e-Science in Geospatial Computation

Eric Yen

9 October 2009

Academia Sinica Grid Computing

Outline

- E-Science and the Landscape
- E-Infrastructure and Applications in Taiwan
- Geospatial Computing Examples by Grid
  - EUAsiaGrid and TWGrid
  - TeraGrid
  - EGEE
- Summary

e-Science Reminder

- Definition
  - "e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it." (by John Taylor, http://www.e-science.clrc.ac.uk)
  - e-Science apps pulls, Information & Communication Technology pushes
- Objectives
  - Support research by e-Science, on data intensive sciences and applications require cross disciplinary distributed collaboration
- Application Driven, Deployment Approach since 2002
  - Only Grid Middleware and Toolkits are not sufficient
  - Data Intensive High Energy Physics, WLCG global collaboration requires Distributed Data Management and Analysis
  - First WLCG Grid Deployment Board (GDB) Meeting on 4 October 2002, Milano
  - Deploy large scale 24x7 production service for HEP application to allow debugging, increase robustness and learn how to operate could drive next generation e-Infrastructure

E-Science for GIScience

- Increasing data in quality and precision
- Sustainable access to geospatial materials
- Expedite mining new knowledge by geospatial resource federation and analysis
- Enable New Data Models and Knowledge Organization System - e.g., Spatial Semantics for Automating Geographic Information Processes
  - Metadata Semantics, Feature Semantics, Geographic knowledge semantics, and Spatial relation semantics
- Interdisciplinary Collaboration
- **BUT!!** We cannot take advantage of these evolution without a scientific knowledge infrastructure for the geosciences

Objectives

- Leveraging e-Science paradigm for Geospatial Information Science
- Facilitating collaboration by virtual science environment
- Providing integrated environment to support variant researches
- Supporting computational capability and performance
- Incorporating workflow management, resource federation and long-term preservation services

Collaborating e-Infrastructures

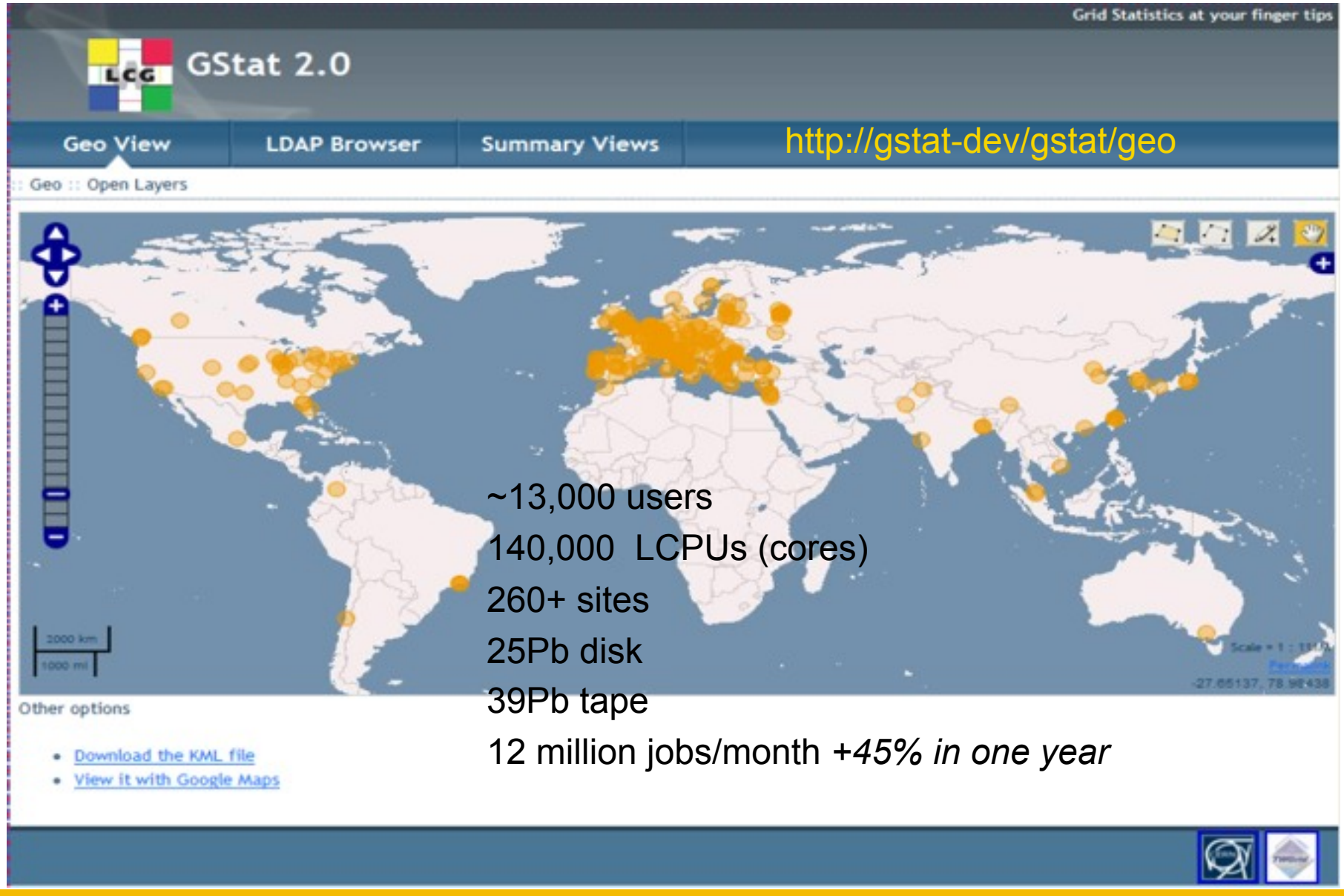
"Production" = Reliable, sustainable, with commitments to quality of service

Collaborating e-Infrastructures

"Production" = Reliable, sustainable, with commitments to quality of service

Potential for linking ~80 countries

Infrastructure Growth



~13,000 users

140,000 LCPUs (cores)

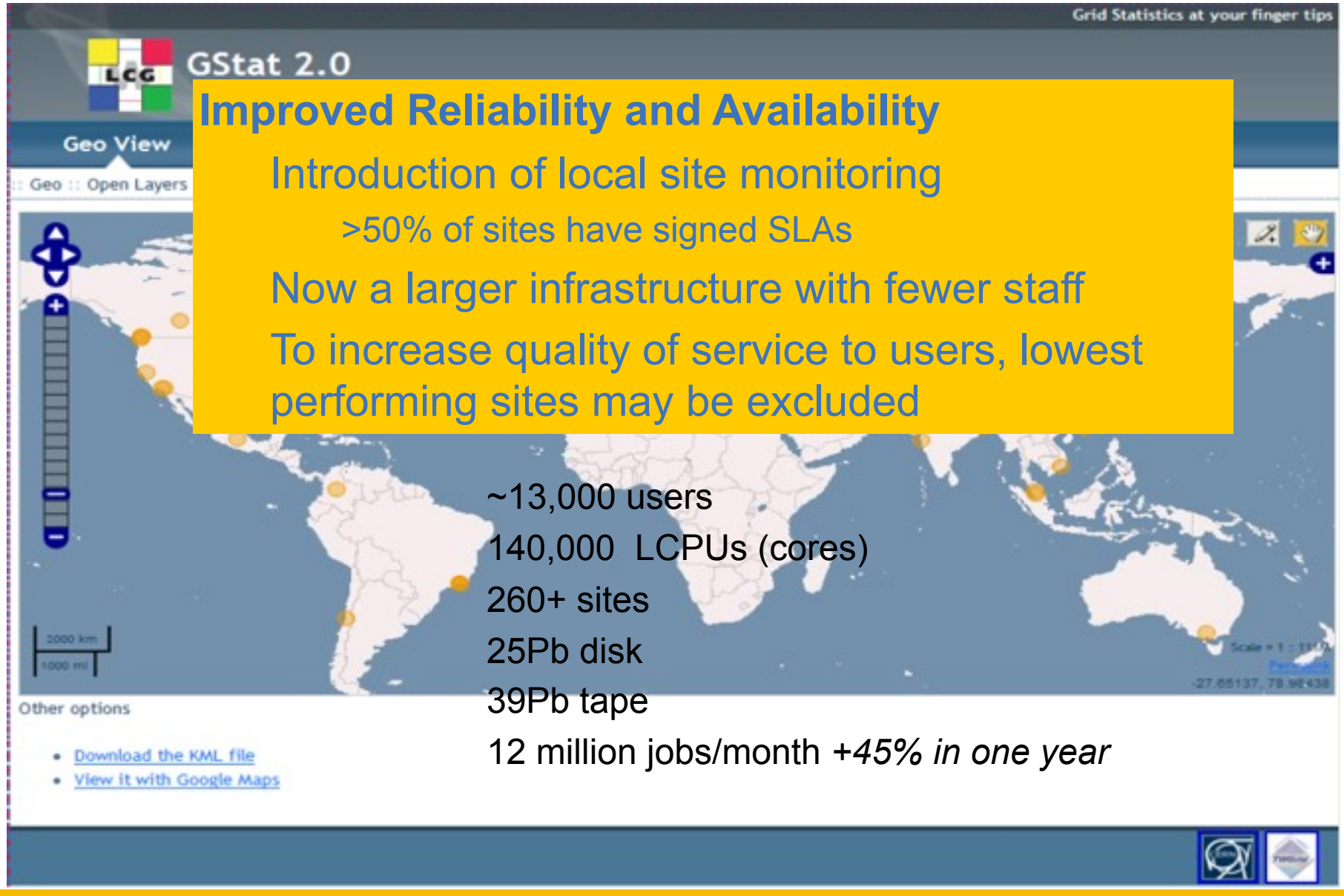
260+ sites

25Pb disk

39Pb tape

12 million jobs/month *+45% in one year*

Infrastructure Growth



Improved Reliability and Availability

- Introduction of local site monitoring
  - >50% of sites have signed SLAs
- Now a larger infrastructure with fewer staff
- To increase quality of service to users, lowest performing sites may be excluded

~13,000 users

140,000 LCPUs (cores)

260+ sites

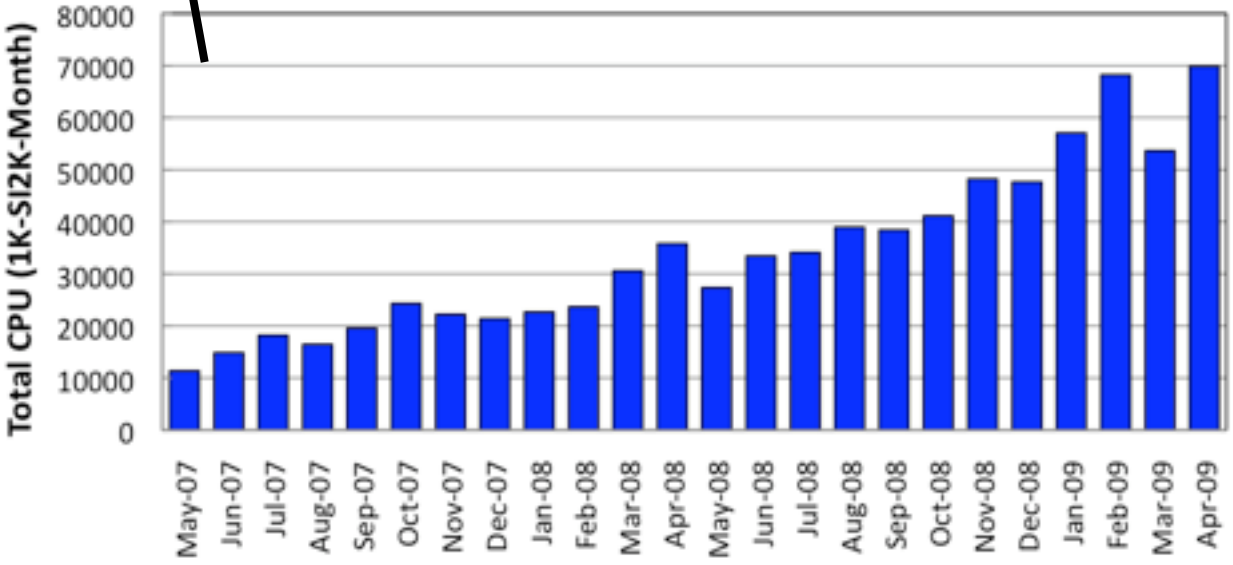
25Pb disk

39Pb tape

12 million jobs/month *+45% in one year*

User Communities

Consistent doubling every 12-18 months.
HEP largest users / contributors
AA/ES/other show strong increase



Domain	VOs	Users
AstroPhy & Astronomy	20	373
Comp Chem	4	347
Comp Sci	4	21
Earth Sci	7	142
Fusion	2	68
High Energy Phys	36	8577
Life Sci	9	379
"Regional"	26	1658
Other	28	1816
TOTAL	136	13381

CIC Portal: http://cic.gridops.org/
Accounting Portal: http://www3.egee.cesga.es/

>13,000
Registered Users

EUAsiaGrid

- Identify and engage scientific communities which can benefit from the use of state-of-art Grid technologies;
- Disseminate EGEE middleware in Asian countries by means of public events and written and multimedia material;
- Provide training resources and organise training events for potential and actual Grid users;
- Support the scientific applications and create a human network of scientific communities by building on and leveraging the e-Science Grid infrastructure.

Challenges of the Project

First Grid project targeting Asia-Pacific region

- ✓ Geographically large and culturally diverse area
- ✓ Uneven levels of adoption of Grids

Wide range of scientific domains addressed

- ✓ Consolidate on traditional areas and engage new communities
- ✓ Dissemination, Training and Support for applications must create a virtuous cycle

Sustainability of the e-infrastructure

- ✓ Need to define accurately a roadmap for the future
- ✓ Need to train the trainers to trigger teaching from local groups

Supported Applications

- **Applications in the DoW**
  - Computational Chemistry
  - Social Science
  - Bioinformatics and Biomedical research
  - High Energy Physics
  - Mitigation of natural disasters
- **Newly identified areas**
  - Digital culture and heritage
  - Weather forecast and climatology
  - Mathematical modelling
  - Biodiversity

e-Science in Asia

- Diversity:
  - Geographically large and culturally diverse in nature
  - Level of scientific collaboration is reflected by the networking connectivity
  - The region as a whole traditionally inexperienced in regional cooperation
- Grids in Asia
  - Disparate Grids with limited operations experience, making collaboration difficult.
- Why e-Science in Asia?
  - The global infrastructure is establishing quickly
  - Take advantage of sharing and collaboration to bridge the gap between Asia and the world
  - To address the challenge of regional cooperation
- EGEE Asia Federation and EUAsiaGrid
  - EGEE AF led by ASGC, Academia Sinica, Taiwan (Dr. Simon C. Lin)
  - EUAsiaGrid is to empower scientific collaboration throughout Asia
  - Demonstrate vigorous synergy with 29 EGEE sites and more than 8,000 CPU Cores and close to 3 Peta Bytes disk space for 12 VOs by early 2010.

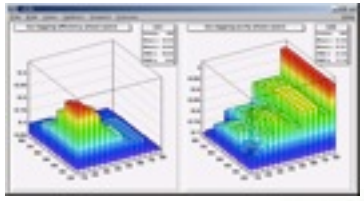
TWGrid Introduction

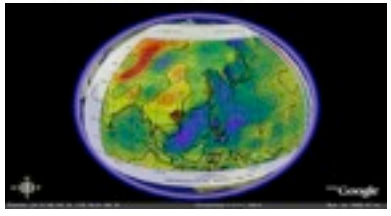
- Consortium Initiated and hosted by ASGC in 2002
- Objectives
  - Gateway to the Global e-Infrastructure & e-Science Applications
  - Providing Asia Pacific Regional Operation Services
  - Fostering e-Science Applications collaboratively in AP
  - Dissemination & Outreach
  - Taiwan Grid/e-Science portal
    - Providing the access point to the services and demonstrate the activities and achievements
    - Integration of Grid Resources of Taiwan
    - VO of general Grid applications in Taiwan
- EUAsiaGrid: alignment for extension of e-Infrastructure and e-Science collaboration in Asia from 2008.

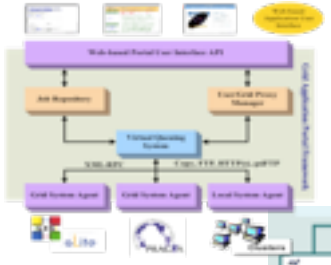
What Do We Deliver ?

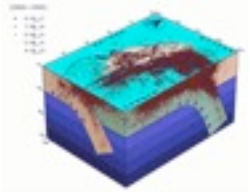
- e-Infrastructure Operation
  - 29 sites across 12 countries in Asia Pacific Region
    - > 8,000 Cores and >2.5 PB storages
  - Continuous monitoring of grid services & automated site configuration management
- Middleware R&D
  - Production quality MW distributed under friendly open source license model
  - Application integration
- E-Science & User Support: Managed process from first contact to production usage
  - Training
  - Expertise in grid-enabling applications
  - online helpdesk
  - Dissemination: attracting more collaborations
- Interoperability: expanding geographical reach and interoperability with collaborating e-infrastructures

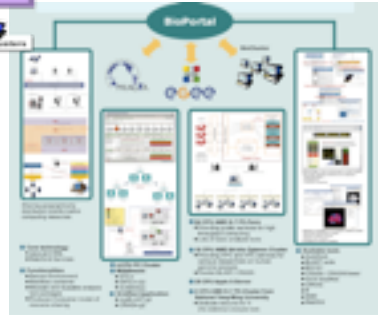
e-Science Applications in Taiwan











- High Energy Physics: WLCG, CDF, Belle
- Bioinformatics: mpiBLAST-g2
- Biomedicine: Distributing AutoDock tasks on the Grid using DIANE
- Digital Archive: Data Grid for Digital Archive Long-term preservation
- Atmospheric Science
- Earth Sciences: SeisGrid, GeoGrid for data management and hazards mitigation
- Ecology Research and Monitoring: EcoGrid
- BioPortal
- Humanity and Social Sciences
- General HPC Services
- Agriculture
- e-Science Application Development Platform

Disaster Mitigation on Earthquake In EUAsiaGrid

Academia Sinica Grid Computing Center and
Academia Sinica Institute of Earth Science
Taiwan

Significance

- Multi-disciplinary Application by e-Science paradigm
  - EDC: Event Details + Simulation + Mitigation
  - Sensor Networking: Federation (local + regional + global)
  - Seismic Wave Propagation Analysis: Source Analysis + Simulation
- Facilitate understanding of ground motion mechanism, rupture process, velocity structure and topographic characteristics (which might be changed after ground shakes)
- Expedite fast response and right protection to earthquake events
- Collaboration by EUAsiaGrid and EGEE
- Toward a sustainable daily services for researchers and general publics

Motivation

- In Asia, natural disasters cause great losses on both economy and human lives. Earthquake is one of the most catastrophic natural disaster.
- Natural disaster is usually not predictable, cross boundary, abrupt and devastating. e-infrastructure boosts new power and new methodology to reduce the hazards.
  - New knowledge: understanding the underlying sciences
  - Much effective to compile and learn from experiences
  - More accurate and effective risk assessment and risk management
- Sharing and collaboration is the best strategy for hazards mitigation.
- Facilitate sustainable e-Infrastructure with real e-Science collaborations.

Application Identification

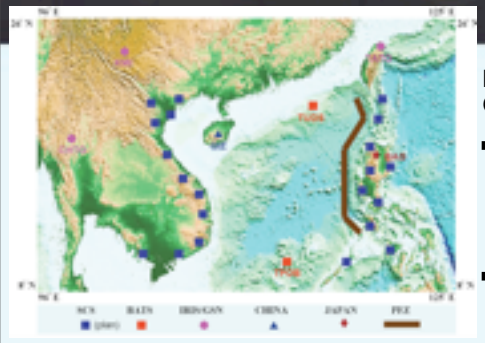
- **Selection Criteria**
  - Partners' common interests
  - Impacts of casualty and economic loss
  - Regional collaboration/support has already been in place
  - Domain Knowledge & Scientific impacts
  - Technology Maturity, in terms of application integration in gLite
- **Target Natural Hazards in EUAsiaGrid**
  - Earthquake and Tsunami
  - Climate and Environmental Change: Carbon Flux Analysis and Climate Simulation
  - Sea Level Rising

Regional Collaboration

	Sensor Network	Seismic Wave Prop. Analysis	Seismic Data Center
Partners	VN, PH, ID, TW		
User Community	Philippine Institute of Volcanology and Seismology (PHIVOLCS), Vietnamese Academy of Science and Technology, The Incorporated Research Institutions for Seismology (IRIS), Global Seismic Network (GSN), Institute of Earth Science & National Central University, Taiwan, Local, Regional, and Global Disaster Mitigation Organization.		
Tech Maturity	TW- Most dense; VN- Sensor Stn ready; PH- expanding	Analysis Model and knowledge available; Cluster and gLite Resources in place; From Global model toward higher resolution regional/country model;	SeisGrid@TW, IRIS, GSN
Exemplar	Integrated Sensor Network by VN, PH and TW	Without local geological data, accurate analysis is not achievable. High resolution historical TW earthquake data sets.	Federation of available Data Centers

e-Science for Earthquake Disaster Mitigation

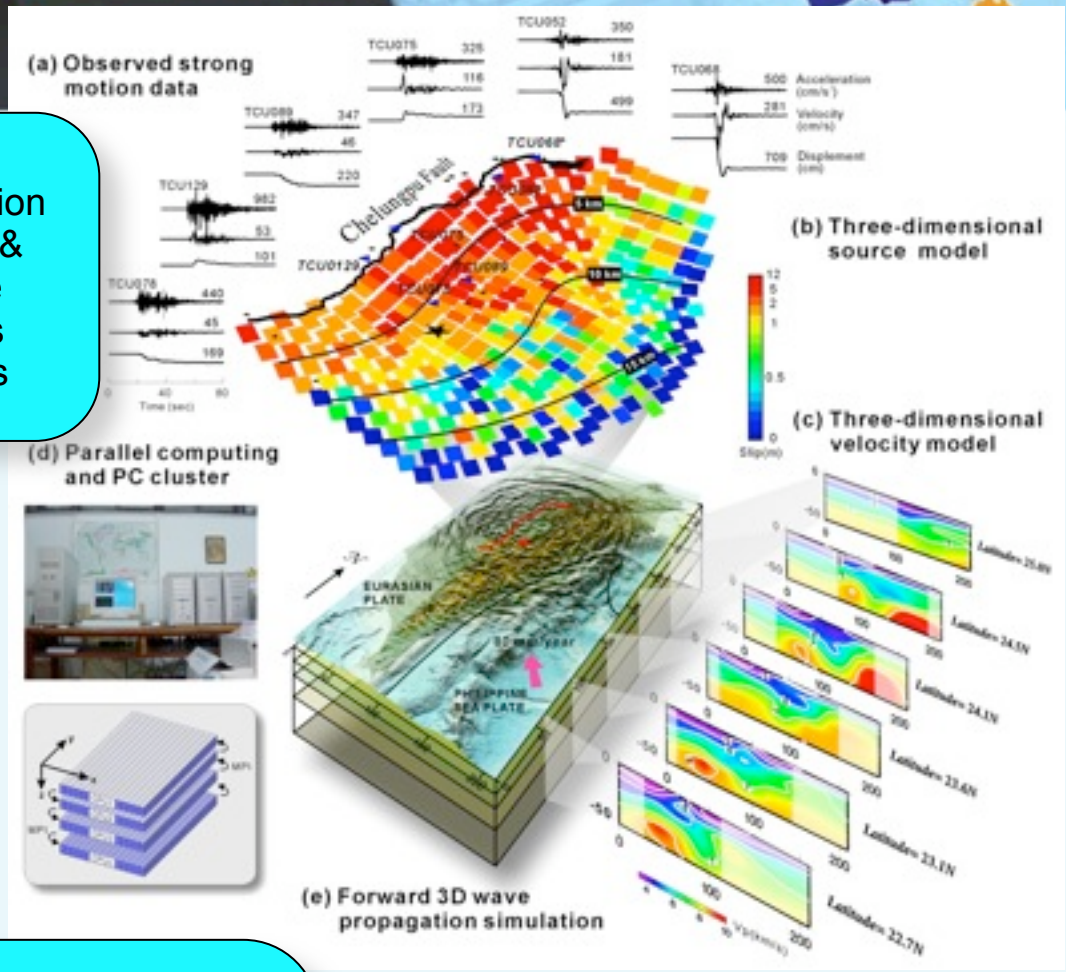
Seismic Sensor Networks



Local Sensor & Observation Data

Global/Regional Sensor Data

High Resolution Source & Rupture Process Analysis



Fast Reporting System

Archive

Earthquake Data Center (SeisGrid)

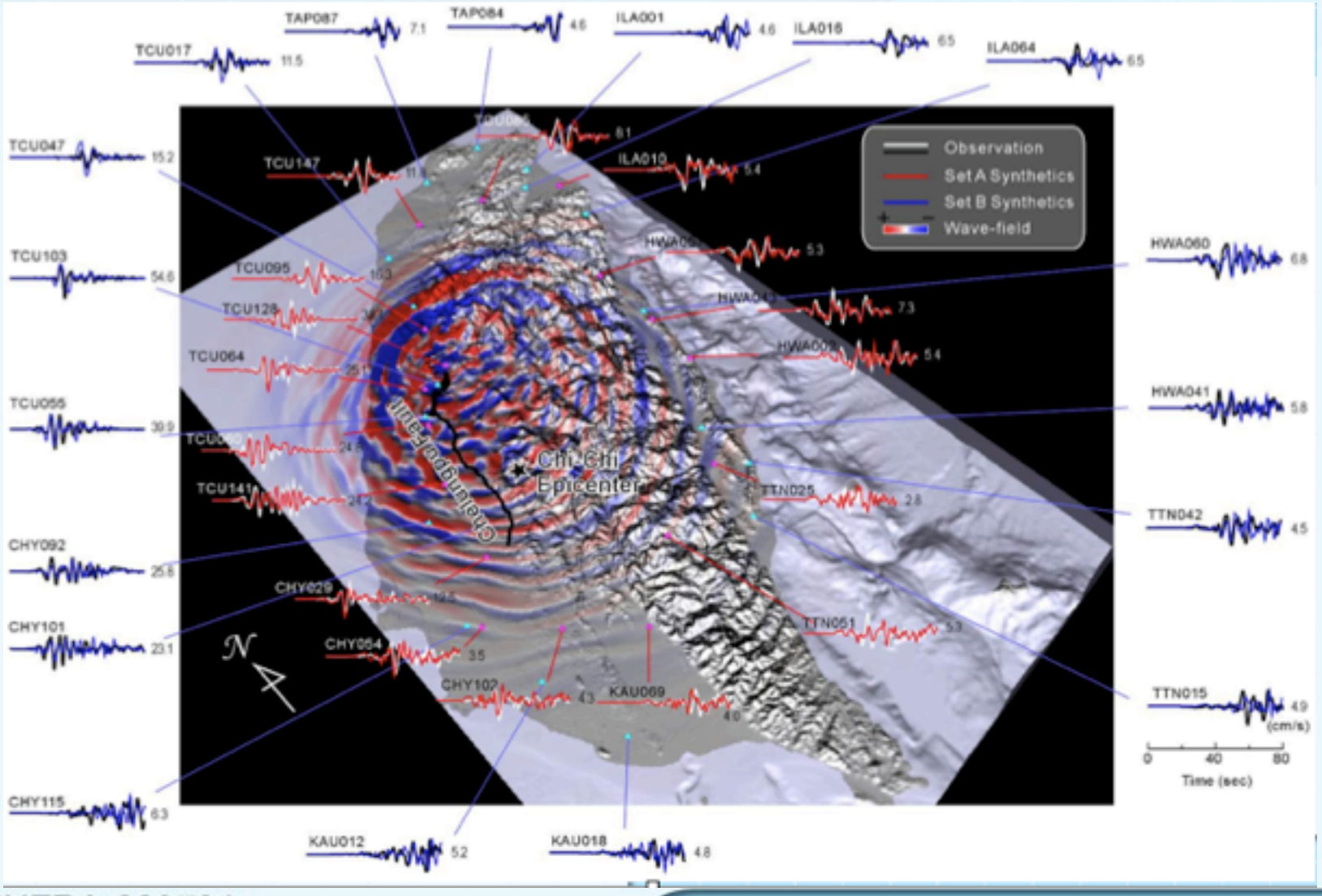
Ref. Historical Events Data

Archive

Forward Simulation & Event Construction on Grid

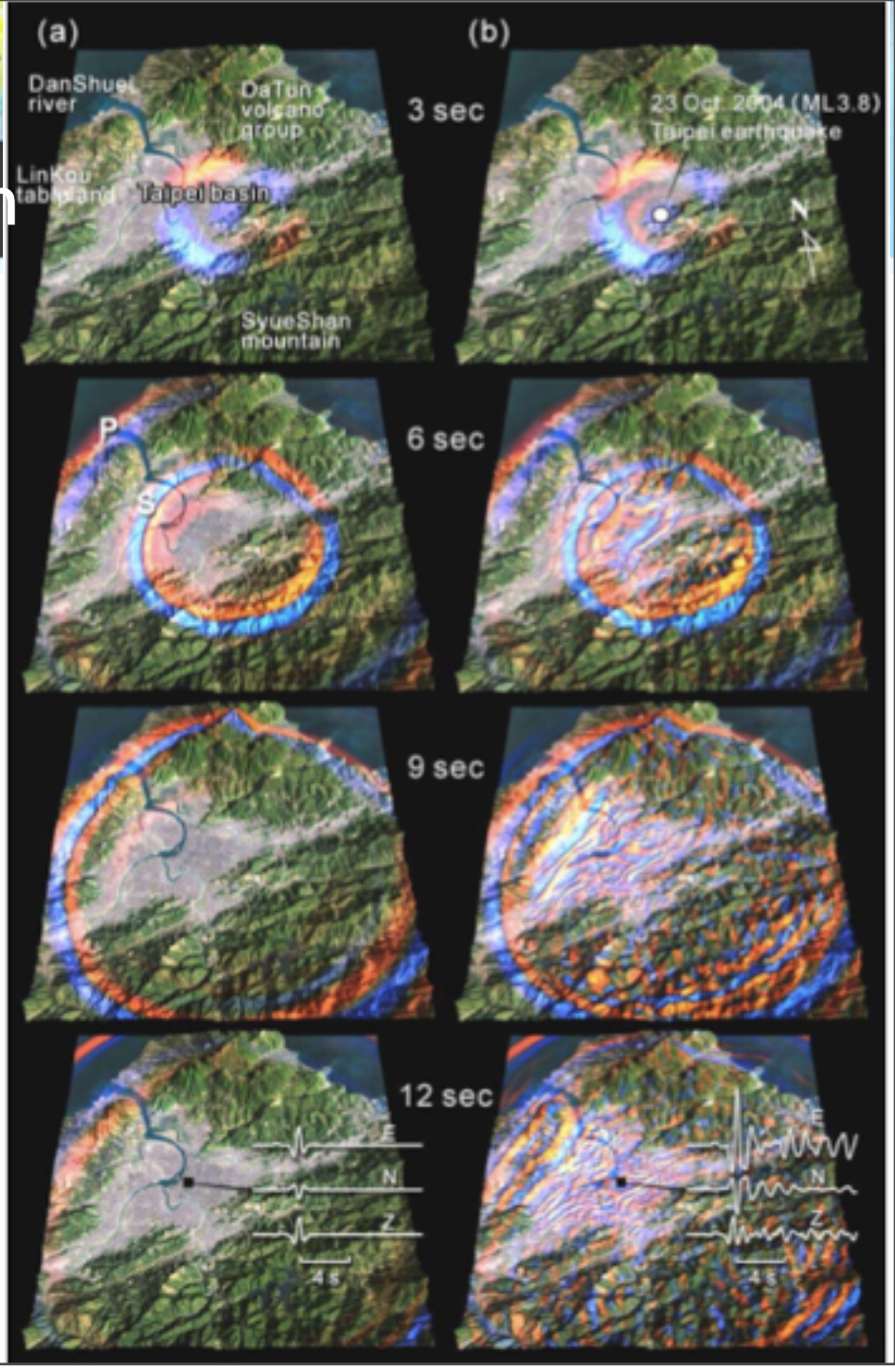
Risk Analysis & Reduction

Seismic Wave Analysis

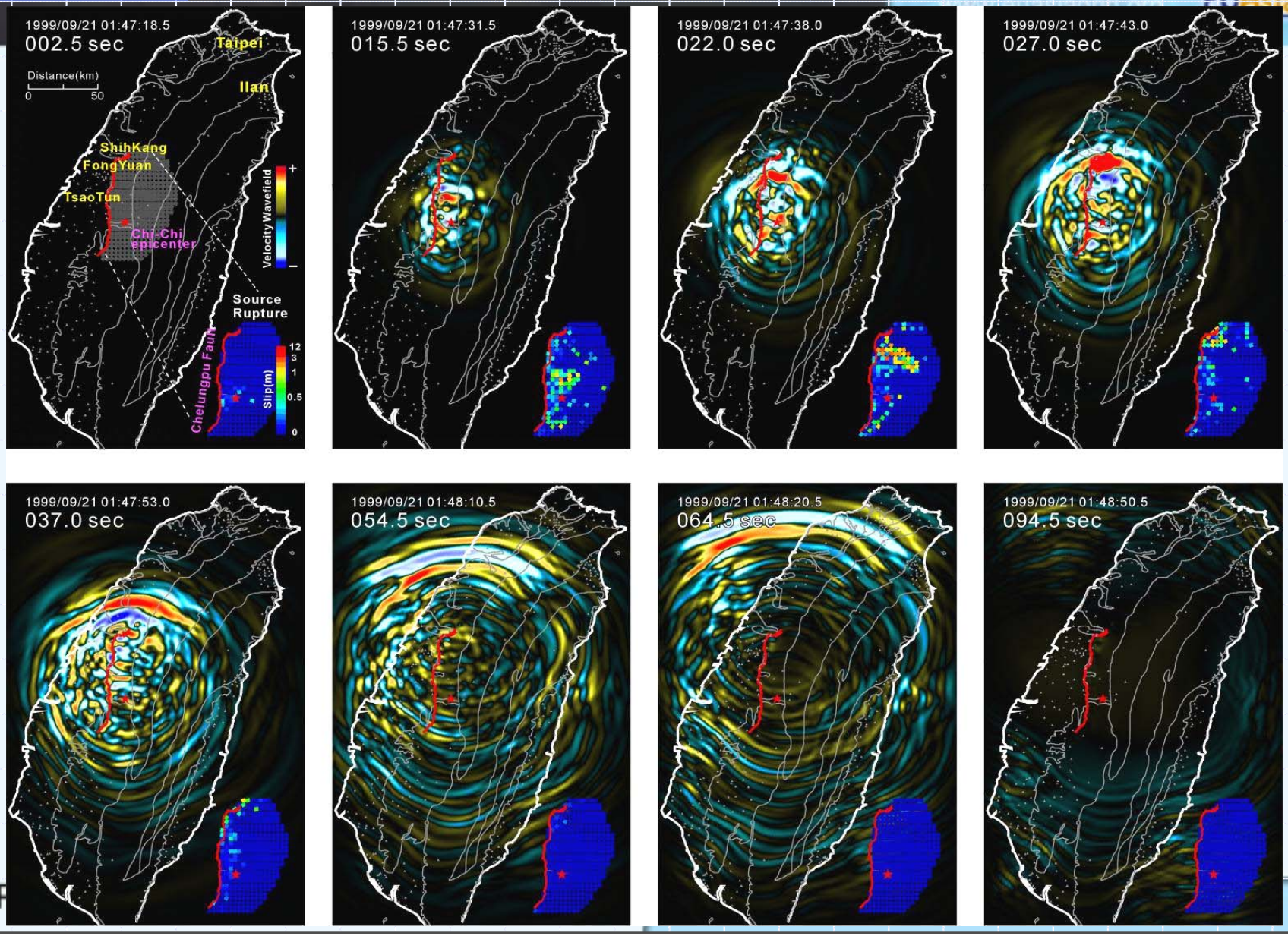


Earthquake Wave Simulation

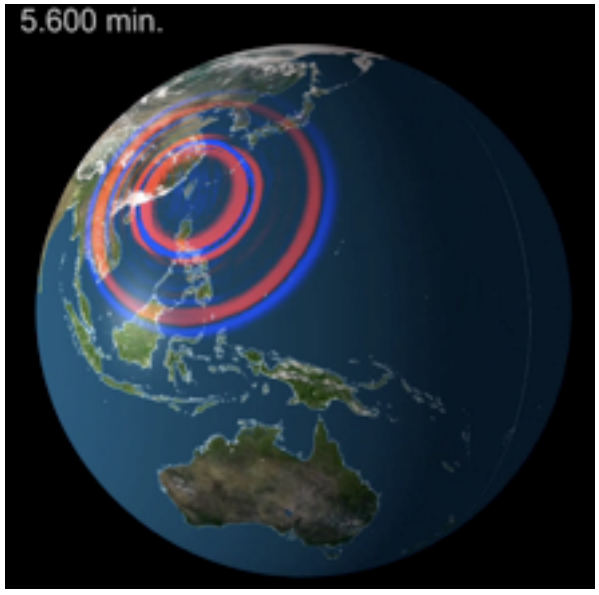
Simulation of 2004 Earthquake in Taipei Basin, w/o Geological Structure consideration

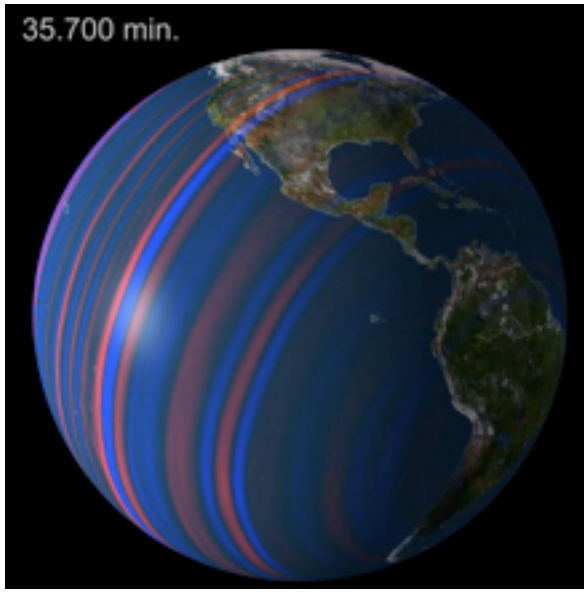


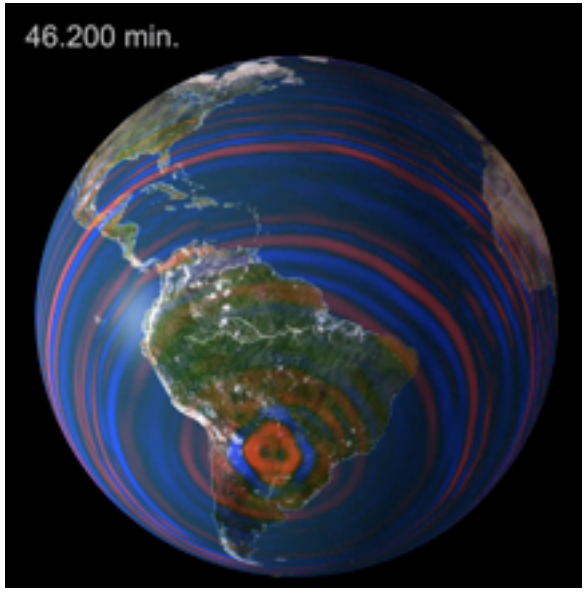
Wave-Field Snapshot

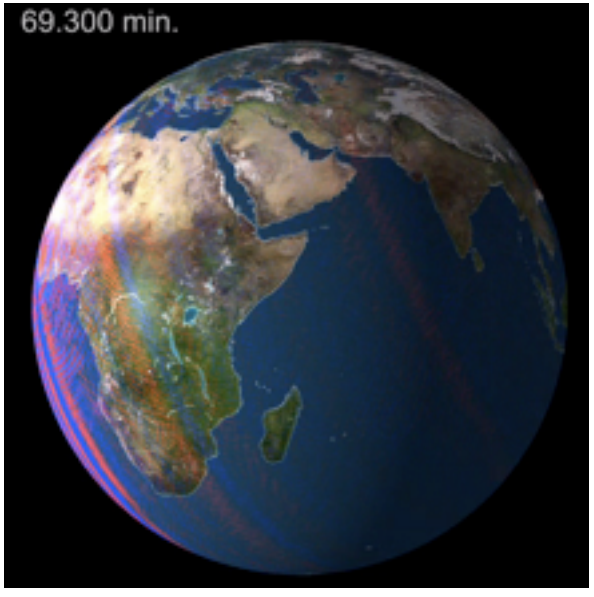


Seismic Wave Propagation Analysis

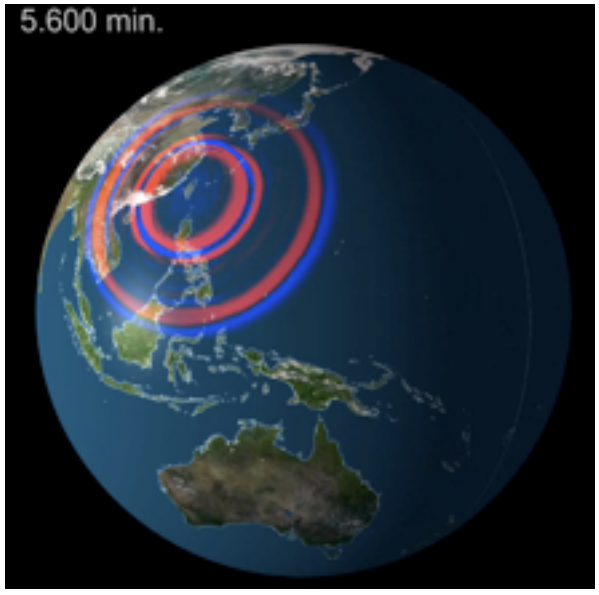


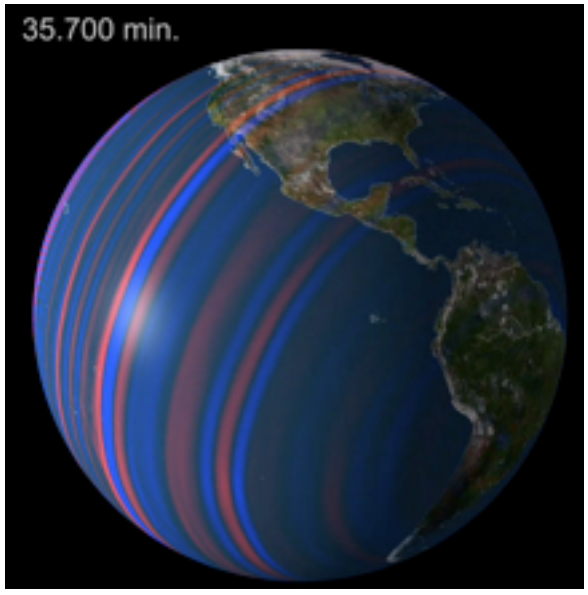


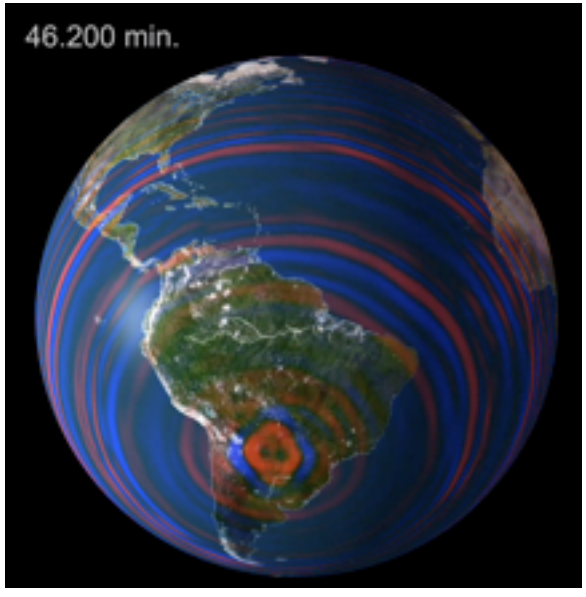


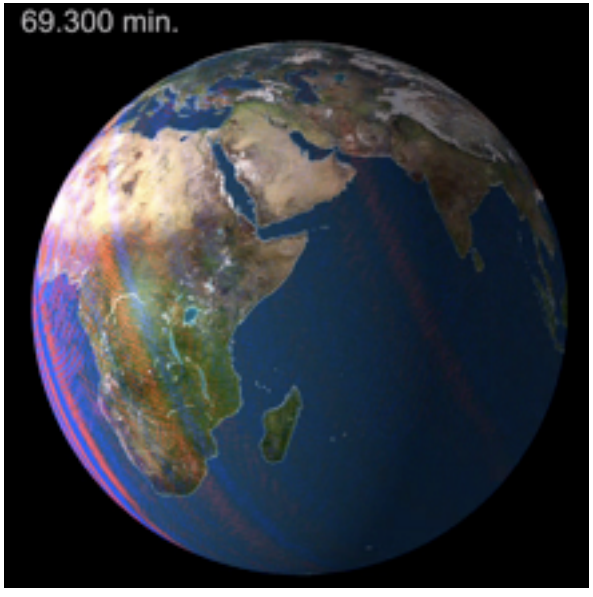


Seismic Wave Propagation Analysis









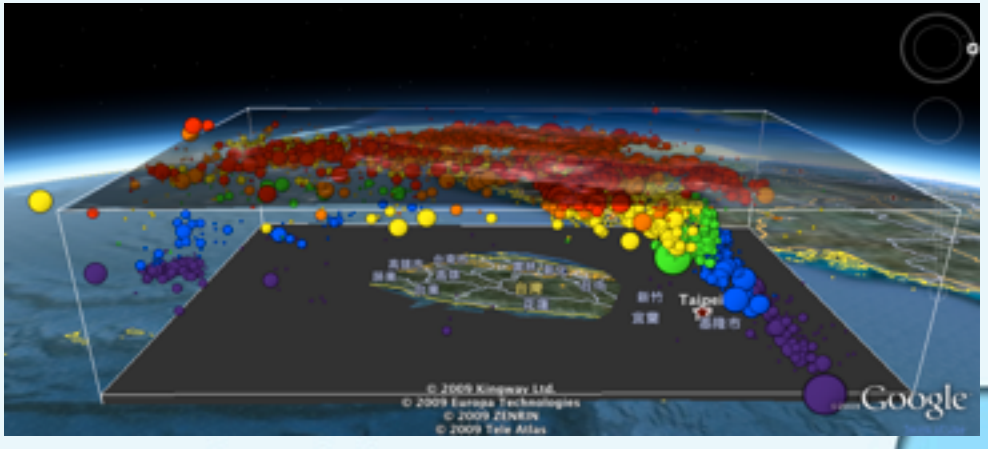
Earthquake Data Center I

- **Data Duration: 1900.1.1 ~ 2009.4.30**
  - 441,100+ records
  - Data Schema
  - Data Statistics
- **Future**
  - Federation
    - With other regional and global EDC
    - With earthquake events/mitigation resources

(Total 9771 Records) Previous Page.2 Next CSV XYZM EventList KMZ- KMZ+ KMZ0 Evt_Out

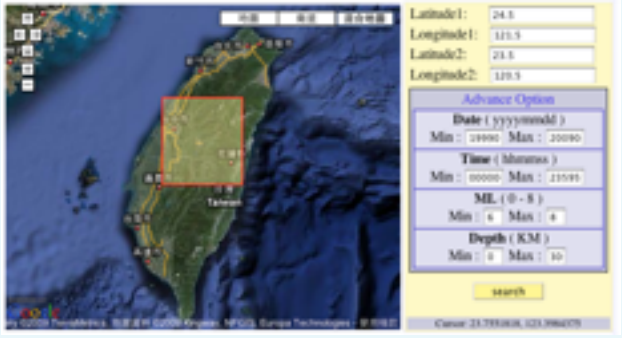
ID	Date	Time	Latitude	Longitude	Depth	ML	NStn	DMin	Gap	Trms	ERH	ERZ	NPH	Quality	Map
93	1910-02-20	14:13:00.00	24.1000	120.7000	0.0000	4.50	3	0.00	0	0.00	0.00	0.00	0		Map
94	1910-02-27	14:28:00.00	21.8000	121.1000	0.0000	4.80	3	0.00	0	0.00	0.00	0.00	0		Map
100	1910-09-07	04:46:00.00	21.8000	120.4000	0.0000	4.60	3	0.00	0	0.00	0.00	0.00	0		Map
102	1910-11-29	02:25:00.00	21.9000	121.2000	0.0000	4.80	3	0.00	0	0.00	0.00	0.00	0		Map
103	1910-12-01	21:42:00.00	23.8000	120.8000	0.0000	4.60	3	0.00	0	0.00	0.00	0.00	0		Map
104	1910-12-30	00:54:00.00	23.9000	121.4000	0.0000	4.60	3	0.00	0	0.00	0.00	0.00	0		Map
105	1910-12-31	07:33:00.00	23.9000	121.6000	0.0000	4.70	3	0.00	0	0.00	0.00	0.00	0		Map

M	Count
<1	3,089
1~2	146,098
2~3	224,854
3~4	55,256
4~5	9,771
5~6	1,871
6~7	199
>7	42
Total	441,180

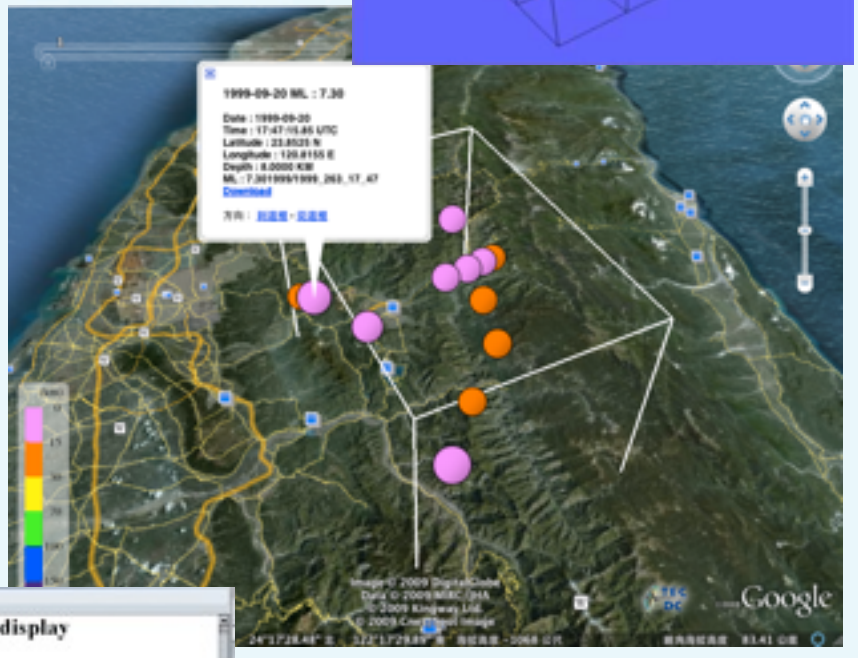


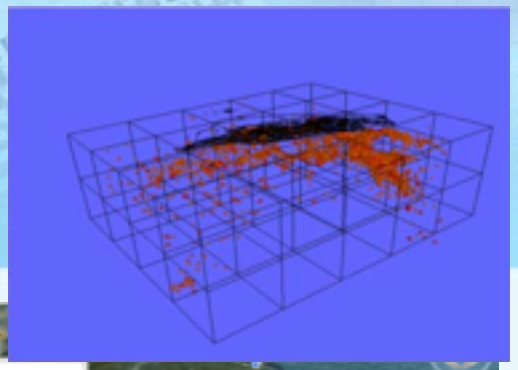
Earthquake Data Center II

Data Selection

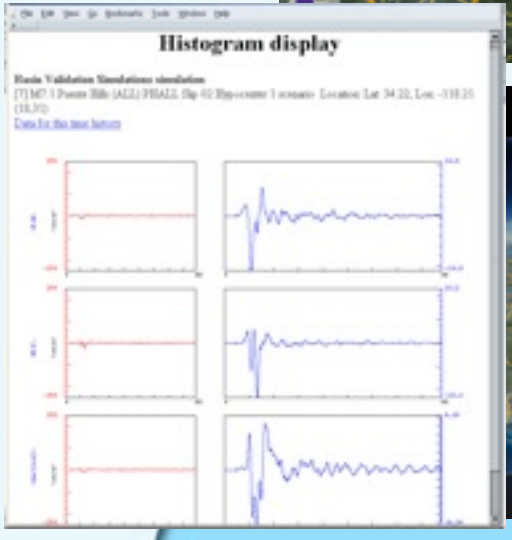


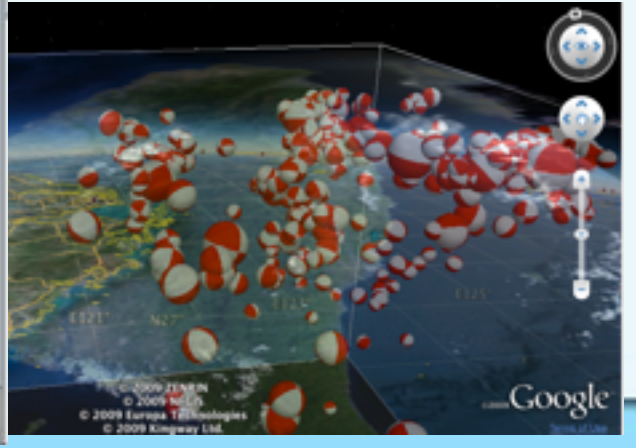
Output





Time History
Seismograms





Seismic Community Library

FP7-INFRA-223791

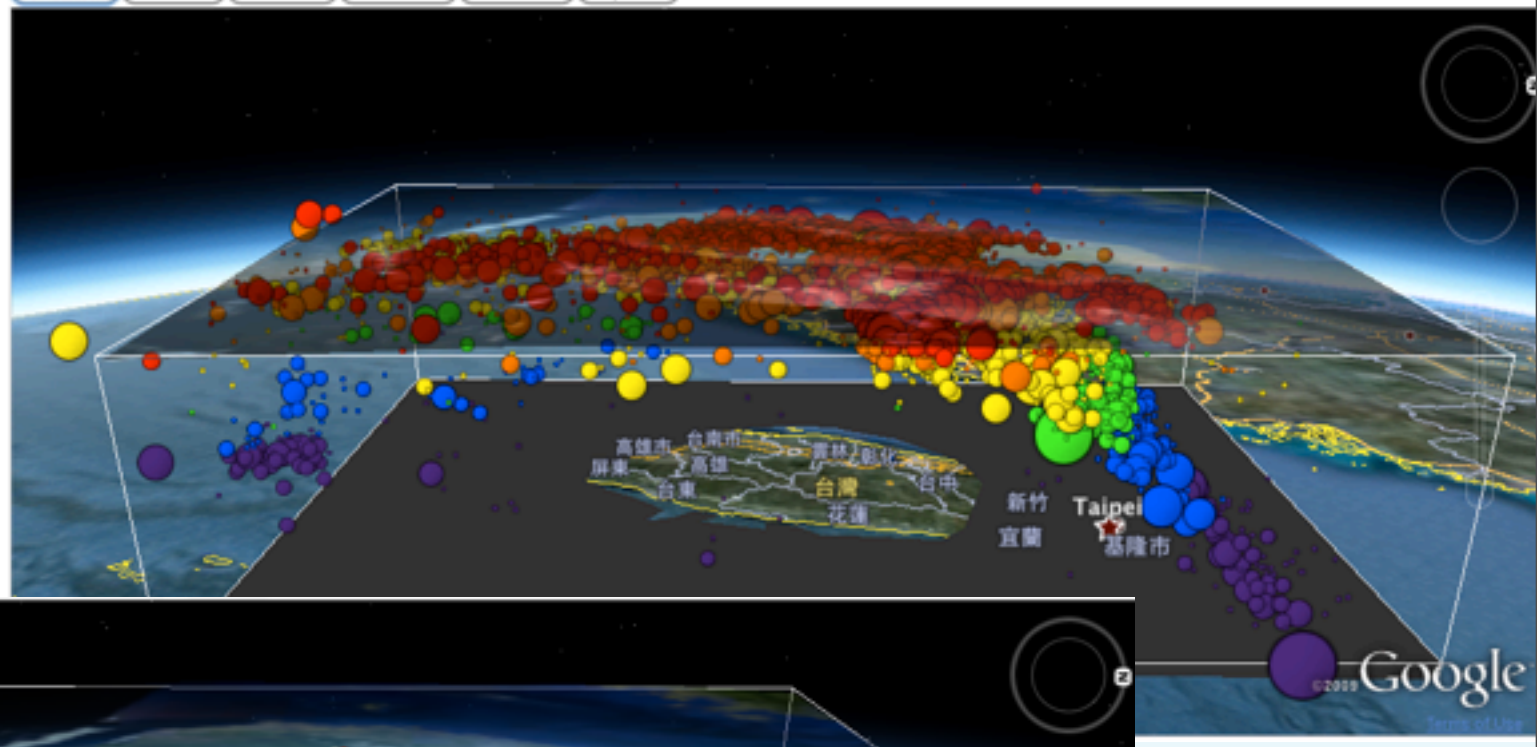
Taiwan Seismicity Events

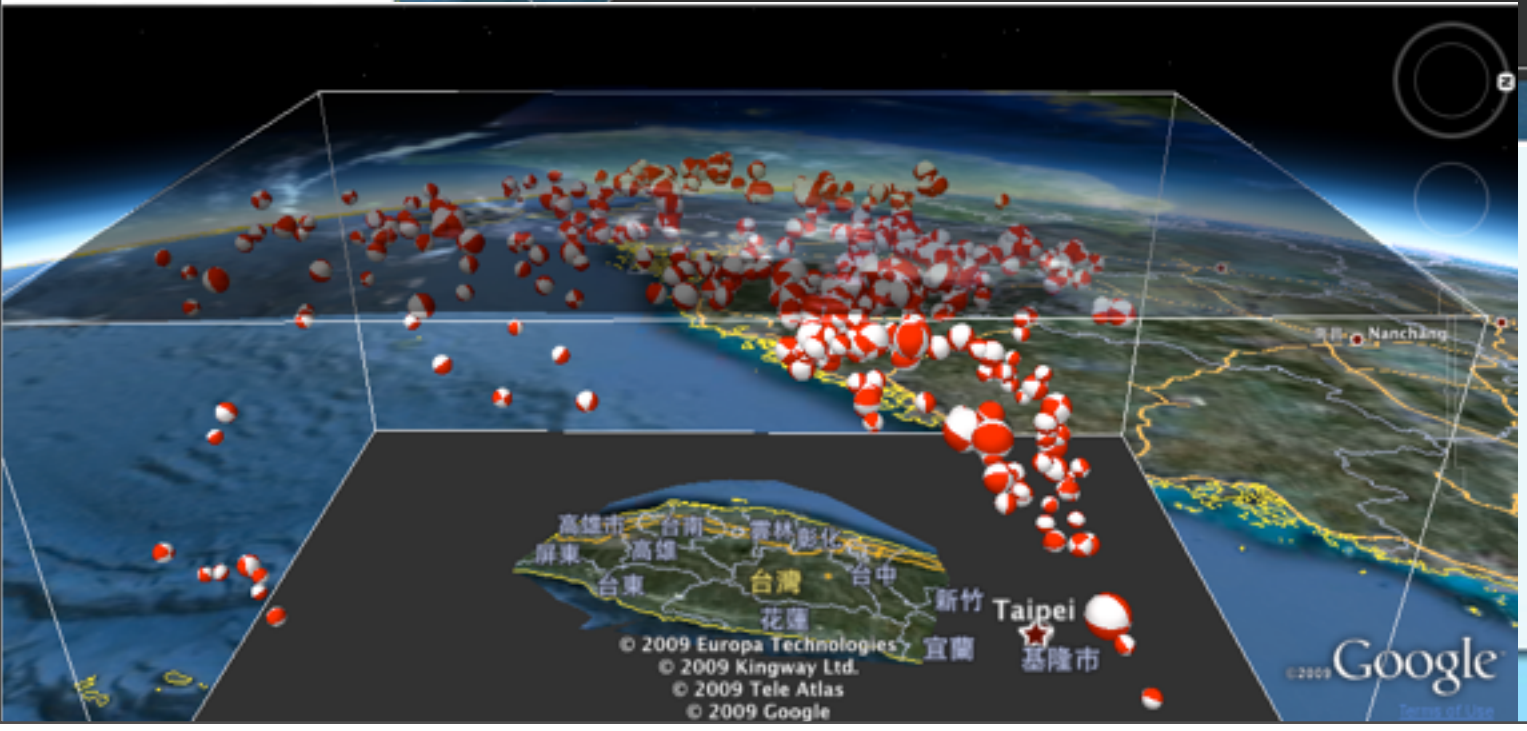
3D Virtualization Viewing

- Seismicity Events
  - Taiwan
    - Beach Ball
  - Tonga
  - Global
- Rising Sea Level Animation

Status Bar | Navigation Control | Grid | Overview Map | Scale Legend | Atmosphere | Mouse Navigation

Start View | East View | West View | South View | North View | Top View





Sy Architecture

Presentation Framework (Portal, GUI)

Job repository

Workflow Management

Virtual Queuing System

Grid Application Platform

Grid-based Digital Library

Metadata

AMGA & Grid Storage

Seismic Station Network

CWB

Sensors

IES,AS

VN, PH, IRIS, ID, etc

Single Server

Computing Cluster

Computing Grid

gLite

Grid-based Computing Pathway

IES or elsewhere

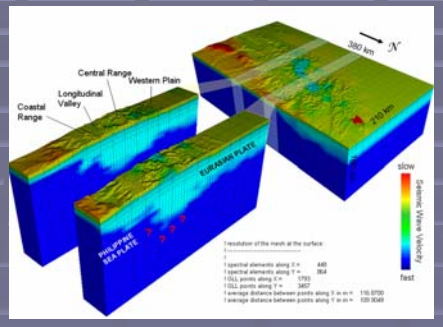
Problem definition

- Rupture model (source)
- Simulation region (Path and Site)
- Physical properties (Maximum frequency, Minimum Velocity and so on)

Input

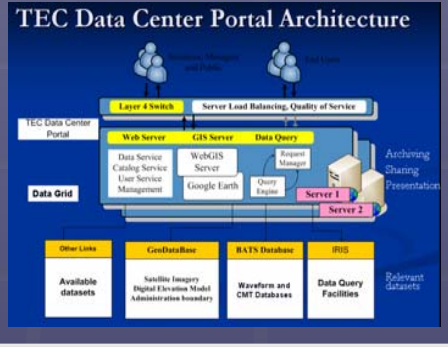
ASGC Grid Resource

Community models



Grid computing

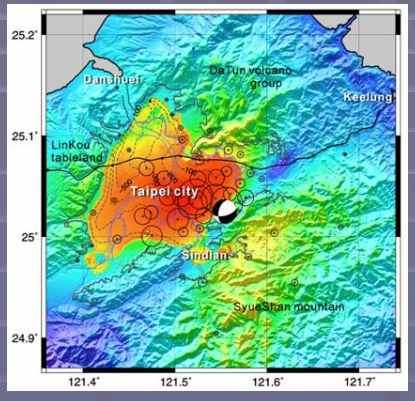
Data Grid



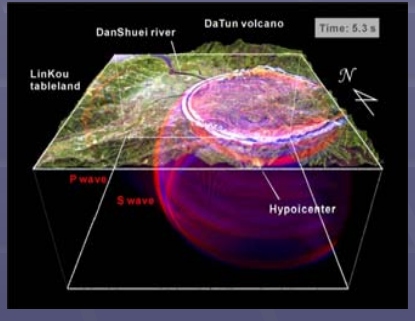
Numerical output

Visualization, Analysis Machines (ASGC, IES)

Hazard map



Numerical visualization

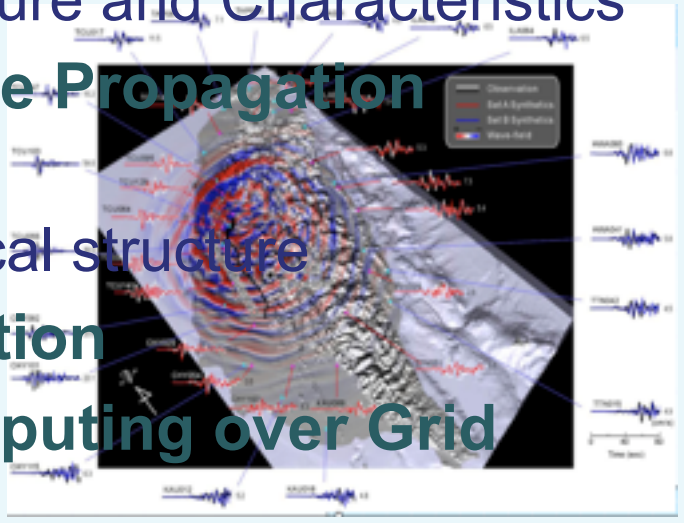


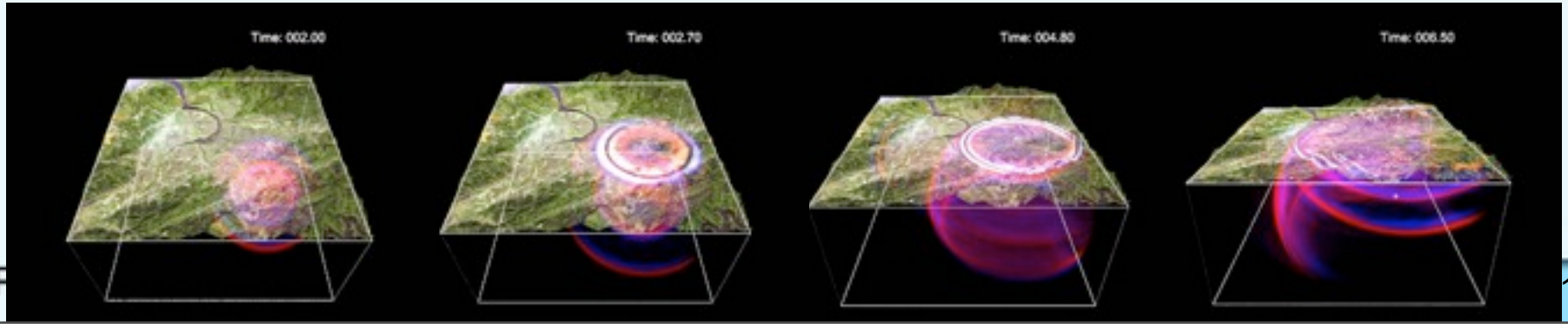
Storage

Result output

Scientific Values

- **Inversion of Earthquake Source Kinematics**
  - Rupture Process Analysis
  - Precise Regional Topographical Structure and Characteristics
- **Large Area Accurate 3D Seismic Wave Propagation Analysis Model**
  - At high freq and with complex geological structure
- **Earthquake Impact Analysis & Mitigation**
- **Optimization of massive parallel computing over Grid**
- **Data Federation**



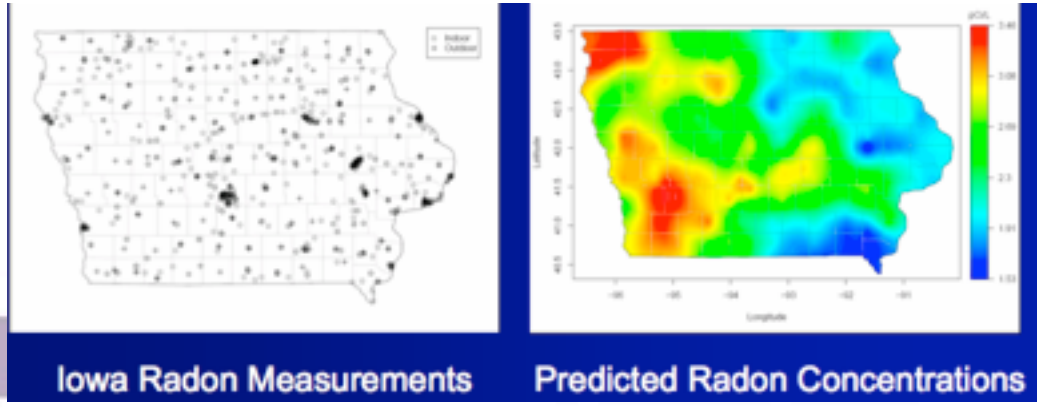


TeraGrid GI Science Gateway

- Features
  - Web Portal
  - Toolkit provides user-friendly capabilities to perform geographic information analysis using computational Grids, and help non-technical users directly benefit from accessing cyberinfrastructure capabilities.
- Modules
  - Random spatial point generator
  - Distance-weighted interpolation of surfaces
  - Cluster detection algorithm (Gi$*$)
  - Bayesian geostatistical spatial model using MCMC
    - Spatiotemporal model with separable correlation structure
- http://www.gisolve.org

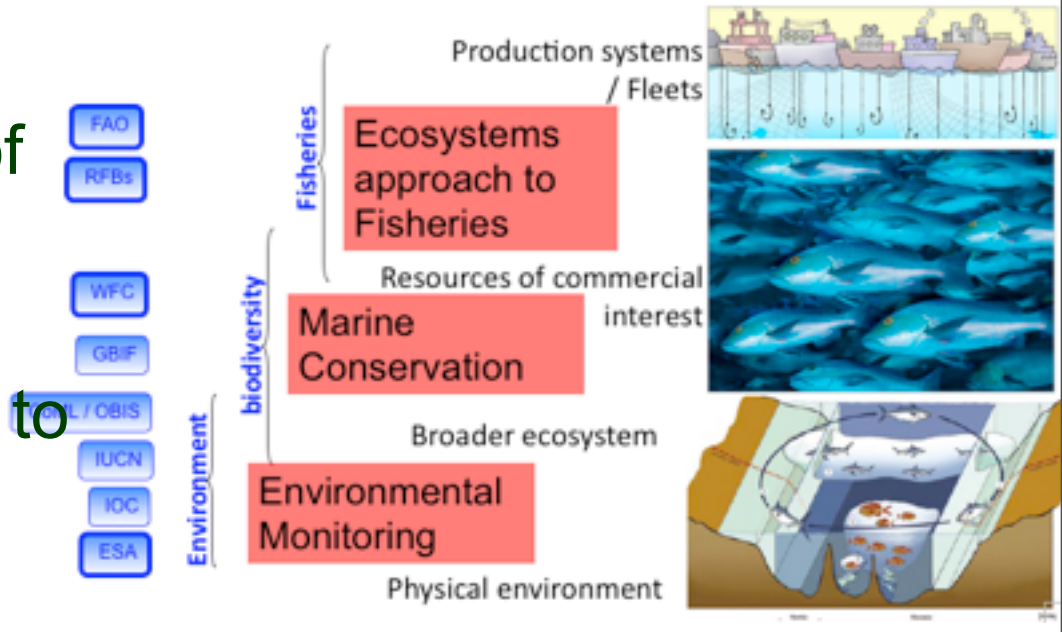
MCMC Case Study

- Parallelize MCMC-based Bayesian inference for geostatistical models
  - Methods
    - Parallelizing matrix calculations within single chains
    - Running parallel chains on Grids
- TeraGrid provides an ideal computing environment for solving this class of problems
- Initial results demonstrate significant performance gains



AquaMaps: Mapping Biodiversity Hotspots and Assessing Impacts of Climate Change

- Goals
  - Long-term protection of marine biodiversity
  - Implementation of Ecosystems Approach to Fisheries
- Solution - Species Distribution Modeling



... produce computer-generated, reproducable species range maps for (evenually) all species using available data and a transparent, easily understandable and modifiable approach, so maps can be reviewed and improved by species experts.

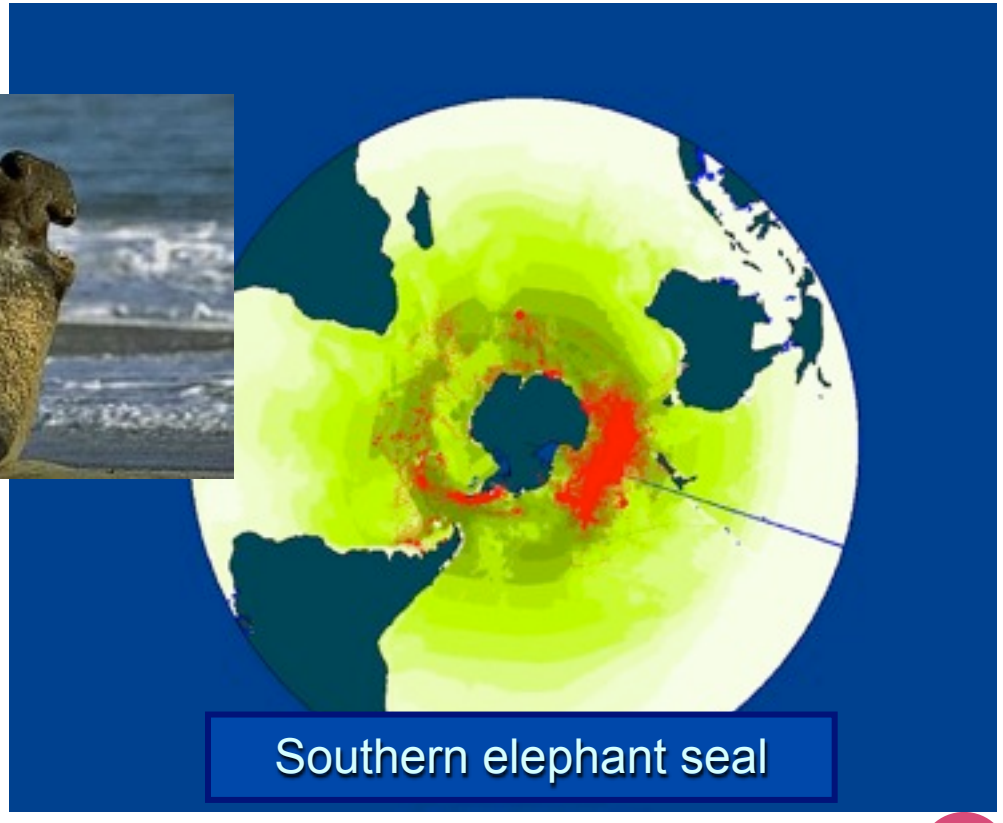
This is where we're going….

D4Science: Collaborative virtual laboratories (VREs) in support to science

- working environment with access to multidisciplinary data sources and chain workflow processes
- Facilitates control of data sharing and collaborative reporting
- Provides access to GRID Infrastructure, storage and computing powers to all regional fisheries bodies

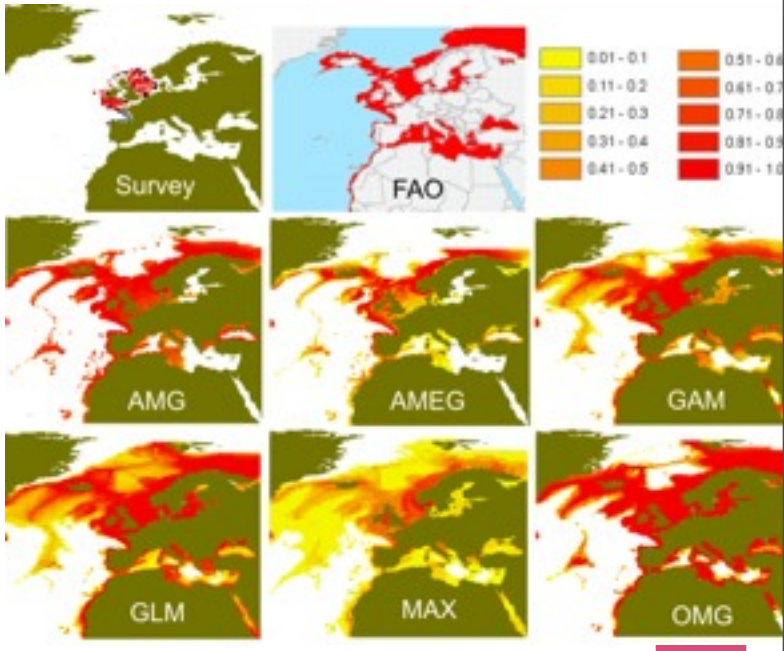
AquaMaps – how good is it?

- Validating individual species range maps
  - Kaschner et al, 2006



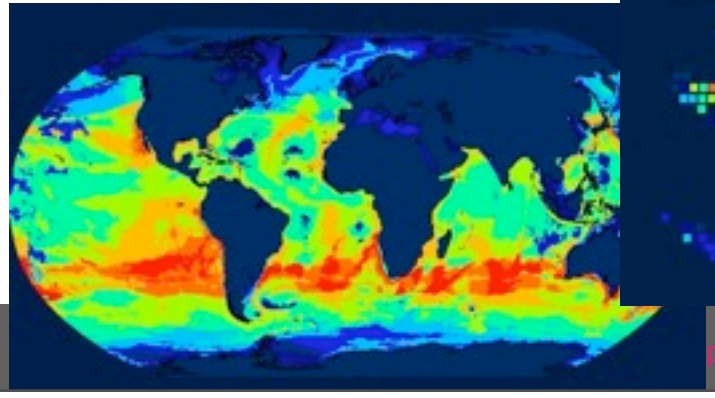
AquaMaps – how good is it?

- Validating individual species range maps
  - Kaschner et al, 2006
- Testing model performance in comparison to other approaches
  - J.Ready, K.Kaschner et al, accepted



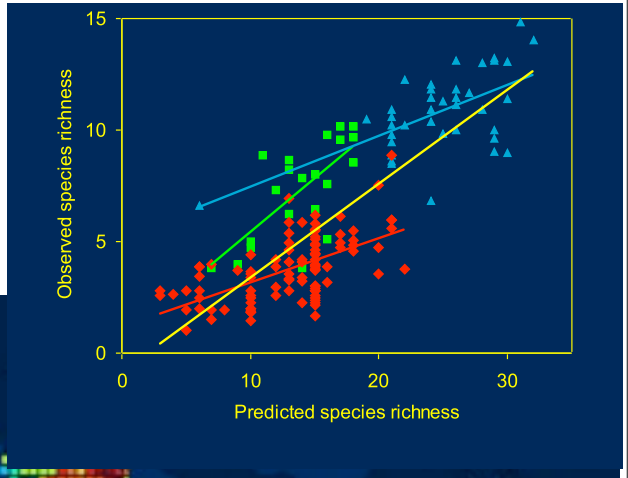
AquaMaps – how good is it?

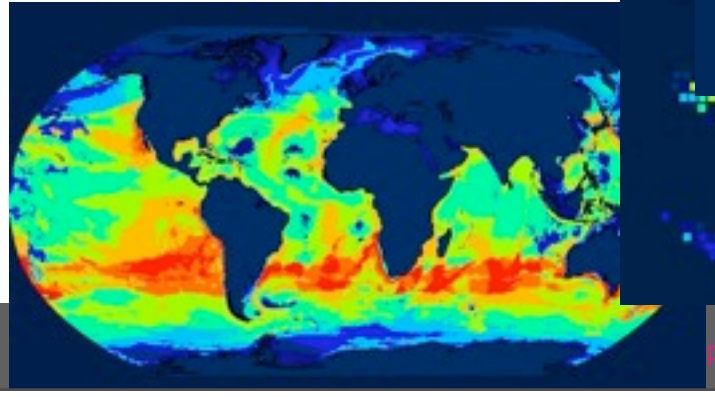
- Validating individual species range maps
  - Kaschner et al, 2006
- Testing model performance in comparison to other approaches
  - J.Ready, K.Kaschner et al, accepted
- Validating species richness maps
  - K.Kaschner et al, in prep

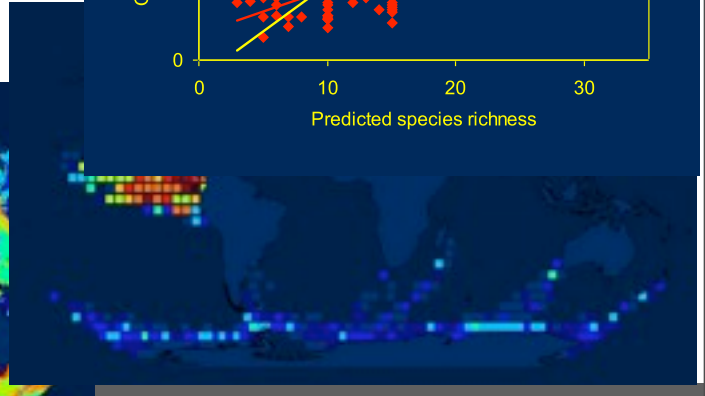


AquaMaps – how good is it?

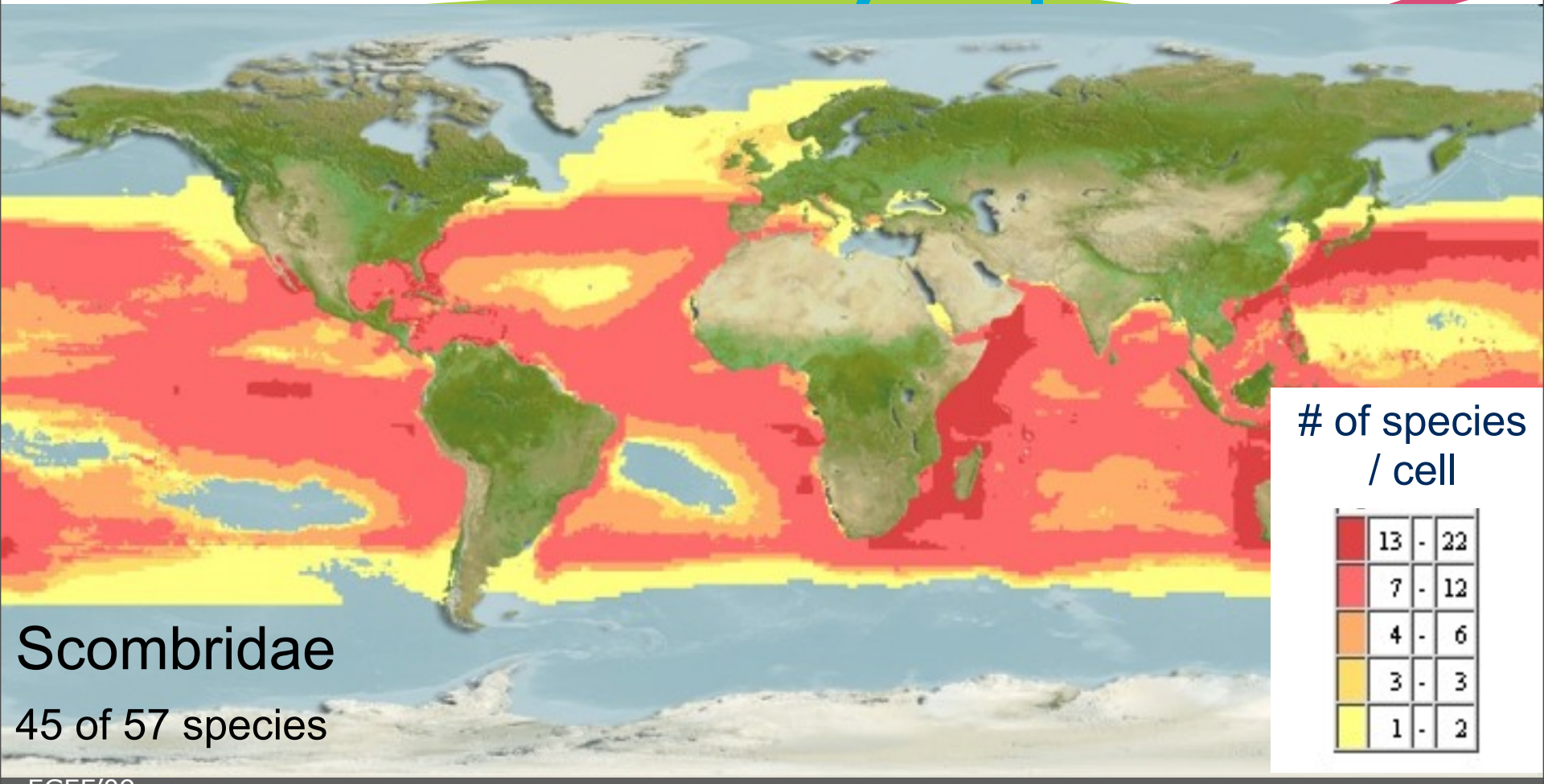
- Validating individual species range maps
  - Kaschner et al, 2006
- Testing model performance in comparison to other approaches
  - J.Ready, K.Kaschner et al, accepted
- Validating species richness maps
  - K.Kaschner et al, in prep



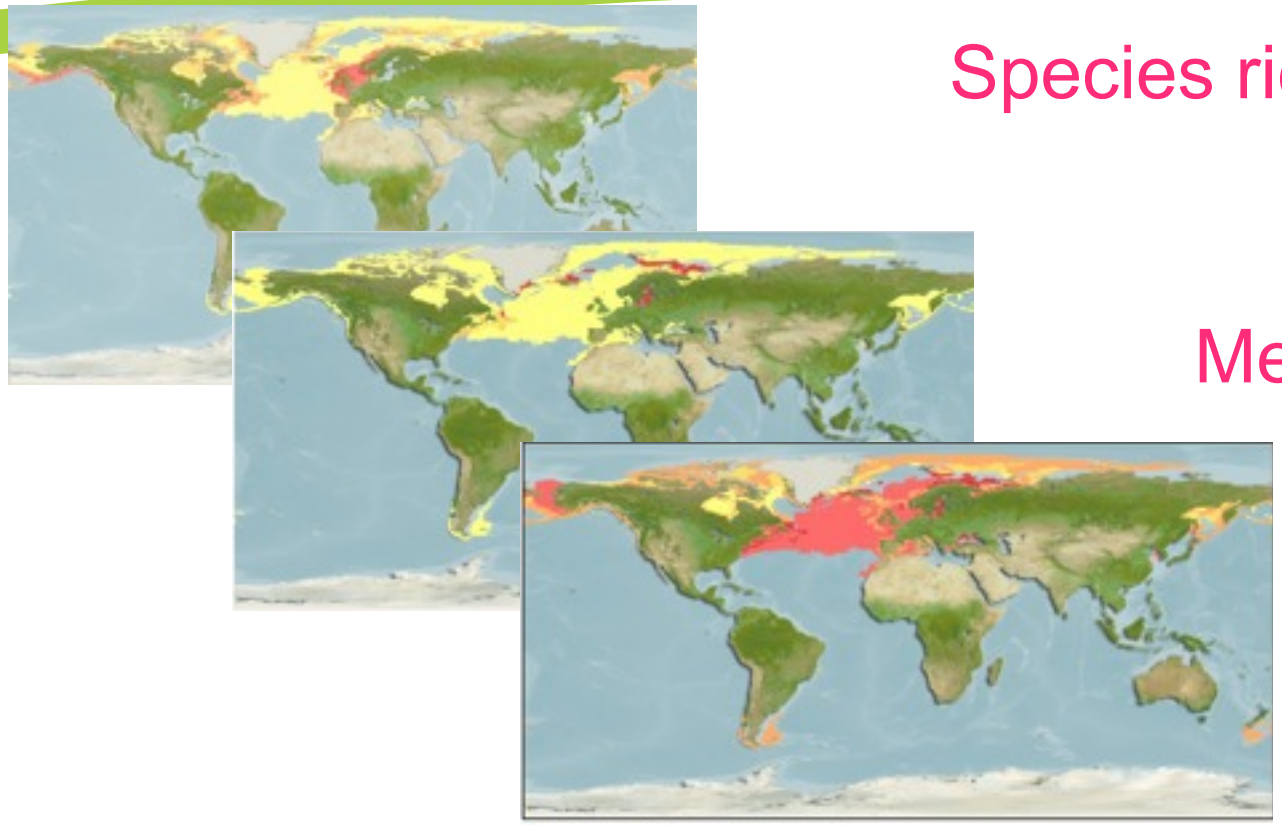




AquaMaps – what can we do with it?
Biodiversity Maps



AquaMaps – what can we do with it?
Biodiversity Maps



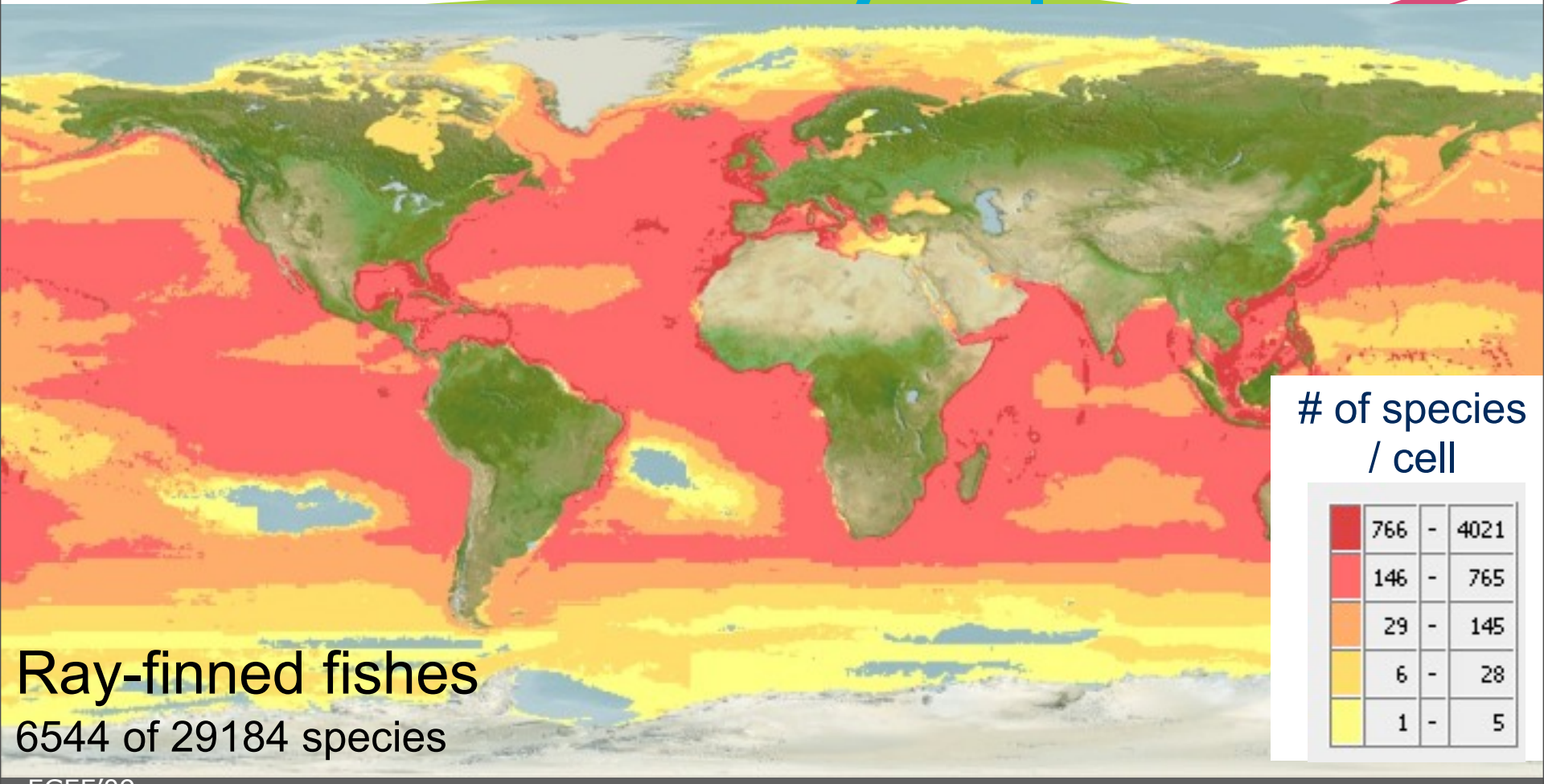
Species richness

Mean length

Mean trophic level

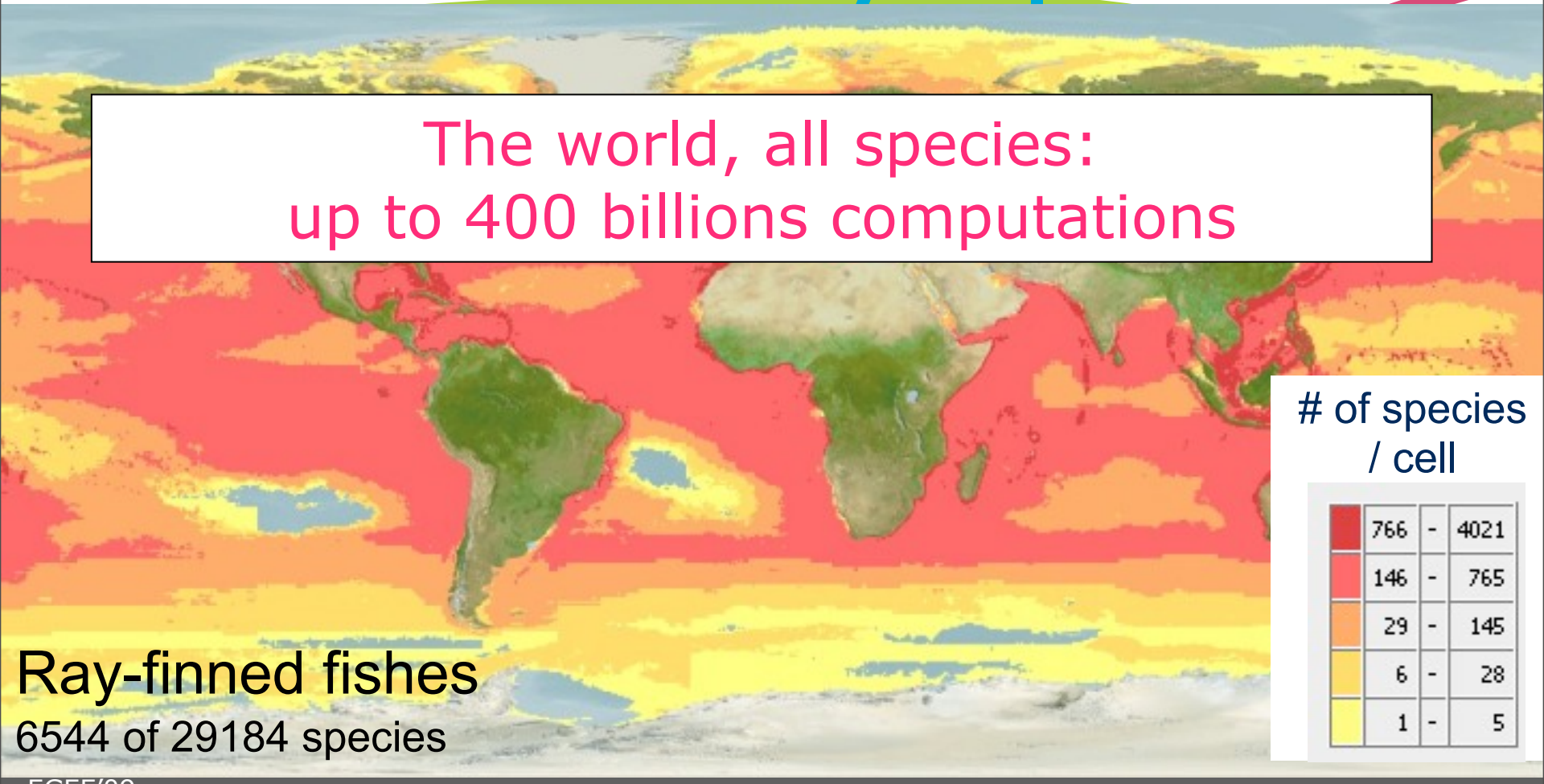
Gadidae: 23 of 25 species

AquaMaps – what can we do with it?
Biodiversity Maps



AquaMaps – what can we do with it?
Biodiversity Maps

The world, all species:
up to 400 billions computations



Ray-finned fishes
6544 of 29184 species

AquaMaps – what can we do with it?
Biodiversity Maps

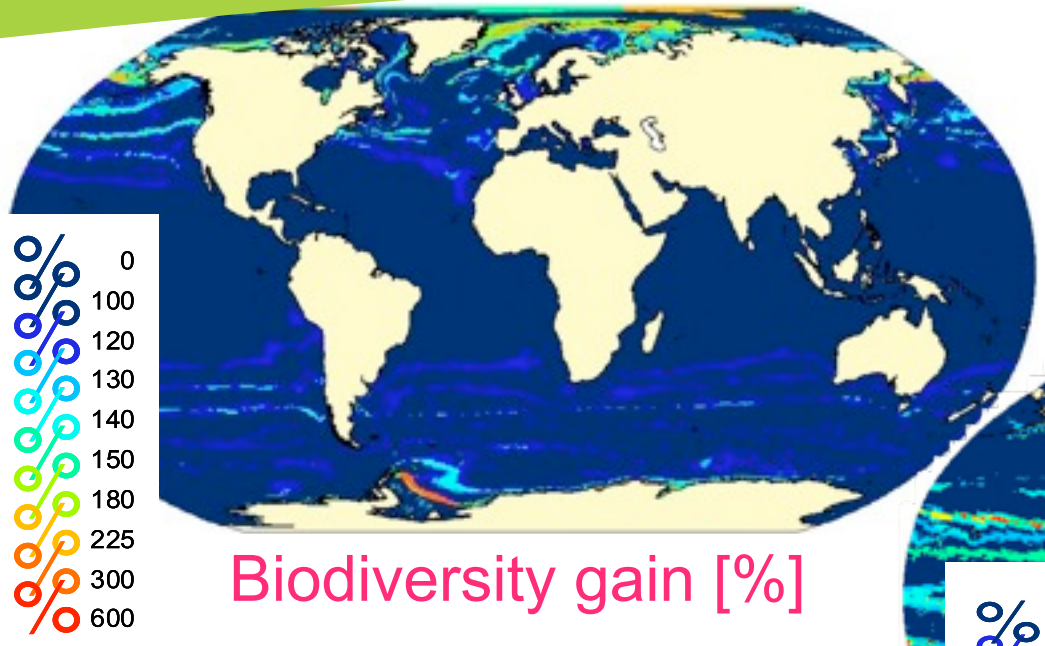
The world, all species:
up to 400 billions computations

# of species / cell

	766	-	4021
	146	-	765
	29	-	145
	6	-	28
	1	-	5

Ray-finned fishes
6544 of 29184 species

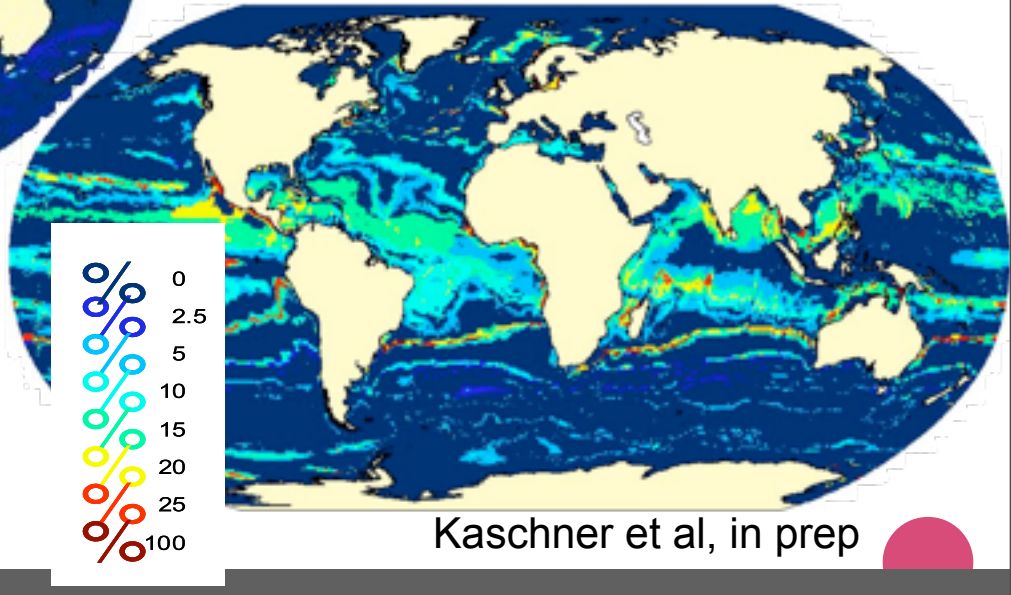
AquaMaps – what can we do with it? Modelling Impacts of Climate Change



Biodiversity gain [%]

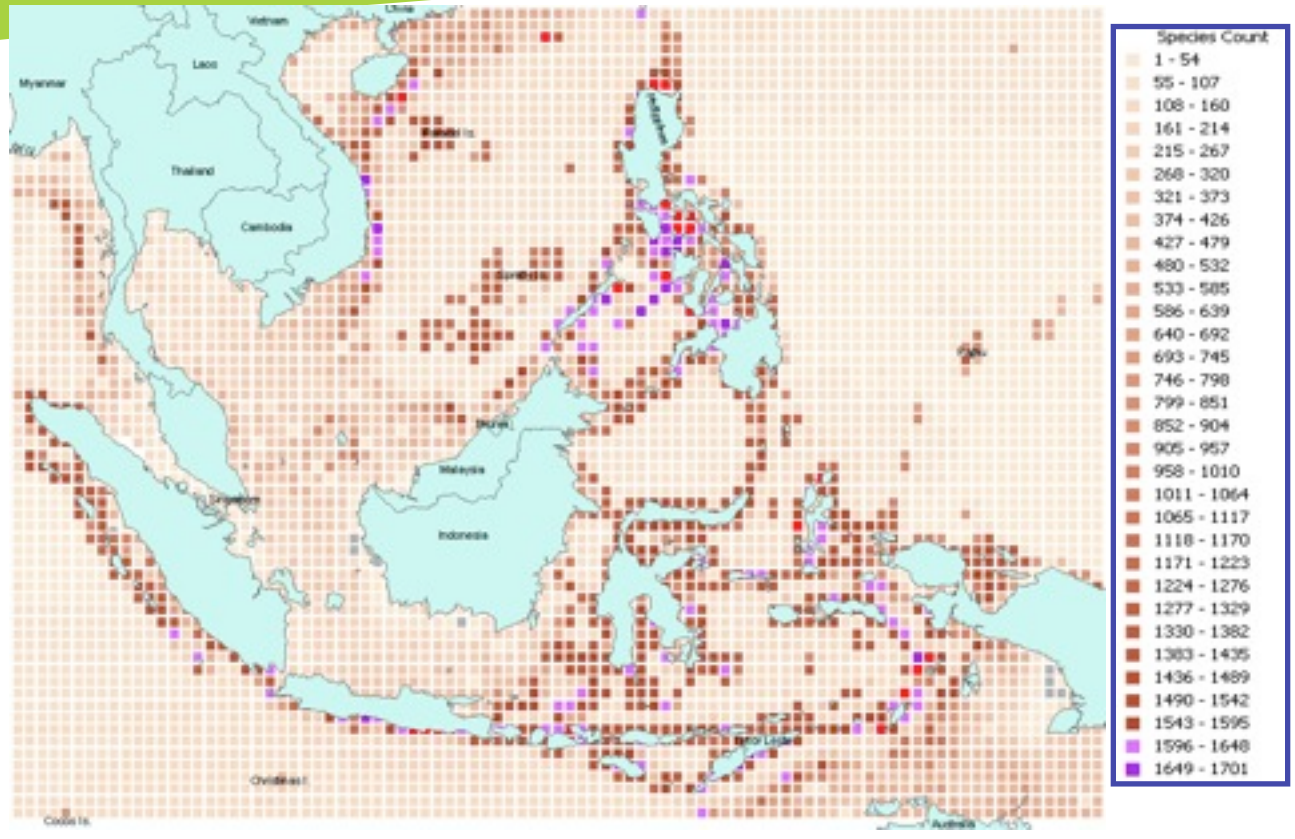
Marine Mammals (n = 115)

Biodiversity loss [%]



Kaschner et al, in prep

AquaMaps – what can we do with it? Modelling Impacts of Climate Change



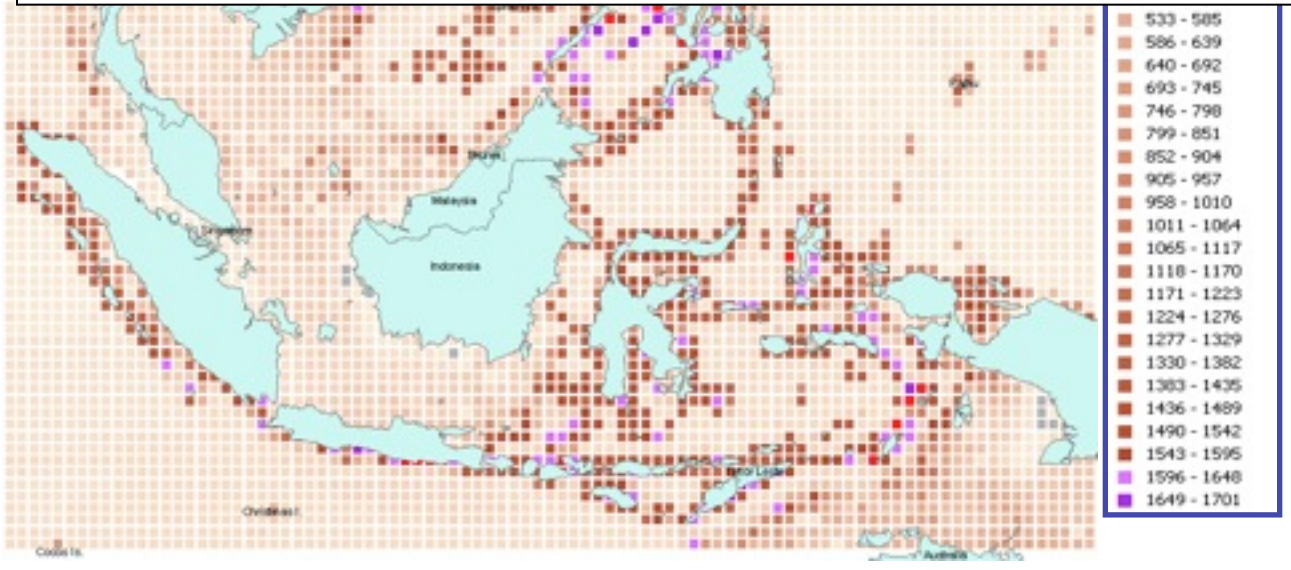
WorldFish Centre Project:

Hotspots, Fisheries and Climate Change in South China Sea

- 6,188 half degree cells
- 2,540 species
- 5+3 environmental parameters

AquaMaps – what can we do with it? Modelling Impacts of Climate Change

Local multispecies map, several climate scenarios:
up to 1 billion computations

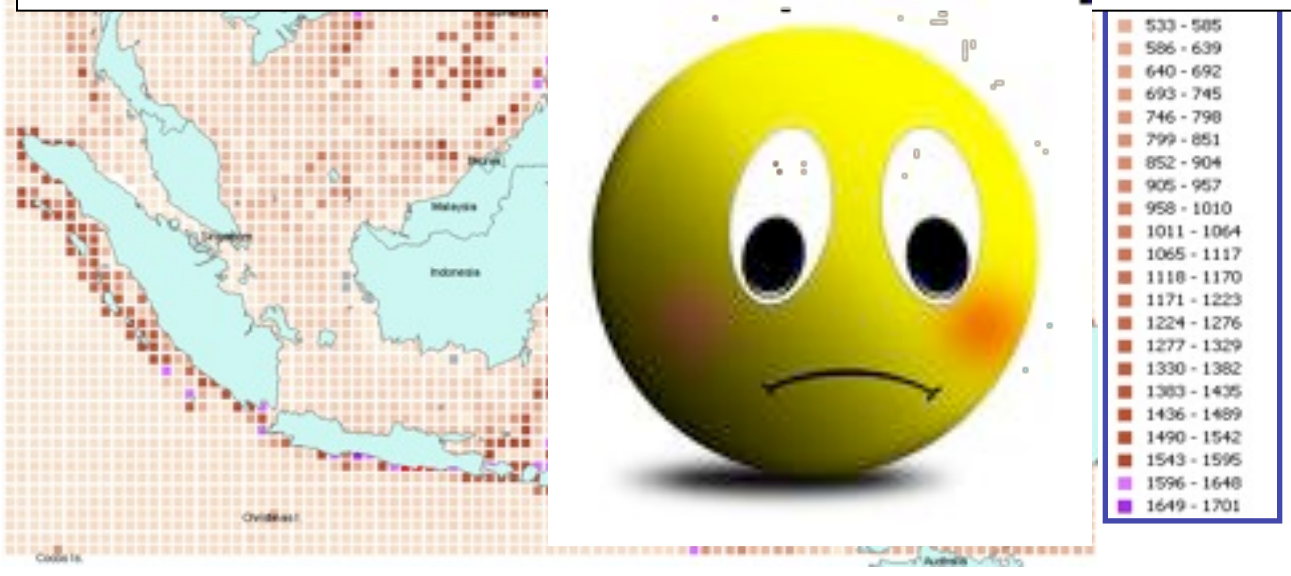


Hotspots, Fisheries and Climate Change in South China Sea

- 6,188 half degree cells
- 2,540 species
- 5+3 environmental parameters

AquaMaps – what can we do with it? Modelling Impacts of Climate Change

Local multispecies map, several climate scenarios:
up to 1 billion computations



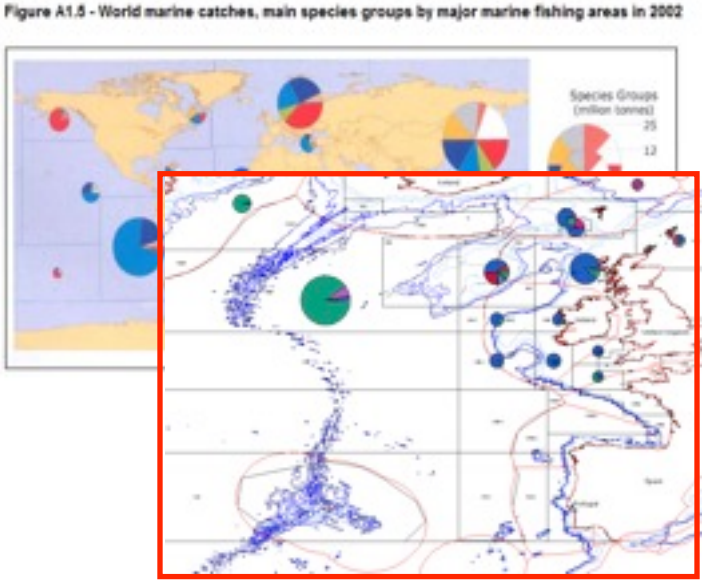
Hotspots, Fisheries and Climate Change in South China Sea

- 6,188 half degree cells
- 2,540 species
- 5+3 environmental parameters

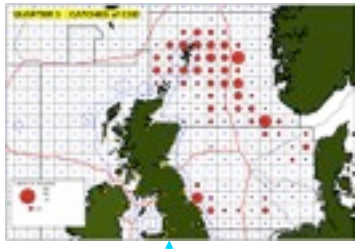
Implementing an Ecosystems Approach to Fisheries

Fishing activity / Catch

Fisheries



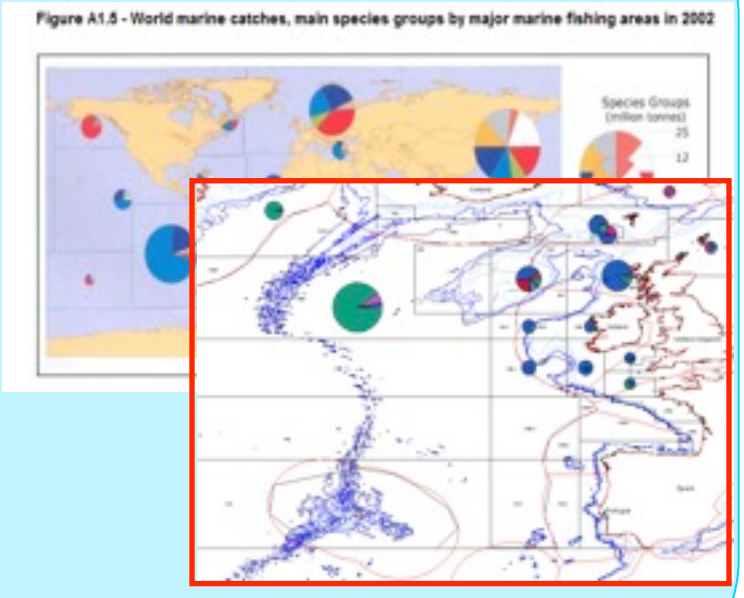
Implementing an Ecosystems Approach to Fisheries



Fishing activity / Catch

Fisheries

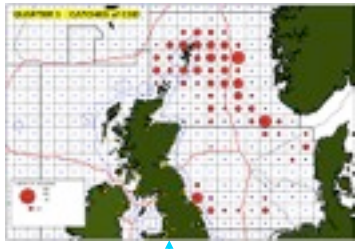
ICIS



Integrated Capture Information System

- product: harmonized and reallocated catch statistics

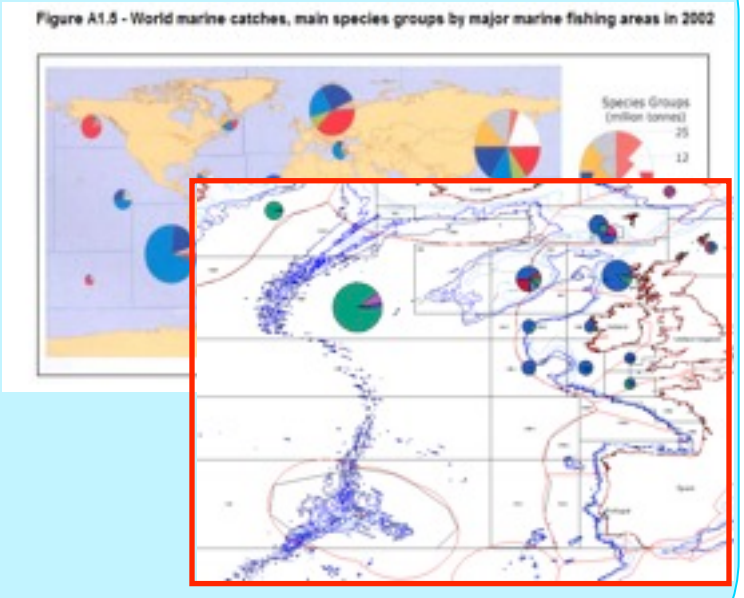
Implementing an Ecosystems Approach to Fisheries



Fishing activity / Catch

Fisheries

ICIS



Integrated Capture Information System

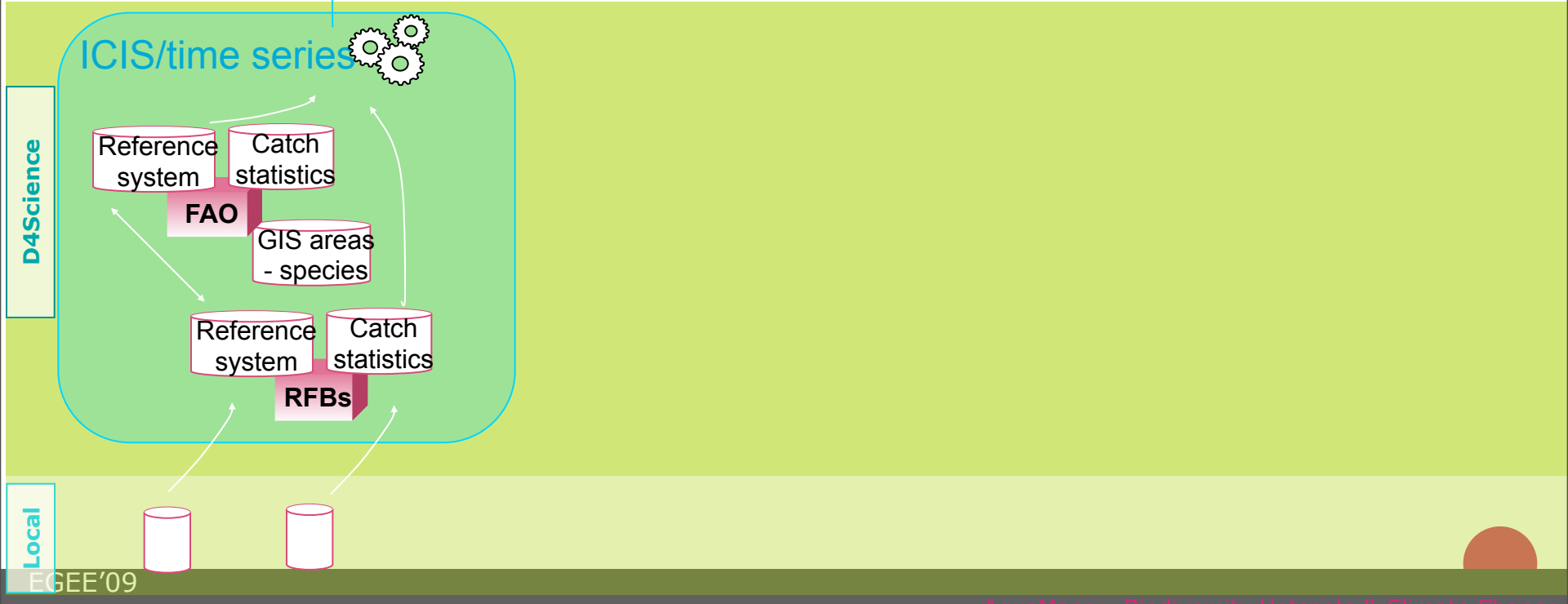
- product: harmonized and reallocated catch statistics

- Requirements
  - harmonization of time series data
  - querying, with aggregation and reallocation rules
  - combining biodiversity information with fisheries Catch time series
  - spatial dimension and mapping (GIS)

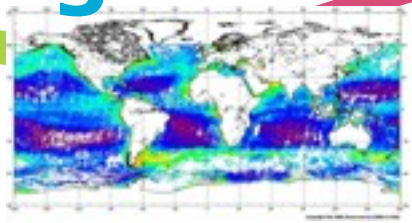
This is where we're going....

Fishing activity / Catch

Fisheries **Biodiversity** **Oceanography**



This is where we're going....



Fishing activity / Catch

Vulnerable Marine Ecosystems

Habitats Geo-forms Hydrography

Fisheries

Biodiversity

Oceanography

D4Science

ICIS/time series

Reference system

Catch statistics

FAO

GIS areas - species

Reference system

Catch statistics

RFBs

Dynamic Aquamaps

Aquamaps

WFC

FishBase

SealifeBase

UBC

Species occurence

OBIS

Environmental monitoring

end-users services

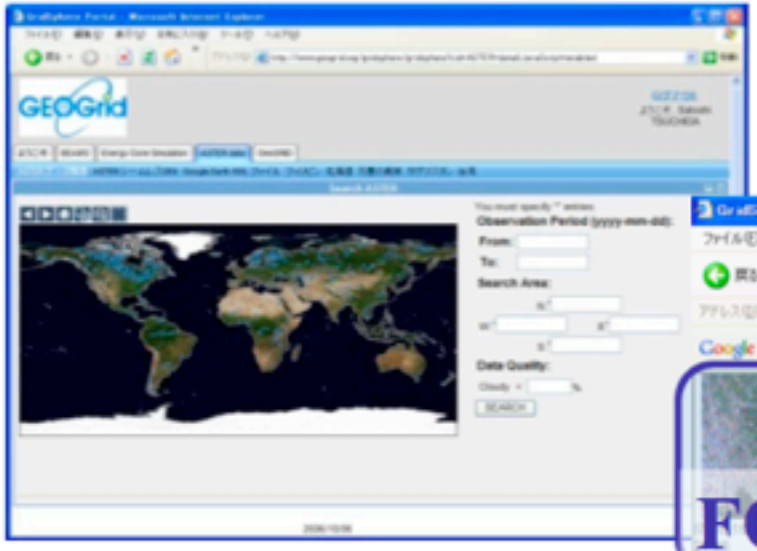
G-POD

ESA

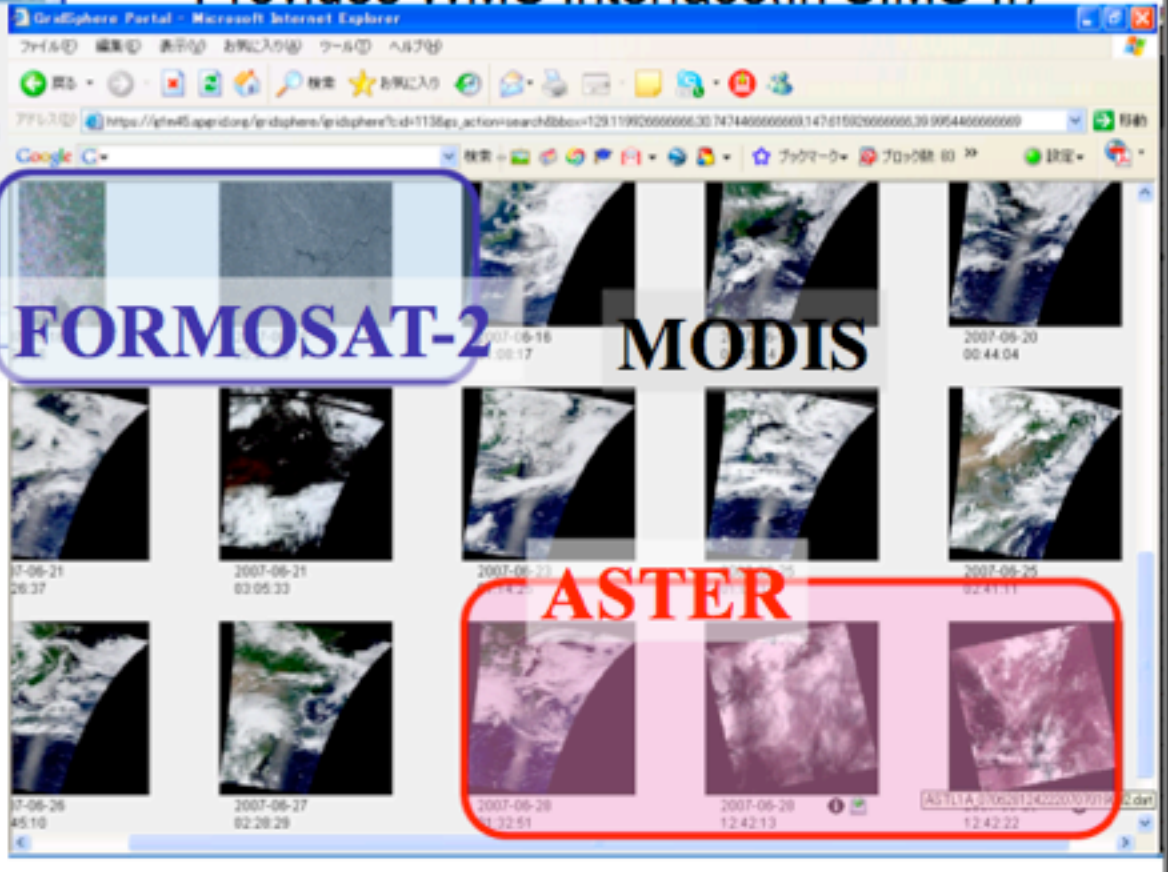
Satellite oceanography

Local

SIMS: VO based User Interface for Federating Distributed Databases



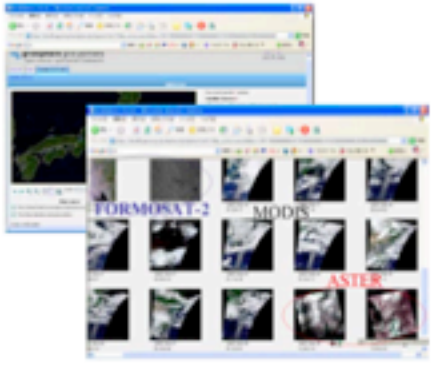
- User interface is constructed as JSR168 porlet on GridSphere
- Provides WMS Interface(in SIMS-II)



Within the Service

- Access database using OGSA-DAI Java API
- Submit image analysis via GRAM
- Retrieve input data from GridFTP server

Database Federation



A Web Map Service

Java Program
Integration Framework with OGSA-DAI
OGSA-DAI Client

SIMS portlet
- query data
- create web page which shows thumbnail images

SQL
SQL
SQL
SQL
SQL

Application Server
Globus
OGSA-DAI
VOMS
SQL w/ JDBC
Database Server (Sybase)
FORMOSAT-2

NSPO@TW

VOMS
Globus
OGSA-DAI
OGSA-DAI
SQL w/ JDBC
Database Server (PostgreSQL)
ASTER
MODIS

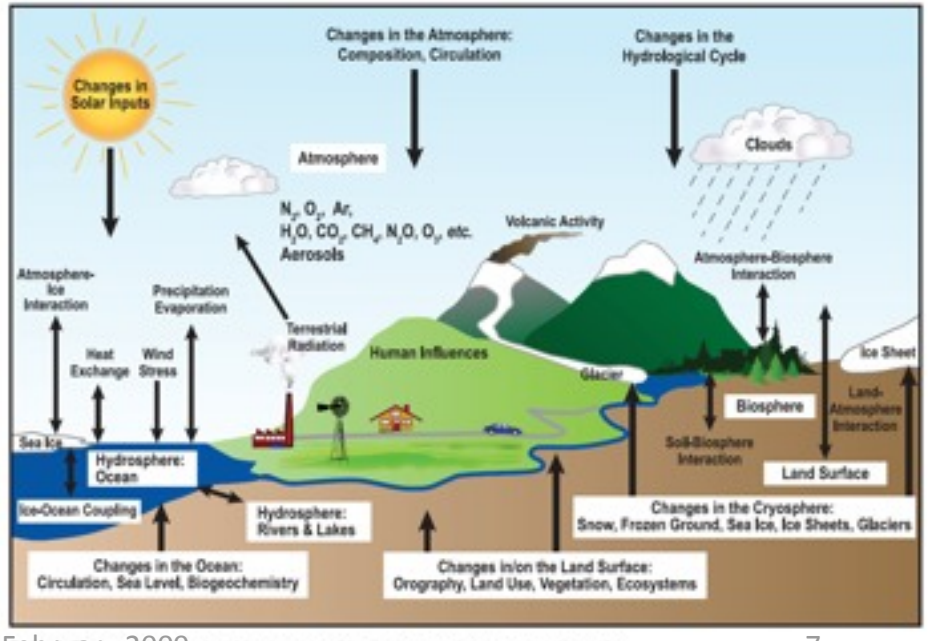
AIST@JP

Nature of Climate Data

- Global
- Long time series (modern and paleo)
- Multi-disciplinary: Atmosphere-Hydrosphere-Lithosphere-Cryosphere- Biosphere
- Observation and Simulation

→Tremendous amount of data

→ Data amount increases astronomically … (giga -> tera bytes)



Nature of Climate Change Research

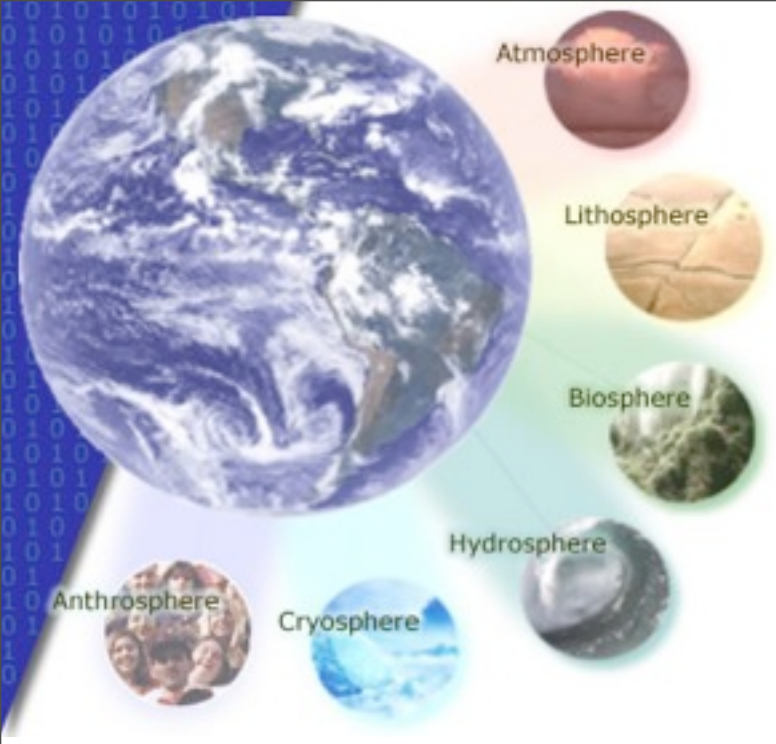
- New Science in Development
- Earth System Science + Human Dimension
- Expanding Interdisciplinary Science

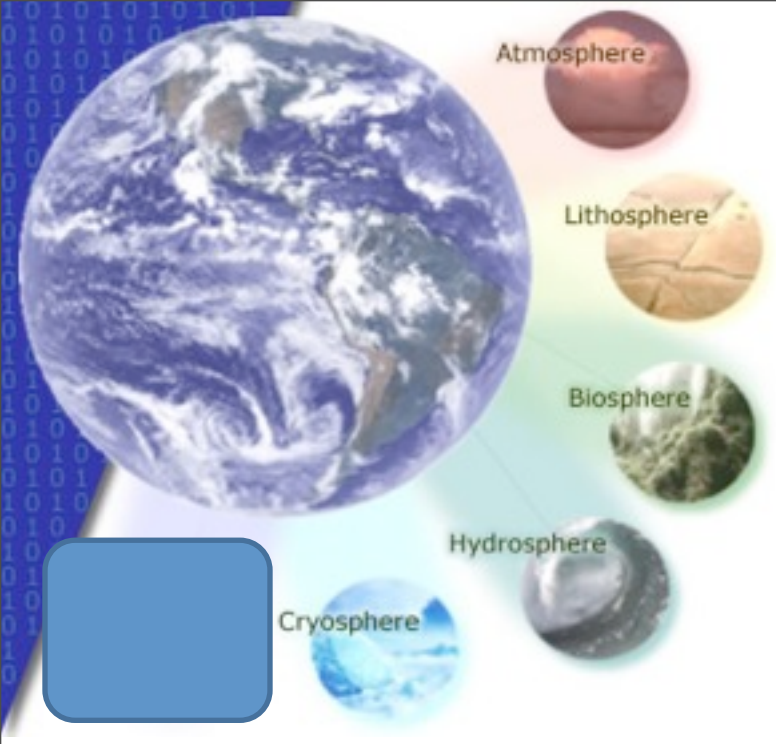
Nature of Climate Change Research

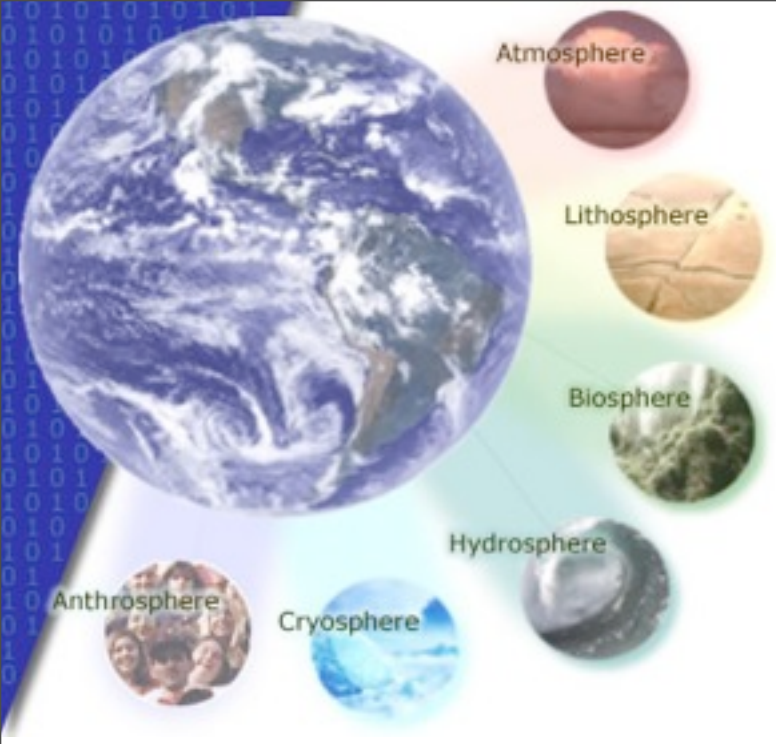
- New Science in Development
- Earth System Science + Human Dimension
- Expanding Interdisciplinary Science

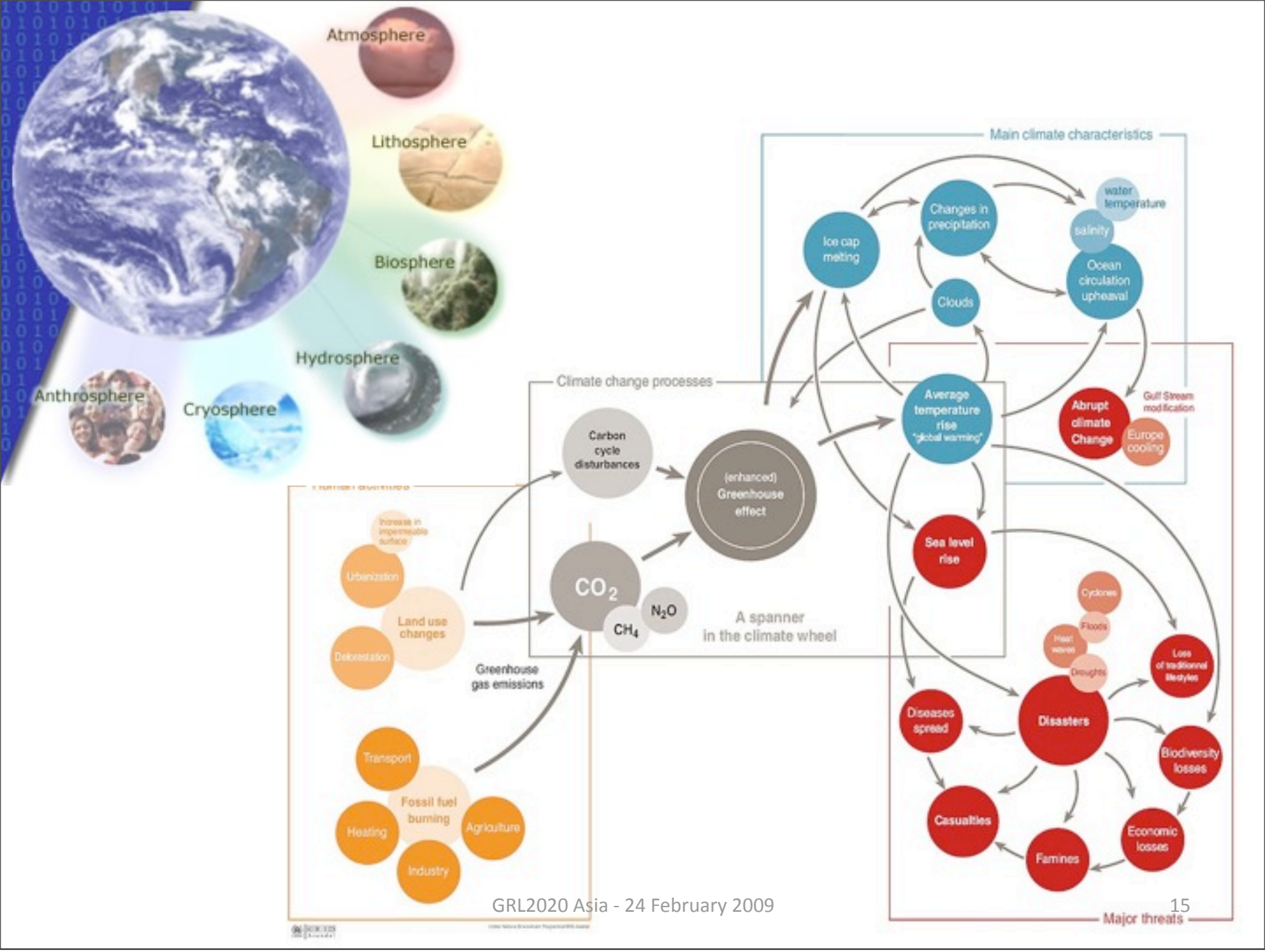
Future Challenge:

- How to combine digital libraries of different disciplines into truly Multi-disciplinary Global Digital Libraries for Earth System/Climate Change Research.


Atmosphere
Lithosphere
Biosphere
Hydrosphere
Anthrosphere
Cryosphere


Atmosphere
Lithosphere
Biosphere
Hydrosphere
Cryosphere


Atmosphere
Lithosphere
Biosphere
Hydrosphere
Anthrosphere
Cryosphere


Atmosphere
Lithosphere
Biosphere
Hydrosphere
Anthrosphere
Cryosphere
Main climate characteristics
Ice cap melting
Changes in precipitation
water temperature
salinity
Ocean circulation upheaval
Clouds
Climate change processes
Carbon cycle disturbances
(enhanced) Greenhouse effect
Average temperature rise "global warming"
Abrupt climate Change
Gulf Stream modification
Europe cooling
Sea level rise
Increase in impermeable surface
Urbanization
Land use changes
Deforestation
CO_2
N_2O
CH_4
A spanner in the climate wheel
Greenhouse gas emissions
Transport
Fossil fuel burning
Heating
Agriculture
Industry
Cyclones
Floods
Heat waves
Droughts
Loss of traditional lifestyles
Diseases spread
Disasters
Biodiversity losses
Casualties
Economic losses
Famines
Major threats

Summary

- e-Science envisages a whole new way of doing collaborative science
- For the sustainable Grid e-Infrastructure, we have to focus more on community building rather than just offering technologies.
- Asia Pacific Region has great potential to adopt the e-Infrastructure :
  - More and more Asia countries will deploy Grid system and take part in the e-Science world
  - However, applications of and for the Asia Pacific scientists are largely in lack which is crucial!!
- Extending from EGEE Asia Federation to EUAsiaGrid, we are widening the uptake of e-Science, by the close collaboration regionally and internationally