An Anomaly Correlation Skill Score for the evaluation of the performance of Hyperspectral Infrared Sounders [1].

Hartmut H. Aumann and Evan Manning
California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA 91109
and
Chris Barnet and Eric Maddy
NOAA/NEDIS, Campsprings, MD
and
William Blackwell
MIT Lincoln Laboratory
Lexington, MA

Abstract

With the availability of very accurate forecasts, the metric of accuracy alone for the evaluation of the performance of a retrieval system can produce misleading results. 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, when the truth differs significantly from the forecast. We illustrate retrieval skill using one day of AMSU and AIRS data with three different retrieval algorithms, which result in retrievals for more than 90% of the potential retrievals under clear and cloudy conditions. Two of the three algorithms have better than 1 K rms "RAOB quality" accuracy on the troposphere, but only one has skill between 900 and 100 mb.

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 CO2 in the thermal infrared for temperature sounding from earth orbit was published in 1959 by Kaplan. Ten years later Chahine (1968) showed that with high spectral resolution measurements Radio Sonde (RAOB) quality temperature and moisture profiles could be obtained with a hyperspectral sounder in the 4.3 micron CO2 band and that retrievals with Radiosonde (RAOB) accuracy should be possible under clear and cloudy conditions using a cloud-clearing technique (Chahine 1974). "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 became quickly limited by the accuracy of the initialization with 2K/2km data. Around that time NASA recognized that the technology for building a hyper-spectral infrared sounder in a polar orbit with 1K/1km accuracy was at hand and that the combination of a GCM initialize 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) (Aumann et al 2003). AIRS was launched on the EOS Aqua spacecraft in May 2002 into a 705 km polar sun-synchronous orbit.

---

[1] 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 (Kalnay et al.1996) 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)

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

This skill is evaluated for large regions, like the entire northern hemisphere, and specific quantities, like the temperature at 500 hP. 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 (LeMarshall 2006). 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, Blumstein et al. 2004), launched in October 2006, and the Crosstrack Infrared Sounder (CrIS, Glumb et al. 2002) is expected to be launch in 2011. AIRS, IASI and CrIS are infrared hyperspectral sounder 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 global clear and cloudy, but carefully quality controlled conditions (Divakarla et al. 2006). 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, but the rms accuracy of the remaining 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)

$$Skill(p) = cor((retrieval(p)-background(p)),(Truth(p)-background(p))) * sqrt(yield) \quad (2)$$

as function of the pressure altitude p.

The sqrt(yield), where yield is the ratio of the number of spectra for which a retrieval was returned to the number of possible retrievals, accounts for algorithms which fail for a significant fraction of the possible retrievals. 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. Optimum Likelihood Retrievals 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 retrieval which returns an extremely accurate results, but only for the 10% of the cases, while returning the background for the remaining 90% would have a yield of 100%, but a skill of only 0.31. The formulation of Eq.2 insures that an algorithm which returns an equally accurate solution for 10% of the cases, but returns no solution for 90% of the attempted retrievals also has a skill of 0.31.

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. (2006) 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.

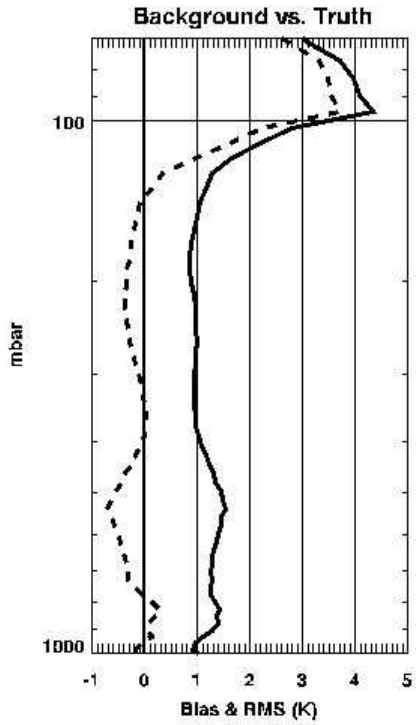

Figure 1. mean (dashed) and rms (solid) difference between the truth for 20020906 in the tropical oceans and the climatology derived from September 2008 ECMWF data.

For our illustration of the retrieval skill metric as a concept we made the following 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 in the troposphere from the tropical oceans from one day of AIRS data, 6 September 2002, with about 54,000 potential retrievals.

We selected 20020906 since this day was previously evaluated extensively and found to be a typical day in terms of clouds. 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.

Figure 1 shows that from the surface to about 120 mb, i.e. in the troposphere, the rms difference between the truth and the climatology, averaged below 120 mb, is about 1.2 K. 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 20020906 and climatology increases to 4 K between 120 and 80 mb due to annual temperature waves. Climatology is a poor proxy for the forecast in the stratosphere.

Making retrievals in tropical oceans 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 ce1231. At 1231 cm^{-1} the atmospheric absorption is very low, i.e. a 300 K sea surface temperature results in typically a 297 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 (Aumann et al. 2006 Validation paper). 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 ce1231 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 reference) created daily by NCEP on a one degree grid as the estimate of the sst. Under almost cloud-free conditions we have shown (Aumann 2006) that ce1231 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 ce1231 for the non-frozen ocean spectra from 20020906. If abs(ce1231)<1) then the spectra are for all intents and purposes minimally effected by clouds. On 20020906 this is the case for 19% of the spectra. About 39% of the spectra have abs(ce1231)<2), which is basically low stratus clouds. For 20020906 ce1231 had a mean of 11.5K. In the presence of the deep convective clouds associated with tropical thunderstorms ce1231 can become very large. On 20020906 ce1231 was larger than 80K for 1% of the spectra.

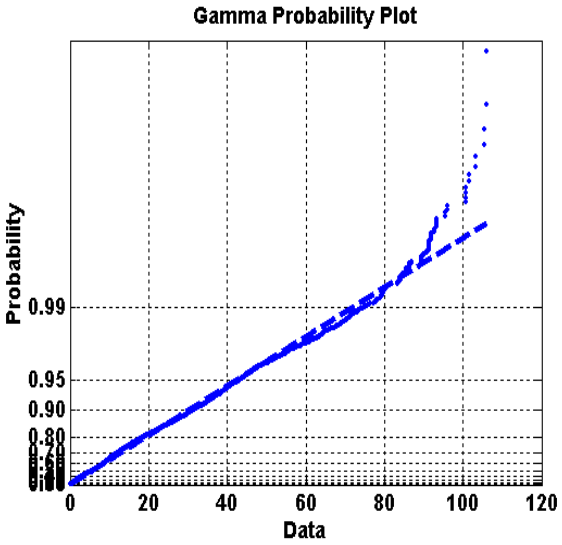

Figure 2 shows the infrared cloud effect, ce1231, for the tropical ocean spectra from 2020906.

Figure 2 shows that the effect of clouds on the infrared brightness temperatures at 1231 cm^{-1} is approximately gamma distributed. The values of ce1231 for 20020906 had a mean of 11.5K, with standard deviation of 15.5K, which are fairly typical. For one percent of the spectra ce1231 is larger than 80 K, typical of deep convective clouds associate with thunderstorms. For 10% of the spectra ce1231>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.

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 retrieved values by the absolute value of the difference from truth, and then perform all statistics over the best 90% of the retrievals. Figure 3. shows the rms accuracy of the retrievals between the surface and 70 mb. Retrieval A has an rms accuracy of about 1 K between 700 mb and 100 mb, which degrades to about 1.8 K rms between 700 mb and the surface. The accuracy of retrieval C is better than 1 K above 800 mb and drops to about 1.4 K between the 800 mb and the surface. The rms accuracy of retrieval B is better than 1 K below 110 mb. Figure 4 shows retrieval skill as function of altitude.

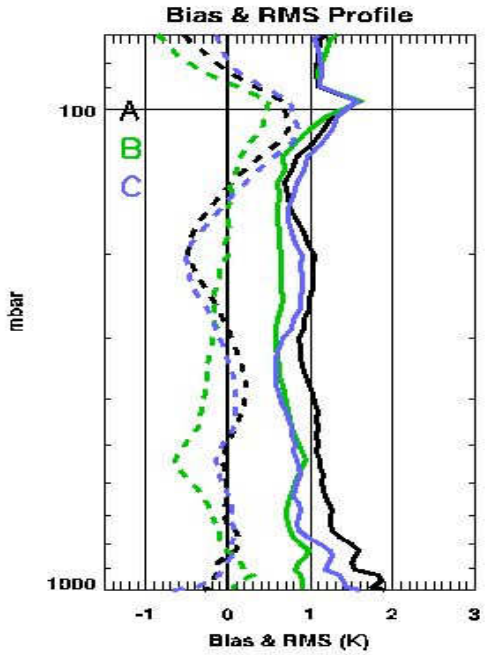

Figure 3 shows the rms accuracy of the retrievals (solid line) and the bias (dashed line) as function of pressure altitude (A=black, B=green and C=light blue).

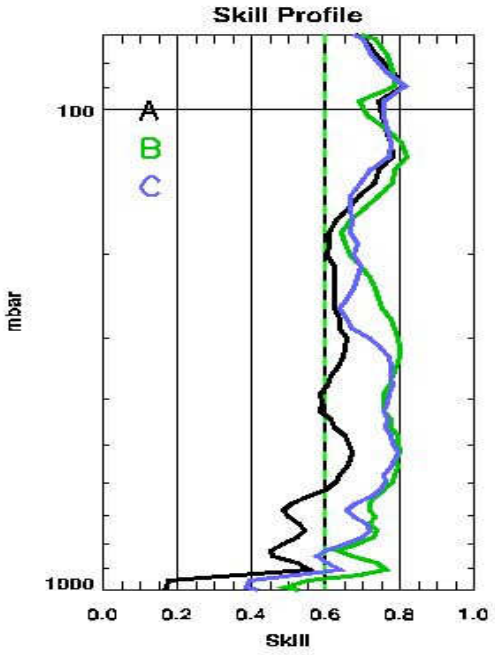

Figure 4 shows the skill of the retrievals as function of pressure altitude (A=black, B=green and C=light blue).

4. Discussion

Given that we evaluate the rms temperature retrieval accuracy for 90% of the possible retrievals in a very difficult area, where the mean cloud effect in the data was 11.5 K, the rms performance of retrievals B and C is at first glance remarkable. However, as shown in Figure 1, the rms difference between the truth and the background was only slightly worse than 1 K in a large fraction of the troposphere. With this 1 K value in mind, the results in Figure 3 are seen from a different perspective. Skill, shown in Figure 4 as function of the pressure altitude, is a much better way to evaluate the performance of the three algorithms. Algorithm A has virtually no skill below 200 mb. Algorithm B has skill from 900 mb throughout the atmosphere, while algorithm C has skill from 800 and 200 mb. Given the accuracy in the knowledge of the sea surface temperature, and its small deviation from the background, it should not be surprising that none of the algorithm showed skill very close to the surface. While this evaluation is only for one day, 20020906, albeit with about 50,000 soundings in the tropical oceans, the evaluation 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.

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 is 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 70% yield based on internal quality indicators designed to improve the rms accuracy, would the skill be improved? What is the retrieval skill for water vapor? What makes the skill of algorithm B so much better than C? How does the retrieval skill of IASI compare to that of AIRS? These evaluations are in progress, but are beyond the scope of this paper.

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 skill, and such a system cannot have impact on the accuracy of the weather forecast. 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 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 close to 1 K rms "RAOB quality" accuracy, but only one has skill between the surface and 100 mb.

Acknowledgments

The work described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration and at MIT/Lincoln lab under Air Force contract FA8721-05-C-0002.

References

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

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

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

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

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

Divakarla M. G., C.D. Barnet, M.D. Goldberg, L.M. McMillin, E. Maddy, W. Wolf, L. Zhou, X. Liu (2006) JGR Vol.111, D09S15, doi:10,1029/2005/JD006166, 2006.

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

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

Kalnay, E.D. et a. (1996) "The NCEP/NCAR 40-year Reanalysis Project" Bull.AMer.Meteor.Soc.77, 437-471.

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 (2006) "Improving G;obal Analysis and Forecasting with AIRS" BAMS, July 2006, pp. 891-894.

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