MIT Open Access Articles

An anomaly correlation skill score for the evaluation of the performance of hyperspectral infrared sounders

An anomaly correlation skill score for the evaluation of the performance of hyperspectral infrared sounders[1].

Hartmut H. Aumann[a], Evan Manning[a], Chris Barnet[b], Eric Maddy[c] andWilliam Blackwell[d]

[a] California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA 91109
[b] NOAA/NEDIS, Camp Springs, MD
[c] Perot Systems Government Service, Inc., Fairfax, VA
[d] MIT Lincoln Laboratory, Lexington, MA

Abstract

With the availability of very accurate six hour forecasts, the metric of accuracy alone for the evaluation of the performance of a retrieval system can produce misleading results: the retrievals may be statistically accurate, but be of little value compared to the accurate forecast. A useful characterization of the quality of a retrieval system and its potential to contribute to an improved weather forecast is its skill, which we define as the ability to make retrievals of geophysical parameters which are closer to the truth than the six hour forecast. We illustrate retrieval skill using one day of AMSU-A and AIRS data with three different retrieval algorithms. In the spirit of achieving global retrievals under clear and cloudy conditions, we evaluated retrieval accuracy and skill for 90% of the covered area. Two of the three algorithms meet the 1 K/1 km "RAOB quality" accuracy requirement and have skill between 900 and 150 hPa, but none have skill between the surface and 900 hPa.

AIRS was launched on the EOS Aqua spacecraft in May 2002 into a 705 km polar sun-synchronous orbit with accurately maintained 1:30 PM ascending node. Essentially un-interrupted data are freely available since September 2002.

Keywords: Atmospheric Infrared Sounder AIRS Temperature retrieval humidity cloud clearing

1. Introduction

The basic physics involved in using the wavelength dependent transmission of CO_2 in the thermal infrared for temperature sounding from earth orbit was published in 1959 by Kaplan [1]. Ten years later Chahine [2] showed that with high spectral resolution measurements Radiosonde (RAOB) quality temperature and moisture profiles could be obtained with a hyperspectral sounder in the 4.3 micron CO2 band and that retrievals with RAOB accuracy should be possible under clear and cloudy conditions using a cloud-clearing technique [3]. "RAOB accuracy" is usually stated as 1 K rms in 1 km thick layers, often referred to as 1K/1km accuracy. By the mid 1980s the accuracy of retrievals from HIRS type low spectral resolution infrared sounders was about 2.5K/2.5km. The first Global Circulation Models (GCM) emerged in the late 1980s and their accuracy quickly became limited by the accuracy of the initialization with 2K/2km data. Around that time NASA recognized that the technology for building a hyperspectral infrared sounder in a polar orbit with 1K/1km accuracy was at hand and that a GCM initialized with RAOB quality global data should result in a major advance in the accuracy of weather forecasting. This led to the approval of what was to become the Atmospheric InfraRed Sounder (AIRS) [4]. AIRS was launched on the EOS Aqua spacecraft in May 2002 into a 705 km polar sun-synchronous orbit.

---

[1] Copyright 2009. All rights reserved. Opinions, interpretations, conclusions, and recommendations are those of the author and not necessarily endorsed by the United States Government.
SPIE Optics-Photonics Meeting 2-6 August 2009, San Diego, California

Around the year 2000, GCMs were initialized using RAOBs, surface reports and HIRS data, and climatology in the form of the National Center for Environmental Predictions (NCEP) reanalysis [5] of the monthly mean state of the atmosphere on a 2.5 degree latitude/longitude grid became available. The national weather forecasting centers soon recognized that accuracy achieved by a GCM with global statistics relative to RAOBs in the presence of a reasonably accurate climatology was not a useful metric for measuring the quality of GCMs and developed the concept of forecast skill. The skill is defined as the ability of a GCM to predict the state of the atmosphere more accurately than the state expected at the same time and location from climatology, i.e. a forecast which was no more accurate than the state of the atmosphere expected from climatology had zero skill. If Forecast is the state of the atmosphere predicted t days in the future at time $t0$ and Truth is the true state of the atmosphere at time $t0$ and location x, then skill is the correlation between (Forecast($t0$-t,x)-climatology($t0$,x)) and (Truth($t0$,x)-climatology($t0$,x)

$$\text{Skill}(t) = \text{cor}((\text{Forecast}(t0\text{-}t,x)\text{-climatology}(t0,x),(\text{Truth}(t0,x)\text{-climatology}(t0,x))) \qquad (1)$$

This skill is evaluated for large regions, like the entire northern hemisphere, and specific quantities, like the temperature at 500 hPa. By definition, Skill(t=$t0$)=1. Skill decreases as the length of the forecast is increased. The time where Skill has dropped to 0.6 is the length of the useful forecast. By 2005 GCMs assimilated data from more than a dozen infrared and microwave sounders, in addition to RAOBs, floating buoys, ship reports and surface reports, and the length of the useful forecast for temperature profiles was about 5 days.

In 2006 the Joint Center for Data Assimilation (JCDA) announced that the assimilation of cloud-free AIRS radiances increased the length of the skilled forecast from 5 days by an additional 6 hours [6]. Less than 1% of the AIRS spectra were used in this assimilation, and these clear spectra were obviously in the meteorologically least interesting conditions.

AIRS was the first hyperspectral infrared sounder used for weather forecasting. It was followed by the Infrared Atmospheric Sounding Interferometer (IASI) [7], launched in October 2006, and the Crosstrack Infrared Sounder (CrIS) [8] is expected to be launch in 2011. AIRS, IASI and CrIS are infrared hyperspectral sounders of comparable spectral coverage, resolution, noise, cross-track spatial coverage and spatial footprint size (12 km at nadir) and all three are expected to be able to make retrievals with "RAOB accuracy".

The global comparison of AIRS retrievals with RAOBs confirmed that AIRS achieved 1K/1km accuracy under fairly global clear and cloudy, but carefully quality controlled, conditions [9]. The same analysis also showed that the accuracy of GCMs approached the 1 K/1km level, with the accuracy of the 6 hour forecast not far behind. In the presence a global forecast with near "RAOB accuracy", the statistical analysis of the accuracy of a retrieval system, particularly when subjected to tight quality control, could produce misleading results: all interesting cases, i.e. due to the passage of storms, where the true state of the atmosphere differed significantly from the state expected from climatology, could be missed, i.e. the retrievals failed or were rejected by quality control. An optimum likelihood retrieval would in these cases return the background. In both cases the rms accuracy of the retrievals may well be better than 1K/1km. A retrieval system is the combination of a hyperspectral sounder, supported by a microwave sounder, and its associated optimized retrieval software. Just as the value of a forecast has to be interpreted relative to the zero skill accuracy of the climatological expected value, we can define retrieval skill as the ability of a sounding system to make retrievals from Earth orbit which are closer to the true state than the value predicted by the 6 hour forecast. The retrieval skill score would allow the comparison of the performance of different retrieval algorithms using data from the same sounders, i.e. AIRS and AMSU in a 1:30 PM orbit, and sounding systems on other spacecraft, such as IASI and AMSU in the 9:30 AM orbit. If it can be shown that the temperature and water vapor profile retrievals in the presence of clouds have skill, then the data contained information which was not captured by the current assimilation of 1% of the clear radiances. If, on the other hand, the retrievals from a retrieval system are statistically accurate, but have no skill, then the usefulness of such a system in support of weather forecasting would have to be re-examined.

2. Method

Our definition of retrieval skill is patterned after the forecast skill. We define retrieval skill as the correlation between (retrieval-background) and (Truth-background)

$$Skill(p) = cor((retrieval(p)-background(p)),(Truth(p)-background(p))) \qquad (2)$$

as function of the pressure altitude p.

Skill(p) is evaluated for all possible retrievals and the background is substituted for those pressure levels where the retrieval did not return a solution of acceptable quality. This procedure emulates optimum likelihood retrievals which return a solution for 100% of the cases, but the solution equals the background where the data contain no information, such as below a solid cloud deck. A perfect retrieval system would have a skill score of unity at all altitudes. The skill of a real retrieval system will be a function of altitude. The pressure altitude where the skill drops below 0.6 is the metric of retrieval quality.

Ideally, the truth should be the state of the atmosphere measured by the RAOBs and the background should be the most accurate representation of the state of the atmosphere available at the time of the observation. At present this is the state predicted by the 6 hour forecast. Under these conditions Eq.2 would produce an absolute skill. Unfortunately, the use of RAOBs requires very large data sets. Divakarla et al. [9] were able to find only 33,000 RAOBs within 50 km and 1 hours of a high quality AIRS retrieval in 2 years of data. For the testing with a smaller data set compromises have to be made.

For our illustration of the retrieval skill metric as a concept we made the three simplifications:

1. We used a monthly climatology for the background. The climatology was generated from data from the European Center for Medium range Forecasting (ECMF) from 2008 in the form of monthly means on a one degree grid for the ascending (day) and descending (night) orbits appropriate for AIRS.

2. We used the ECMWF analysis interpolated to the time of the retrieval as the truth.

3. Based on simplifications 1. and 2. we limited the evaluation of retrieval skill to temperature retrievals from the tropical oceans from one day of AIRS data, 6 September 2002, with about 54,000 potential retrievals. However, the evaluation of rms accuracy of retrieval algorithms for hundreds of days indicates that the statistical flavor from this day is representative for a much larger data set, including global data.

Strictly speaking, with the use of climatology as background (instead of the forecast) and the use of the ECWMF analysis as the truth (instead of RAOBs) the skill calculated from Eq.2 is not an absolute skill, but is a relative skill.
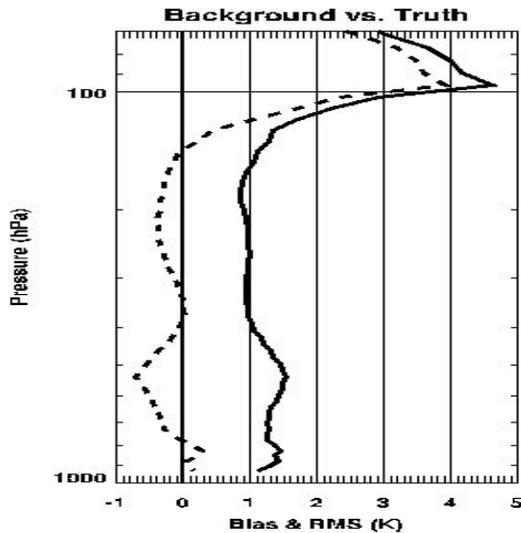
Background vs. Truth

From Figure 1 it can be seen that from the surface to about 400 hPa, i.e. in the lower troposphere, the rms difference between the truth and the climatology is about 1.5K, decreasing to about 1 K rms between 400 hPa and 150 hPa. Use of the tropical ocean climatology thus emulates the accuracy of the six hour forecast in the troposphere. The rms difference between the truth for 2002-09-06 and climatology increases to 4 K between 120 and 80 hPa due to quasi biannual temperature waves in the stratosphere. The use of climatology as background in our definition of skill is a poor proxy for the six hour forecast in the stratosphere.

Figure 1. The mean (dashed) and rms (solid) lines show the difference between the truth for 2002-09-06 in the tropical oceans and the climatology derived from September 2008 ECMWF data.

Making retrievals in the tropical oceans with 1 K rms accuracy and high yield is challenging, since almost all cases are under considerably cloudy conditions.

We illustrated the magnitude of the cloud effect on the infrared brightness temperatures using the parameter c1231. At 1231 cm[-1] the atmospheric absorption is relatively low, i.e. a 300 K sea surface temperature results in typically a 296 K brightness temperature. By using two channels, one at 1231 cm[-1], the other at 1227.75 cm[-1], we can account for the atmospheric absorption [10]. If we also have a reliable value for the true surface temperature, SST, we can calculate the brightness temperature which we should measure in the absence of clouds. We define c1231 to be the difference between the expected brightness temperature under cloud-free conditions and the brightness temperature observed at 1231 cm[-1]. We use the Real Time Global SST (RTGSST) [11] created daily by NCEP on a one degree grid as the estimate of the SST. Under almost cloud-free conditions we have shown [10] that c1231 has a bias of less than 0.3 K and a standard deviation of 0.5 K. These conditions occur for less than 1% of the spectra. Figure 2 shows the distribution of c1231 for the non-frozen ocean spectra from 2002-09-06. The mean c1231 was 11.5 K, but 19% of the spectra were minimally effected by clouds (abs(c1231)<1). For one percent of the spectra c1231 is larger than 80 K, typical of deep convective clouds associate with thunderstorms.

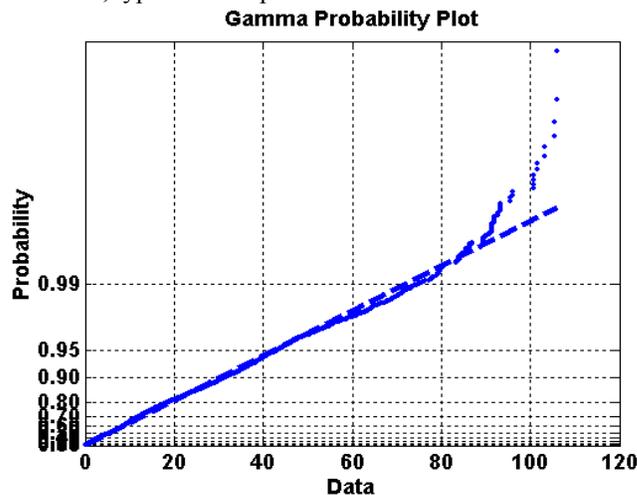Figure 2 shows that the effect of clouds on the infrared brightness temperatures at 1231 cm[-1] is approximately gamma distributed. The values of c1231 for 2002-09-06 had a mean of 11.5K, with standard deviation of 15.5K, which are fairly typical. For 10% of the spectra c1231>32K. Obtaining accurate retrievals under these conditions would be unrealistic. Requiring retrievals in the tropical oceans for 90% of the cases is consistent with a claim of global retrievals under clear and cloudy conditions.

Figure 2. C1231, for the tropical ocean spectra from 2020906.

3. Results

There are a number of techniques capable of making retrievals under cloudy conditions. We evaluate skill for three different retrievals, A, B and C, all using the combination of AMSU and AIRS data, and all returned solutions with a variety of quality control (QC) indicators for about 98% of the possible retrievals. Rather than attempting to normalize the QC from the three retrieval algorithms we ranked the quality of the retrievals by the mean rms difference from truth averaged for the atmospheric levels between 930 hPa and 100 hPa, and then perform all statistics over the best 90%. The retrieval skill as function of pressure altitude is shown in Figure 4.
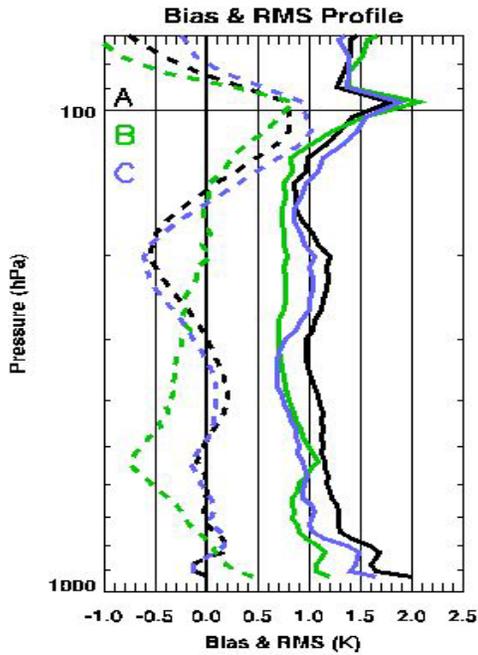

Figure 3 The rms accuracy of the retrievals (solid line) and the bias (dashed line) as function of pressure altitude (A=black, B=green and C=light blue).
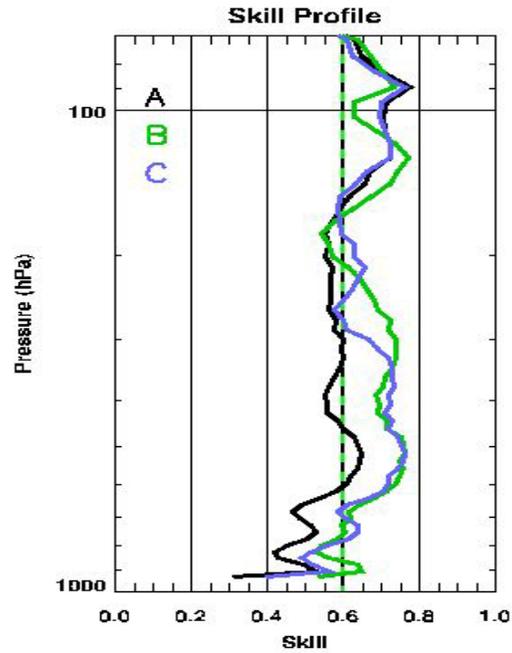
Figure 4 shows the skill of the retrievals as function of pressure altitude (A=black, B=green and C=light blue).

4. Discussion

We evaluated the rms temperature retrieval accuracy in a difficult area, where the mean cloud effect in the data was 11.5 K. In addition, and in the spirit of achieving retrievals globally under cloudy conditions, we demanded retrievals for 90% of the area where data were taken. From Figure 3. it can be seen that the mean accuracy of the retrievals between the surface and 100 hPa for all retrievals is close to 1.2 K rms, i.e. does not meet the 1K/1km requirement. Algorithms B and C meet or exceed the 1K/1km requirement in a mean sense throughout the troposphere. Retrieval A reaches an rms accuracy of 1 K only near 300 hPa, and is worse than 1.5 K rms below 800 hPa. The rms accuracy of retrieval B is better than 1 K between 800 and 130 hPa., and reaches an rms accuracy close to 1.1 K from 800 hPa to the surface, but with a curious 0.5 K warm bias near the surface, and a cold bias throughout most of the troposphere. The accuracy of retrieval C is better than 1 K above 800 hPa, but the rms drops to about 1.4 K between the 800 hPa and the surface. Algorithms A and B have close to zero bias throughout the troposphere. The retrieval skill as function of

pressure altitude, gives new insight into the performance of the three algorithms, in view of the fact that the ECMWF 6 hour forecast in this area is very good.

Based on its rms accuracy alone it is no surprise that algorithm A has virtually no skill below 150 hPa, certainly non below 600 hPa, presumably due to heavy reliance on the background and due to the way clouds are handled. Retrievals B and C did better in terms of rms accuracy than the forecast between 800 hPa and 150 hPa, and from Figure 4 we can see that B and C have significant skill above 800 hPa. Below 800 hPA algorithm C and A show an equal lack of skill, while algorithm B still has some skill at 900 hPa. None of the three algorithm has significant skill near the surface. This should not be surprising, given the accuracy of the knowledge of the sea surface temperature in the forecast, and the difficulty of handling clouds near the surface.

Retrieval skill is the result of a complex interaction between the capability of a sounder to make measurements in the presence of clouds with high signal-to-noise, radiometric and spectral fidelity and retrieval software. There are a large number of permutations of this evaluation even for a given instrument or instrument combination, such as AIRS and AMSU, and retrieval algorithm details. For example: If we were to accept a 50% yield based on internal quality indicators, the rms accuracy would certainly improve, but would the skill increase or decrease? Temperature and water vapor retrievals for land are much more difficult and both are much less accurate in the forecast. What is the water vapor retrieval skill? Is the skill over land better than over ocean? What makes the skill of algorithms B and C better than A? How does the retrieval skill with IASI compare to that of AIRS? These evaluations are in progress, but are beyond the scope of this paper.

The AIRS instrument with algorithms B or C has skill in most of the troposphere and assimilation of AIRS data has provided positive impact on the weather forecast. This begs a larger question: If a retrieval system were to be shown to have no skill, could it still have a positive impact on the forecast, and if not, would such a retrieval system be marketable?

5. Conclusion

With the availability of accurate six hour forecasts, the metric of accuracy alone for the evaluation of the performance of a retrieval system can produce misleading results. The design of a retrieval system in terms of accuracy alone, in particular the use of tight quality control, can have the undesirable result of creating a system which is accurate, but has no or little real skill. A more relevant characterization of the quality of a retrieval algorithm and its potential to contribute to an improved weather forecast is its skill, which we define as the ability of an algorithm to get closer to the truth than the forecast, when the truth differs significantly from the forecast. We illustrate retrieval skill using one day of AMSU-A and AIRS data from non-frozen oceans with three different retrieval algorithms, all with more than 90% yield under tropical ocean cloudy conditions. Two of the three algorithms have better than 1 K/1km "RAOB quality" accuracy between 800 and 150 hPa and have skill. None of the algorithms have skill between the surface and 900 hPa.

Acknowledgments

References

[1] Kaplan, L.D "Inference of atmospheric structure from satellite remote radiation measurements", J. Opt. Soc. Amer. Vol 49, pp1004-1007 (1959).

[2] Chahine, M.T "Determination of the temperature profile in the atmosphere from its outgoing radiances", J. Opt. Soc. Amer. Vol 58, pp. 1634-1637 (1968).

[3] Chahine, M.T "Remote Sounding of Cloudy Atmospheres. The single cloud layer", J. Atmos. Sci. Vol 31, pp.233-243 (1974)

[4] Aumann, H. H. et al. "AIRS/AMSU/HSB on the Aqua Mission: Design, Science Objectives, Data Products, and Processing Systems", IEEE Trans. Geo Sci.&Remote Sensing Vol 41, 2, pp.253-264 (2003).

[5] Kalnay, E.D. et al., "The NCEP/NCAR 40-year Reanalysis Project" Bull.Amer.Meteor.Soc.77, 437-471 (1996)

[6] LeMarshall, J., J. Jung, J. Derber, M. Chahine, R. Treadon, S.J. Lord, M. Goldberg, W. Wolf, H. C. Liu, J. Joiner, J. Woollen, R. Todling, P.VanDelst and Y. Tahara "Improving G;obal Analysis and Forecasting with AIRS" BAMS, July 2006, pp. 891-894 (2006)

[7] Blumstein, D. et al. "IASI instrument: Technical overview and measured performances", Proc. SPIE Vol. 5543, p.196, DOI:10.1117/12.560907 (2004)

[8] Glumb, R. J., D.C. Jordan, and P. Mantica, "Development of the Crosstrack Infrared Sounder (CrIS) sensor design", Proc. SPIE Vol. 4486, p. 411-424, Infrared Spaceborne Remote Sensing IX, (2002)

[9] Divakarla M. G., C.D. Barnet, M.D. Goldberg, L.M. McMillin, E. Maddy, W. Wolf, L. Zhou, X. Liu, "Validation of Atmospheric Infrared Sounder temperature and water vapor retrievals with matched radiosonde measurements and forecasts" JGR Vol.111, D09S15, doi:10,1029/2005/JD006166, (2006).

[10] Aumann, H.H. et al. " Three years of AIRS radiometric calibration validation using sea surface temperatures", JGR 111,D16S90, doi:10.1029/2005/JD006822, (2006)

[11] Thiebaux, J.E., E.Rogers, W.Wang and B.Katz, "A new high resolution blended real-time global sea surface temperature analysis", Bull.Amer.Meteor.Soc. 84(5),645-656 (2003).