THE PIECEWISE LINEAR REGRESSION RETRIEVAL OF TEMPERATURE, HUMIDITY AND OZONE WITHIN THE EUMETSAT IASI L2 PPF VERSION 6

Tim Hultberg, Thomas August

EUMETSAT, Eumetsat Allè 1, 64295 Darmstadt, Germany

Abstract

PWLR (piecewise linear regression) is a fast statistical all sky retrieval scheme, which has been developed for version 6 of EUMETSAT's operational IASI L2 processor. The retrieval uses IASI as well as co-located AMSU and MHS radiances as inputs and serves two purposes within the IASI L2 processor: 1) to provide a high quality all sky temperature and water vapour profile product and 2) to provide a-priori information used in the clear sky optimal estimation retrievals. The retrieval is based on linear regression, but in order to capture non-linear relationships between the inputs and the outputs better, the input space is divided into several intervals. For each of these intervals a separate set of linear regression coefficients is computed from the corresponding subset of the training set, such that overall a piecewise linear function from the input space into the output space is obtained. The piecewise linear regression retrieval errors are substantially smaller than the corresponding retrieval errors from linear regression. In the current configuration a total of 36 different input intervals (retrieval classes) have been used. The training set consists of 953242 IASI/AMSU/MHS measurements paired with co-located ECMWF analysis profiles. In this paper we first describe this training set in further detail. This is followed by a discussion of the definition of the retrieval classes (i.e. how the training set is divided into 36 disjunctive subsets); the introduction of the quality indicators associated with the retrievals; a presentation of the retrieval error statistics obtained from a comparison with ECMWF analysis in both profile and radiance space and finally conclusions and possibilities for improvement.

THE TRAINING SET

The training set has been constructed from 12 days of real measurements (as opposed to forward model simulated radiances) from three Metop-A instruments: IASI, AMSU and MHS. To include seasonal variations, the twelve days were chosen one month apart over a one year period: 1st November 2011, 1st December 2011, 1st January 2012 and so on until 1st October 2012. As estimate of the atmospheric “truth” corresponding to the measurements, time and space co-located ECMWF analysis fields have been used. Clearly this “truth” has its own associated errors originating both from the co-location and the analysis itself, but as long as the training set is sufficiently big, random (unbiased) errors do not affect the regression negatively. On the other hand any systematic biases found in the co-located ECWMF analysis fields are retained in the regression retrievals. This problem is likely to affect mainly the ozone profiles, the highest stratospheric atmospheric temperatures as well as the surface skin temperature in certain continental areas (mostly desert) at daytime.

The measurements were co-located in time and space with ECMWF analysis fields at 6 hourly time resolution, 0.5 degree longitude/latitude grid and 91 vertical model levels for the profiles. The set of predictors is composed of the surface altitude, the secant of the satellite zenith angle, AMSU radiances in 14 channels (channel 7 was excluded due to its unavailability on Metop-A), MHS radiances in all 5 channels as well as the IASI radiance spectra represented by the 30 leading principal component scores in each of the three IASI bands obtained with the eigenvector files used for EUMETSATs operational IASI Principal Component (PC) Compression product. In a first step, so called IASI_PRP files were generated for all orbits from the selected 12 days. These IASI_PRP files, which are available from the authors on request, are in HDF5 format and contain the IASI L1C measurements represented as PC scores together with collocated data extracted from AMSU and MHS 1B products, ECMWF grib files (analysis in this case) and global databases of land/sea coverage and surface topography. In a second step the training set was extracted from the PRP files by considering only every fourth IASI scan-line. From each of these scan lines 30 instantaneous fields of view (IFOV), the warmest of the four IFOV's within each effective field of view (EFOV), were used to form the training set consisting of a total of 953242 cases.

The retrieved parameters consist of 4 surface parameters, surface pressure, surface air temperature, surface air dew point temperature and surface skin temperature as well as atmospheric profiles of temperature, humidity and ozone. For the regression step we have chosen to keep the profiles at the 91 (surface pressure dependent) model levels found in the ECMWF analysis files. The water vapour mixing ratios at each profile level, w, have been converted to dew point temperature (using the formula $\frac{243.5\log(0.263P_a w)}{17.67-\log(0.263P_a w)}$, where P_a is the level pressure) as we expect to find a better linear relationship with the predictors using this unit. Also the ozone profile mixing ratios have been converted to "dew point temperature" using the same formula as for water vapour. The retrieved surface pressure is used to assign pressures to the model levels such that the profiles can be interpolated to a fixed pressure grid after the regression retrieval on model levels.

DEFINITION OF RETRIEVAL CLASSES

The purpose of having different retrieval classes is to divide the training set into several parts in such a way that the type of atmosphere and surface within each class is relatively homogeneous and a linear function can provide a good approximation of the relationship between the inputs and the outputs. Of course the class definition must be based on the value of the inputs only and it can be done in many different ways. We have used a hierarchical approach. In the first step the input space is divided into three parts based on the surface height, z, in kilometres and the AMSU radiance in channel 2 and 4, a$_2$ and a$_4$, in mW/m^2/sr/cm^{-1}. These three parts are A: high elevation land ($z > 1$), B: open sea ($z = 0$ and $a_4 > 370 + 1.5a_2$) and C: low elevation land or sea ice (if not (A or B)). We note that some fields of view over sea with high clouds are classified as C, which is not a problem since the purpose of the classification is not to make a distinction between land and sea, but rather to provide a subdivision of the training set. Each of the three top level subsets are further subdivided in three based on a_4, the AMSU radiance in channel 4. For class B and C this subdivision is obtained using the thresholds 635 and 665 and for class A, 600 and 665 are used as thresholds. The nine subsets obtained so far are then further divided into four, first by applying a threshold on the radiance in MHS channel 2 and then within each of the 18 resulting subsets applying a threshold on the first PC score in Band 2 to arrive at the final 36 classes. The values of these two thresholds are different within each of the relevant subsets and are chosen such that the number of cases within each of the four low level subsets derived from the same high level subset is approximately the same. Figure 1 is based on the subset of the training set consisting of class B (open sea) with $a_4 > 665$and illustrates the regression retrieval improvement which is obtained by further subdivision as described above. The dotted black lines show the overall standard deviation of the regression fit using a single set of regression coefficients. The four coloured dotted lines show the standard deviation obtained by the same regression but computed individually for each of the four subsets. If a different set of regression coefficients is computed and applied for each of the four subsets, standard deviations as illustrated by the solid coloured lines are obtained, which results in an overall standard deviation as shown by the solid black line.

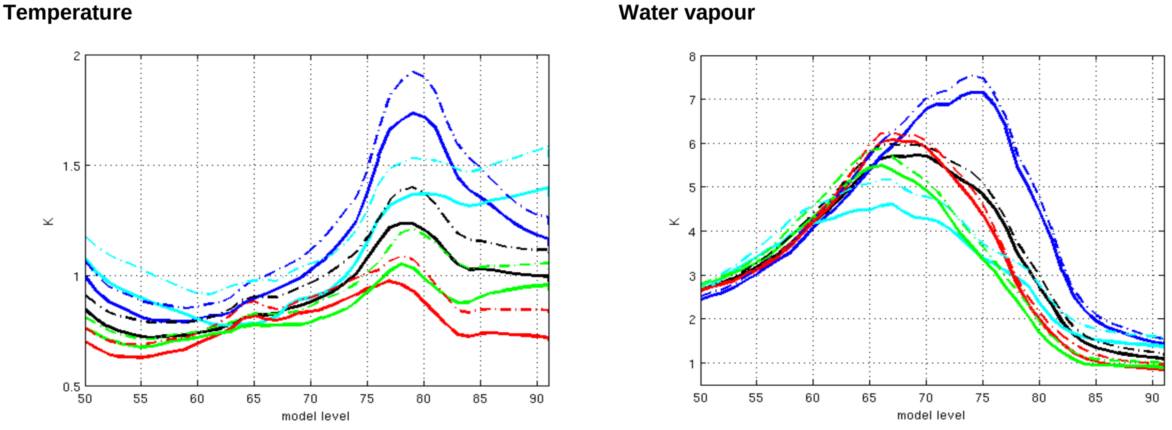

Temperature **Water vapour**

Figure 1 **Standard deviation of training set regression fit before (dotted) and after (thick) subdivision.**

To check that the improvements are not a result of over-fitting, an experiment with a random subdivision in four was performed, which as expected did not lead to any improvements.

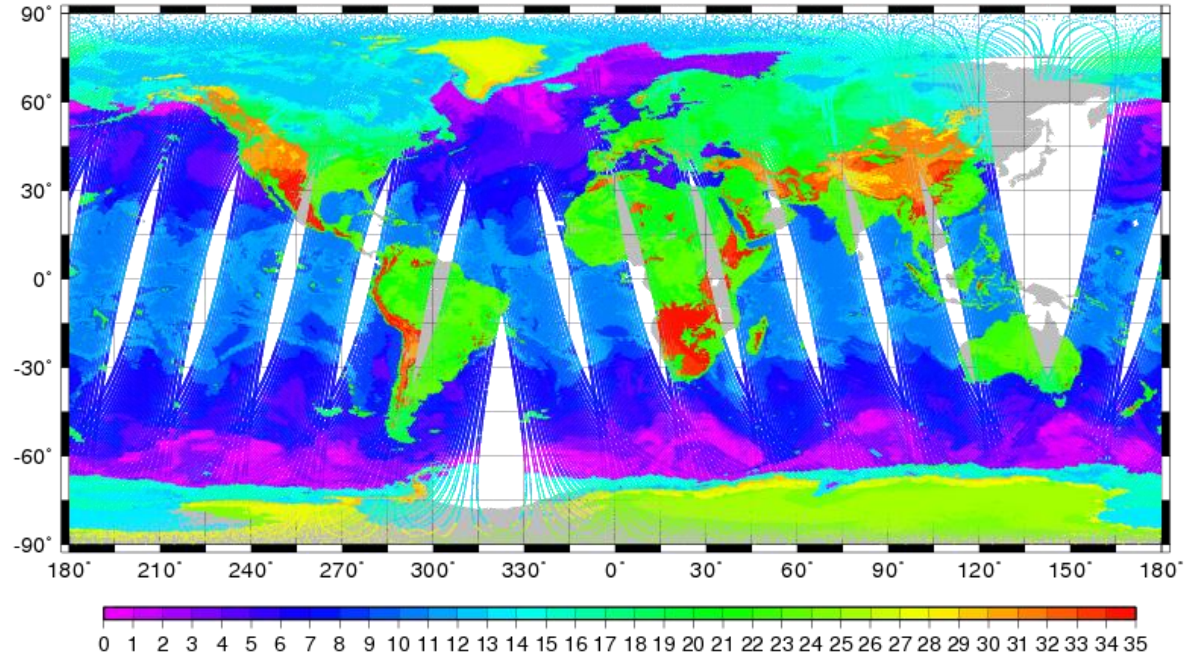


Figure 2 **MWIR PWLR retrieval classes 20130417PM.**

A version of the PWLR scheme for IASI only was also developed to serve as a back-up in case AMSU or MHS data are not available. For this version 32 retrieval classes were defined based on a successive dichotomy based on the following five predictors: PC1 Band1, PC1 Band2, PC2 Band1, PC2 Band2 and PC3 Band1.

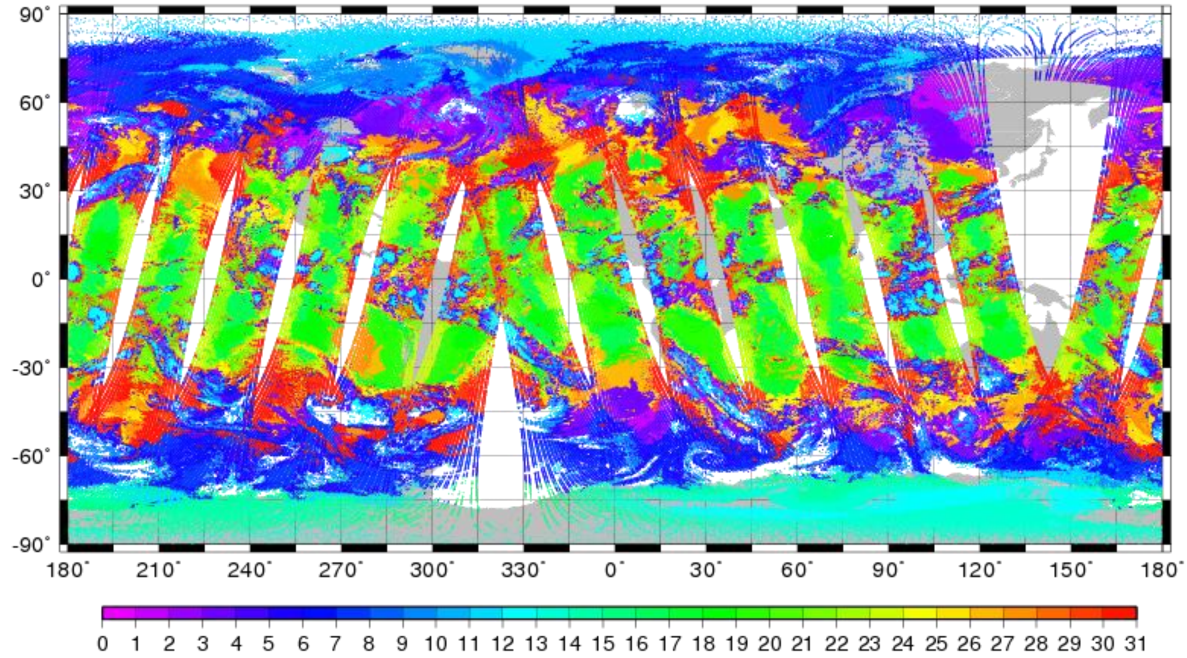


Figure 3 **IR PWLR retrieval classes 20130417PM.**

THE QUALITY INDICATORS

For each retrieved parameter a simple but reliable quality indicator can be obtained by regression against the absolute value of the retrieval error. This regression retrieval of the absolute value of the primary retrieval errors uses the same predictors as the primary retrieval. An individual quality indicator for each of the 277 retrieved values is possible, but in order to keep the product small we have settled on 7 quality indicators: one for each of the 4 surface parameters and one for each of the 3 profiles. For the profiles an average of the absolute errors at a subset of the individual model levels was used as the basis for regression. For T and W, model levels 32 to 91 were used in the average whereas for O, model levels 2 to 62 were used.

The quality indicator for a parameter is an estimate of the absolute retrieval error as compared against ECWMF analysis. The random error of this estimate is quite high but averaged over large datasets the error estimate is very accurate. The two following plots show the geographical distribution of the surface air temperature quality indicator for both the MWIR and IR versions of the PWLR. In both cases the negative impact of certain cloud patterns on the retrieval quality can be observed. We also note the better retrieval quality in the MWIR case in general and especially for cloudy situations.

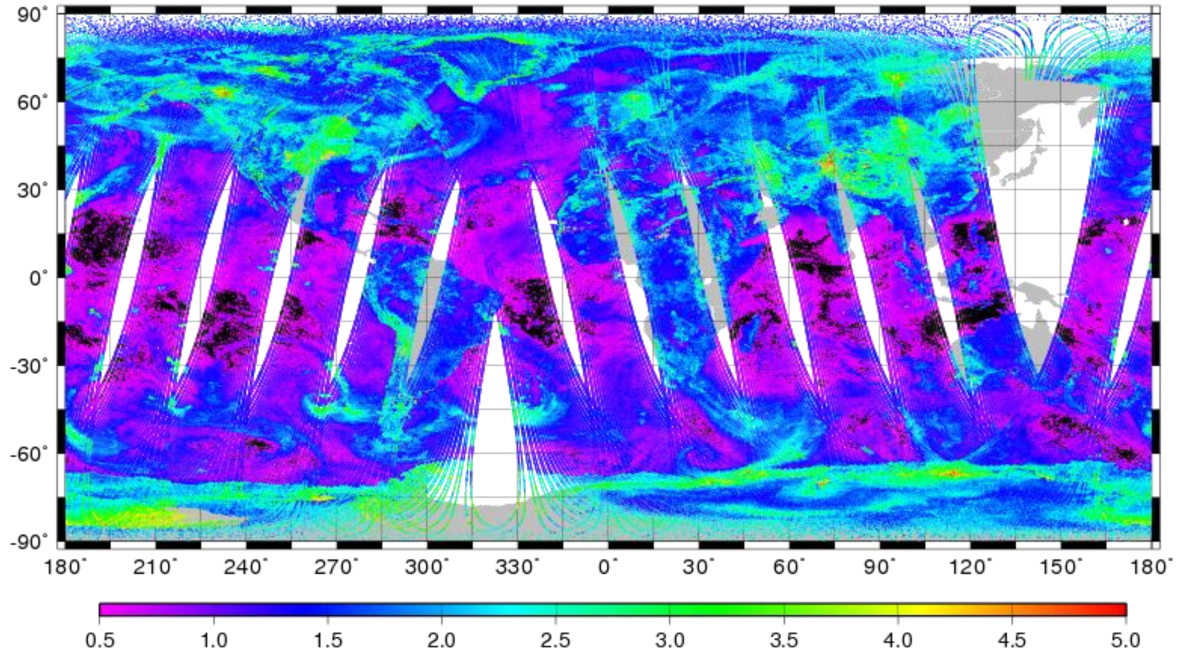


Figure 4 **MWIR PWLR quality indicator for Ta, 20130417PM.**

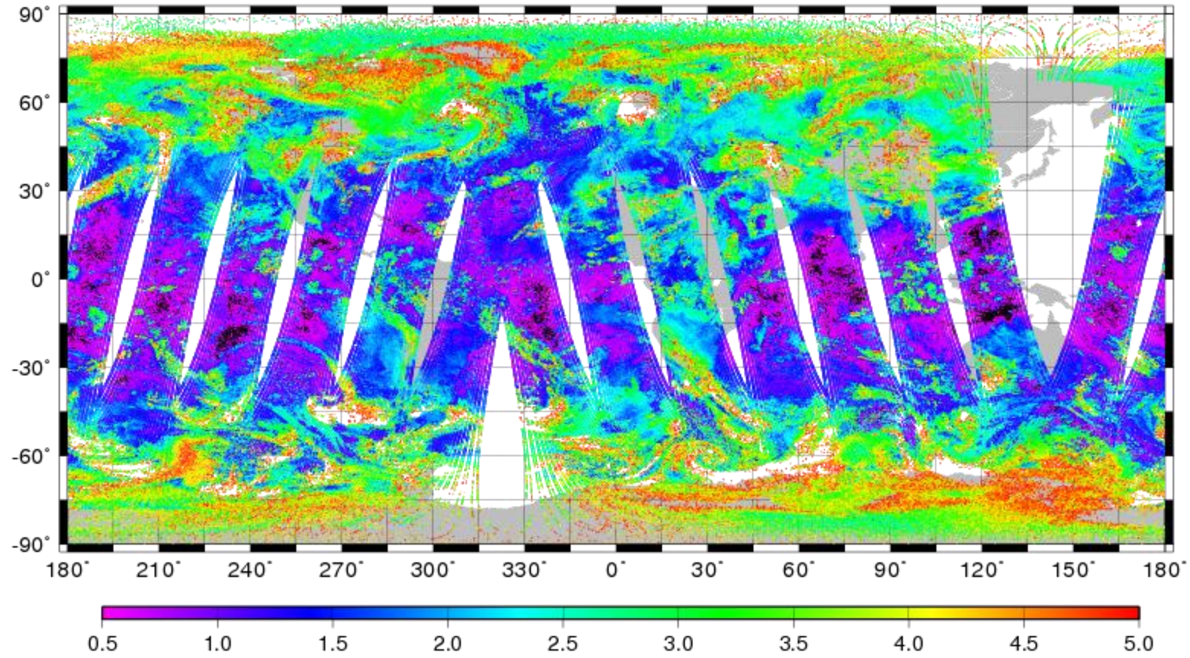


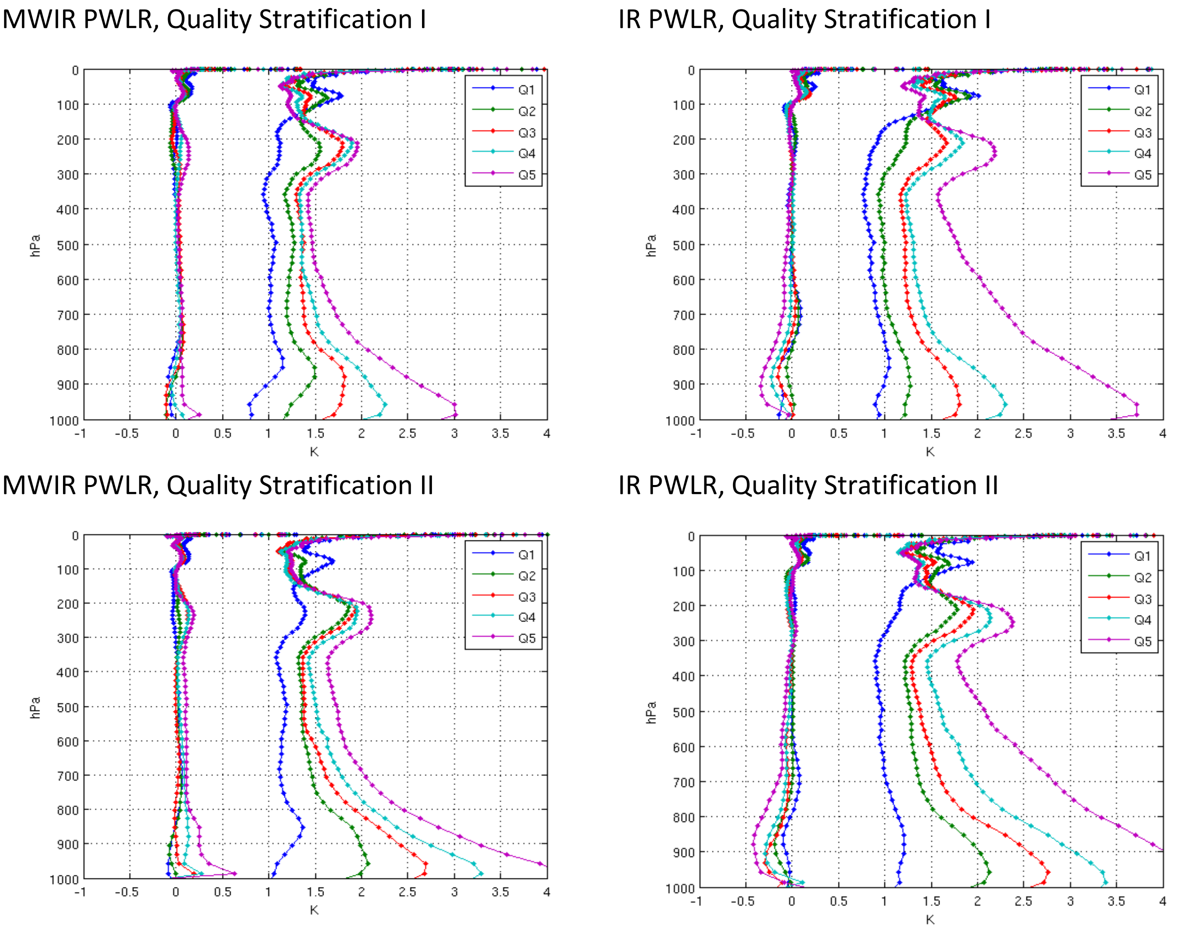
Figure 5 **IR PWLR Quality indicator for Ta, 20130417PM.**

RETRIEVAL RESULTS AND ERROR STATISTICS

In this section we look at the retrieval performance as characterised by the mean and standard deviation of the difference between the retrieval results and collocated ECMWF analysis. The first set of statistics was obtained from the complete set of measurements from the 12 days used in the training set. This corresponds to more than 15 million cases including the training set as a subset. A second set of statistics from 3 independent days is also shown. Instead of stratifying the statistics in land/sea, different latitude bands, etc., in this paper we will show statistics stratified based on the quality indicators alone. To keep it simple we have considered only the quality indicator for surface air temperature (QTa). The thresholds applied to this quality indicator, to get a separation into 5 different quality classes, were chosen such that approximately 200000 cases per day (about 15.6% of the total) in each of the first four quality classes were obtained. Based on QTa from MWIR this approach resulted in the following thresholds: 0.703, 0.978, 1.285 and 1.612. As the IR retrieval quality is generally worse (and therefore QTa higher) than for MWIR, using these threshold to classify the IR retrievals leads to smaller yields in the high quality classes. So in order to get a fair comparison of the MWIR and IR retrieval quality, we have also used a second quality stratification chosen such that distribution in the quality classes for the IR case resembles the distribution in the MWIR case with the first stratification (the thresholds on QTa used to obtain this are 1.028, 1.629, 2.172 and 2.827. A summary of the overall yields in the quality classes using the two stratifications is given in the Table 1.

	Q1	Q2	Q3	Q4	Q5
MWIR PWLR, Quality Stratification I	15.6 %	15.6 %	15.6 %	15.6 %	37.9 %
IR PWLR, Quality Stratification I	5.7 %	8.6 %	7.8 %	8.6 %	69.3 %
MWIR PWLR, Quality Stratification II	34.0 %	29.2 %	20.2 %	12.1 %	4.5 %
IR PWLR, Quality Stratification II	15.6 %	15.6 %	15.6 %	15.6 %	37.9 %

Table 1 **Yield of quality classes with two different set of thresholds on QTa .**


MWIR PWLR, Quality Stratification I

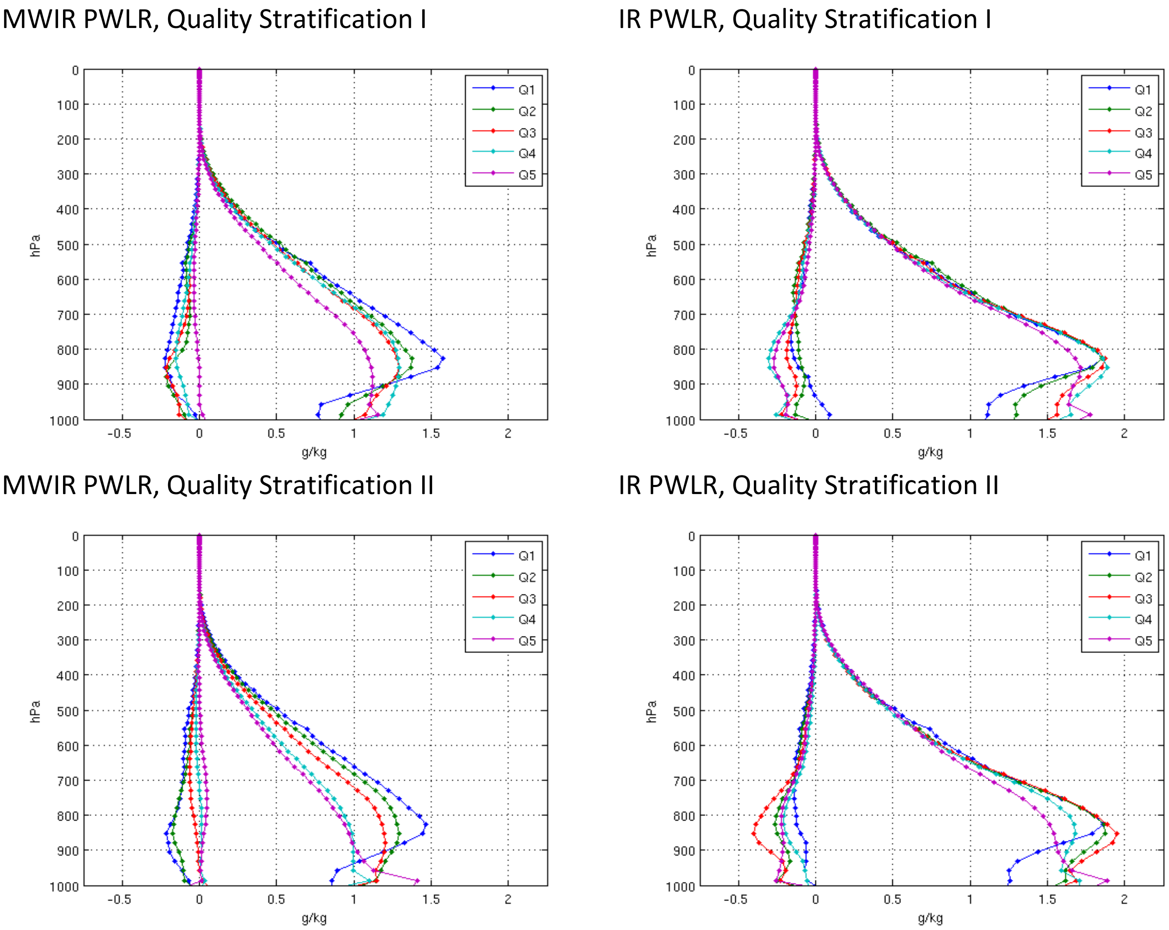
IR PWLR, Quality Stratification I

MWIR PWLR, Quality Stratification II

IR PWLR, Quality Stratification II

Figure 6 **Temperature profile statistics of retrieval minus ECMWF analysis. 12 days from Nov 2011 to Oct 2012 .**

In Figure 6 we see that the classification based on the quality indicator of Ta reflects the actual quality of the retrieved temperature profile up to about 150 hPa. For the water vapour profiles (Figure 7) this is true only up to about 900 hPa. When using the same quality thresholds, similar retrieval error statistics, but with considerable lower yield, are obtained for IR PWLR as for MWIR PWLR. Comparing the two cases with equal yields we get a feeling for the improvement due to the inclusion of microwave radiances in the retrieval, We see that for temperature this improvement is mainly seen below 850 hPa in Q1, below 700 hPa in Q3 and below 500 hPa in Q3.


MWIR PWLR, Quality Stratification I

IR PWLR, Quality Stratification I

MWIR PWLR, Quality Stratification II

IR PWLR, Quality Stratification II

Figure 7 **Water vapour profile statistics of retrieval minus ECMWF analysis, 12 days from Nov 2011 to Oct 2012.**

For water vapour we see a general improvement below 600 hPa when microwave information is included in the retrieval. The water vapour retrieval performance is relatively similar in the 5 quality classes defined by the value of QTa. By defining quality classes in terms of the water vapour profile quality indicator, very distinct figures for water vapour retrieval error standard deviation is obtained (not shown here). Notice the very low retrieval biases. This is guaranteed by the linear regression method, which produces an unbiased estimate of the retrieved parameters (provided the reference truth used in the training set is unbiased). The small negative bias of the water vapour retrieval observed in the best quality classes is explained by the sampling strategy employed for the selection of the training set, where the warmest IFOV in each EFOV is chosen. This strategy was originally chosen with the idea in mind only to make a single retrieval of the profiles within each EFOV based on the warmest (and likely clearest) IFOV. But experiments have shown that better retrievals are obtained when individual retrievals for each IFOV are performed and the training set sampling strategy should be changed accordingly. In fact there is no reason not to make the training set as big as possible and include, for example, all 15 million measurements from the 12 selected days.

As the retrieval performance statistics presented above were based on the same days as were used for the training we repeat some of the statistics for three independent days in 2013 for both Metop-A and B. This is shown for MWIR PWLR using quality stratification I only. In Figure 8 and 9 we see that the retrieval statistics remain essentially unchanged for Metop-A for both the temperature and humidity profiles, but that considerable biases appear in the Metop-B retrievals using the regression coefficients based on Metop-A data. Interestingly this is not the case for IR PWLR (not shown here), which indicates that these biases are caused by different characteristics of AMSU/MHS on Metop-A and B.

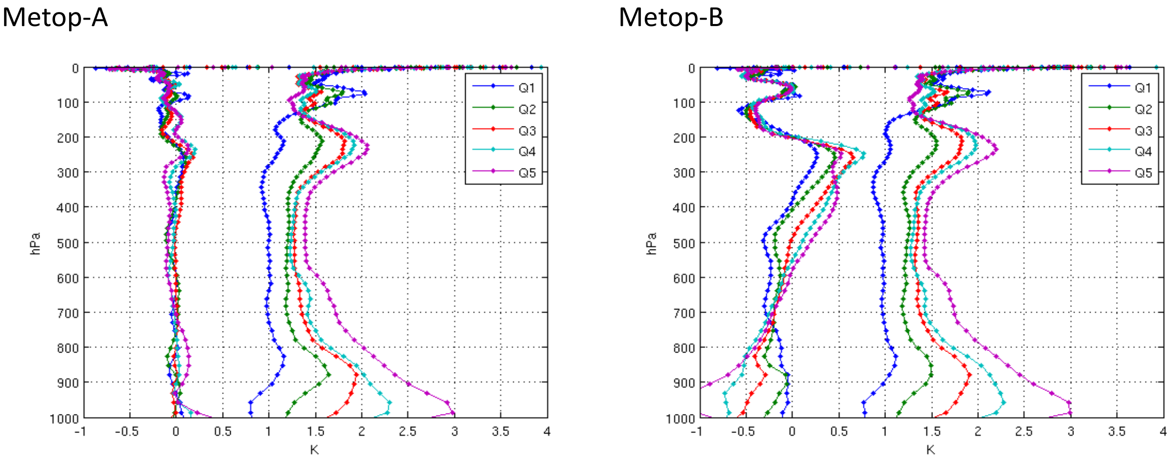


Figure 8 Temperature profile statistics of retrieval minus ECMWF analysis. 20130417 + 20130717 + 20131017.

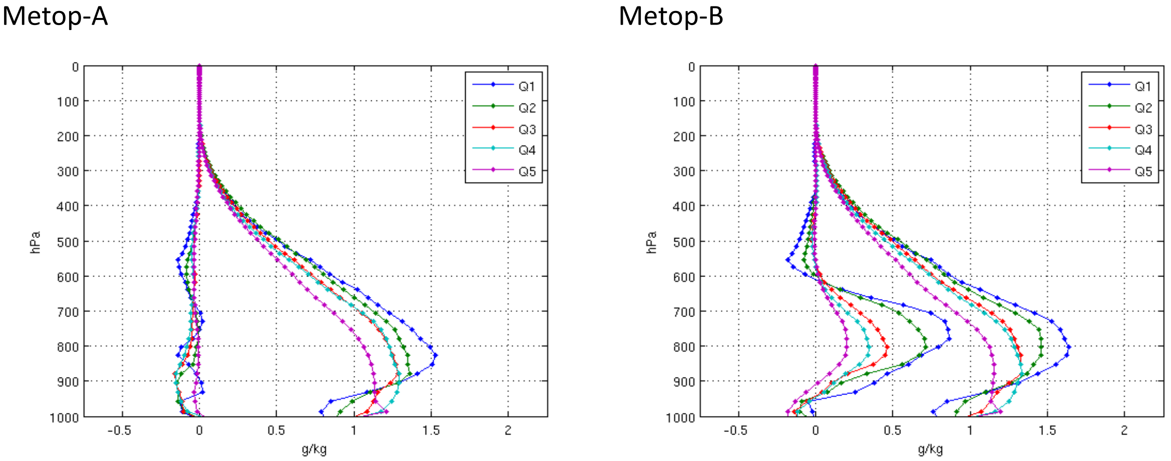


***Figure 9* Water vapour profile statistics of retrieval minus ECMWF analysis. 20130417 + 20130717 + 20131017.**

We now look at the retrieval error statistics in radiance space. For this we have used the clear fields of view over ocean and computed the standard deviation of the difference between the noise filtered IASI L1C measurements and the corresponding forward model spectra computed with RTTOV 10.2 taking the retrieved profiles and surface skin temperature as input. For comparison the same statistics were also performed taking profiles and surface skin temperature from ECMWF analysis as input. In Figure 10 and 11 the results for Band 1 and 2 respectively are shown. It is seen that the residual standard deviation is quite similar in the two cases and is actually lower for the retrievals than for ECMWF in most of Band 2. This is very encouraging when you recall that the retrieval does not involve any explicit minimization of the residual and suggests that for certain pressure levels the retrieved humidity is closer to the truth than the collocated ECMWF analysis profiles. Whereby we must keep in mind that is probably due to the representativeness errors of the latter, which is expected to be high for water vapour since it is highly variable both temporally and spatially.

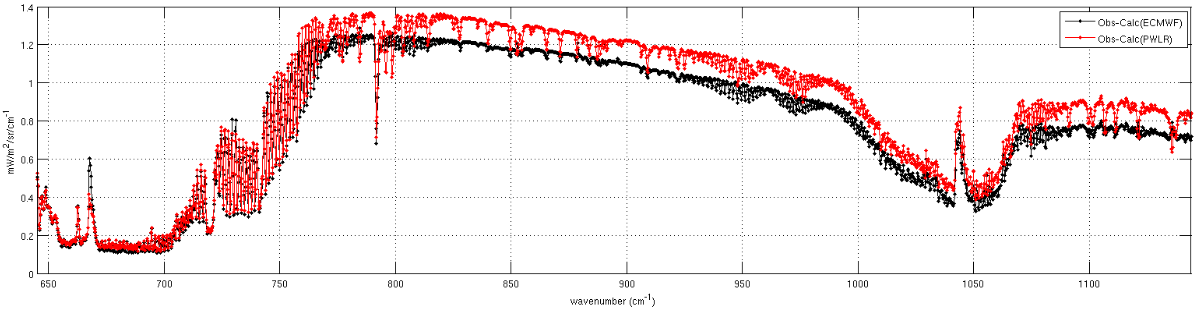


***Figure 10* Band 1 OBS minus CALC for MWIR PWLR and ECMWF analysis. Clear sky, ocean, 20121001.**

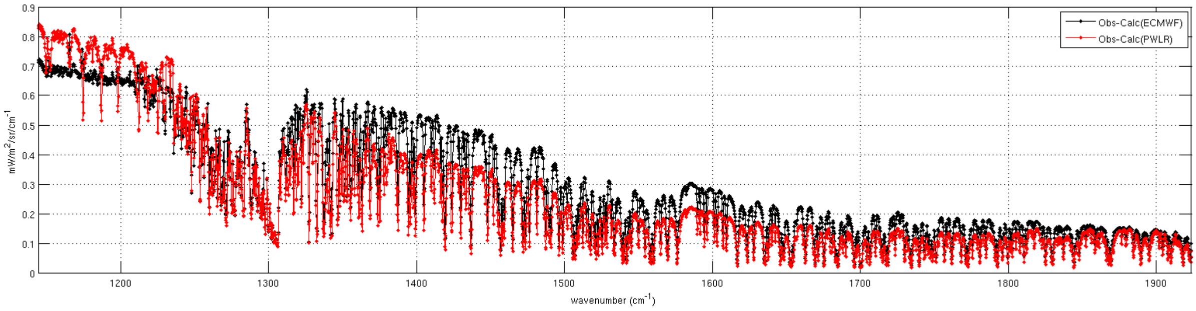


Figure 11 **Band 2 OBS minus CALC for MWIR PWLR and ECMWF analysis. Clear sky, ocean, 20121001**.

CONCLUSIONS AND POSSIBILITIES FOR IMPROVEMENT

We have developed a fast regression retrieval of atmospheric and surface parameters based on co-located infrared and microwave observations, with an associated estimate of the retrieval error. The retrieval is designed for all sky situations and has been shown to perform well as measured against ECMWF analysis. Besides the synergistic use of IASI, AMSU and MHS radiances, which enables much better retrievals than possible with any one of the instruments alone, other factors contributing to the good performance are: the use of a large training set using real satellite observations, the choice of the representation of the parameters to retrieve as well as the simple and flexible method to account for nonlinearities offered by the piecewise linear regression method. This method is known as regression trees in the classical statistical literature and could be complemented with other modern nonlinear techniques such as kernel ridge regression to improve the performance further.

The use of ECMWF analysis to build the training set has the advantage that it is easy to create training sets of very large size, which makes them robust with respect to random errors, but other and better (in terms of biases) sources of “truth” could also be considered, if available. We note that the ECWMF model has recently been improved and that the analysis is now available at 137 model levels. This update has resulted in significant biases in the stratospheric temperatures between the new and old models. To take advantage of these recent improvements, both the MWIR and IR PWLR will be retrained with ECMWF analysis fields from the new model. In the scope of this retraining we will derive separate coefficients for Metop-A and B in order to get rid of the retrieval biases observed for Metop-B.

The current scheme uses (a subset of) the IASI PC scores used for dissemination as predictors. This has two drawbacks: 1) the PC scores are separated into three bands, which mean that inter-band correlations are not exploited and that there is some co-linearity in the predictors and 2) instrument artefacts affecting the IASI radiances are included in the predictors. These drawbacks could be avoided by using an alternative set of basis vectors for the PC score generation, based on the similarity of directions between the signal and forward model spaces. It is important to emphasize that the IASI PC scores used for dissemination contain (extremely close to) the full amount of atmospheric signal within the measurements and that the second set of PC scores can therefore be computed directly from the dissemination PC scores.

For cloud free cases the PWLR retrievals offer excellent statistically based a-priori information for filling in the null space within an optimal estimation method, which improves the retrieval performance even further. Due to the difficulties of cloudy forward modelling, it will likely be hard to get improved retrieval performance from optimal estimation compared to PWLR in cloudy situations, but studies along this direction will be undertaken.