Pitfalls of cross validation

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

Main ideas

- Cross-validation is useful for estimating the performance of a model on independent data.
- Few assumptions.
- Computationally expensive.
- Can be misused.

Outline

- Cross validation and linear regression
  - Computational method
  - Bias
- Pitfalls
  - Not including model/predictor selection in the cross-validation.
  - Not leaving out enought data so that the training and validation samples are independent.

Cross-validation is expensive.

Cross-validation of linear regression

Cross-validated error can be computed without the computational cost of cross-validation.

$$y(i) - \hat{y}_{cv}(i) = \frac{e(i)}{1 - v_{ii}}$$

where

- $\hat{y}_{cv}(i)$ is the prediction from the regression with the i-th sample left out of the compuation of the regression coefficients .
- $e(i) = y(i) - \hat{y}(i)$ in-sample error
- v_{ii} is the i-th diagonal of the "hat"-matrix $V = X(X^T X)^{-1} X^T$.

(Cook & Weisberg, *Residuals and Influence in Regression*, 1980)

Cross-validation for linear regression

What is the correlation of a climatological forecast?

What is the cross-validated correlation of a climatological forecast?

- X = column of n ones.
- $X^T X = n$, $(X^T X)^{-1} = 1/n$,
- $V = X(X^T X)^{-1} X^T = n \times n$ matrix with values $1/n$.
- $e(i) = y(i) - \overline{y}$.

$$\hat{y}_{cv}(i) = y(i) - \frac{e(i)}{1 - v_{ii}} = y(i) - \frac{y(i) - \overline{y}}{1 - 1/n}$$

$$= \overline{y}\frac{n}{n-1} - y\frac{1}{n-1}$$

What is the correlation between y and $\hat{y}_{cv}$? Why?

Cross-validation for linear regression

What is the correlation of a climatological forecast?
What is the cross-validated correlation of a climatological forecast?

- X = column of n ones.
- $X^T X = n$, $(X^T X)^{-1} = 1/n$,
- $V = X(X^T X)^{-1} X^T = n \times n$ matrix with values $1/n$.
- $e(i) = y(i) - \overline{y}$.

$$\hat{y}_{cv}(i) = y(i) - \frac{e(i)}{1 - v_{ii}} = y(i) - \frac{y(i) - \overline{y}}{1 - 1/n}$$

$$= \overline{y}\frac{n}{n-1} - y\frac{1}{n-1}$$

What is the correlation between y and $\hat{y}_{cv}$? Why?

Cross-validation for linear regression

What is the correlation of a climatological forecast?
What is the cross-validated correlation of a climatological forecast?

- X = column of n ones.
- $X^T X = n$, $(X^T X)^{-1} = 1/n$,
- $V = X(X^T X)^{-1} X^T = n \times n$ matrix with values $1/n$.
- $e(i) = y(i) - \overline{y}$.

$$\hat{y}_{cv}(i) = y(i) - \frac{e(i)}{1 - v_{ii}} = y(i) - \frac{y(i) - \overline{y}}{1 - 1/n}$$
$$= \overline{y}\frac{n}{n-1} - y\frac{1}{n-1}$$

What is the correlation between y and $\hat{y}_{cv}$? Why?

Cross-validation for linear regression

What is the correlation of a climatological forecast?
What is the cross-validated correlation of a climatological forecast?

- X = column of n ones.
- $X^T X = n$, $(X^T X)^{-1} = 1/n$,
- $V = X(X^T X)^{-1} X^T = n \times n$ matrix with values $1/n$.
- $e(i) = y(i) - \overline{y}$.

$$\hat{y}_{cv}(i) = y(i) - \frac{e(i)}{1 - v_{ii}} = y(i) - \frac{y(i) - \overline{y}}{1 - 1/n}$$
$$= \overline{y}\frac{n}{n-1} - y\frac{1}{n-1}$$

What is the correlation between y and $\hat{y}_{cv}$? Why?

Cross-validation for linear regression

What is the correlation of a climatological forecast?
What is the cross-validated correlation of a climatological forecast?

- X = column of n ones.
- $X^T X = n$, $(X^T X)^{-1} = 1/n$,
- $V = X(X^T X)^{-1} X^T = n \times n$ matrix with values $1/n$.
- $e(i) = y(i) - \overline{y}$.

$$\hat{y}_{cv}(i) = y(i) - \frac{e(i)}{1 - v_{ii}} = y(i) - \frac{y(i) - \overline{y}}{1 - 1/n}$$
$$= \overline{y}\frac{n}{n-1} - y\frac{1}{n-1}$$

What is the correlation between y and $\hat{y}_{cv}$? Why?

Cross-validation for linear regression

What is the correlation of a climatological forecast?
What is the cross-validated correlation of a climatological forecast?

- X = column of n ones.
- $X^T X = n$, $(X^T X)^{-1} = 1/n$,
- $V = X(X^T X)^{-1} X^T = n \times n$ matrix with values $1/n$.
- $e(i) = y(i) - \overline{y}$.

$$\hat{y}_{cv}(i) = y(i) - \frac{e(i)}{1 - v_{ii}} = y(i) - \frac{y(i) - \overline{y}}{1 - 1/n}$$
$$= \overline{y}\frac{n}{n-1} - y\frac{1}{n-1}$$

What is the correlation between y and $\hat{y}_{cv}$? Why?

Cross-validation for linear regression

What is the correlation of a climatological forecast?
What is the cross-validated correlation of a climatological forecast?

- X = column of n ones.
- $X^T X = n$, $(X^T X)^{-1} = 1/n$,
- $V = X(X^T X)^{-1} X^T = n \times n$ matrix with values $1/n$.
- $e(i) = y(i) - \overline{y}$.

$$\hat{y}_{cv}(i) = y(i) - \frac{e(i)}{1 - v_{ii}} = y(i) - \frac{y(i) - \overline{y}}{1 - 1/n}$$
$$= \overline{y}\frac{n}{n-1} - y\frac{1}{n-1}$$

What is the correlation between y and $\hat{y}_{cv}$? Why?

Cross-validation for linear regression

More intuitively

- Leaving out a wet year, gives a drier mean
- Leaving out a drier year, gives a wetter mean.

Negative correlation.

Problem solved if somehow the mean did not change. How?

Cross-validation for linear regression

More intuitively

- Leaving out a wet year, gives a drier mean
- Leaving out a drier year, gives a wetter mean.

Negative correlation.

Problem solved if somehow the mean did not change. How?

Cross-validation for linear regression

More intuitively

- Leaving out a wet year, gives a drier mean
- Leaving out a drier year, gives a wetter mean.

Negative correlation.

Problem solved if somehow the mean did not change. How?

Cross-validation for linear regression

More intuitively

- Leaving out a wet year, gives a drier mean
- Leaving out a drier year, gives a wetter mean.

Negative correlation.

Problem solved if somehow the mean did not change. How?

Cross-validation for linear regression

More intuitively

- Leaving out a wet year, gives a drier mean
- Leaving out a drier year, gives a wetter mean.

Negative correlation.

Problem solved if somehow the mean did not change. How?

Cross-validation for linear regression

More intuitively

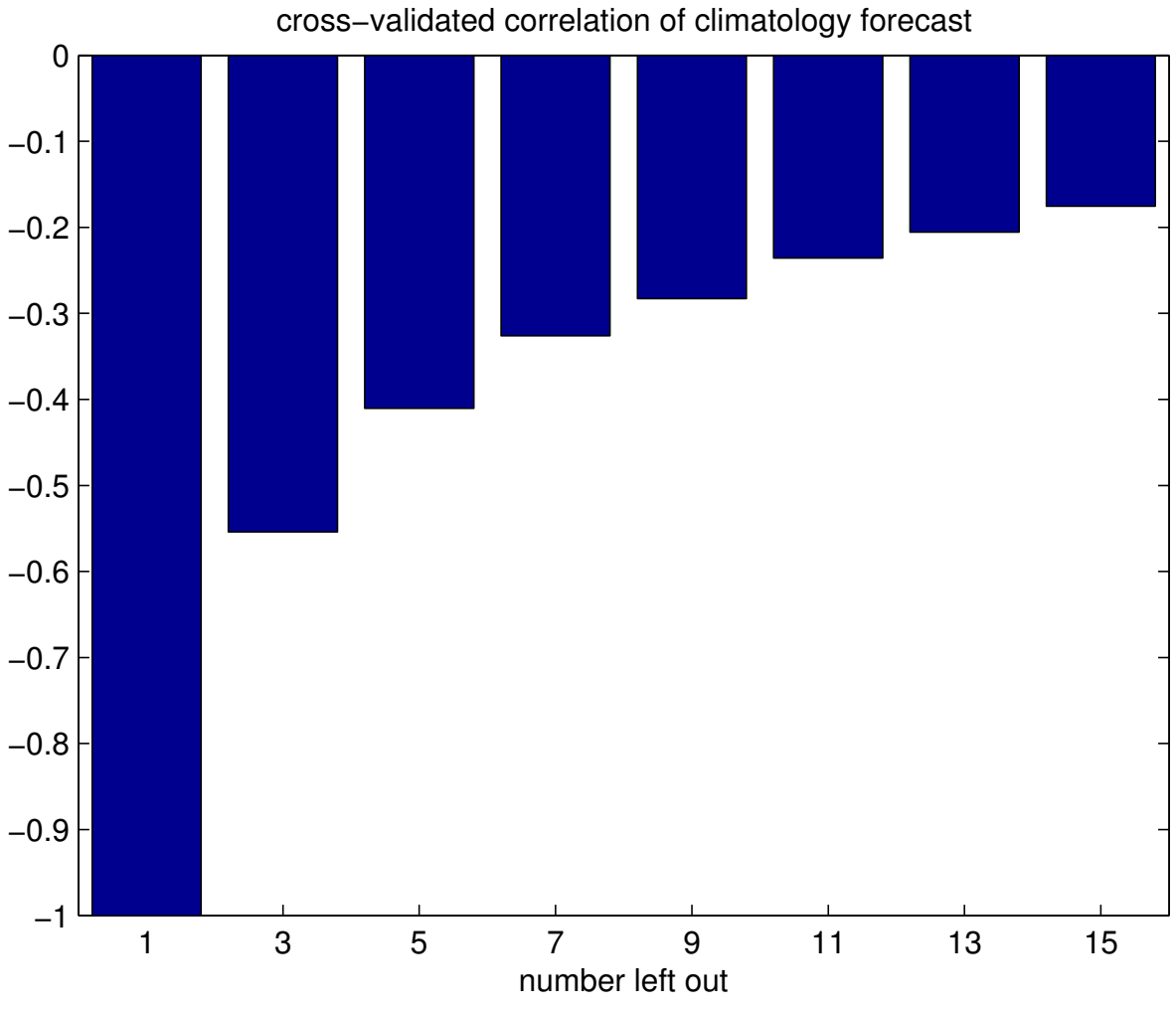
- Leaving out a wet year, gives a drier mean
- Leaving out a drier year, gives a wetter mean.

Negative correlation.

Problem solved if somehow the mean did not change. How?

Cross-validation for linear regression

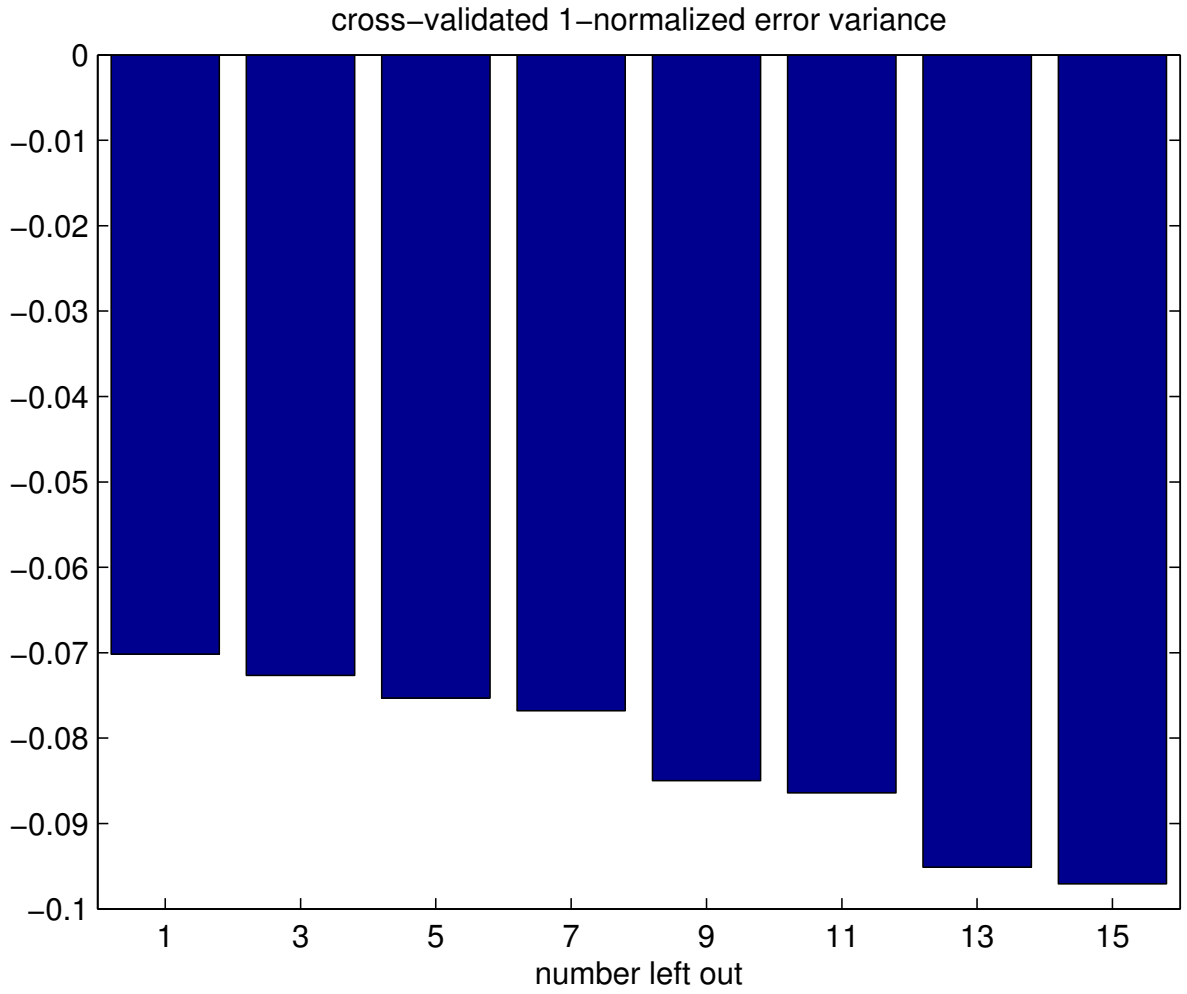
Some benefit to leaving out more years and predicting middle year. ($n = 30$)



Problem with trend?

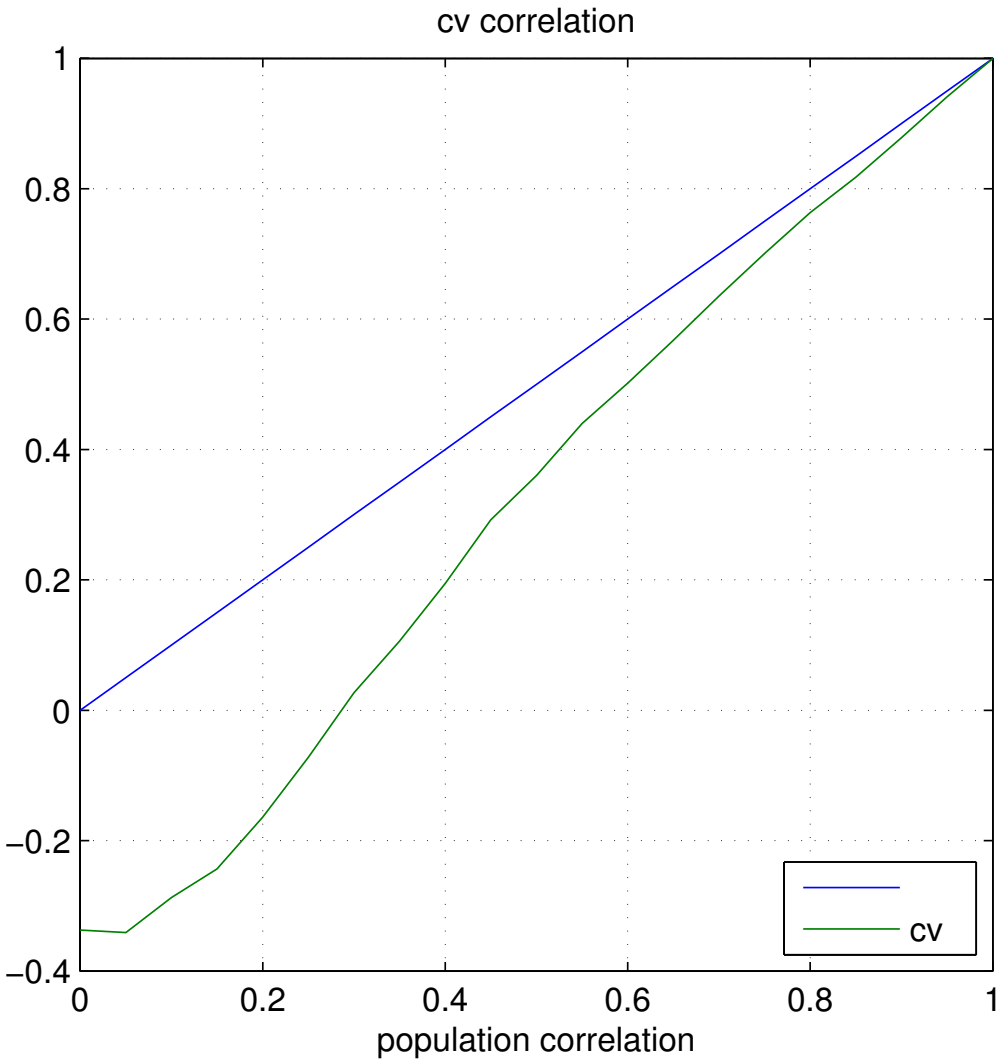
Cross-validation for linear regression

Some benefit to leaving out more years and predicting middle year but increase in error variance. ($n = 30$)



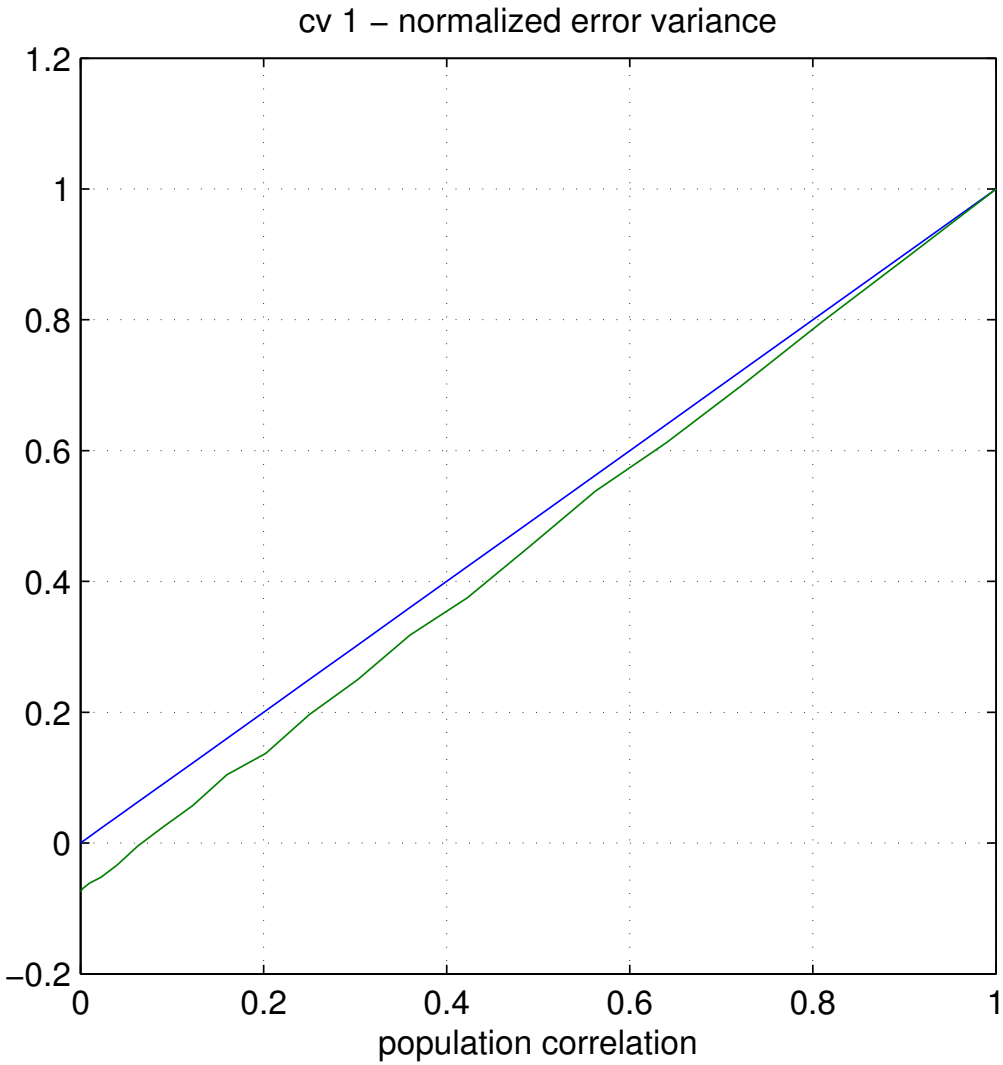
Bias of cross validation

n=30. $y = ax + b$.



Bias of cross validation

n=30. $y = ax + b$.



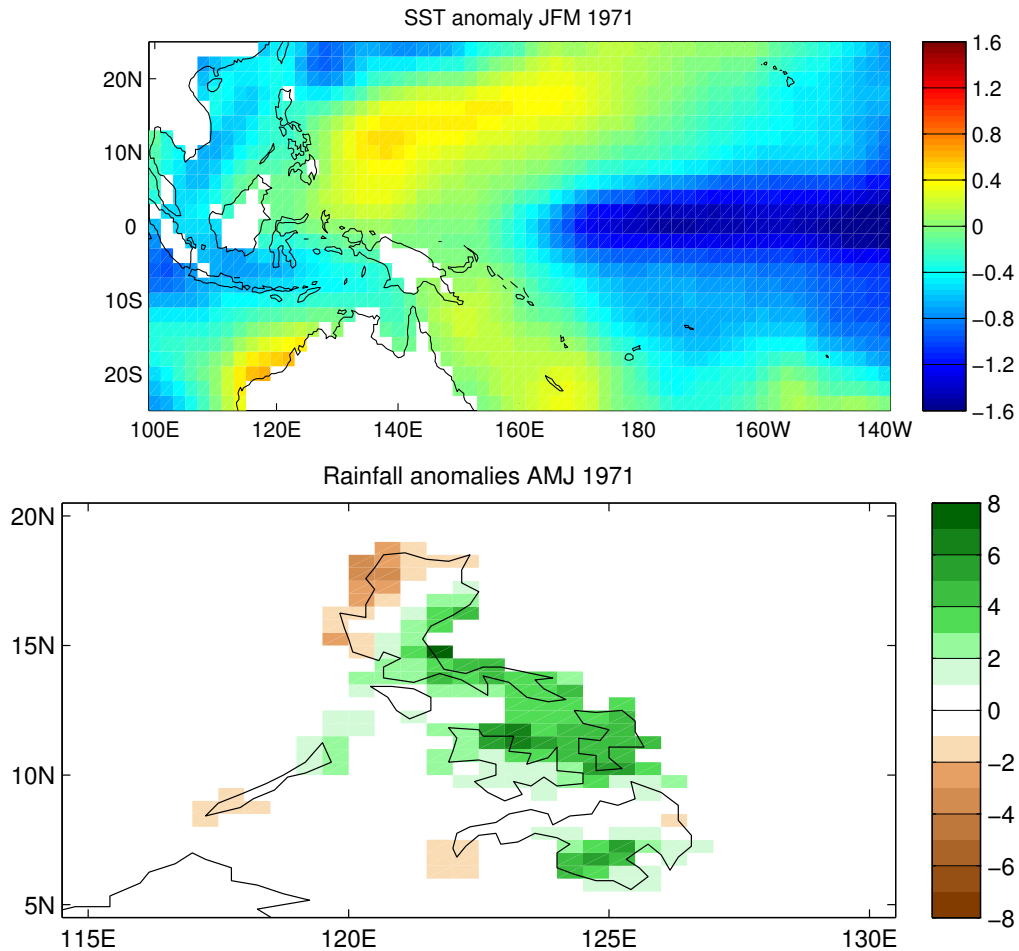
Pitfalls of cross validation

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

From wikipedia

Example: Philippines

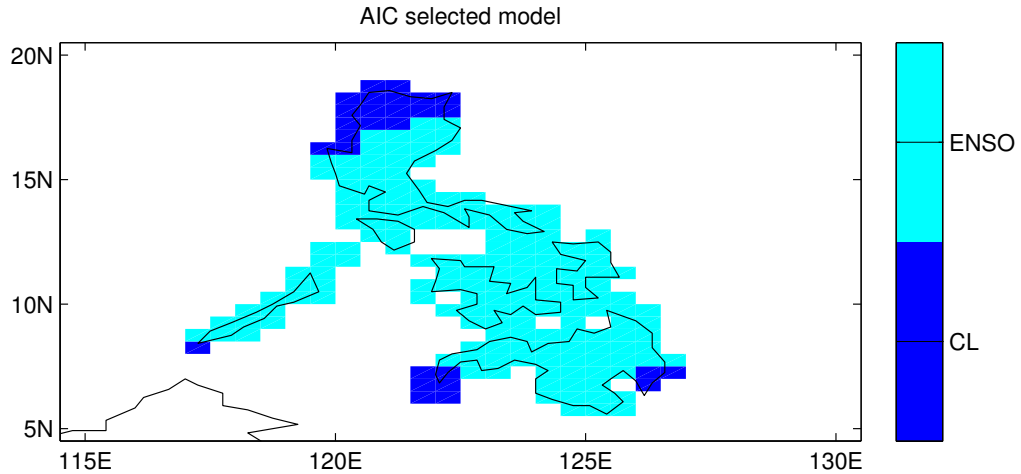
Problem: predict gridded April-June precipitation over the Philippines from proceeding (January-March) SST.



Example: Philippines

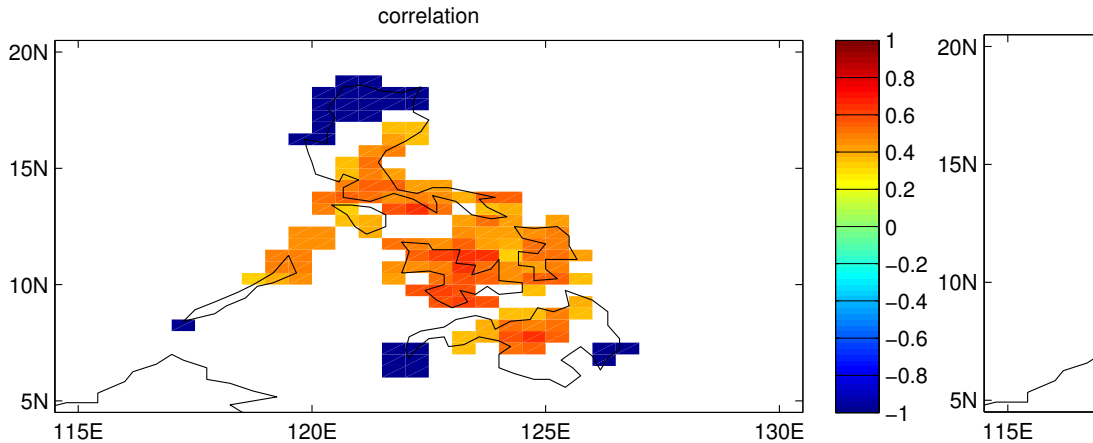
Simple regression model: Either climatology or ENSO as predictors.

Use AIC to choose.



Example: Philippines

- Some skill (cross-validated)
- Why the negative correlation?



Pitfall!

- Showed the model selected.
- Presented the cross-validated skill of that model.

What's wrong?

The entire dataset was used to select the predictors.
Solution?

Pitfall!

- Showed the model selected.
- Presented the cross-validated skill of that model.

What's wrong?

The entire dataset was used to select the predictors.

Solution?

Pitfall!

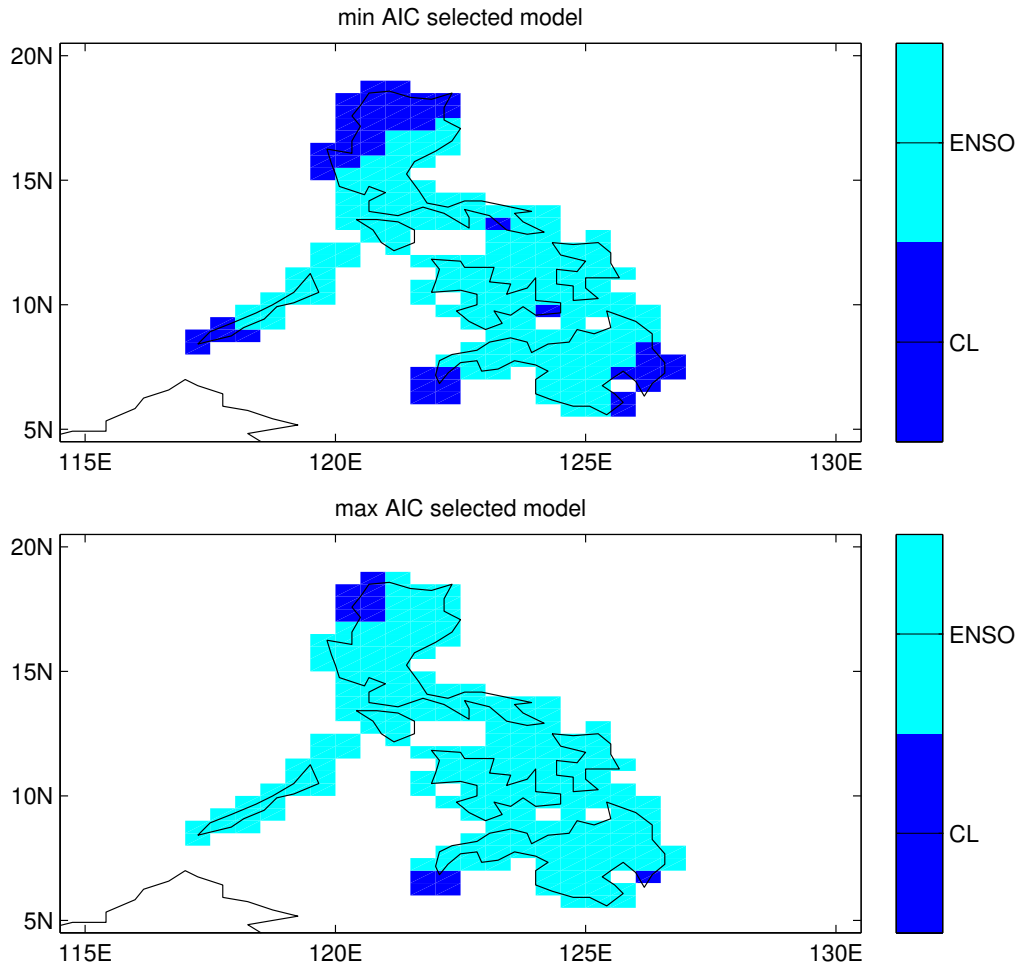
- Showed the model selected.
- Presented the cross-validated skill of that model.

What's wrong?

The entire dataset was used to select the predictors.
Solution?

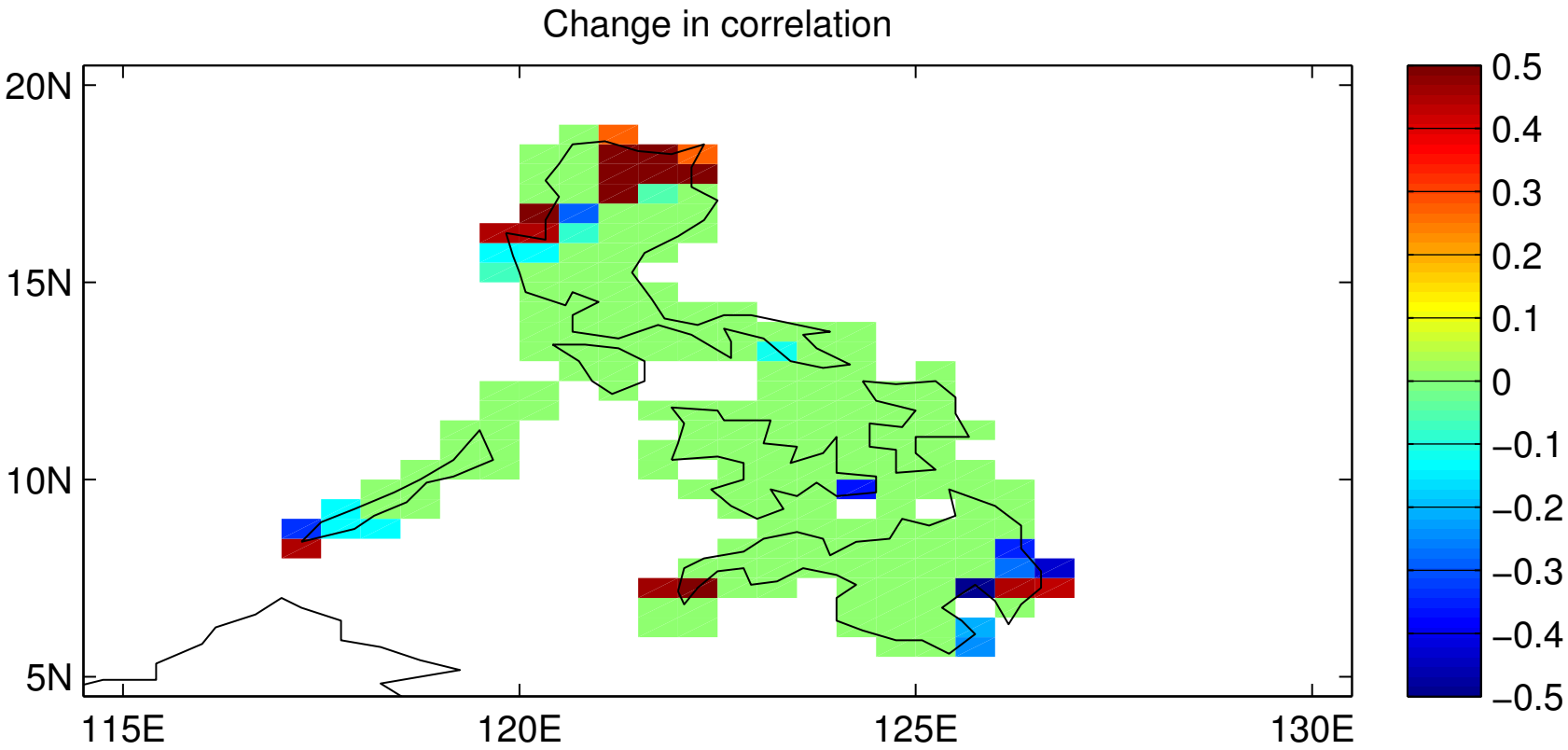
Pitfall!

Solution: include predictor selection in cross-validation.



Pitfall!

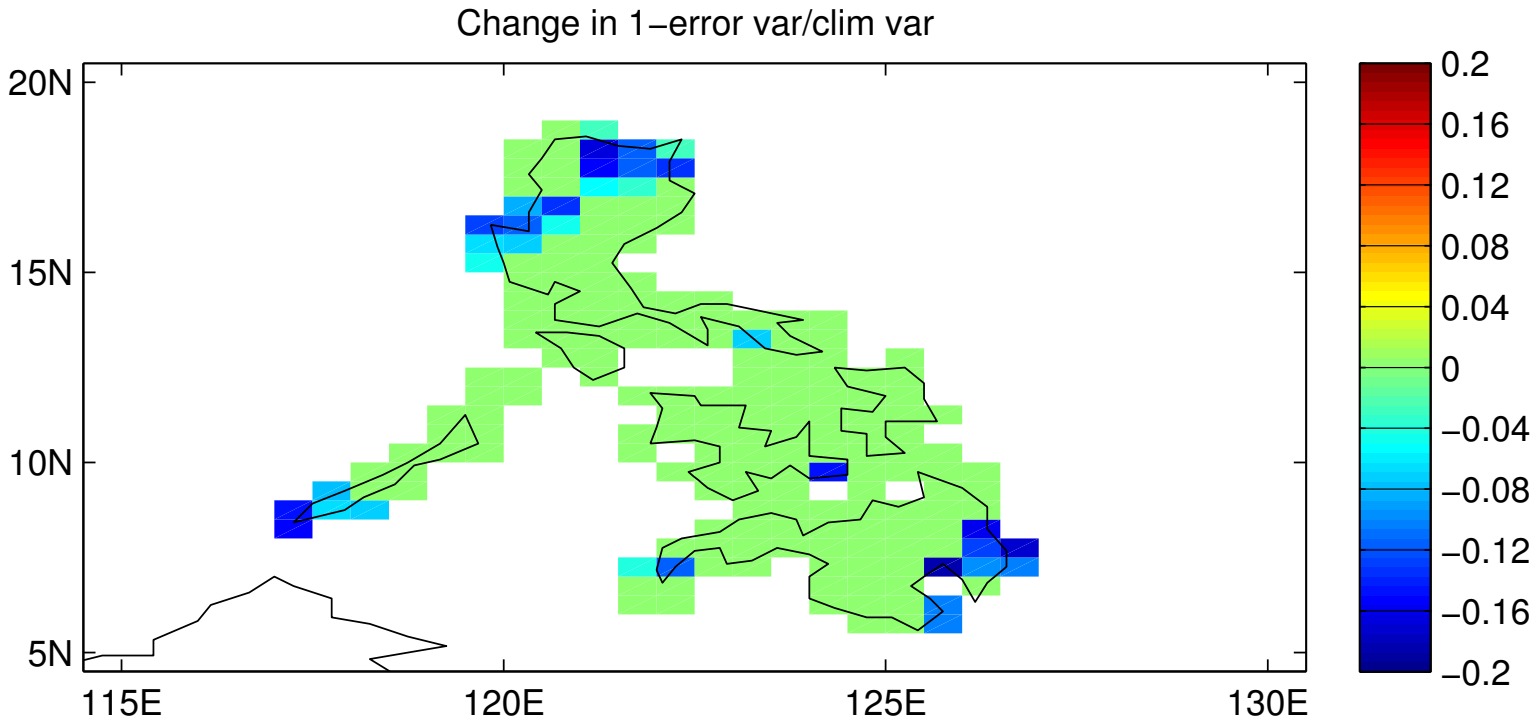
Negative impact on correlation in places with skill?



Why is the impact positive in some areas?

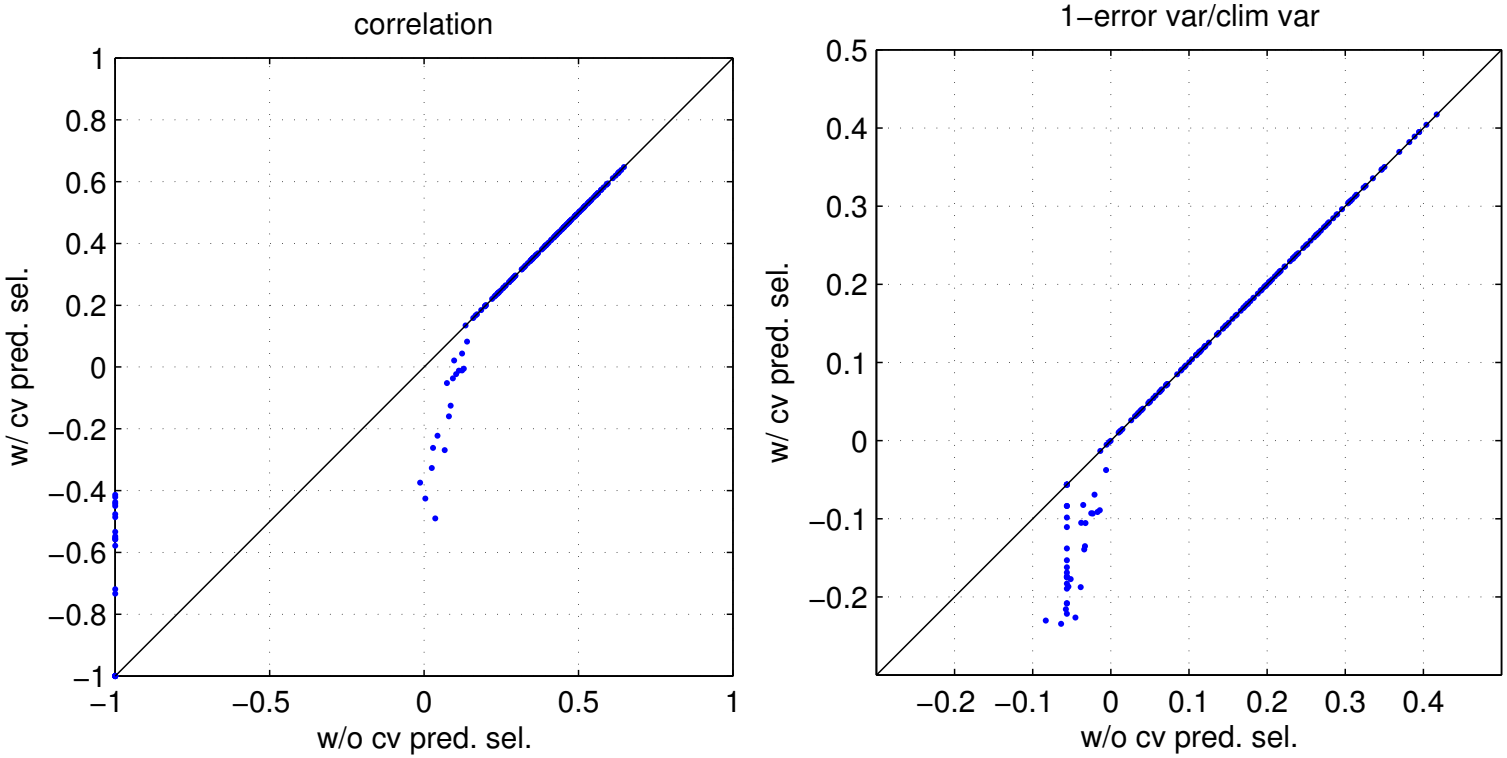
Pitfall!

Negative impact on normalized error



Pitfall!

Points with little skill are most affected.



Idea: convervative models don't go too far bad . . .
What if the model had used many PCs?

A more nefarious (but real) example

- 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 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?

A more nefarious (but real) example

- 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 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?

A more nefarious (but real) example

- 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 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?

A more nefarious (but real) example

- 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 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?

A more nefarious (but real) example

- 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 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?

A more nefarious (but real) example

- 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 and retain the PCs with skill exceeding some threshold as your predictors.
- Estimate skill using cross-validation.

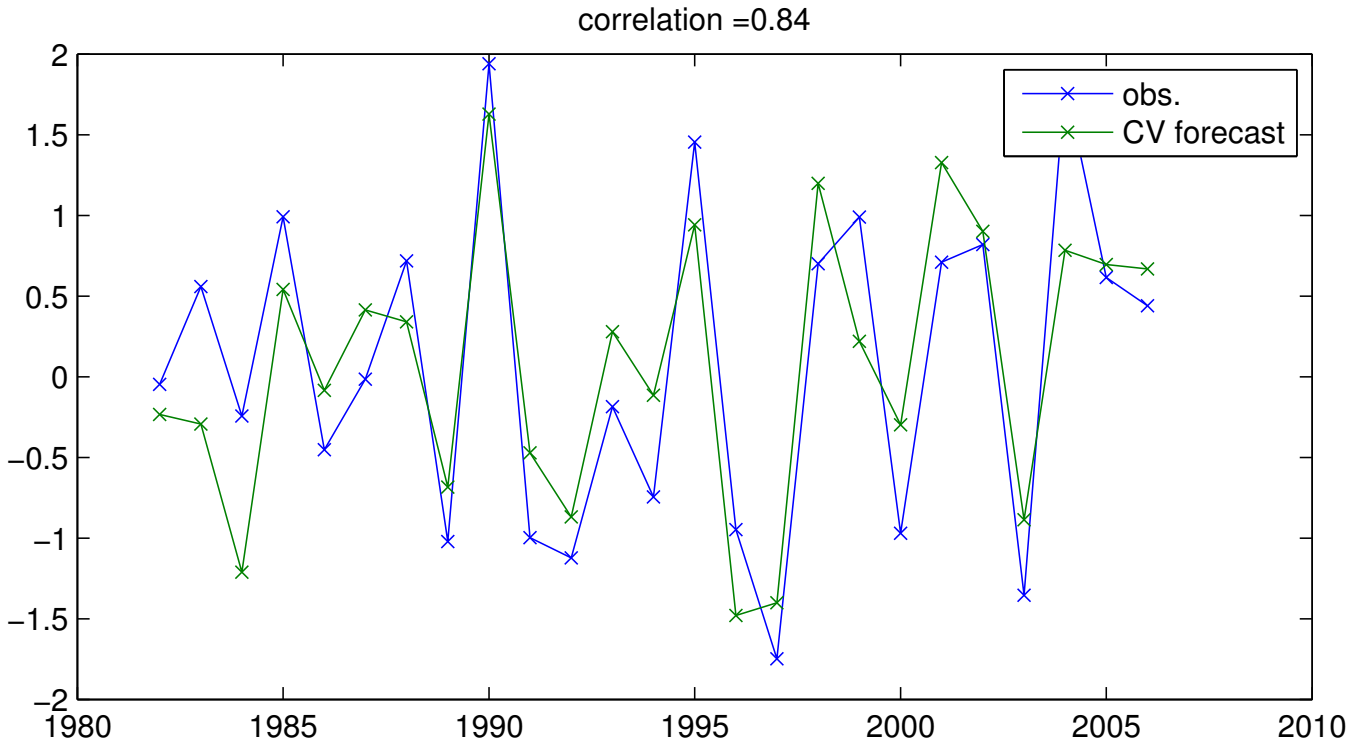
Sounds harmless, maybe even clever.

What is the problem?

What is the impact?

A more nefarious (but real) example

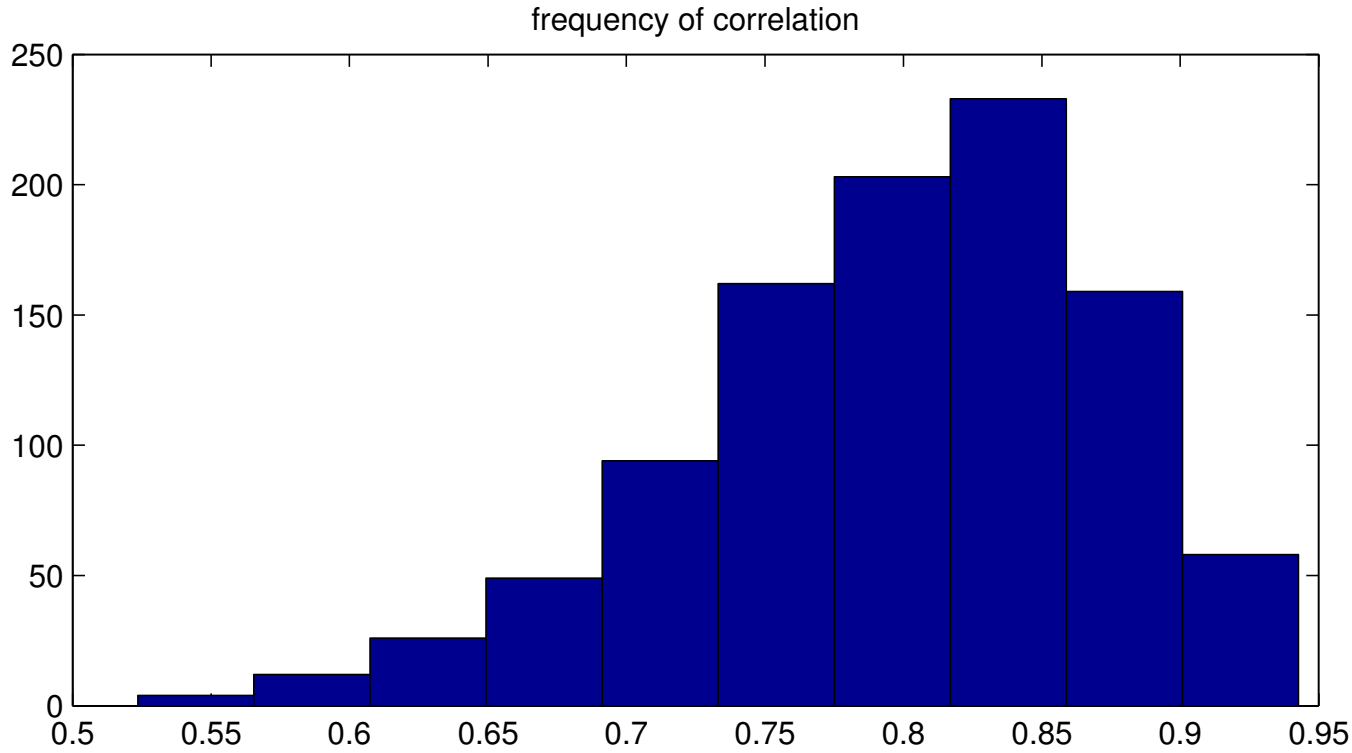
Cross-validated forecasts show good skill.



What is the real skill?

A more nefarious (but real) example

Apply this procedure 1000 times to random numbers

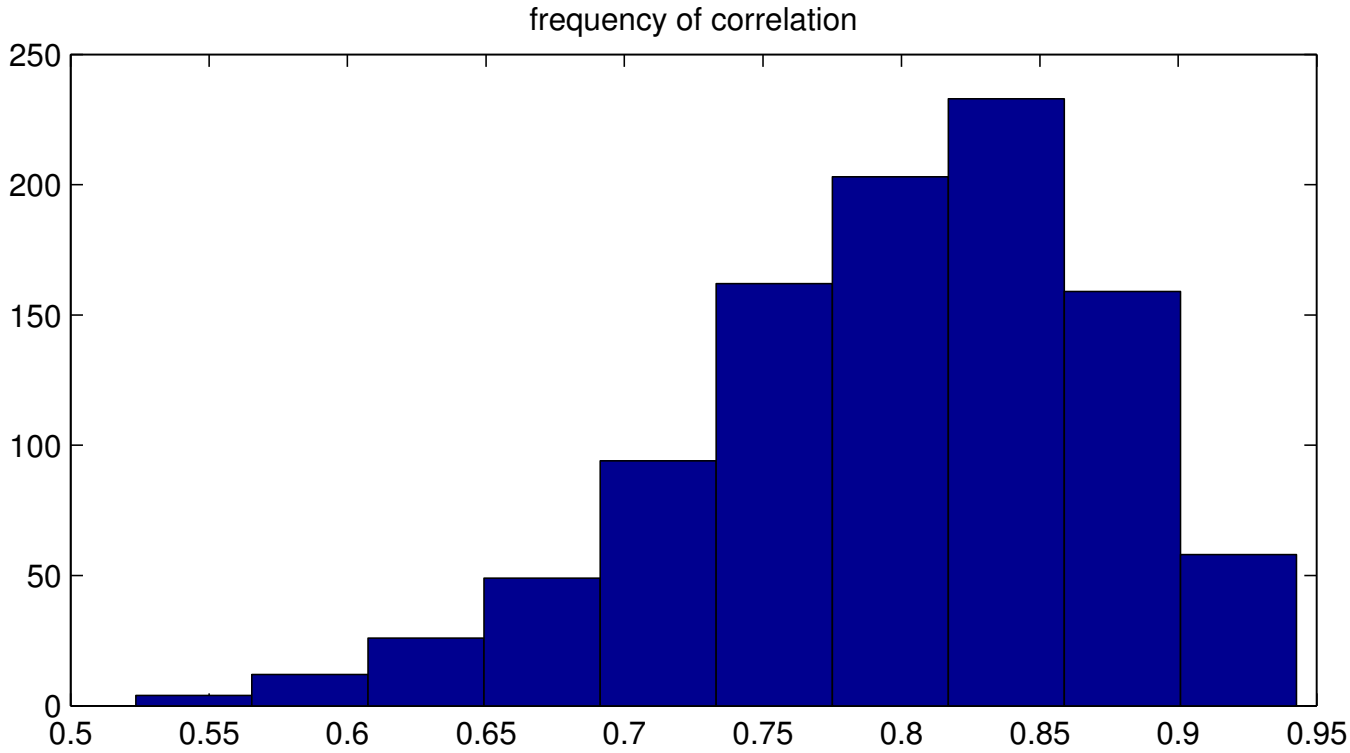


mean correlation = 0.8

Never underestimate the power of a screening procedure.

A more nefarious (but real) example

Apply this procedure 1000 times to random numbers



mean correlation = 0.8

Never underestimate the power of a screening procedure.

Cross-validation and Predictor selection: Bad

```
predictor_selection(x,y)
ypred = y+NA
for(ii in 1:N) {
   out = (ii-1):(ii+1)
   training = setdiff(1:N,out)
   xcv = x[training]
   ycv = y[training]
   model.cv = lm(ycv ~ xcv)
   ypred[ii] = predict(model.cv,list(xcv=x[ii]))
}
```

Cross-validation and Predictor selection: Good

```
ypred = y+NA
for(ii in 1:N) {
    out = (ii-1):(ii+1)
    training = setdiff(1:N,out)
    xcv = x[training]
    ycv = y[training]
    predictor_selection(xcv,ycv)
    model.cv = lm(ycv ~ xcv)
    ypred[ii] = predict(model.cv,list(xcv=x[ii]))
}
```

Summary

- Predictor selection needs to be included in the cross-validation.
- Impact varies.

Example: PCA and regression

We asked:
Is there any benefit to predicting the PCs of y rather than y?

Compared regression at each gridpoint to regression between patterns.

- Compute PCs of SST
- Compute PCs of rainfall.
- Skill from cross-validated regression between the PCs.

What is the problem?

Cannot do PCA of "future" observations. PCA of y needs to go in the CV-loop

Example: PCA and regression

We asked:
Is there any benefit to predicting the PCs of y rather than y?

Compared regression at each gridpoint to regression between patterns.

- Compute PCs of SST
- Compute PCs of rainfall.
- Skill from cross-validated regression between the PCs.

What is the problem?

Cannot do PCA of "future" observations. PCA of y needs to go in the CV-loop

Example: PCA and regression

We asked:
Is there any benefit to predicting the PCs of y rather than y?

Compared regression at each gridpoint to regression between patterns.

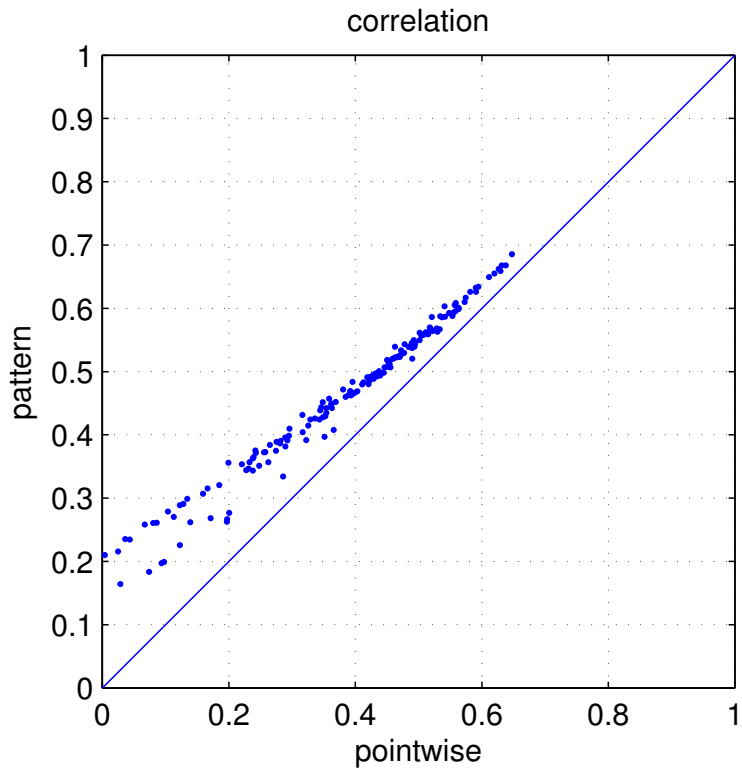
- Compute PCs of SST
- Compute PCs of rainfall.
- Skill from cross-validated regression between the PCs.

What is the problem?

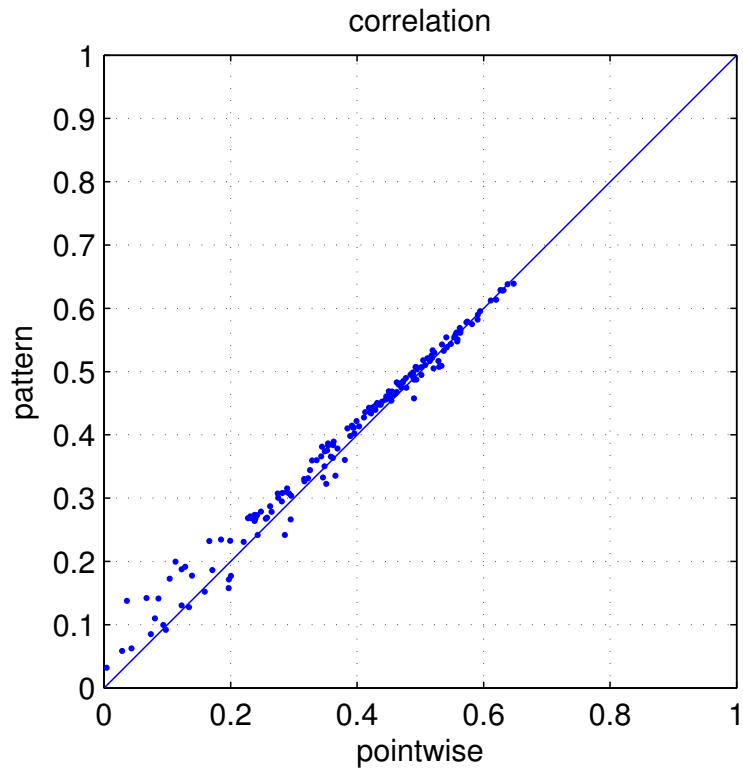
Cannot do PCA of “future” observations. PCA of y needs to go in the CV-loop

Example: Philippines

Y PCA outside CV

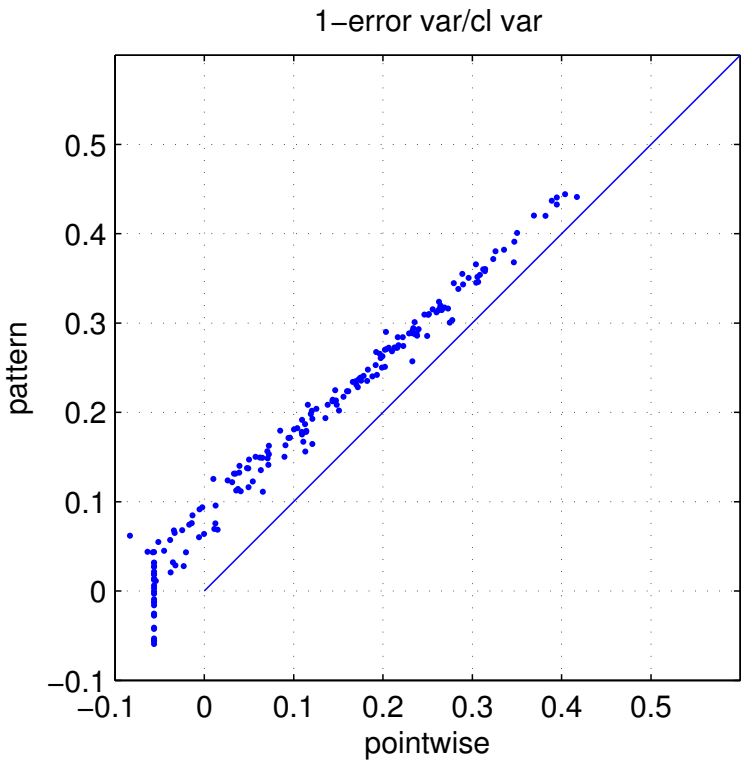


Y PCA inside CV

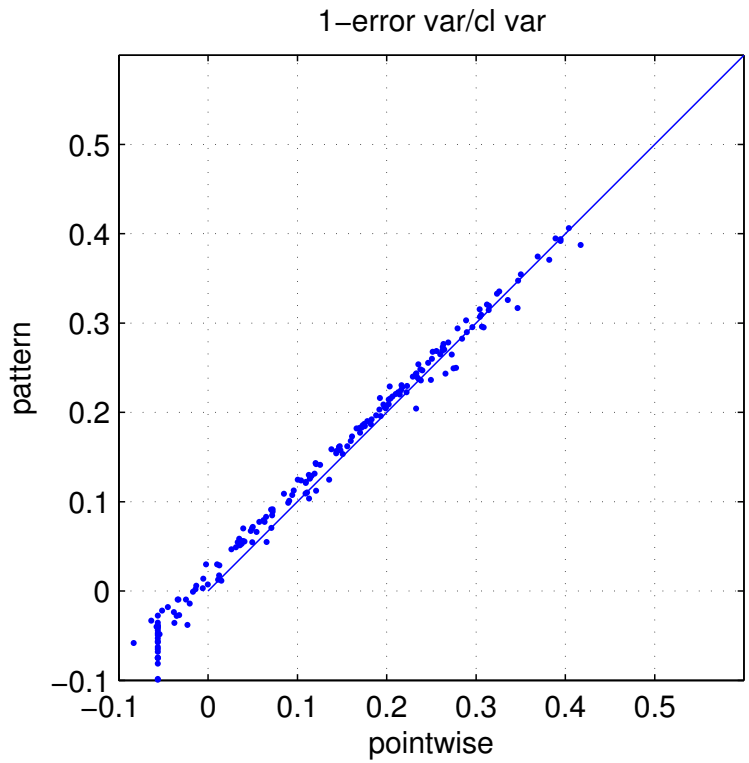


Example: Philippines

Y PCA outside CV



Y PCA inside CV



Similar problem.

- Do CCA.
- Find patterns and time-series.
- Use time-series in a regression.
- Check skill using cross-validation.

What is a problem?

What is a solution?

Similar problem.

- Do CCA.
- Find patterns and time-series.
- Use time-series in a regression.
- Check skill using cross-validation.

What is the problem?

What is a solution?

CPT

Climate prediction tool.

- 3 consecutive years are left out.
- CCA is applied to the remaining years.
  - CCA depends on the number of PCs retained.
- Middle year of the left out years is forecast.
- Repeat until all years are forecast.
- Cross-validated forecast depends on the number of PCs retained.
- Select number of PCs that optimizes cross-validated skill. This skill is the forecast skill.

What is the problem? Solution?

CPT

Climate prediction tool.

- 3 consecutive years are left out.
- CCA is applied to the remaining years.
  - CCA depends on the number of PCs retained.
- Middle year of the left out years is forecast.
- Repeat until all years are forecast.
- Cross-validated forecast depends on the number of PCs retained.
- Select number of PCs that optimizes cross-validated skill. This skill is the forecast skill.

What is the problem? Solution?

CPT

Climate prediction tool.

- 3 consecutive years are left out.
- CCA is applied to the remaining years.
  - CCA depends on the number of PCs retained.
- Middle year of the left out years is forecast.
- Repeat until all years are forecast.
- Cross-validated forecast depends on the number of PCs retained.
- Select number of PCs that optimizes cross-validated skill. This skill is the forecast skill.

What is the problem? Solution?

Three data sets

- Data set 1 = estimate parameters of the model.
- Data set 2 = select predictor/model.
- Data set 3 = estimate skill of model

Why are two data sets not enough?

"Example"
Suppose many models are compared.
Same skill except for sampling differences. $s \pm \delta$
Pick model with largest skill $s + \delta$, larger than real skill s.

Three data sets

- Data set 1 = estimate parameters of the model.
- Data set 2 = select predictor/model.
- Data set 3 = estimate skill of model

Why are two data sets not enough?

"Example"
Suppose many models are compared.
Same skill except for sampling differences. $s \pm \delta$
Pick model with largest skill $s + \delta$, larger than real skill s.

Three data sets

- Data set 1 = estimate parameters of the model.
- Data set 2 = select predictor/model.
- Data set 3 = estimate skill of model

Why are two data sets not enough?

"Example"
Suppose many models are compared.
Same skill except for sampling differences. $s \pm \delta$
Pick model with largest skill $s + \delta$, larger than real skill s.

Three data sets

- Data set 1 = estimate parameters of the model.
- Data set 2 = select predictor/model.
- Data set 3 = estimate skill of model

Why are two data sets not enough?

"Example"
Suppose many models are compared.
Same skill except for sampling differences. $s \pm \delta$
Pick model with largest skill $s + \delta$, larger than real skill s.

SST prediction

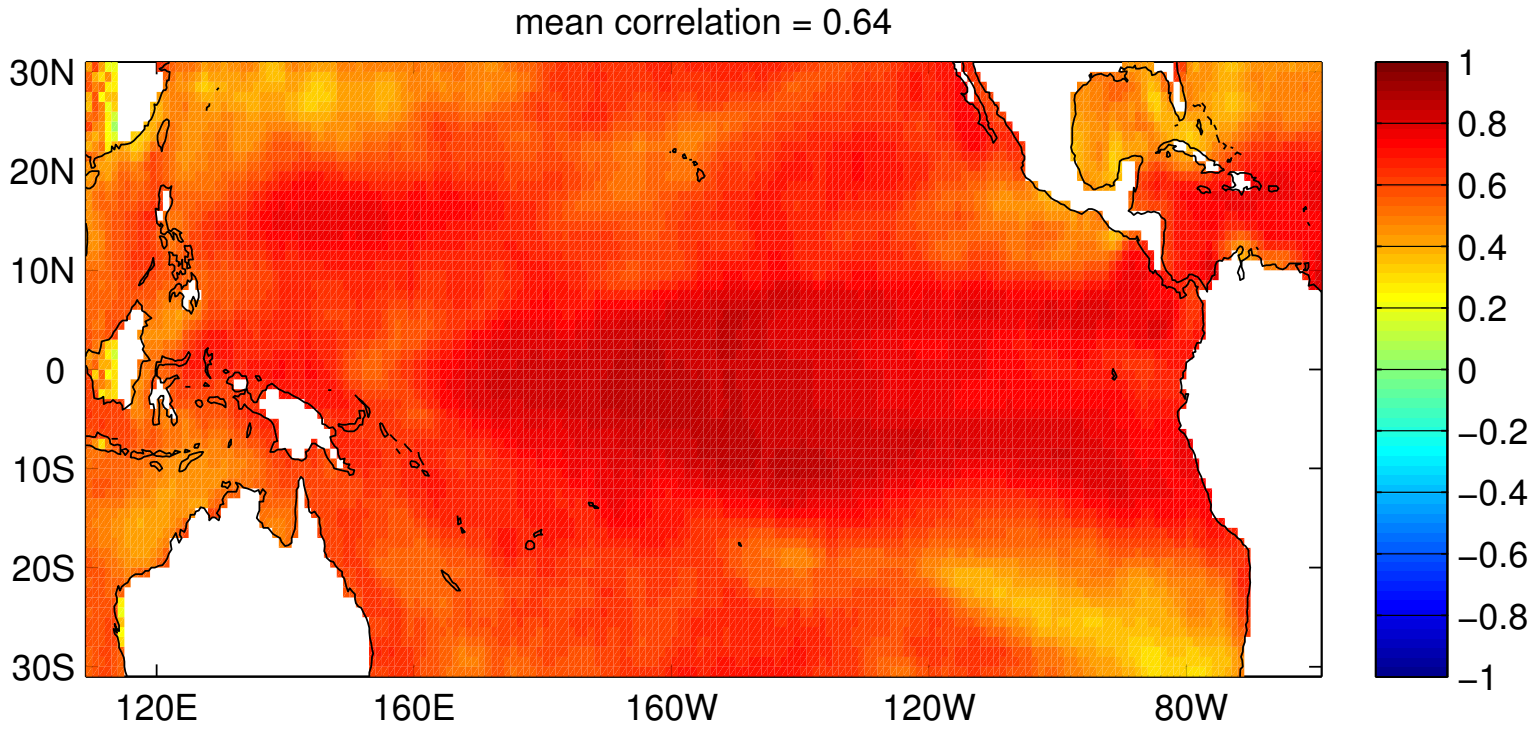
- Predict monthly SST anomalies from monthly SST anomalies six months before.
- 1982-2009. 28 years.
- PCA of monthly SST anomalies.
- Cross validation to pick the number of PCs.
- Look at the correlation of those cv'd forecasts.

What are the problems?

SST prediction

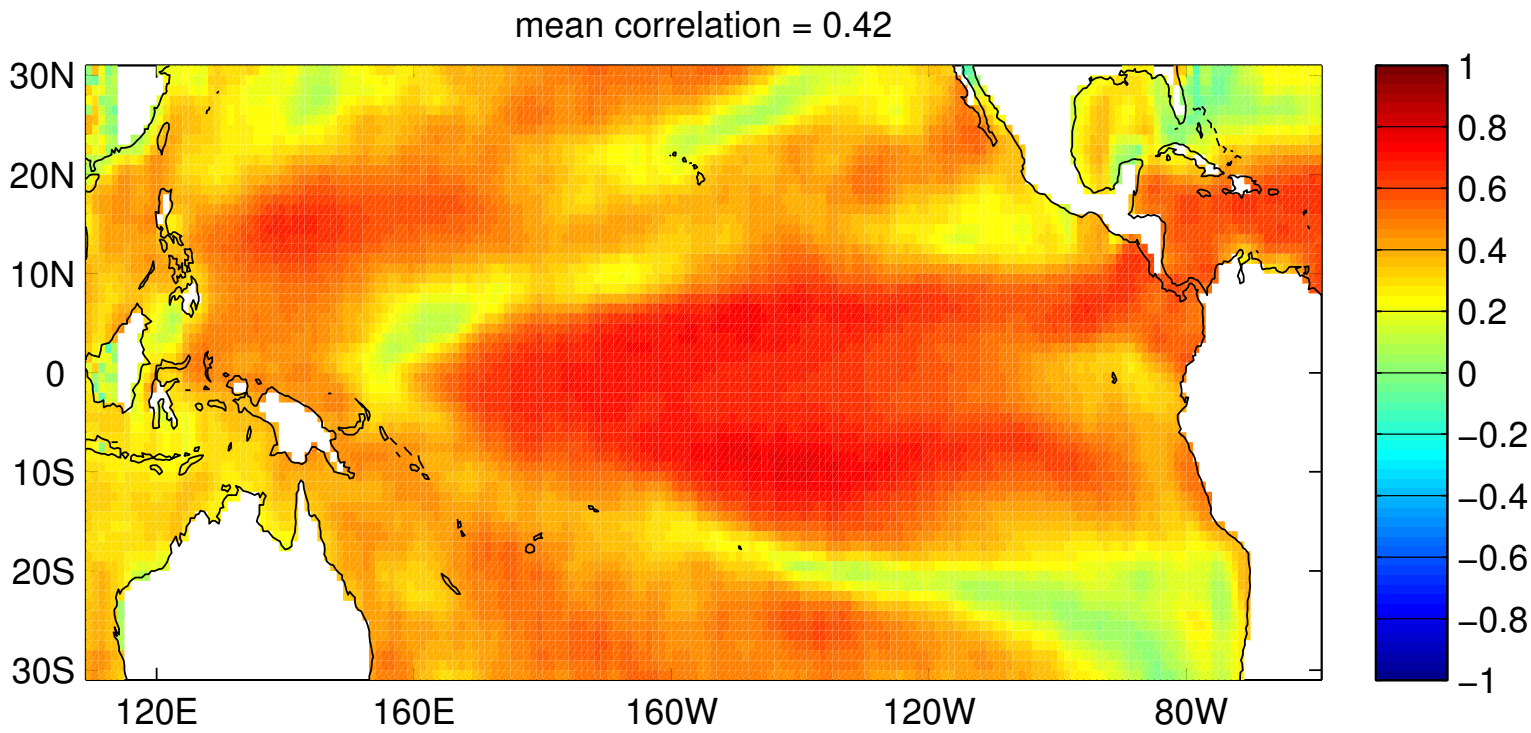
- Leave-one-month out cross-validation.
- Checked truncations up to 75 PCs.
- Lowest cross-validated error with 66 PCs.

What's wrong with this picture?



SST prediction

- Leave-one-year-out cross-validation.
- Checked truncations up to 75 PCs.
- Lowest cross-validated error with 6 PCs.



Summary

- Efficient method to compute leave-one-out cross-validation for linear regression.
- There are some biases with CV. Climatology forecasts have negative correlation.
- Include model/predictor selection in the CV.
- Left-out data must be independent.