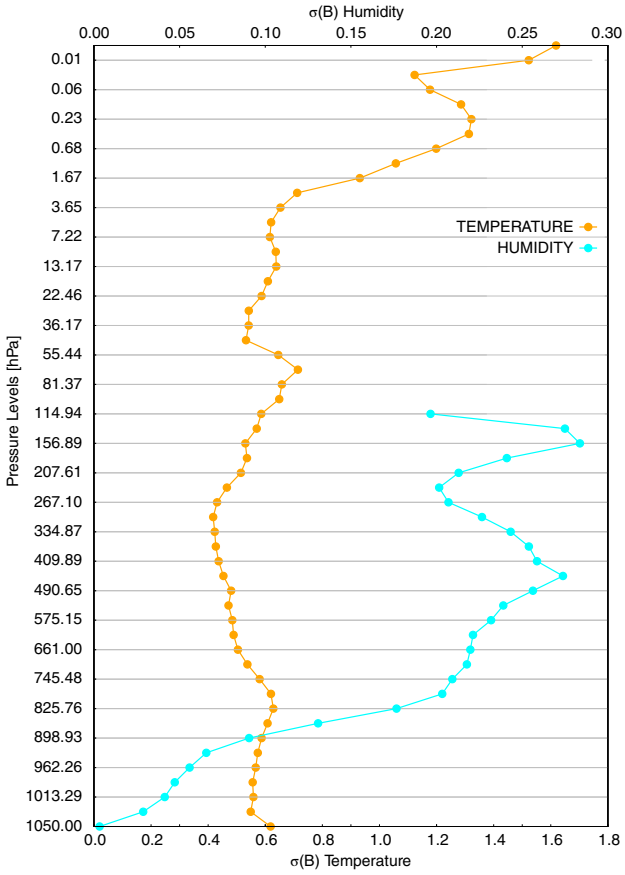

FIGURE 5 Background-error stardard deviation on the 54 used pressure levels for temperature (K) and 29 levels for humidity (g·kg^{-1}, expressed as ln(q)) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Detail of the profiles chosen as a case-study for the channel selection step through an agglomerative hierarchical technique

	February	May	August	November	Total
Day	12	7	18	6	43
Night	10	9	10	5	34
Total	22	16	28	11	**77**

covariance information on 54 atmospheric levels, one for the 29 bottom levels of humidity (represented as ln q) error covariances, one for surface temperature, one for surface humidity and finally one with information on skin temperature. Covariances exists among and between the errors of the different variables, but these latter are small compared to auto-covariances.

For this matrix we show in Figure 5 the background-error standard deviation σ^B for temperature and humidity. It is remarkable that the temperature values are higher in the stratospheric and higher tropospheric levels, while humidity values are very low for the lower levels. As regards the skin temperature, the σ^B value is equal to 2.03 K. The values of standard deviation that we show are similar to those employed in some global NWP models (e.g., Pereira and Berre 2006).

On the other hand, we do not yet have a **R** matrix at this stage. Consequently, we designed a diagonal matrix containing the standard deviations of *innovations* $\sigma(O-B)$ to be used for the initialization of the process only.

We executed 1D-Var on the 6,267 profiles illustrated in Subsection 4.1, by using all the 6,049 preselected channels in the early stage. With the retrievals thus obtained, we diagnosed a full **R** matrix through Equation 3.

As already mentioned, the sought **R** matrix will be used to carry out the selection procedure via DFS. Moreover, at a later time, it will also be used to perform 1D-Var assimilation experiments in order to test and validate the selection. However, to be used in 1D-Var, all the covariance matrices must be both symmetric and positive definite and this diagnosed matrix did not fulfil these two conditions. So, in anticipation of use in the assimilation process, the diagnosed **R** matrix had to be symmetrized (by taking the mean of the matrix itself and its transpose) and its negative eigenvalues had to be modified to become positive.

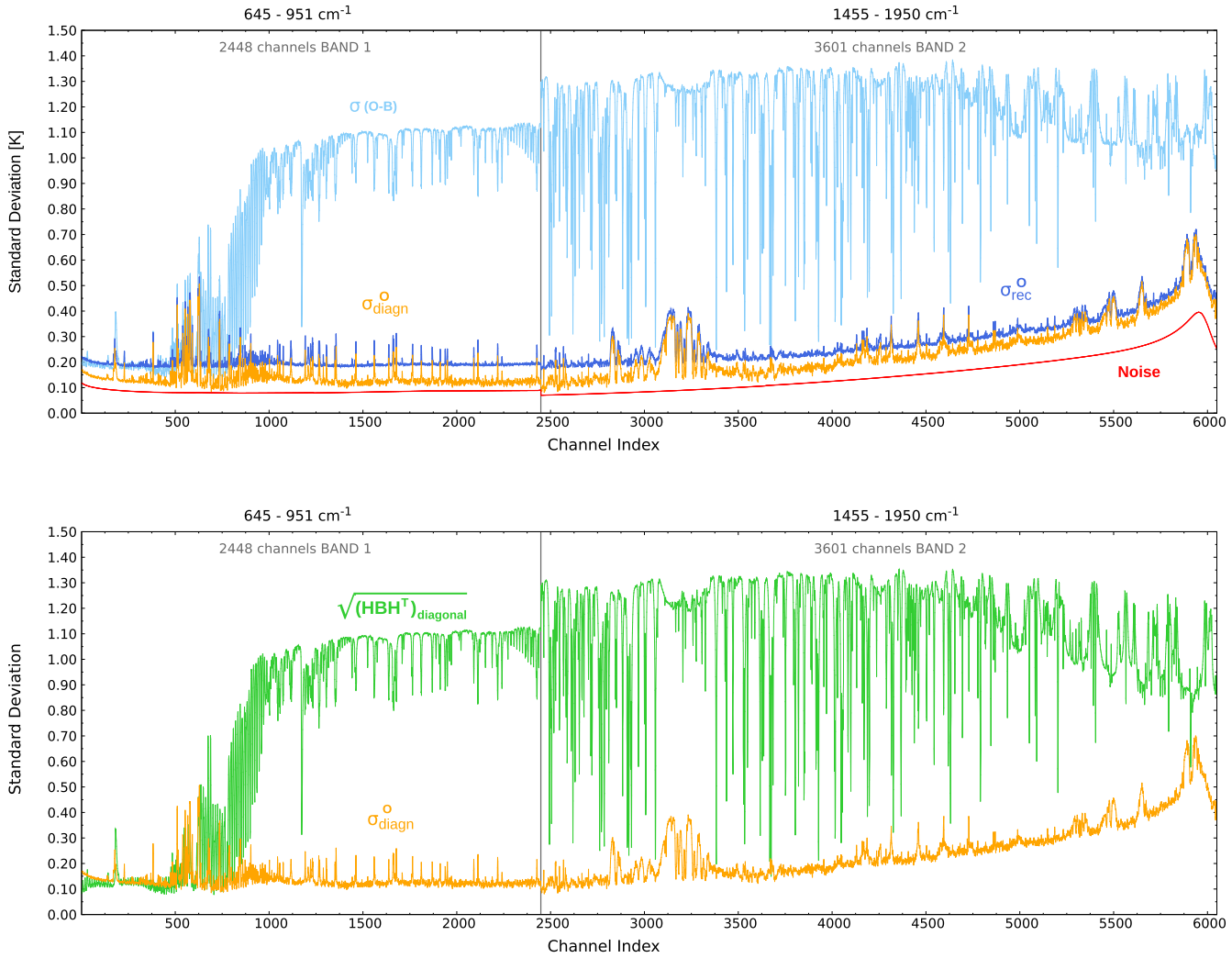

FIGURE 6 Diagnosed IASI-NG observation-error standard deviations from 1D-Var output (orange), standard deviations from 1D-Var output after reconditioning (dark blue), standard deviations of the innovations (light blue), instrument noise (red) and square root of the diagonal elements of the diagnosed $\mathbf{HBH^T}$ matrix (green) [Colour figure can be viewed at wileyonlinelibrary.com]

Modifying the eigenvalues of this kind of matrix can be tricky since the risk is to change too much the overall characteristics of the matrix itself, or rather the inter-channel correlation errors. Weston *et al.* (2014) suggest a couple of methods aiming to adjust the eigenvalues in a context where reconditioning the **R** matrix would make the minimization in the assimilation process more efficient. The most efficient method consists of modifying the eigenvalues by adding them an increment λ_{inc}:

$$\lambda_{\text{inc}} = \frac{\lambda_{\max} - \lambda_{\min} k_{req}}{k_{\text{req}} - 1}, \tag{12}$$

where $\lambda_{\max}$ is the diagnosed largest eigenvalue, $\lambda_{\min}$ the smallest and k_{req} a proper reconditioning number.

The approach chosen for this study is inspired by the above proposal. However, instead of directly applying an increment to all the eigenvalues, we first put all negative eigenvalues to zero. The increment was then computed through Equation 12, where $\lambda_{\min} = 0$ and $k_{\text{req}} = 2{,}000$. This value for the reconditioning number has been chosen empirically after performing several tests. Indeed, the comparison of the diagnosed observation-error standard deviations σ^O_{diagn} to standard deviations after reconditioning σ^O_{rec} and standard deviations of the innovations, showed that the smaller the reconditioning number, the less consistent the error estimation. In other words, using $k_{\text{req}} < 10^3$ results in σ^O_{rec} too high compared to σ^O_{diagn} and even higher than $\sigma(O-B)$. On the other hand, the σ^O_{rec} obtained with values $k_{\text{req}} > 10^3$ appear to converge. Thus the choice fell on the value $k_{\text{req}} = 2{,}000$.

Figure 6a shows the σ^O_{rec} compared to σ^O_{diagn}, $\sigma(O-B)$ and instrument noise. The diagnosed error standard deviation curve is shifted compared to the instrument noise by almost 0.05 K. This is expected since the

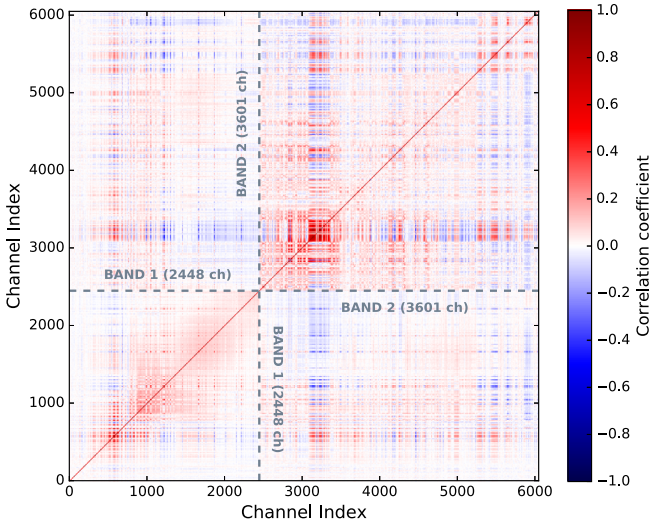

FIGURE 7 Diagnosed IASI-NG error correlations from 1D-Var output. The boxes outlined by the dashed lines highlight the results corresponding to the 2,448 channels of band 1 (bottom left) and to the 3,601 channels of band 2 (top right). The bottom right and top left boxes refer to the diagnosed cross-correlations [Colour figure can be viewed at wileyonlinelibrary.com]

instrument noise is not the only contribution to the overall observation error. Actually, the inter-channel correlations, which we aimed to build through this process, are introduced by a combination of other errors such as forward model and representativeness errors.

On the other hand, the standard deviation of the reconditioned matrix appears even more shifted upwards. It deviates by a factor 0.1 K from the instrumental noise. Also, although the reconditioning number is the best that could be chosen after careful consideration, for the first approximately 900 channels, the $\sigma^{O}_{\mathrm{rec}}$ shows values that sometimes are somewhat higher than the standard deviations of innovations. This behaviour is definitely due to the reconditioning method.

In order to use the same matrix both for the channel selection (DFS computation) and the subsequent evaluation through 1D-Var experiments, the reconditioned **R** matrix is the one that will be used from now on.

Figure 7 displays the diagnosed observation-error correlation matrix after reconditioning. Two specific boxes are highlighted. The bottom-left box represents the band 1-related part, from channel 1 to 2,448 (645.000–950.875 cm^{-1}). In this box, the highest correlation values are around channels 600 and 700 and all around the surface sensitive area. In the band 2 box (top right), the variability in correlations is higher than in the other part. The highest values are reached between indices 3,000 and 3,400 which correspond to channels from 7,392 to 7,792. This kind of behaviour matches with the actual variability of the absorption line in the corresponding spectral area.

6 | CHANNEL SELECTION

6.1 | Optimal selection

The selection method described in Subsection 3.3 was applied separately to each of the 77 representative profiles selected for the case-study. Actually, the more channels are chosen, the larger the **R** matrix to be inverted during the selection process to compute the DFS is, with a significant increase in the computational time as a result. Therefore, we decided to restrict the selection to 500 channels for each profile. This can be seen as a threshold considering that 500 is the current number of IASI channels distributed to the operational centres through the global telecommunication system (GTS). Indeed, if the principal component scores are also to be used for the IASI-NG data distribution, the number of disseminated channels will probably remain unchanged. For this reason, 500 is also the number of channels that we would like to provide to the community at the end of this study.

The Jacobians used in Equation 8 to compute the DFS, and throughout the rest of the work, have been obtained through the use of the RTTOV radiative transfer model.

Since the channel selection relies on the variation of the DFS values, it is interesting to observe how this parameter evolves with the number of selected channels. This will also give an information about the contribution of the different DFS components, as well as the input used in the process, to the channel choice.

Figure 8 shows the DFS evolution with the number of selected channels, averaged on the 77 case-study profiles. One can remark that the Total DFS growth is very fast in the very first part of the selection and it becomes slower and slower as the number of selected channels increases. The skin temperature, which largely contributes to the very first selected channels, stops providing a steady value of almost 1, which is its maximum value, just after these very few selection steps have taken place. This can be explained by considering the background error associated with this variable, which is the largest (2 K) when compared with the one associated with temperature in the rest of the atmospheric column (0.5 K on average in the tropospheric levels up to 1.6 K in the stratosphere, as shown in Figure 5). The IASI-NG spectrum has many surface-sensitive channels that have a strong skin temperature Jacobian. Consequently, just a very few channels are required in order to reduce the uncertainty of the skin temperature variable.

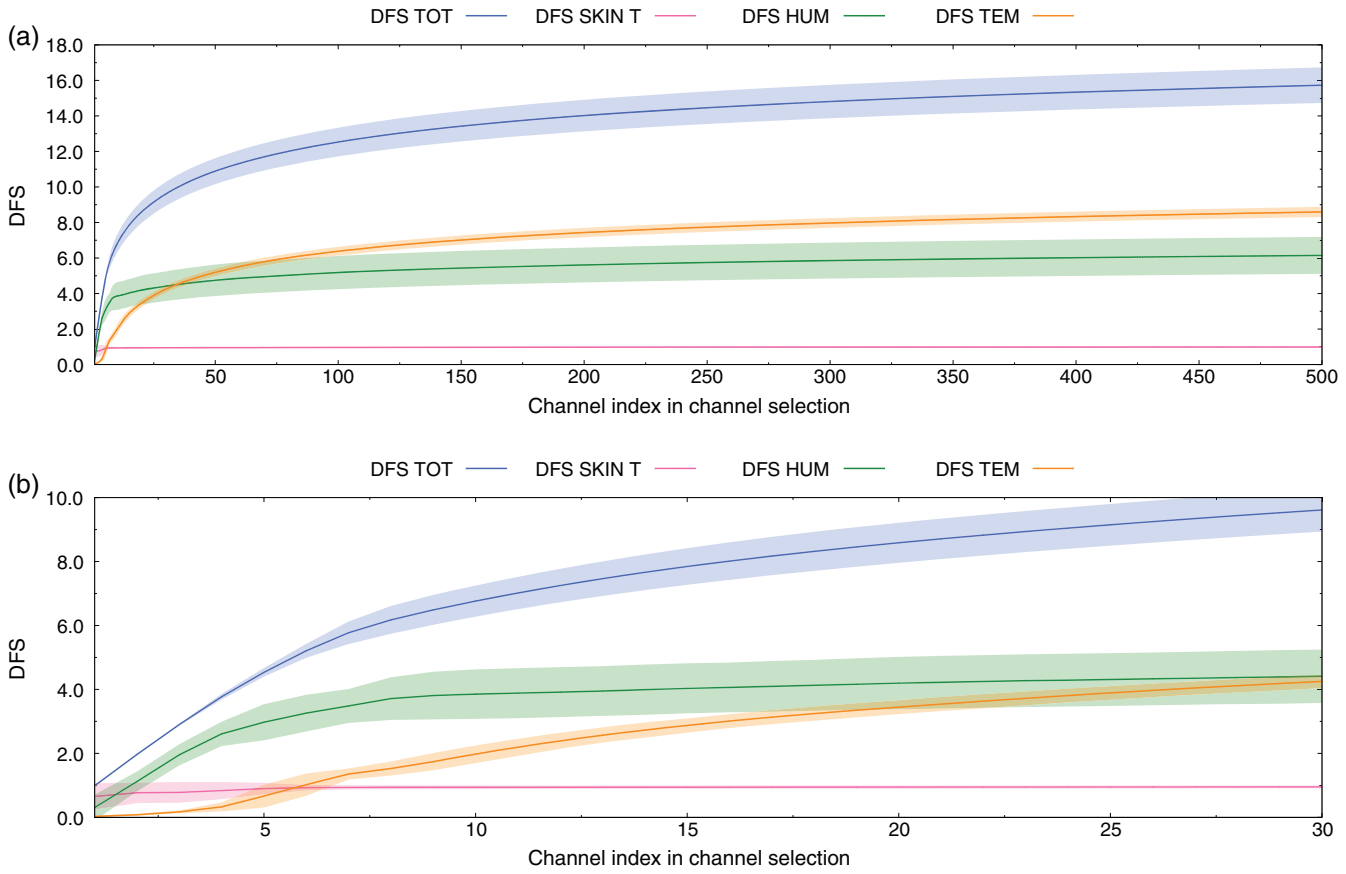

FIGURE 8 DFS trend values in channel selection averaged on the 77 case-study profiles. The solid blue line denotes the total DFS. The pink, green and orange lines refer to the the skin temperature, humidity and temperature terms, respectively. The shaded areas associated with each curve indicate the standard deviations for that specific term. (a) shows the complete result, while (b) shows a zoom on the first 30 steps in the process [Colour figure can be viewed at wileyonlinelibrary.com]

After that, until approximately the first 30 channels are selected, the greatest contribution to the Total DFS comes from the humidity term. The humidity background errors are larger than the temperature ones, in proportion, which is consistent with what is observed in NWP global models. Moreover, Figure 6b shows that the diagnosed background errors projected onto the observation space are larger for the water vapour channels (band 2) and surface channels (end of band 1). Many IASI-NG channels have strong humidity Jacobians in the mid to upper troposphere where the humidity background errors are the largest (Figure 5). The channel selection tries to reduce these errors by picking up channels mainly sensitive to humidity in those atmospheric levels.

The temperature DFS, on the other hand, starts providing a weak contribution to the Total DFS growth only after a couple of channels have already been selected. The predominant contribution becomes the temperature one just after the humidity gives way to it (i.e., after the 30 first selected channels). This is consistent with the trend of the background error, which has a minor influence compared with the other two variables. The Total DFS value reaches 15.8 at the end of the selection process (with DFS_{SkTem} of 1, DFS_{Hum} of 6 and DFS_{Tem} of 8.8). These specific results, as just explained, are influenced by the characteristics of the chosen background error and of the specific profiles considered and, of course, they would be affected by the choice of another input.

A study carried out on the channels selected in the first stages of the DFS selection process proves that these channels are primarily sensitive to surface or humidity. For readability purposes, Figure 9 summarizes the results obtained by focusing on the first three channels selected during each optimal selection. If three different channels were chosen for each profile, the total number of single channels would have been 231. However, there are channels in common to the selections and this leads to 49 unique channels, which are listed on the *y*-axis. The bar lengths illustrate how many times the channel of interest has been selected among the three first steps of the DFS selection process. Among these 49 channels, 4 are sensitive to temperature, 12 to the surface and the

remaining 33 to humidity. The channel the most frequently selected is 1,382 (33 times). This is a channel showing a double sensitivity to both surface temperature and humidity. This specific feature, associated with the characteristics of the background error already exposed, makes of it the first choice in 43% of the cases where the selection has been performed. This channel is then followed by 2,390 (selected 18 times), which is located in the atmospheric window spectral area as the previous one, but it is predominantly surface sensitive. Humidity-sensitive channels 6,862 and 6,863 follow, selected in 15 and 14 cases respectively. Their Jacobians exhibit large values in the region between 300 and 400 hPa. Similarly, these channels are sensitive to areas where the background errors for humidity are the largest.

As previously mentioned, the leading idea is to have, at the end of this study, a selection of 500 channels, since this number would be consistent with the one currently distributed to NWP centres from IASI. At this stage, however, we do not have yet a total number of 500, but rather an optimal selection of 500 channels per profile.

Since the 77 selections are all different from each other, we performed a statistical study in order to understand how many channels are in common and how often they appear. The histogram in Figure 10 summarizes the results thus obtained. Each bar shows the percentage of the initial 6,049 channel group which has been chosen and how often, or rather in how many cases, those channels have been selected. From these data one can deduce that 65% of the whole 6,049 initial group is never selected. Among the remaining 35 % channels that have been selected at least in one case (namely 2090), only 109 are always selected (1.8 % of the initial set). Through this study it was also possible to identify the 500 most frequently selected channels, or rather the channels selected in at least 37.7% of cases. It is this last group that we decided to evaluate as the sought channel selection. The complete list is provided in Table A1.

Among these 500, 300 channels come from band 1 and the remaining 200 from band 2. Their spectral positions are shown in Figure 11. Spectral coverage is generally homogeneous. The less populated area appears to be the surface-sensitive one. However, we still select 24 channels for this spectral region. The selected channels are preferably channels whose weighting functions are peaking in the higher stratosphere, where the background error is higher. For this reason, their contribution to the DFS is very strong, and they end up being the prevailing channels.

Figure 12 shows the percentage and number of channels in bands 1 and 2 for each selection frequency. The majority of channels never selected belong to band 2 (2,329 versus 1,630 in band 1). The same applies to the

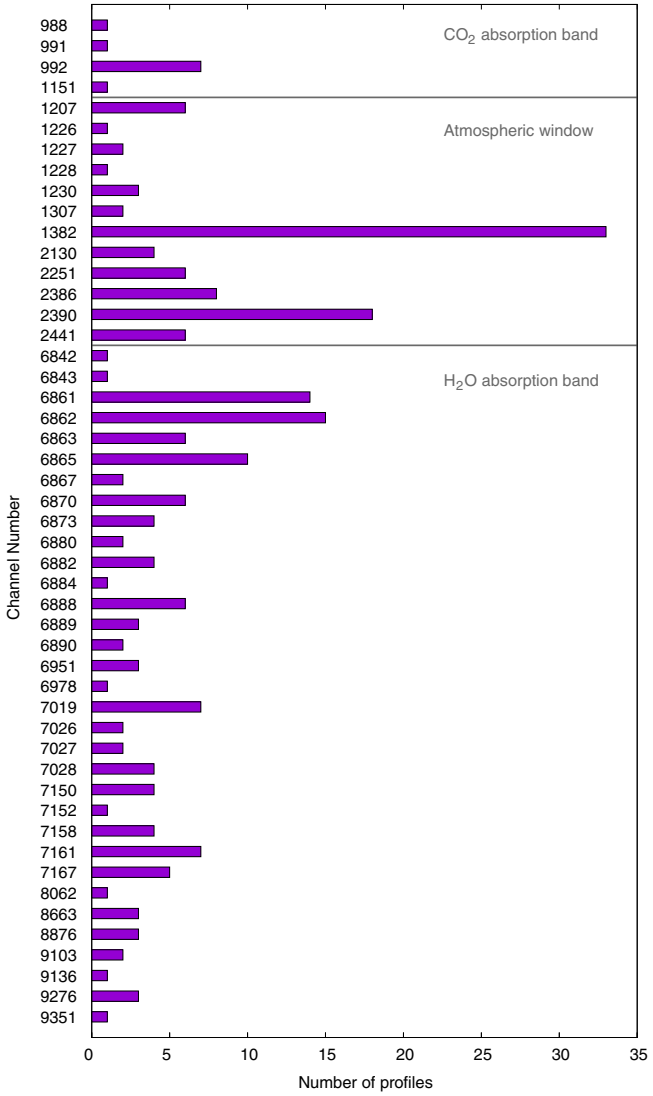

FIGURE 9 Channels selected among the first three on each of the 77 profiles. The *y* axis lists channels satisfying this characteristic. The bar width denotes the number of cases the channel in question has been selected among the first three. Dashed lines delimit the subsets of temperature, surface and water vapour sensitive channels [Colour figure can be viewed at wileyonlinelibrary.com]

channels selected in at least one case (1,272 band 2 versus 818 band 1). However, the more the selection frequency increases, the more the trend reverses and the prevailing contribution to the total comes from band 1. Indeed, the channels always selected mostly belong to band 1 (96 versus 13 in band 2). The 13 humidity channels have large sensitivity to the upper troposphere/lower stratosphere with Jacobians peaking at 200 hPa (not shown). These results are consistent with the channels most frequently selected among the first three from band 2. The 96 temperature channel Jacobians show large values in the upper troposphere and mainly in the stratosphere.

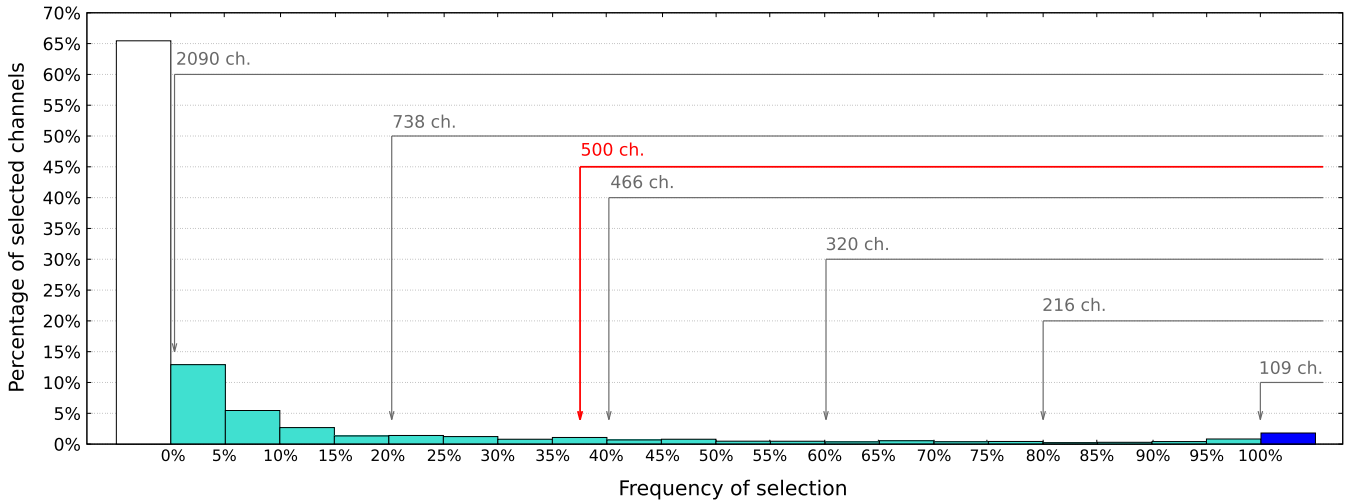

FIGURE 10 Percentage of the initial 6,049 channel group that has been selected per number of profiles (frequency of selection). The white column represents the percentage of never-selected channels, and the blue one the channels selected in all cases. All the intermediate groups are depicted in cyan. The grey arrows indicate the number of channels selected with a frequency of 100, 80, 60, 40, 20% and those selected at least once (frequency >0%). The red arrow highlights that the 500 most frequently chosen channels correspond to the channels selected in 37.7% of cases [Colour figure can be viewed at wileyonlinelibrary.com]

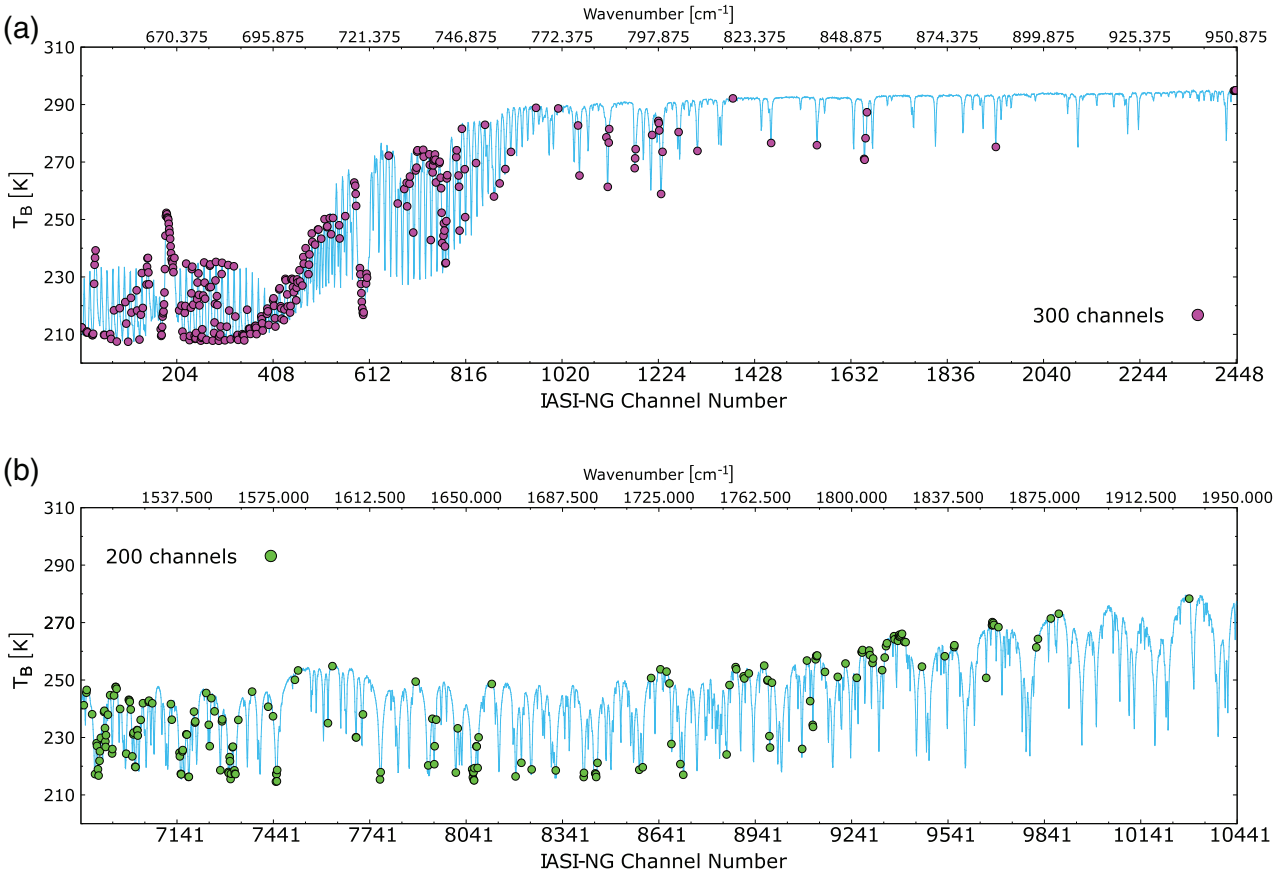

FIGURE 11 Spectral location of the 500 selected channels: 300 in band 1 (magenta dots) and 200 in band 2 (green dots). (a) represents the preselected band 1 area (from 645.000 to 950.875 cm^{-1}), and (b) the band 2 area (from 1,500.000 to 1,950.000 cm^{-1}) [Colour figure can be viewed at wileyonlinelibrary.com]

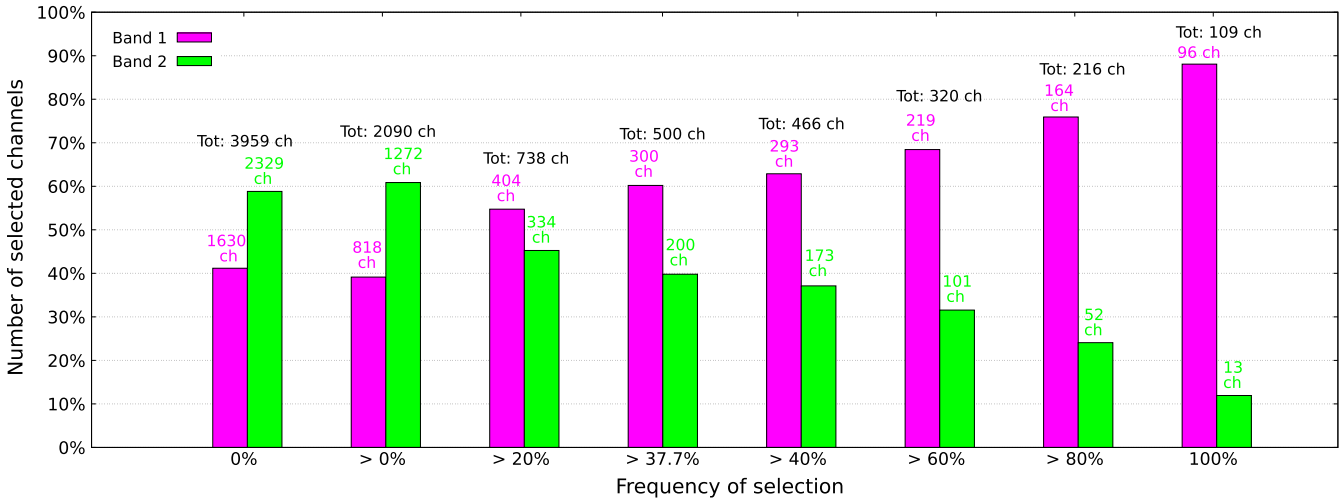

FIGURE 12 Band 1 (magenta) and band 2 (green) contributions to the channels picked up in eight main subsets from the study as in Figure 10. The groups at issue correspond to channels: never selected (0% of selection frequency), selected at least once (frequency >0%), selected with frequencies >20%, >40%, >60%, >80%, always selected (frequency 100%) and selected in >37.7% of cases (namely the 500 most frequent channels). For each group, the total number on which the percentage has been calculated is also shown [Colour figure can be viewed at wileyonlinelibrary.com]

6.2 | Impact of the channel selection on the retrievals

In order to test the selection contribution to the assimilation process, 1D-Var assimilation experiments have been performed on the 6,267 pre-selected profiles (Subsection 4.1). The idea was to test the selection on a representative sample of profiles, which is why we used this larger subset than only the 77 profiles on which the selections were made.

In addition to the 500 most frequently selected channels, other groups have also been tested in the same way. The groups chosen are those containing channels selected at least once, in 20, 60 and 100% of cases.

The results have been evaluated statistically by computing, for each atmospheric level in all profiles, the *background-minus-truth* $(\mathbf{x}_b - \mathbf{x}_t)$, or background error, and the *retrieved-minus-truth* $(\mathbf{x}_a - \mathbf{x}_t)$, or analysis error. For these parameters, an average and a standard deviation have been computed on the 6,267 cases.

Figure 13 displays the results in terms of Rate of Improvement (ROI) of the retrievals compared to the background. This parameter has been computed for each atmospheric level j as follows:

$$\mathrm{ROI}(j) = \frac{\sigma\left(\mathbf{x}_a - \mathbf{x}_t\right)(j) - \sigma\left(\mathbf{x}_b - \mathbf{x}_t\right)(j)}{\sigma\left(\mathbf{x}_b - \mathbf{x}_t\right)(j)}, \qquad (13)$$

where $\sigma\left(\mathbf{x}_a - \mathbf{x}_t\right)(j)$ and $\sigma\left(\mathbf{x}_b - \mathbf{x}_t\right)(j)$ represent the analysis- and background-error standard deviations respectively, computed on the 6,267 profiles and for each atmospheric level j. The atmospheric levels where the ROI parameter is computed range between 1 and 54, except for humidity where the 1D-Var performs the minimisation over the 29 lowest levels.

The overall outcome always shows an improvement of the retrievals compared to the background when using the various channel selections. When examining the results in more details, it appears that the more channels are added, the better the results are.

The use of the always selected channels allows an improvement of the temperature and humidity retrieval all along the profiles. The results obtained using the 109 set of always selected channels show, for temperature, a stronger improvement around 200 hPa where the ROI reaches values up to −25%. Two other secondary maxima are present at 950 hPa and in the upper atmosphere. For humidity, on the other hand, the best improvement occurs again in the higher troposphere, with maxima up to 40% (Figure 13). Another maximum is observed at around 700 hPa. These results are consistent with the sensitivities of the selected channels.

Moving from 109 to 320 channels, that is, to those channels chosen in 60% of cases, the ROI values decrease further. This is especially true for humidity, for which the improvements are higher in the middle/high troposphere, where a maximum of 50% is reached at 450 hPa. If we look at the temperature, adding channels almost uniformly improves results up to 200 hPa altitude. Above this level, the improvement is drastically reduced until it stops at pressures lower than 50 hPa. At the same time, we find a very good impact in the lower atmospheric layers (from 960 to 990 hPa approximately).

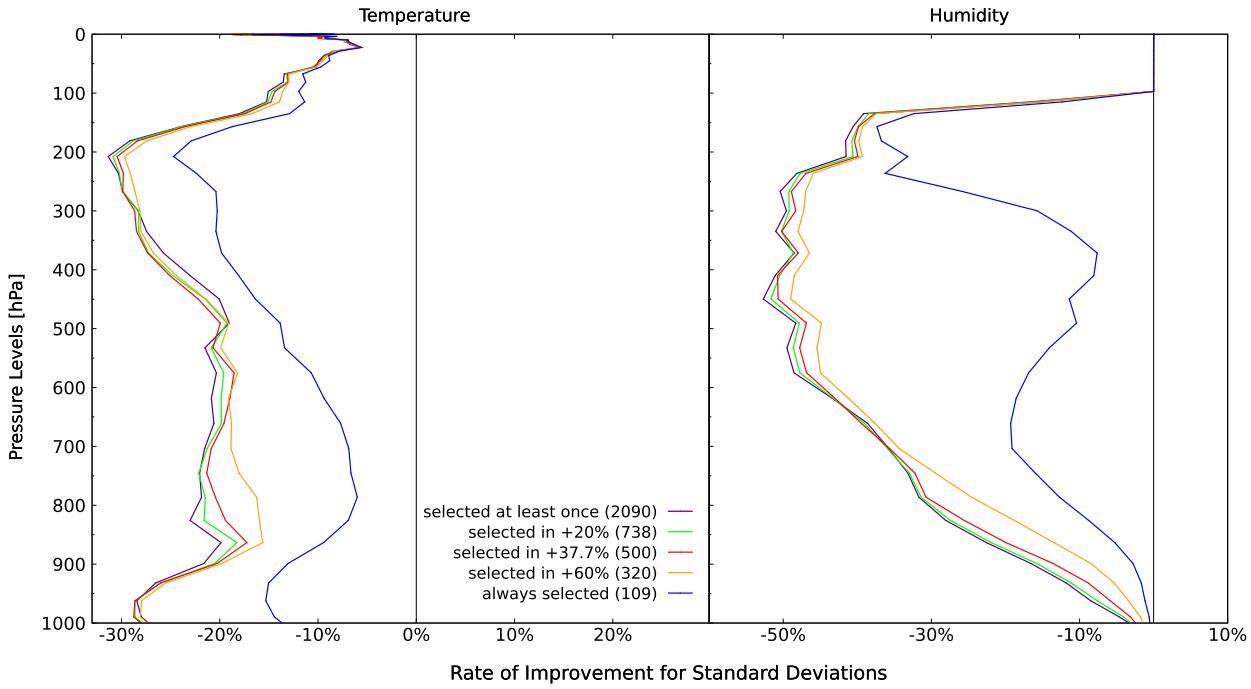

FIGURE 13 Rate of Improvement (ROI) of analysis-error standard deviations compared to the background ones averaged on 6,267 profiles, for temperature (left) and humidity (right). The figure compares the results obtained for the groups of channels chosen at least once, and in 20, 60, and 100% of cases and the ROI obtained through our 500 channel selection. Negative values mean retrievals were improved relative to the background [Colour figure can be viewed at wileyonlinelibrary.com]

Using our selection of the 500 channel selection further improves 1D-Var outcome. For temperature the improvement especially occurs between 600 and 900 hPa, while for humidity it is smaller but spread over almost all of the minimised atmospheric layers. Adding channels beyond this threshold leads to some small improvements, but mostly the performance saturates, even though it is sometimes observed that the retrievals obtained through the use of a larger number of channels can be slightly worse than those obtained with a less populated group. More specifically, this behaviour is noticeable between 260 and 500 hPa (degradation of less than 2%) for temperature and between 600 and 700 hPa (degradation less than 1%) for humidity.

The general ROI trend can be explained by examining the Jacobians of the involved channels. This is true for all the channel groups, however we will show those concerning the 500 channel selection only for a midlatitude profile (Figure 14) as an example, and since it is the most interesting for our purposes. Band 2 channels have a strong temperature sensitivity between 250 and 500 hPa, which is exactly where the ROI degradation occurs when increasing the number of channels. Band 1 channels have almost no sensitivity to humidity. A low sensitivity (with values of the order of 0.1 K) is shown by the channels with indices between 180 and 200, which are spread in the spectral range between 703.000 and 950.875 cm^{-1}. These same channels are those showing sensitivity to skin temperature. Among them are the 24 channels selected in the atmospheric window and which, for this reason, have the highest sensitivity values to skin temperature. Non-zero values are also present for the last six channels in the 500 list, which lie between 1,850.000 cm^{-1} and the end of the band.

Figure 15 presents a cumulative ROI averaged on the atmospheric column. It is evident that the best value is obtained with the largest number of channels (channels selected at least once). It is also remarkable that, by adding channels after the 500 threshold, some sort of saturation is reached. In other words, adding channels after this threshold brings very small improvement to the overall result (less than 0.6% for the total ROI). As shown in Figure 13, if computed only on the 29 lowest atmospheric levels, the improvement of the average ROI is larger (27.6%) than the one computed over the whole atmospheric column (22.9%) showing that the largest improvement provided by our selection takes place in the troposphere.

7 | RANDOM SELECTION

In order to evaluate the optimality of the already well-known channel selection methodology applied in this work, a completely random selection was performed as

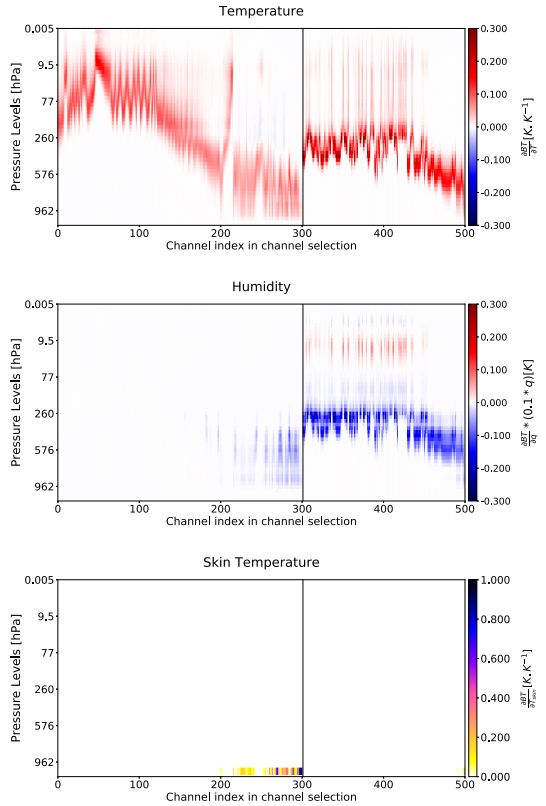

FIGURE 14 Temperature, humidity and skin temperature Jacobians for the 500 selected channels. The values are displayed for a midlatitude profile (44°31′15.6″N, 19°18′57.6″W) [Colour figure can be viewed at wileyonlinelibrary.com]

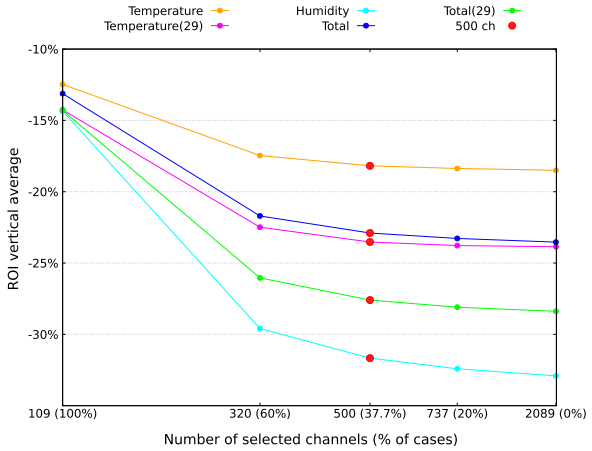

FIGURE 15 Rate of Improvement (ROI) of retrieval standard deviations compared to the background ones averaged over the 6,267 profiles and over the atmospheric column: humidity data obtained by averaging the 29 levels for which the parameter is minimized in 1D-Var (cyan); the temperature data obtained using the same 29 atmospheric levels as for humidity (pink); and the average of the temperature values over the 54 1D-Var atmospheric levels (orange). The blue line represents the average for both temperature and humidity and the green line is the same but for only the 29 levels in the troposphere. The values are calculated for the five groups of channels studied: channels selected in 100%, 60%, 37.7% (red dots) and 20% of cases and channels selected at least once [Colour figure can be viewed at wileyonlinelibrary.com]

in Martinet *et al.* (2014). Indeed, assessing the impact of this further selection on the analysis process has been considered to be a good sanity check for this purpose.

First of all, 77 different 500 random channel selections were produced without any constraints or conditions except that the aforementioned channels belonged to the spectral areas covered during this study (Subsection 2.2).

After that, 1D-Var assimilation experiments have been performed by using a different random channel selection for each of the 77 case-study profiles. The aim was to compare these results with those from the assimilation experiments carried out using for each of the 77 cases its optimal selection obtained through the DFS method (Subsection 6.1). Figure 16 shows the average improvement of standard deviations of the retrievals compared to the background for random and DFS selection. The use of a different random selection for each profile leads to the same overall improvement trend as for the optimal selection. However, using selections obtained through the DFS method almost always leads to a better result. This is particularly true in the upper part of the atmosphere above 400 hPa for temperature. The only exceptions are between 450 and 600 hPa (difference lower than 1.5%) for temperature and between 350 and 550 hPa (lower than 2%) for humidity.

In order to more specifically evaluate the 500 channels most frequently selected via DFS, another test has been carried out. Only one 500 random selection among the previous 77 has been chosen and used to carry out 1D-Var assimilation experiments on all of the 77 profiles. The average result is shown in Figure 17 together with the average improvement obtained by exploiting the aforementioned 500 DFS channel selection for all the case-study. As well as for the optimal selections, the improvement trends are quite similar, although the DFS selection provides better results than the random one.

This random selection consists of 90 channels belonging to the CO_2 absorption band, 115 surface sensitive and 295 sensitive to humidity. The large number of surface channels may also be beneficial to the retrieval in the boundary layer.

Figure 18 shows the different average improvement obtained with the four channel selections. The rate between an optimal selection and a random one is comparable for both cases when a fixed selection or a variable one is used. However for all the five parameters, we find a better score if using a specific channel selection for each profile.

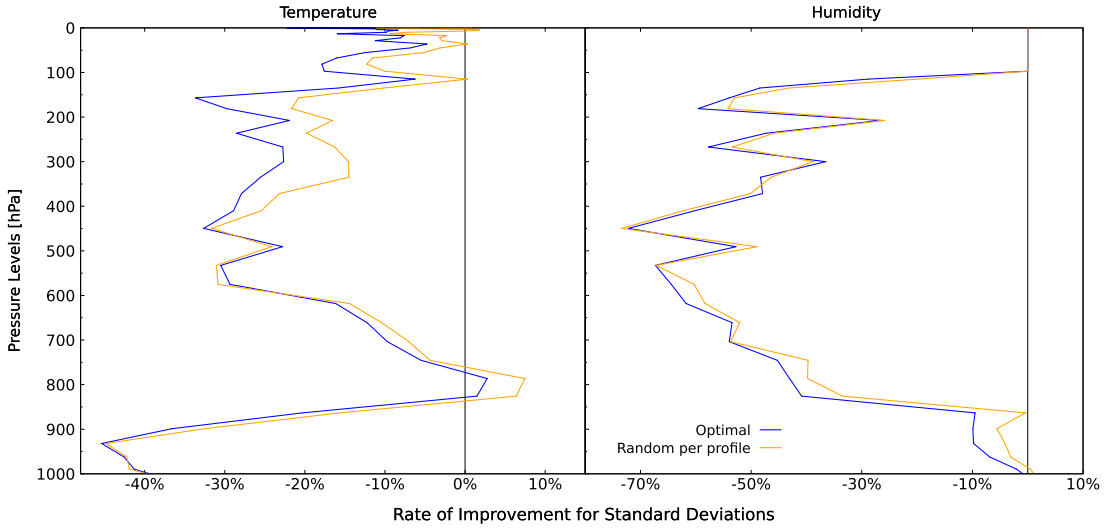

FIGURE 16 ROI of retrieval standard deviations compared to the background ones, for temperature (left panel) and humidity (right). The plot shows in blue the average of the ROI obtained using the optimal channel selection (DFS) method for each of the 77 case-study profiles. The orange curve represents the averaged ROI for the 1D-Var retrievals obtained using a different random selection for each of the 77 profiles [Colour figure can be viewed at wileyonlinelibrary.com]

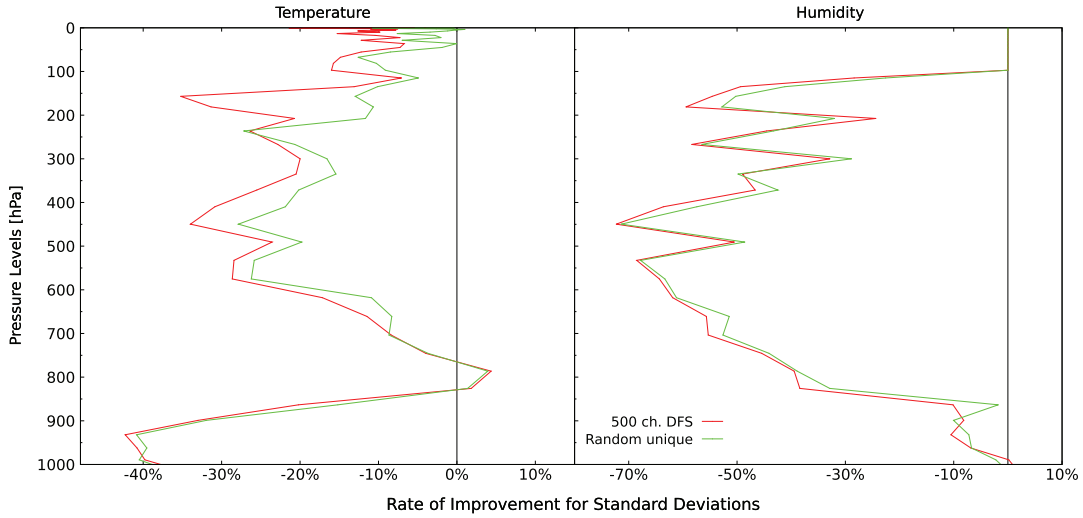

FIGURE 17 ROI of retrieval standard deviations compared to the background ones, for temperature (left panel) and humidity (right). In red is the average of the 77 case-study profiles of the ROI obtained each using the 500 channels most frequently selected through the DFS method. The green curve shows the results produced using the same random channel selection applied to all 77 profiles [Colour figure can be viewed at wileyonlinelibrary.com]

8 | CONCLUSIONS AND PERSPECTIVES

An IASI-NG channel selection was performed with the aim of serving as a starting point for future assimilations in the NWP context. The work was carried out on a database of atmospheric temperature and humidity profiles simulated with the specific purpose of serving as a basis for the preparation of this new instrument.

The method used to perform the selection is the iterative procedure by Rodgers (1996) for clear conditions, using DFS as a figure of merit for the implementation. This method has been applied on 77 representative profiles over the globe for various seasons.

Different groups of channels have been chosen at the same time, but the selection proposed in this paper consists of the 500 most frequently selected channels, or rather the channels chosen in 37.7% or more of the evaluated cases. The threshold of 500 was chosen because this number is currently used for the distribution of IASI channels to NWP centres. The selection consists of 300 channels belonging to band 1 (277 temperature and 23

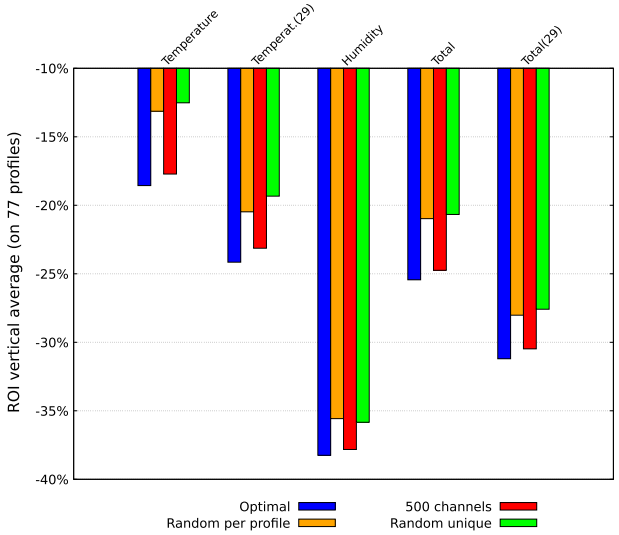

FIGURE 18 Histograms of ROI averaged over the 77 profiles and the vertical (total or partial) column, as a function of temperature, humidity and for the total. Blue bars refer to the optimal selection, the orange ones to a different random selection for each profile, the red ones to the 500 channel selection and the green bars to a unique random selection applied over the whole 77 profile subset [Colour figure can be viewed at wileyonlinelibrary.com]

surface-sensitive channels) and 200 from the water vapour band.

This selection, plus other groups of channels retained more or less frequently, have been tested on a larger set of 6,267 atmospheric profiles. The results obtained always show an improvement in the retrievals compared to the background profiles when using the selected channels. The more channels are used, the better the results. Regarding temperature, the improvement is especially pronounced in the very low layers (960 to 990 hPa approximately) and in the upper troposphere, with improvements reaching up to 30% at 200 hPa. For water vapour, on the other hand, the improvement is more pronounced in the medium/high layers (between 600 and 200 hPa) with improvements up to 50%. The selection of 500 channels proposed here completely follows this same trend. By examining the Jacobians of the channel selection on the case-study profiles, it has been observed that this behaviour is due to the sensitivity which is maximal for these regions.

The selection was also compared to a completely random selection. It has been found that the optimal selection and the proposed 500 channel selection perform better than random ones. The gain is particularly important for temperature in the upper part of the troposphere/stratosphere.

Each channel selection method has advantages and disadvantages. In further studies, it could be interesting to compare several methods using the same inputs (**B**, **R**, observations) and to assess the impact on the metrics we used, such as the DFS statistics and the rate of improvement. On the other hand, the choice of the **B** matrix may also have an impact on the selection. An intercomparison exercise using several **B** matrices provided by different NWP centres in the selection process may be of interest to evaluate the overall sensitivity of the selection to the background errors.

In conclusion, the 500 channel selection has been tested and it provides pretty good results. The selected channels depend on many components, such as the atmospheric profiles and the errors prescribed to the background and the observations. However, we have to remember that the selection has been performed on two limited areas of the spectrum. Future work may focus on spectral areas not yet explored in this study. Scanning and evaluating channels sensitive to other chemical species would make the selection even more complete and optimal. Cloudy and over-land case-studies could also be analysed. Moreover, it would be interesting to evaluate the IASI-NG selection in a more realistic context such as in a specific global model (e.g., the Météo-France Action de Recherche petite Echelle Grande Echelle, ARPEGE; Courtier *et al.*, 1991) through Observing System Simulation Experiments (OSSEs).

AUTHOR CONTRIBUTIONS

Francesca Vittorioso: Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing - original draft; Writing – review and editing. **Vincent Guidard:** Conceptualization; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – review and editing. **Nadia Fourrie:** Methodology; Project administration; Supervision; Writing – review and editing.

ACKNOWLEDGEMENTS

This research has received funding from the Centre National d'Études Spatiales (CNES) in the framework of the IASI-NG project. The authors would also like to thank Dr. Jean-François Mahfouf for his valuable advice which helped improve the quality of the manuscript.

ORCID

F. Vittorioso https://orcid.org/0000-0001-5483-2468

REFERENCES

Anderberg, M.R. (2014) *Cluster Analysis for Applications*. Cambridge, MA: Academic Press.

Andrey-Andrés, J., Fourrié, N., Guidard, V., Armante, R., Brunel, P., Crevoisier, C. and Tournier, B. (2018) A simulated observation

database to assess the impact of the IASI-NG hyperspectral infrared sounder. *Atmospheric Measurement Techniques*, 11(2), 803–818.

Aumann, H.H., Chahine, M.T., Gautier, C., Goldberg, M.D., Kalnay, E., McMillin, L.M., Revercomb, H., Rosenkranz, P.W., Smith, W.L., Staelin, D.H., Strow, L.L. and Susskind, J. (2003) AIRS/AMSU/HSB on the Aqua mission: design, science objectives, data products, and processing systems. *IEEE Transactions on Geoscience and Remote Sensing*, 41(2), 253–264. https://doi.org/10.1109/TGRS.2002.808356.

Bermudo, F., Rousseau, S., Pequignot, E. and Bernard, F. (2014). IASI-NG program: a new generation of infrared atmospheric sounding interferometer, pp. 1373–1376 in 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, Canada. New York, NY: IEEE, DOI 10.1109/IGARSS.2014.6946690, (to appear in print).

Blumstein, D., Chalon, G., Carlier, T., Buil, C., Hebert, P., Maciaszek, T., Ponce, G., Phulpin, T., Tournier, B., Simeoni, D., Astruc, P., Clauss, A., Kayal, G. and Jegou, R. (2004). IASI instrument: technical overview and measured performances, pp. 196–208 in Infrared Spaceborne Remote Sensing XII, Denver, CO. Bellingham, WA: International Society for Optics and Photonics, DOI 10.1117/12.560907.

Bormann, N., Collard, A.D. and Bauer, P. (2009). Estimates of spatial and inter-channel observation error characteristics for current sounder radiances for NWP. Technical Memorandum 600, Reading, UK: ECMWF.

Capelle, V., Chédin, A., Siméon, M., Tsamalis, C., Pierangelo, C., Pondrom, M., Crevoisier, C., Crepeau, L. and Scott, N. (2014) Evaluation of IASI-derived dust aerosol characteristics over the tropical belt. *Atmospheric Chemistry and Physics*, 14(17), 9343–9362.

Chalon, G., Cayla, F. and Diebel, D. (2001). IASI – an advanced sounder for operational meteorology, in Proceedings of 52nd Congress of IAF. Toulouse, France.

Chevallier, F., Di Michele, S. and McNally, A.P. (2006). Diverse profile datasets from the ECMWF 91-level short-range forecasts. NWP SAF Technical Report 10, Reading, UK: ECMWF.

Clerbaux, C., Boynard, A., Clarisse, L., George, M., Hadji-Lazaro, J., Herbin, H., Hurtmans, D., Pommier, M., Razavi, A., Turquety, S., Wespes, C. and Coheur, P.-F. (2009) Monitoring of atmospheric composition using the thermal infrared IASI/MetOp sounder. *Atmospheric Chemistry and Physics*, 9(16), 6041–6054.

Collard, A.D. (2007) Selection of IASI channels for use in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 133, 1977–1991.

Courtier, P., Freydier, C., Geleyn, J.-F., Rabier, F. and Rochas, M. (1991). The Arpege project at Météo-France, pp. 193–231 in Seminar on Numerical Methods in Atmospheric Models, 9–13 September 1991. Reading, UK: ECMWF.

Crevoisier, C., Clerbaux, C., Guidard, V., Phulpin, T., Armante, R., Barret, B., Camy-Peyret, C., Chaboureau, J.P., Coheur, P.F., Crépeau, L., Dufour, G., Labonnote, L., Lavanant, L., Hadji-Lazaro, J., Herbin, H., Jacquinet-Husson, N., Payan, S., Péquignot, E., Pierangelo, C., Sellitto, P. and Stubenrauch, C. (2014) Towards IASI–new generation (IASI-NG): impact of improved spectral resolution and radiometric noise on the retrieval of thermodynamic, chemistry and climate variables. *Atmospheric Measurement Techniques*, 7, 4367–4385. https://doi.org/10.5194/amt-7-4367-2014.

Desroziers, G., Berre, L., Chapnik, B. and Poli, P. (2005) Diagnosis of observation-, background- and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, 131, 3385–3396.

Fourrié, N. and Thépaut, J.-N. (2003) Evaluation of the AIRS near-real-time channel selection for application to numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 129, 2425–2439. https://doi.org/10.1256/qj.02.210.

Gambacorta, A. and Barnet, C.D. (2012) Methodology and information content of the NOAA NESDIS operational channel selection for the Cross-track Infrared Sounder (CrIS). *IEEE Transactions on Geoscience and Remote Sensing*, 51(6), 3207–3216.

George, M., Clerbaux, C., Hurtmans, D., Turquety, S., Coheur, P.-F., Pommier, M., Hadji-Lazaro, J., Edwards, D.P., Worden, H., Luo, M., Rinsland, C. and McMillan, W. (2009) Carbon monoxide distributions from the IASI/MetOp mission: evaluation with other space-borne remote sensors. *Atmospheric Chemistry and Physics*, 9(21), 8317–8330. https://doi.org/10.5194/acp-9-8317-2009.

Glumb, R.J., Williams, F.L., Funk, N., Chateauneuf, F., Roney, A. and Allard, R. (2003). Cross-track Infrared Sounder (CrIS) Development Status, in Proceedings vol. 5152, Infrared Spaceborne Remote Sensing XI; San Diego, CA. Bellingham, WA: International Society for Optics and Photonics, DOI 10.1117/12.508544, (to appear in print).

Hilton, F., Armante, R., August, T., Barnet, C.D., Bouchard, A., Camy-Peyret, C., Capelle, V., Clarisse, L., Clerbaux, C., Coheur, P.F., Collard, A.D., Crevoisier, C., Dufour, G., Edwards, D., Faijan, F., Fourrié, N., Gambacorta, A., Goldberg, M., Guidard, V., Hurtmans, D., Illingworth, S., Jacquinet-Husson, N., Kerzenmacher, T., Klaes, D., Lavanant, L., Masiello, G., Matricardi, M., McNally, A.P., Newman, S., Pavelin, E., Payan, S., Péquignot, E., Peyridieu, S., Phulpin, T., Remedios, J., Schl˝ ussel, P., Serio, C., Strow, L., Stubenrauch, C., Taylor, J., Tobin, D., Wolf, W. and Zhou, D. (2012) Hyperspectral earth observation from IASI: Five years of accomplishments. *Bulletin of the American Meteorological Society*, 93(3), 347–370.

Hocking, J., Rayer, P., Rundle, D., Saunders, R., Matricardi, M., Geer, A.J., Brunel, P. and Vidot, J. (2015). RTTOV v11 Users Guide, NWP SAF. Darmstadt, Germany: EUMETSAT.

Karagulian, F., Clarisse, L., Clerbaux, C., Prata, A.J., Hurtmans, D. and Coheur, P.F. (2010) Detection of volcanic SO_2, ash, and H_2SO_4 using the infrared atmospheric sounding interferometer (IASI). *Journal of Geophysical Research: Atmospheres*, 115(D2). https://doi.org/10.1029/2009JD012786.

Kaufman, L. and Rousseeuw, P.J. (2005) *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons.

Martinet, P., Lavanant, L., Fourrié, N., Rabier, F. and Gambacorta, A. (2014) Evaluation of a revised IASI channel selection for cloudy retrievals with a focus on the Mediterranean basin. *Quarterly Journal of the Royal Meteorological Society*, 140, 1563–1577.

Matricardi, M. and McNally, A.P. (2014) The direct assimilation of principal components of IASI spectra in the ECMWF 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 140, 573–582.

Noh, Y.C., Sohn, B.J., Kim, Y., Joo, S., Bell, W. and Saunders, R. (2017) A new Infrared Atmospheric Sounding Interferometer channel selection and assessment of its impact on Met Office NWP forecasts. *Advances in Atmospheric Sciences*, 34(11), 1265–1281.

Pereira, M.B. and Berre, L. (2006) The use of an ensemble approach to study the background error covariances in a global NWP model. *Monthly Weather Review*, 134(9), 2466–2489.

Rabier, F., Fourrié, N., Chafäi, D. and Prunet, P. (2002) Channel selection methods for Infrared Atmospheric Sounding Interferometer radiances. *Quarterly Journal of the Royal Meteorological Society*, 128, 1011–1027.

Rodgers, C.D. (1996). Information content and optimization of high-spectral-resolution measurements, pp. 136–148 in Volume 2830, Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research II. Bellingham, WA: International Society for Optics and Photonics.

Siméoni, D., Singer, C. and Chalon, G. (1997) Infrared Atmospheric Sounding Interferometer. *Acta Astronautica*, 40(2-8), 113–118.

Smith, F. (2016). 1D-Var User Manual. NWP SAF, Darmstadt, Germany: EUMETSAT. https://nwpsaf.eu/site/download/documentation/1dvar/nwpsaf-mo-ud-032_NWPSAF_1DVar_Manual.pdf.

Susskind, J., Barnet, C.D. and Blaisdell, J.M. (2003) Retrieval of atmospheric and surface parameters from AIRS/AMSU/HSB data in the presence of clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 41(2), 390–409.

Ventress, L. and Dudhia, A. (2014) Improving the selection of IASI channels for use in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 140, 2111–2118.

Weston, P.P., Bell, W. and Eyre, J.R. (2014) Accounting for correlated error in the assimilation of high-resolution sounder data. *Quarterly Journal of the Royal Meteorological Society*, 140, 2420–2429.

How to cite this article: Vittorioso, F., Guidard, V. & Fourrié, N. (2021) An Infrared Atmospheric Sounding Interferometer – New Generation (IASI-NG) channel selection for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 147(739), 3297–3317. Available from: https://doi.org/10.1002/qj.4129

APPENDIX A. CHANNELS SELECTED THROUGH THE DFS METHOD

Table A1 provides the 500 IASI-NG channel selection obtained through this study.

TABLE A1 List of the 500 channels most frequently selected with the DFS method

1	2	3	13	14	15	25	26	27
29	30	31	32	50	51	63	65	71
77	83	95	97	101	107	108	119	120
125	128	130	132	140	141	142	143	144
145	170	171	172	173	174	175	176	177
178	179	180	181	182	183	184	185	186
187	188	189	190	191	192	193	194	195
196	197	199	209	214	216	219	221	222
224	231	232	234	235	236	237	239	242
243	244	245	247	248	250	251	252	255
256	257	258	260	262	263	264	267	273
274	275	276	280	283	284	285	286	287
289	290	293	294	296	297	299	300	306
310	313	315	318	325	328	331	332	333
334	343	345	346	347	353	357	358	359
360	369	370	371	372	373	383	384	385
386	395	396	397	399	408	409	410	412
413	421	422	424	425	426	431	432	434

(Continues)

TABLE A1 Continued

435	438	439	444	445	447	448	450	451
452	457	458	459	460	464	465	470	472
473	477	483	484	485	490	491	497	503
504	510	517	522	523	528	531	535	548
549	561	580	582	583	584	592	593	594
595	596	597	598	599	600	605	606	607
653	672	686	691	692	697	698	705	710
712	713	714	726	727	739	740	742	746
747	748	749	751	752	753	760	761	763
764	765	766	769	770	771	772	773	774
775	776	777	796	797	801	802	803	808
814	815	838	857	876	888	900	912	965
1,012	1,054	1,057	1,114	1,117	1,119	1,120	1,174	1,175
1,176	1,211	1,224	1,225	1,226	1,230	1,233	1,267	1,307
1,382	1,463	1,560	1,660	1,661	1,663	1,666	1,939	2,443
2,445	2,446	2,447	6,851	6,859	6,860	6,877	6,886	6,891
6,892	6,895	6,897	6,900	6,901	6,904	6,914	6,915	6,917
6,918	6,919	6,920	6,926	6,933	6,939	6,940	6,945	6,950
6,951	6,953	6,964	6,983	6,991	6,993	6,994	6,997	7,002
7,005	7,006	7,011	7,012	7,017	7,018	7,019	7,028	7,035
7,053	7,054	7,064	7,122	7,125	7,149	7,150	7,153	7,154
7,158	7,159	7,170	7,172	7,177	7,178	7,194	7,197	7,198
7,232	7,240	7,243	7,248	7,258	7,276	7,281	7,282	7,301
7,302	7,305	7,306	7,308	7,309	7,314	7,315	7,320	7,321
7,323	7,332	7,375	7,425	7,440	7,449	7,450	7,451	7,452
7,453	7,508	7,518	7,611	7,625	7,698	7,699	7,720	7,773
7,775	7,884	7,923	7,936	7,942	7,943	7,948	8,009	8,015
8,060	8,062	8,063	8,064	8,065	8,066	8,067	8,073	8,074
8,077	8,080	8,121	8,195	8,213	8,245	8,320	8,407	8,408
8,443	8,445	8,446	8,450	8,580	8,591	8,617	8,646	8,665
8,674	8,680	8,708	8,717	8,852	8,861	8,880	8,882	8,905
8,907	8,921	8,969	8,979	8,980	8,985	8,987	8,993	9,087
9,102	9,112	9,120	9,121	9,129	9,131	9,132	9,134	9,158
9,198	9,222	9,257	9,272	9,274	9,275	9,296	9,298	9,306
9,308	9,336	9,344	9,348	9,351	9,373	9,375	9,385	9,388
9,389	9,390	9,392	9,395	9,398	9,405	9,409	9,460	9,531
9,560	9,561	9,660	9,678	9,679	9,680	9,682	9,684	9,698
9,816	9,821	9,861	9,886	10,292				

Note: This table reports the channel numbers, n. To compute the corresponding wave number ν, apply the formula: $\nu = 0.125(n - 1) + 645.000$.