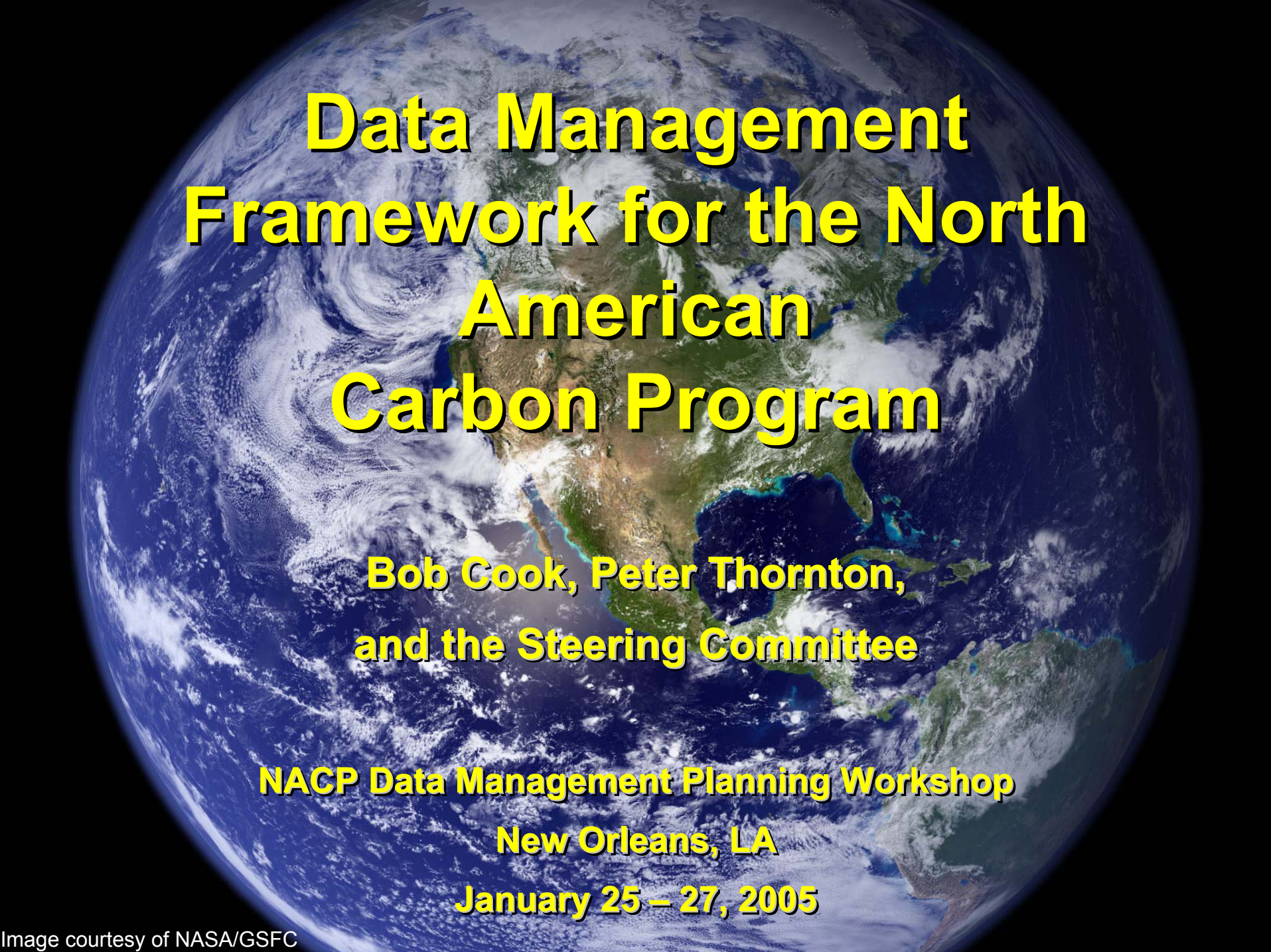


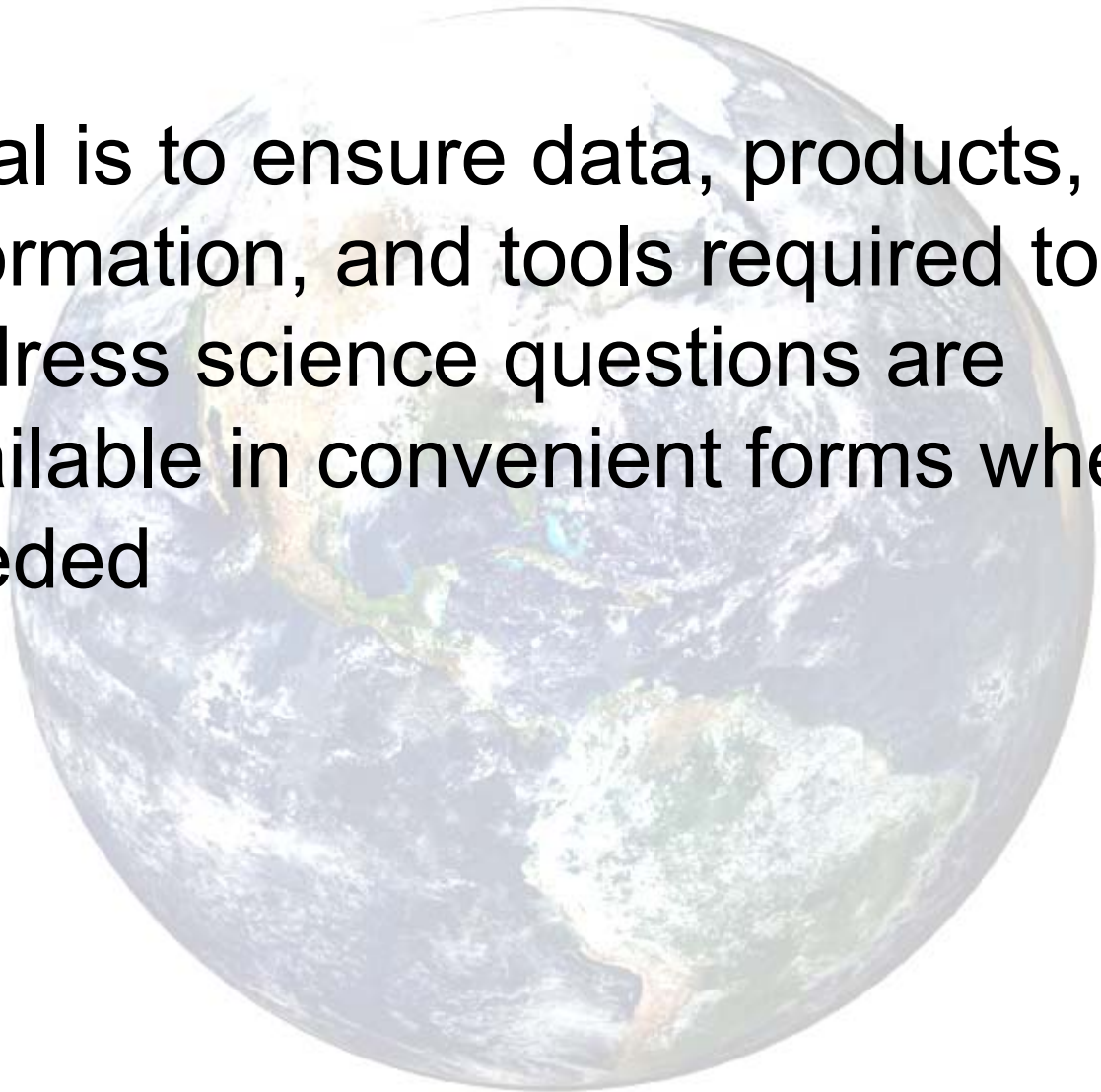
# Data Management Framework for the North American Carbon Program

Bob Cook, Peter Thornton,
and the Steering Committee

NACP Data Management Planning Workshop
New Orleans, LA
January 25 – 27, 2005

Image courtesy of NASA/GSFC

# NACP Data and Information Management



- Goal is to ensure data, products, information, and tools required to address science questions are available in convenient forms when needed

# NACP Data Management Vision

- Data System that enables
  - discovering, accessing, browsing, and comparing data with standard tools (GIS, visualization/analysis systems) without concern for format, location, or volume
- Free users from the productivity losses associated with
  - lack of a central clearinghouse,
  - incompatible formats,
  - unwieldy file sizes, and / or
  - large non-aggregated collections

# Challenge for Workshop

- Data system will most likely be funded with base-funding or reprogrammed funding
  - “Right-size” the data management system
  - Make optimum use of existing infrastructure
- Developing data management capability that
  - Reflects the needs of the user community,
  - Is created in a reasonable time-frame, and
  - Is universally accepted as a value-added capability to the those doing work (researchers, decision makers, policy makers, etc.).
- Make full use of existing technology and anticipate using beneficial future technologies
  - Dynamic / flexible design

# Characteristics of the Data System

- Integrated system that delivers a suite of diverse products
- Facilitate rapid exchange of large amounts of data, products, and information
- Close coordination between the data group(s) and the producers and users of the data products
- Process to assure data quality and uncover errors early, to allow proper interpretation and use
- Ensure that finalized NACP data and documentation are transferred to an appropriate archive

# Data Policy for NACP

- Full and open access to data
  - Consistent with international / agency policies
- For programs that allow initial periods of exclusive data use, data should be made openly available as soon as they become widely useful
  - along with documentation on data quality and potential usefulness
- Define provisions for accessing key data that are copyrighted or proprietary
- Must be effective
  - Enforceable and investigators willingly follow the policy

# Identify Data Requirements

- Data and information management requirements and schedules should be identified early
  - Draft list of required data products included in *Science Implementation Strategy Plan*
  - Received input from existing and recently funded NACP investigators and other data users and producers (November 2004 Survey)
  - SSG needs to identify priority data products required to address NACP research questions

# Data Required for NACP (from *SIS Plan*)

- Model output products
  - Integrated fluxes
- Atmospheric constituents
  - Flask, tower, and buoy networks, airborne, remote sensing
- Flux products
  - Ecosystem exchange, fossil fuel emissions, fire emissions
- Land cover / land use products,
  - Biomass, vegetation structure, historical and current land cover, land use / change
- Carbon inventory products
  - Above and below-ground biomass, soil C
- Data required for diagnostic and prognostic models
  - Climate, meteorology, radiation, ecosystem attributes, DEM, water flow, soil properties

# Key Data Dependencies (U.S. Agencies)

**USDA**

- Forest & soil inventories
- Agricultural, forest and range management
- Carbon sequestration

**NOAA**

- Meteorological observations
- Ocean surface temperature and land cover observations
- Atm. $CO_2$ flask/tall tower network
- Weather models (NCEP)
- Air-Sea $CO_2$ exchange studies
- Integrated carbon modeling
- Ship-based ocean $CO_2$ surveys

**DOE**

- Fossil fuel emissions
- AmeriFlux
- FACE and other $CO_2$ expts.
- Carbon databases (CDIAC)
- C modeling & C sequestration
- ARM Project

**USGS**

- Landsat data & data products
- Topography & land cover maps
- Stream gauge network
- Hydrography

**NASA**

- Remote sensing: satellite time series (Landsat, SeaWiFS, and EOS); airborne sensors
- Remote sensing research
- Field campaigns— LBA, BOREAS
- Ocean, land, atmosphere and coupled carbon-climate modeling
- ESTO & DIS

**NSF**

- Earth science research
- Ocean campaigns
- Process studies
- NCAR, NCEAS, LTER
- CyberInfrastructure Initiative

*Slide courtesy of Lisa Dilling*

# Data Required - 1

- What are the priority existing data products required for NACP?
- How do the data need to be modified for NACP use? (if at all)
  - Revise the measurements?
    - Additional parameters; changes in spatial resolution or temporal frequency; improved measurement techniques, precision, or accuracy; inter-site calibration, etc.
  - Create value-added products?
    - Change format, projection, or gridding; gap-filling; inclusion of uncertainty and other quality information
    - Will NACP modify the data products or will the agencies' data centers?
- Establish data coordinating agreements between NACP and agencies / data centers
  - Unprecedented levels of cooperation across current data management institutions and programs
  - Expand partnerships with foreign governments, intergovernmental agencies, and international scientific bodies and data networks

# Data Required - 2

- Identify new data products that will be generated by recently funded NACP investigators
    - Model output products, others?
- Are there integrated data sets that need to be assembled for data fusion / data assimilation?
    - Will these integrated data sets be assembled by NACP or individual modeling groups?

# Data Management Functions

- Facilitate data acquisition
  - central or distributed
  - protocols for submission, data formats, and metadata content
- Quality assurance
  - document data quality for end users
- Distribution and access
  - data and metadata standards
  - Central, distributed, or a combination
    - option: centralized Web-based metadata clearinghouse with links to distributed data
  - tools to facilitate data discovery, subsetting, visualization, and analysis
- Preparation of value-added data products
  - products that are essential or would enhance analyses, and that are not already being generated by funded research teams or data centers

# NACP Models

- Make numerical models available, along with model parameters and example input and output data
- Model source codes, when made publicly available, can be used
  - to understand the uncertainty of model results relative to results from other model or observations
  - to enable others to see how models treat individual processes
  - ultimately to improve modeling approaches

# Archive Data

- NACP data products need to be archived
  - When the data sets are finalized
  - After completion of NACP
- Many data products used for NACP are currently being archived by agency or national data centers
- Plans should be made for the long-term archival and distribution of key NACP data products, including value-added products generated by the project

# Data Management: Concluding Thoughts

- Dedicated financial support for data management is essential for both individual projects and for NACP-wide data activities
- Team leading NACP needs to establish
  - Priority data products needed to address NACP research questions
  - Data coordinating agreements between NACP, NACP investigators, and contributing agencies and their data centers
  - NACP Data Policy
- Communication among data managers, those making the measurements, modelers, and other data users is critical
- Data management coordination through
  - Annual or semi-annual modeling / analysis workshops (subgroups of NACP or NACP-wide)
  - "*User Working Group*" that provides oversight and guidance

Image courtesy of NASA/GSFC

# Extra slides

# Period of Exclusive Use

- In the past, some Principal Investigators have retained data for indefinite periods, and this has inhibited their widespread use.
  - This practice should be eliminated through active consideration of the tradeoffs between widespread distribution of data sets and the need to assure data quality and validity.
  - The guiding principle is that as soon as data might be useful to other researchers the data should be released, along with documentation which can be used by the other researchers to judge data quality and potential usefulness.
  - users can then determine for themselves if they want to proceed with data of questionable quality or wait for additional developments.

- Access should be seamless, platform independent, timely, and open

- Organizations make observations, for a wide range of purposes
  - Lack of homogeneous observations hinders many applications and limits capacity to monitor the changes relevant to carbon cycle