Pitfalls of Statistical Prediction: Screening

Timothy DelSole

George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD

February 1, 2010

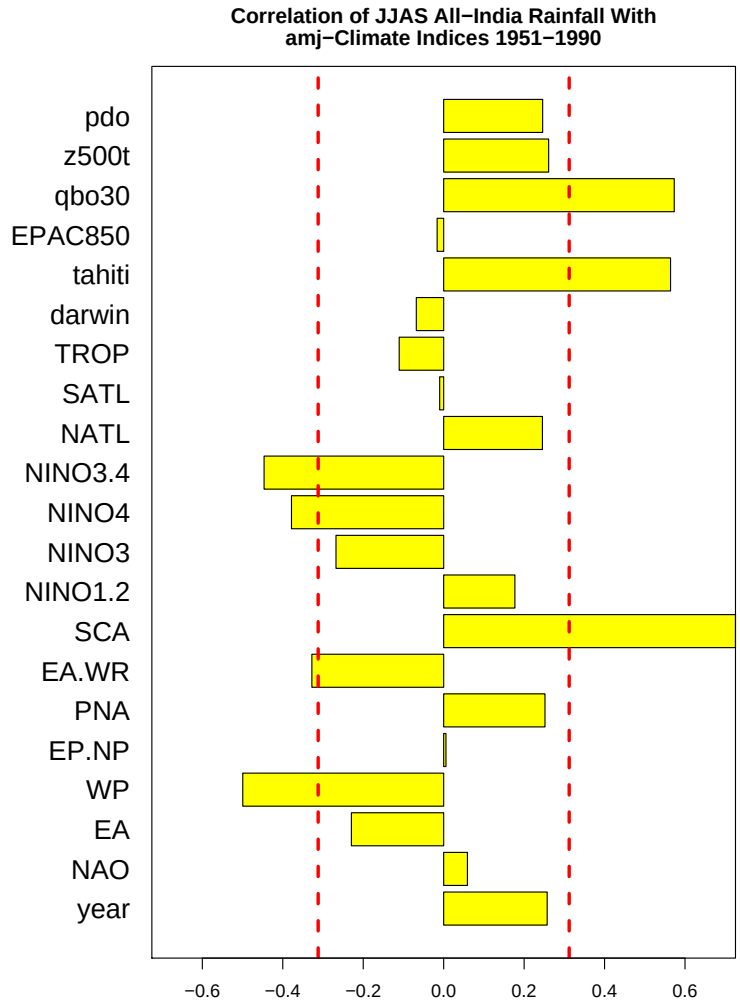


Figure: Correlation between AMJ-climate indices and all-India Monsoon Rainfall for 1951-2008. Red dashed line is the 5% significance level.

Gilbert Walker

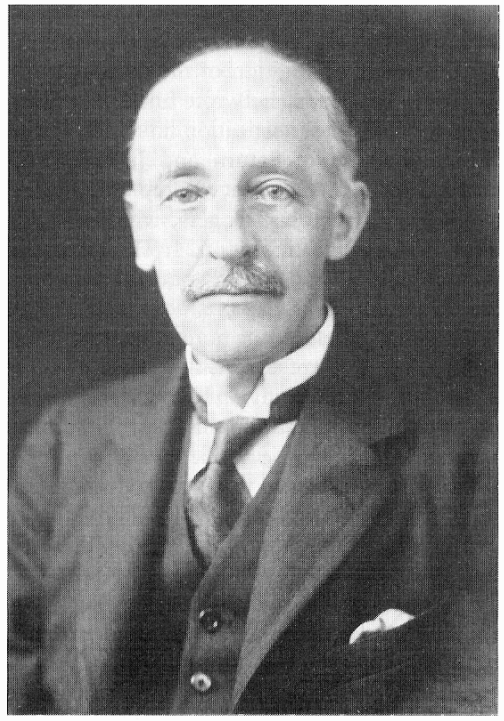


Figure: Sir Gilbert Walker

"[Let c be the probability that the correlation between independent quantities is less than p.] Then the chance of all coefficients [between m pairs of independent quantities] being less than p will be c^m." -Walker 1914

Experimentwise Error Rate

The 5% significance level is the absolute correlation below which sample correlations will fall 95% of the time, for independent data.

However, if the sample correlation is calculated for M different indices, then the probability that at least one correlation out of M exceeds the α-significance level is

$$prob = 1 - (1 - \alpha)^M$$

M	1	2	3	4	5	10	20
prob	5%	10%	14%	19%	23%	40%	64%

Table: Probability that event occurs at least once in M trials when probability of the event occurring in one trial is 5%

The probability of at least one false rejection of the null hypothesis over multiple comparisons is called the experimentwise error rate.

Multiple Comparisons

The comparisonwise $\alpha_c = 5\%$ significance level should NOT be used for multiple comparisons.

For multiple comparisons, one should use the experimentwise significance level:

$$\alpha_e = 1 - (1 - \alpha_c)^{1/M}$$

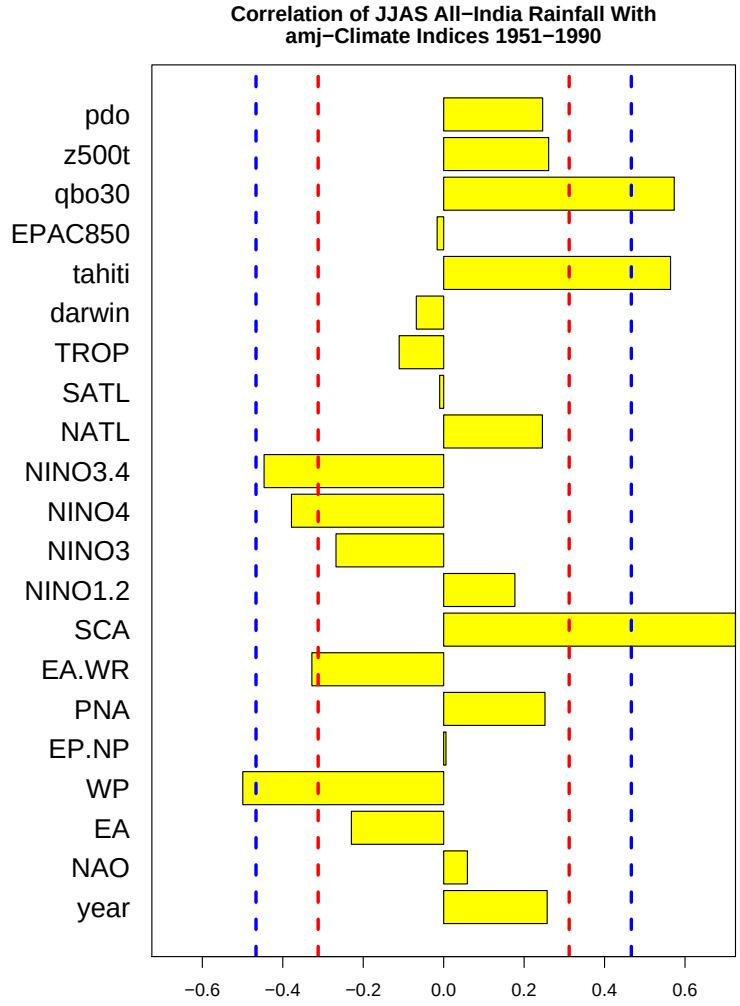
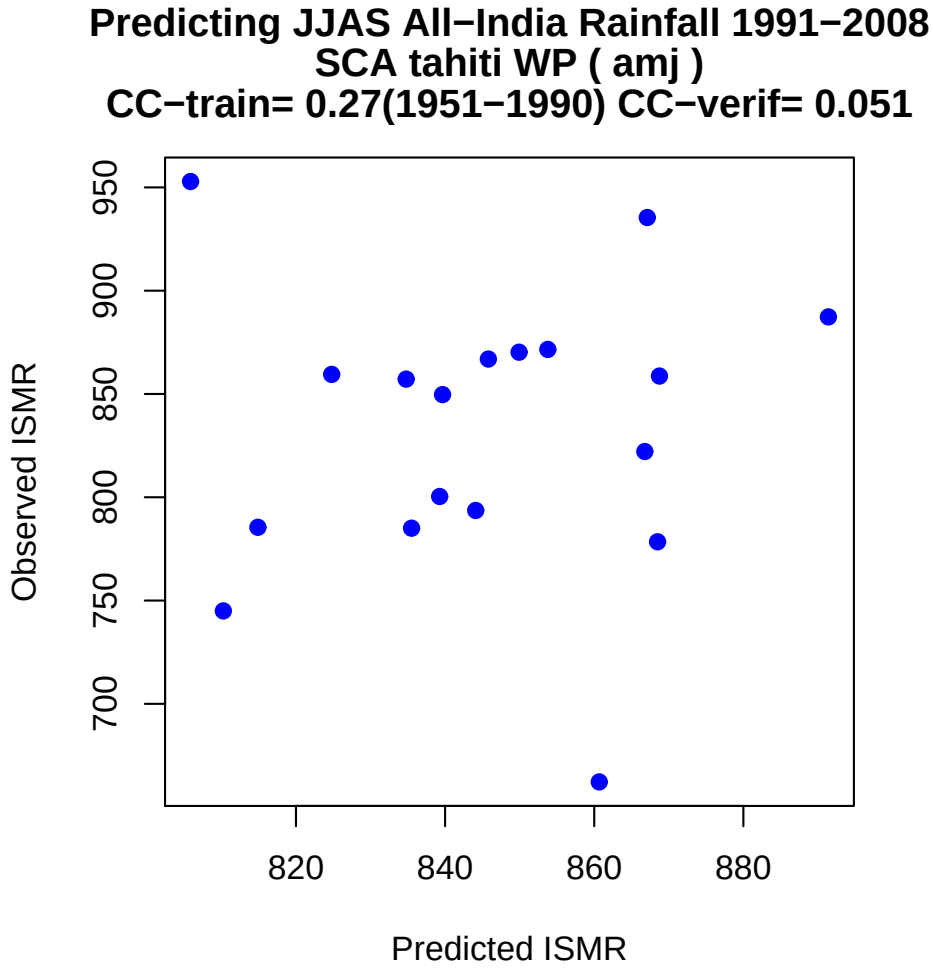


Figure: Correlation between AMJ-climate indices and all-India Monsoon Rainfall for 1951-2008. Red and blue dashed lines are the 5% significance level for comparisonwise and experimentwise comparisons, respectively.

Independent Verification



Which Predictors Should Be Chosen?

In statistical climate prediction, the available data is about 50-150 years whereas the number of possible predictors is thousands:

- Sea surface temperature
- Sea level pressure
- Winds (at different levels)
- Precipitation
- Soil moisture
- Snow cover

Predictors depend on spatial location and lag (months to years).

The availability of many more predictors than years presents special problems in statistical analysis.

Screening

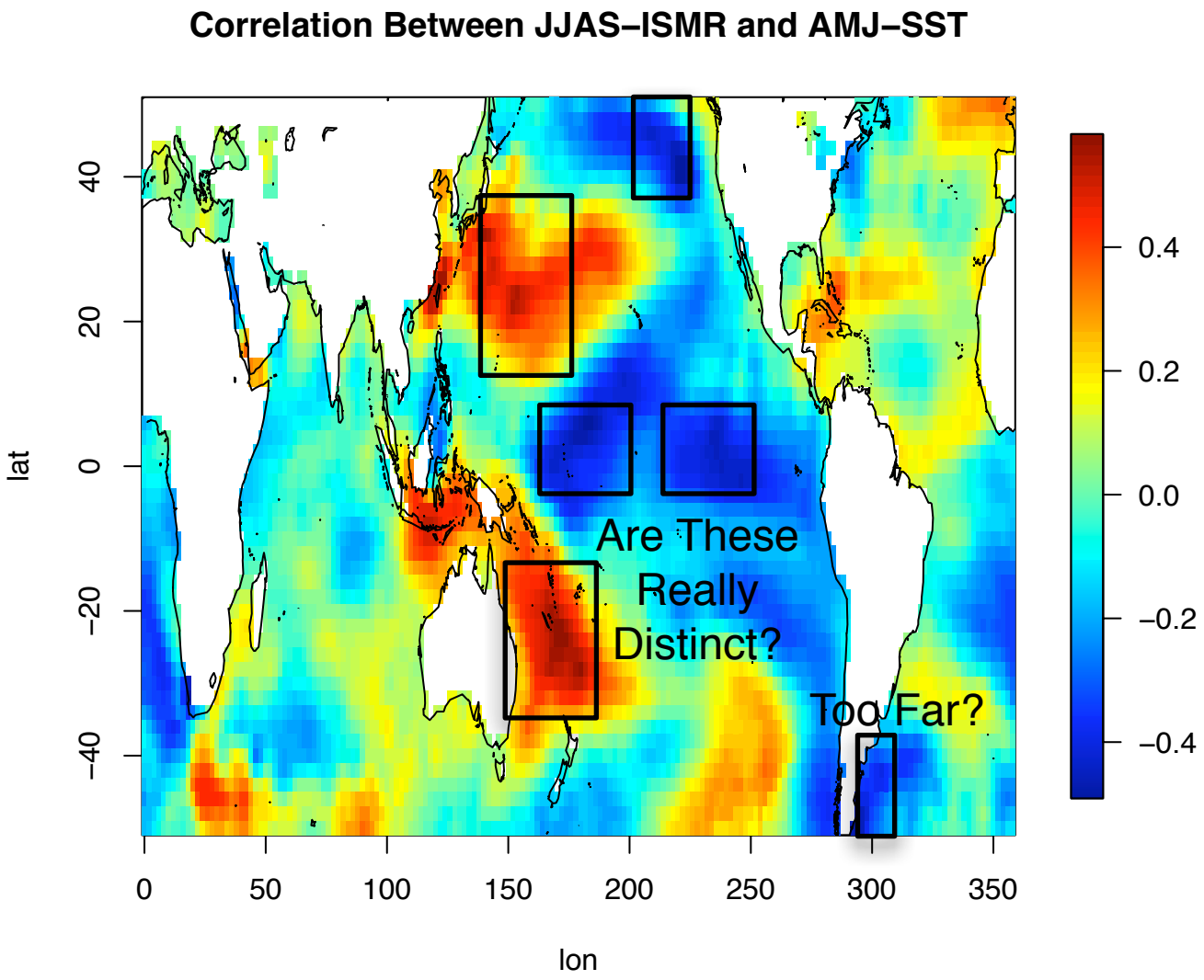
Screening: the process of preferentially selecting variables because of their strong correlation with a prediction variable.

Screening is used by many forecasters, including

- Indian Meteorological Department (for predicting monsoons)
- Klotzbach and Gray (for predicting hurricanes)

Example: Correlation Maps

A typical application of screening is to construct correlation maps and then average over regions with large correlation.



Example: Selecting Monsoon Predictors

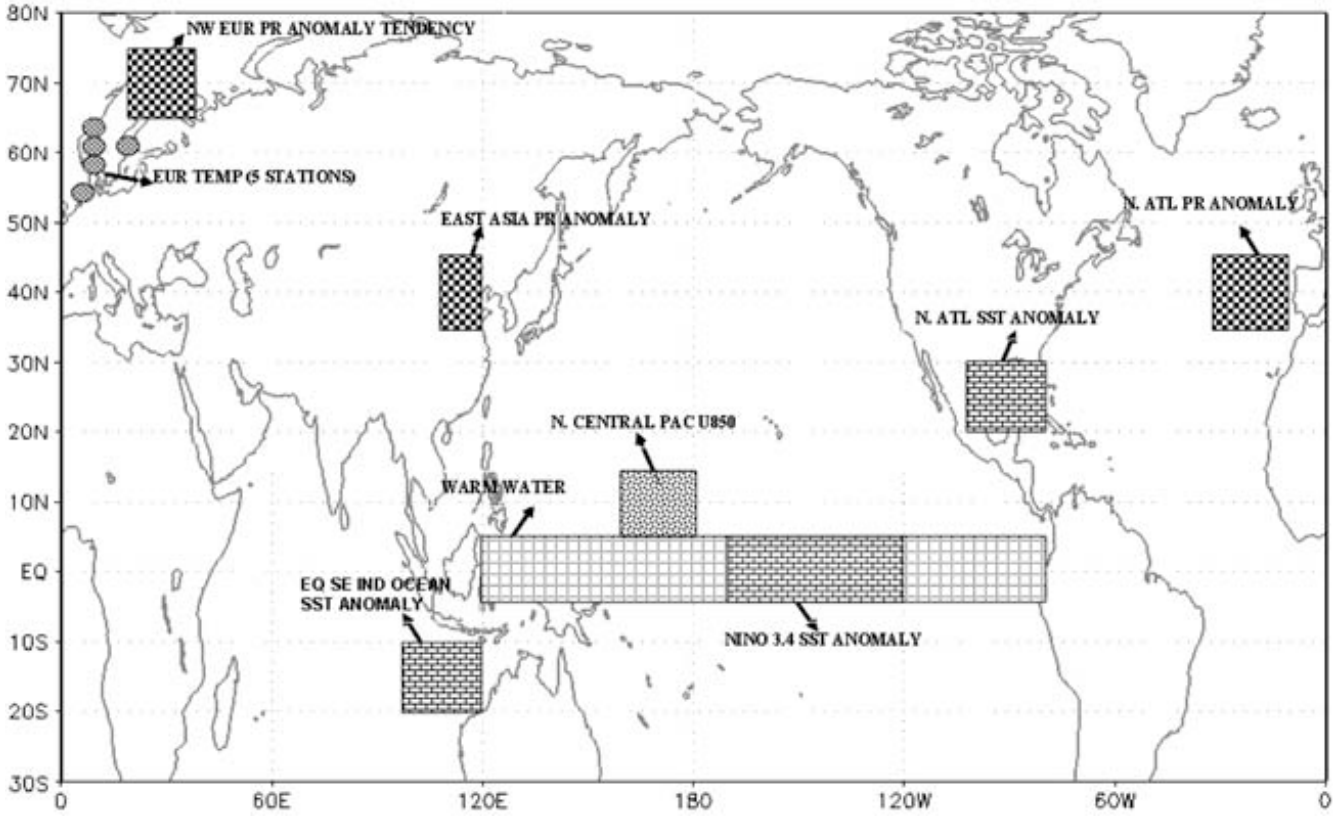
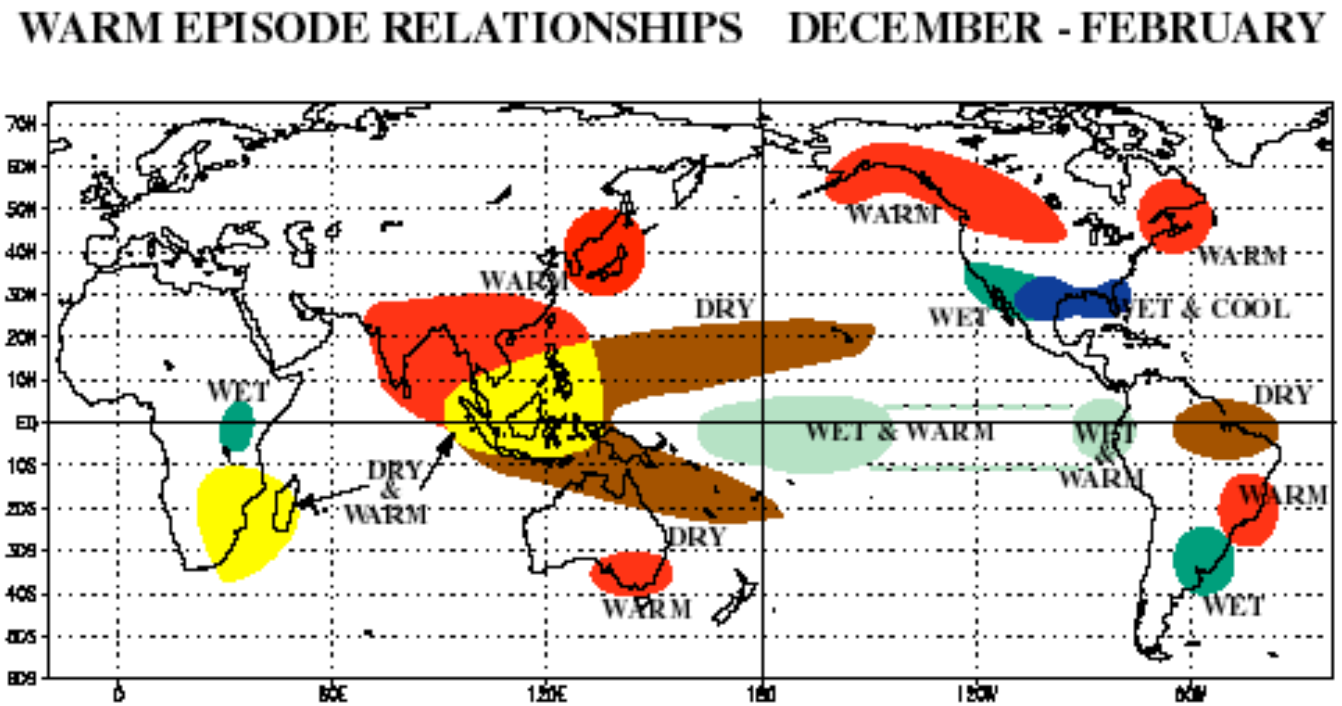


Figure: Location of predictors for Indian Meteorological Department May statistical prediction of monsoon for 2007.

Verification Is Not Just For Predictions
It's Also For Understanding



Correlation maps like this are used to identify relations between ENSO and Precipitation and Temperature. Are they correct, or just spurious results due to examining too many variables?[1]

[1] figure from http://www.cpc.noaa.gov/products/analysis_monitoring/impacts/enso.html

Getting Time Series From Patterns

Given a correlation map,

$$\underset{[M \times 1]}{\mathbf{p}}$$

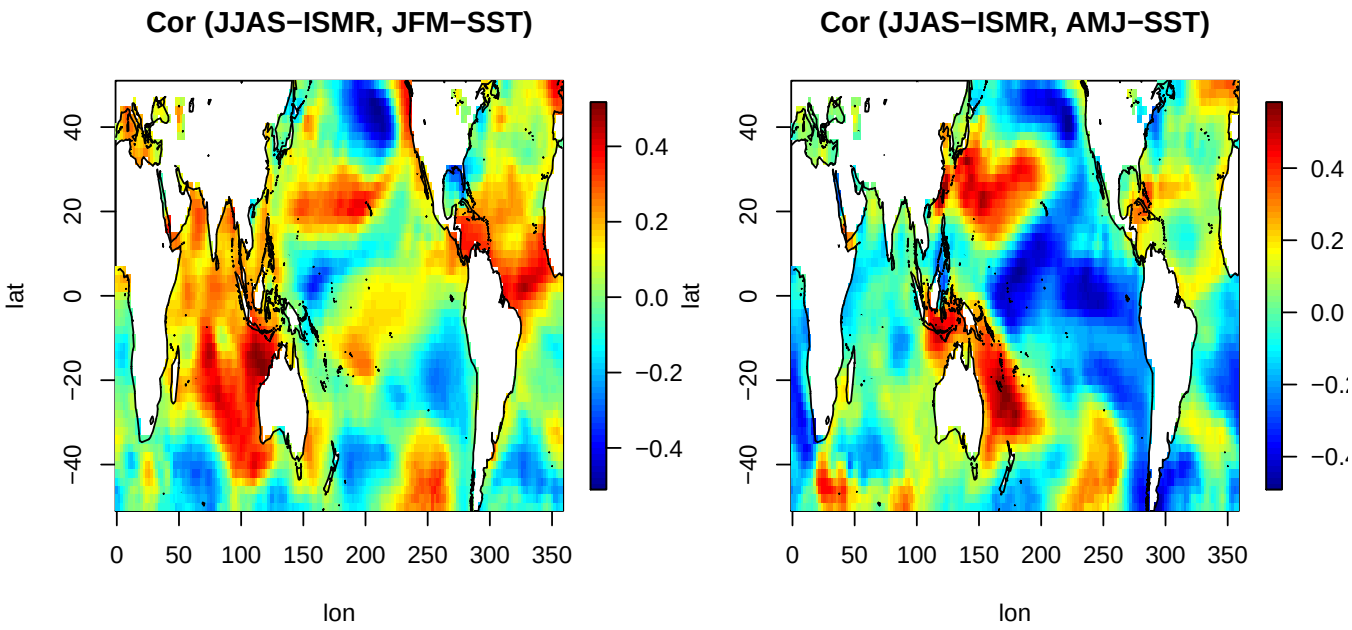
what is the time series $\mathbf{r}$ that goes with it, in the sense that

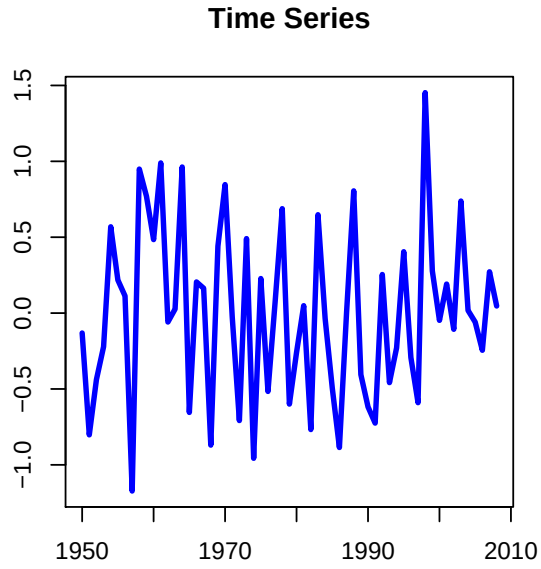
$$\underset{[N \times M]}{\mathbf{X}} = \underset{[N \times 1]}{\mathbf{r}} \underset{[1 \times M]}{\mathbf{p}^T} + \underset{[N \times M]}{\text{other stuff}} .$$

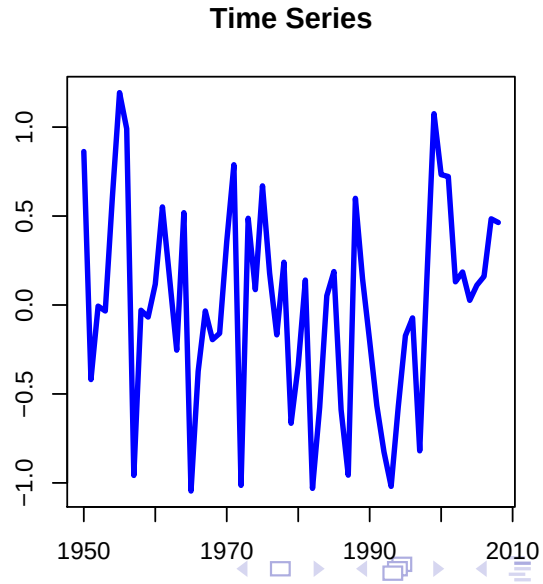
A sensible answer is that you find the time series $\mathbf{r}$ that best approximates $\mathbf{X}$ in a least squares sense, which is

$$\hat{\mathbf{r}} = \mathbf{X}\mathbf{p}\left(\mathbf{p}^T\mathbf{p}\right)^{-1}$$

Two Correlation Maps and Their Projection on SST



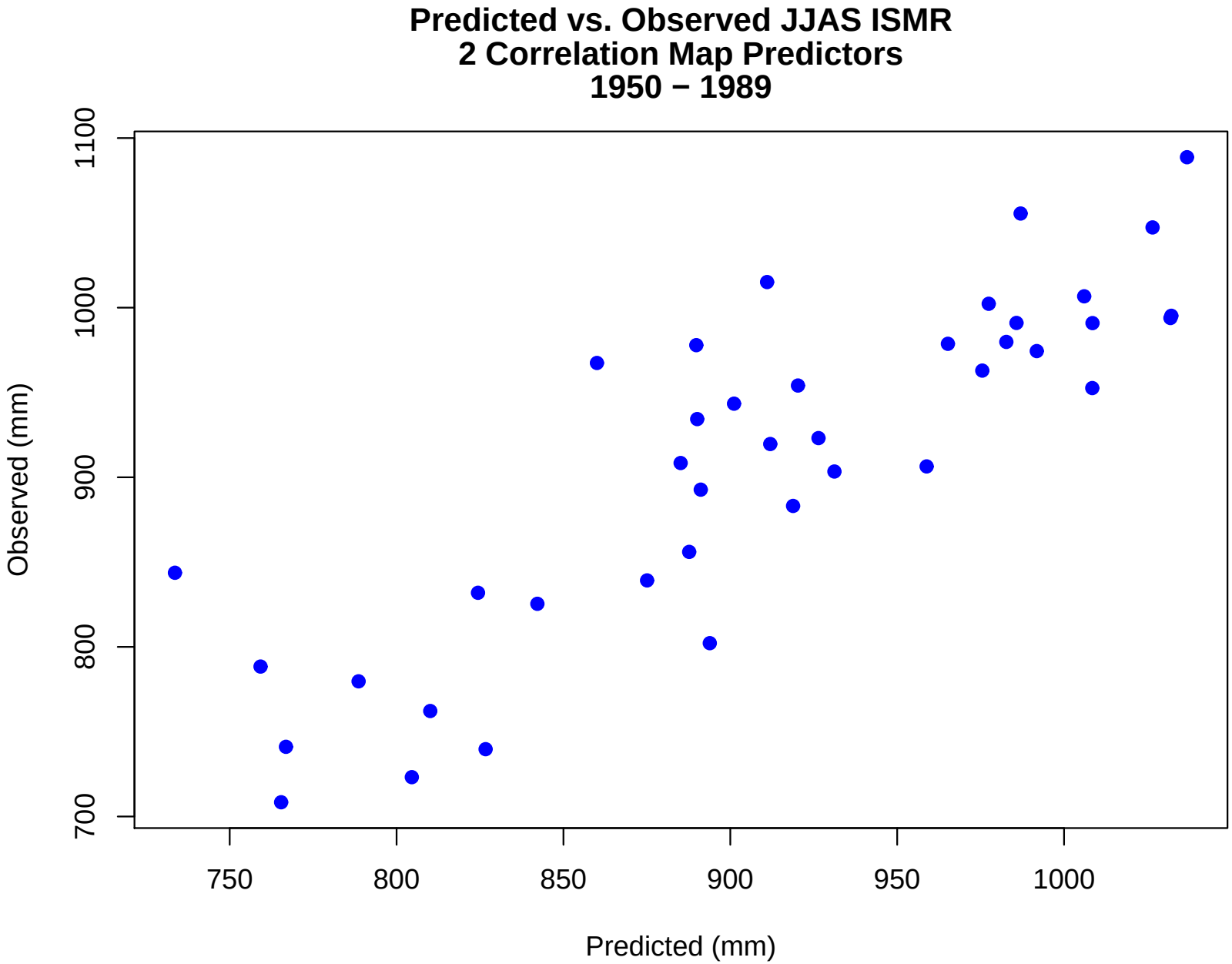




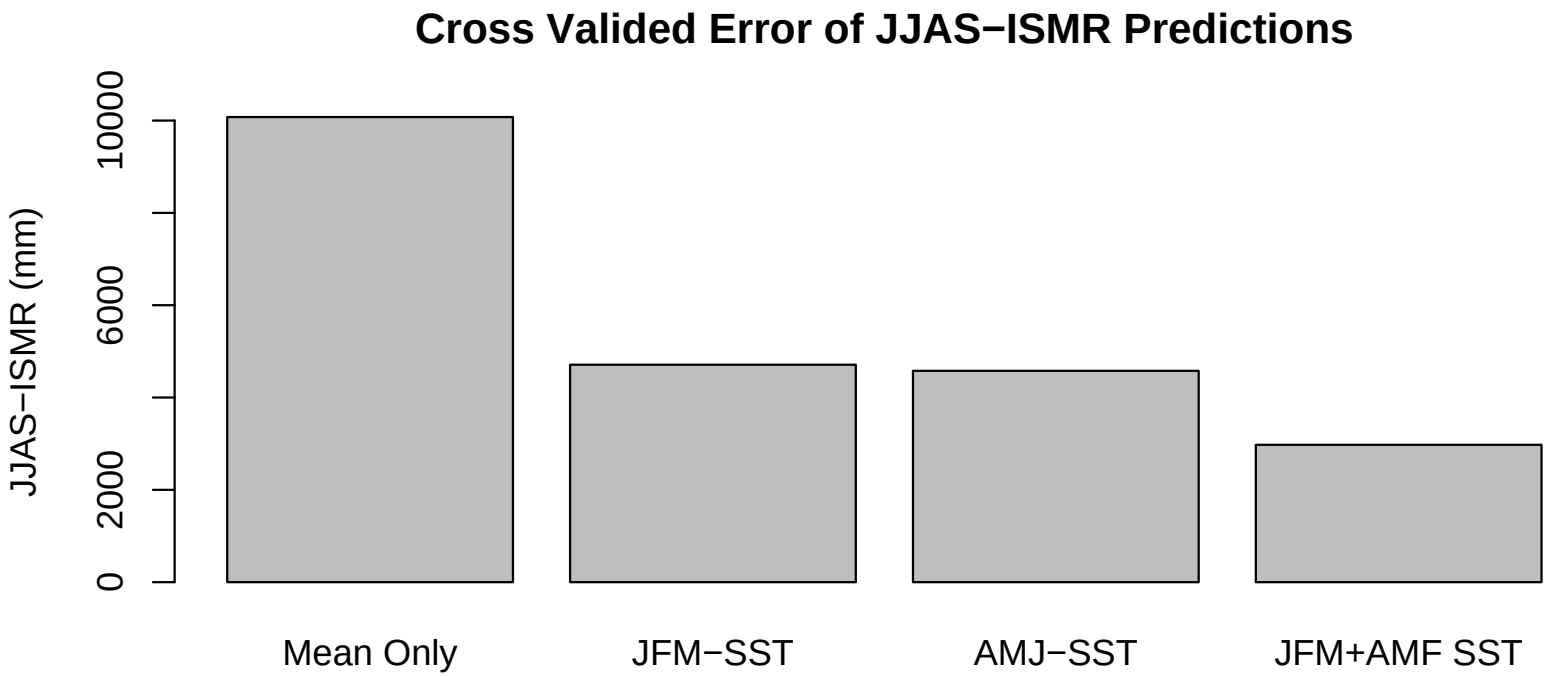
Use Projections from Two Correlation Maps as Predictors

```
> goodyears = which(!is.na(ismr.jjas))
> summary(ismr.jjas[goodyears] ~ cor.jfm.time[goodyears]
+ cor.amj.time[goodyears])

Call:
lm(formula = ismr.jjas[goodyears] ~ cor.jfm.time[goodyears] +
    cor.amj.time[goodyears])

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               910.274      8.231 110.594  < 2e-16 ***
cor.jfm.time[goodyears]    79.071     16.081   4.917 1.82e-05 ***
cor.amj.time[goodyears]    87.657     16.918   5.181 8.04e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 52.05 on 37 degrees of freedom
Multiple R-squared: 0.7385, Adjusted R-squared: 0.7244
F-statistic: 52.26 on 2 and 37 DF,  p-value: 1.667e-11
```

Compare Forecasts With Observations

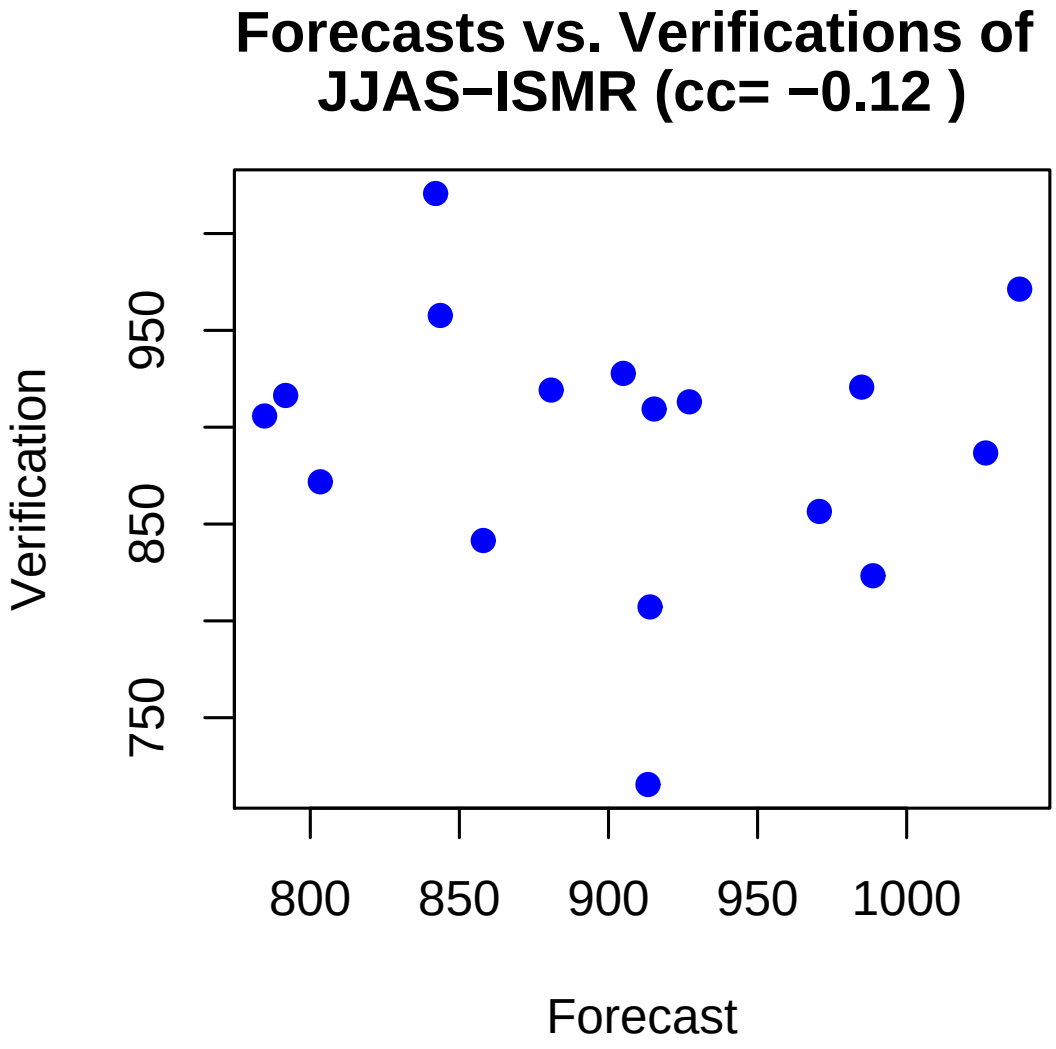


What if the Model is Cross Validated?



What if the Model is Tested on Independent Data?

Withheld years: 1990 - 2006



How Can Cross-Validation Be So Inconsistent From Independent Verification?

How Can Cross-Validation Be So Inconsistent From Independent Verification?

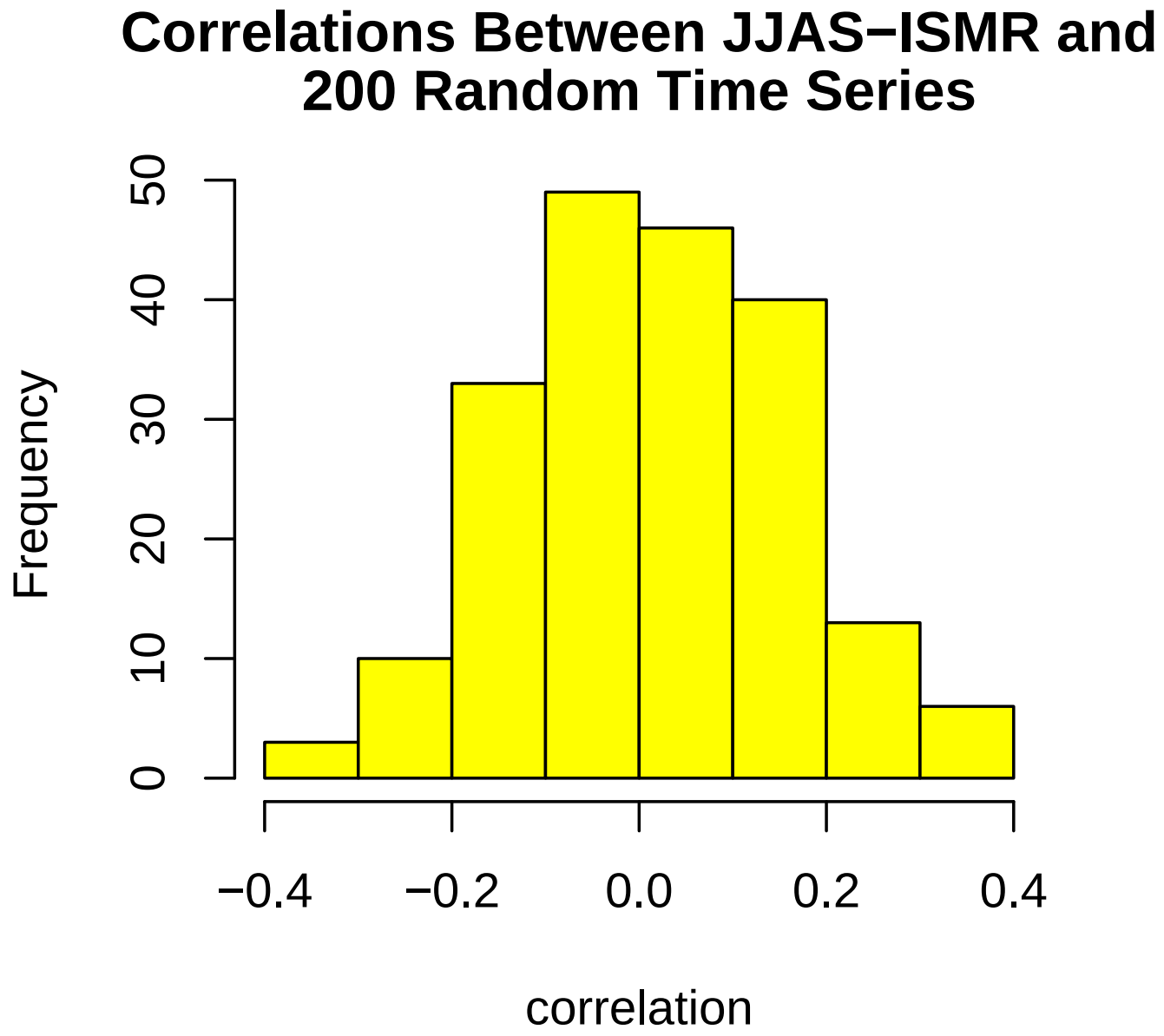
The correlation maps were generated using all the data.

This means the predictors were selected based on all data.

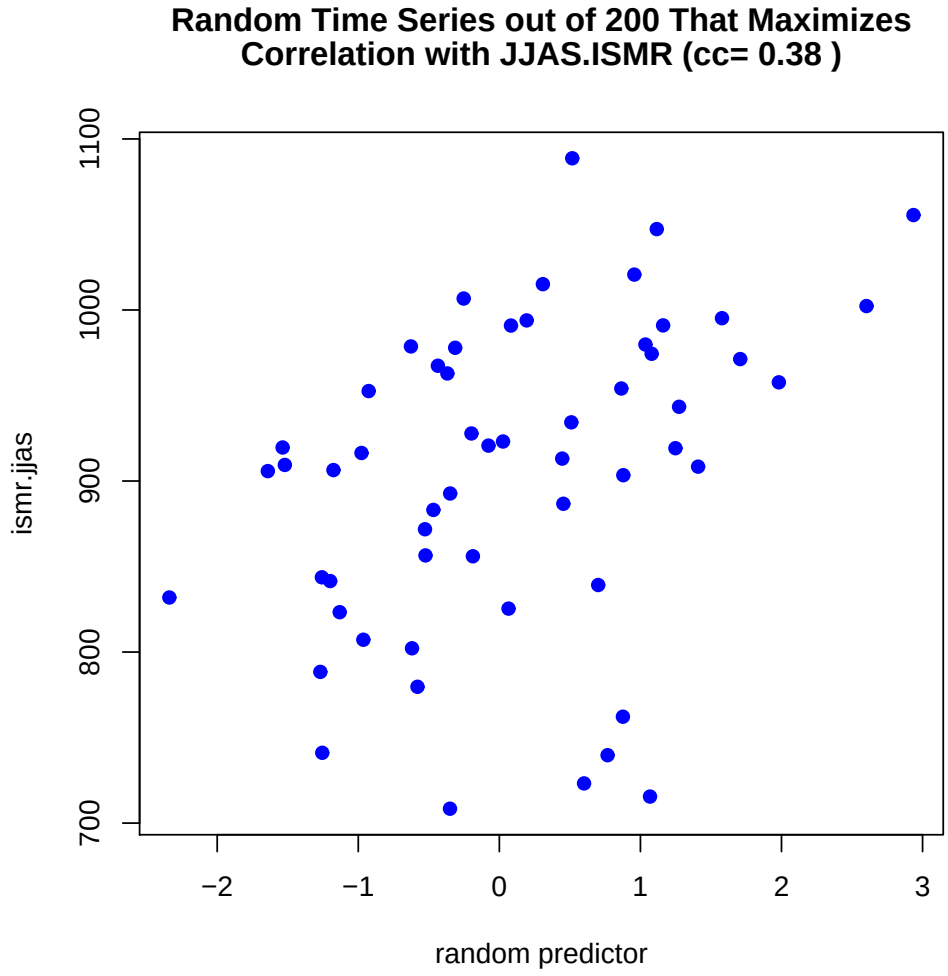
This is not true cross-validation– data from withheld year should not be used for any purpose, especially for choosing predictors.

If the screening procedure is not taken into account in significance testing or cross validation, then the skill estimate is biased.

Example With Random Data

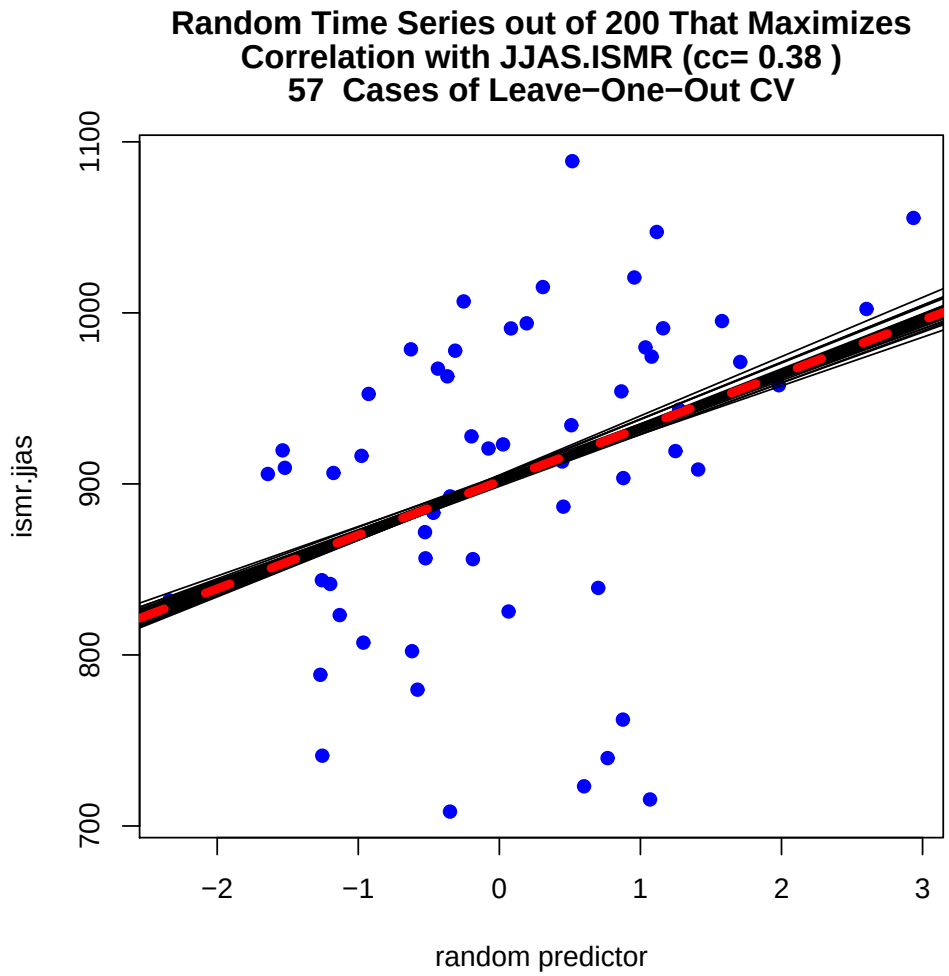


Example With Random Data



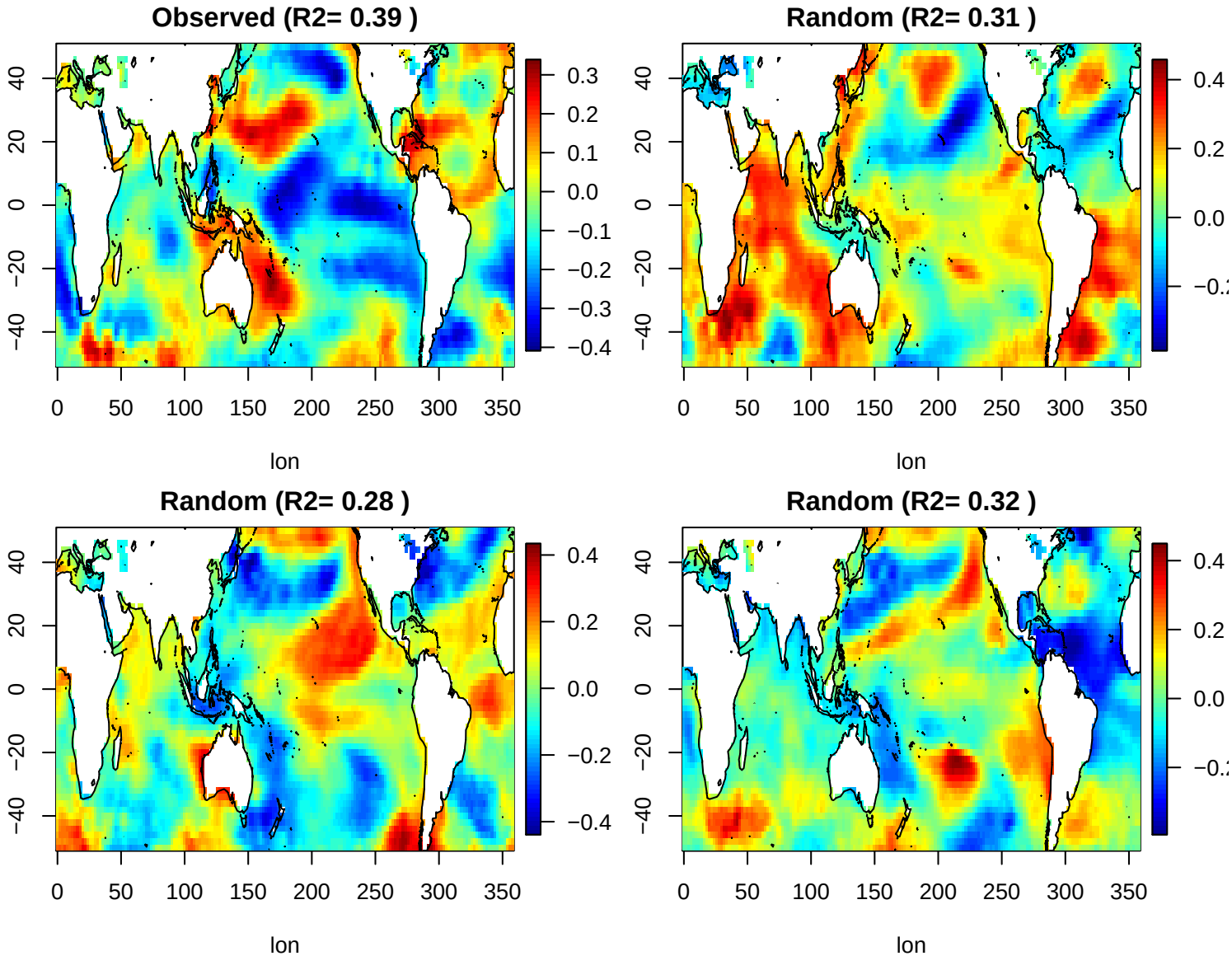
Now perform cross validation on THIS time series.

Example With Random Data



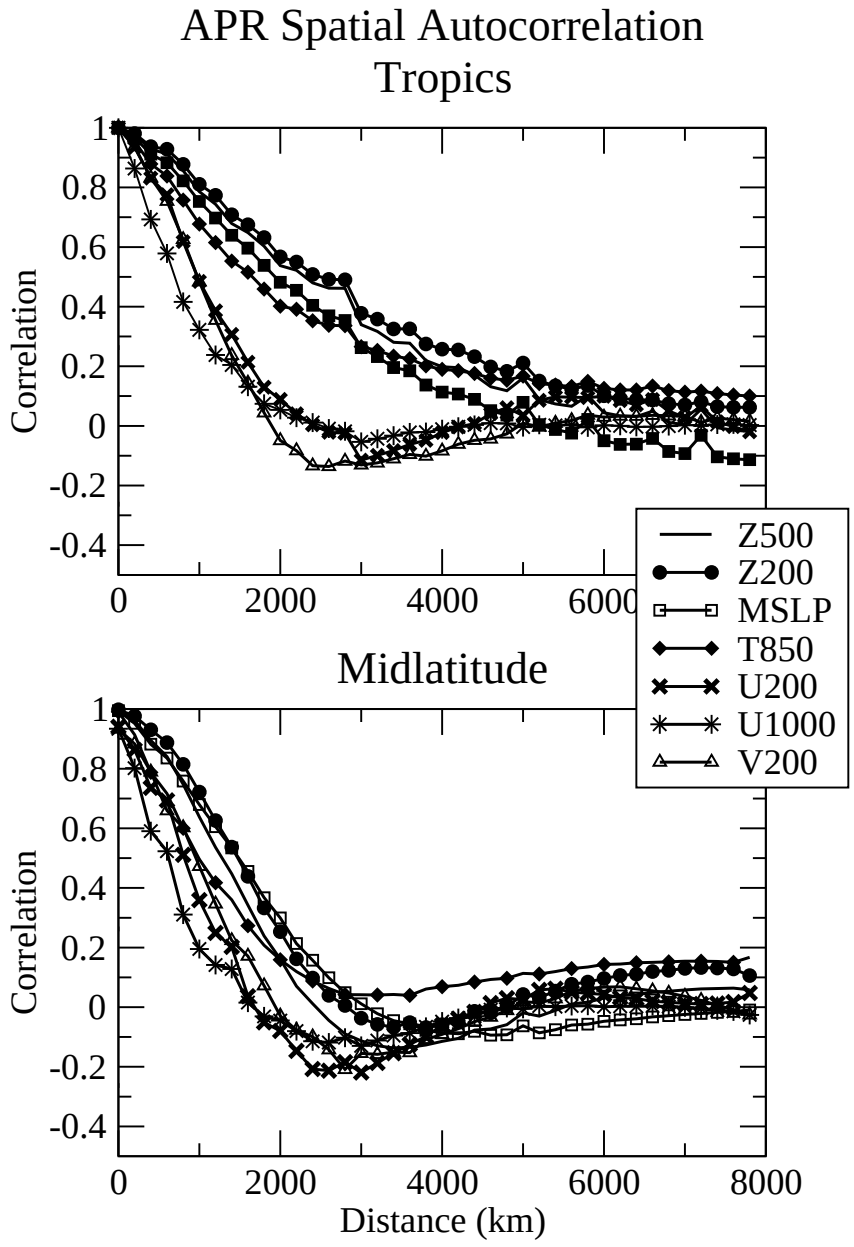
If predictor is selected because it is strongly correlated with the prediction variable, then leave-one-out cross-validation also is biased toward strong correlation.

Random Selection



SST field selected randomly and then correlated with JJAS-ISMR.

How Many Independent Time Series Are There?



How Many Independent Time Series Are There?

Table: Length scale and effective number of spatial degrees of freedom (dof) for April mean fields selected from the NCEP-NCAR Reanalysis.

	Midlatitude		Tropics		Total
Variable	Length (km)	dof	Length (km)	dof	dof
500-hPa geopotential height	1900	22	3800	6	28
200-hPa geopotential height	2200	17	4200	5	21
Mean sea level pressure	2100	18	2900	10	28
850-hPa temperature	1600	32	2700	11	43
200-hPa zonal wind	1200	56	1500	36	92
1000-hPa zonal wind	800	127	1000	81	208
200-hPa meridional wind	1300	48	1500	36	84
Total degrees of freedom		321		184	505

Comparison of Operational and Random Forecasts

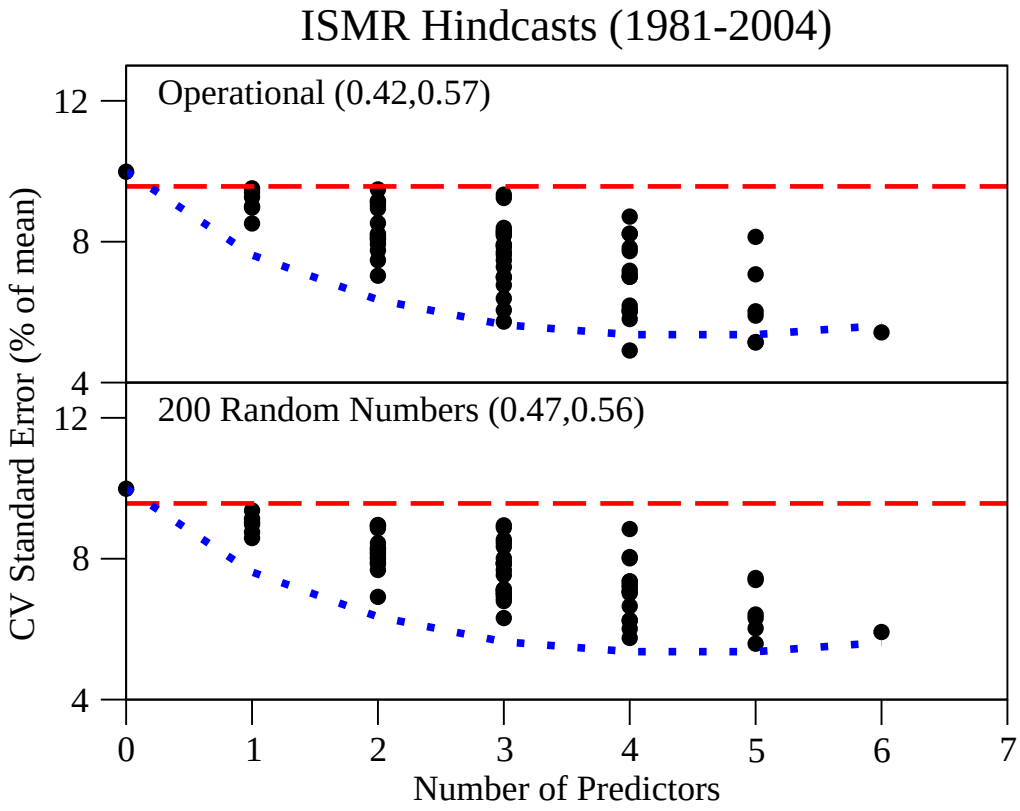


Figure: CV standard error of ISMR forecasts using the actual predictors of the May 2006 IMD operational forecast, and (bottom) a regression model derived from the 6 out of 200 random time series that are most correlated with ISMR. The dotted curve gives the 5% significance threshold when screening is taken into account.

Comparison of Operational with Other Forecasts

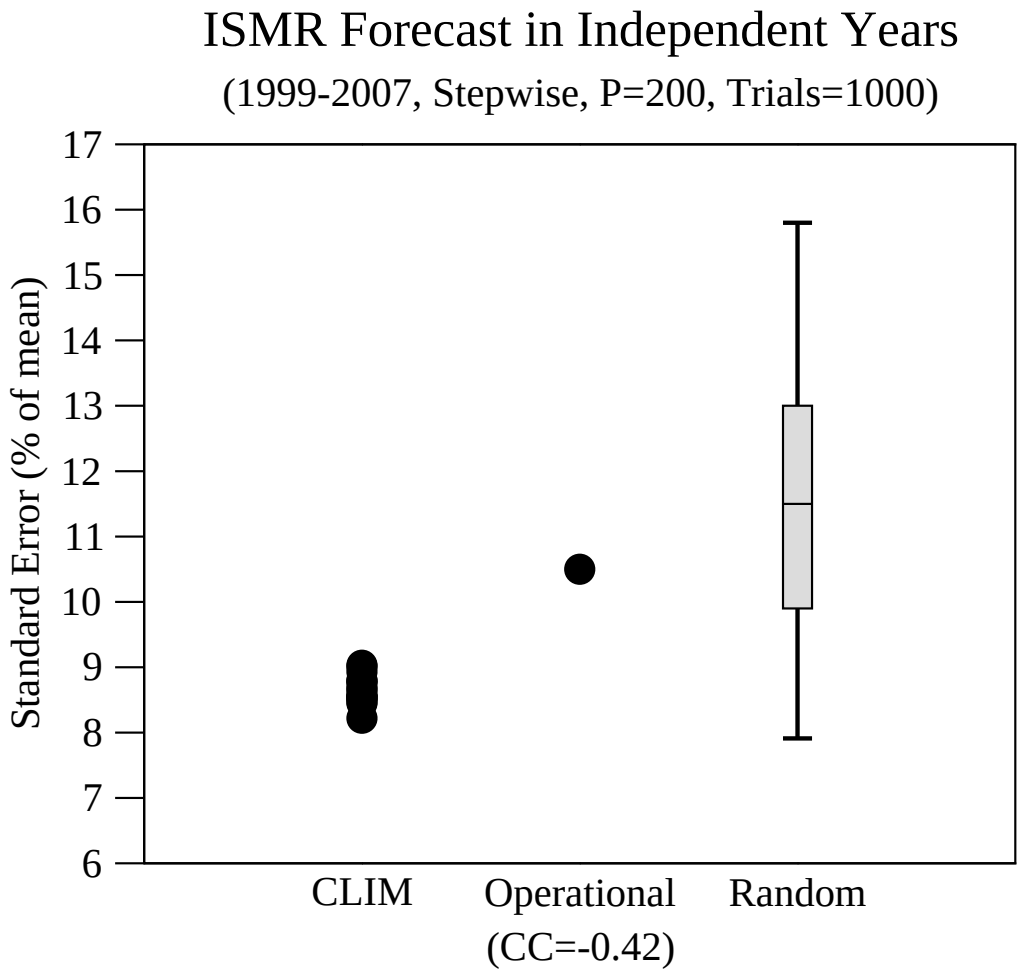


Figure: Standard error of forecasts of JJAS All-India Rainfall by: the climatology of the preceding data, from 15-30 years ("CLIM"); the IMD ("Operational"); models with random predictors, as selected by the stepwise regression procedure, using the predictand only from the preceding 30 years ("Random"). The box-plot shows the median (center-line in rectangle), the first and last quartile (as the ends of the rectangle), and 5% and 95% values as error bars

Summary

- Identifying predictors by screening biases forecast skill.
- This bias exists even with cross validation.
- The skill of some operational forecasts are consistent with a no-skill forecast when this bias is taken into account.

How Should Predictors Be Chosen?

There is no general answer to this question.

- If you search many predictors, you must account for the resulting bias in your forecast assessment.
- Constrain the predictors based on independent principles (e.g., physics or dynamical model results).
- Include all predictors, but apply shrinkage methods (e.g., ridge regression).
- Restrict the number of predictors a priori (e.g., say 1/10 of the number of independent samples).