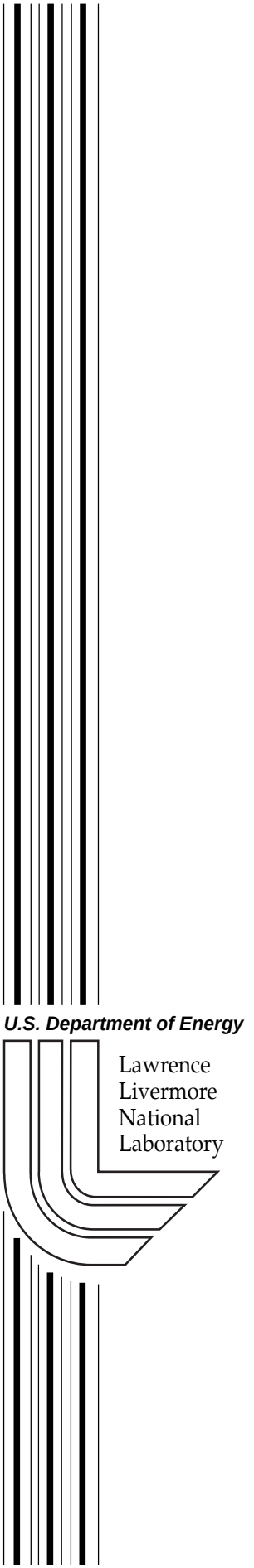
Preprint
UCRL-JC-147225

Multi-Resolution Modeling of a Large Scale Scientific Simulation Data

C. Baldwin, G. Abdulla, T. Critchlow

This article was submitted to: *15th International Conference on Scientific and Statistical, Cambridge, MA USA 07/09/2003 – 07/11-2003*

January 31, 2003



Approved for public release; further dissemination unlimited

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.

This work was performed under the auspices of the United States Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

This report has been reproduced directly from the best available copy.

Available electronically at http://www.doc.gov/bridge

Available for a processing fee to U.S. Department of Energy
And its contractors in paper from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-mail: reports@adonis.osti.gov

Available for the sale to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/ordering.htm

OR

Lawrence Livermore National Laboratory
Technical Information Department's Digital Library
http://www.llnl.gov/tid/Library.html

Multi-Resolution Modeling of Large scale Scientific Simulation Data

Chuck Baldwin, Ghaleb Abdulla, Terence Critchlow
Lawrence Livermore National Laboratory
baldwin5@llnl.gov, abdulla1@llnl.gov, critchlow@llnl.gov

Abstract

This paper discusses using the wavelets modeling technique as a mechanism for querying large-scale spatio-temporal scientific simulation data. Wavelets have been used successfully in time series analysis and in answering surprise and trend queries. Our approach however is driven by the need for compression, which is necessary for viable throughput given the size of the targeted data, along with the end user requirements from the discovery process. Our users would like to run fast queries to check the validity of the simulation algorithms used. In some cases users are welling to accept approximate results if the answer comes back within a reasonable time. In other cases they might want to identify a certain phenomena and track it over time. We face a unique problem because of the data set sizes. It may take months to generate one set of the targeted data; because of its shear size, the data cannot be stored on disk for long and thus needs to be analyzed immediately before it is sent to tape. We integrated wavelets within AQSIM, a system that we are developing to support exploration and analyses of tera-scale size data sets. We will discuss the way we utilized wavelets decomposition in our domain to facilitate compression and in answering a specific class of queries that is harder to answer with any other modeling technique. We will also discuss some of the shortcomings of our implementation and how to address them.

1. Introduction

Multi-resolution techniques, specifically wavelets, have been used for many years as effective modeling tools for data derived from signal and image processing applications [5]. Multi-resolution based paradigms have been shown to be a great promise in knowledge discovery and data mining applications for data obtained from astronomical observation, specifically clustering objects in large scale sky surveys [6]. Another application domain and for fast responses to range sum queries, researchers in [2] have developed a wavelet based approach for approximate query processing. They mapped the data to a relational table, which is compressed and used to resolve select, project, and join operations. A progressive technique, which maps the query, along with the data, to the wavelet domain for query resolution, has been introduced by Shahabi, Chung and Safar [8]. This technique is more like our work but does not a-priori compress the data set to an approximation.

We are using wavelets to model and compress large-scale spatio-temporal scientific simulation data and enabling queries over the resulting compressed model. The targeted data is large-scale multivariate field quantities gathered from scientific simulations [1]. Typical quantities found in these simulations are fundamental or derived physical quantities such as temperature, pressure, velocity, or entropy. We are directly querying the compressed wavelets transform data rather than the original data itself. This means that our queries are posed with regard to the wavelet transform of a temperature field is as opposed to the temperature field itself, for instance. Focusing on the latter would necessitate either mapping equivalent queries to the domain of the wavelet transform or reverse the transform in some intelligent fashion to obtain approximations to or subsets of the field data itself. Querying the wavelet transform data itself has a practical problem associated with it, namely understanding how to think in the domain of the wavelet transform data rather than the intuitive domain of the field data. To address this issue we define specific queries (with associated semantics) that will allow a particular coupled reconstruction from the wavelet transform data. The identified class of queries are motivated by the users' needs to explore some irregularities (or outliers) in the data. For example the user might want to know for what regions of the mesh the temperature changes mostly. It should be noted that we are exploring other modeling techniques which lend themselves to resolving other types of queries, such as range based queries, however for this work we will concentrate on this wavelet based modeling and querying framework.

2. Data Compression

We use an intuitive yet very effective method of compressing data (measured with an l^2 norm), namely keep the coefficients with largest absolute value, weighted by a factor involving their level. A more rigorous development of this idea can be found in [3,4]. This insight into the relation to the coefficients and their

individual contribution to the global error actually gives two methods to store compressed models, see table 1. During the model construction time, we can specify either the target data size or relative error as the criterion for compression.

	Coefficient Selection Scheme
1	Choose sorted coefficients until a calculated specified total number of coefficients is reached, thereby assuring that a prescribed model size is achieved.
2	Choose sorted coefficients until a user specified relative error is achieved, thereby assuring that a prescribed model relative error is achieved.

Table 1: Methods for Compressing a Wavelet Transform

It should be noted that the sorting procedure used in the two methods above has complexity *O(N lg(N))* in the number of coefficients N . This is larger than the *O(N)* time complexity of the wavelet transform itself but an acceptable cost for construction of the wavelet model in our pre-processing stage. From talking to our users, we concluded that they want to be able to analyze the data reasonably fast but not necessarily very accurate. Hence, we identified two important parameters that they can experiment with (at query time) in order to be able to do that, namely, error and time. Error in reconstruction and time of reconstruction are inversely proportional. In order to enable the users to do the analysis effectively we utilized the idea that reconstructing the data using large value coefficients will reduce the error in a nonlinear fashion, hence, we can achieve a reasonable accuracy with a small subset of the total coefficients. This also means that the reconstruction time will be quick, as a result sorting coefficients and allow queries that can be mapped directly to which subset of the sorted coefficients to use, will address our users' needs.The l^2 error is easily computable as the weighted size of the coefficients left out of the selection set.

For multivariate functions, with which we are concerned, there is a dearth of research on the general subject of multivariate or vector multi-resolution analysis. Our solution approach is to incorporate the multivariate analysis solely into the sorting and selection procedure rather than research and develop new multivariate wavelet transforms. To do this we first perform a standard single variable transform on each variable of the data as with regular wavelets. Then if we label the individual transform coefficients with their multivariate component $c = 1 \ldots. m$ as : $\{\{d^l_{c_{i,j}}\}^l_{j=1}\}^L_{l=1}$

and form an equivalent to the real multivariate transform coefficient as : $d^l_j = (d^l_{1;j},....,d^l_{m;j})$

for each level $l = 1,...,L$ and level index $j = 1,...,J_l$. We then use a weighted multivariate norm:

$$|| (v_1,...,v_m) ||^2 = \sum_{i=1}^{m} w_i \mid v_i \mid^2$$

To find a size (or importance) estimate for the coefficients and use that as a sort key. The weights w_i are positive and
have the property that :

$$\sum_{i=1}^{m} w_i = 1$$

The simple weights $w_i = \frac{1}{m}$, giving equal weight to each component, is the most natural choice and currently the one we use. However it is not illogical or difficult to use some statistical measures of the coefficients themselves to derive a more appropriate non-linear weighting scheme. Once the coefficients are chosen the resulting coefficients along with their significance ordering obtained from the sorting are saved to disk. This compressed model represents a starting point upon which ad-hoc queries are performed.

3. Queries on the Compressed Data

The reconstruction of an approximate representation of the original data in the query resolution phase in AQSIM [1] is performed under more interactive time constraints. As mentioned we store the wavelet coefficients in the model with their sorting order and utilize this information to provide some additional query processing for the end user. This information can also be incorporated into a progressive reconstruction which is controlled by the user and provides a more visual metric to conclude when a reconstructed approximate is "good enough". The queries that we provide are ultimately queries about the quality, quantity, or possibly spatial location of the stored wavelet coefficients themselves. This approach makes the data compression a discovery process, where the compression hopefully removes unwanted noise or homogenizes redundant information so that discovery of useful facts can be achieved. This connection between compression and knowledge discovery has been noted by Ramakrishnan and Grama [7].

The collection of wavelet queries that we are currently working on is shown in table 2. The figure describes in words the semantics of the queries we are interested in. The first query will use the complete model of the data to build the best approximate to the original data. The second query in uses the pre-sorted coefficients to reconstruct an approximate to the original data with a user

specified percentage of the available data. The third query uses the pre-sorted coefficients to reconstruct an approximate to the original data with a specified relative error (as measure against the original data). The second and third queries can be implemented in a progressive fashion; namely, the coarse scale smooth data can be displayed to the user and as the sorted coefficients are added back to the approximate data, the display of the data can be updated to reflect this gained accuracy. This process can also be interruptible. The progressive display or interruptibility is another of our long-range research goals for the ad-hoc query system. The fourth query uses the coefficients from the most important levels to reconstruct an approximate to the original data. By computing the combined total of the weighted coefficients on the different levels of the compressed model a relative merit for adding each levels coefficients can be compared and used. The fifth query in represents a point wise reconstruction, again using the representation formula

	Description of the Wavelet Model Queries
1	Reconstruct using all the coefficients from the stored wavelet decomposition.
2	Reconstruct by further choosing the most significant wavelet coefficients based on a user-supplied percentage.
3	Reconstruct by choosing the wavelet coefficients that produce an approximate with a user supplied relative error.
4	Reconstruct using only the most significant levels of the wavelet decomposition in the model file.
5	Reconstruct using coefficients that effect a given spatial location.

Table 2: Queries Relevant to the Wavelet Model

4. Examples

Our first example is a simple univariate time series. Recall that our target data is multivariate but the same ideas and algorithms hold in the univariate case. Figure 1 shows the values of the Standard & Poors 500 for the year 2001 with respect to the stock market trading day. We first perform a wavelet transform using a simple Haar orthonormal wavelet. We next create the compressed data file; a 50 percent compression ratio is established by choosing to store only half of the wavelet coefficients. This results in a compressed approximation with 0:337% global relative (l^2 error), figure 2 shows what that compressed time series looks like by simply uncompressing it. Figure 3 is using about 33% of the original coefficients (about 66% of the compressed coefficients) and the resulting reconstruction has a global relative error of .590%. All of these simple univariate time series examples show that it is not difficult to achieve good compression with the approximation and still retain much of the important characteristics of the data, with respect to variation and change.

In order to test this procedure on actual data we used simulation data of a can being crushed. We again show results using the familiar Haar orthonormal wavelet system. This data is time dependent and has about 13 independent variables per grid point. We show only a pressure field from the procedure due to space constraints although other fields show similar behavior. Figures 4 shows the original uncompressed can at the 1st time-step of the simulation. Figure 5 shows the compressed can at the same time-step using 33% compression. Figure 6 shows the compressed can at the same time-step using 66% compression. The results show that while simulation data does not have the same simple reconstruction behavior of the univariate example above, it is still possible to reconstruct data values and keep the kind of variational character in the results.

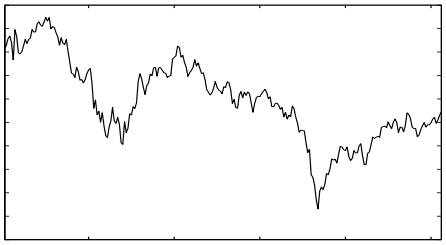


Figure 1: Original Data set

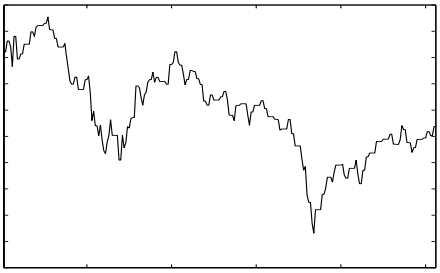


Figure 2: Compressed approximation using 50% of the coefficients.

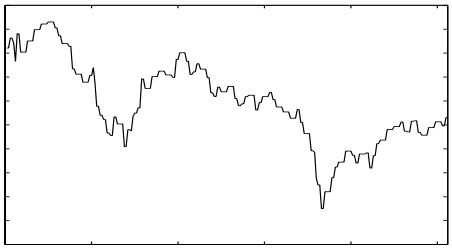


Figure 3: Reconstruction Using 33% of the coefficients.

4. Comments And Conclusions

Our research adapts and extends ideas of wavelet theory to multivariate data, and formulates methodologies and algorithms for compressing the wavelet coefficients resulting from that work. We also devise ways in which users can effectively query the resulting compressed data in an intuitive fashion without understanding too many of the wavelet specific details. In our initial experimentation we have found that using wavelets to decompose, compress, and reconstruct data yields results that are helpful in analyzing the dynamic portions of simulation data. The wavelets are attuned, in various degrees, to smoothness in data. Compressing by keeping only the largest coefficients implies that the reconstruction will be accurate around areas where the data is not smooth, i.e. highly dynamic. Currently we are working on the scalability of our system to speed some of the described algorithms in order to work with tera-scale data.

5. Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405- Eng-48. The authors would like to thank Kevin Durrenburger, Tina Eliassi-Rad, Roy Kamimura, Edward Smith, and Nu-Ai Tang from the Center for Applied Scientific Computing at the Lawrence Livermore National Laboratory for their advice and assistance regarding this research.

6. References

[1] G. Abdulla, C. Baldwin, T. Critchlow, R. Kamimura, I. Lozares, R. Music, N. A. Tang, B. S. Lee, and R. Snapp. Approximate ad-hoc query engine for simulation data. In *Joint Conference on Digital Libraries JCDL-01*, pages 255-256, June 2001.

[2] K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim. Approximate query answering using wavelets. *VLDB Journal*, 3, 2001.

[3] R. DeVore, B. Jawerth, and B. Lucier. Image compression through wavelet transform coding. *IEEE Trans. Image Processing*, 38:719-746, 1992.

[4] R. DeVore, B. Jawerth, and V. Popov. Compression of wavelet decompositions. *American Journal of Mathematics*, 114:737-785, 1992.

[5] S. Mallat. A Wavelet *Tour of Signal Processing*. Academic Press, 1998.

[6] F. Murtagh, J.-L. Starck, and M. W. Berry. Overcoming the curse of dimensionality in clustering by means of the wavelet transform. *Computer Journal*, 43(2):107{120, 2000.

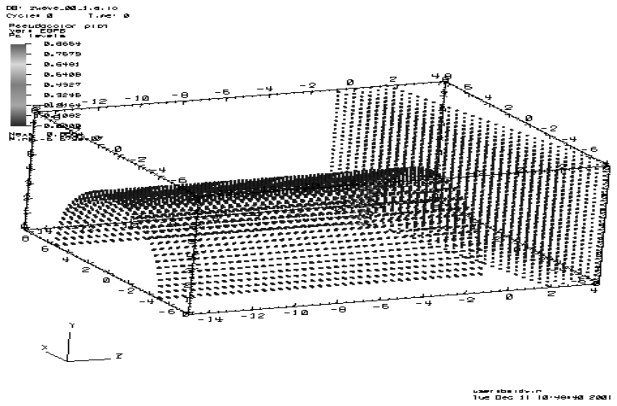


Figure 4: Original mesh from the can data

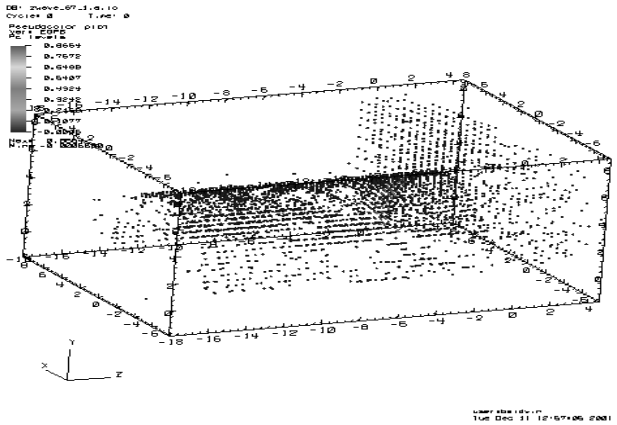


Figure 5: Can data with 33% compression

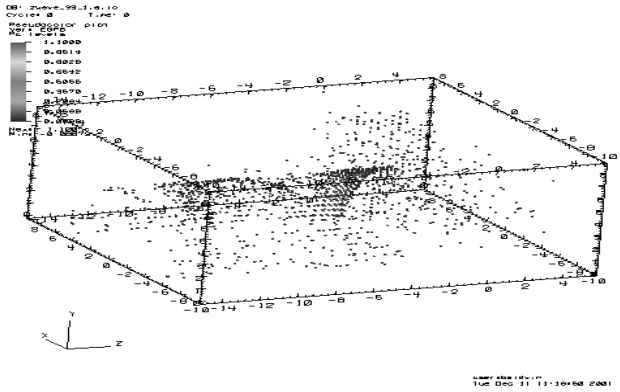


Figure 6: can data with 66% compression

[7] N. Ramakrishnan and A. Grama. Mining scientific data, *In Advances in Computers*, pages 119-169. Academic Press, 2001.

[8] C. Shahabi, S. Chung, and M. Safar. A wavelet-based approach to improve the effieciency of multi-level surprise mining. In *PAKDD International Workshop on Mining Spatial and Temporal Data 2001*, 2001.