Issues that need to be addressed on data management for AMY

-- Towards more active sharing of hydro-climatic information in Asia --

Kooiti MASUDA
Frontier Research Center for Global Change
Japan Agency for Marine-Earth Science and TEChnology
3173-25 Showa, Kanazawa-ku, Yokohama 236-0001, Japan
http://www.jamstec.go.jp/frcgc/research/p2/masuda/

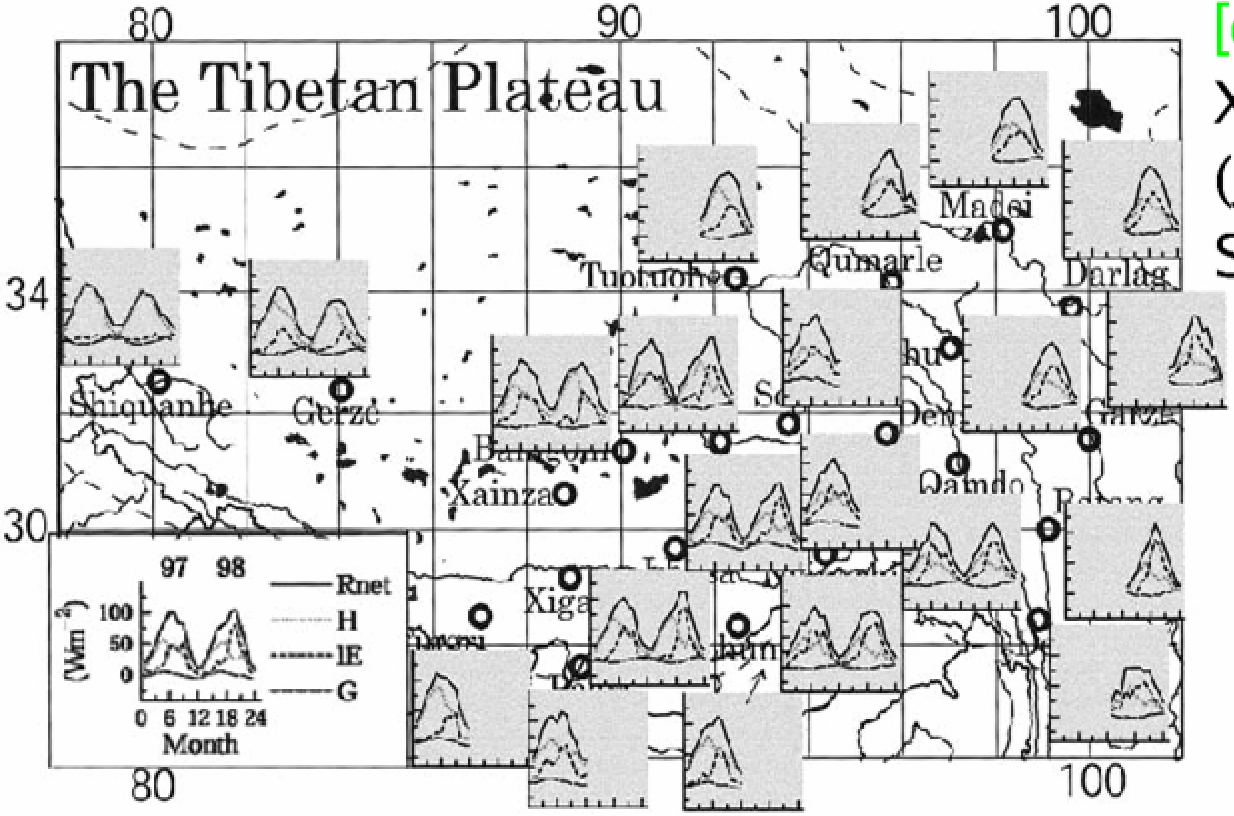
MAHASRI - Asian Monsoon Year workshop
Tokyo, Japan, 10 January 2007

Types of data

- Experimental observations
- Operational observations
- Observation-based composite data (e.g. radar-raingauge)
- Satellite products
- Data assimilation product
- Model (intercomparison) experiment input/output
- Model output (e.g. climate change projection)

We need data from both experimental and operational observations.

- Experimental observations give accurate, detailed information.
- Operational observations are useful
  - to prepare for the experiment;
  - to know the situation of the experiment in the larger context;
    e.g. How abnormal was the season?
  - to extend knowledge obtained by the experiment
    with "models" are calibrated with exp. obs. and applied with op. obs.



[example from GAME-Tibet]

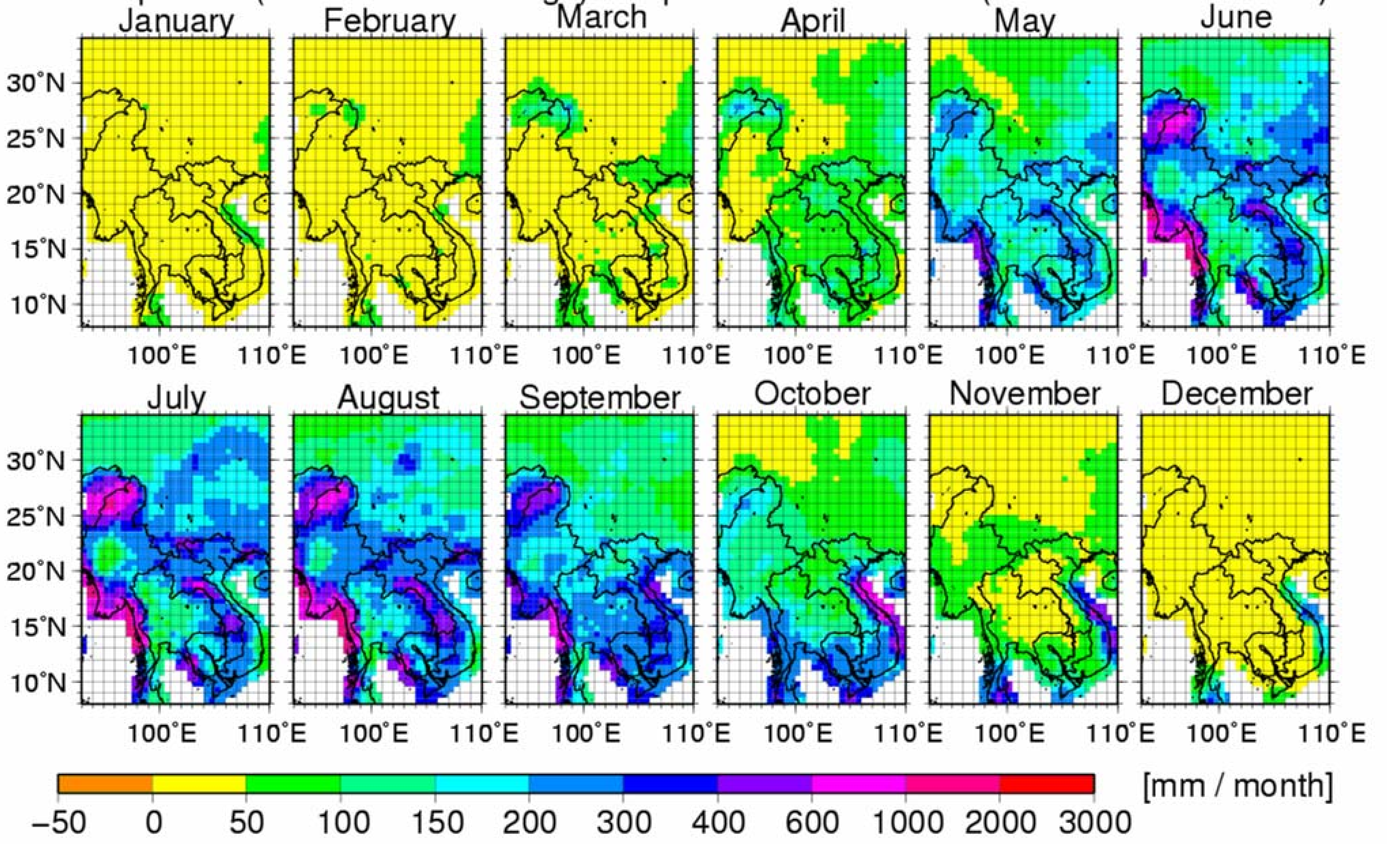
XU Jianqing et al. 2005
(J. Meteorol. Soc. Japan)
Surface energy balance

- a model for soil water
- + empirical formulas
  for radiative fluxes
- input <- operational obs.
  temp., humidity,
  sunshine duration
- calibration <- exp. obs.

The natural environment is borderless. We need combine data across borders.

[example from GAME-Tropics] Masuda et al., paper in preparation

Precipitation (1981 – 2000 average) interpolated from stations (GAME+MRC+national)



January, February, March, April, May, June, July, August, September, October, November, December

Consideration about sharing data

- (I think) Environmental data are (or should be) "public goods".
  - They should be exchanged freely without restriction.

    Matter of human survival ... IMO(1873), "Essential" data of WMO 1995

    Academic tradition ... IGY(1957-58), GARP(1970s), WCRP etc.
- In practice, costs must somehow be accounted for.
  - Restriction may be applied to some data ("Additional" data of WMO 1995)

- For disaster prevention etc., data should be released in real-time.
- In case of experimental data, the principal investigator may want delay.

- Probably we will have several different classes of data together:
  - Open to public without any restriction
  - Non-commercial use only (notified but not checked strictly)
  - Non-commercial use only (checked strictly)
  - Shared among the participants (of MAHASRI or of AMY)
- We need a rule for possible delay (e.g. 1 year if requested)
- We should cooperate with operational agencies for real-time data exchange.

Some consideration about tasks of data management

- Data quality is important.

- Feedback between users and producers is important for data quality.
  - Data problems are often discovered by users.
  - Correction of data problems are usually possible only by the producers.

- Documentation of data (metadata) is important.
  - Users may not have the same background knowledge as producers do.

- We learn from sharing experience (incl. experience of failure).

- Persons enthusiastic for sharing of data/knowledge are important.

- Observations in Asia are less well represented in global data sets than those in N. Amarica or Europe.

[example] Data quality issues about raingauge data (GAME-tropics plus extra)

- Distinction between "value missing", "no precip." and "trace precip."
  - Insufficient description in some source
  - Monthly sum 0 mm in rainy season
- Very large (but physically possible) values
  - Real heavy rain?
  - Mistake? (Decimal column shift? Extra character inserted?)
- Data at the same station at the same time do not match each other
  - Parallel observation?
  - Mistake in transcription and/or digitization?
- Unrealistic seasonal cycle
  - Mistake between rows and columns of a table
  - Mistake between calendar year and hydrological year
- Problem of boundary of day (for daily precipitation)
  - 00 UTC for Thailand; 12 UTC for China; sometimes undocumented
  - Is observation today called "precip. today" or "precip. yesterday"?
- Problems of identification
  - Location (latitude/longitude) wrong or unknown
  - Duplicate? (more than one data sets may include the same source)

Proposed structure of data management in MAHASRI / AMY

- We do not expect a single "data center" to manage all relevant data.
- Distributed Active Archive Centers (DAAC)
  - MAHASRI participants-(data)→ DAAC-(data collection)→users
  - Quality check, standardization, documentation, portal to knowledge

- MAHASRI Data Management "Council" ... e-mailing list (?)
  - Representing institutions (data providers, DAACs, major data users)

- Digital bulletin board on the WWW for information and discussion about data
  - Voluntary contribution of individuals (scientists, administrators, students...)
  - To exchange information about data, techniques, ideas of application
  - To promote the idea of sharing data and to attract support to it

- MAHASRI Data Management "Core Team" ... meeting frequently (?)
  - Drafting the policy to be approved by the "Council"
  - Coordination between DAACs (esp. common inventory, documentation)
  - Management of digital bulletin board and e-mailing lists

Immediate tasks

- Nominating (candidates for) DAACs and/or Core Team members
  - e.g. JAMSTEC, MRI/JMA, U.Tokyo, APCC, ...
- Deciding (tentatively) role(s) of each DAAC and/or Core Team member
- Nominating the "Council" members
  - national commitee or agencies/projects -> Core Team [ ]
- Organizing mailing lists and web site(s) by Core Team [ ]
- Establishing the data policy for AMY
  - Drafting by Core Team [ ] -> discussion -> approval by "Council"
  - Candidate starting point: policy for CEOP Reference Sites
    http://www.eol.ucar.edu/projects/ceop/dm/documents/ceop_policy.html
- Listing up offers and/or request of data sets (types, amounts) for AMY
  - Core Team [ ] will ask members of "Council" and/or IMASSC

Data exchange guidelines:

(1) To comply with WMO Resolutions 40 (CG-XII) and 25 (CG- XIII) in particular: <u>No financial implications.</u>

(2) CDA and *data users*: Commercial exploitation of CEOP data is prohibited.

(3) *Data users*: No transfer to third parties.

(4) Data release to *data users*: Turn-around period.
Category 1 data: 6 months *Category 2* data: 15 months

(5) Acknowledgement and citation

(6) Co-Authorship for Reference Sites' PIs recommended, collaboration base required if PI requests co-authorship (in particular for *category 2* data)

(7) CEOP Publication Library at CDA