Sequential Data Assimilation: Online Information Fusion Platform for Simulation and Observation Data

Research Organization of Information and Systems

*The Institute of Statistical Mathematics/JST CREST**

Tomoyuki Higuchi

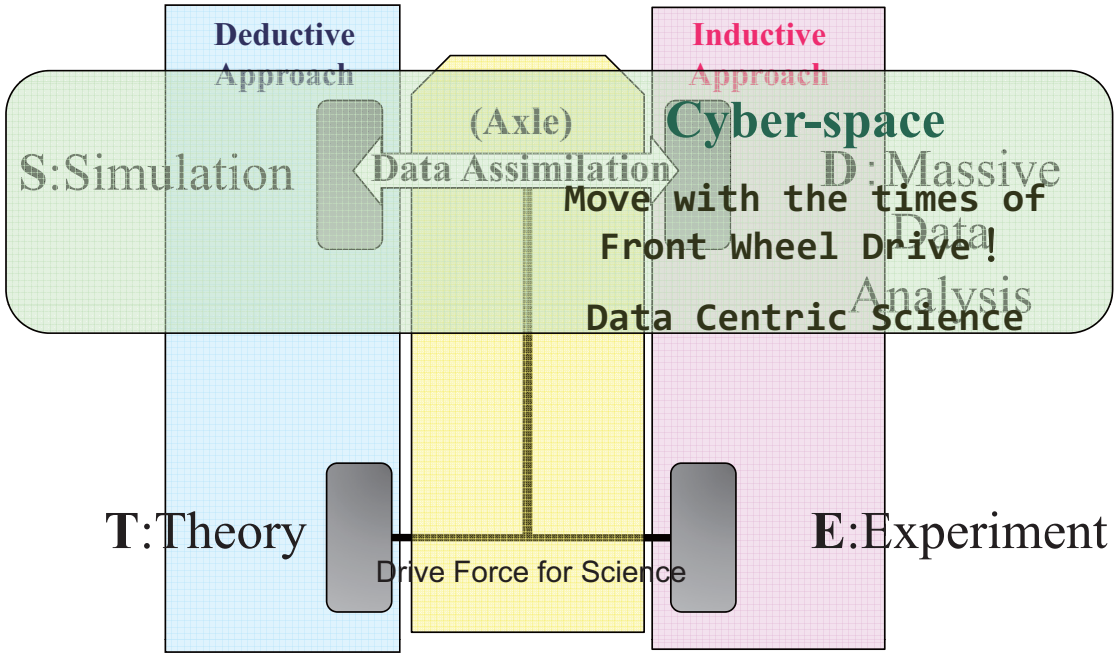
独立行政法人 科学技術振興機構

*Japan Science Technology Agency

Core Research for Evolutional Science and Technology

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

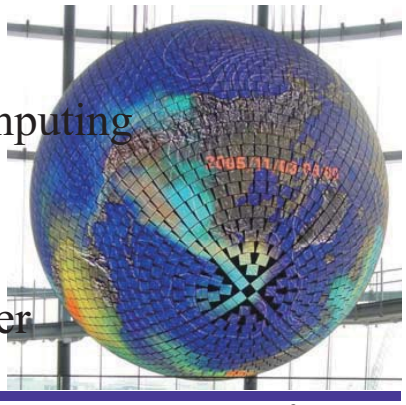
TESD: Four Kind of Methodology of Science



大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Outline

1. Simulations with uncertainties
2. Data Assimilation (DA)
3. Modeling uncertainties
4. Sequential DA and generalized state space model
5. Ensemble-based nonlinear filtering methods
   1. Ensemble Kalman filter
   2. Particle filter
6. Applications with peta-scale computing
   1. Tsunami Simulation model
   2. Ocean Tide Simulation
   3. Genome Science
7. Next generation of supercomputer
8. Conclusions



統計数理研究所

Construction of Simulation Model

（simplified meteorological model around Japan）

PDE to approximate real ph
(continuous time/s

$\frac{\partial x}{\partial t} = cx^2 + \cdots$

PDE : Partial d

Discrete simulation
(discrete time/spac

$x_t = F(x_{t-1})$

(time v
Boun

Nonlinear state spac
(discrete time/space, stoc

$x_t = F(x_{t-1}, v_t)$

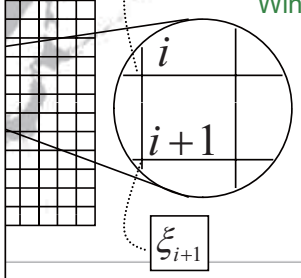
$$x_t = \begin{bmatrix} \xi'_1 \\ \vdots \\ \xi'_i \\ \xi'_{i+1} \\ \vdots \\ \xi'_K \\ \theta \end{bmatrix}$$

variable vector ξ_i is assigned
rid point.

temperature

$\xi_i = (T_i, U_i, V_i)$

Wind vector

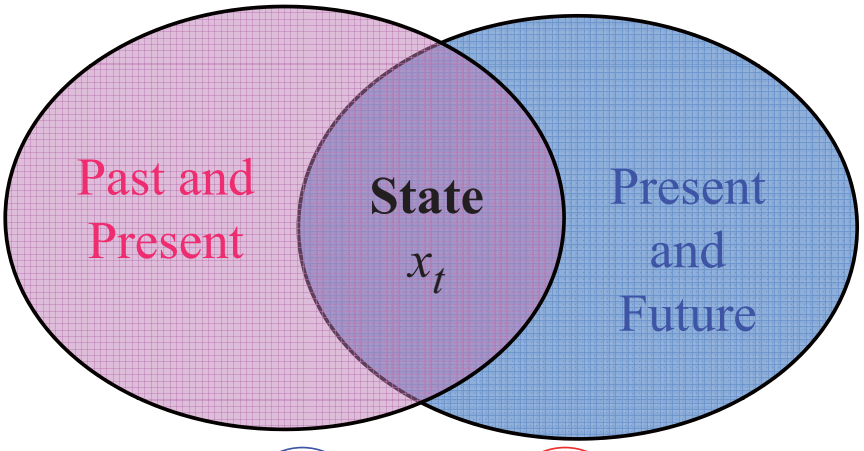


n points and observed
e limited.

4

統計数理研究所

State Vector : Contact point between past and future



State of time t ← $x_t = F(x_{t-1})$ → State of time t-1

Simulation Model

統計数理研究所

Simulation

Simulation model

$$x_t = F_t(x_{t-1})$$

time $n \cdot \delta t \to t$

$t = 0,1,2,\ldots,n,\ldots$

δt : simulation time step

x_t: State vector (simulation variables)

x_0 : Initial condition, given

When $x_0 = \alpha$,

$$x_1 = F_1(\alpha)$$
$$x_2 = F_2(x_1)$$
$$\vdots$$
$$x_T = F_T(x_{T-1})$$

Sequence $(x_1, x_2, \cdots x_T)$ is obtained deterministically.

統計数理研究所

Simulation including uncertainty

Simulation model

$$x_t = F_t(x_{t-1})$$

x_0 : Initial condition

When $x_0 = \alpha$,

$$x_1 = F_1(\alpha),$$
$$x_2 = F_2(x_1),$$
$$\vdots$$
$$x_T = F_T(x_{T-1})$$

with uncertain evolution

$$x_t \approx F_t(x_{t-1})$$ BC, parameters, …

with uncertain IC

When $x_0 \approx \alpha$,

$$x_1 \approx F_1(\alpha),$$
$$x_2 \approx F_2(x_1),$$
$$\vdots$$
$$x_T \approx F_t(x_{T-1})$$

Sequence $(x_1, x_2, \cdots x_T)$ should be evaluated probablistically.

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

What is Data Assimilation ?

- Emerging subject in meteorology and oceanography.
- Methodology to synthesize numerical simulation model and observed data
  - Simulation model can not represent real phenomena accurately.
    - (e.g.) Accurate weather forecast needs good initial conditions.
    - Uncertainty in the model (boundary condition, initial condition, unknown parameters, unknown dynamics...) exists.
  - Observation data have some physical/budgetary restrictions.

Correct variables in numerical simulation model using observation data.

= Data Assimilation

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Objects of Data Assimilation from a viewpoint of Meteorology and Oceanography

[1] To produce the best (better) **initial condition** for forecasting. It is actually realized in the real weather forecast (ex., Japan Meteorological Agency).

[2] To find the best (better) **boundary condition** in constructing a simulation model. This procedure includes a setting of appropriate boundary conditions necessary for dealing with a coupled phenomena.

[3] To attain an optimal **parameter** vector that appears in an empirical law (scheme) employed for describing complicated phenomena which possesses the different time and spatial scales. A **validation** of the empirically given values is regarded as this problem.

[4] To inter/extrapolate (estimate) an physical quantity at times and locations without observations based on a numerical simulation model. This procedure is called "a **generation of re-analysis dataset** (product)". This dataset is used to discover a new scientific findings by general geophysical researchers.

[5] To conduct an experiment with a virtual observation network and perform a **sensitivity analysis** in an attempt to construct an effective observation network system with less budgetary cost and less consuming time.

(ex. Kamachi et al., 2006)

Modeling uncertainties

•Represent a wide variety of uncertainty in a research target by distribution function.

•Understand a complex targets, NOT from its simple statistics such as mean, BUT from its distribution directly.

Notion of Probability: The machinery of probability theory is used to describe the uncertainty in model parameters or choice of model itself.

Probability theory provides a framework for quantification and manipulation of uncertainty. We will introduce a basic concept of probability theory next.

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Bayesian View
Central Role in Pattern Recognition and Machine Learning

It expresses how probable the observed dataset is for different settings of the parameter vector X.

Data dist.（likelihood function） Prior dist.

Posteiror dist.

$$p(x \mid y) \equiv \frac{p(y \mid x) \cdot p(x)}{p(y)} \propto p(y \mid x) \cdot p(x) = p(x, y)$$

Bayes' Theorem

It is independent of data Y, and describes a conviction degree against X numerically.

X: Parameter vector

Y: Data

Probability of data (Since data is given (the actually observed one), it takes some value.)

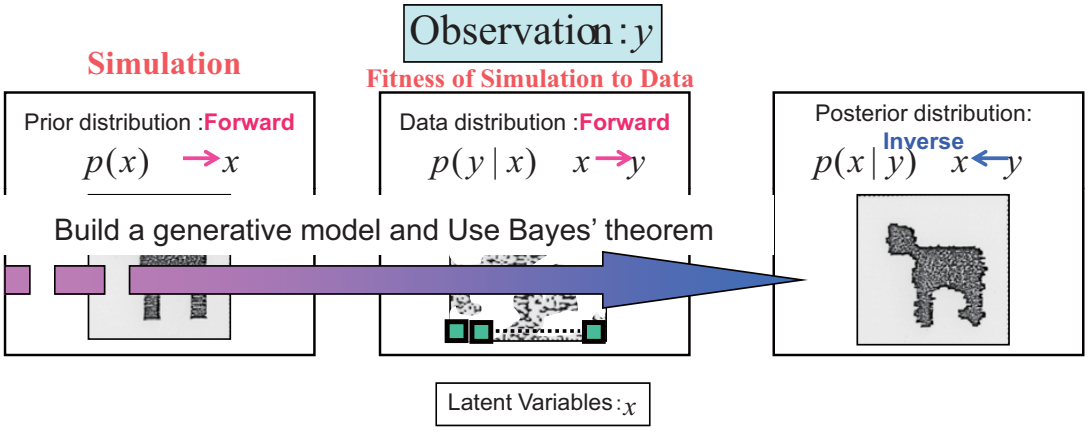
Joint dist.

We are interested in estimating a posterior distribution in most of circumstances.

We would like to be able to quantify our expression of uncertainty and make a precise revisions of uncertainty in the light of new evidence, as well as subsequently to be able to take optimal actions/decisions as a consequence.

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Generative Model, Inversion with Bayes' theorem, and Data Assimilation



大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Data Assimilation in Generalized State Space Model

State Vector (Simulation variables)

$L \Rightarrow L$: nonlinear map

Δt : sampling time of observations

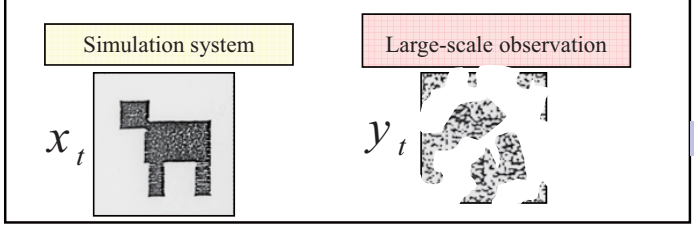
δt : simulation time step

$\Delta t = 1 >> \delta t$

$$x_t = F(x_{t-1}, v_t)$$ **Stochastic simulation model**

$$y_t = Hx_t + w_t$$ Observation model (Measurement model)

Bayesian Approach



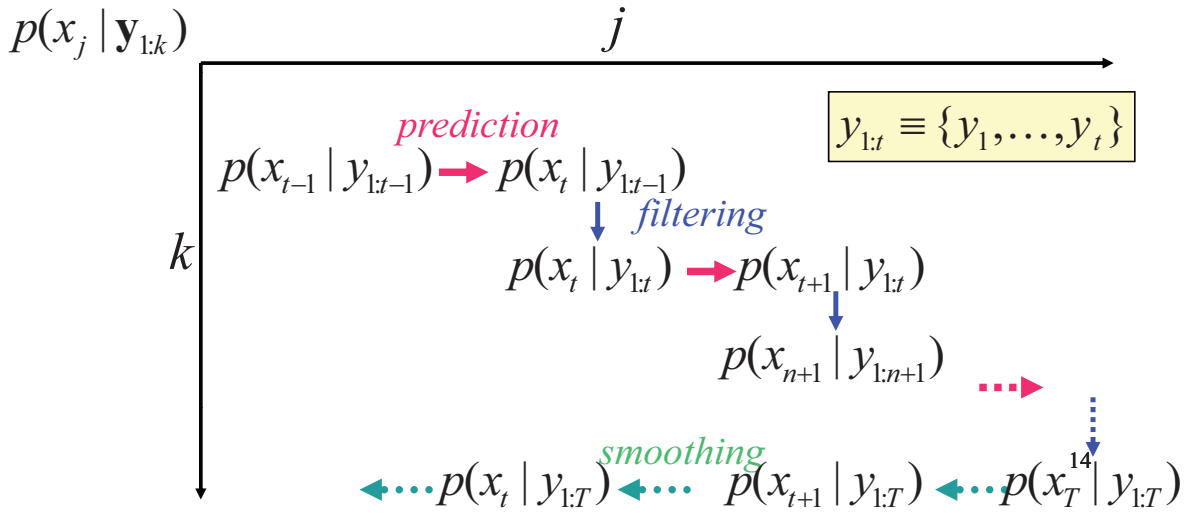
統計数理研究所

Conditional Distribution Recursive formula

predictive density: $p(x_t \mid \mathbf{y}_{1:t-1})$ Today's economic situation given yesterday's stock market data

filter density: $p(x_t \mid \mathbf{y}_{1:t})$ Today's economic situation estimated by the stock market data up to today

smoother density: $p(x_t \mid \mathbf{y}_{1:T})$ Today's economic situation analyzed by using all available data when we look back on the today in future



Prediction

$$p(x_t \mid y_{1:t-1})$$

$$= \int p(x_t, x_{t-1} \mid y_{1:t-1}) dx_{t-1}$$

$$= \int p(x_t \mid x_{t-1}, y_{1:t-1}) p(x_{t-1} \mid y_{1:t-1}) dx_{t-1}$$

$p(x_t \mid x_{t-1}, y_{1:t-1}) = p(x_t \mid x_{t-1})$ Markov property（1）

$$= \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid y_{1:t-1}) dx_{t-1}$$

Filter pdf at time *t-1*

15

filtering

$$p(x_t \mid y_{1:t})$$

Posterior, Belief

$$= p(x_t \mid y_t, y_{1:t-1})$$

$$= \frac{p(x_t, y_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})}$$

$$= \frac{p(y_t \mid x_t, y_{1:t-1}) \cdot p(x_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})}$$

Markov Property（2）

$p(y_t \mid x_t, y_{1:t-1}) = p(y_t \mid x_t)$

$$= \frac{p(y_t \mid x_t) \cdot p(x_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})}$$

$$= \frac{p(y_t \mid x_t) \cdot p(x_t \mid y_{1:t-1})}{\int p(y_t \mid x_t) \cdot p(x_t \mid y_{1:t-1}) dx_t}$$

16

Smoothing

$$p(x_t \mid y_{1:T})$$

$$= \int p(x_t, x_{t+1} \mid y_{1:T})dx_{t+1}$$

$$= \int p(x_t \mid x_{t+1}, y_{1:T}) \cdot p(x_{t+1} \mid y_{1:T})dx_{t+1}$$

$$= \int p(x_t \mid x_{t+1}, y_{1:t}) \cdot p(x_{t+1} \mid y_{1:T})dx_{t+1}$$

$$= \int \frac{p(x_t, x_{t+1} \mid y_{1:t})}{p(x_{t+1} \mid y_{1:t})} \cdot p(x_{t+1} \mid y_{1:T})dx_{t+1}$$

$$= \int \frac{p(x_t \mid y_{1:t}) \cdot p(x_{t+1} \mid x_t, y_{1:t})}{p(x_{t+1} \mid y_{1:t})} \cdot p(x_{t+1} \mid y_{1:T})dx_{t+1}$$

$$= p(x_t \mid y_{1:t}) \cdot \int \frac{p(x_{t+1} \mid x_t)}{p(x_{t+1} \mid y_{1:t})} \cdot p(x_{t+1} \mid y_{1:T})dx_{t+1}$$

Filter Dist. Smoothing Dist. Prediction Dist.

17

Sequential Data Assimilation

Estimate PDF of state vector x_t or its moments (mean, variance, …) sequentially on each observation

y_{t-1}

$p(x_{t-1} \mid y_{1:t-1})$

$\left(p(x_i \mid y_{1:k}) = p(x_i \mid y_1, y_2, \cdots, y_k) \right)$

Challenging problem: Huge dimension and inversion

- Data Assimilation = Estimation problem of state vector x_t:

(system model) $x_t = F_t(x_{t-1}, v_t | x_0)$

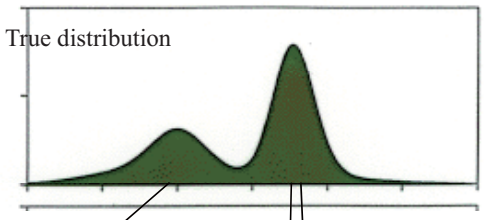
(observation model) $y_t = H_t x_t + w_t$ or $y_t = h_t(x_t) + w_t$

  - x_t : All variables in simulation model
  - y_t : All observed variables
  - v_t : Stochastic part to represent uncertainty of model (boundary condition, …)
  - w_t : Observation error
  - v_t, w_t : Normally Gaussian x_0: Initial condition

dimension x_t :10^4~10^6 y_t :10^2~10^5 $\dim(x_t) >> \dim(y_t)$

Numerical representation of distribution

$p(x_t \mid y_{1:t-1}), p(x_t \mid y_{1:t}), p(x_t \mid y_{1:T})$



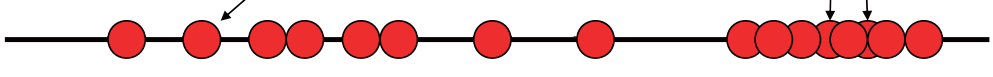
Monte Carlo approximation

Represent pdf by the actual realizations.

N: # of particles

$$p(x_t \mid y_{1:t-1}) \cong X_{t|t-1} \equiv \left[x_{t|t-1}^{(1)}, x_{t|t-1}^{(2)}, \ldots, x_{t|t-1}^{(N)}\right]$$

$$p(x_t \mid y_{1:t}) \cong X_{t|t} \equiv \left[x_{t|t}^{(1)}, x_{t|t}^{(2)}, \ldots, x_{t|t}^{(N)}\right]$$

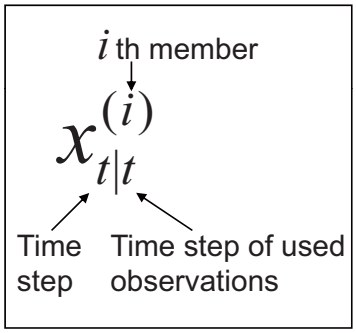


Sequential DA Methodology

- Ensemble Kalman Filter (EnKF) is widely used.
  - Conditional PDF is approximated by a set (ensemble) of realizations.
  - Kalman Filter is used for filtering.
- Application of Particle Filter (PF) is rare.
  - This is also ensemble based.

$$p(x_t \mid y_{1:t-1}) \cong \frac{1}{N}\sum_{i=1}^{N}\delta(x_t - x_{t|t-1}^{(i)}) \qquad \left\{x_{t|t-1}^{(i)}\right\}_{i=1}^{N}$$

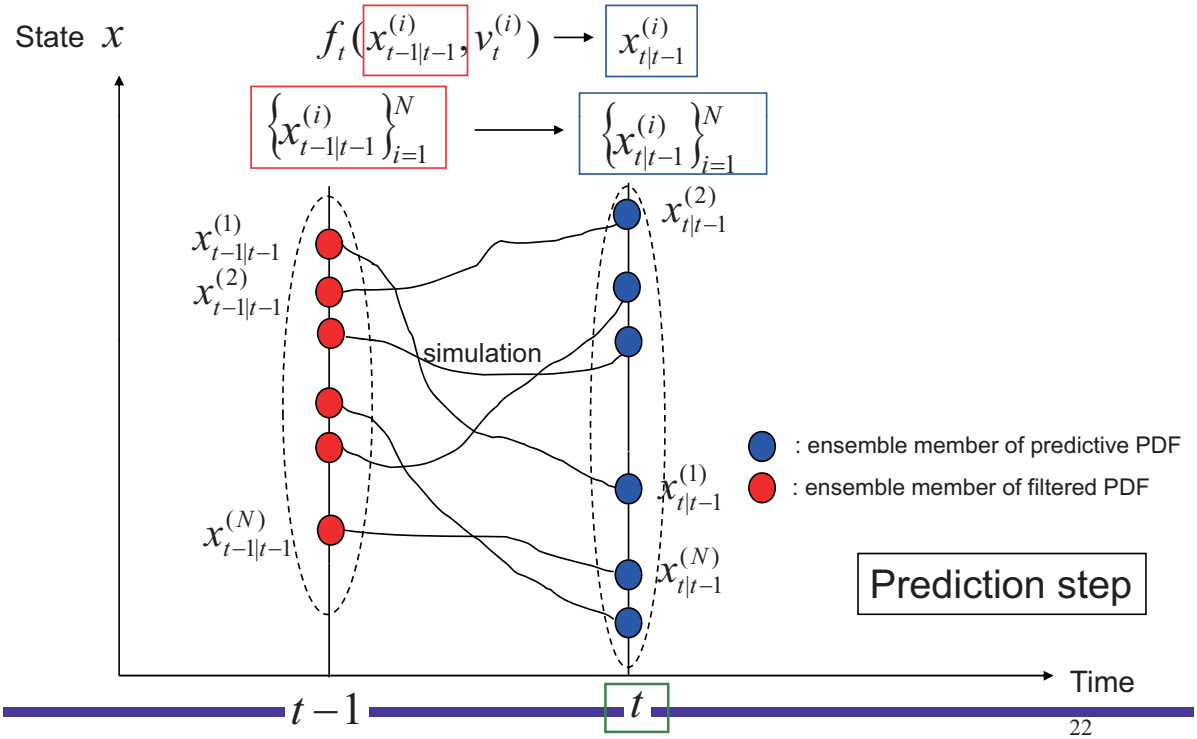
$$p(x_t \mid y_{1:t}) \cong \frac{1}{N}\sum_{i=1}^{N}\delta(x_t - x_{t|t}^{(i)}) \qquad \left\{x_{t|t}^{(i)}\right\}_{i=1}^{N}$$



21

ISM

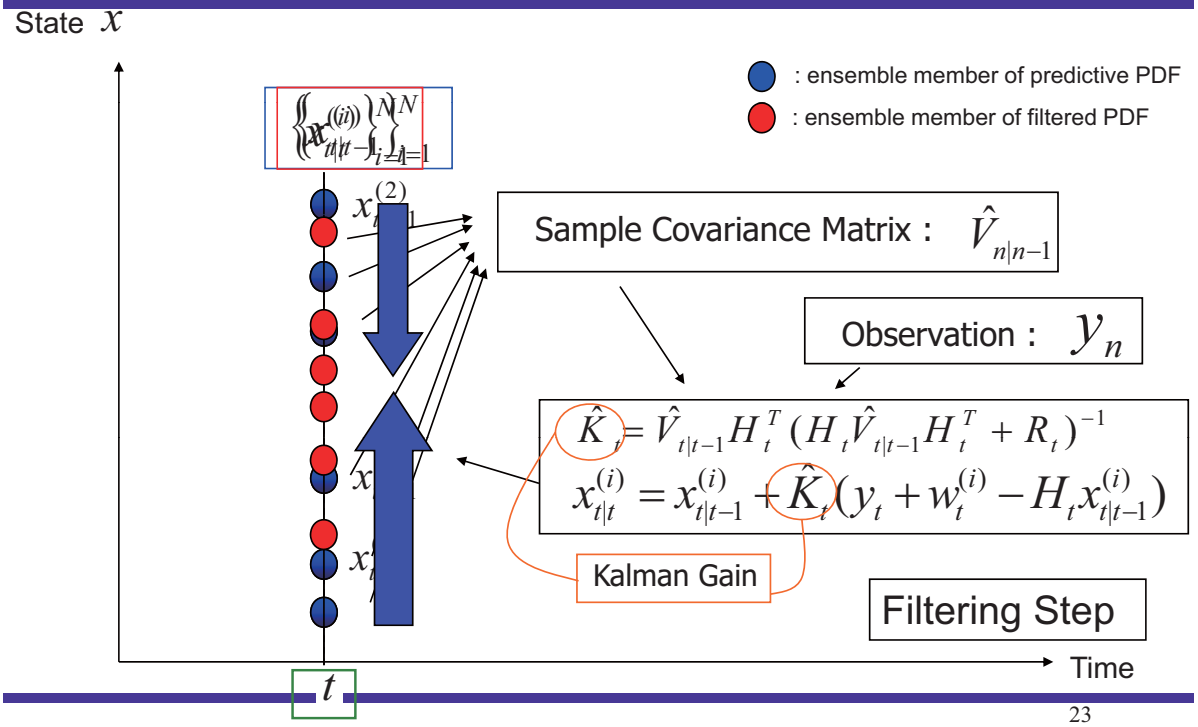
Prediction Step (Common in EnKF and PF)



22

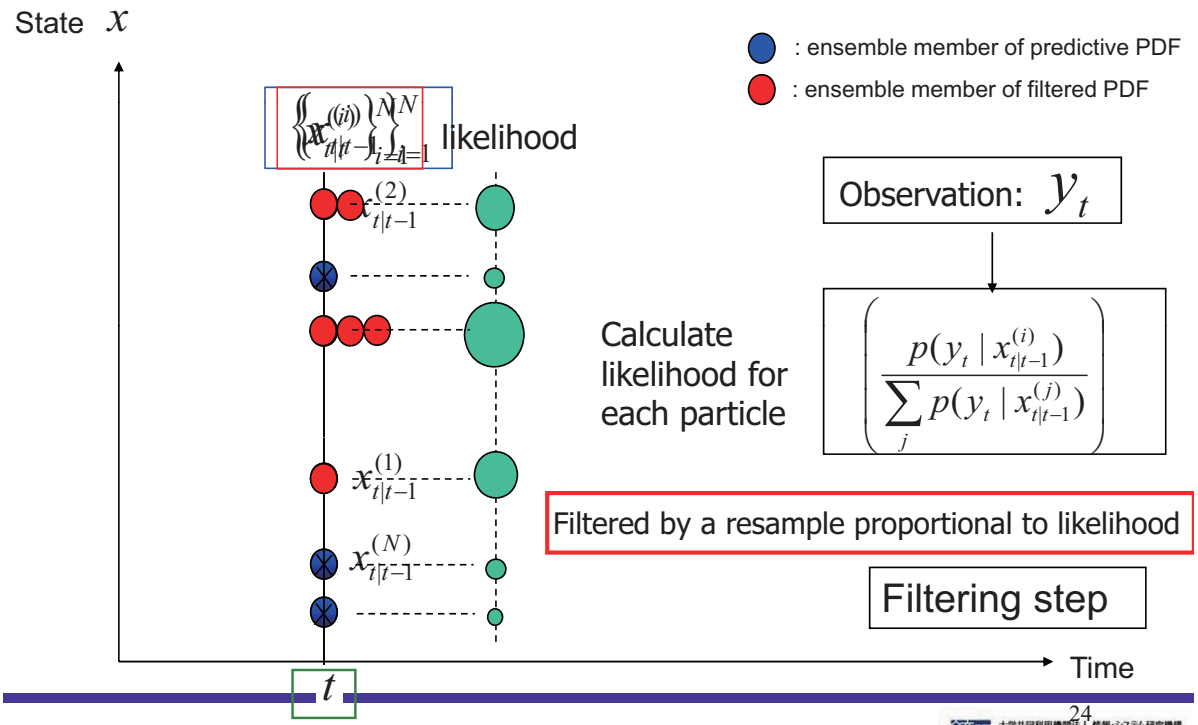
ISM

Filtering step of EnKF



ISM

Filtering Step of PF



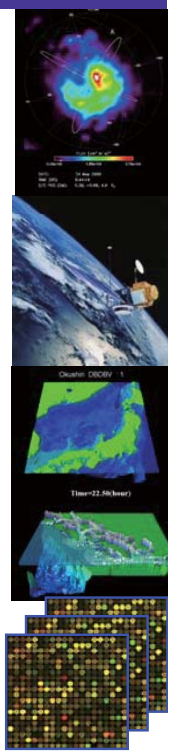
統計数理研究所

Projects in progress

- Coupled Ocean-Atmosphere model
  Genta **Ueno** (ISM/JST CREST)
  T. Kagimoto (JAMSTEC, FRCGC), N. Hirose (Kyushu Univ., RIAM)
- Tsunami model
  Kazuyuki **Nakamura** (JST CREST)

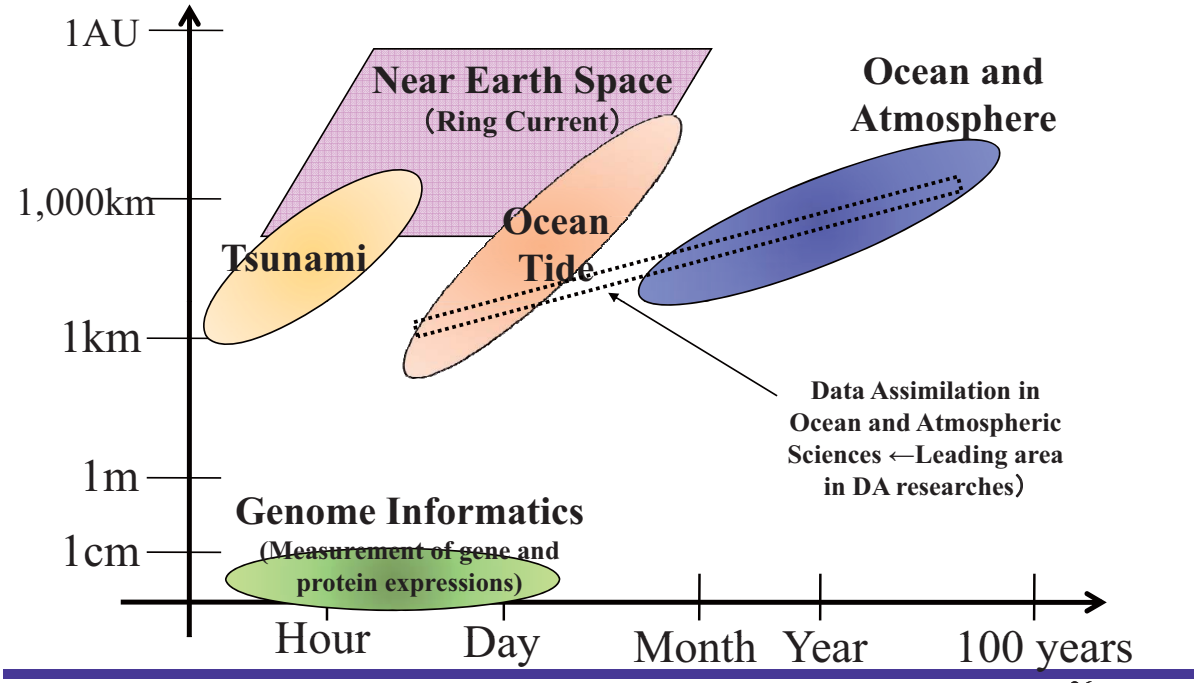
N. Hirose (Kyushu Univ., RIAM)

- Ocean tide
  Daisuke **Inazu** (JST CREST)
  T. Sato, S. Miura (Tohoku Univ.), and others (Alaska Univ.)
- 3D structure of ring current
  Shin'ya **Nakano** (JST CREST),
  Y. Ebihara (Nagoya Univ.) , M.-C Fok (NASA)
  S.-I. Ohtani, P.C.Brandt（Johns Hopkins Univ.)
- Genome informatics
  Ryo **Yoshida** (ISM/JST CREST)
  Miyano lab. (Tokyo Univ./IMS)



大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Time and Spatial Scale

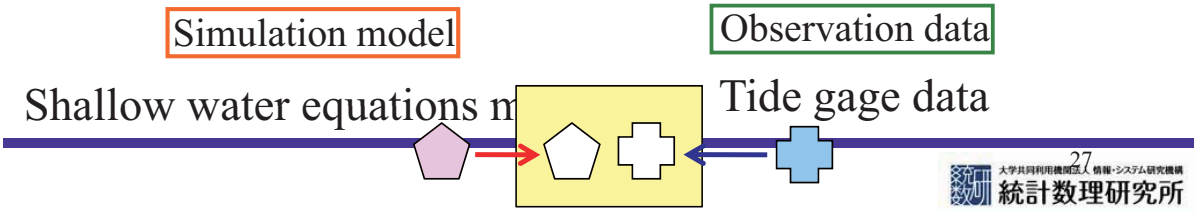


大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Revisit : What is Data Assimilation?

- Emerging subject in meteorology and oceanography.
- Methodology to synthesize numerical simulation model and observed data
  - Simulation model can not reflect real physics accurately.
    - (e.g.) Accurate weather forecast needs good initial conditions.
    - Uncertainty in the model (boundary condition, initial condition, unknown parameters, unknown dynamics...) exists.
  - Observation data have some physical/budgetary restrictions.

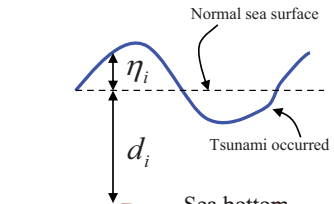
Correct variables in numerical simulation model using observation data. = Data Assimilation

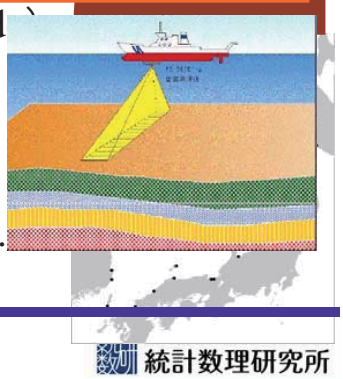


統計数理研究所

Tsunami Simulation Model

- Based on PDE→Shallow water equations [Choi *et al.* 01]
- Discretized temporally and spatially.
  - 4 physical variables at each grid i .
    - Flow vector (longitudinal/latitudinal) : (U_i, V_i)
    - Displacement of sea surface height : η_i
    - Water depth at each grid : d_i ← Uncertainty in measured water depth!
  - # of grid points: 192 (longitude) × 240 (latitud
    - Half of them are on the sea.
    - Dimension of state vector is about 9×10^4.
- Propagation speed depends on water depth.
  - Deep water makes tsunami propagation faster.



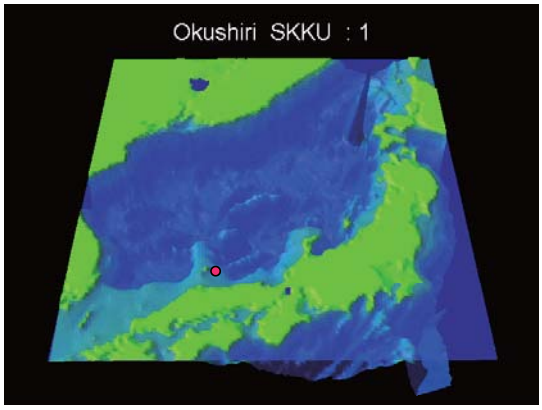


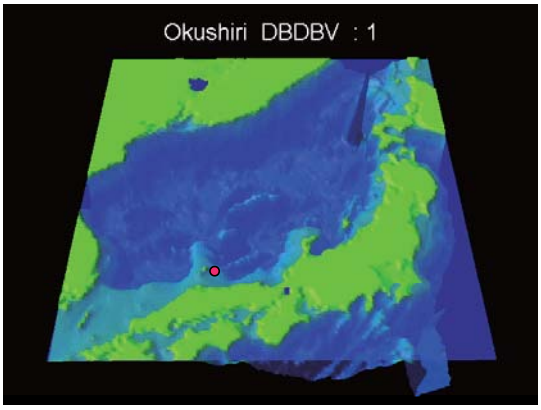
統計数理研究所

Numerical Simulation (Not DA)

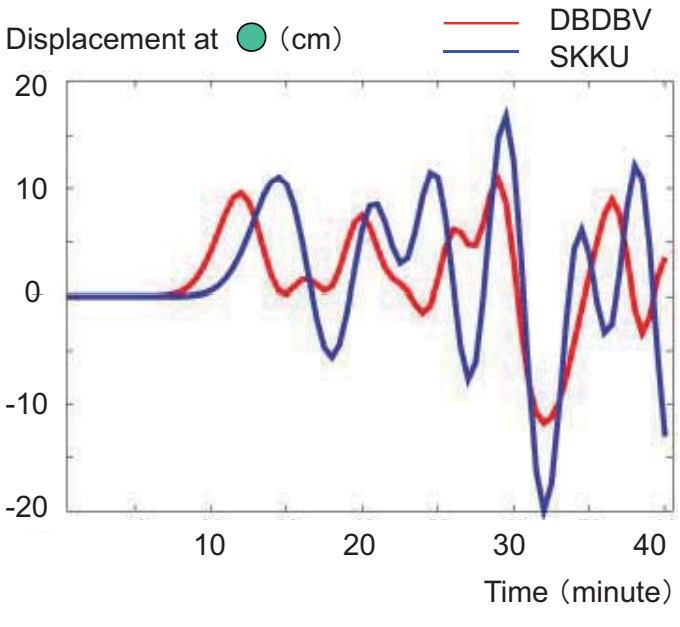
- Simulation of Okushiri Tsunami
  - Simulation based on topographies made by different organizations.
  - It looks similar, but time series of sea surface displacement at a point (●) is ...

(From Japan Coast Guard)



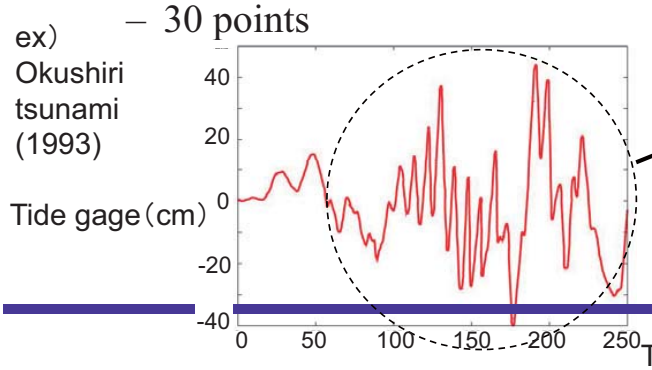


Comparison of Sea Surface Displacement



Observation Data

- Tide gage data:
  - (Linear/Nonlinear) Response to sea surface displacement at instrument installation site.
  - One dimensional time series.
- Number of tide gage stations:
  - 30 points



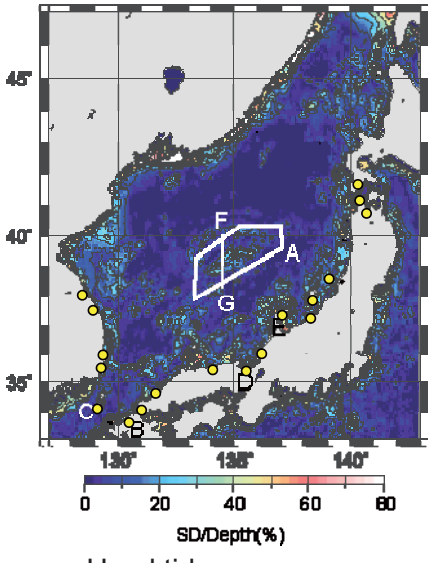
統計数理研究所

Application to Real Data

- Analysis by real tsunami occurred in the Japan Sea in 1993.

•The depths in and around Yamato Rises (area A) varies among 4 bottom topography data set.

•Uncertainty is introduced into South Rises and around area as linear combination of 4 data sets.

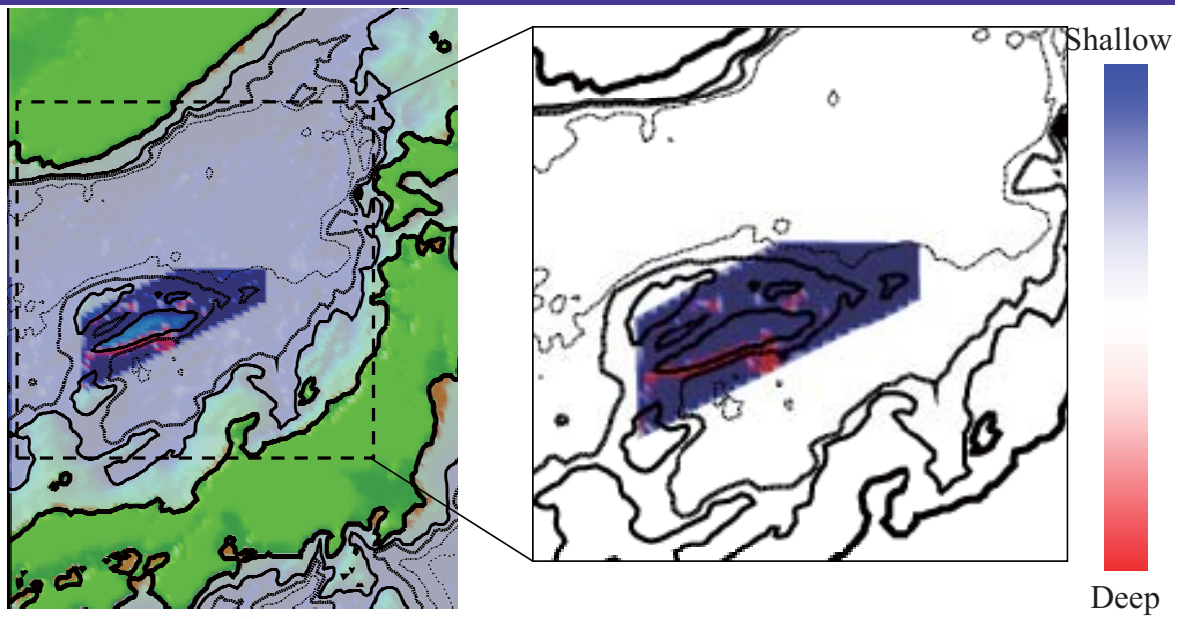


j : number of data sets

$$d_m^{(i)} = \sum_j w_j^{(i)} d_{m,j} \quad , \quad w_j^{(i)} \sim N(0.25, \sigma^2)$$

統計数理研究所

DA result



•South Rise might be shallower than public data.
•Deeper area exists in south slope.

Personalized Simulation :

A boundary condition is assimilated to local information.

We introduce a local/personal information into a numerical simulation model and personalize the simulation for each location/person.

Motion Eq. : $\mathbf{v}$

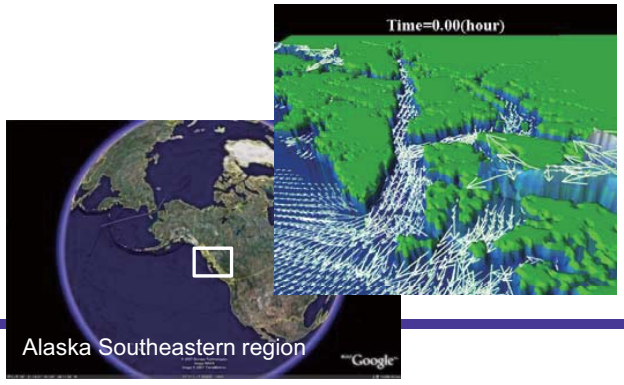
$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} + \mathbf{f}\times\mathbf{v} = -g\nabla\eta - \gamma_b \frac{\mathbf{v}|\mathbf{v}|}{H} + A_H\nabla^2\mathbf{v}$$

Sea Bottom friction coefficient

Sea depth

Continuous Eq. :

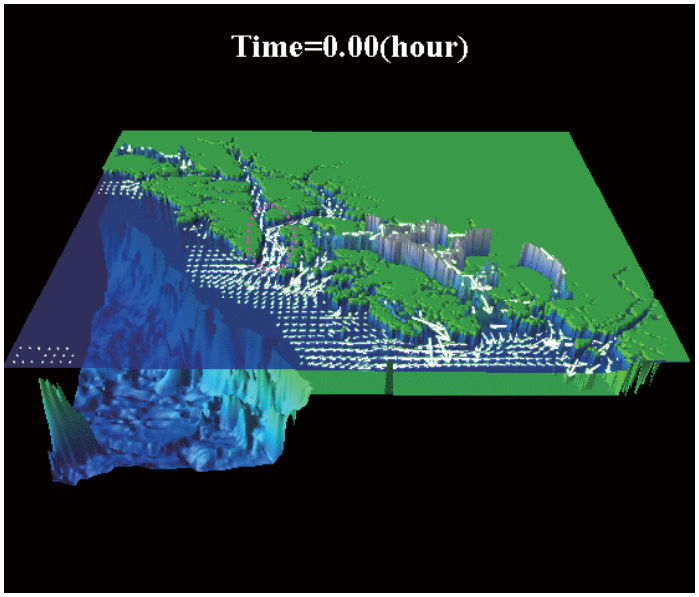
$$\frac{\partial \eta}{\partial t} + \nabla\cdot(\mathbf{v}H) = 0$$



$\mathbf{v}$: 2-dimensional flow vector
η : Water surface height
H : Water depth, $\mathbf{f}$: Coriolis parameter

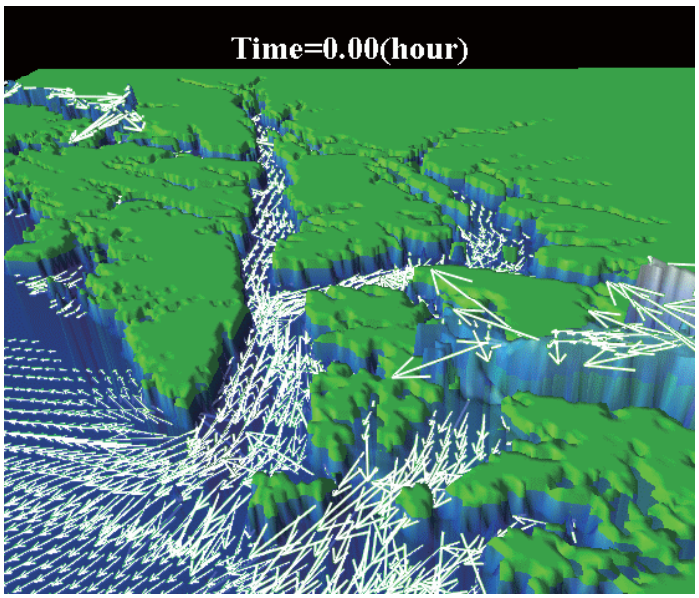
Ocean tide simulation by our CREST project

Water level and Flow vectors



Sea level range is About ±5 m. Current speed is (much) more than 1 m/s at the mouth of Chatham Strait.

Water level and flow vectors（Closeup）

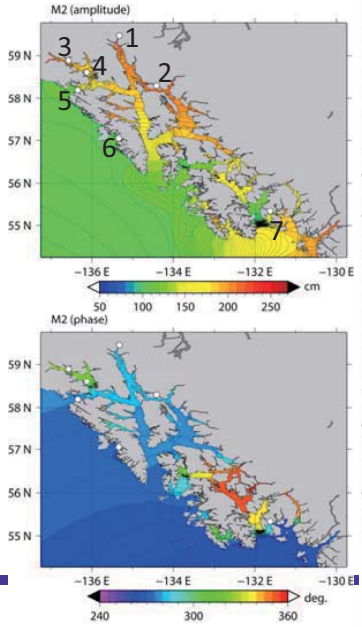


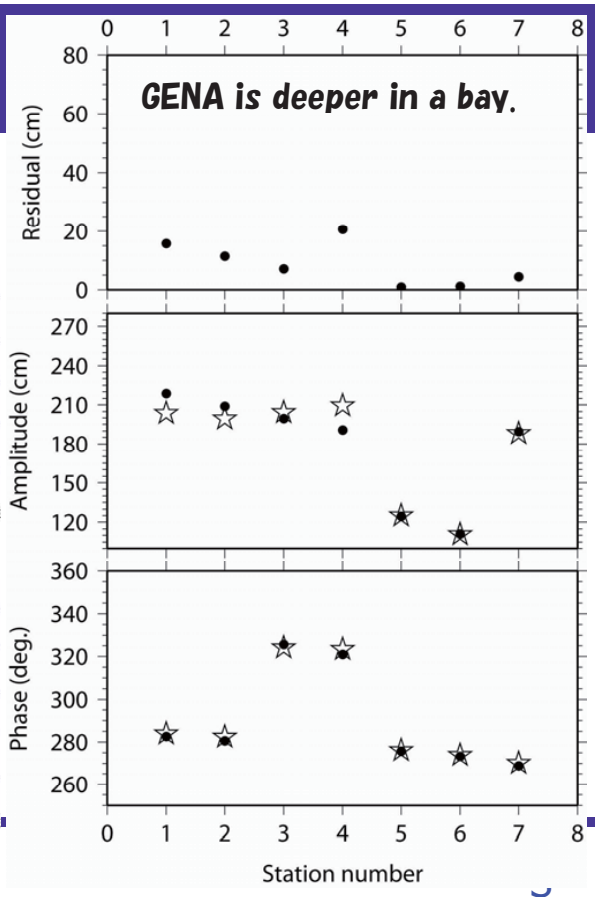
M_2component tide
Result with GINA

GENA shows a great performance in representing an ocean tide.

Amplitude

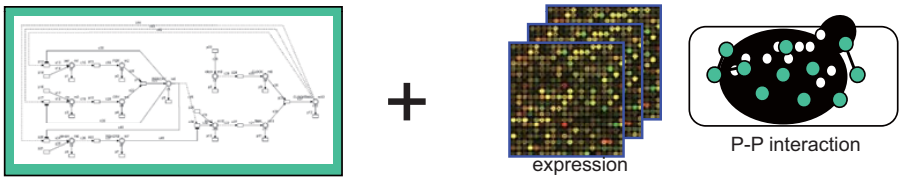
Phase





Genomic Data Assimilation

Statistical framework to link simulation model and data



Simulation model　　Biological data

Formulated by the generalized State Space Model

$$x_t = f(x_{t-1}, v_t, \theta)$$ System model

$$y_t = Hx_t + w_t$$ Observation model

x_t : state vector at time t,　f : simulation devise,　$t = 1, \ldots, T$

v_t : system noise,　θ : parameter vector,

y_t : observation vector at time t,　H : observation matrix,

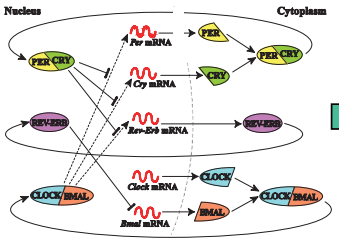
$w_t \sim N(0, \sigma^2)$: observation noise

統計数理研究所

Circadian Rhythm Model with HFPN

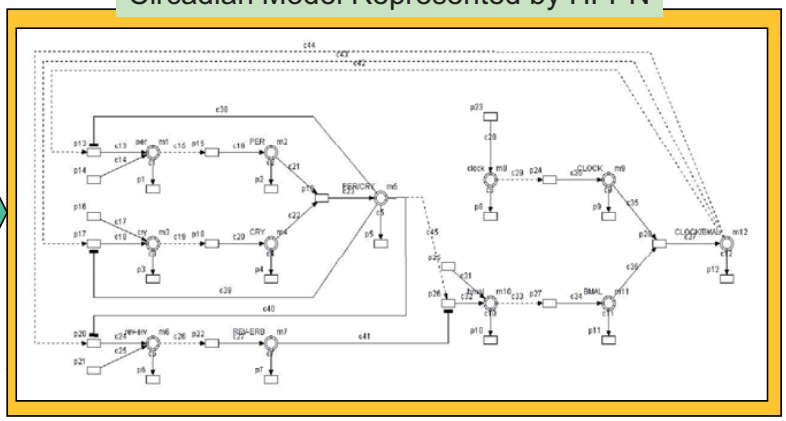
HFPN (Hybrid Functional Petri Net) : A graphical programming language suitable to model biological pathways and can be used for simulations

Circadian Rhythm Model of Mouse



Fujii et al. 2005

Circadian Model Represented by HFPN



45 parameters (12 states),

4 observations

Parameters

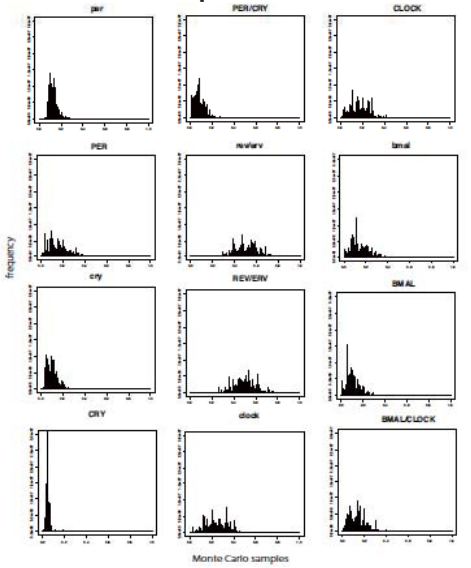
$m_1(0),\ldots,m_{12}(0)$	:Initial values
k_1,k_2,k_3,k_4	:Speeds
s_1,s_2,s_3,s_4	:Thresholds
τ^2,σ^2	:Noise variances

時不変パラメータ(速度定数(14)と敷居値(7)及び初期値(12))数33, 状態が12

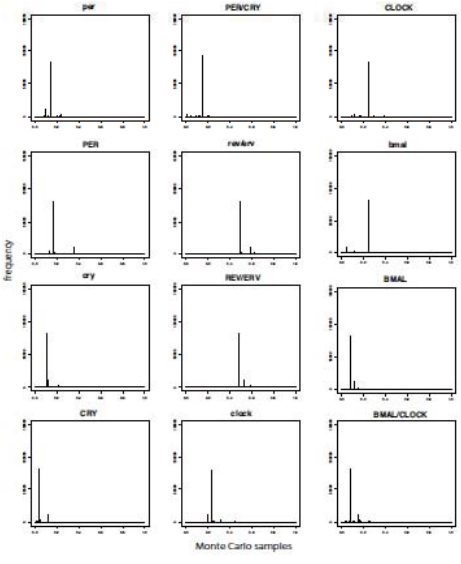
大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Toward Peta-scale Computing

100,000,000 particles（1億）



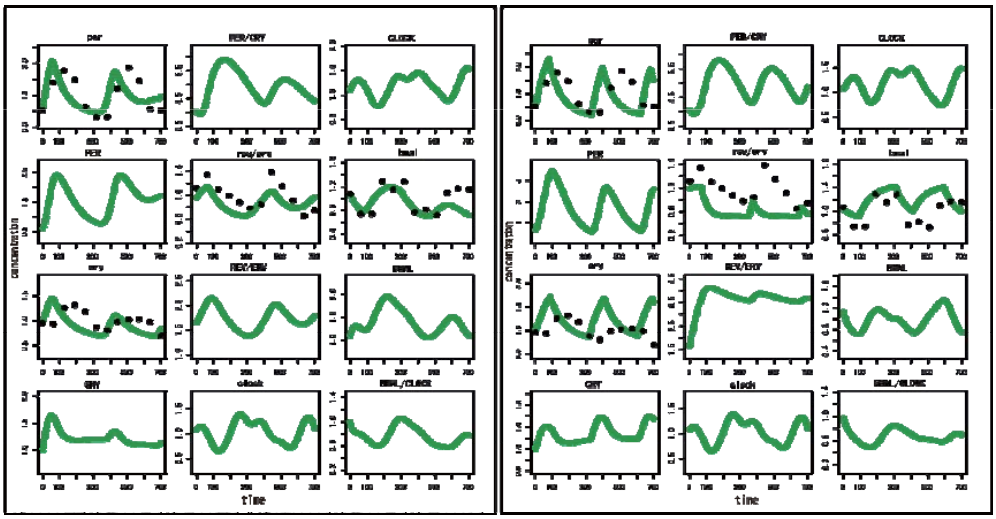
100,000 particles:10万



大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Prediction

100,000,000 particles（1億）　100,000 particles:10万



大学共同利用機関法人 情報・システム研究機構 統計数理研究所

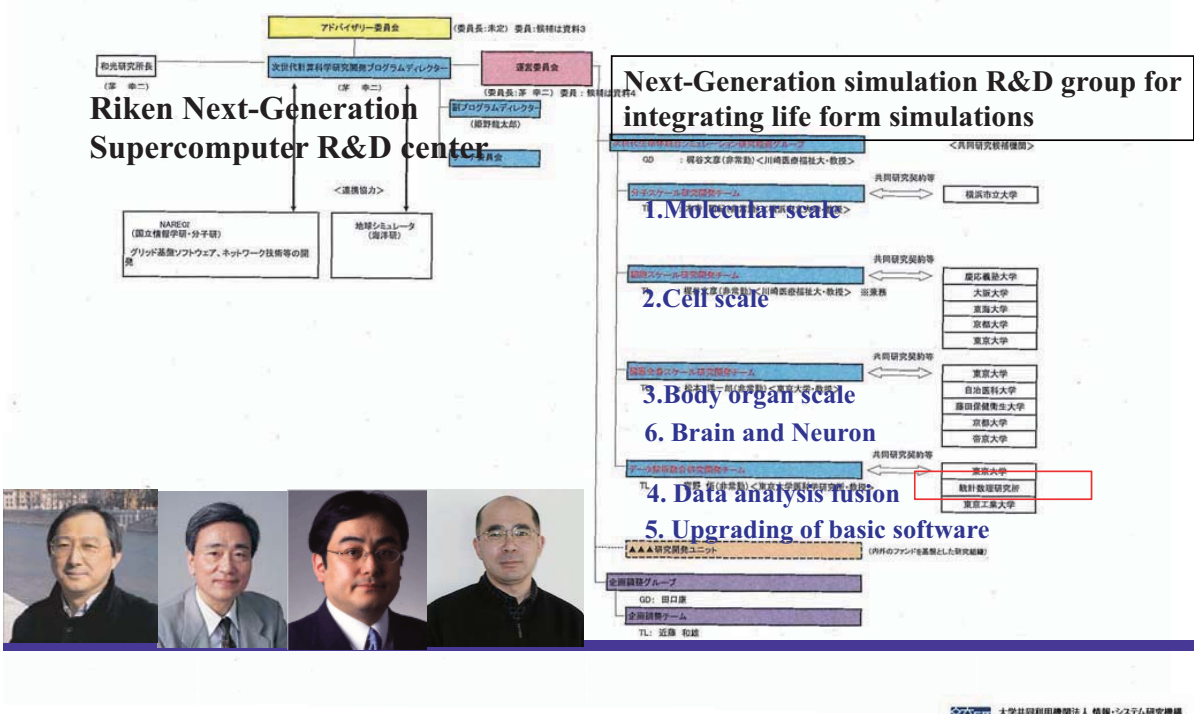
Next-Generation of Supercomputer in Japan at Kobe

Japanese Government will spend more than 1 billion US$ for this national project.

■Grand Challenge:
-- Nanotech (Institute for Molecular Science)
-- Life Science (RIKEN)

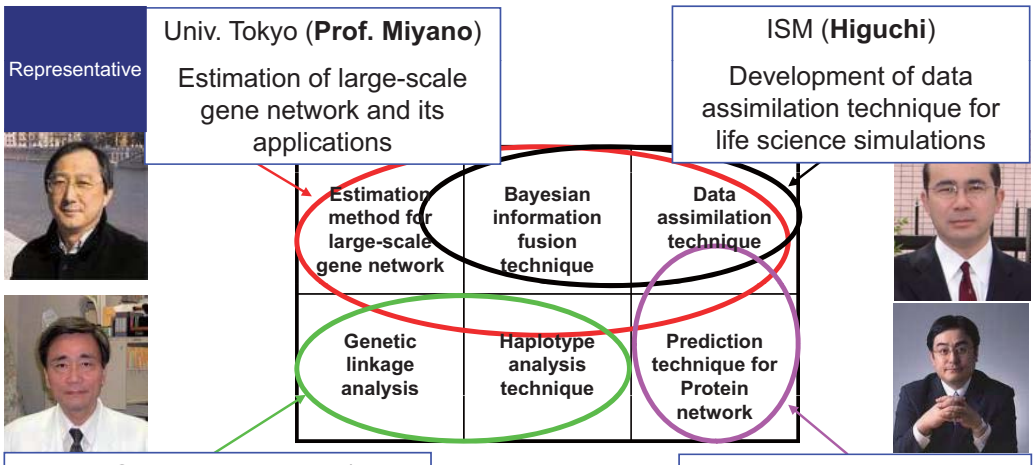
大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Development for next-generation simulation software for whole human body



統計数理研究所

Data analysis fusion Team

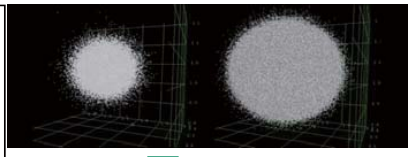


統計数理研究所

Attempt to realize personalization technique

Making a parallel computation scale larger enables us to carry out a data-dependent simulation, and results in drawing a scenario and in making a risk assessment.

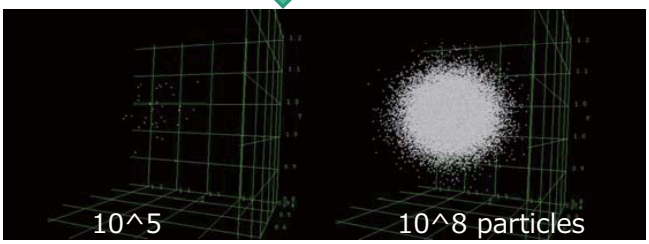
Prior and posterior distributions for three parameters among parameters estimated PF are demonstrated in 3-dimensional space.



Prior distribution of parameters
（left：10^5 right: 10^8）

Estimation by PF

Although the PF with 10^5 particles results in the PDF with a small number of particles, the PF with 10^8 particles leaves many particles.



Posterior distribution of parameters

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

Perspective of our Project

"Creation of meta-simulation model"

1. We automate a procedure searching for better simulation model to describe real phenomena.

2. We develop a procedure to generate a new simulation model that has greater ability of predictive performance than existing ones.

3. We give consistent view to assessment of simulation model that is said to be subsidiary problem in simulation science; Maximum Likelihood Principle.

4. We give a platform to design a measurement system in an attempt to enhance a scientific return together with reducing a total budgetary cost.

大学共同利用機関法人 情報・システム研究機構 統計数理研究所