Processors, Pipelines, and Protocols for Advanced Modeling Networks

Joseph C. Coughlan
Mail Stop 242-4
NASA Ames Research Center
Moffett Field, CA 92035-1000

Eric P. Bjorkstedt
Santa Cruz Laboratory, NOAA Fisheries
110 Shafer Road
Santa Cruz, CA 95060

***Abstract*— NASA's Earth Science Enterprise has established the goal of developing a predictive capability for the Earth System. NASA uses the vantage point of space to provide information about Earth's land, atmosphere, ice, oceans, and biota that is obtainable in no other way. To enhance predictive capabilities, NASA is planning a sensor web to collect data across a range of spatio-temporal scales. The end-to-end process of data collection, data assimilation, biogeophysical modeling and prediction is inseparable and predominately enabled by software. Software transforms the raw data into usable products and information and software disseminates these products to end-users. New information system technologies are needed to enable better prediction, flexible data assimilation and model coupling to build integrated Earth system models. Advancement of our modeling capabilities will require not only faster processing, but new programming methods, new algorithms, high-speed data pipelines, and interoperable architectures that allow the networking of diverse Earth System models.**

I. INTRODUCTION

NASA's Earth Science Enterprise Missions contribute to the multi-agency United States Global Change Research Program (USGCRP) and integrate with international scientific activities such as the World Climate Research Program and the International Geosphere-Biosphere Programme. The overall mission of NASA's Earth Science Enterprise (ESE) is to enable improved prediction capability for the highly integrated and dynamic Earth system by developing a scientific understanding of the Earth system and its response to natural or human-induced changes. The USGCRP is also supported through NASA's establishment of a long term commitment to improved Earth Science monitoring and prediction that relies heavily on air and space borne observations (data) of the Earth as well as field validation campaigns.

These data are spatio-temporally heterogeneous, collected by a variety of instruments and of immense volume and dimensionality. The raw data require extensive processing prior to their use as biophysical parameters in support of scientific inquiry, policy formulation and resource management. The processing of large data holdings requires significant computational resources and scarce expertise in the interdisciplinary Earth and computational sciences. There are significant information system challenges posed by mission science requirements.

In this new century there will be increasing societal needs for seasonal and interannual climate predictions, for environmental assessments, and for sophisticated model-assimilated data sets to aid in the quantification of the biogeochemical processes that determine the balance of environmental parameters (i.e. temperature, water, winds, ozone, productivity, etc.). Many projects meeting these goals have grown out of basic research activities that have expanded to take on the responsibility of providing products to a broader community [1]. Sensor-web observations (data) and science models will need to be rapidly and cost effectively transferred into networked production systems to build products and assessments for public and private use.

II. MODELING

Models are essential tools for the development of scientific understanding. Modeling refers to those activities involved in building, applying, and validating biogeophysical models in software. The on-going refinement and development of new computational models that can simulate the dynamics of the Earth system is critical for the USGCRP. Models can be used in hind casting experiments to test hypothesis of how the Earth system behaves; and, models can be run in predictive mode to simulate the response of the Earth system to scenarios of future forcing and feedbacks on these forces by Earth system responses. These simulation activities are critical to providing the environmental assessments used to synthesize Earth science results and provide information to policy and decision makers [1].

For purposes of discussing information technology, modeling can be roughly categorized as discovery or production. Discovery modeling typically is PI led, often funded with grants, and oriented towards discovery and scientific inquiry of fundamental mechanisms within a discipline. Production modeling is community led, interdisciplinary, and driven by the need to generate standard products, such as forecasts and monitoring the productivity of the Earth. Often called high-end modeling, it requires larger

teams, budgets and computing resources, and often undertakes the unconstrained modeling of long-term scenarios.

A. Discovery Modeling

Discovery modeling finds new relationships and interactions by dividing, isolating and understanding specific components of the Earth System. This incremental approach allows scientists to cope with the complexity of environmental phenomena, target specific important processes, and adapt solutions to available computational resources. Standard data products from other domains are used to interface with the system being studied. Data may be obtained with a network link to a source, but more often input data reside locally and there is little need to access current data in real-time. The models are mechanistic and are built to convey understanding. A mechanistic model computes intermediate variables and states that correspond to the measurable system, and when compared to observations, help verify model behavior. The system being studied is decomposed into basic empirical or geophysical relationships. These basic algorithms reproduce behavior observed in the data and, when integrated, implicitly represent hypotheses about the system.

1) Data volume and dimensionality

The process of analyzing data to discover relationships for modeling is labor intensive and costly. Scientists are overwhelmed by the sheer volume of data collection enabled by modern technology [2]. A sensor–web will increase the potential for more data collection therefore, software is essential to aid the scientist and reduce data volume. Any algorithms used to aid the scientist and reduce data volume must originate in the discovery process before they can be implemented in data production or moved on-board.

2) Knowledge discovery in databases

Knowledge Discovery in Databases (KDD) is an interdisciplinary activity that automates the identifying of patterns and structure in large databases for discovering novel features and algorithms [3]. KDD involves visualization, artificial intelligence, knowledge management, and data mining, which in turn involves database statistics and machine learning. This process is incremental and begins with (1) selection of hypotheses or phenomena to explore, (2) cleansing and preprocessing data to fill in missing observations, account for noise in the data and cope with time series, (3) the often necessary reduction of the size and dimensionality of data, which may include extracting high-level features in place of using raw data (e.g. extract and mine the trajectory of cyclones instead of mining the imagery). Subsequent data mining is a complex process where the science goals must match the data, the mining method and the desired form of the new algorithm (regression, decision tree, classification rules, etc). After mining, patterns must be interpreted (animated, plotted, visualized, etc.), and the KDD steps repeated to refine, concisely summarize and catalog the results in relation to existing knowledge (and check for consistency with existing knowledge) [3].

Feature extraction can transform massive, low-level data into smaller, higher-level feature descriptions such as replacing a series of high-resolution images with the (x, y) tracking of a cyclone eye over time. Feature extraction methods are candidates for space-based processing when the goal is to reduce the data volume prior to transmission or to task the sensor-web to autonomously collect higher resolution data in the path of the cyclone to improve the inputs to storm-tracking models. Therefore, specialized algorithms or biased sampling schemes may be needed to ensure capture of rare but important classes of data that may occur with very low probability and appear as noise to more general algorithms [3]. These algorithms can be run as data are collected, as data stream into the archive or as data are serendipitously extracted from the archive to fulfill an order. Scalable, crafted algorithms are necessary to operate on large databases and must be understandable since findings need to be interpreted as knowledge or explained [3].

3) Automated programming

Low-level languages (e.g. FORTRAN, C++, JAVA) impede progress because they require the problem description to be expressed with the problem solution and detailed control structures such as loops and parallelism [4][5]. Graphical based programming reduces program complexity but is still problematic. Even with dedicated libraries and middleware, there are limits to the degree these kinds of methods can facilitate the construction of software without further aids [6].

Researchers have developed prototype techniques that can automate the construction of efficient data analysis software from a concise problem description expressed in code [7]. Given a problem where data sources are intermixed and signals to be identified, separated and modeled, such systems can automatically separate the data and find optimal algorithms, either symbolically or numerically, reproducing data from each source [8]. These systems also simultaneously generate compliant, detailed documentation and efficient, maintainable code implementing these solutions. One line of code describing the problem translates, on average, to 30 lines of C describing the algorithm solution [4].

B. Production models

The technologies that support the discovery process are germane to production modeling however production systems pose additional challenges. As successful discovery activities mature, they can provide products to assess impacts on the system of study. Today there are serious deficiencies in the science community's ability to provide the necessary products for climate assessment [1]. Barriers include processing limitations, the difficulty of integrating new scientific models, competing for scarce talent with the commercial sector, costly

Software-Practice and Experience, 30:1541-1570, 2000.

[6] J. Thomas, D. Batory, V. Singhal, and M. Sirkin. “A Scalable Approach to Software Libraries”. In *Proceedings of the 6th Annual Workshop on Software Reuse* , Owego, NY,, Nov. 1993.

[7] B. Fischer and J. Schumann. “AutoBayes: A System for Generating Data Analysis Programs from Statistical Models”, Submitted for publication in J. Functional Programming, 2001.

[8] W. L. Buntine, B. Fischer, and T. Pressburger. ``Towards Automated Synthesis of Data Mining Programs". In S. Chaudhuri and D. Madigan, (eds.), *Proc. 5th Intl. Conf. Knowledge Discovery and Data Mining*, pp. 372-376, San Diego, CA, August 15-18 ACM Press. 1999.

[9] D. Cooke, “SequenceL for the Information Power Grid”, *1999 IEEE Symposium on Application-Specific Systems and Software Engineering & Technology*, 24 - 27 March, 1999, Richardson, TX, 1999.

[10] I. Foster and C. Kesselman, “*The Grid: Blueprint for a New Computing Infrastructure”*, Morgan Kaufmann, Nov. 1998.