

## CHAPTER 1B

### A RATIONAL FOR A CENTRALIZED DATA REPOSITORY FOR MODELING

(Task 3.1B)

Vibhas Aravamuthan<sup>1</sup>

<sup>1</sup>*Louisiana State University*

#### 1B.1 Overview

Mathematical models are an abstraction of reality and as such range from simple box models to sophisticated models based on partial differential equations. All these models require field data for their calibration and verification, and the accuracy and predictability of these models rely heavily on the availability of such data. The data requirements for mathematical modeling are summarized below:

1. **Driving forces:** All mathematical models have physical variable, which in turn may directly or indirectly be dependent on other physical variables. For example in a hydrodynamic model the dependent variables modeled include water levels, currents, salinity, sediments etc. These variables in turn depend on other physical variables like wind speed and direction, rainfall, temperature, evaporation etc. A model's accuracy and predictability would then depend on the availability of data for all these variables. Although some of these variables can be parameterized one would still require some data to check the validity of these parameterizations.
2. **Spatial distribution:** Spatially distributed models of a dependent variable also need field data at various locations within the model domain. These gages should be spatially distributed in such a way that major variations within the model domain are captured accurately.
3. **Boundary conditions:** Spatially distributed models need field data as boundary conditions. The accuracy of these models largely depends on the extent and accuracy of the boundary conditions that are used to drive them. Most model inaccuracies can be attributed to inadequate or inaccurate boundary condition data.
4. **Sampling interval:** All mathematical models resolve only certain frequencies occurring naturally because of the spatio-temporal approximations used. Mathematically the frequencies resolved for a sampled variable depends on the sampling interval and the length of the sample. If a variable in the field is sampled at an interval of  $dt$  and if there are  $N$  samples then the highest frequency resolved is given by  $1/(2*dt)$  and the lowest frequency resolved is given by  $1/(N*dt)$ . It is therefore important that the field sampling of a particular variable capture the predominant frequencies occurring in nature. This then translates to a sampling interval fine enough to capture frequency ranges where large variations in a variable occur.
5. **Sampling duration:** Data should be sampled for a relatively long period of time so that meaningful statistical properties of the data can be obtained. Ideally the variable should be sampled during both low and high frequency events so that the data can be used to study the ability of a model for predicting these events.

6. **Missing Data:** Mathematical models especially unsteady models use a forward in time integration method whereby the time domain is discretized using small time increments. This is especially true in models, which are based on the discretization of partial differential equations. Large gaps in the data would degrade the modeling accuracy.
7. **Accuracy:** The accuracy of the sensor collecting the data needs to be known and should be able to capture typical variations in that variable. For example a gage measuring water level should be accurate at least to the nearest centimeter. If it is not, then it would be useless from a modeling perspective.
8. **Standardized data formats:** Normally large data sets are stored in binary formats in order to conserve space. Raw binary formats are architecture dependent as in whether a particular hardware architecture is little endian or big endian. Recognizing this problem there are a number of machine independent data formats for the storing, and dissemination of scientific data have been developed. The most popular formats are the NetCDF (Network Common Data Format) developed by NCAR (National Center for Atmospheric Research) and HDF (Hierarchical Data Format) developed by NCSA (National Center for Supercomputing Applications). NetCDF is the format adopted by the oceanographic, climate and meteorological modeling communities.

A number of federal and state organizations are involved in collecting environmental data in the coastal areas of Louisiana. Most of these efforts are targeted towards monitoring projects proposed by these organizations in order to gage their effectiveness. Although the data gathered for these efforts would be useful to environmental modelers to design, calibrate and verify new or existing models, a number of drawbacks exist in the effective utilization of these data for modeling. These include:

1. **Critical environmental variables are not sampled:** Since a mathematical model requires as driving forces a number of environmental factors a lack of data for a primary variable used as a driving force would result in poor predictions. For example a critical driving force for a coastal hydrodynamic model are wind speed and direction. Currently there are very few stations which record wind speeds and direction in the coastal area. Most data is available only at airports which are located on land far from the coastal areas. Since most of the near shore dynamics is driven by wind and tidal forces the lack of wind data has proved to be a big stumbling block in developing effective coastal hydrodynamic models.
2. **Spatial distribution of gages:** Most of the gages in the coastal zone have been deployed from a management perspective of specific projects or for flood control purposes. This has then resulted in either data rich areas or extremely data poor areas. From a modeling perspective a uniform spatial distribution would help in the development of models covering large spatial domains.
3. **Sampling duration:** Since most gages have been deployed for specific management perspective, these gages are decommissioned as soon as the mandate expires. This then results in large data gaps in particular areas.
4. **Sampling rate:** Although the sampling rate for a particular variable may be adequate from a management perspective, it may not be sufficient from a modeling perspective as it may not capture the predominant time scales of a particular variable.
5. **Lack of uniform metadata and formats:** Different agencies use different metadata formats for the description of the data sets they provide. This then results in modelers spending considerable amounts of time bringing the data sets from different organizations into a uniform format which can be then used in their models.

6. **Lack of a central repository:** Since datasets are collected by several agencies modelers are sometimes not even aware that a particular dataset exists.
7. **Lack of dialog between modelers and agencies:** Currently there is no mechanism that exists by which modelers can coordinate with different agencies for their data requirement needs.

An effective dialog between the data gathering entities and the modeling community would enhance the usefulness of these data sets to modeling community. A feedback mechanism between the modelers would then help the monitoring organizations to optimize further their data collection operations. In order to achieve these objectives we propose the development of a centralized data repository under CLEAR to which modelers can turn to for their data needs

## 1B.2 Objectives of CLEAR

The objectives of CLEAR are to establish an archive of data collected by different federal and state agencies. The data collected by CLEAR would augment the data collected by these agencies in the following ways:

1. **Data Portal:** Develop a data portal from which modelers can obtain their data over the Web.
2. **Unified data formats:** All data collected by CLEAR would be organized into a database and a unified data format would be adopted.
3. **Missing data:** Any missing data would be clearly flagged and an option would be provided to the modelers to fill in these missing observations by interpolation. It would also be made clear to the modelers that the data has been modified from the original source and that the data can only be used for modeling purposes. Any legal disclaimers that data collecting organizations wish to include would be included.
4. **Metadata standards:** A standard metadata format would be designed and developed from the viewpoint of modelers and provided along with the datasets.
5. **Data format standards:** Data would also be provided as standard ASCII text files and in a self describing machine independent scientific data format like NetCDF (Network Common Data Format) developed at the National Center for Atmospheric Research (NCAR). NetCDF has been adopted as a standard for disseminating data by the Oceanographic and meteorological communities.
6. **Data management utilities:** Utilities in the form of a subroutine library would be developed by CLEAR so that modelers can call these in their codes and thus speed up their model development.
7. **Model output repository:** CLEAR would also act as a central repository for model outputs that have been quality checked and peer reviewed by a committee.
8. **Data solicitation from other non-governmental organizations:** CLEAR would also encourage other educational institutions, private organizations and individuals to contribute data to the repository. These data would then be standardized and made available to the modeling community as a whole.
9. **Promoting dialog between agencies and modeling community:** CLEAR would act as a conduit between the modelers and the data collecting agencies and any suggestions for improvement, enhancement etc. for current and future data gathering efforts would be summarized and forwarded to the different agencies.