Combining Information Sources to Develop Bayesian Predictions

Mark Berliner
Ohio State University
mb@stat.osu.edu

Chicago ASA, March 6, 2015

Outline

- Bayesian Prediction
- Hierarchical Modeling
- Selected Issues
- Examples

Bayesian Prediction

- **Predict a variable *X* based on data *y***
- **Answer: find the**

<u>predictive distribution</u> $p(x|y)$

- **Seems all we need is**

<u>data model</u> $p(y|x)$

and **<u>prior</u>** $p(x)$

- **Then** $p(x|y) = \mathrm{p}(\mathrm{y}|\mathrm{x})\ \mathrm{p}(\mathrm{x}) / \mathrm{p}(\mathrm{y})$
- **Any Questions?**

Bayesian Prediction

Find $p(x|y) = p(y|x)\, p(x) \,/\, p(y)$

- **In practice, obtaining these inputs is difficult and can be perilous**
- **Finding $p(y)$ can be infeasible**

Pause: Why endure the difficulties and seemingly complicated methods I'll show?

1) **Enable input of information from various sources and of various types**
   a) **Mechanistic models are prior information. They contribute scientific basis for prediction**
   b) **Enriched model classes; e.g. space-time parameters**
   c) **Treat multiple scales & multiple variables**
2) **Uncertainty quantification:**
   a) **Predictive *distribution***
   b) **Risk analysis**
   c) **Decision making**

Bayesian Hierarchical Models (BHM)

Y are data; X is the predictand;

θ unknown parameters

BHM Skeleton:

1. <u>**Data Model**</u>: $p(y \mid x, \theta)$
2. <u>**Process Model Prior**</u>: $p(x \mid \theta)$
3. <u>**Prior on Parameters**</u>: $p(\theta)$

Bayes' Theorem:

<u>**posterior distribution**</u>: $p(x, \theta \mid y)$

<u>**posterior predictive**</u>: $p(x \mid y) = \int p(x, \theta \mid y)\, d\theta$

Selected Issues

1) **Incorporating diverse datasets:**

$$p(\, y_1, y_2 \mid x\,) = p(\, y_1 \mid x\,)\, p(y_2 \mid x\,) \quad \text{(if OK)}$$

2) **Several related process variables**

$$p(x_1 \mid y) = \int p(x_1 ,\, \theta \mid y)\, d\theta$$

versus

$$p(\, x_1 \mid y) = \int p(x_1 ,\, x_2 ,\theta \mid y)\, dx_2\, d\theta$$

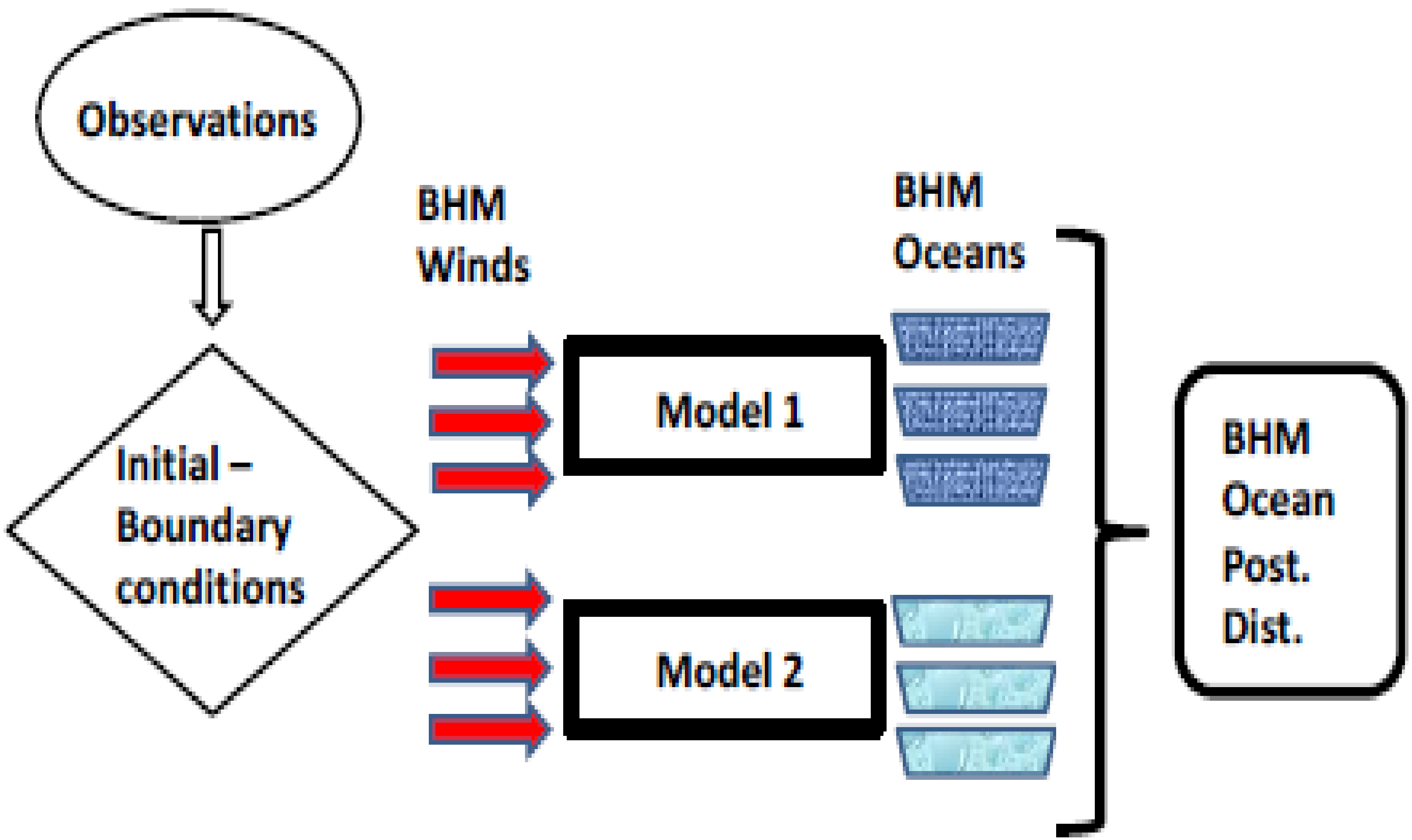
Which depends on how well can we

- **model and learn about x_2**
- **model the relationship between x_1 and x_2**

3) **Mechanistic models are useful but**
    a) **Subject to error; unknown parameters**
    b) **Computationally hard (nonlinear PDE)**
    c) **Massive supercomputer models eg: climate system models**
        1. **spatially gridded at regional, not local, levels**
        2. **various ad hoc approximations needed**
        3. **different models give different results: “multi-model ensembling”**
        4. **too large to obtain large samples (ensembles)**
4) **Different methods for models we can embed in the BHM versus using output from supercomputer models**

Example 1: Mediterranean Ocean Forecasting

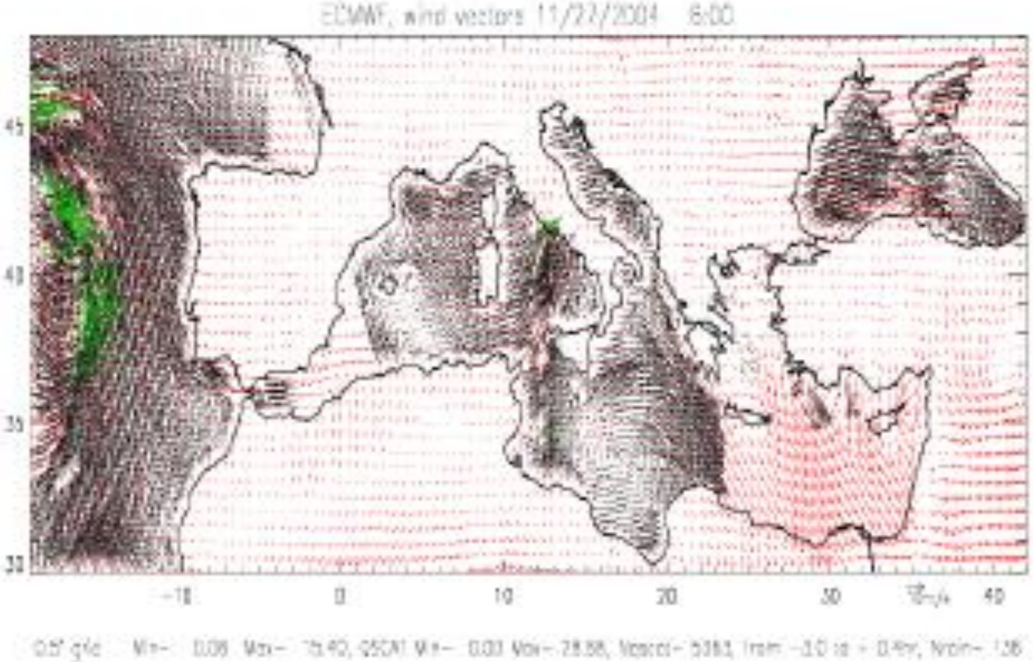
1. **BHM surface wind model to drive ocean model**
2. **BHM to do multi-model ensembling**



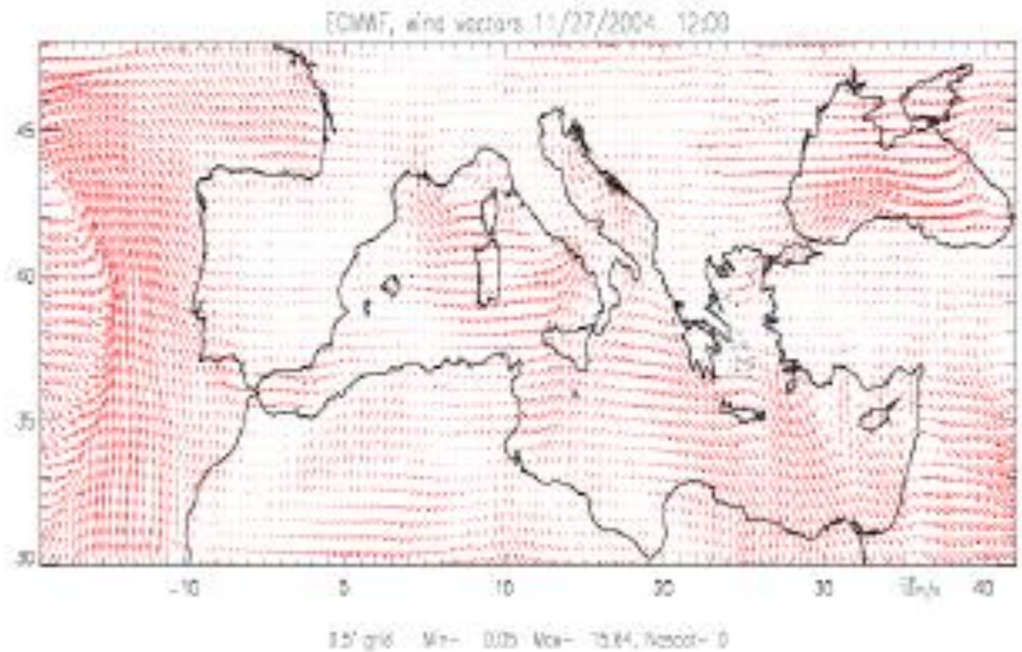
Building wind dist. (BHM-SVW)

1. Data Stage

Satellite (QSCAT) and Numerical Weather Pred. Analyses (ECMWF)



QSCAT



ECMWF

2. *Process Model:* Rayleigh Friction Model (Linear Planetary Boundary Layer Equations)

Theory

$$\begin{aligned}\frac{\partial u}{\partial t} - fv &= -\frac{1}{\rho}\frac{\partial p}{\partial x} - \gamma u \\ \frac{\partial v}{\partial t} + fv &= -\frac{1}{\rho}\frac{\partial p}{\partial y} - \gamma v\end{aligned}$$

(neglect second order time derivative) discretize:

$$\begin{aligned}V_t &= \left[1 + \frac{2\gamma}{f^2\triangle} + \frac{\gamma^2}{f^2}\right]^{-1} \\ &\left[(\frac{2\gamma}{f^2\triangle})V_{t-1} + (\frac{1}{f})D_xP_t + (-\frac{1}{f^2\triangle} - \frac{\gamma}{f^2})D_yP_t + (\frac{1}{f^2\triangle})D_yP_{t-1}\right] \\ U_t &= \left[1 + \frac{2\gamma}{f^2\triangle} + \frac{\gamma^2}{f^2}\right]^{-1} \\ &\left[(\frac{2\gamma}{f^2\triangle})U_{t-1} + (-\frac{1}{f})D_yP_t + (-\frac{1}{f^2\triangle} - \frac{\gamma}{f^2})D_xP_t + (\frac{1}{f^2\triangle})D_xP_{t-1}\right]\end{aligned}$$

Our model

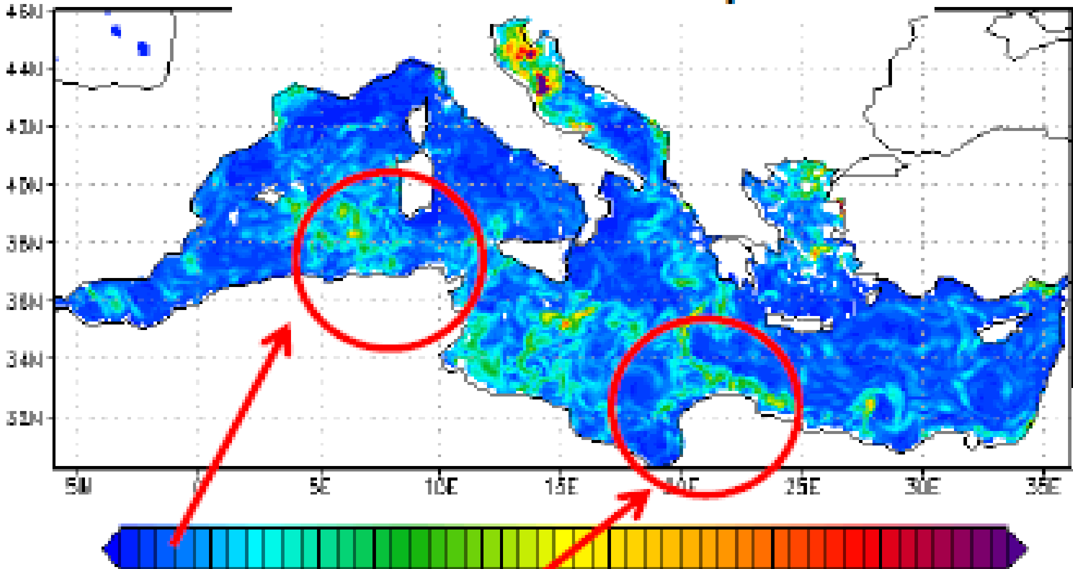
$$V_t = -L_{v|v(1)}V_{t-1} + c_{v|p_x}D_xP_t + c_{v|p_y}D_yP_t + c_{v|p_y(1)}D_yP_{t-1} + \epsilon$$

$$U_t = -L_{u|u(1)}U_{t-1} - c_{u|p_y}D_yP_t + c_{u|p_x}D_xP_t + c_{u|p_x(1)}D_xP_{t-1} + \epsilon$$

BHM-SVW-OEF initial condition spread:

Sea Surface Temperature

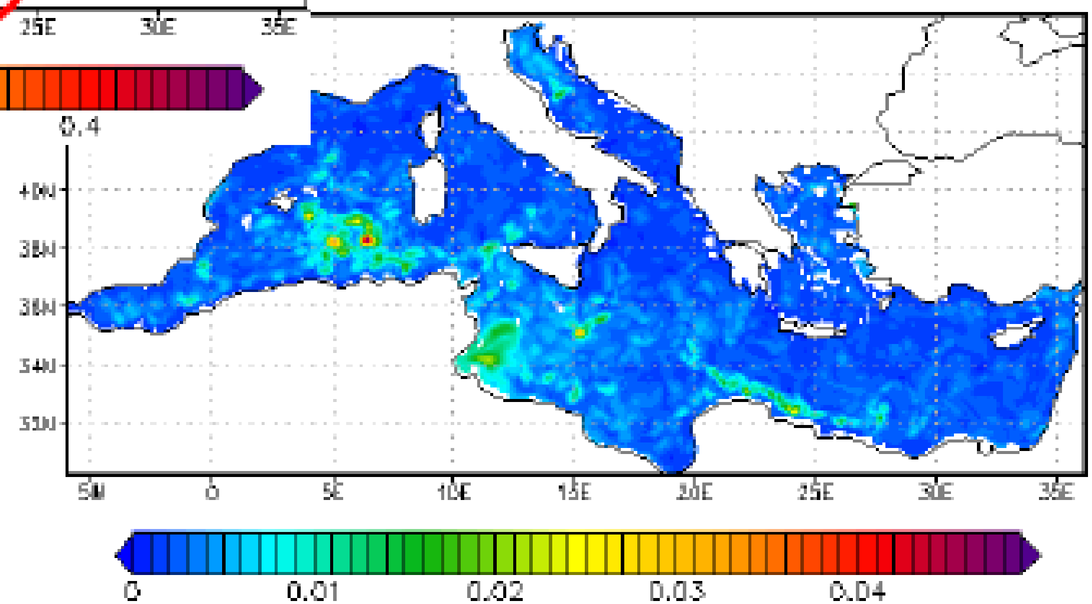
Initial condition spread

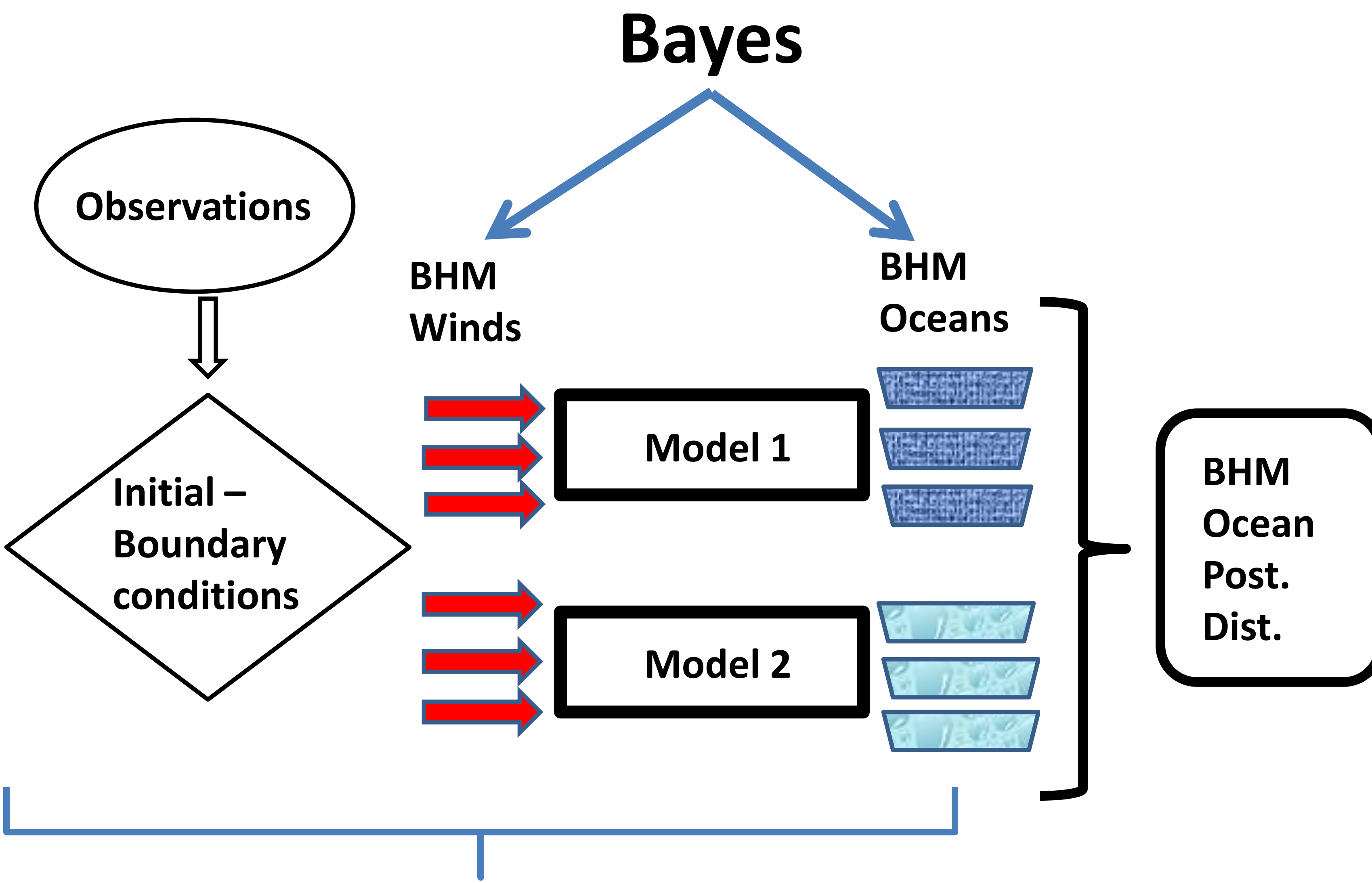


Uncertainty is concentrated at mesoscales

Sea Surface Height

Initial condition spread





Milliff et al (2011); Pinardi et al (2014);
Dobricic et al (2014): *Q. J. R. Meteor. Soc.*

Multi-model Ocean Modeling

Berliner et al. (2015)

- **Process: profiles of temperature $X(z,t)$**
- **16 vertical levels from 0m to 300m**
- ***t*=1,…,60 days**
- **Bayes wind-model gives ensembles of boundary conditions for ocean models:**

1) **Ocean Parallelized (OPA)** $\tilde{X}_{\mathrm{OPA}}$
2) **Nucleus for European Modeling of the Ocean (NEMO)** $\tilde{X}_{\mathrm{NEMO}}$

Multi-model Ensembles

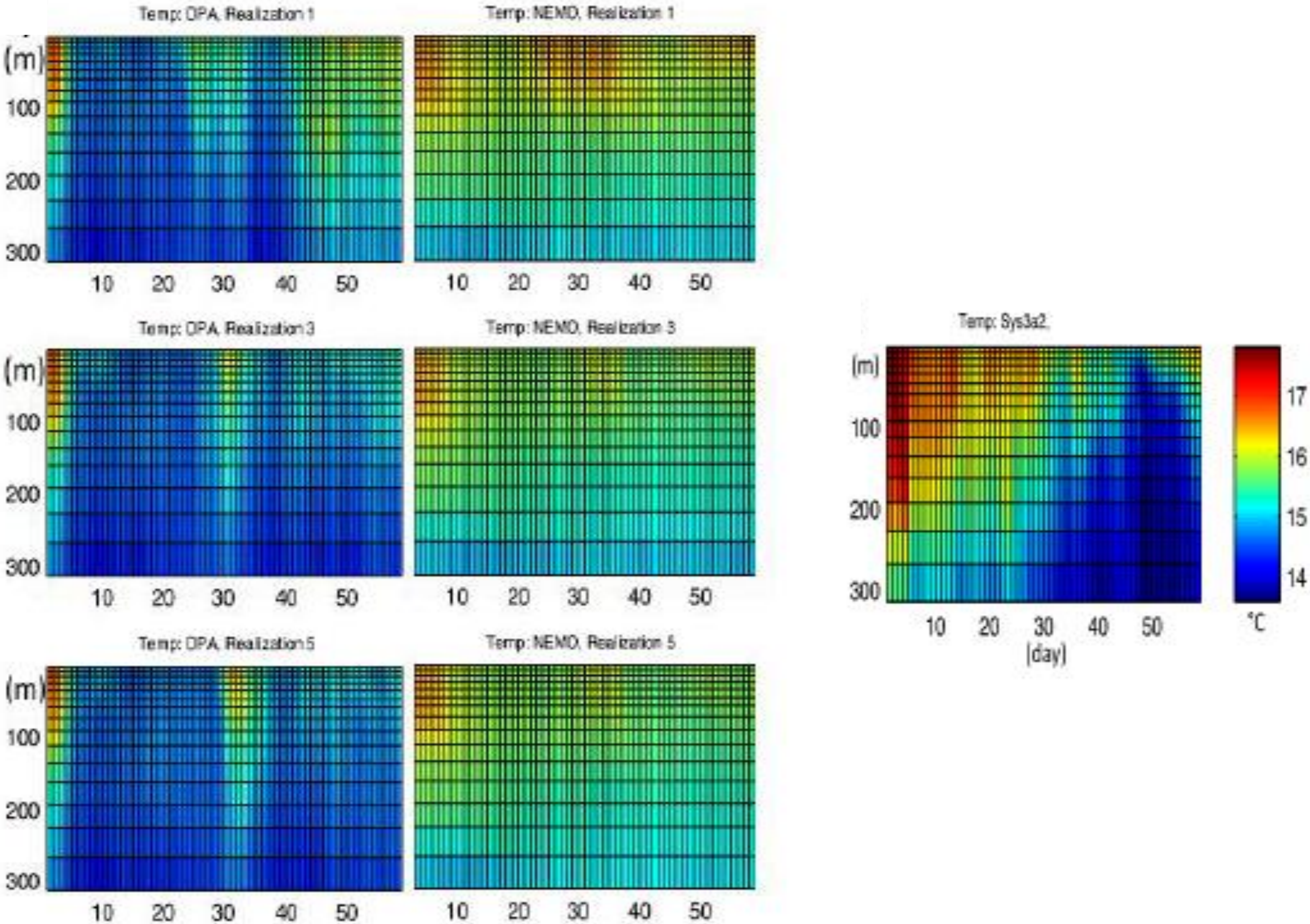
Act as if model output are biased observations of the process

- **Model output data model:**

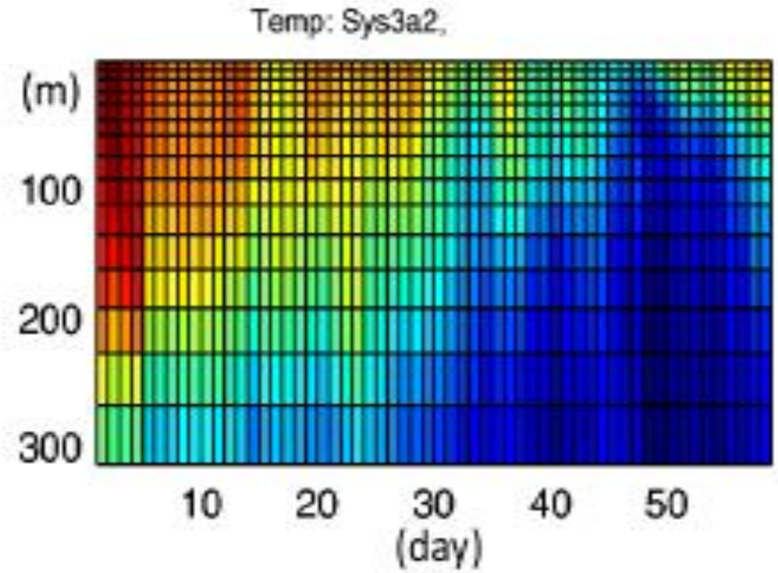
$$\widetilde{X}_{\mathrm{OPA\,j}} = X + b_{\mathrm{OPA}} + \xi_{\mathrm{j(O)}} \qquad \mathrm{j=1,\ldots,10}$$

$$\widetilde{X}_{\mathrm{NEMO\,j}} = X + b_{\mathrm{NEMO}} + \xi_{\mathrm{j(N)}} \qquad \mathrm{j=1,\ldots,10}$$

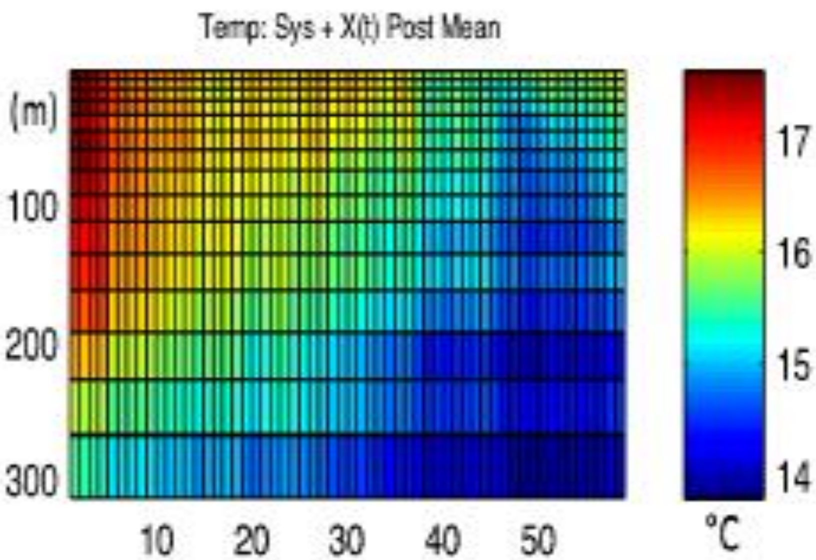
- **Prior for *X*: some technical thinking**
- **Prior for *b*'s: prior mean 0 and slowly varying in time.**
- **Model covariance of ξ's; …**



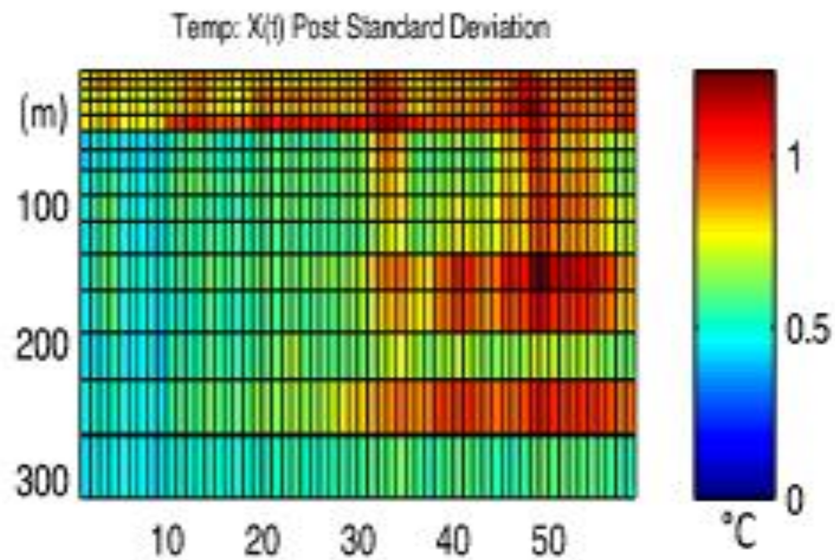
Temp: OPA, Realization 1
Temp: NEMO, Realization 1
Temp: OPA, Realization 3
Temp: NEMO, Realization 3
Temp: OPA, Realization 5
Temp: NEMO, Realization 5
Temp: Sys3a2
(m)
100
200
300
10
20
30
40
50
(day)
17
16
15
14
°C



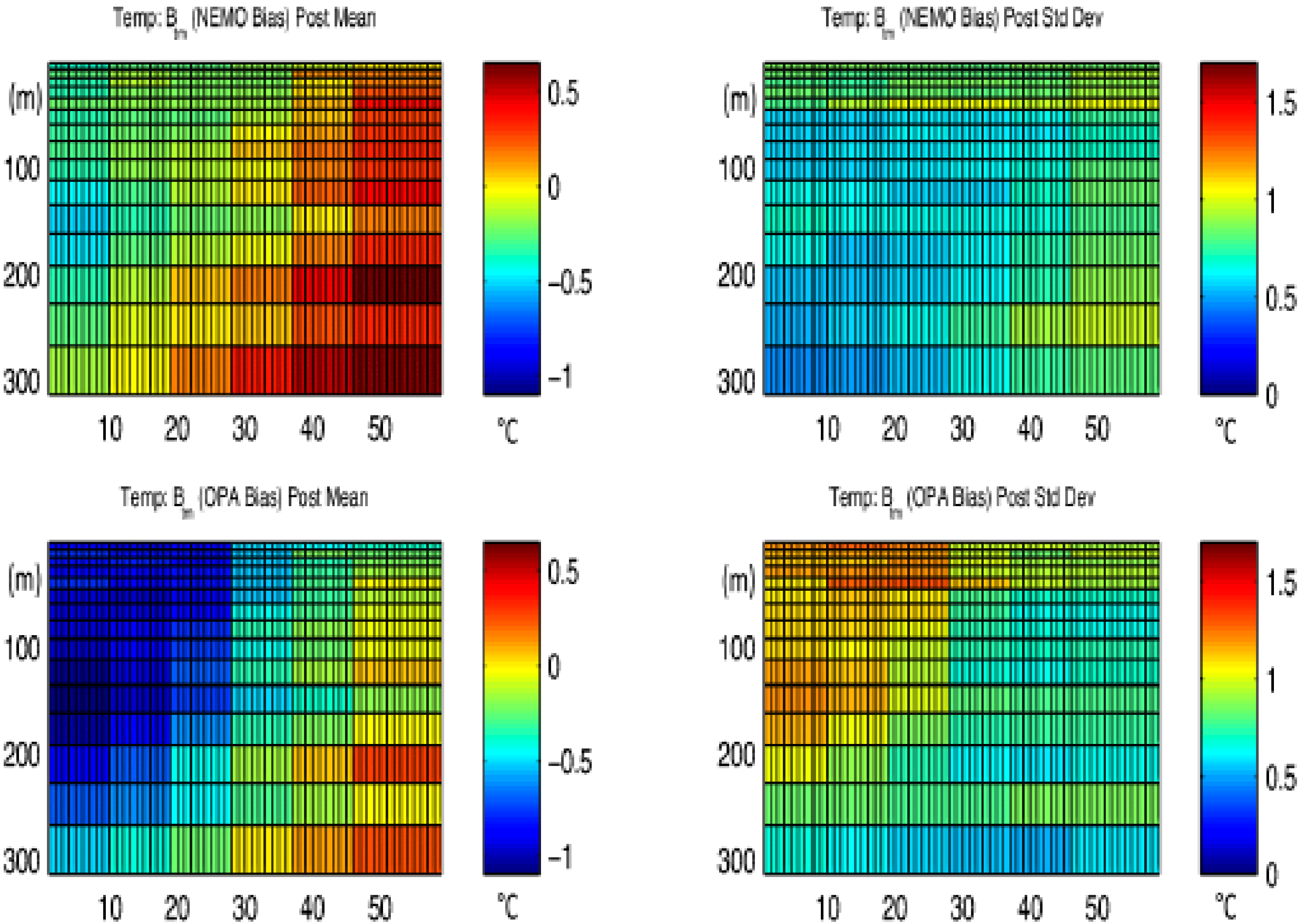
Temp: Sys3a2,
(m)
100
200
300
10
20
30
40
50
(day)



Temp: Sys + X(t) Post Mean
(m)
100
200
300
10
20
30
40
50
17
16
15
14
°C



Temp: X(t) Post Standard Deviation
(m)
100
200
300
10
20
30
40
50
1
0.5
0
°C

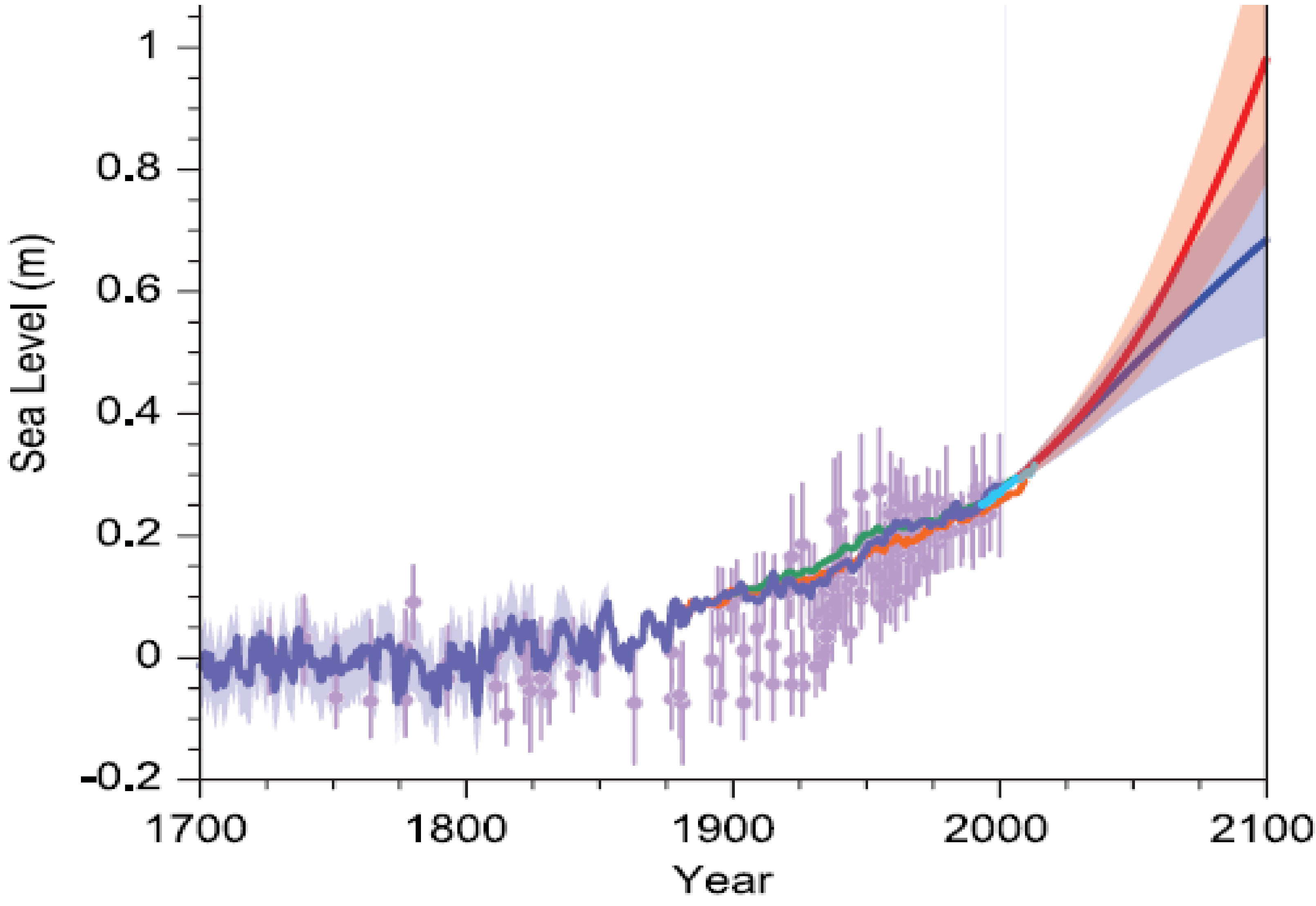

Temp: B (NEMO Bias) Post Mean
Temp: B (NEMO Bias) Post Std Dev
Temp: B (OPA Bias) Post Mean
Temp: B (OPA Bias) Post Std Dev
(m)
100
200
300
10
20
30
40
50
0.5
0
−0.5
−1
1.5
1
°C

Example 2: Predicting *Local* Sea Levels

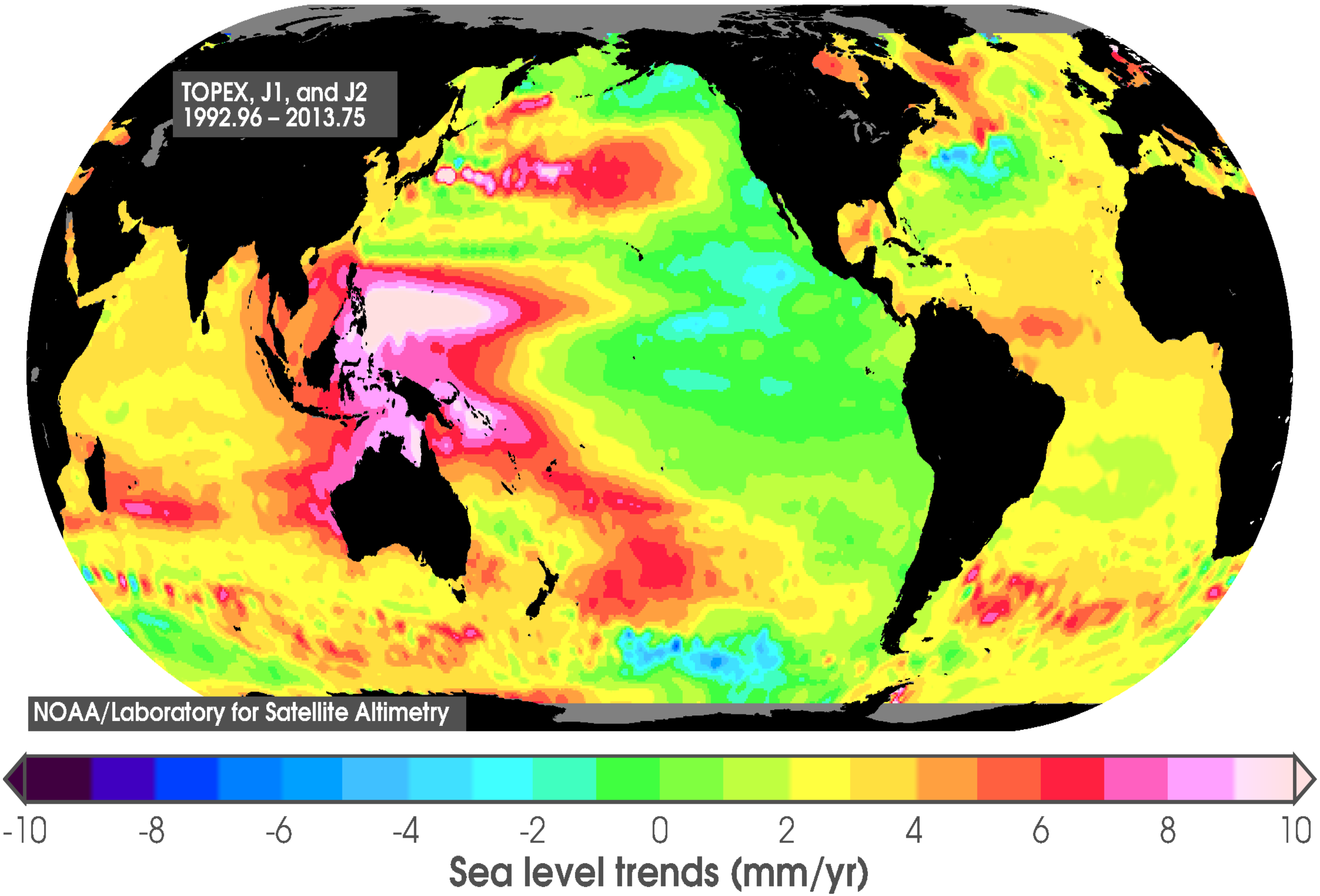
- **Global warming leads to**
  - **Ice melting**
  - **Warmer oceans leads to thermal expansion**
- **Next 100 years: 3-6 ft rise possible**
- **Impacts**
  - **3.2 Billion people live within 200 km of a coast**
  - **Pop. centers: New York, London, Netherlands,…**
  - **New infrastructure: eg., ports**

- **Data: tide gauges at numerous locations**
- **Focus: local & regional coastal sea levels**

IPCC

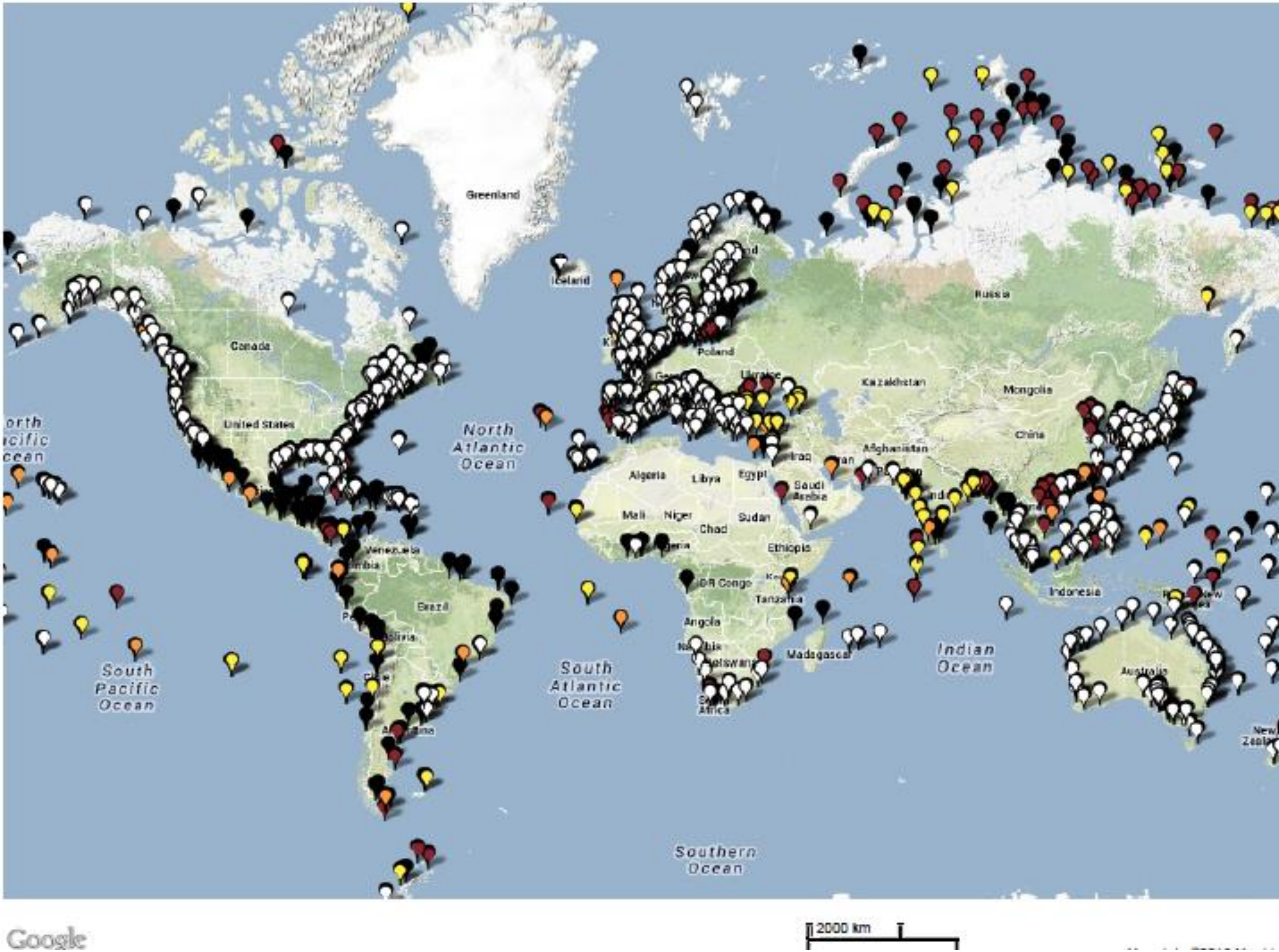


Spatial variation in sea level rise



Plan

- Manage substantial regional variations
- Use information at local scales (targets of prediction analysis)
- Form time series models that
  - Borrow strength across spatial scales
  - Incorporate temperature data
- Use climate model temperature projections to project local sea level


Greenland
Iceland
Canada
United States
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean
Russia
Poland
Ukraine
Kazakhstan
Mongolia
China
Afghanistan
Iraq
Iran
Saudi Arabia
Egypt
Libya
Algeria
Mali
Niger
Chad
Sudan
Ethiopia
DR Congo
Tanzania
Angola
Madagascar
Venezuela
Brazil
Bolivia
Chile
Indonesia
Australia
New Zealand
Google
2000 km

Model overview

- Each site has it’s own time series model: at month t
  - AR(2) in sea level at t-1, t-2
  - Linear terms in hemispheric temperature at t, t-1, t-2
  - Its own parameters!!!
- Site-wise parameters are samples from regional model with regional parameters
- Region-wise parameters are samples from hemispheric model with hemispheric parameters
- Hemispheric parameters have priors

Stage 1 Model: Each site has its own parameters

Sea level $S(t,s(r))$ (month t, site s in region r)

Temperature $T(t, h(s))$ for hemisphere containing s

$\alpha(m(t), s(r))$: monthly intercept

$\beta(s(r))$: coefficient of temp

$\phi(1, s(r))$; $\phi(2, s(r))$ autoreg coeff for lags 1; 2

$$
\begin{aligned}
S(t, s(r)) = {} & \alpha(m(t), s(r)) + \beta(s(r))\, T(t, h(s)) \\
& + \phi(1, s(r))\, [S(t-1, s(r)) - \{\alpha(m(t-1), s(r)) \\
& \qquad + \beta(s(r))\, T(t-1, h(s))\}] \\
& + \phi(2, s(r))\, [S(t-2, s(r)) - \{\alpha(m(t-2), s(r)) \\
& \qquad + \beta(s(r))\, T(t-2, h(s))\}] \\
& + e(t, s(r))
\end{aligned}
$$

Stage 2 Model

Local Parameters

$(\alpha(1,s(r)), \ldots, \alpha(12,s(r)))$

$\sim^{iid}$ MVN[$(\alpha(1,r), \ldots, \alpha(12,r))$, $\Sigma(r)$]

$\beta(s(r))$ $\sim^{iid}$ N[$\beta(r), \sigma(r)$]

(similar priors for local ϕ ‘s)

Stage 3 Model

Regional Parameters

$(\alpha(1,r), \ldots, \alpha(12,r))$

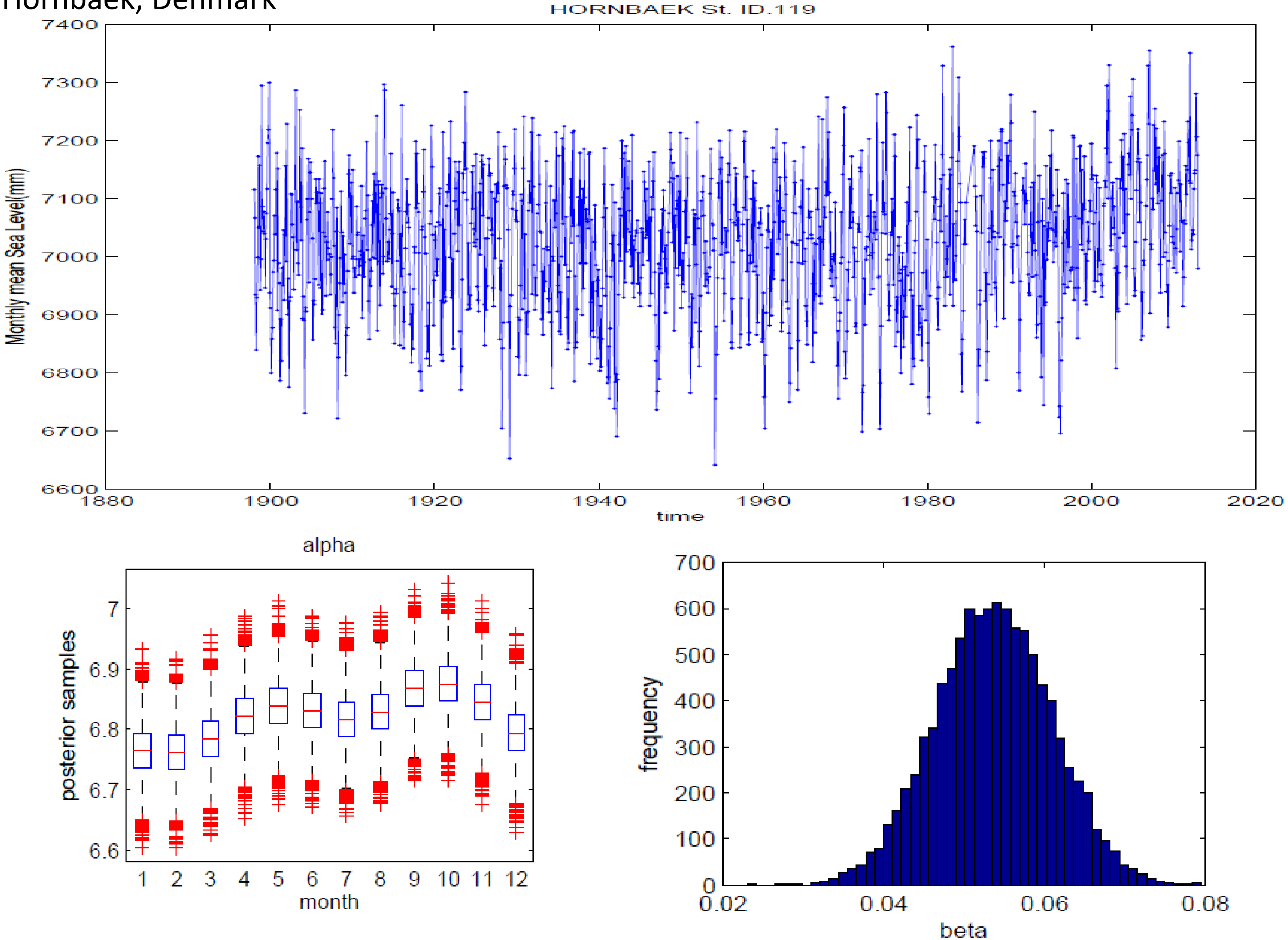
$\sim^{iid}$ MVN[$(\alpha(1,h), \ldots, \alpha(12,h))$, $\Sigma(h)$]

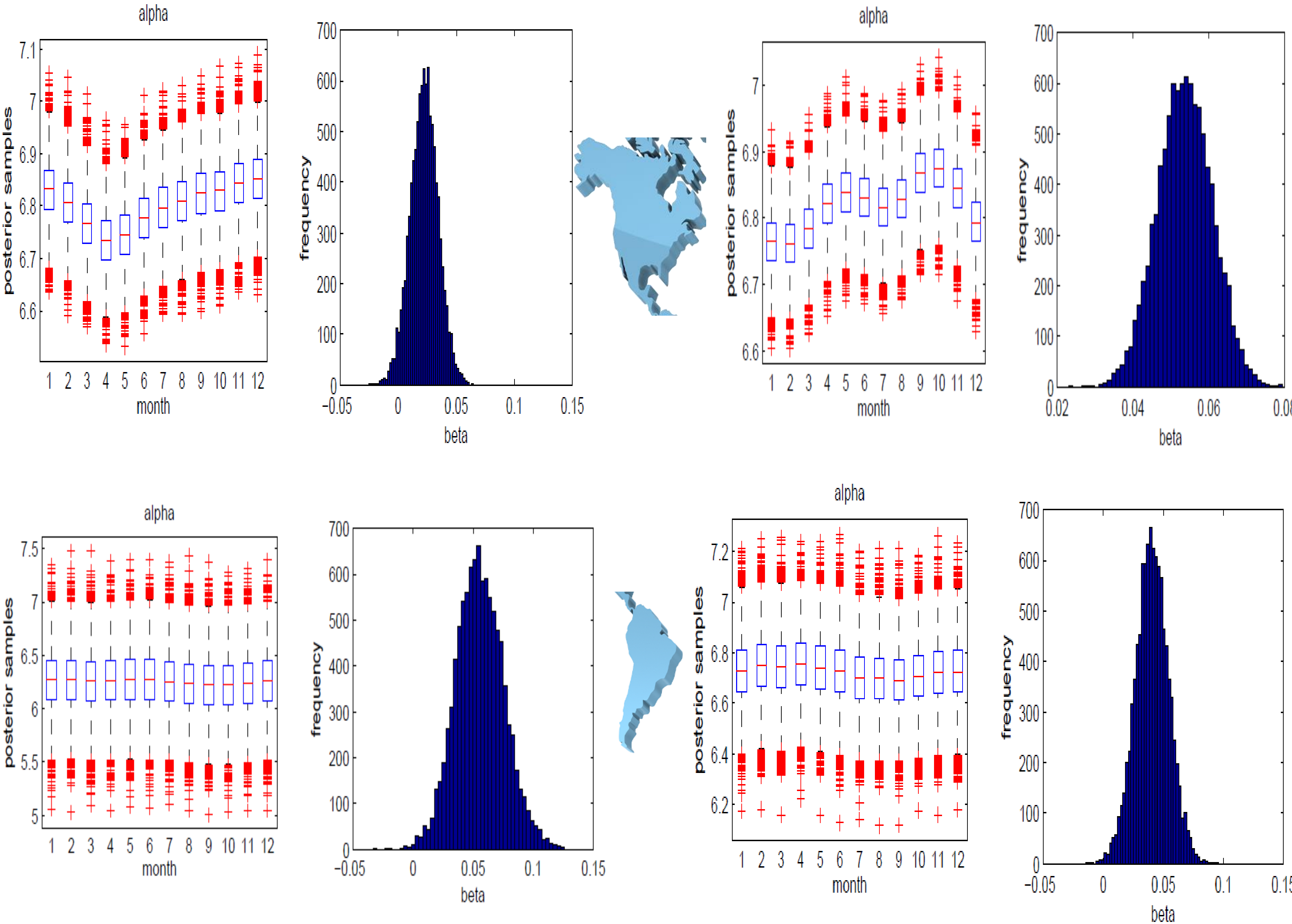
$\beta(r)$ $\sim^{iid}$ N[$\beta(h)$, $\sigma(h)$]

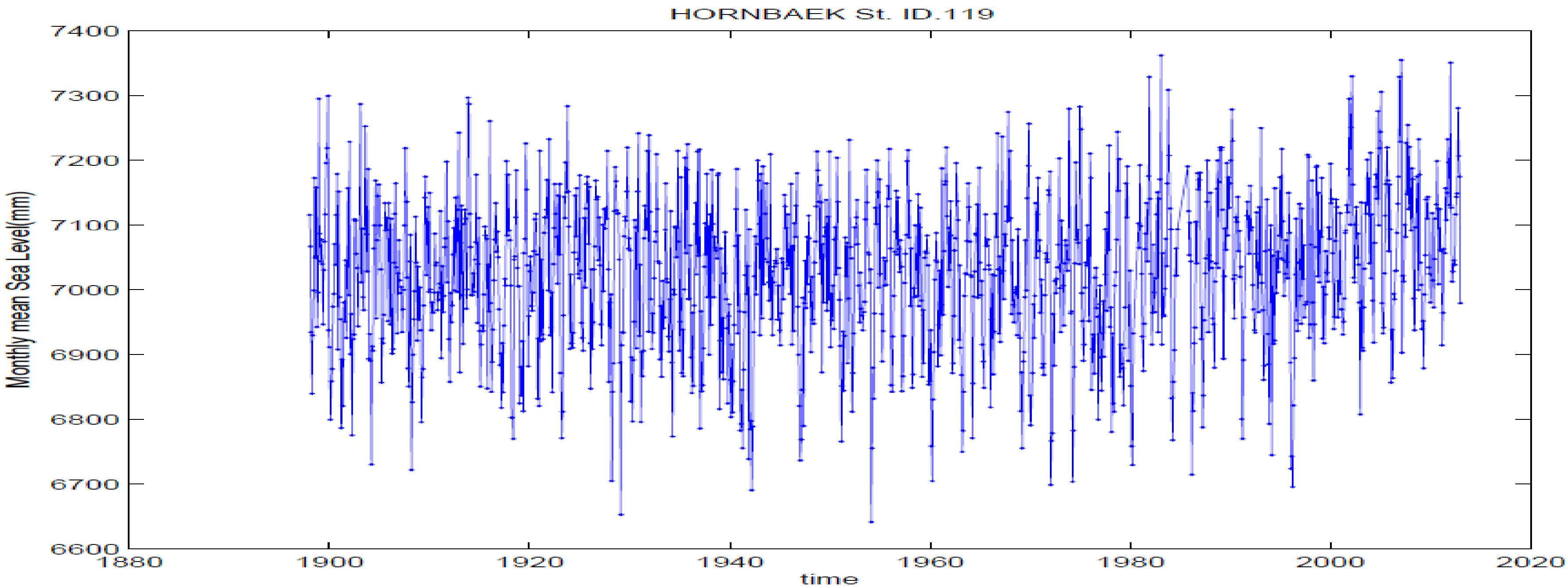
(similar priors for regional ϕ 's)

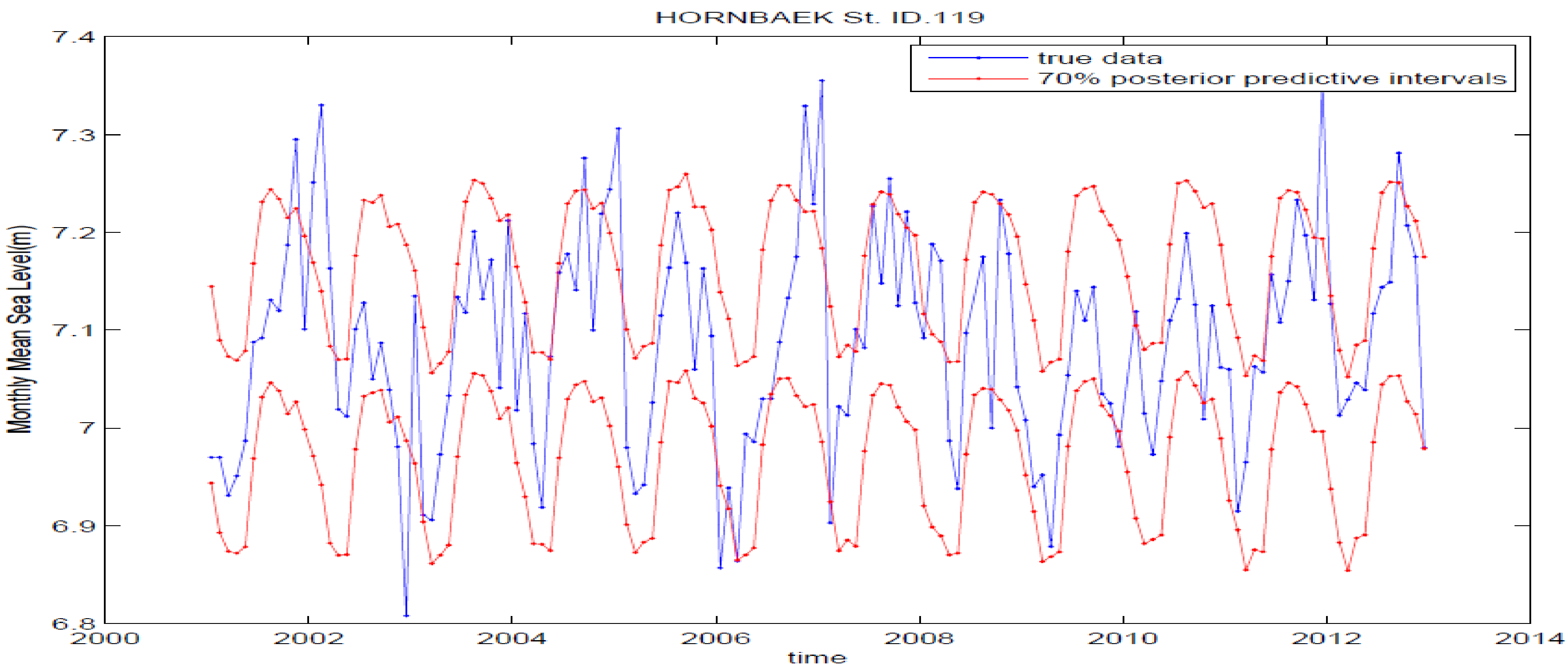
Prior on variances…..

Hornbaek, Denmark




alpha
posterior samples
month
frequency
beta
alpha
posterior samples
month
frequency
beta
alpha
posterior samples
month
frequency
beta
alpha
posterior samples
month
frequency
beta


HORNBAEK St. ID.119
Monthly mean Sea Level(mm)
7400
7300
7200
7100
7000
6900
6800
6700
6600
1880
1900
1920
1940
1960
1980
2000
2020
time


HORNBAEK St. ID.119
true data
70% posterior predictive intervals
Monthly Mean Sea Level(m)
7.4
7.3
7.2
7.1
7
6.9
6.8
2000
2002
2004
2006
2008
2010
2012
2014
time

Next steps

- Use climate projections of temperature to develop projections of local sea levels
- Remark: “medium-range” forecasting

  (ie, maximum horizon of 5-10 years)
- Issues
  - Use more localized temperatures?
  - Why use temp? Climate models produce regional scale sea levels
- Attribution of change to anthropogenic inputs
  - Important for local variables and impacts

  eg) causal relationships for agriculture, disease, etc.
  - Decision support
- Thank you!!!