Data Compression:
A NOAA view of its Data Reduction Benefits for Distribution and Archiving

NOAA-WP-27 provides an overview of the data compression activities NOAA is investigating to support data dissemination and archiving in the future. Considering the increase in data rates and data volumes, new advances in data compression are needed to meet the demands for real-time distribution of environmental information. NOAA has undertaken some four plus years of research on Satellite sensor Earth science data compression to explore how to accurately reduce data volumes. NOAA sees Earth science data compression, in the emerging near future of more powerful satellite sensors and enlarging data rates, as a tool to support CGMS members' data distribution goals. This research was undertaken to explore how to implement new data compression techniques to Earth science sensor data, and how a level of reduction in data volume was achievable. NOAA believes the higher the reduction rate, the more data satellite operators can distribute to world nations.

NOAA discovered the compression of satellite Earth science data to be a unique problem generated by unique sensor data patterns. As a rule, new classes of compression algorithms, such as that coming out of NOAA's funded research, are outperforming the ability of conventional compression algorithms by almost a two to one factor when applied to this problem. Where conventional algorithms that might achieve a 1.5 to 1 or 2 to 1 "Lossless" reduction in data volume when applied to satellite Earth science data; new classes of algorithms, such as those emerging from NOAA research, are yielding reductions or 3;1, 4:1 and 5:1.

Considering the impact of the growth in future satellite sensor data volumes, NOAA promotes the exploration and adoption of new compression algorithms to Earth science satellite data.

Data Compression: A NOAA view of its Data Reduction Benefits for Distribution and Archiving

1. INTRODUCTION

NOAA sees Earth science satellite sensor data compression, through its reduction in data levels, being able to serve a number of important purposes for the Coordination on Geostationary Meteorological Satellites (CGMS) member nations. Data compression can do the following:

- Reduce costs of data distribution, be it by satellite or terrestrial land lines.
- In cases of limited RF spectrum availability, for meteorology and Earth exploration satellite data transmission use, enable a greater level of Earth science data be distributed to world nations.
- Reduce Earth science data archive costs.

Distributing maximum levels of satellite Earth science data and doing so at low cost to world nations is consistent with many policy goals. Such a data distribution is consistent with the data distribution goals of the World Meteorological Organization (WMO); and Agreements reached at the Earth Observation Summit (EOS) and by EOS signatory nations to establish the Global Earth Observation System of Systems, known as GEOSS.

2. Compression of Satellite Earth Science Sensor Data

NOAA believes its compression research program on satellite Earth science data reduction has achieved important results. It has shown compression of satellite Earth science sensor data to be a unique program, defined by the uniqueness of the data structure itself. It has shown that new compression algorithms, introducing specialized fields of mathematics to this problem, can yield significantly greater data reductions then possible from adopting conventional popular standardized compression algorithms (e.g. JPEG, JPEG 2000, GZIP, EZW).

Data compression is a mathematical way to represent digital data with as few bits as possible. Compression algorithms thus reduce the volume of the data; with the data then being restored either fully or partially before use depending upon the algorithm design (Lossless, near Lossless, Lossy compression approaches). Compressed data must be decompressed before it can be used. Compression of digital data is not a new concept. It is widely used in many elements of today's digital era and in a way has been brought about by today's digital era of Personal computers, internet, and global communications. It is widely used in the internet through the widely known JPEG and JPEG 2000 compression standards developed for compressing two-dimensional picture imagery for internet and PC use. Compression standards such as JPEG and JPEG 2000 rely on the Discrete Cosine Transform (DCT) and wavelet transforms of mathematics, respectfully. Compression of data volumes has become widely used in the medical field to reduce the burden of storing digital imagery records as well as readily enabling digital imagery to be transmitted for viewing, diagnosis, and consultations over long distances. Compression algorithms for reducing medical digital data volumes have traditionally relied on the mathematics of vector quantization.

NOAA satellite Earth science sensor data compression research to date has been directed at two types of widely used visible and IR sensors. The first type is the satellite multispectral imager such as Japan's MTSAT imager, the NASA MODIS imager, and the future GOES R satellite series ABI imager. The second type is the hyperspectral sounder, being of two classes. The first class of sounder is a grating spectrometer such as the 2378 channel NASA EOS Aqua AIRS sounder. The second sounder class is the Michelson Interferometer such as the CNES developed IASI sounder on the EUMETSAT METOP satellite.

To date all of our NOAA research has centered on what is categorized as "Lossless" data compression. By definition, compression is "Lossless", if on reconstruction or decompression, the original data has been exactly restored. There are no bit errors, at all. NOAA may eventually in its research explore a "near Lossless" compression, depending upon funding availability. With near Lossless compression some of the original data is lost on decompression. In some cases, this might be judged by the science community acceptable depending upon the impact.

We wish to highlight some key findings to the CGMS member nations coming out of our NOAA research of compression of satellite Earth science sensor data. We do this with the 12 bullet list below as well as 5 abstracts from recent journal publications of NOAA sponsored compression research:

1. We find compression of Earth science satellite sensor data to be a unique problem. To achieve maximum possible data volume reductions through compression, new unique data compression algorithms are required.
2. Satellite Earth science sensors are found each have unique data patterns; which directly relate the ability for compression.
3. Earth science sensor data patterns vary among sensor types such as between sounder (hyperspectral) and imager (multispectral) type sensors
4. Data patterns vary between sounder types such as between Michelson Interferometer type sounders or grating spectrometer type sounders. To date we have been exploring separate compression algorithms for Michelson interferometer sounders and Grating spectrometer sounders.
5. Data patterns vary between instrument designs and within a type. To date we have been designing separate compression algorithms for each type of multispectral imager- JMA's MTSAT, U.S. NASA MODIS, Meteosat-8 SEVERI 12 channel imager.
6. Each varying sensor data pattern can lead to the need for a unique new compression algorithm to achieve maximum data compression, with varying possible levels of Lossless compression.
    - NOAA in its research has developed distinct algorithms for each class of sensor type (e.g. Michelson Interferometer hyperspectral sounder, grating spectrometer hyperspectral sounder, multispectral imager. Further the algorithms at present are directed at a given sensor design.
7. Compressing Earth science sensor data for any satellite transmission requires that additional mathematics be introduced to provide robustness against the relative high electronic noise environment of wireless communications and that for satellites this be synergistic with the "Forward Error Corrections"

coding typically used in a satellite communication system such a Turbo Product code or Low Density Parity Check (LDPC) code. Failure to do so, could lead to loss of all data, partial loss, or striping on data decompression at the receive site. We find for example the JPEG 2000 compression standard is very susceptible to data stream errors in satellite use; a vulnerability that can lead to loss of all data.

8. Lossless Compression results to date expressed in terms of compression ratio from NOAA research on satellite Earth science sensor data, with developed custom designed algorithms are:

   - New working Lossless compression algorithms (two have been developed) for **AIRS** grating spectrometer **sounder** sensor compression - 3.7 and 4.0 to 1 for two separate algorithms. At this compression ratio neither result contains error containment for satellite use.
     i. With robust error containment a CR of 3.1 to 1.
   - New working Lossless compression algorithm for Michelson interferometer sounder sensor using aircraft flown NASTI sensor yielding a CR of 5.0 to 1.
   - Multispectral Imager compression:
     i. Meteosat-8 SEVERI using 11 or its 12 channels - Research program “Shannon entropy” estimate of 3.1 to 1 for maximum possible Lossless CR.
     ii. NASA Moderate Resolution Imaging Spectroradiometer (MODIS) 36 channel imager- NOAA R&D program’s “Shannon entropy” estimates defining arrange of potential Lossless compression ratios (CR) ranging between 2.25 and 3.25 to 1.
     iii. Developed a new working Lossless compression algorithm tailored to the JMA MTSAT imager 600Kbps data stream level yielding a Lossless CR of 2.9 to 1.

9. The poor performance of conventional compression algorithms on sounders can be seen by their Lossless compression results when applied to Earth science sensors such as the AIRS sounder and then contrasted with the NOAA results in the immediate above paragraph 8:

Conventional Compression Algorithm type Lossless compression Aaplied to NASA AIRS sounder data	**Compression Ratio (CR) when applied to AIRS data**
JPEG 2000 standard applied 3 dimensionally	**2.37 to 1**
JPEG LS	**2.46 to 1**
NASA-GSFC “USES Algorithm	**1.97 to 1**
GZIP	**1.89 to 1**
SPIHT	**1.97 to 1**
EZW	**1.95 to 1**

These above tabularized compression research results (CRs) are published in a journal paper by the SPIE society, titled “Wavelet Lossless Compression of Ultraspectral Sounder Data”. The authors are: Joan Serra-Sagristà and Fernando García-Vílchez, Department of Information and Communications Engineering ETSE, Universitat Autònoma Barcelona, SPAIN

Julià Minguillón and David Megías, Estudis d'Informàtica i Multimedia, Universitat Oberta de Catalunya, SPAIN
Bormin Huang and Alok Ahuja, CIMSS/SSEC, University of Wisconsin, Madison,USA

The paper is available from both SPIE and Amazon within a published book of edited papers or singularly as a paper from SPIE: "Satellite Data Compression, Communications And Archiving" (Proceedings of SPIE, August 2005 annual meeting, San Diego , CA), by Bormin Huang (Editor), Roger W. Heymann (Editor), Charles C. Wang (Editor

10. Undocumented but reported Lossless compression ratio of NASA "USES" compression algorithm applied to GOES –R series ABI data 1.4 to 1.

11. We want to show that compression of Satellite Earth science data is a new and unique problem. Yes, Lossless compression and its data reduction can be obtained by applying conventional algorithms to satellite Earth science data developed for other data types such as JPEG 2000 for 2-dimensional picture imagery. However, we find through our research that a compression ratio almost double that of conventional algorithms can be obtained from algorithms designed around the Earth science data patterns. That is what has led to date to Lossless compression ratios of ~ 3.0 to 1 for Multispectral imagers, 4 to 1 for grating spectrometer sounders, and ~5 to 1 for Michelson Interferometer sounders.

12. Below are presented 5 abstracts for 5 presentations and journal papers submitted to SPIE for its Aug 17-22, 2006

Below 5 selected abstracts associated with 5 published papers identifying recent 2006 results of NOAA's funded satellite Earth science sensor data compression research are provided. The results of the NOAA funded research reflected in the abstracts were presented at the 2006 August 13-17, 2006 SPIE annual meeting, in its OEI106 conference. An edited book of all papers at the OEI106 conference, including the journal for the 5 below abstracts can be obtained through both SPIE and Amazon.com in the following published book: " Satellite Data Compression, Communications and Archiving", by Roger W. Heymann (Author), Charles C. Wang (Author), Timothy J. Schmit (Author).

"A lossless compression algorithm for hyperspectral data"

I. Gladkova [1]**, M. Grossberg** [1]
1 CCNY, NOAA/CREST, 138th Street and Convent Avenue, New York, NY 10031
SPIE Meeting August 13-17, 2006. Meeting Journal paper **6300-1**

ABSTRACT
In this paper, which is an expository account of a lossless compression techniques that have been developed over the course of a sequence of papers and talks, we have sought to identify and bring out the key features of our approach to the efficient compression of hyperspectral satellite data. In particular we provide the motivation for using our approach, which combines the advantages of a clustering with linear modeling. We will also present a number of visualizations which help clarify why our approach is particularly effective on this dataset. At each stage, our algorithm achieves an efficient grouping of the data points around a relatively small number of

lines in a very large dimensional data space. The parametrization of these lines is very efficient, which leads to efficient descriptions of data points. Our method, which we are continuing to refine and tune, has to date yielded compression ratios that compare favourably with what is currently achievable by other approaches. A data sample consisting of an entire day's worth of global AQUA-EOS AIRS Level 1A counts (mean 12.9 bit-depth) data was used to evaluate the compression algorithm. The algorithm was able to achieve a lossless compression ratio on the order of 3.7 to 1.

"Adaptive VQ-based Linear Prediction for Lossless Compression of Ultraspectral Sounder Data"

Bormin Huang 1, Alok Ahuja 1, and Mitchell D. Goldberg 2
1 CIMSS, University of Wisconsin-Madison
2 NOAA/NESDIS, Satellite Applications and Research
SPIE Meeting August 13-17, 2006. Meeting Journal paper **6300-2**

ABSTRACT
Contemporary and future ultraspectral sounders represent a significant technical advancement for environmental and meteorological prediction and monitoring. Given their large volume of spectral observations, the use of robust data compression techniques will be beneficial to data transmission and storage. In this paper, we propose a novel Adaptive Vector Quantization (VQ)-based Linear Prediction (AVQLP) method for ultraspectral data compression. The method is compared with several state-of-the-art methods such as CALIC, JPEG-LS and JPEG2000. The compression experiments show that our AVQLP method is the first to surpass the 4 to 1 lossless compression barrier for a selected set of AIRS ultraspectral sounder test data.

"Preliminary Lossless Compression Results with Michelson Interferometer Data"

Timothy J. Schmit
NOAA/NESDIS, STAR (Satellite Applications and Research),
Advanced Satellite Products Branch (ASPB)
Bormin Huang, Y. Sriraja and Hung-Lung Huang
CIMSS, University of Wisconsin-Madison
SPIE Meeting August 13-17, 2006. Meeting Journal paper **6300-4**

ABSTRACT
The next-generation GOES-R (Geostationary Operational Environmental Satellite) HES (Hyperspectral Environmental Suite) Sounder will be either a grating or interferometer design. The HES will be able to provide hourly atmospheric soundings with spatial resolutions of 5 ~ 10 km with higher accuracy than the current geostationary sounder. A number of GOES-R products will be made from the HES data, this information will help both in forecasting and numerical model initializations. Extensive research has been done with lossless data compression with data from a grating-type ultraspectral instrument. NAST-I aircraft data is chosen for testing data from interferometers until IASI (Infrared Atmospheric Sounding Interferometer) and CrIS (Cross-track Infrared Sounder) are available. Preliminary work at CIMSS with

lossless data compression of Michelson Interferometer data achieves compression ratios (CR) above 5.

"Priority-Based Error Correction Using Turbo Codes for Compressed AIRS Data"

I. Gladkova,[1] **M. Grossberg,**[1]
E. Grayver,[2] **D. Olsen,**[2]
N. Nalli,[3] **W.Wolf,**[3] **L. Zhou,**[3]
M. Goldberg[4]
1 CCNY, NOAA/CREST, 138th Street and Convent Avenue, New York, NY 10031
2 The Aerospace Corporation, El Segundo, CA 90245
3 QSS Group, Inc. NOAA/NESDIS, E/RA1 5211 Auth Road, Camp Springs, MD 20746
4 ORA NOAA/NESDIS, E/RA1 5200 Auth Road, Camp Springs, MD 20746
SPIE Meeting August 13-17, 2006. Meeting Journal paper **6300-8**

ABSTRACT
Errors due to wireless transmission can have an arbitrarily large impact on a compressed file. A single bit error appearing in the compressed file can propagate during a decompression procedure and destroy the entire granule. Such a loss is unacceptable since this data is critical for a range of applications, including weather prediction and emergency response planning. The impact of a bit error in the compressed granule is very sensitive to the error's location in the file. There is a natural hierarchy of compressed data in terms of impact on the final retrieval products. For the considered compression scheme, errors in some parts of the data yield no noticeable degradation in the final products. We formulate a priority scheme for the compressed data and present an error correction approach based on minimizing impact on the retrieval products. Forward error correction codes (e.g., turbo, LDPC) allow the tradeoff between error correction strength and file inflation (bandwidth expansion). We propose segmenting the compressed data based on its priority and applying different-strength FEC codes to different segments. In this paper we demonstrate that this approach can achieve negligible product degradation while maintaining an overall 3-to-1 compression ratio on the final file. We apply this to AIRS sounder data to demonstrate viability for the sounder on the next-generation GOES-R platform.

"An analysis of optimal compression for the advanced baseline imager-based on entropy and noise estimation" _

M. Grossberg,[1] **S. Gottipati,**[1] **I. Gladkova,**[1]
M. Goldberg, [2] **L. Roytman.**[1]
1 CCNY, NOAA/CREST, 138th Street and Convent Avenue, New York, NY 10031
2 ORA NOAA/NESDIS, E/RA1 5200 Auth Road, Camp Springs, MD 20746
SPIE Meeting August 13-17, 2006. Meeting Journal paper **6300-22**

ABSTRACT
As new instruments are developed, it is becoming clear that our ability to generate data is rapidly outstripping our ability to transmit this data. The Advanced Baseline Imager (ABI), that is currently being developed as the future imager on the Geostationary Environmental Satellite (GOES-R) series, will offer more spectral

bands, higher spatial resolution, and faster imaging than the current GOES imager. As a result of the instrument development, enormous amounts of data must be transmitted from the platform to the ground, redistributed globally through band-limited channels, as well as archived. This makes efficient compression critical. According to Shannon’s Noiseless Coding Theorem, an upper bound on the compression ratio can be computed by estimating the entropy of the data. Since the data is essentially a stream, we must determine a partition of the data into samples that capture the important correlations. We use a spatial window partition so that as the window size is increased the estimated entropy stabilizes. As part of our analysis we show that we can estimate the entropy despite the high dimensionality of the data. We achieve this by using nearest neighbor based estimates. We complement these a posteriori estimates with a priori estimates based on an analysis of sensor noise. Using this noise analysis we propose an upper bound on the compression achievable. We apply our analysis to an ABI proxy in order estimate bounds for compression on the upcoming GOES-R imager.

Selected Compression results extracted from the 6300-22 paper are:

- “We also analyzed the limits of compression based on computing the entropy of the Meteosat-8 imager.”
- “We also obtained a lower bound of 2.6-to-1 on the optimal lossless compression using a linear model applied to this data.”
- “With this novel tool we have estimated that 10x10 windows of 11 channel Meteosat-8 data have a theoretical compression bound of 3.1 to 1. Although this number cannot be directly applied to the ABI, the similarity of a significant portion of the data sets to 11 channel Meteosat-8 should be helpful in accessing the risk associated with ABI designs dependent on significantly higher lossless compression.”

.