Reviewer 1 :

We thank Reviewer 1 for his/her comments which helped to improved, we hope, the quality of the manuscript. Reviewer 1's comments are in bold font, our answers are written with normal font.

This manuscript looks at an interesting area and some new results are presented but without getting into much depth of analysis. There are gaps to be filled to support the conclusions. In particular it is hard to make an informed choice between the different inhomogeneity screening options based on this work, as only one of the viable options was tested in an NWP system (the other tested option already showed clear defects even before NWP system testing). There is also no independent validation of whether the screening achieves its goal, which is homogeneous scenes.

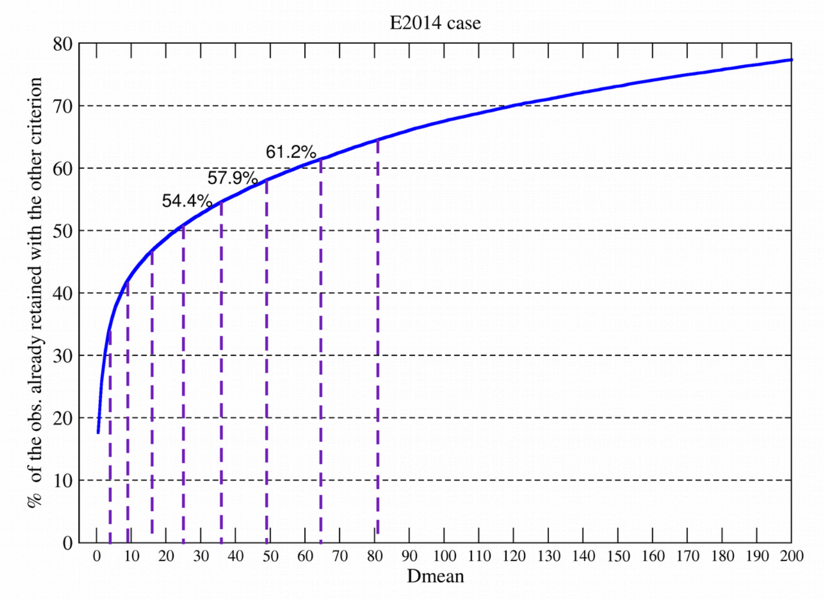
An independent validation was added. The cloud homogeneity was compared with an homogeneity criterion based on cloud types retrieved from SEVIRI observations. We found similar results as those obtained with the AVHRR cloud cover. In addition we have removed one of the data assimilation experiment in order to keep only the experiment with the COMPR criteria.

Major points

1) P6 L9: "we plan to assimilate clear or cloudy observations that are completely covered in the IASI FOV .... discarding fractional cloud observations". This still allows the possibility of fully clear obs being assimilated in a fully cloudy model (or vice-versa). Is that the intention? How do these screening methods treat cloudy modeled scenes if at all? The text should explain. In the context of the preparation of all-sky assimilation, we plan to assimilate indiscriminately clear or cloudy observations that are completely covered in a homogeneous way, discarding the cases of fractional cloud observations : the clear cases would be assimilated as in the current operational version and the cloudy ones in manner to be determined. The comparison to modeled scenes is ensured with the background departure check. In this case it will not be possible to keep a clear observation with a fully cloudy model.

2) The choice of 49Kˆ2 departure threshold in 3.2.2 is unsupported and uninvestigated in the text. Important questions are what this threshold means in terms of retained cloudy scenes, and how do its effects differ from those of the 7K AVHRR departure check in the Martinet et al. (2013) approach? Ultimately it should be investigated why the adapted (it is not the original) Eresmaa (2014) technique provides poor cloud screening here. Maybe it is the adaptation of this departure check? Finally, it is probably a case of poor wording rather than science, but it seems incorrect to claim the departure threshold as a check on homogeneity rather than just on cloud (P10 L19).

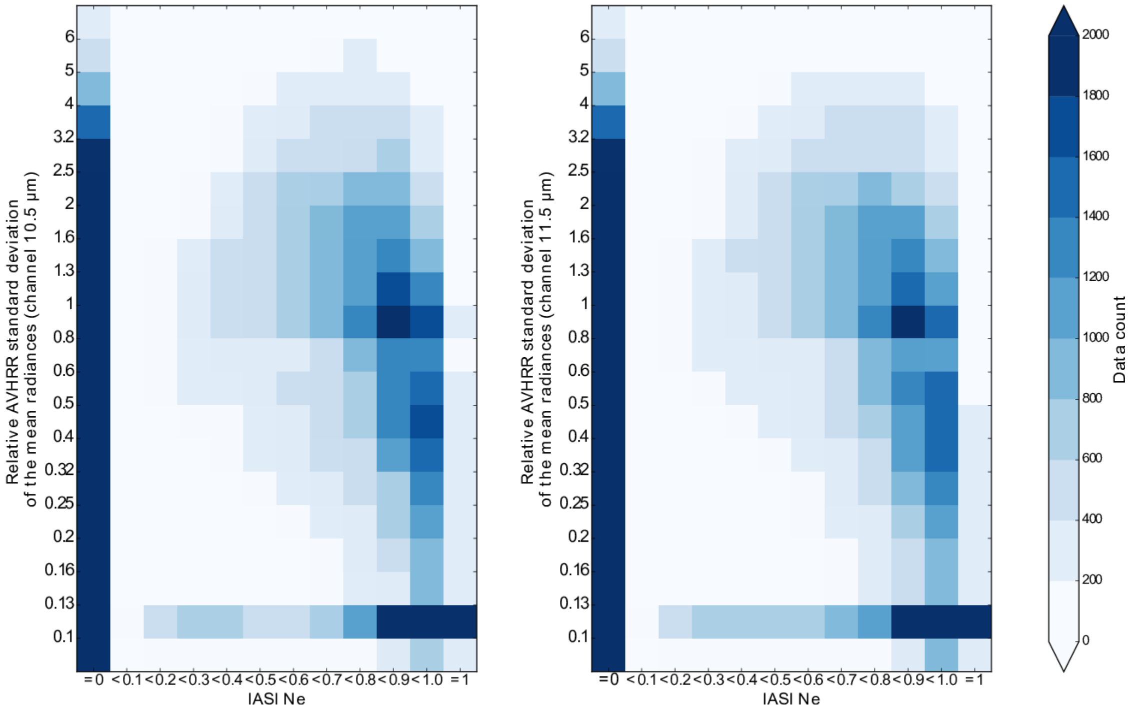
The choice of 49Kˆ2 departure was studied with the graph of leaving observations as a function of the departure. As shown below, this threshold allows to keep more than 50 % of the observations. That is why this threshold was kept. In addition it fits the 7K AVHRR departure check of M2013.
It is proposed in the text the following sentence « is less than 49K^2 . This particular value of threshold allows to keep more than 50 % of the observations compared to the initial threshold of 1K^2 by Eresmaa (2014) which retains only 10 % of the observations. In addition this threshold compares well with the one applied by M2013, but it is applied over the 2 IR AVHRR channels. »
The compromise selection method is an adaptation of the E2014 and is strongly based on it as we used the 2 IR AVHRR channels and the distance D_{mean} proposed by E2014 is used for the background check.



We agree that the D_{mean} based check cannot be considered as a homogeneity check and the sentence (initially P10 l19) has changed : « In this test, we used the D_{mean} proposed by Eresmaa (2014) to perform a kind of cloudiness consistency check between the observation and the model simulation »

3) The choice of the 0.8% threshold on p11 is barely supported in the text or by Figure 2. It may be the colour scale but there seem to be no highly inhomogeneous scenes according to the IASI Ne (e.g. between 0.3 and 0.7 on Fig. 2) and there certainly seems no correlation between the relative cluster standard deviation and the Ne.

We recognise that the original Figure 2 was not clear with the interpolation. Figure 2 has been plotted with another color scale, a logarithm scale on y axis and as a function of data count.



As can be seen in the Figure, the threshold of 0.8% allows to remove 39,8% and 42,0% of observations for each AVHRR channel.

4) P11 L17 "Similarly in model space the D_mean..." Since D_mean is based on observation minus background it is not in model space and neither is it a direct indication of the model cloudiness.
We agree that D_mean is computed in the observation space. We propose the following rewording : « Similarly, we used the background departure check in the observation space D_mean (presented in section 3.2.2).

5) "The percentage of cloudy AVHRR pixels" - if this is a good enough indicator of fractional cloud and/or inhomogeneity to use it to validate the screening criteria, why is it not used as part of the screening criteria? We agree that this percentage of cloudy AVHRR pixels is not so good indicator of the presence of homogeneous clouds to assess the screening criteria. As said above, we propose here an independent evaluation against SEVIRI cloud type. The results obtained with these data are similar to those obtained with the AVHRR cloud cover.

6) Section 5, the intercomparison of selection criteria, does not fully make the case for the proposed selection method. M2013 keeps 29% of data in table 2 compared to 21% in the compromise method, with only slightly higher standard deviations. That could be a good choice, but it has been rejected at this stage. The balance between a slight increase in std. dev. and gaining extra data has hence not been properly explored. It seems odd to instead test the "Obs_HOM" approach in data assimilation as already from the intecomparison it is clear it does not work well. Further, without an exploration of its sensitivity to threshold choices, the E2014 test does not have much of a chance in this intercomparison. Finally, each of the previous techniques M2013 and E2014, as well as the newly proposed compromise technique are composed of two tests, and it would be good to know how many rejections each is responsible for and how much overlap there is between the two tests.
We agree with the reviewer that the issue here is to find a compromise between good statistics and a sufficient number of observations for the assimilation. Here the Obs_HOM approach clearly keeps too much observations for the assimilation. In section 5, this experiment with the Obs_Hom Criteria was removed and we keep only the COMPR criteria for the data assimilation experiment. The COMPR method results from the E2014 as the D_mean criterion was used.

Please find below a table summarizing the impact of each criteria for each method.

	% of remaining observations	When 2 criteria applied	When all criteria applied
M2013			
Sigma inter 1	89,8%	84,8%	53,9%
Sigma intra 1	87,8%		
Background check 1	59,4%		
E2014			
Crit 1 can 1	39,8%	39,4%	22,8%
Crit1 can 2	42,9%		
Dmean<49K	42,7%		
COMPR			
obs_hom 1	68,2%	67,3%	36%
Obs hom 2	69,6%		
Dmean<49K^2	42,7%		
OBS_Hom 1+Dmean 1	43,3%	36%	
obs_Hom 2+Dmean 2	44,7%		

As can be seen in the previous table, if tests based on a single channel allow to keep more or less the same percentage of observations, the combination of both channels lead to a more restrictive choice.

7) Much of the conclusions will need to be updated to reflect a more thorough comparison of the different methods but particularly problematic is the statement P19 L4-6, saying that the M2013 and E2014 techniques were unsatisfactory due to a large loss of observations. Since M2013 provides more observations than the proposed method according to table 2, this statement is wrong. Also E2014 was not well explored in terms of thresholds and other possible adaptations to make it work in the current framework, using all-sky forward radiative transfer. Any rejection of this technique needs to be carefully qualified.

We have removed the sentence. The conclusions of this study were updated.

A new method using of collocated AVHRR cluster information to improve the selection of homogeneous IASI observation scenes within the numerical weather prediction ARPEGE model has been developed at Météo-France for data assimilation purposes and has been presented in this study.

The first step consisted in adapting the IASI observation operator based on the RTTOV radiative transfer model by using the RTTOV-CLD module with cloudy microphysical parameters (liquid water content (ql), ice content (qi) and cloud fraction) for the simulation of cloudy radiances. A qualitative evaluation of such module showed realistic simulated cloud structures at various locations around the globe with a quite good agreement against IASI observations.

The second and main step of this work was to assess the impact of several methods used to select homogeneous IASI observations using AVHRR clusters. Two selection methods (derived from the literature : Martinet et al. (2013) and Eresmaa (2014))) were preliminarily evaluated. Despite a good improvement in terms of biases and standard deviations of the FG departures, it was found that these two methods were not satisfactory in an operational context (in assimilation) to a large IASI observation reduction. the criteria from the Martinet et al. (2013)'s method favours the homogeneous cloudy observations and retains more than a half of the observations while the Eresmaa (2014)'s method gives priority to clear observations and keeps only 22% of the observations. Then, two new sets of criteria were defined from these two previous methods in order to have a more balanced choice of clear and cloudy observations and good statistics in terms of bakcground departures and implemented within the ARPEGE model :

– The first criterion derived from Martinet et al. (2013) method looks for the consistency between different clusters occupying the same IASI FOV by examining this homogeneity relative to the weighted average brightness temperature of the AVHRR clusters; it is only based on observations and computed for both infrared AVHRR channels as in Eresmaa (2014). This criterion allows to retain 67% of observations.

– In addition, the second criterion is derived from Eresmaa (2014)'s test and assesses the coherence of each cluster compared to the background brightness temperature simulation ; it is in fact a good compromise between the previous criterion and the two "historical" ones with accurate statistics and a sufficient number of observations 36% that passed the check. It also allows to retain the same proportion of homogeneous clear and cloudy observations contrary to the derived Martinet et al. (2013) and Eresmaa (2014) methods.

Therefore, assimilation experiments were conducted to assess the impact of these new selecting homogeneous IASI observation features in the current clear sky assimilation. This revised check was added to the McNally and Watts (2003) cloud detection. The results obtained in this case show that the scenes categorization has been facilitated and cloudy observations can be better filtered out compared to what is done in the operational ARPEGE version. 3% of all observations are rejected with the compromise method for the assimilation. The impacts on the first guess and analysis departures (showing more Gaussian shape) are generally low but with a beneficial reduction on the standard deviation of first guess departures mainly on the IASI and AMSU-A observations. Regarding the forecasts scores, neutral impact is reported when these selection criteria are taken into account on top of the McNally and Watts (2003) algorithm.

Minor points

1) Introduction: an up-to-date reference on homogeneity critetria for all-sky is the following. It should be discussed: Okamoto, K. (2017), Evaluation of IR radiance simulation for all-sky assimilation of Himawari-8/AHI in a mesoscale NWP system. Q.J.R. Meteorol. Soc., 143: 1517-1527. doi:10.1002/qj.3022 A reference to the work by Okamoto (2017) was added in the text, after the section on the selection of cloudy scenes based on cloud homogeneity. « Okamoto (2017) studied the impact of the super-observation homogeneity quality control on the Advanced Himawari Imager brightness temperature simulation. He concluded that for larger size of super-observations, the standard deviation threshold should be relaxed in order to keep sufficiently low brightness temperatures associated with high-level cloud because of the presence of more cloud heterogeneity in large size observations»

2) P3 L31 "the first level at 10m" sounds odd; surely the lowest 10m of the atmosphere is also included in the model? The first level of the ARPEGE is well at 10m. Fields below this level are interpolated from the model and the surface.

3) P5 L28 "Stratus Continental and Stratus Maritime" are cloud microphysical options in RTTOV; this should be stated; also it should be made clear how the choice is made, even if it is as obvious as using the land-sea mask. The cloud type chosen depends on the land-sea mask of the model. Over land, the cloud type stratus continental is chosen, the stratus maritime is used over sea. The sentence was reworded as follows : « To simulate the radiances observed in cloudy conditions using RTTOV-CLD, we use two main cloud types : firstly liquid water cloud which corresponds to two RTTOV-CLD cloud microphysical options depending on the land sea mask of the model (Stratus continental over land and Stratus Maritime over the sea), secondly the ice water cloud of the Cirrus type, using Baran parameterisation (Baran et al 2014 and Vidot et al 2015) to define the optical properties

4) P10 L1 "Background brightness temperature for AVHRR" - for clarity explain if this is clear-sky or all-sky. Here this background brightness temperature is simulated with clear sky as in Eresmaa (2014). This was clarified in the text « where R_i^{BG} is the clear-sky background brightness temperature for AVHRR channel i ».

5) P10 L7 If f^j is different from C_j please explain how; otherwise use consistent notation. f^j and C_j are the same quantity, the fractional coverage of the cluster. f_j was changed into C_j.

6) P11 L10 Instead of "relationship" is "ratio" intended? At present the text is imprecise. Relation ship was changed into ratio.

7) P12 L15 "Mean standard deviation" is always a confusing phrase and needs explanation of what samples got meant or standard deviated and in what order. The paragraph was rewritten and there is no more mean standard deviations.

8) Table 2 should additionally include statistics for the observations that are fractionally cloudy according to AVHRR. This column was added in Table 2.

9) P12 L15 to P14 L10 is hard to read as it is overloaded with numerical results and bereft of much interpretation. Many of the numerical results are already listed in table 2 and do not need exhaustive restating in the text, which needs to be rewritten with a higher level of analysis for the reader. Where numerical results given in this text are not already in tables, they should be.

This part of the paper was rewritten and numbers scattered along the text were removed. “The percentage of cloudy AVHRR pixels in the IASI field was also used to assess the choice of homogeneity criteria (Table 3).

Our global data set is made of 50% of the observations entirely covered by clouds and 12% of clear observations according to the AVHRR cloud cover. These results obtained over the globe set agree well with the ones obtained with SEVIRI data over the Atlantic Ocean. Except for M2013 method, the percentage of selected observations for each selection method is larger (+2/4 %) with the SEVIRI data evaluation than with the AVHRR cloud cover. The bias and standard deviation of observations minus simulations (O-G), are shown in Figure 3.(a) for the 314 IASI channels. As expected, the best statistics are obtained for channels less affected by clouds (e. g. CO2 and water vapour high peaking channels).

For the whole dataset, window channels present a bias of around -0.6K. The standard deviations are larger (around 12 K) for window channels sensitive to the surface. With the M2013 selection method (figure 3. (b)), the standard deviation of window channels is reduced to around 4 K as well as the bias close to zero. The standard deviation of the other channels (680-780cm^{-1} is also well decreased. The E2014 selection method (figure 3. (c)) improves the bias and the standard deviation (2.0 K for window channels) for all the channels. As expected, the impact is larger for surface sensitive (and thus cloud sensitive) channels than for the tropospheric channels (680-780cm^{-1}). On the contrary, with the Obs_HOM method (figure 3. (d)), small statistics improvement is obtained for the standard deviation and the bias. The statistics obtained with the COMPR method (Fig. 3.(e)) are reduced compared to the whole data set and slightly less good than with the initial E2014 method (for window channels the standard deviation is around 2.2 K instead of 2 K for E2014).”

10) P14 L22 "Gaussian" is overly strong here and is not backed up by any statistical tests of Gaussianity. We propose the following modification « the O-G distributions become symmetrical and get closer to the gaussian distribution »

11) P14 L30 "we keep more observations" - which method is referred to by "we"? This corresponds to E2014 method and it was changed in the text.