The interpretation of model output: making sense of it all…

A. L. Mila

North Carolina State University

NC STATE UNIVERSITY

Use an example

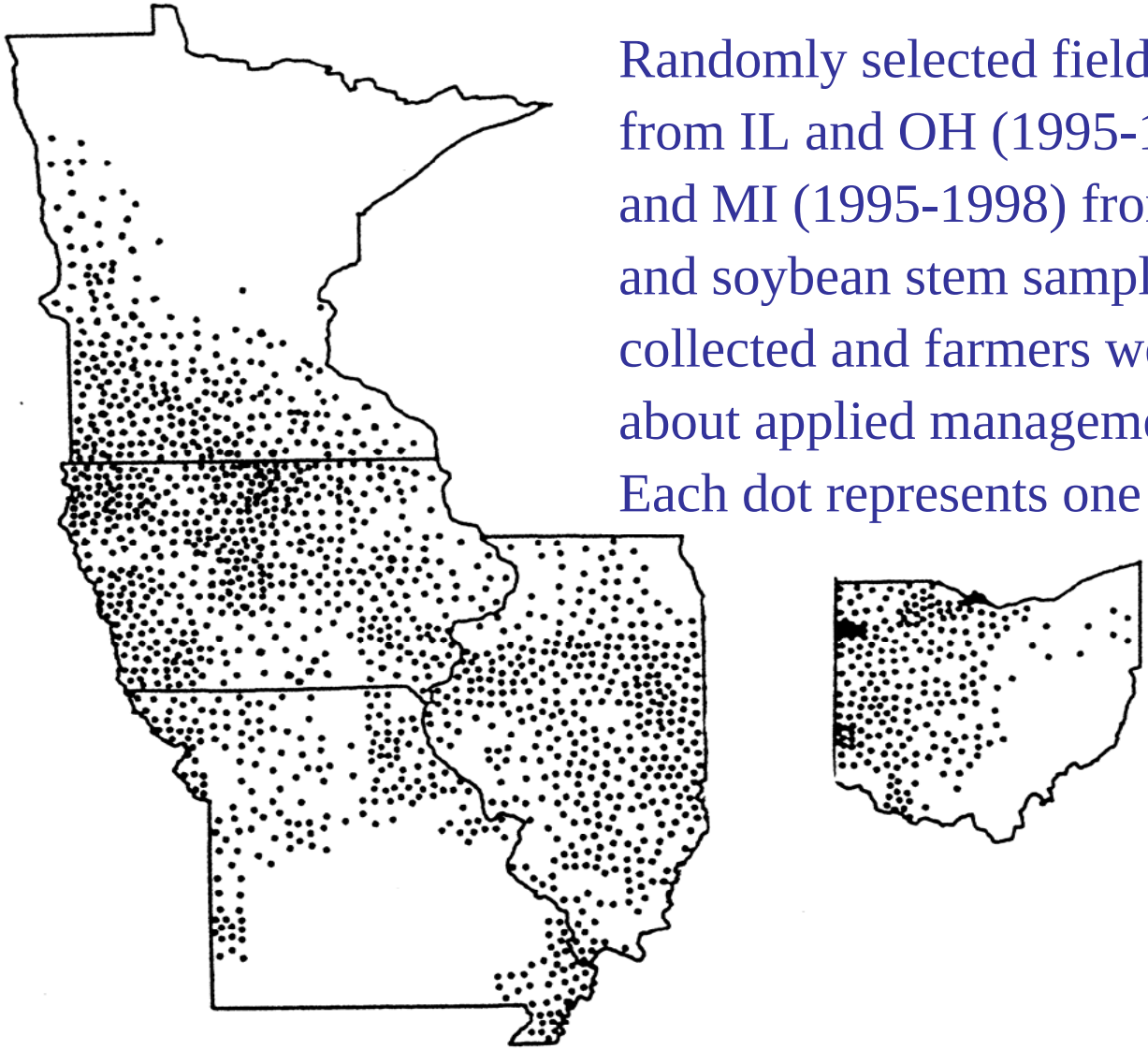
Prevalence of Sclerotinia Stem Rot of soybeans

in North Central US

Define scale of modeling

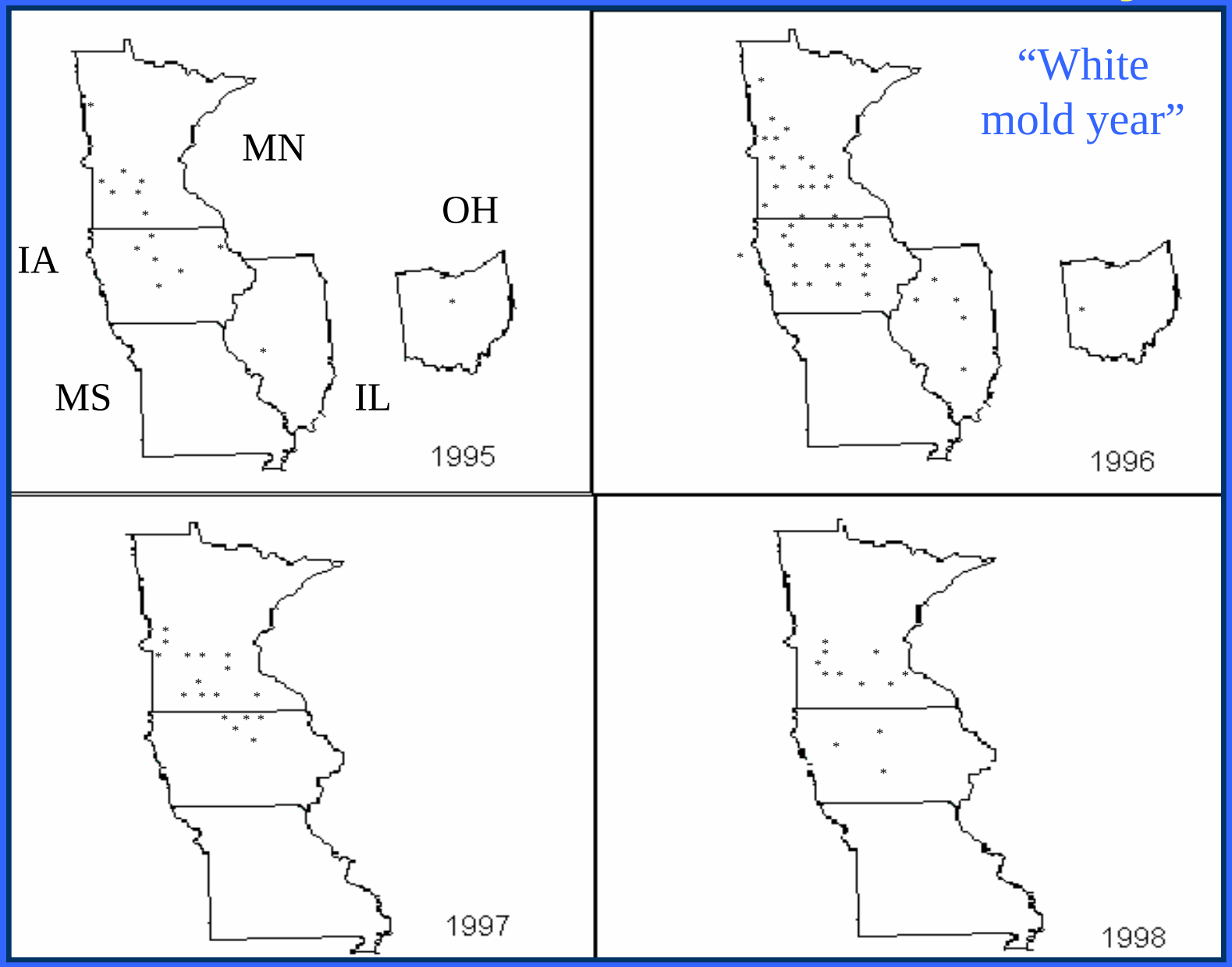
- Large
- Multiple counties
- Or multiple states

Sample collection



Randomly selected fields by NASS from IL and OH (1995-1996), IA, MN and MI (1995-1998) from which soil and soybean stem samples were collected and farmers were interviewed about applied management practices. Each dot represents one field.

Fields with Sclerotinia stem rot of soybeans



Question to answer:

What triggers *prevalence* of SSR in North Central US?

Prevalence = presence or absence of SSR

in a field (Yes or No)

Model development

Logistic regression analysis

Model Input

Input variables

- Monthly average air temperature for the months of April through August
- Monthly total precipitation for the months of April through August

- Tillage practices
- Regional effect

Model output

- *Summer weather*

  Temperature (July-August)

  Precipitation (July)

- *Spring weather*

  Temperature (April-May)

  Precipitation (April)

Which output is the one to use?

Which output is the one to use?

- Literature available on effect of summer weather on SSR occurrence
- High correlations between spring and summer weather variables

Continue with summer weather variables…

Model output:

A. Important explanatory variables

- Air temperature of July and August
- Precipitation of July
- Geographic region

B. Importance of air temperature depends on:

- Tillage practices

How certain are we about the singificance of these variables?

….. using Bayesian methods

Bayesian inference: Key ideas

- parameters as random variables

- probability used to describe uncertainty about parameters

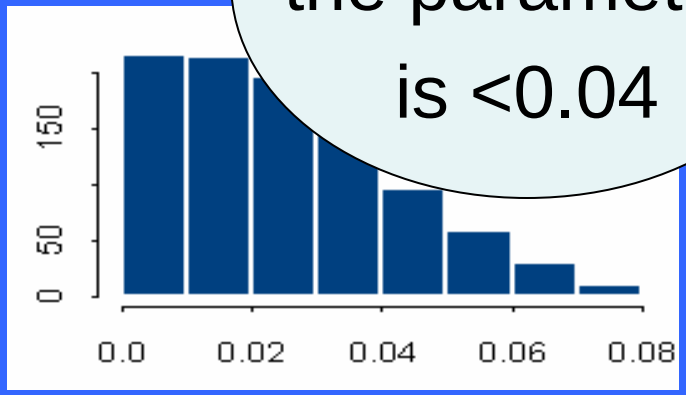
Usually:

Mean

St. Error

Confidence Interval

Baye



75% chances the parameter is <0.04

Bayesian inference (con't)

- combine information (prior distributions and current data)

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)$$

Posterior Distribution

What I know at the end…

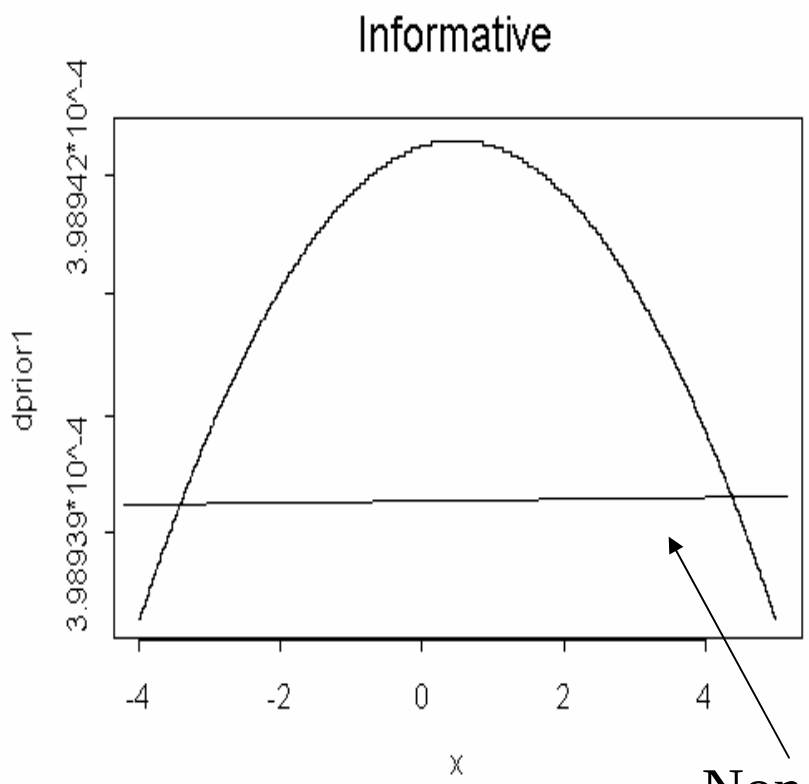
Data

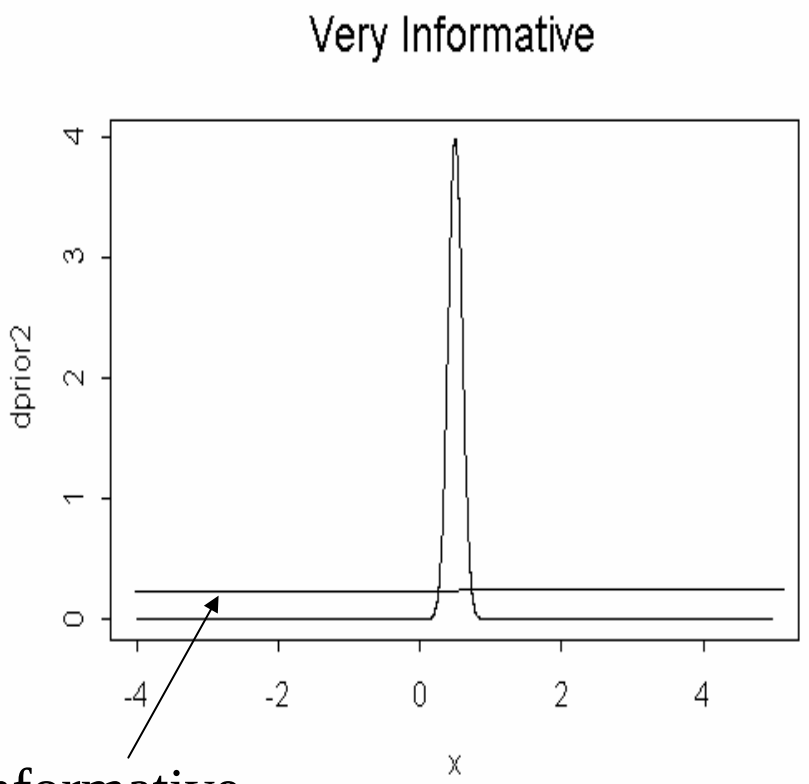
Prior distribution

Prior knowledge?

Expertise?

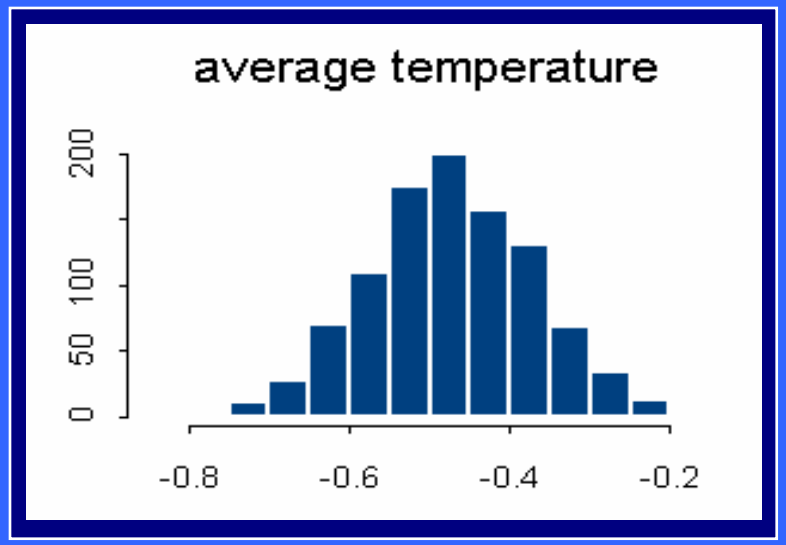
Non-informative, informative and very informative priors

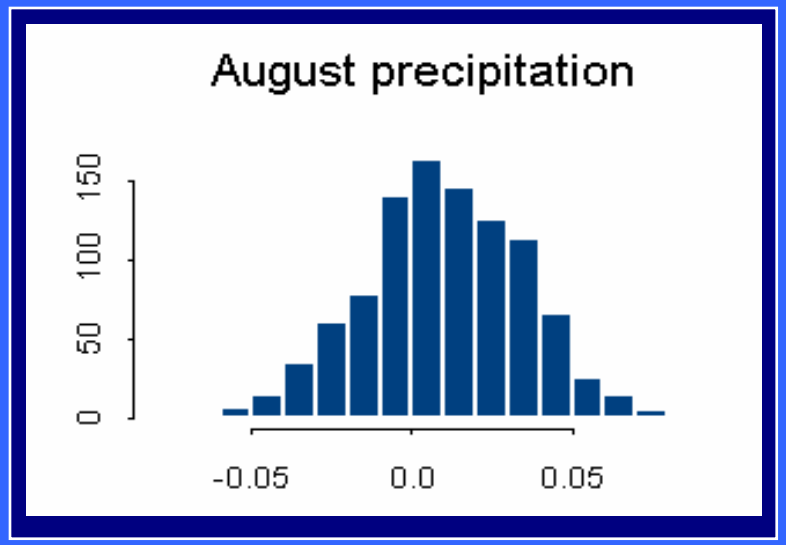




Non-informative

Back to SSR example
(Non informative priors)



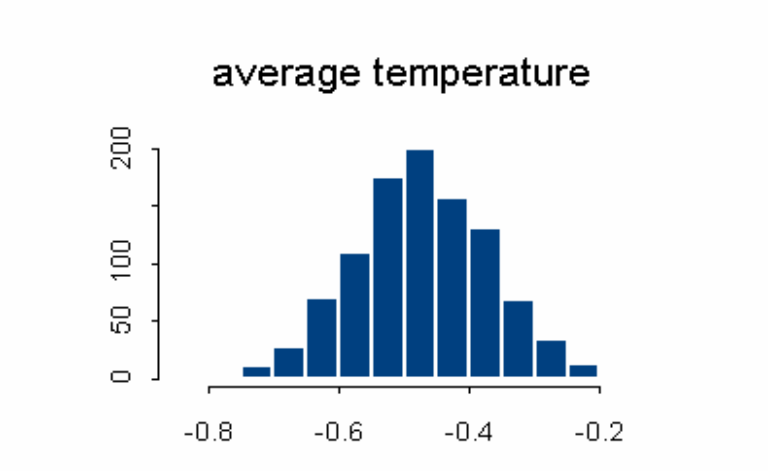


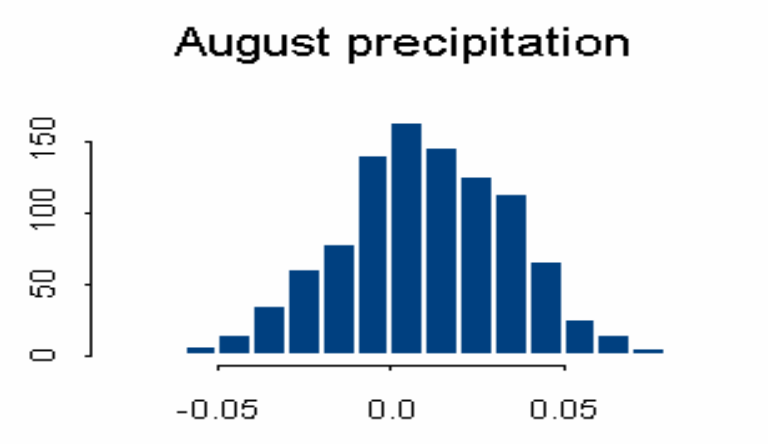
Significant (P>.05)

Not Significant (P>.05)

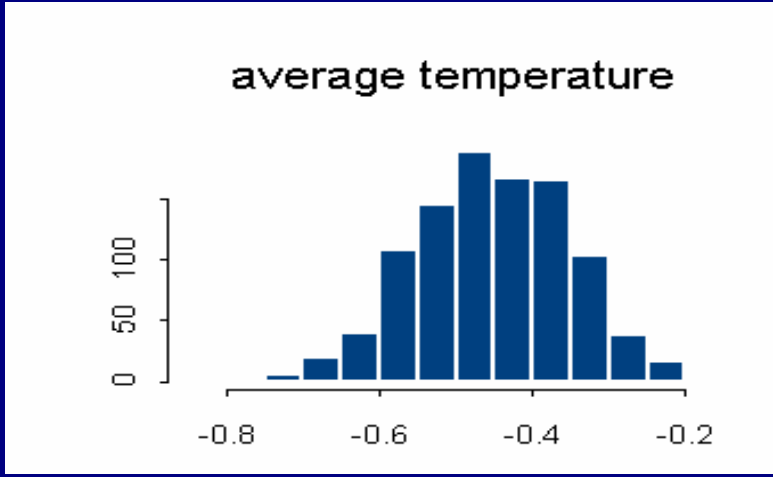
Informative priors

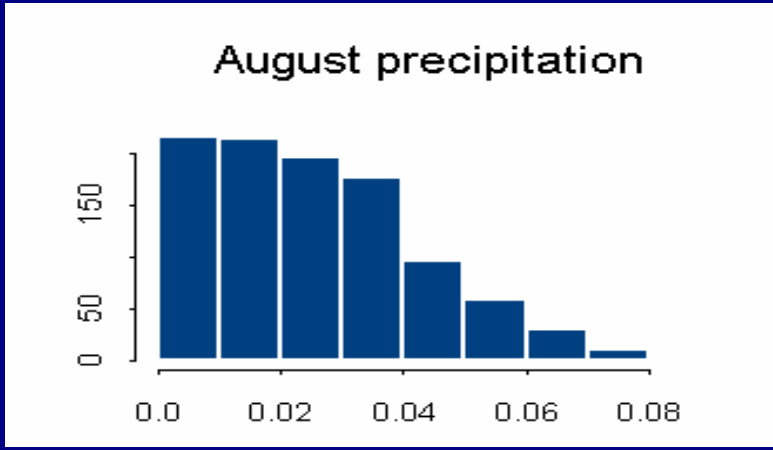
Non-informative





Informative





Use of prior distributions

To detect “sensitivity to the priors” for parameter estimates

A. If parameters **are not** sensitive, current data sufficient for parameter estimation

B. If parameters **are** sensitive, current data are not very informative

Model output

- *Summer weather*

  Temperature (July-August)

  Precipitation (July)

- *Spring weather*

  Temperature (April-May)

  Precipitation (April)

Which output is the one to use?

Which output is the one to use?

Should we choice one of the two outputs?

OR

Use both outputs?

(i.e. by using Bayesian Model Averaging, BMA)

Thank you…