PREPARING THE ASSIMILATION OF IASI - NEW GENERATION IN NWP MODELS: A FIRST CHANNEL SELECTION

Francesca Vittorioso, Vincent Guidard, Nadia Fourrié

CNRM UMR 3589, Météo-France & CNRS, GMAP/OBS, Toulouse, France

Abstract

As the EUMETSAT Polar System-Second Generation (EPS-SG) is being prepared, a new generation of the IASI instrument is being designed. The IASI - New Generation (IASI-NG) will measure at spectral resolution and a signal-to-noise ratio improved by a factor 2 compared to its predecessor. Measurement precision will be improved as well.

The high amount of data resulting from IASI-NG will present many challenges in the areas of data transmission, storage and assimilation and the number of individual pieces of information will be not exploitable in an operational Numerical Weather Predictions (NWP) context. For these reasons, an appropriate IASI-NG channel selection to be used in global and mesoscale NWP models is needed.

With this purpose, the standard iterative channel selection methodology, which is based on the optimal linear estimation theory, has been applied to a set of simulated data of the spectrum that IASI-NG will be able to characterize. However, the procedure has been adjusted so as to allow spectrally correlated errors to be properly evaluated.

This paper will illustrate the preliminary preparatory study, the analysis and selection processes and, eventually, the results thus obtained.

INTRODUCTION

In the framework of the preparation for the next European polar-orbiting program, i.e. EPS-SG, the New Generation of the hyperspectral Infrared Atmospheric Sounding Interferometer (IASI) has been designed. IASI-NG, which will be launched on board the Metop-SG series allegedly in 2022, will measure at 16921 wavelengths (or channels) in each sounding pixel, benefiting of a spectral resolution and a signal-to-noise ratio improved by a factor 2 compared to its predecessor [*Crevoisier et al.* (2014)]. Measurement precision will be improved as well starting from the 1 K in temperature and 10% in humidity IASI precision. This will lead to huge improvements in detection and retrieval of numerous chemical species and aerosols, and in thermodynamic profiles retrievals.

On the other hand, the high amount of data resulting from IASI-NG will present many challenges in the areas of data transmission, storage and assimilation. Moreover, the number of individual pieces of information will be not directly exploitable in an operational NWP context. For these reasons, an appropriate IASI-NG channel selection is needed, aiming to provide the most informative subset of data/channels to be used in global and mesoscale NWP models.

The work so far has been carried out on a simulated observation database, containing simulated data for IASI and IASI-NG [*Andrey-Andrés et al.* (2018)]. One-dimensional variational (1D-Var) assimilation experiments have been carried out to serve as a tool for analyzing and testing.

The need to take into account the error correlations among channels has required a study for the evaluation of full covariance matrices of the observation errors (**R** matrices).

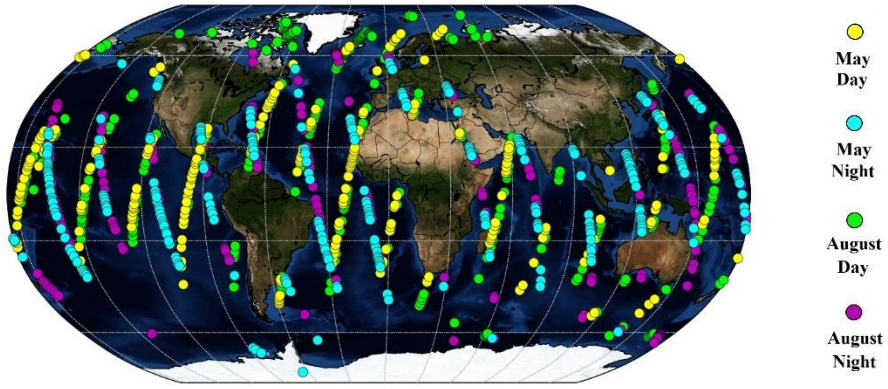

Figure 1: **Subset of selected observations for preparing the IASI-NG channel selection. The data are here represented in different colors related to the days of the year picked up and to the day/night illumination conditions.**

Finally, relying on the data thus obtained, a complete selection of channels was performed by assessing the contribution of each individual channel to improve the analysis errors over the background ones. The methodology employed and the results hence obtained will be shown below.

SIMULATED OBSERVATION DATABASE – CASE STUDY

The data that IASI-NG will provide will only be available in a few years. Therefore, in order to carry out an a-priori work of selection on its channels, a carefully simulated data set was necessary. With this purpose, during a previous project by *Andrey-Andrés et al.* (2018), a database of simulated observations has been build, containing simulated profiles for both IASI and IASI-NG. The database covers four dates in the middle of each season from 2013 (February the 4th, May the 6th, August the 6th and November the 4th) and it contains the full IASI orbit for each one of these dates. A total of 5 242 047 simulations is thus available for each instrument (IASI scan geometry is used for IASI-NG).

At this stage, the selection has been planned to be performed on *nadir – over sea – clear sky* conditions. With this in mind, we derived from the total initial database a subset of data to be used as a case study. More in details, 1099 of the 318 391 observations matching these criteria have been judged to be a representative sample of all the geographical regions. This smaller subset contains observations for:

- *May and August*
- *polar, mid-latitudes and tropical regions*
- *day and night*

located as in Figure 1.

PREPARATORY STUDY TOWARDS THE CHANNEL SELECTION

As an investigative tool in carrying out the work, one-dimensional variational (1D-Var) assimilation experiments have been employed. This method deals with minimizing the cost function:

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

where **x** is the model state vector (which in our case will contain Temperature, Humidity and Skin Temperature), $\mathbf{x}^{\mathrm{b}}$ is the background state vector, **y** is the vector of observations, **B** is the background-error covariance matrix, **R** the observation-error covariance matrix and *H* is the observation operator. In other words we look for the best compromise between background and observations, giving to each of them the proper weight.

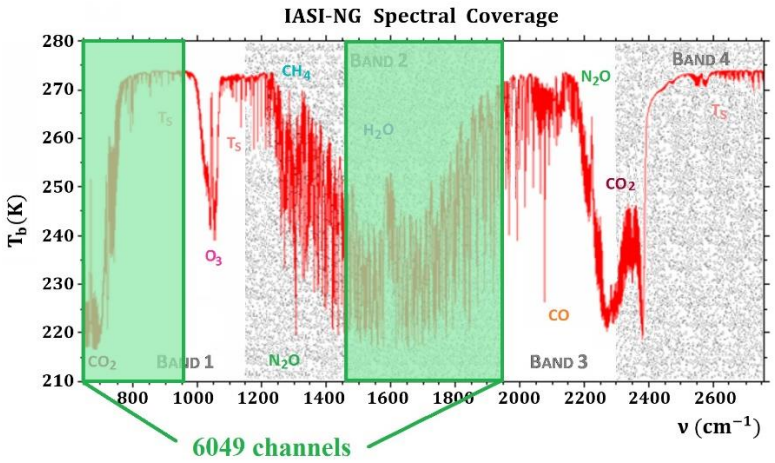

***Figure 2*: highlighted in green are Band 1 (2448 consecutive channels) and Band 2 (3601 consecutive channels) portions of the IASI-NG spectrum, which have been examined during this study.**

In this process it's very important to take into account the errors associated with the data and, in our case, since we work with simulated observations, it will be essential to build good observation errors which are represented in the **R** matrix term.

Using a diagonal **R** matrix that only contains information about the instrument noise error can be quite unrealistic, since it does not take into account all the other sources of observation errors, such as forward model, representativeness and pre-processing errors. This significantly contributes to correlations between different channel errors, which should be taken into account in this kind of study. For this reason, a diagnostic procedure introduced by *Desroziers et al.* (2005) has been used to estimate the structure of a full **R** matrix.

A code has been developed in order to implement the diagnostic in an iterative way. Through this tool, different regions of the spectrum that IASI-NG will be able to characterize have been explored. The attention has been mainly focused on the Band 1 and Band 2, which are the most relevant for the assimilation in the NWP context. Eventually, the diagnostic procedure has been realized on a total amount of 6049 channels, lying in the areas highlighted in Figure 2. In more detail, the first 2448 contiguous channels of Band 1 (645.000 up to 950.875 cm^{-1}) are pre-selected, leaving out the ozone-related part of the band. For what concerns Band 2, the last 3601 consecutive channels are taken, which are still able to provide important water vapour information, uncontaminated by trace gases (1500.000 up to 1950.000 cm^{-1} - channels from 6841 to 10441).

The results of this procedure carried out on the 6049 channels just mentioned are shown in Figure 3. On the left hand side, in red is the instrument noise and in blue the diagnosed error standard deviations from the fifth iteration in the diagnostic. Since the instrument noise is not the only contribution to the overall observation error (among the others the radiative transfer model used to simulate), the two curves are shifted. On the right hand side of Figure 3, the matrix of correlations is displayed. The bottom-left box represents the Band 1 related part, where we can observe highly correlated values (mainly corresponding to surface sensitive channels). In the top-right box, the Band 2 part, there is more variability in correlations caused by the same type of variability in the corresponding spectrum area.

METHOD

The methodology applied is the one suggested by *Rodgers* (1996) and proved to be a good a priori method for 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 time on a scalar figure of merit reflecting the improvement of the analysis error over the background one.

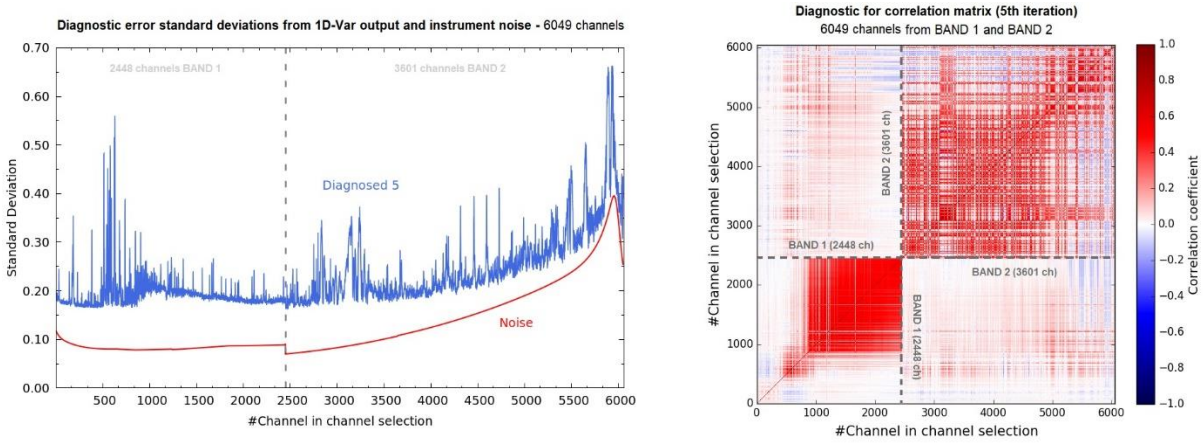

Figure 3: **Diagnosed error Standard Deviations from 1D-Var output and instrument noise (left); diagnostic for inter-channel observation error correlation matrix from 1D-Var output (right).**

The figure of merit we chose to iterate the procedure is the *Degrees of Freedom for Signal,* $\mathrm{DFS} = Tr\,(\mathbf{I} - \mathbf{AB}^{-1})$, where $\mathbf{A} = (\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}$ is the analysis error covariance matrix and **I** the identity matrix. As a result, through some mathematical manipulation, the DFS can also be expressed, and thus computed, as $\mathrm{DFS} = Tr(\mathbf{I}) - Tr[(\mathbf{I} + \mathbf{BH}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}]$.

The **B** matrix used in this study is the one provided in the NWP SAF package containing the 1D-Var software [*Smith and Havemann* (2016)], the **H** matrix contains the Jacobians of our case study profiles and the **R** matrix is the one just obtained through the Desrozier's diagnostic procedure.

More to the point, the DFS parameter used in the selection will be the one that, from now on, will be named *Total DFS.* It will be the sum of the DFS of Temperature, Humidity and Skin Temperature, namely our state variables.

The first step in selecting channels with this method consists in computing the Total DFS for every single channel among the N (6049 for us) pre-selected. The one among them providing the highest value of Total DFS will represent the first element in our selection.
We proceed by computing Total DFS for every couple of channels consisting of the first one just selected and all the remaining N-1 not yet taken. The second element in our selection is the one that, in couple with our first choice, provides the maximum value of Total DFS.
In a similar way, we will proceed for the successive 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 Total DFS.

However, selecting until all the N available channels have been taken can be very expensive in terms of time and resources. Therefore, in order to stop the procedure before all channels are exhausted, a reasonable criterion has to be devised. We decided, indeed, to narrow the choice down to the amount of channels showing an absolute difference of Total DFS greater than 0.001 between one iteration and the previous. This criterion leads to selecting a specific number of channels for each one among the different observations.

This procedure has been applied to the whole case study dataset. However, to get an idea of the evolution of the DFS during the selection, we show in Figure 4 its trend for one representative case among the 1099 taken into account. We can observe how, the DFS growth appears very fast in the very first part of the selection where the greatest contribution to the Total DFS is given by the Humidity term. On the other hand, the growth becomes slower and slower as the amount of selected channels increases and the predominant contribution becomes the one associated with the Temperature term. On the basis of our criterion, we select 998 channels for this observation.

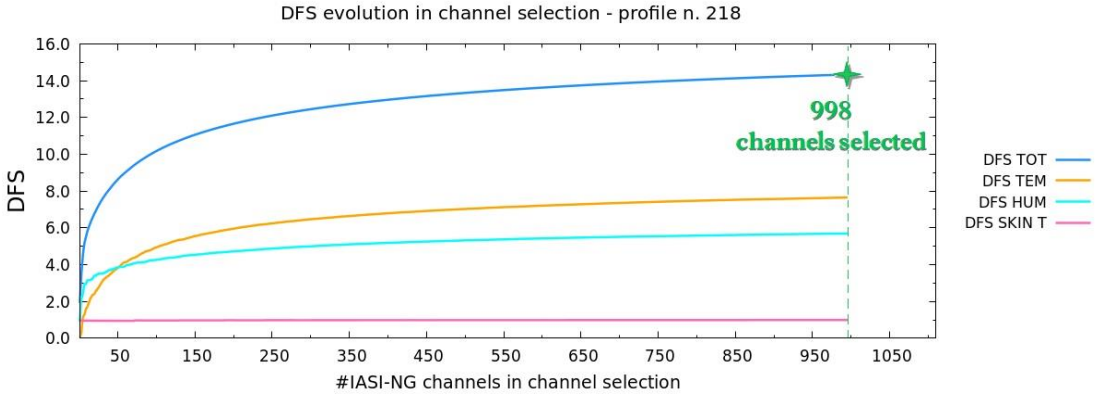

***Figure 4*: DFS growth trend in channel selection performed on profile number 218 (mid-latitude, nocturnal and summer observation). The star indicates the point at which the selection was stopped and the number of channels selected at that step.**

RESULTS

By extending the procedure illustrated in the previous section to the whole set of case study, we carried out our channel selection. As previous mentioned, according to our criterion for stopping the process, we have a different amount of channels that is picked up for each observation. In more than 80% of cases (i.e. profiles), we select an amount of channels that is between 1000 and 1150 channels. Besides, the average value of selected channels per profile is 1074, the median value 1079.

As regards the values of DFS reached at the end of the selection for each profile, few representative results are summarized in Table 1. Among the reported values, it is interesting to note a maximum of Total DFS of almost 17 and a minimum of 13. On the other hand, the average of all the DFS values reached in every single case is ~15.

DFS	Total	Temperature	Humidity	Skin Temp.
Average	15.2	7.9	6.3	0.99
Maximum	16.9	8.9	8.0	1.00
Minimum	13.0	7.2	4.1	0.96

***Table 1:* Averaged, Maximum and Minimum values of DFS reached through the channel selection.**

Another way to visualize the selection results is by examining the histogram in Figure 5, where the amount of channels selected per groups of profiles is expressed as a percentage (and as an actual amount of channels). In white we have the ***never selected*** and in blue the ***always selected*** channels (or selected in 100% of cases). We decided to consider as negligible all channels taken less than 5% of the time (3680 channels) and therefore to reject them. Consequently, our selection of channels consists of the 2369 remaining channels (among which 211 are always selected).

In order to be able to better understand our selection, it can be very interesting to evaluate the contribution of the two different bands to the overall results shown in Figure 5. In other words, we wondered: how many among the rejected, retained or always selected channels belong to Band 1 and how many to Band 2?

In Figure 6 we can have a look at the Band 1 contribution. We see that among the 2369 channels held in selection carried out on both bands simultaneously, 945 are those belonging to Band 1. Even rejecting 1503 channels out of the 2448 taken in Band 1, we preserve a good coverage of the different spectrum areas of interest (Figure 7), with a discrete group of channels in the surface sensitive area. From this study arises also that the always selected channels are predominantly high picking ones. This is because the background error profile shows more significant values in the upper troposphere.

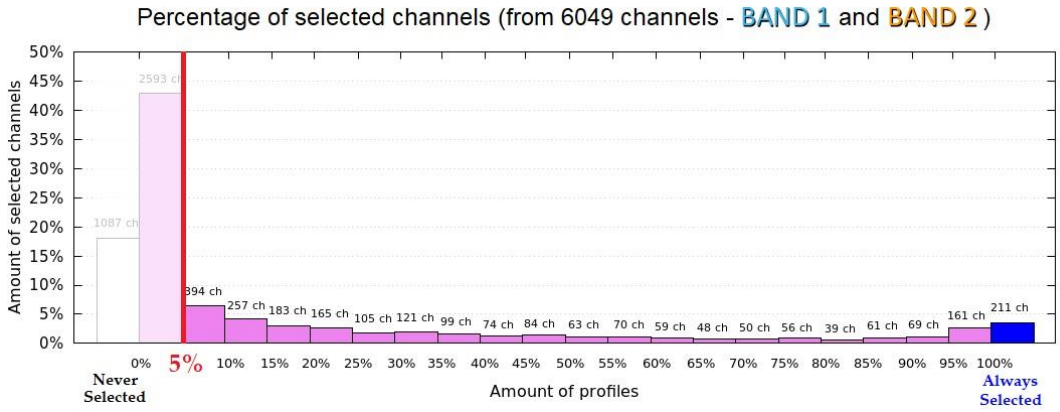

***Figure 5*: percentage of selected channels per amount of profiles. In white is the amount of never selected channels, in blue the channels selected in the 100% of cases and all the rest in between. The red line marks the threshold below which we consider the results obtained to be negligible and therefore to be rejected.**

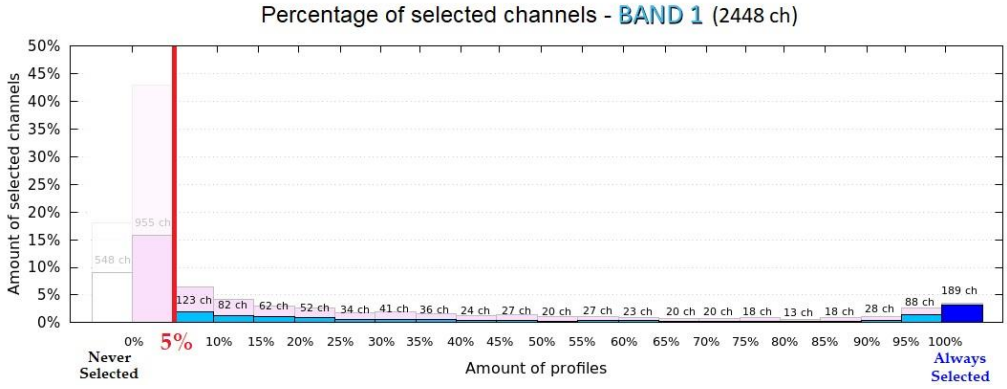

***Figure 6*: percentage of selected channels per amount of profiles – Band 1 contribution.**

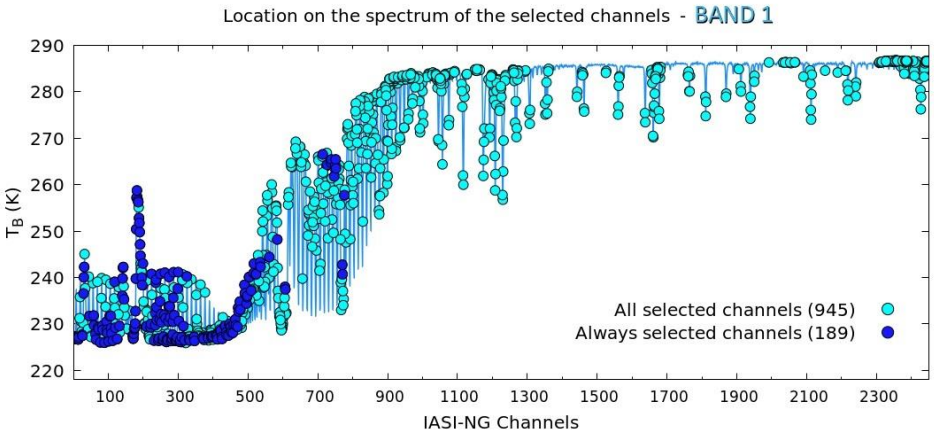

***Figure 7*: spectral location of the selected channels – Band 1.**

The same kind of data but for Band 2 can be found in Figure 8 and 9. In this part of the spectrum, we reject 2177 and we retain 1424 channels. As for Band 1, we have a good global and homogeneous coverage of the spectral area, even if the amount of channels always selected is quite smaller. Only 22 channels in the 211 are selected in Band 2.

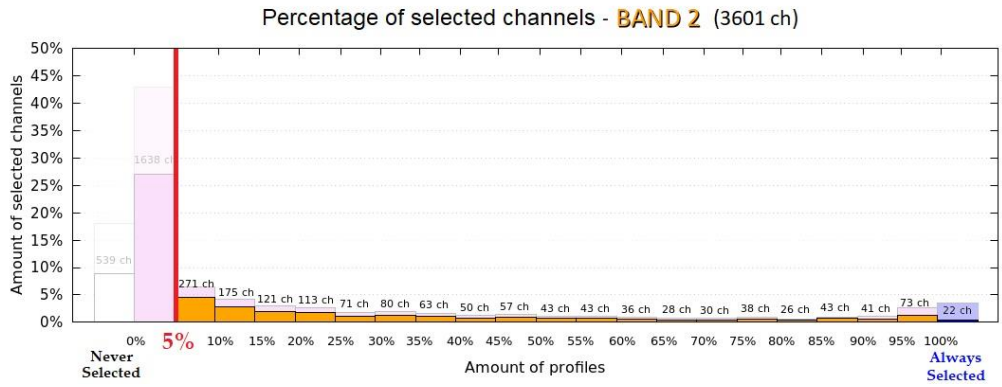

***Figure 8*: percentage of selected channels per amount of profiles – Band 2 contribution.**

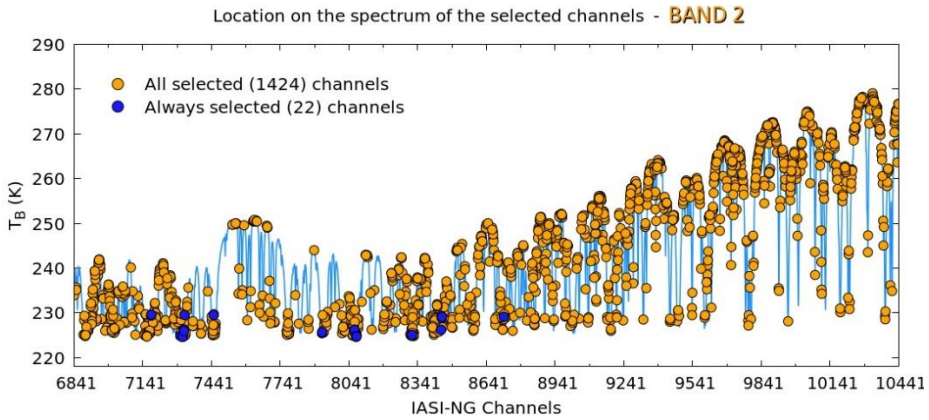

***Figure 9*: spectral location of the selected channels – Band 2.**

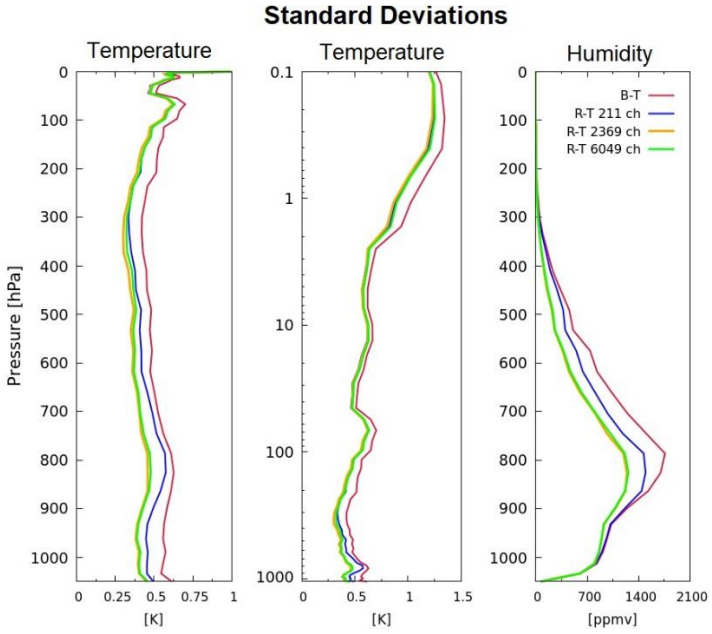

***Figure 10*: background (B-T) and analysis (R-T) error standard deviations averaged on the 1099 case study profiles, for Temperature (linear scale on the left and log scale in the middle) and Humidity (right). The blue, orange and green curves are associated to 1D-Var experiments using always selected, all selected and all channels respectively.**

Eventually a 1D-Var experiment has been run for the 1099 case study profiles, using:

- the 6049 channels considered in the beginning;
- “all channels” selected at least in 5 % of times (2369);
- “always selected” channels (211).

A statistical study on the output thus obtained has been performed. The background and analysis errors standard deviations, for Temperature and Humidity are shown in Figure 10. The red curve represents the *background error.* In green, orange and blue we have instead the *analysis errors* associated to the three aforementioned groups of channels. Looking at these results, we notice immediately that using only the always selected channels we produce an improvement in retrievals with respect to the background, even if small (blue curve). This improvement is about a half of what we achieve by using all the 6049 channels available (green curve). On the other hand, using the 2369 channels of our channel selection (orange curve) the reached results are very similar to those obtained with the totality of channels. Actually, if we look at the rate of improvement in this specific case we reach maximum values greater than 30% for Temperature and greater than 50% for Humidity.

FUTURE WORKS: IASI-NG CHANNEL SELECTION FOR THE NWP

The next step of this project will aim at refining the selection by applying another method. It will be based on picking up the most relevant channels relying on the characteristics of their weighting functions [*Gambacorta and Barnet* (2012)]. In our projects there is also to bring the selection in the context of the French global model ARPEGE, using all the simulated data at our disposal.

AKNOWLEDGEMENTS

This research has received funding from the Centre National d'Études Spatiales (CNES) in the framework of the IASI-NG project.

REFERENCES

Andrey-Andrés, J., Fourrié, N., Armante, R., Brunel, P., Crevoisier, C., Guidard, V., and Tournier, B., (2017) A simulated observation database to assess the impact of IASI-NG hyperspectral infrared sounder. Atmos. Meas. Tech. Discussion, **2017**, pp 1-29.

Crevoisier, C., and Coauthors, (2014) Towards IASI-New Generation (IASI-NG): impact of improved spectral resolution and radiometric noise on the retrieval of thermodynamic, chemistry and climate variables. Atmos. Meas. Tech., **7**, pp 4367-4385.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P., (2005) Diagnosis of observation, background and analysis-error statistics in observation space. Q. J. R. Meteorol. Soc., **131**, pp 3385-3396.

Gambacorta, A., and Barnet, C.D., (2012) Methodology and Information Content of the NOAA NESDIS Operational Channel Selection for Cross-Track Infrared Sounder (CrIS). IEEE Trans. Geosci. Remote Sens., **51**, pp 3207-3216.

Rabier, F., Fourrié, N., Chafai, D., and Prunet, P., (2002) Channel selection methods for Infrared Atmospheric Sounding Interferometer radiances. Q. J. R. Meteorol. Soc., **128**, pp 1011-1027.

Rodgers, C. D., (1996) Information content and optimisation of high spectral resolution measurements. Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research II, SPIE, **2830**, pp 136-147.

Smith, F., Havemann, S., (2016) NWPSAF 1D-Var User Manual Version 1.1.1, NWPSAF-UD-032.