Start with a CPT example.

`http://iri.columbia.edu/ tippett/CPT/`

Two files:

Y: `SSTa.tsv`
X: `station.tsv`

Save somewhere. Open CPT and read them.

Run CPT. Describe the results.

# Pitfalls of statisical forecasting: Screening and Predictor selection

Michael K. Tippett

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

Center of Excellence for Climate Change Research, Department of Meteorology,
King Abdulaziz University, Jeddah, Saudi Arabia

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*

# Question: What makes a forecast useful?

Skill.

How can we tell the difference between skill and luck? (Why?)

Significance testing.

# Question: What makes a forecast useful?

Skill.

How can we tell the difference between skill and luck? (Why?)

Significance testing.

# Question: What makes a forecast useful?

Skill.

How can we tell the difference between skill and luck? (Why?)

Significance testing.

# Question: What makes a forecast useful?

Skill.

How can we tell the difference between skill and luck? (Why?)

Significance testing.

# Significance testing

Could the observed skill of a forecast occurred by **chance**?

Chance = variations particular to the sample

- Forecast really has more skill.
- Forecast really has less skill.
- Forecast really has no skill.

First two are not clearly defined.
Third is used in significance testing.

# Significance testing

How likely is the observed skill to have come from a forecast model with no skill?

Need to know the distribution of the skill of a no-skill model.

- Analytic methods – F-test, t-test.
- Monte Carlo – simulate no-skill forecasts and compute skill

# Significance testing: Example

Correlation. 30 years of forecasts. $n = 30$

$r = 0.3$
Significant?
Analytic solution:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

No-skill model $t$ has a $t$-distribution with $(n-2)$ dof.

Prob($r \geq 0.3$ | no-skill) = Prob ($t \geq 1.66$ | no-skill) = 5.4%

Good. But does not pass at “95% level”

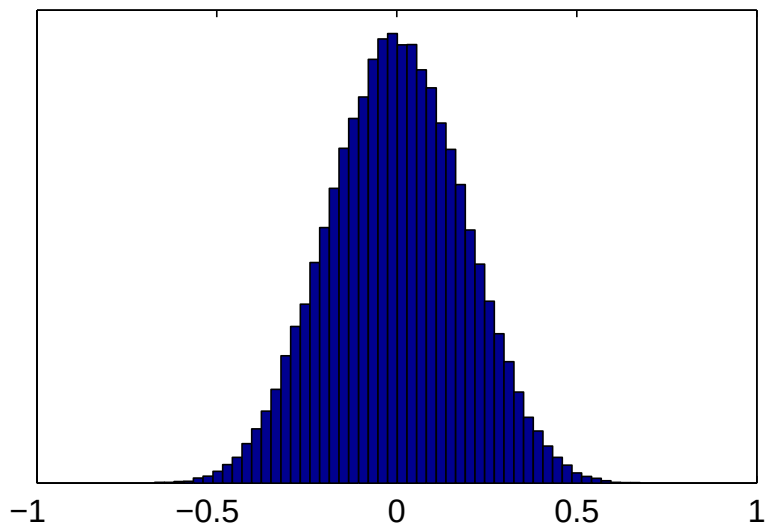
**Assumes that variables are Gaussian.**

# Significance testing: Example

Correlation.
30 years of forecasts. $n = 30$
Correlation: $r = 0.3$ Significant?
Monte Carlo

```
m=10000; n=30;
F = randn(m,n);
O = randn(m,n);
c = correlation(O,F);
hist(c)
mean(c>0.3)
ans = 0.0538

prctile(c,95)
ans = 0.3062
```




Histogram of the 30-year sample skill of “no-skill” forecasts.

# Screening

Process of selecting variables because of their strong correlation with target variable.

Sounds like looking for good predictors.

Problems arise when many possible candidate variables are screened.

Can changes the no-skill distribution dramatically.

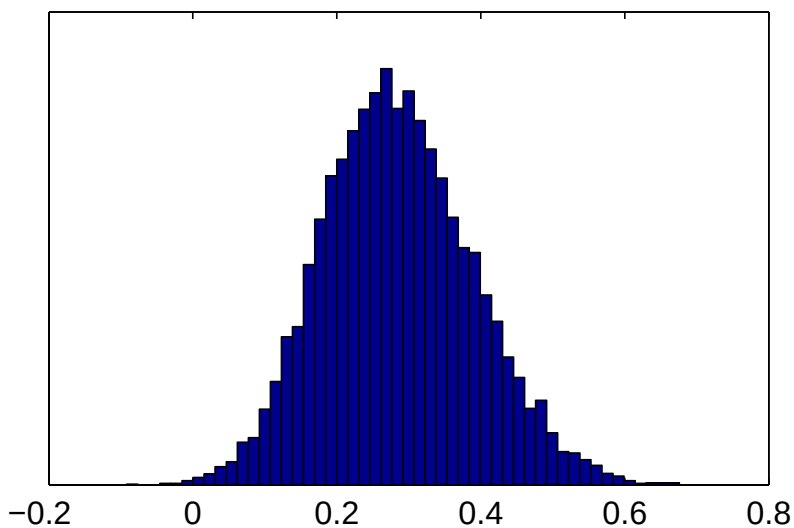
# Screening: Example 1

Suppose we look at many forecasts and choose the best one.
How does that change the significance test?

Analytical answer. Not standard.
Monte Carlo. Easy!

```
k=10;
c_screen = reshape(c,k,m/k);
cbest = max(c_screen);
hist(cbest,100)
mean(cbest>0.3)
ans = 0.4237

prctile(cbest,95)
ans = 0.4486
```



# Screening: Example 2

If the likelihood of a no-skill forecast appearing significant (by chance) is $p$,

then the likelihood of at least one of $k$ independent forecasts appearing significant (by chance) is

$$1 - \text{Prob(none sig.)} = 1 - (1 - p)^k \approx kp$$

k = 10, p = 5%
5% $\rightarrow$ 40%

# Screening implications

Screening, looking at many forecasts and just reporting the best, can make a non-skill forecast appear skillful.

Correct significance testing of procedures with excessive screening, can lead to skillful forecasts appearing insignificant.

# Screening and predictor selection

# Predictor selection

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.

# Predictor selection

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.

# Predictor selection

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.

# Cross validation

Cross-validation gives a skill estimates on **independent** data. Independent of data used to estimate the model parameters.

Leave-$k$-out cross-validation:

- Leave out $k$ consecutive years.
- Estimate the statistical model on the remaining years.
- Predict the middle of the $k$ years ($k$ odd).
- Repeat until all years predicted.

Leave-1-out

- Estimate model from years 2-$N$.
- “Predict” year 1.
- And so on

# Pitfalls of cross validation

- Performing an initial analysis using the entire data set to identify the most informative features. (Screening)
- Using cross-validation to assess several models, and only stating the results for the model with the best results. (Selection bias)
- Allowing some of the training data to be (essentially) included in the test set. (Cheating)

From wikipedia

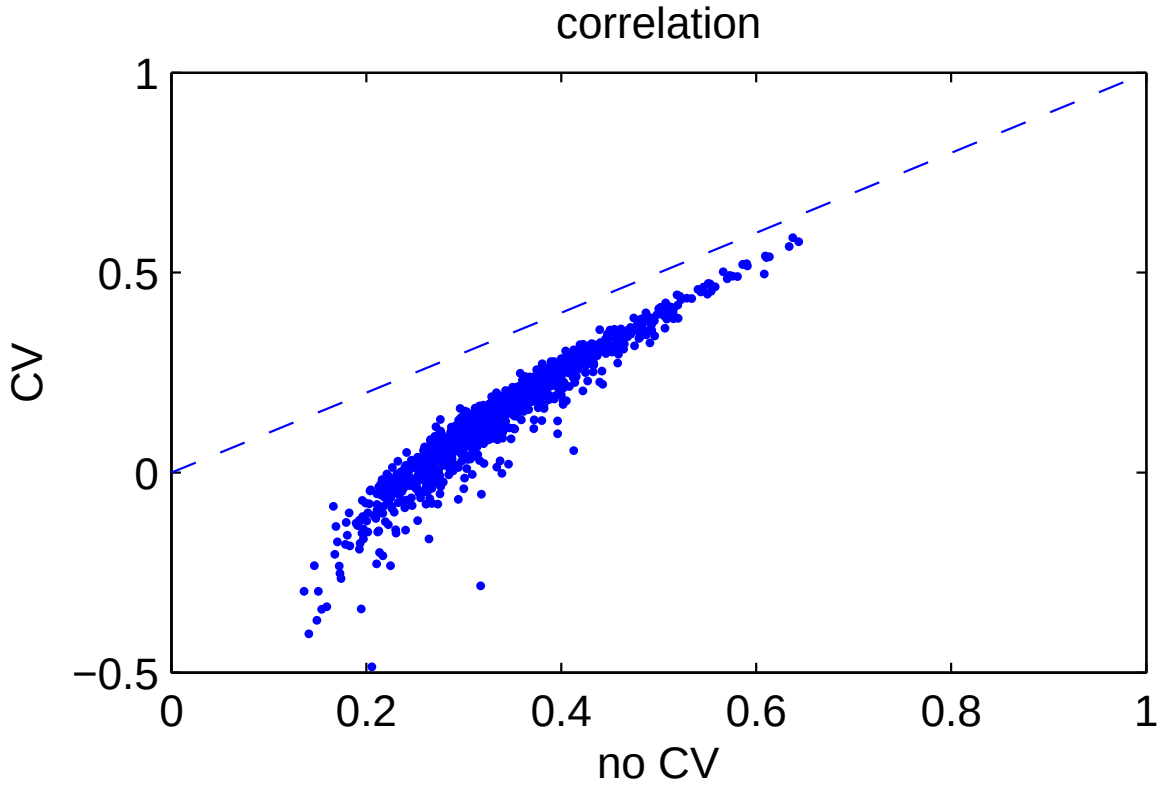
# Screening and cross-validation

If the predictors are chosen using the entire data set, the cross-validated skill will be larger than in an independent data.

# Screening and cross-validation: Example 1

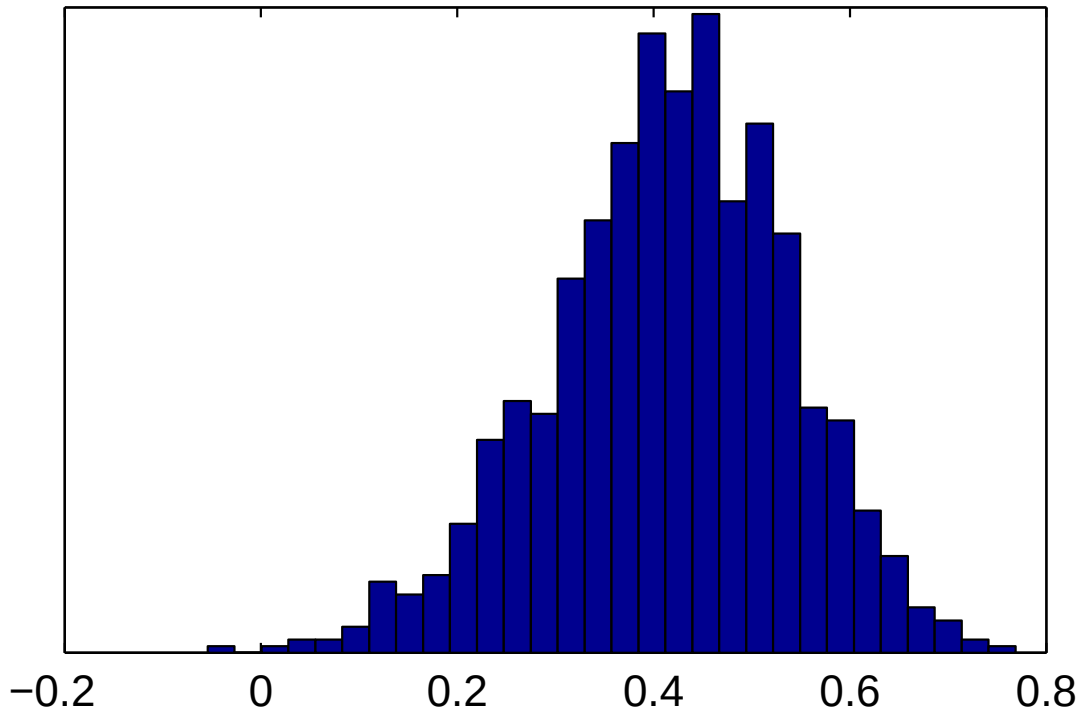
The problem is that screening finds predictors with high correlation and cross-validation only slightly reduces it

Single predictor problem.



# Screening and cross-validation: Example 2

Suppose you have a 40-member ensemble and you pick the 3 members with best correlation as predictors.
(Why is this wrong?)
And then compute the cross-validated correlation.




Zero skill distribution

# Screening and cross-validation: Example 3

- Observe that in a 40-member ensemble of GCM predictions some members have more skill than others.
- Pick the members with skill exceeding some threshold.
- Perform PCA on those members and retain the PCs with skill exceeding some threshold as your predictors.
- Estimate skill using cross-validation.

Sounds harmless, maybe even clever.

What is the problem?

What is the impact?

# Screening and cross-validation: Example 3

- Observe that in a 40-member ensemble of GCM predictions some members have more skill than others.
- Pick the members with skill exceeding some threshold.
- Perform PCA on those members and retain the PCs with skill exceeding some threshold as your predictors.
- Estimate skill using cross-validation.

Sounds harmless, maybe even clever.

What is the problem?

What is the impact?

# Screening and cross-validation: Example 3

- Observe that in a 40-member ensemble of GCM predictions some members have more skill than others.
- Pick the members with skill exceeding some threshold.
- Perform PCA on those members and retain the PCs with skill exceeding some threshold as your predictors.
- Estimate skill using cross-validation.

Sounds harmless, maybe even clever.

What is the problem?

What is the impact?

# Screening and cross-validation: Example 3

- Observe that in a 40-member ensemble of GCM predictions some members have more skill than others.
- Pick the members with skill exceeding some threshold.
- Perform PCA on those members and retain the PCs with skill exceeding some threshold as your predictors.
- Estimate skill using cross-validation.

Sounds harmless, maybe even clever.

What is the problem?

What is the impact?

# Screening and cross-validation: Example 3

- Observe that in a 40-member ensemble of GCM predictions some members have more skill than others.
- Pick the members with skill exceeding some threshold.
- Perform PCA on those members and retain the PCs with skill exceeding some threshold as your predictors.
- Estimate skill using cross-validation.

Sounds harmless, maybe even clever.

What is the problem?

What is the impact?

# Screening and cross-validation: Example 3

- Observe that in a 40-member ensemble of GCM predictions some members have more skill than others.
- Pick the members with skill exceeding some threshold.
- Perform PCA on those members and retain the PCs with skill exceeding some threshold as your predictors.
- Estimate skill using cross-validation.

Sounds harmless, maybe even clever.

What is the problem?

What is the impact?

# Screening and cross-validation: Example 3

- Observe that in a 40-member ensemble of GCM predictions some members have more skill than others.
- Pick the members with skill exceeding some threshold.
- Perform PCA on those members and retain the PCs with skill exceeding some threshold as your predictors.
- Estimate skill using cross-validation.

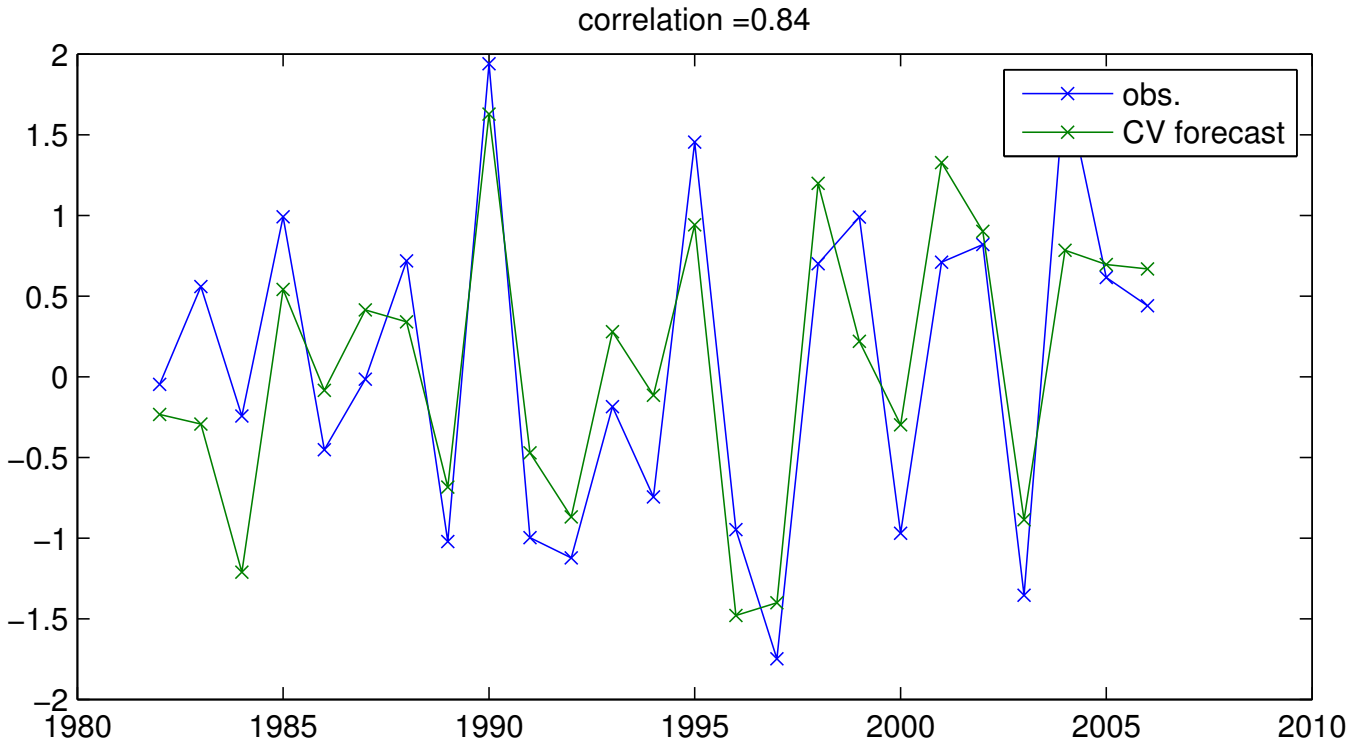
Sounds harmless, maybe even clever.

What is the problem?

What is the impact?

# Screening and cross-validation: Example 3

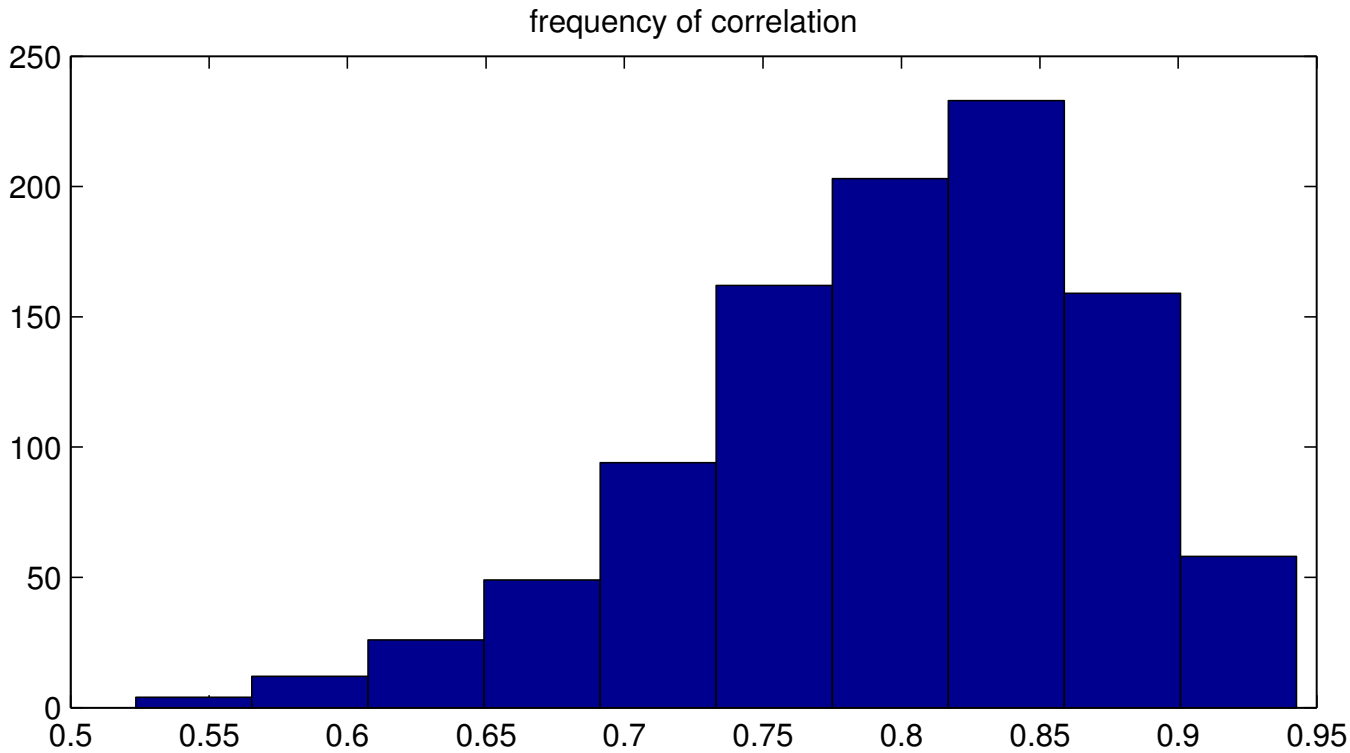
Cross-validated forecasts show good skill.




What is the real skill?

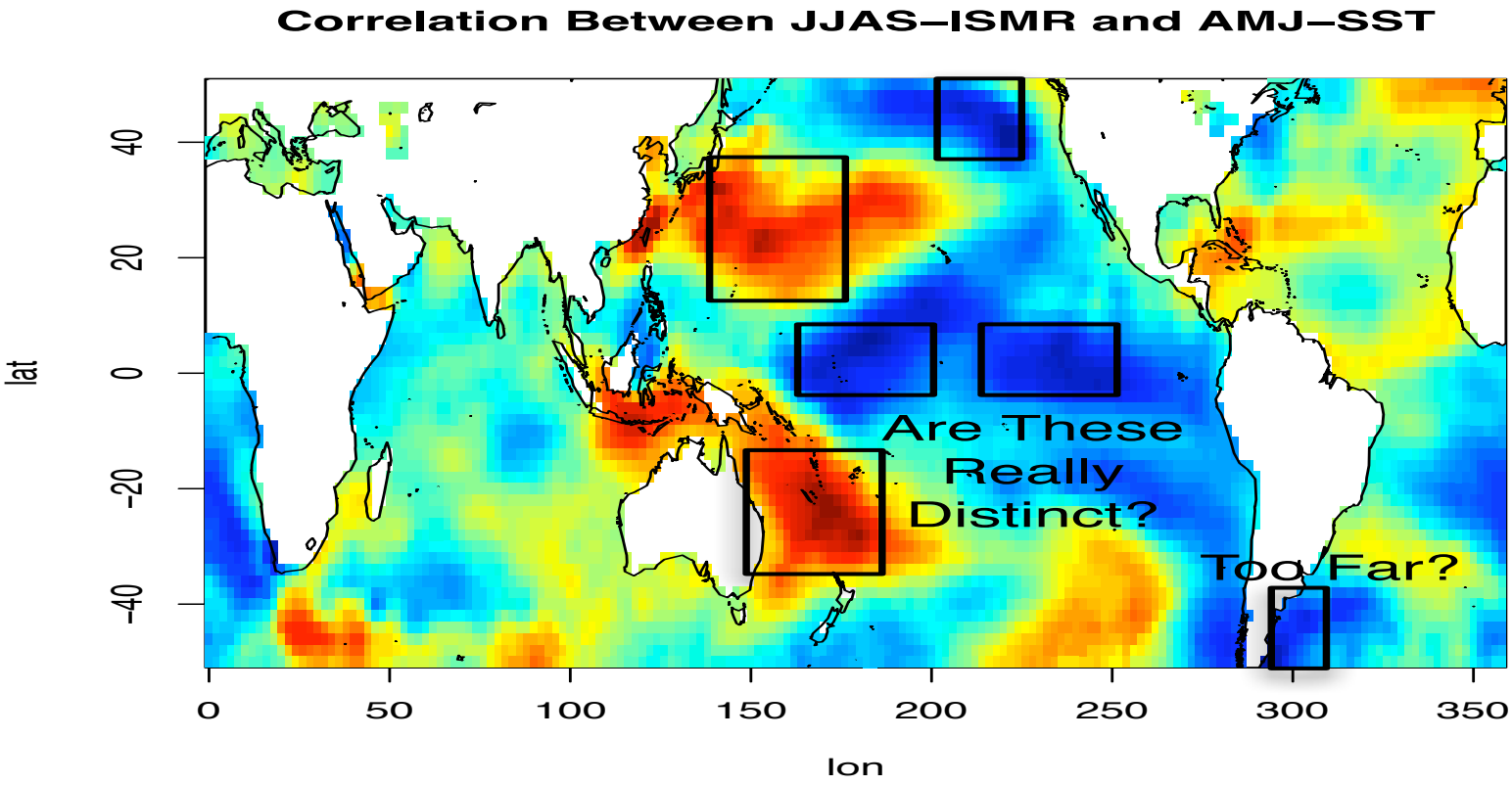
# Screening and cross-validation: Example 3

Apply this procedure 1000 times to random numbers




mean correlation = 0.8

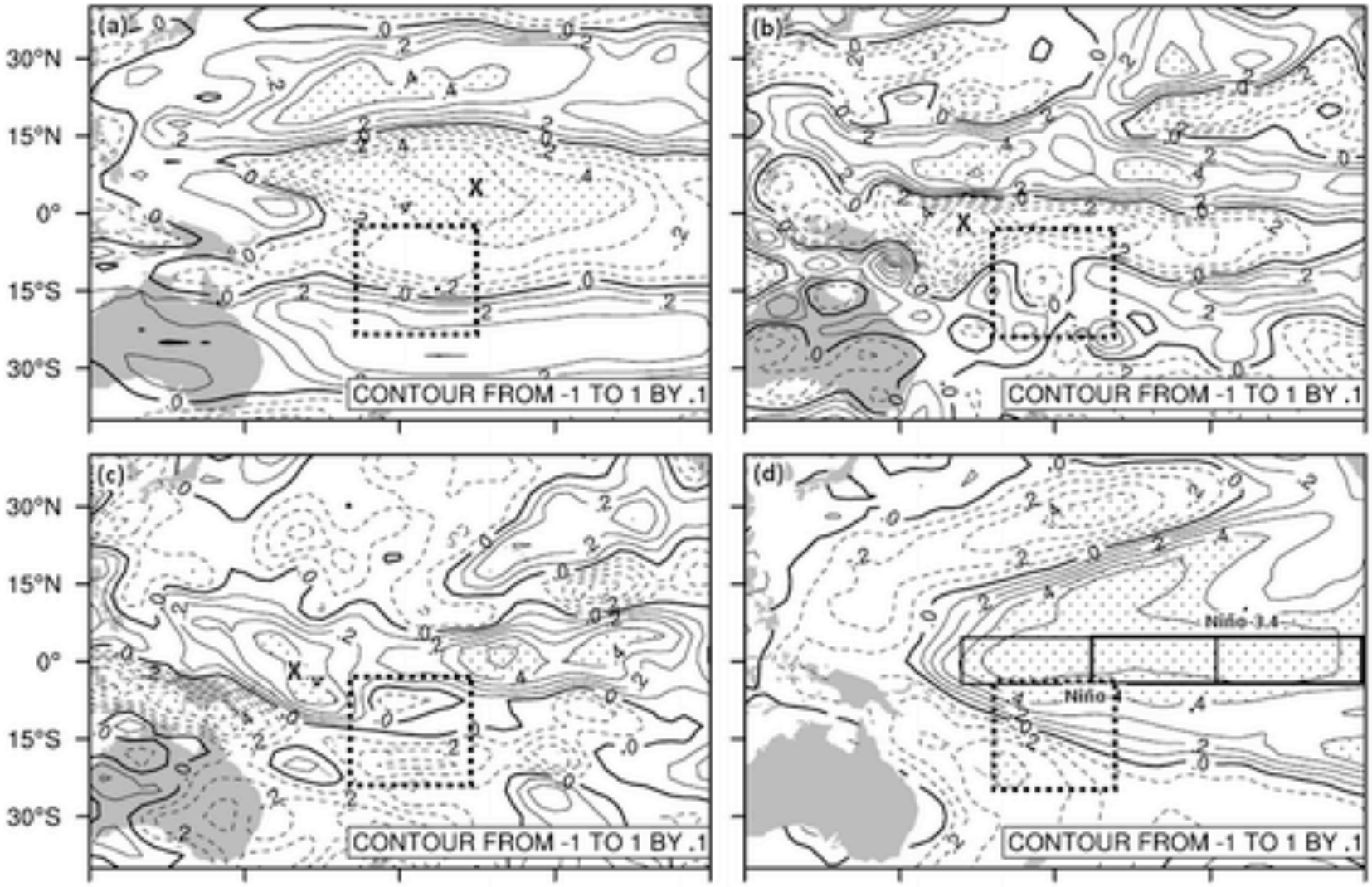
# Correlation maps and boxes



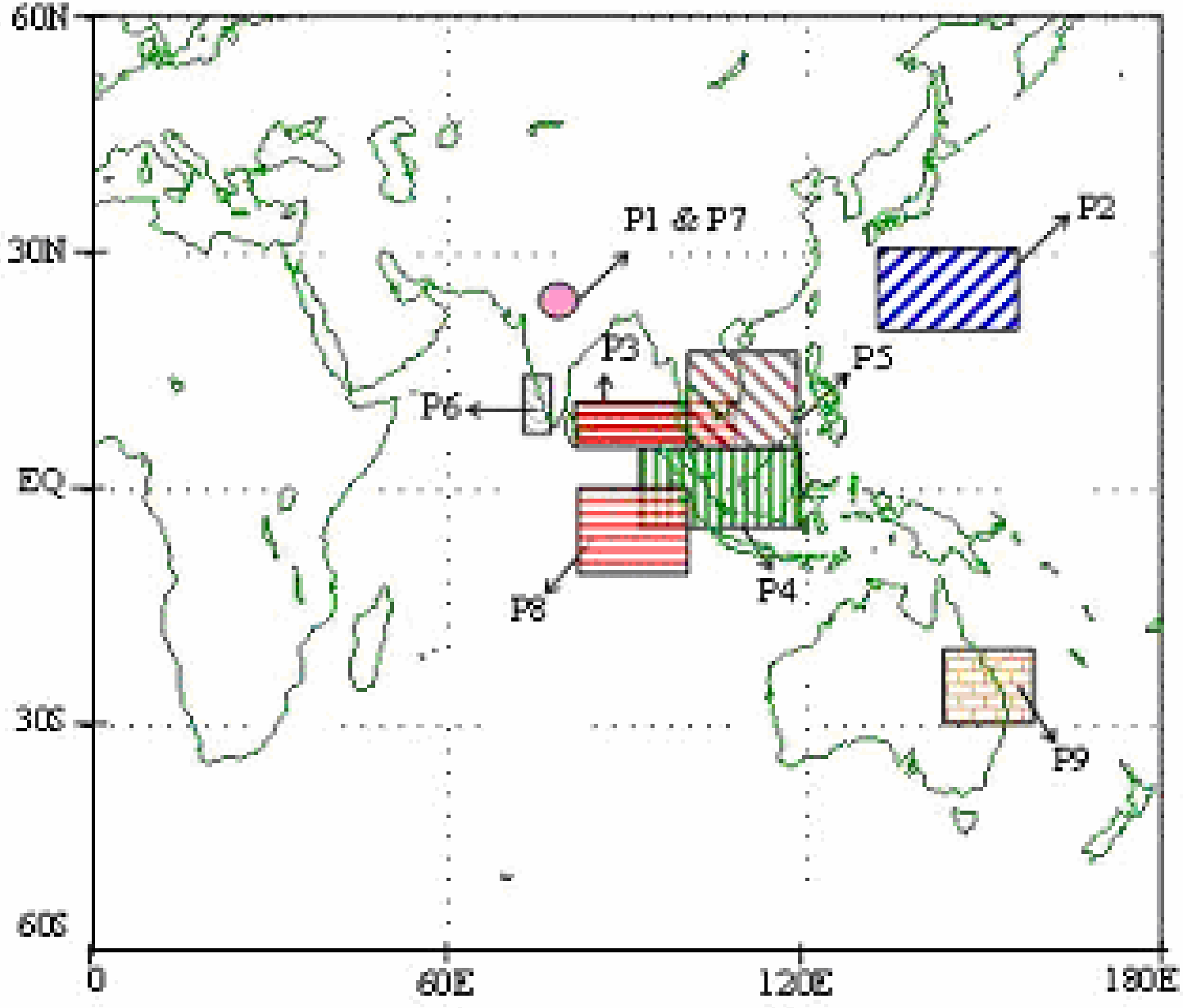

Common screening method

- Correlate time series with field (SST or . . . )
- Draw boxes around regions with significant correlations.
- Average over boxes
- Screen predictors

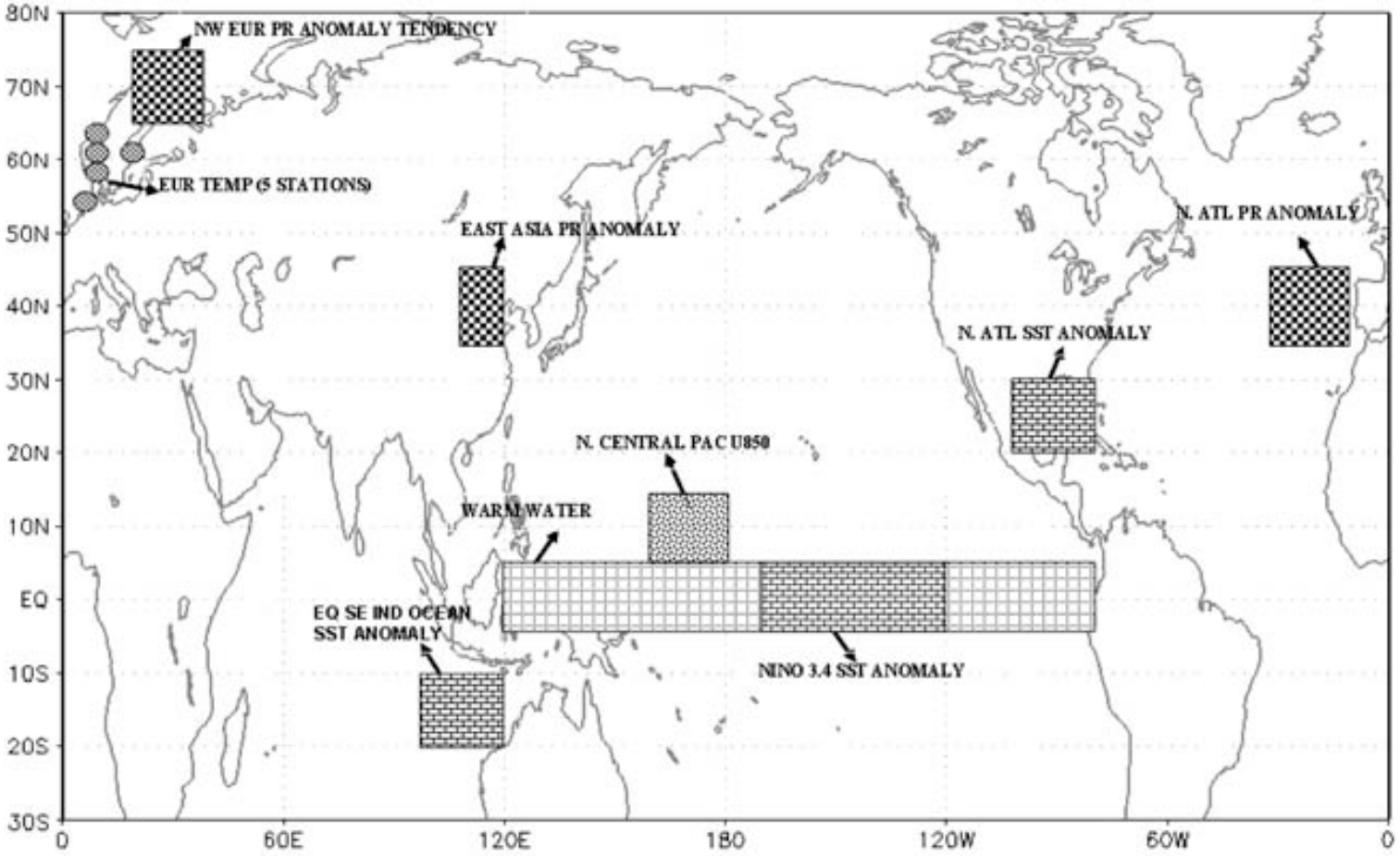
# Correlation maps and boxes



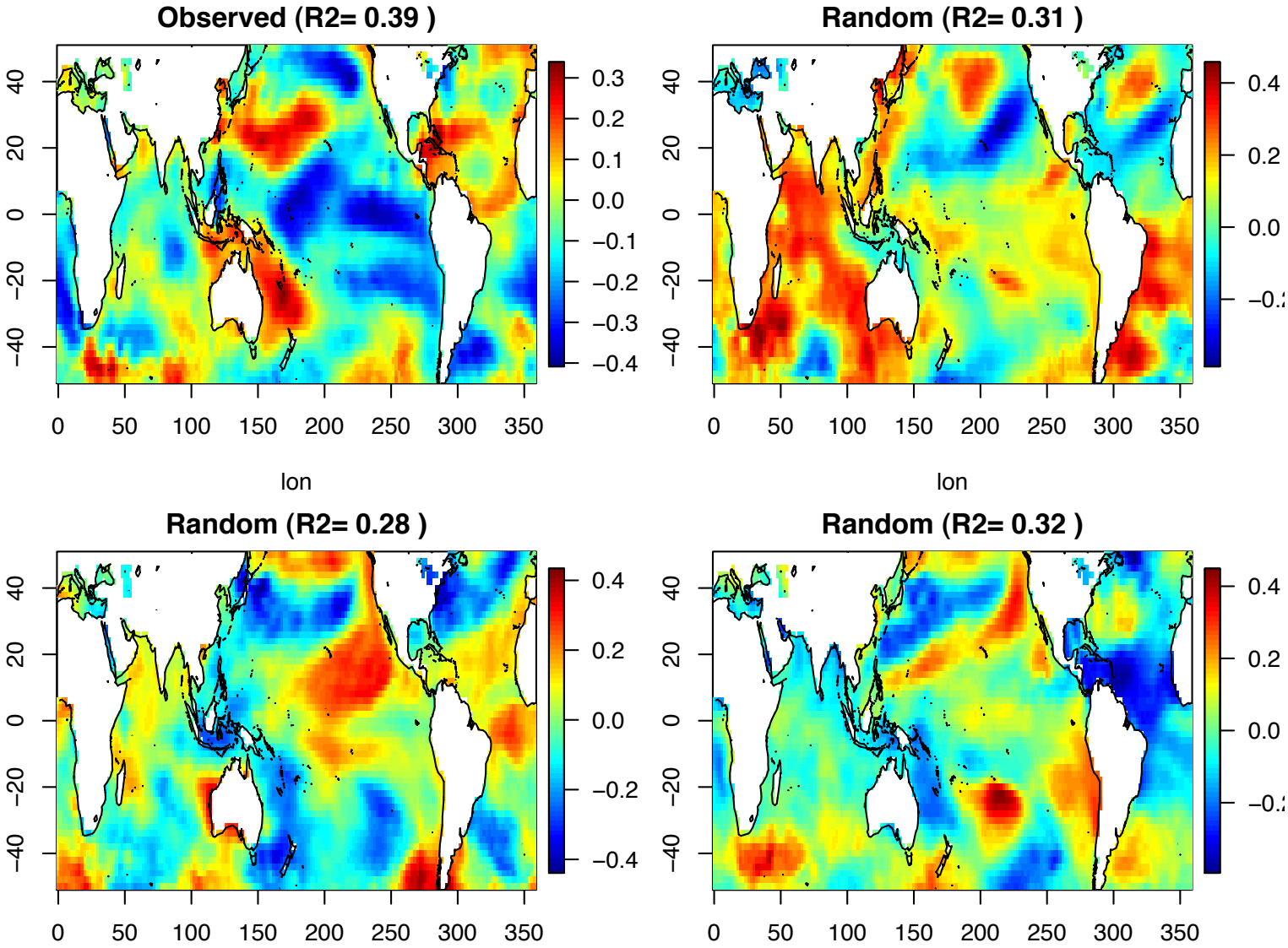
# Correlation maps and boxes



# Correlation maps and boxes



# Correlation maps and boxes



# Selection bias

*Using cross-validation to assess several models, and only stating the results for the model with the best results.*

CPT

- Optimizing number of EOFs
- Optimizing EOF domain

# Summary

- Screening leads to overestimates of skill.
  - Poor skill in operation.
- Screening invalidates standard significance tests as well as cross-validation.
- The degree of overestimation can be minor or large.
- Selection bias leads to overestimates of skill.

# Recommendations

- Don't look too hard for "good" predictors
- Good to have a physical explanation. But not enough.
- Don't use correlation maps and boxes.
- Use model data.
  - Pick predictors based on relations in models.
  - Does not work if model is poor.