Predictor selection

Michael K. Tippett

International Research Institute for Climate and Society
The Earth Institute, Columbia University

Statistical Methods in Seasonal Prediction, ICTP
Aug 2-13, 2010

The problem of how best to select which predictors to include in a model is a nontrivial, unsolved one.

> *"All models are wrong but some are useful."*
>
> *–George Box*

The difficulty comes from having to estimate future performance from past behavior.

> *"Past performance is no guarantee of future results."*
>
> *– Any investment document'*

As a forecaster, it is better to know a model has poor skill than to mistakenly think a poor model has good skill.

> *"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so"*
>
> *– Mark Twain*

The problem of how best to select which predictors to include in a model is a nontrivial, unsolved one.

"All models are wrong but some are useful."

–George Box

The difficulty comes from having to estimate future performance from past behavior.

"Past performance is no guarantee of future results."

– Any investment document'

As a forecaster, it is better to know a model has poor skill than to mistakenly think a poor model has good skill.

"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so"

– Mark Twain

The problem of how best to select which predictors to include in a model is a nontrivial, unsolved one.

"All models are wrong but some are useful."

–George Box

The difficulty comes from having to estimate future performance from past behavior.

"Past performance is no guarantee of future results."

– Any investment document'

As a forecaster, it is better to know a model has poor skill than to mistakenly think a poor model has good skill.

"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so"

– Mark Twain

Problem

Given a pool of candidate predictors, how to do select those to include in a prediction model?

(Why not the model that best fits the data?)

Goal: a model which skillfully predicts *independent* data.

Problem

Given a pool of candidate predictors, how to do select those to include in a prediction model?

(Why not the model that best fits the data?)

Goal: a model which skillfully predicts *independent* data.

Problem

Given a pool of candidate predictors, how to do select those to include in a prediction model?

(Why not the model that best fits the data?)

Goal: a model which skillfully predicts *independent* data.

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic method.
- All can be tricked by screening
- Avoid methods that can construct spurious methods.
- Include screening in predictor selection procedure.

Indirect methods (no use of independent data):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic method.
- All can be tricked by screening
- Avoid methods that can construct spurious methods.
- Include screening in predictor selection procedure.

Indirect methods (no use of independent data):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic method.
- All can be tricked by screening
- Avoid methods that can construct spurious methods.
- Include screening in predictor selection procedure.

Indirect methods (no use of independent data):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic method.
- All can be tricked by screening
- Avoid methods that can construct spurious methods.
- Include screening in predictor selection procedure.

Indirect methods (no use of independent data):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Example: DJF temperature

Predictand (y)

- Average Dec-Feb 1962-2003 temperature over land. (42 years)

Predictors (x)

- Climatology
- Sep-Nov NINO 3.4.
- Trend

Consider 4 possible sets of predictors.

- Climatology
- Climatology & Sep-Nov NINO 3.4.
- Climatology & Trend
- Climatology & Sep-Nov NINO 3.4.& Trend

Example: DJF temperature

Predictand (y)

- Average Dec-Feb 1962-2003 temperature over land. (42 years)

Predictors (x)

- Climatology
- Sep-Nov NINO 3.4.
- Trend

Consider 4 possible sets of predictors.

- Climatology
- Climatology & Sep-Nov NINO 3.4.
- Climatology & Trend
- Climatology & Sep-Nov NINO 3.4.& Trend

F-test

Reference forecast = “climatology” (1-parameter model).
Compare the SSE of a P-predictor model with that of the 1-predictor reference model.

$$f = \frac{\frac{SSE_1 - SSE_P}{P-1}}{\frac{SSE_P}{N-P}}$$

where

- $SSE_1 = \sum_{i=1}^{N}(Y_i - \overline{Y})^2$ is the sum of squared error for the climatology forecast.
- $SSE_P = \sum_{i=1}^{N}(Y_i - Y_{Pi})^2$ is the sum of squared error for the model with P predictors,
- N is the sample size.

Nested.

F-test

$$f = \frac{\frac{SSE_1 - SSE_P}{P-1}}{\frac{SSE_P}{N-P}}$$

- Under the null hypothesis that the P-parameter model is not better than the 1-parameter model, f has an F distribution with parameters $(P-1, N-P)$.
- Compute the associated $\alpha = Prob(F > f)$ probability value.
- Find the model with the lowest α.
- Check that α is smaller than some limit (5%). If α exceeds the limit, use climatology forecast.

F-test

A correction is needed for multiple comparisons.

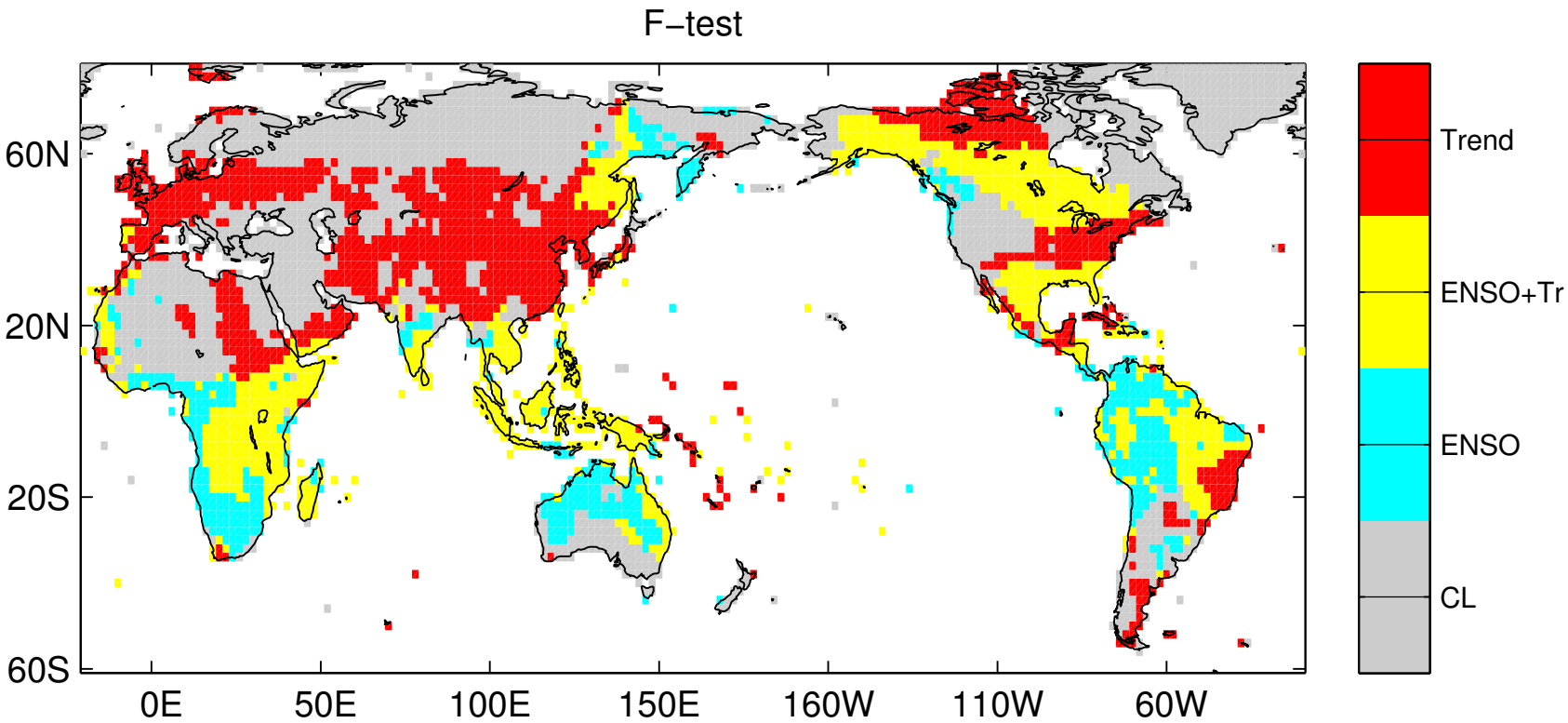
$$\alpha \to \alpha/(m-1)$$

Not quite right (not independent).

Modest values of m lead to very strict requirements on the significance level.

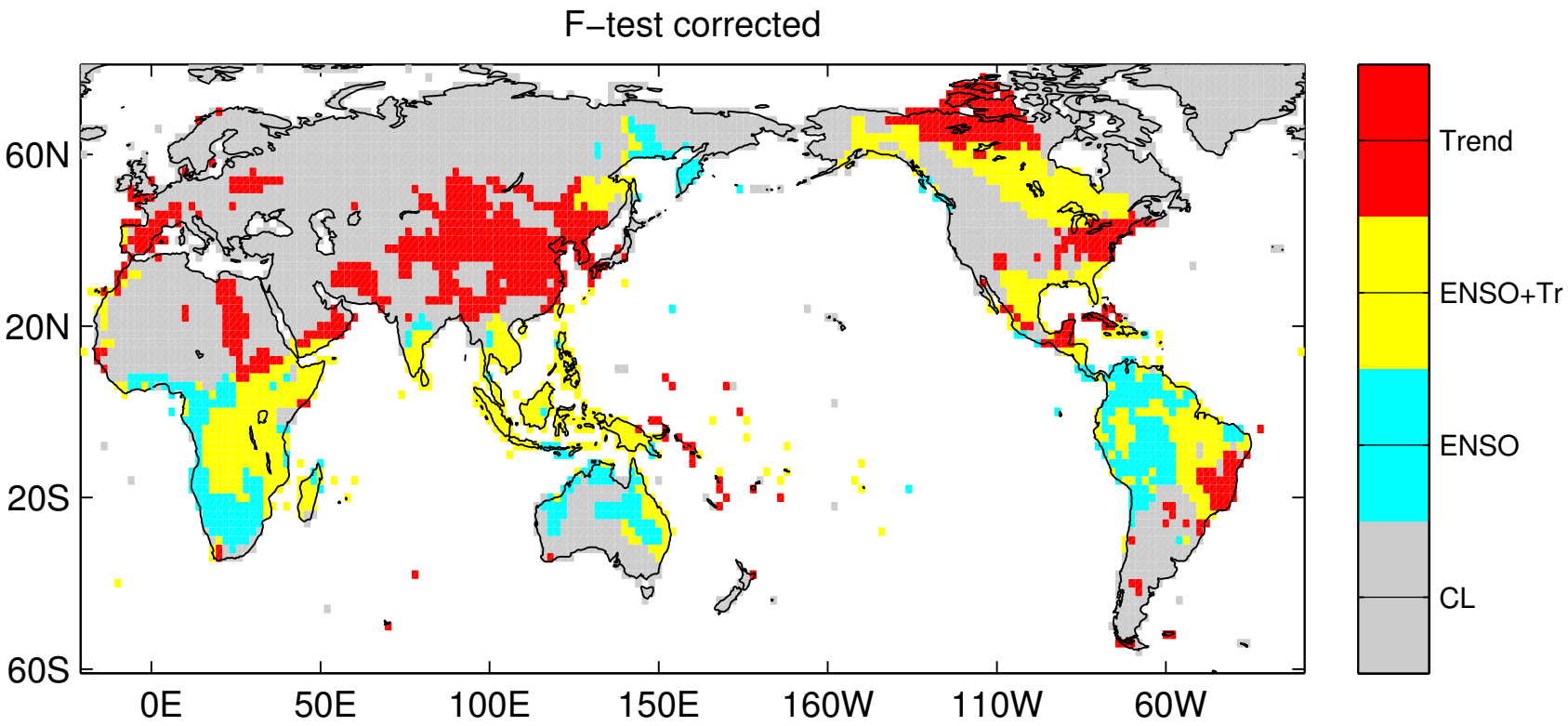
Example: DJF temperature

Models selected at each gridpoint using the F-test ($\alpha \leq 0.05$)



Example: DJF temperature

Models selected at each gridpoint using the F-test ($\alpha \leq 0.05/3$)



Mallow's C_P

$$C_P = \frac{SSE_P}{MSE_{full}} - N + 2P$$

where

- $SSE_P = \sum_{i=1}^{N}(Y_i - Y_{Pi})^2$ is the sum of squared error for the model with P predictors,
- Y_{pi} is the predicted value of the i-th observation of Y from the model with P predictors.
- $MSE_{full} = \frac{1}{N-K}\sum_{i=1}^{N}(Y_i - Y_{Ki})^2$ is the residual mean square of the model using the complete set of K predictors
- N is the sample size.

Mallow's C_P

$$C_P = \frac{SSE_P}{MSE_{full}} - N + 2P$$

If the extra variables are noise (no more variables needed)

$$E\left[\frac{SSE_P}{MSE_{full}}\right] = (n-p)\frac{\sigma_P}{\sigma_{full}} = n - p$$

and

$$E\left[C_P\right] = p$$

If the extra variables are useful (not enough variables in model), $\sigma_P > \sigma_{full}$ and

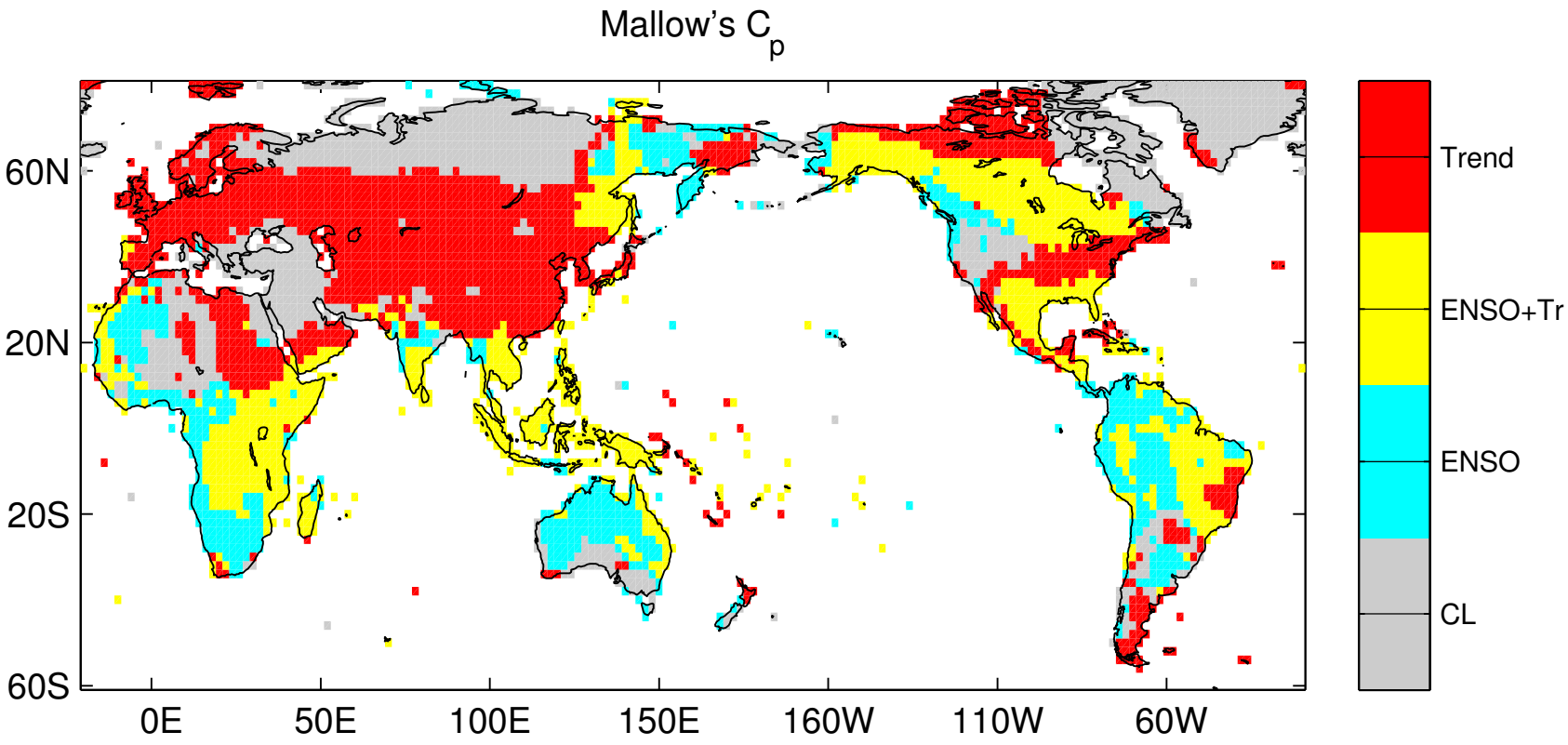
$$E\left[C_P\right] > p$$

The model with the lowest C_P value approximately equal to P is the most "adequate" model.
Strategies:

- Minimize C_p.
- Graphical

Example: DJF temperature

Models selected at each gridpoint using Mallow's C_P.



AIC

Information theory measure of the difference between model and truth.

- Maximize the likelihood of the model.
- Maximized likelihood is biased.
- Likelihood increases as the number of predictors increases. AIC corrects for this bias.

General case

$$AIC = -2 \log L + 2P$$

where L is the maximized likelihood of a model with P parameters.

Can be applied to any model where L is known.

AIC

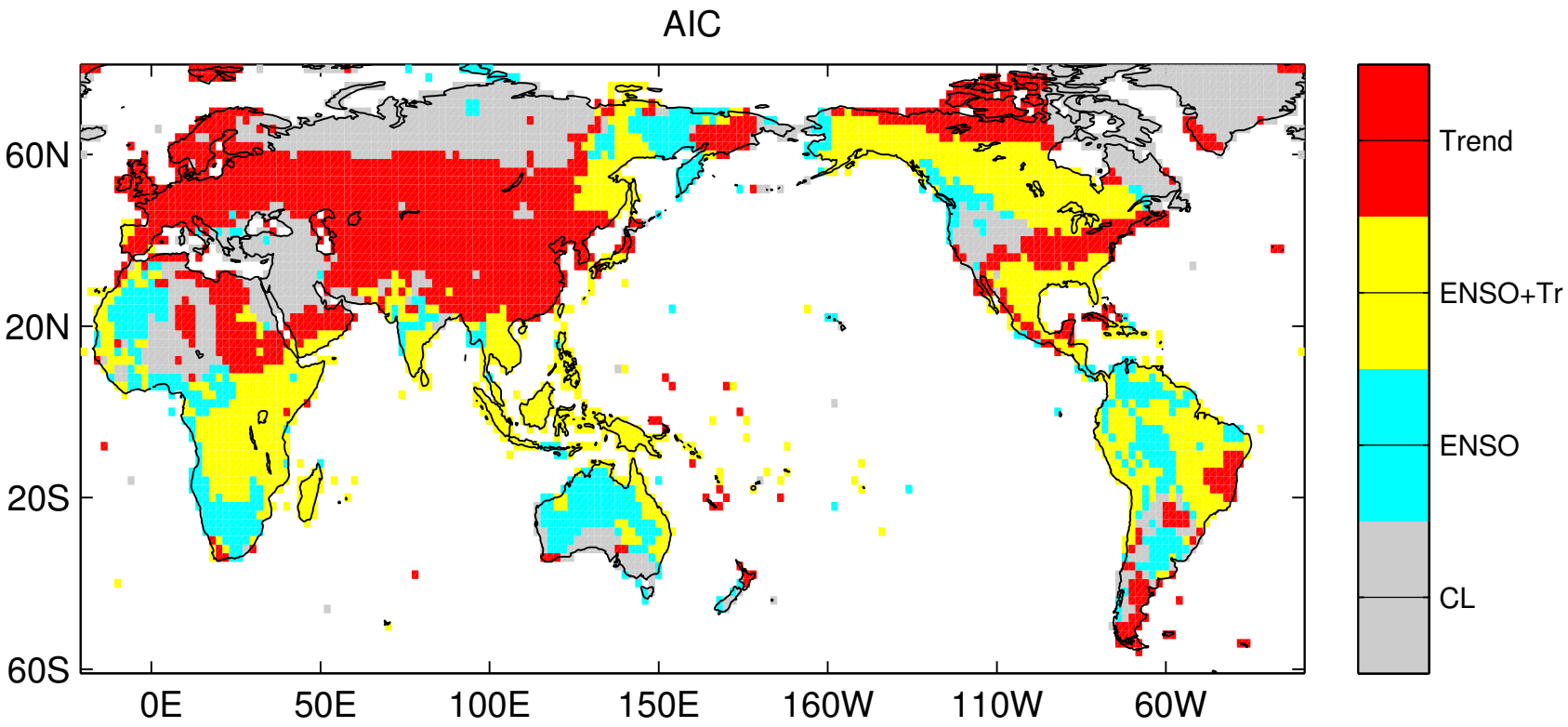
For linear regression (neglecting some constants),

$$AIC = N \log SSE_P + 2P$$

- $SSE_P = \sum_{i=1}^{N} (Y_i - Y_{Pi})^2$ is the sum of squared error for the model with P predictors,
- Rewards fit, penalizes complexity.
- Choose model that minimizes AIC.
- Differences in AIC are relevant.
  $\Delta < 2$ small.
  $4 < \Delta < 7$ large.
  $\Delta > 10$ very large.

Example: DJF temperature

Models selected at each gridpoint using AIC.



Corrected AIC

Correction for small sample size.
AIC is an approximation.
AICc is more accurate for small sample sizze.

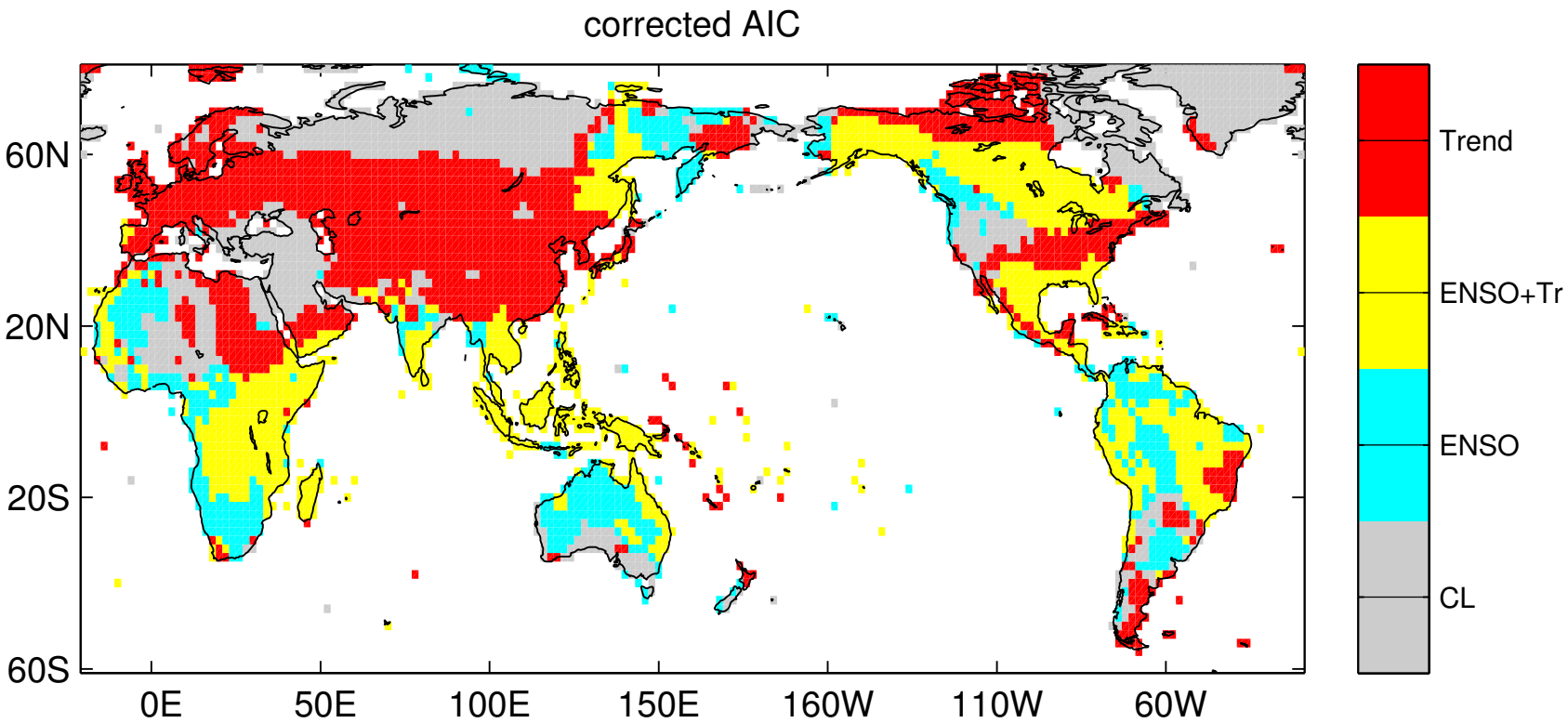
Should be used always (especially. for $N/P < 40$)

$$\text{AICc} = N \log SSE_P + 2P + \frac{2P(P+1)}{N-P-1}$$

Rewards fit, penalizes complexity a little more.

Example: DJF temperature

Models selected at each gridpoint using AICc.



BIC

Approximation to Bayes factor with equally likely priors.
(AIC = Bayes factor with "savvy" prior).

General case

$$BIC = -2 \log L + P \log N$$

where L is the maximized likelihood of a model with P parameters.

$$[AIC = -2 \log L + 2P]$$

Which picks simpler models? Why?

BIC

For linear regression (neglecting some constants),

$$BIC = N \log SSE_P + P \log N$$

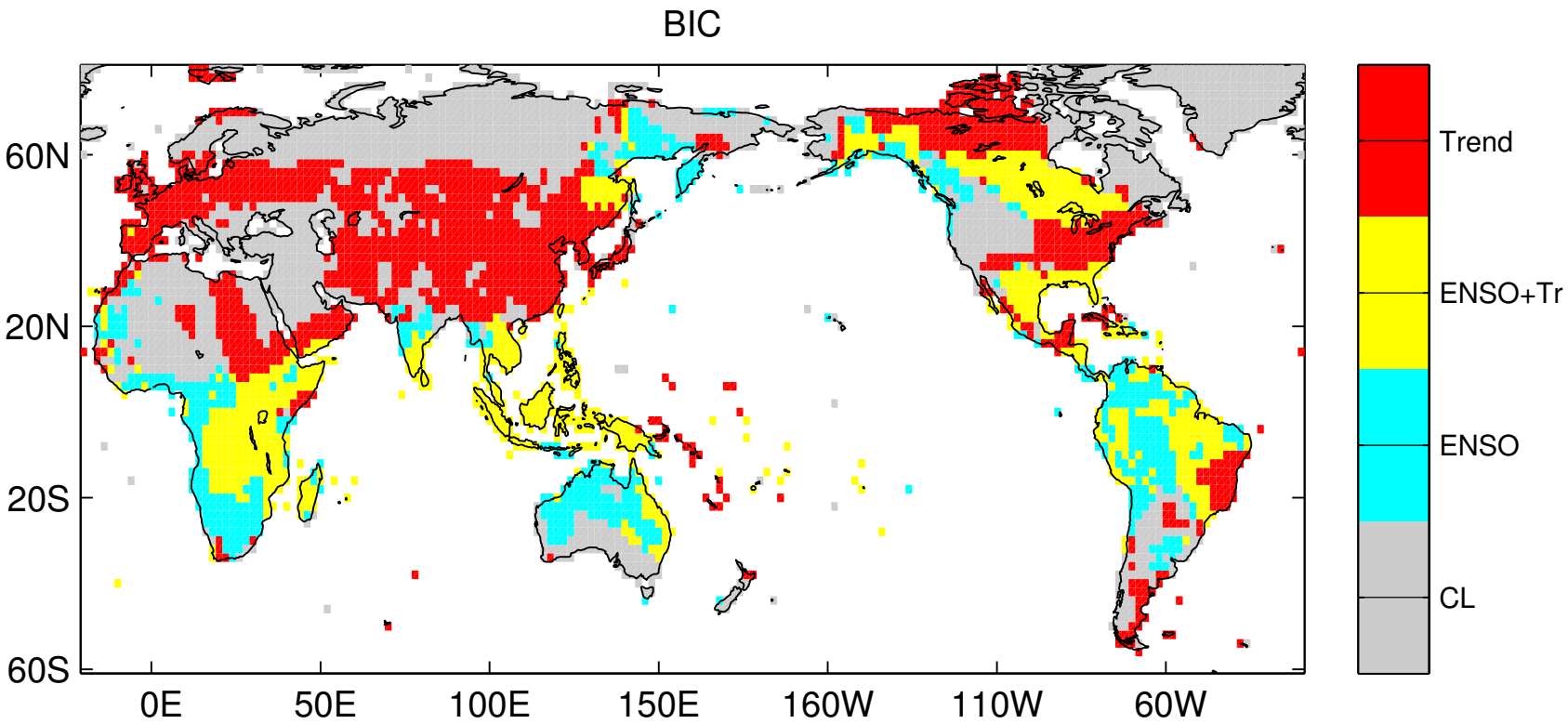
Rewards fit, penalizes complexity more than AIC.

May under-fit in small-moderate sample sizes.

AIC vs. BIC? Unsettled.

Example: DJF temperature

Models selected at each gridpoint using BIC.



Data splitting method

- Train the model on one data set.
- Apply it to a second independent data set.
- Choose the model that performs best on the second data set.

Note: a third data set is need to evaluate the skill of the selected model. Why?

The third data set . . .

A screening example.

- Your model = 20 random numbers.
- Generate such models.
- Check how well they predict the last 20 years.
- Pick the one that does best.

Skill in the selection data set is high.

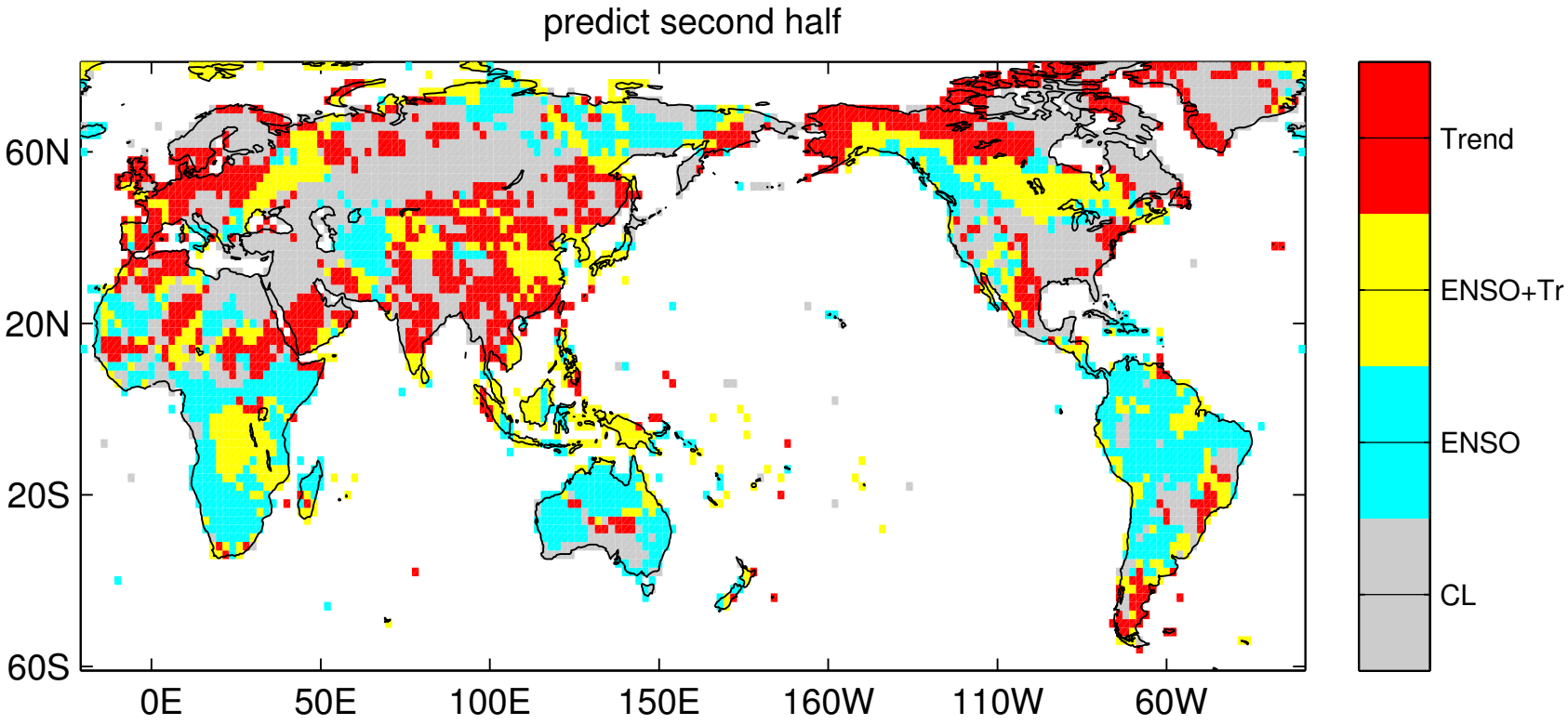
Skill in an independent data set (and real skill) would be low.

1. Avoid looking at many models.
2. Should separate model selection and skill estimation.

One approach is to avoid procedures that lead to the skill in the "third data set" being very different from that in the selection data set.

Example: DJF temperature

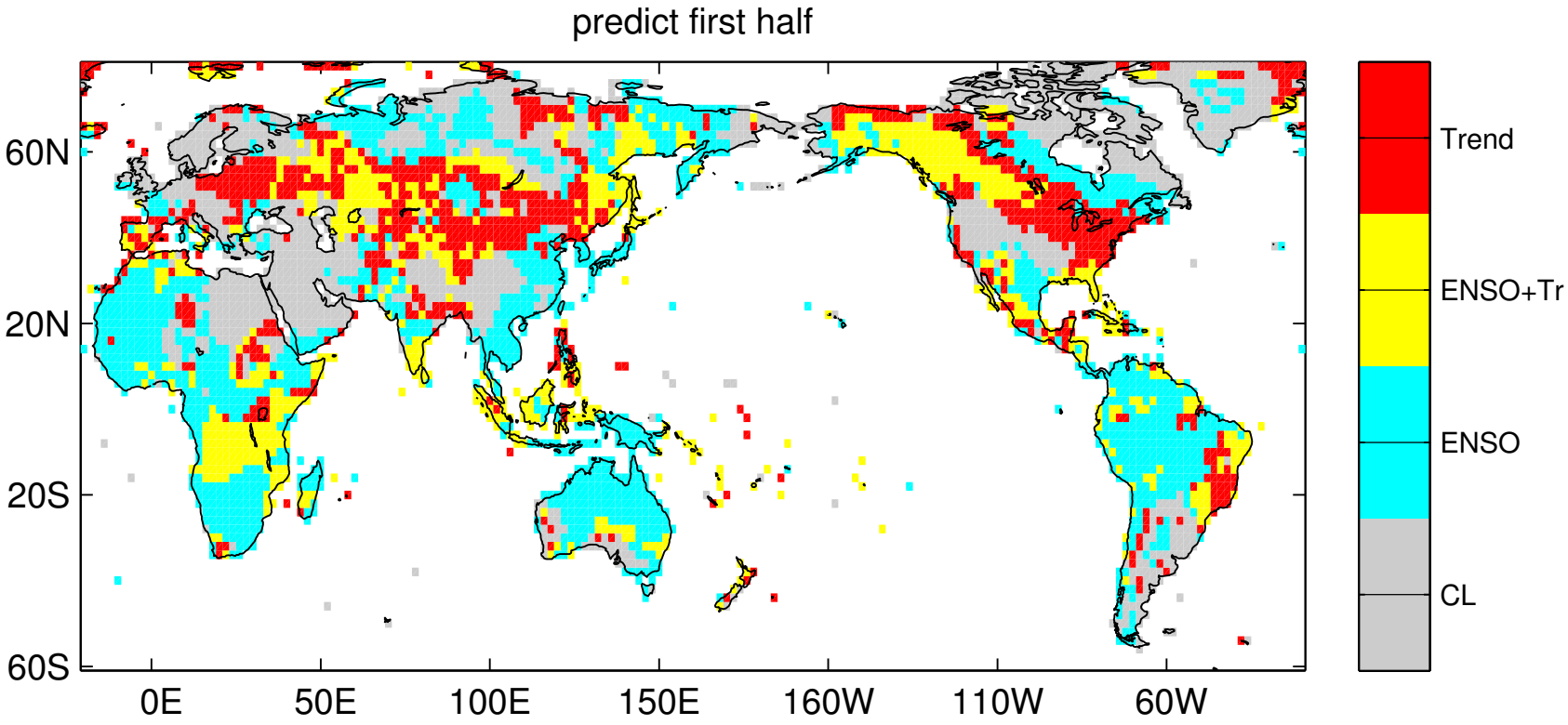
Models trained using 1962-1982 and selected at each gridpoint using skill 1983-2003.



Why noisier?

Example: DJF temperature

Models trained using 1983-2003 and selected at each gridpoint using skill 1962-1982.



Why noisier?

Cross validation

A method for mimicking actual forecasting.

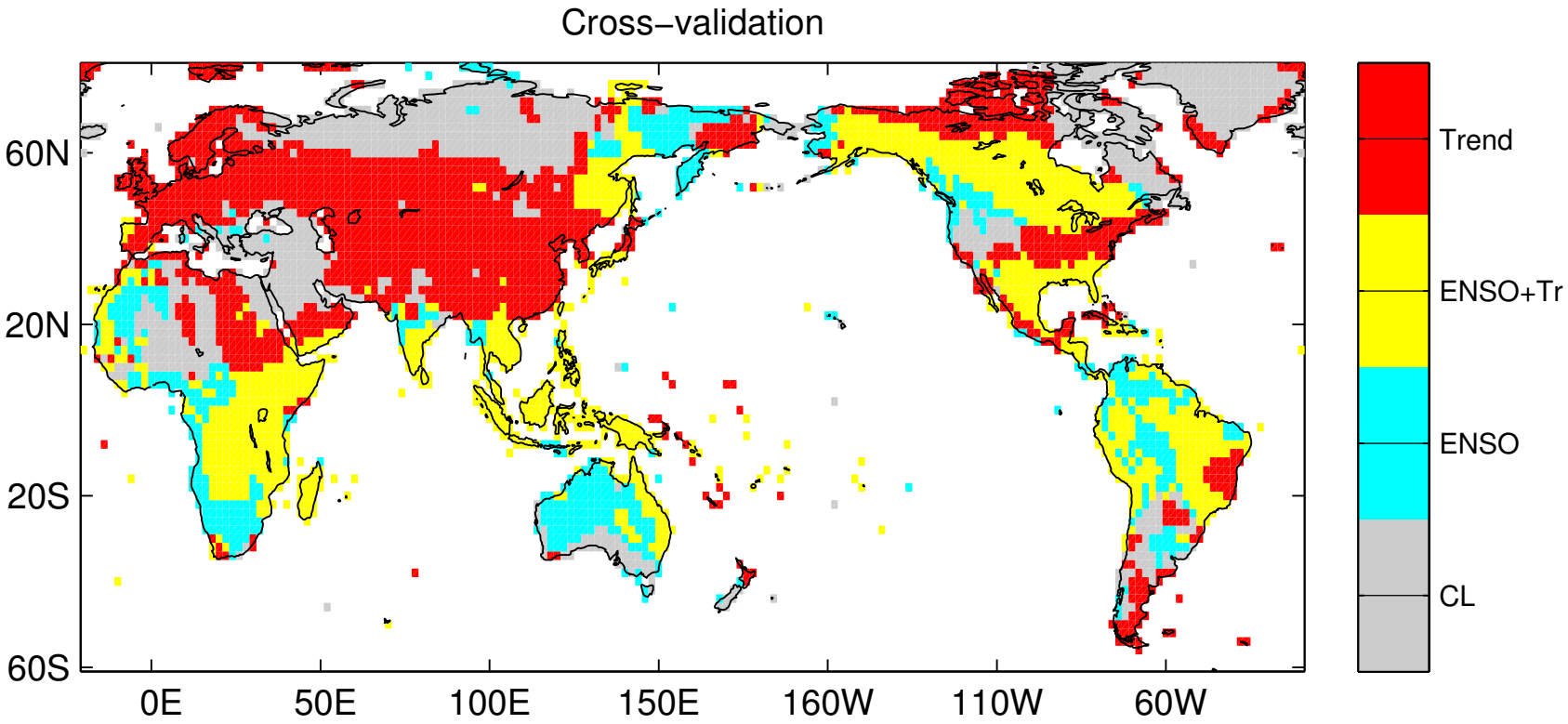
An alternative to splitting the data.

- Remove some number K of samples from the data set.
- Estimate the model on the remaining $N - K$ samples.
- Use that model to predict the K left-out samples.
  - Sometimes a set of K contiguous in time samples are left out and only the middle one is predicted to deal with temporal correlation.[More later]
- Repeat.

Often $K = 1$. Leave-one-out cross-validation.

Example: DJF temperature

Models selected at each gridpoint using leave-one-out cross-validation.



Summary of methods

Two types of methods

Balance between fit and number of predictors.

- F-test
- Mallow's C_P
- AIC (corrected), BIC

Apply model to independent data:

- Split data
- Cross-validation

Summary of methods

Two types of methods

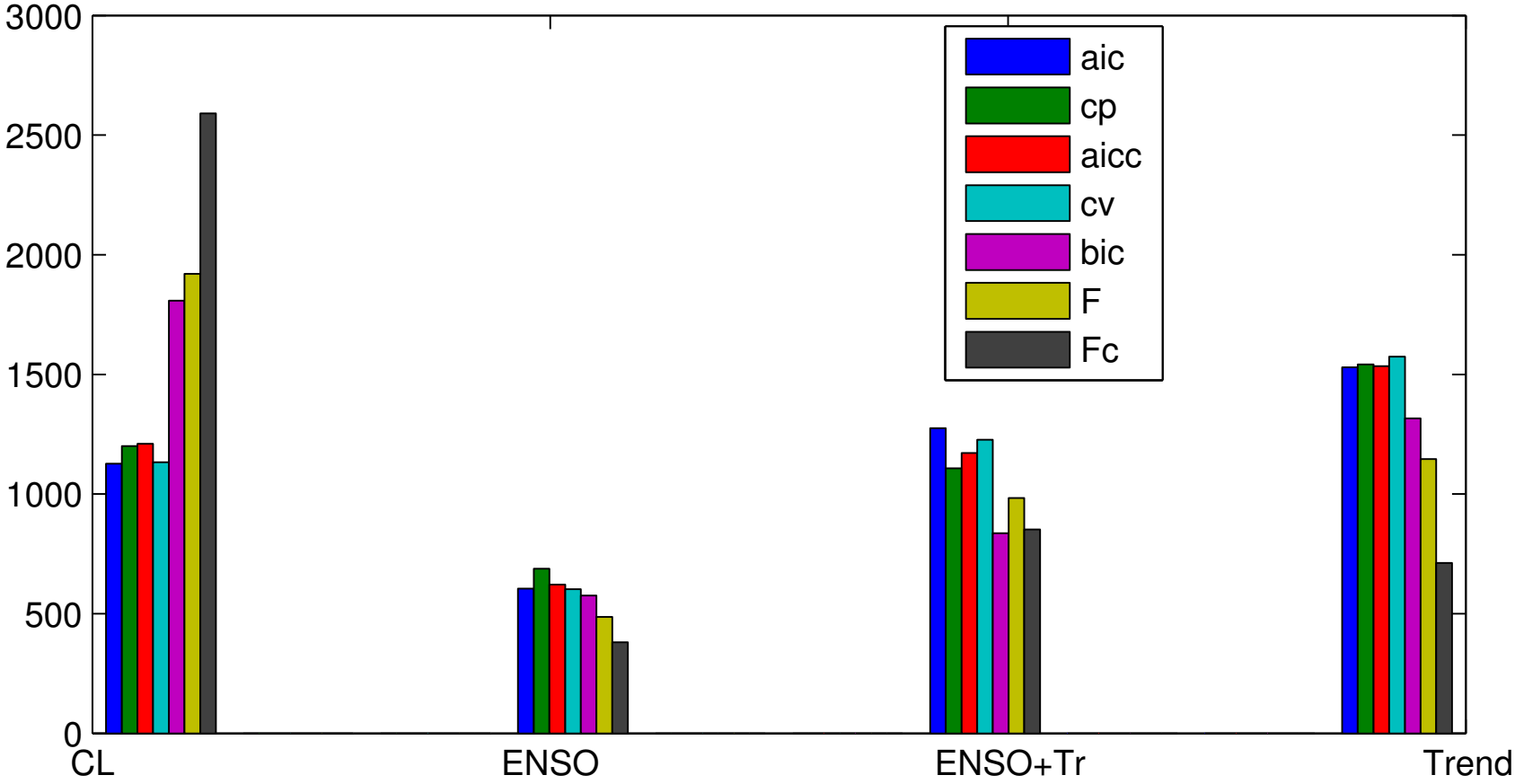
Balance between fit and number of predictors.

- F-test
- Mallow's C_P
- AIC (corrected), BIC

Apply model to independent data:

- Split data
- Cross-validation

Frequencies of the models selected



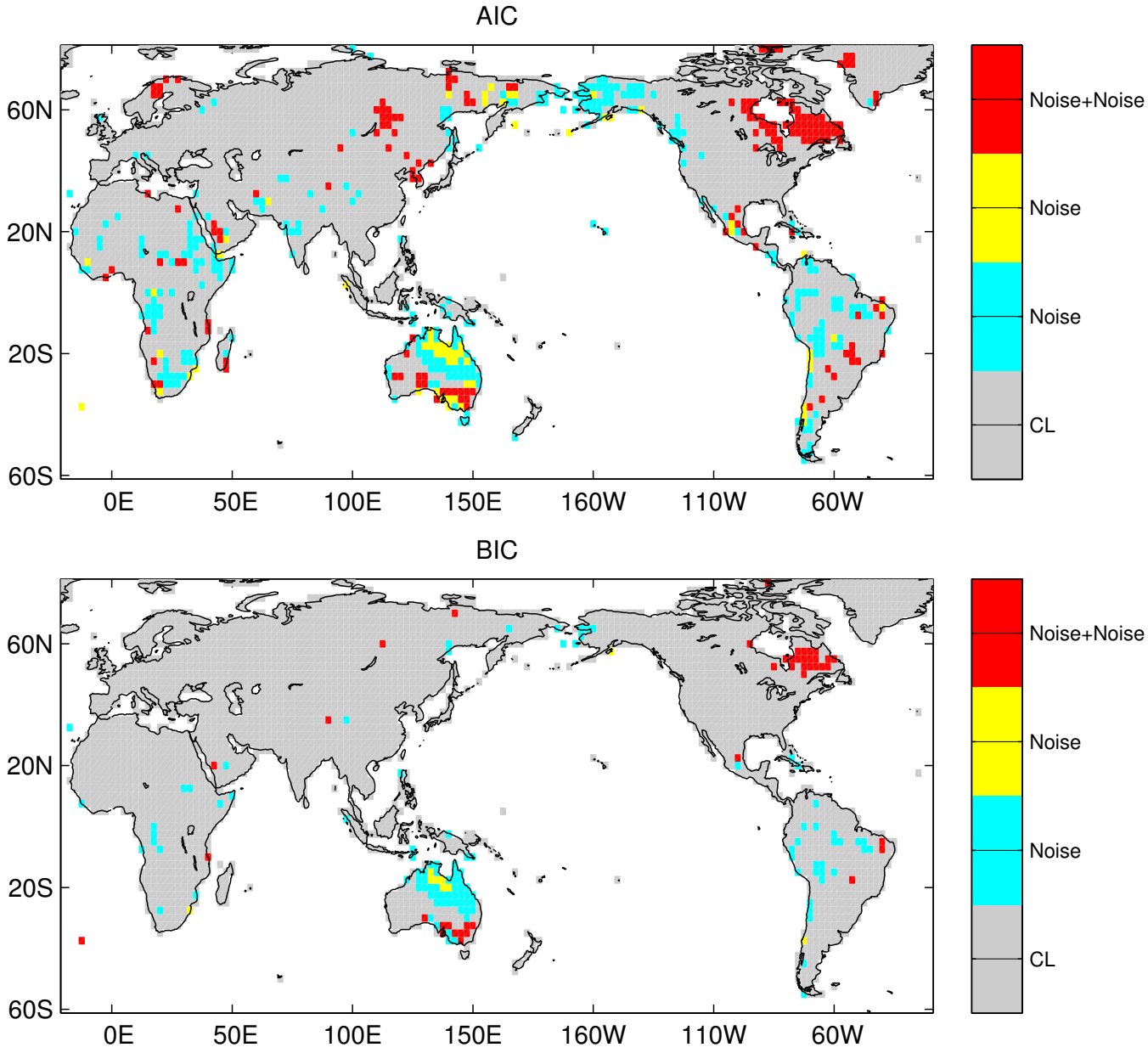
- AIC, AICc, C_p and cross-validation agree at 90% of the gridpoints.
- BIC and F-test agree in 93% of the gridpoints.
- F-test “corrected” for multiple comparisons is very strict.

How effective are the methods?

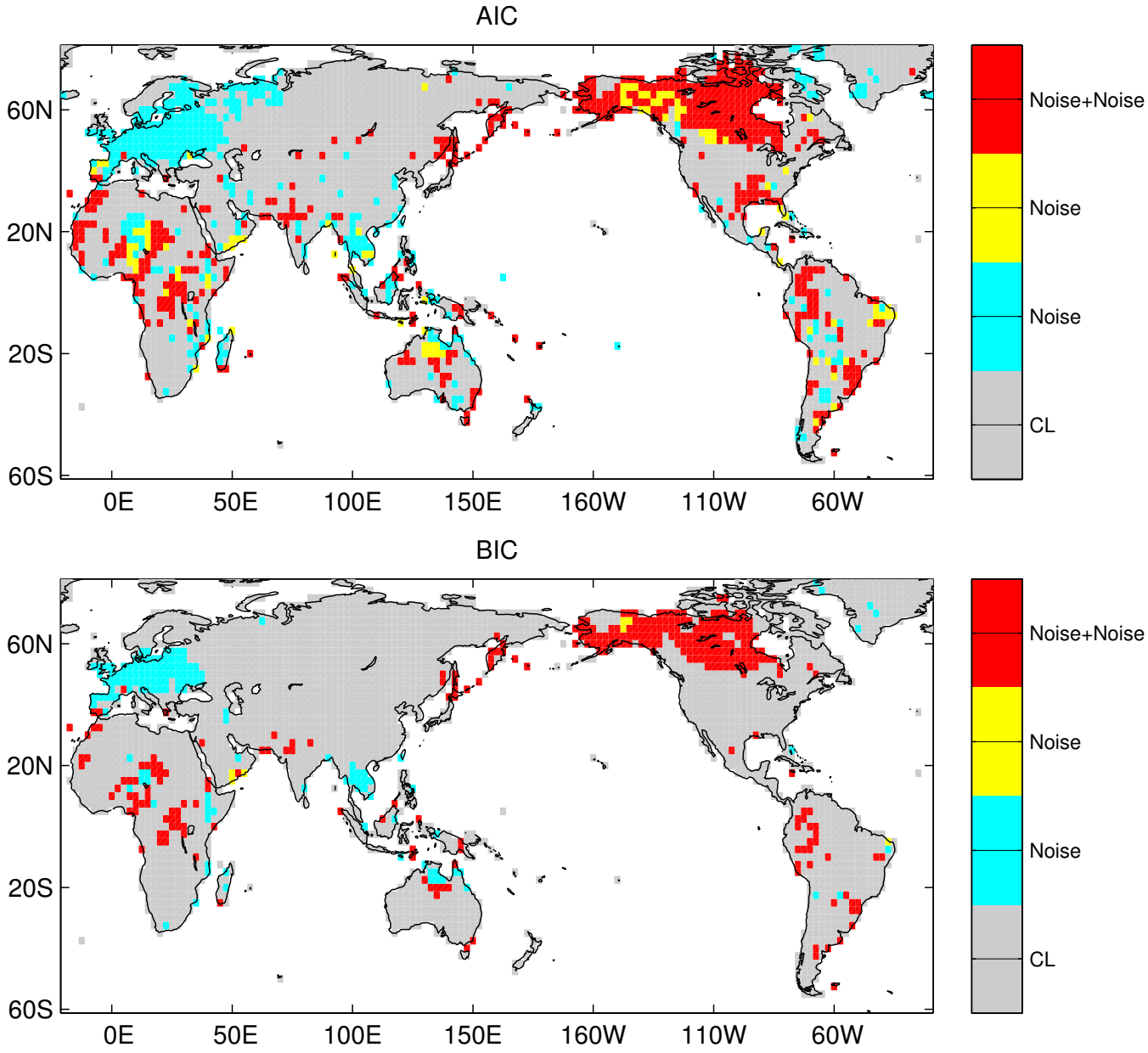
Apply them to models with random predictors.

Perforamnce across methods is more similar than different.

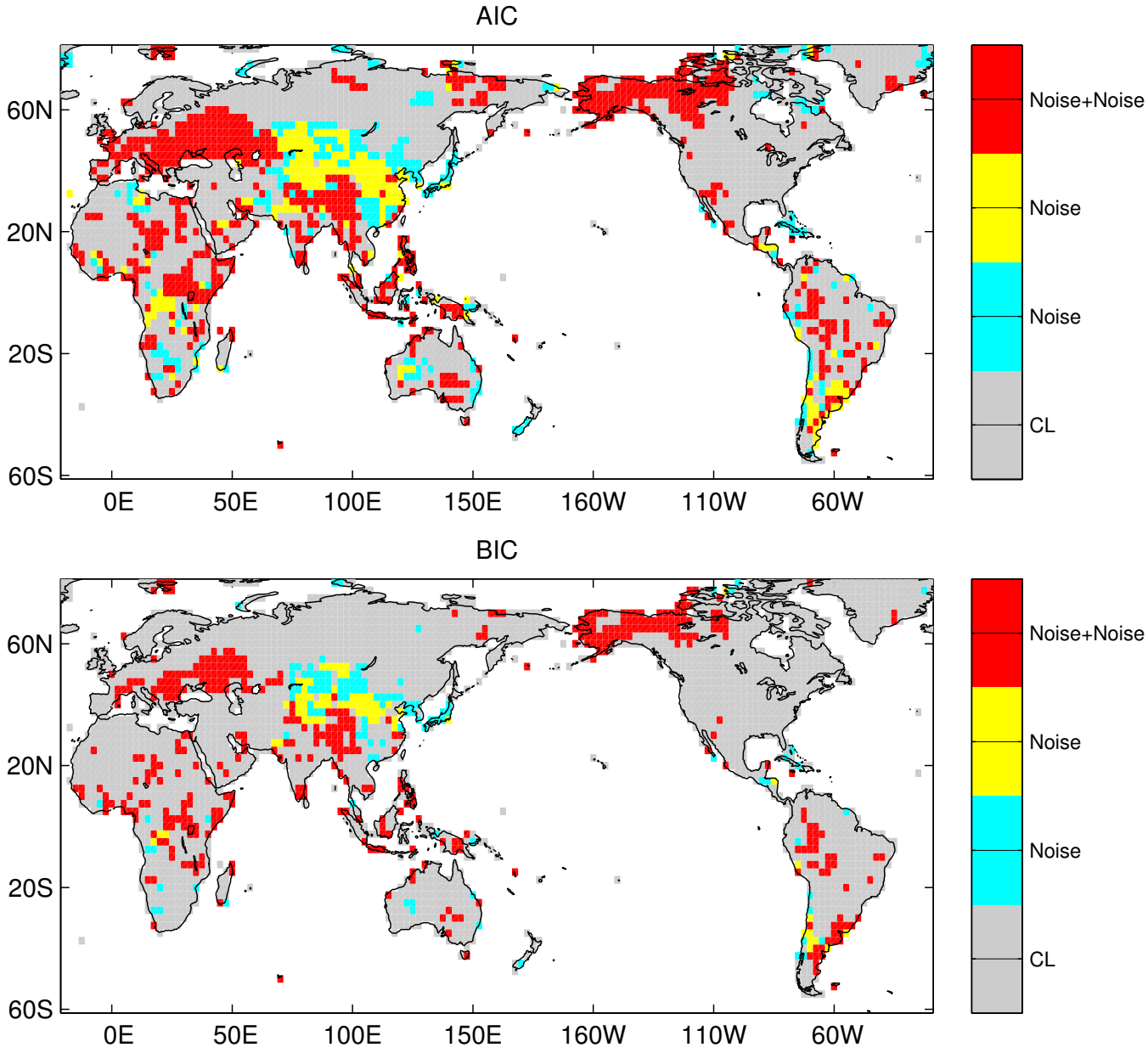
Example: DJF temperature



Example: DJF temperature



Example: DJF temperature



Moral

- Many predictor selection methods.
- All can be fooled given enough chances.