Statistical Modelling
Potholes

Willem A. Landman & Tony Barnston

Lecture 10

Finley's Tornado Predictions

	O_{yes}	O_{no}
F_{yes}	28	72
F_{no}	23	2680

- Hite Rate = proportion correct
- HR = (28+2680)/2803
  = 0.966

"No Tornados" Always Predicted

	O_{yes}	O_{no}
F_{yes}	0	0
F_{no}	51	2752

- Hite Rate = proportion correct
- HR = (0+2752)/2803
  = 0.982

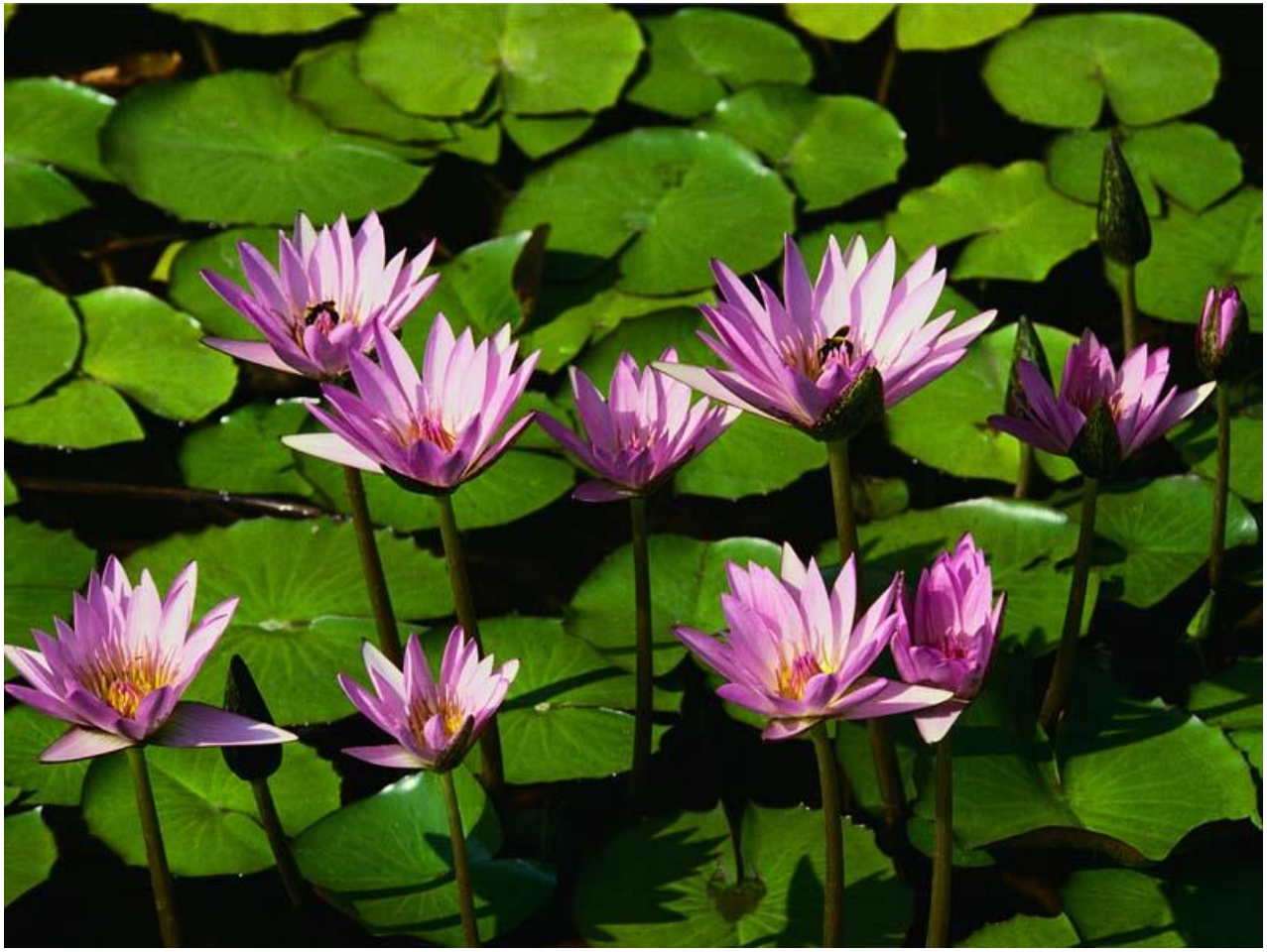
Finley lessons

- Finley's scheme has the advantage that it predicted more than half of the tornado cases successfully
- 'No tornados' scheme never did
- ***Hit rate*** may not be the best way to summarize the value of this forecast scheme
- Because hits in upper left box ($\mathbf{F}_{yes}$; $\mathbf{O}_{yes}$) are extremely crucial – hit rate misses the point!

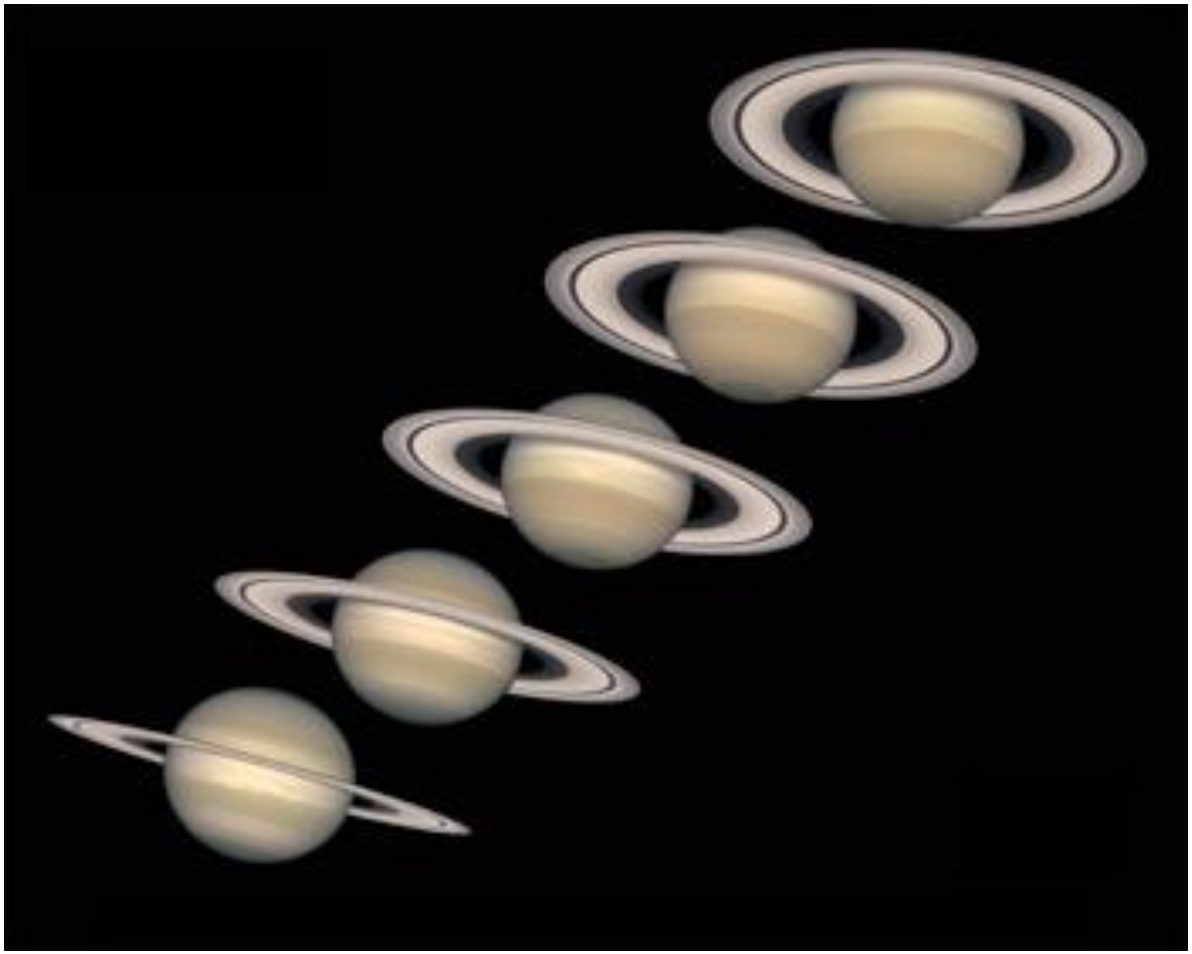
Predictor Selection

- Many potential predictors available
- Do not simply add potential predictors to the regression
- Dangers in too many predictors in forecast equation
- Collinearity (more detail later…)

A useful predictor…?

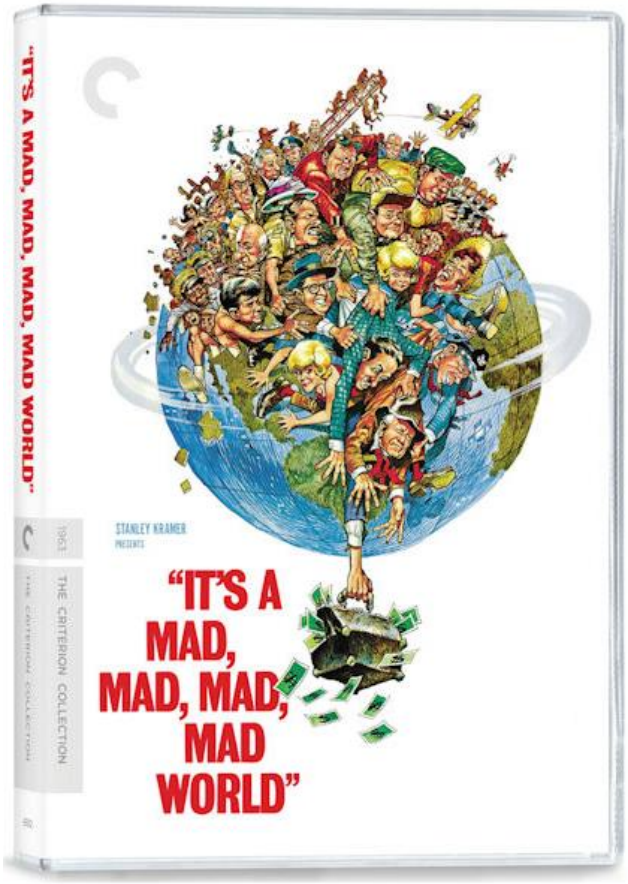


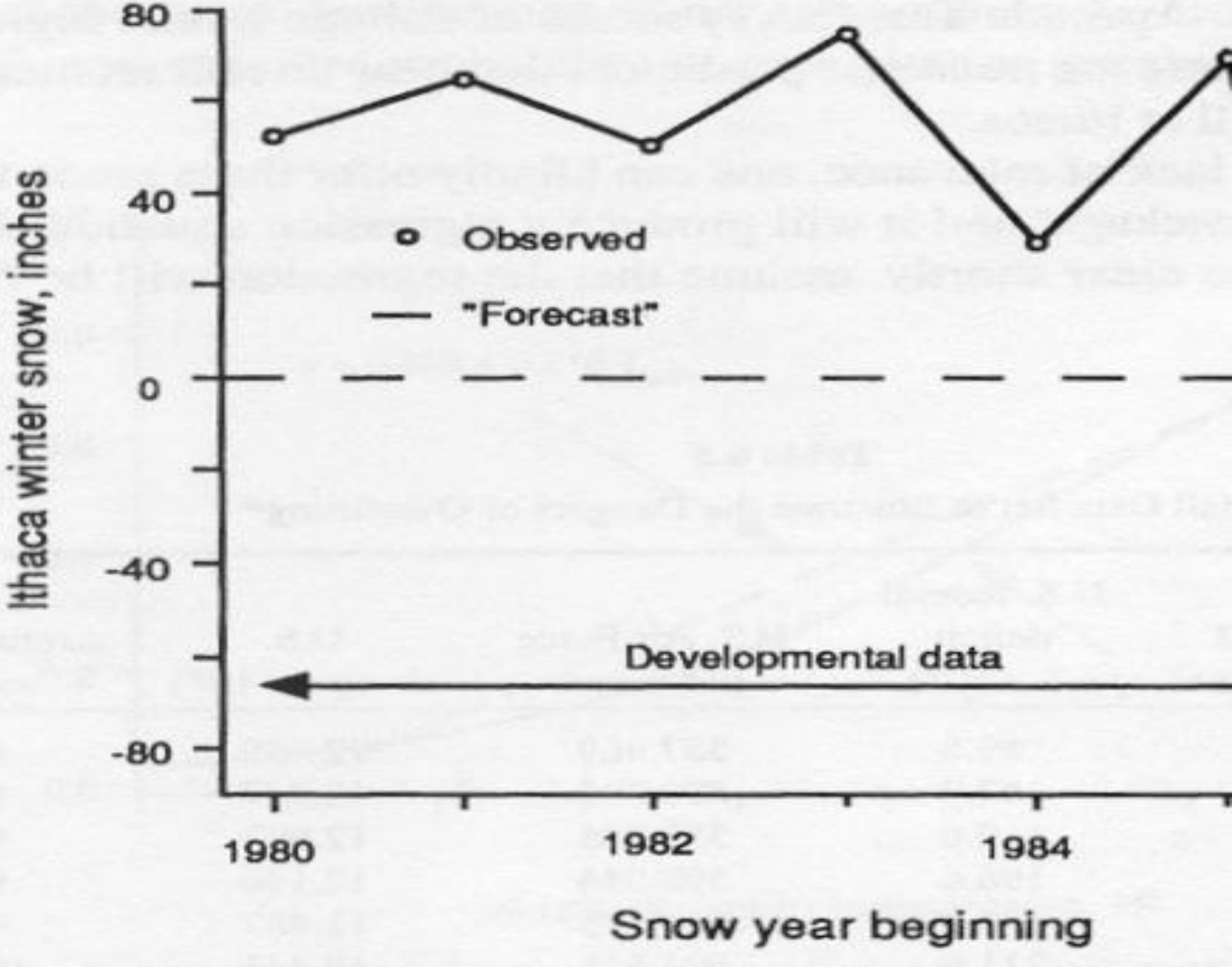
…what about celestial objects as predictor…?

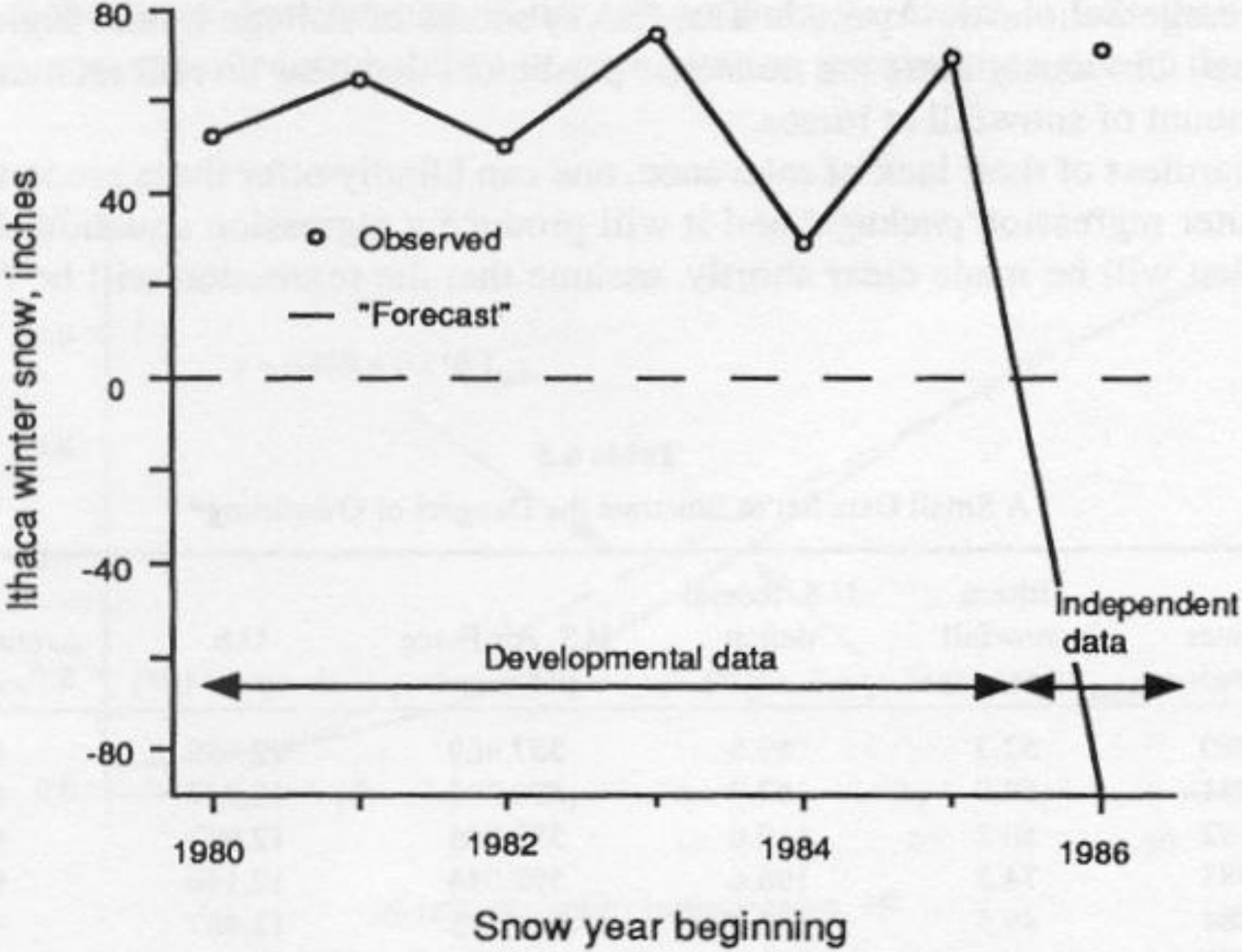


It's a Mad, Mad, Mad, Mad World…

- Predictand: Snowfall (inches)
- Predictors (!!!!)
  - USA Federal deficit
  - US Air Force personnel
  - US sheep




80
40
0
-40
-80
Ithaca winter snow, inches
Observed
"Forecast"
Developmental data
1980
1982
1984
Snow year beginning


80
40
0
-40
-80
Ithaca winter snow, inches
Observed
"Forecast"
Developmental data
Independent data
1980
1982
1984
1986
Snow year beginning

Lessons from MMMMW example:

- Choose only physically reasonable or meaningful potential predictors
- Test prediction equations on sample of data not involved in its development
- Large skill difference between dependent and independent samples may suggest overfitting
- Large independent sample necessary to ensure stability of regression
- Small sample size → chance of sampling error

Also be aware of…

- ***Autocorrelation*** in predictand: might have to exclude additional adjacent years in cross-validation process

- Predicting a value that is ***contained*** in the training period of an empirical model (cross-validation: the value that is predicted, is ***omitted*** from the training period, which is the preferred method)

Cross-Validation

	Year1	Year 2	Year 3	Year 4	Year 5	Year 6
Model 1	omitted					
Model 2		omitted				
Model 3			omitted			
Model 4				omitted		
Model 5					omitted	
Model 6						omitted
Model 7						

AVOID this model!!!

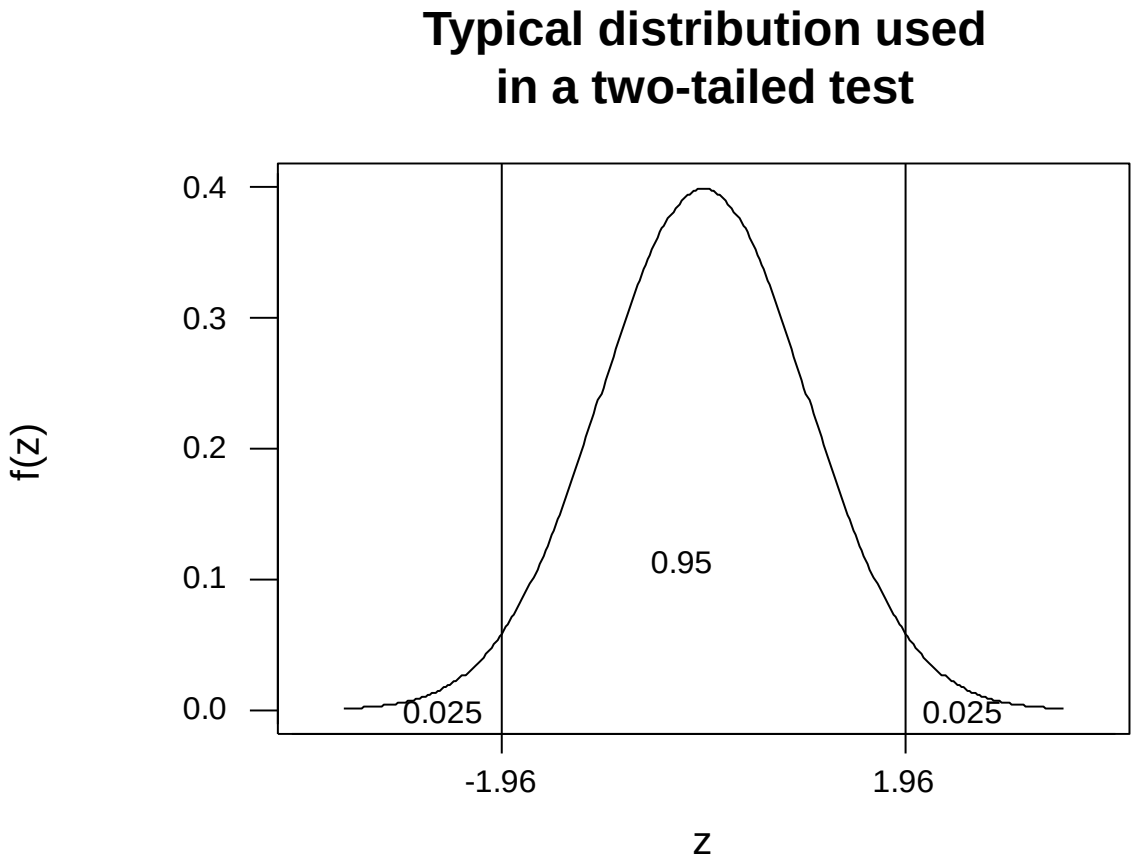
What is Autocorrelation?

- A series of numbers set besides itself will have correlation of 1
- Shifting the series upward or downward by one value, each value paired with preceding value
- Correlating these *lagged* values determines if dependencies exist among successive values – correlation value referred to as autocorrelation
- Effective sample size decreases by 1 for each lag
- No autocorrelation: series of observations are independent of one another

One-Tailed vs. Two-Tailed Tests

- A statistical test can be either one-tailed (-sided) or two-tailed (-sided)
- Probabilities on the tails of the distribution govern whether a test result is significant or not
- Whether a test is one- or two-tailed depends on the nature of the hypothesis being tested:
  - just interested in positive correlations: one-tailed (i.e., skill of a forecast time series)
  - interested in both positive and negative correlations: two-tailed (i.e., association between any two time series)

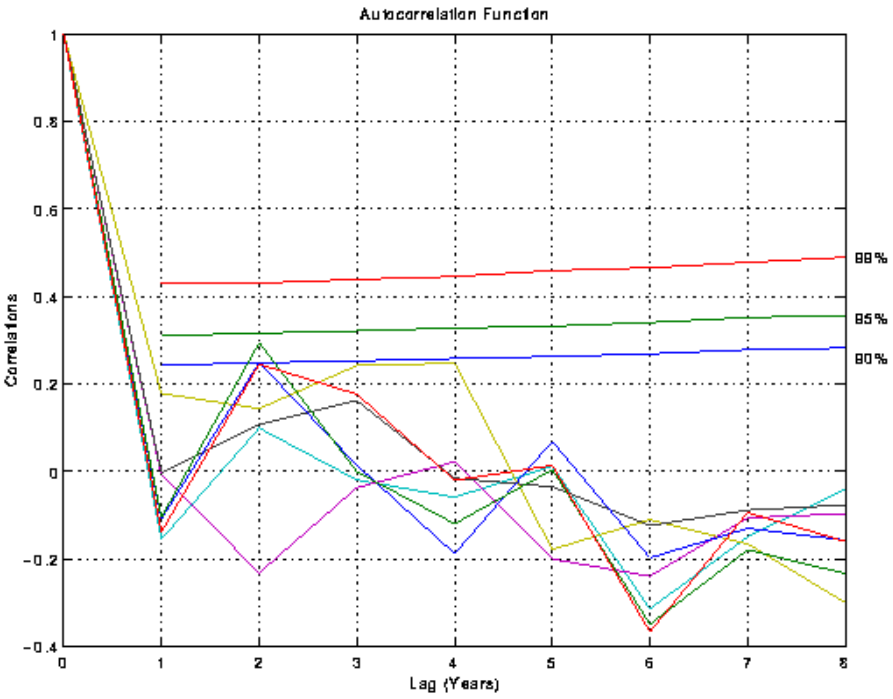
Probabilities for N(0,1) – 95% interval

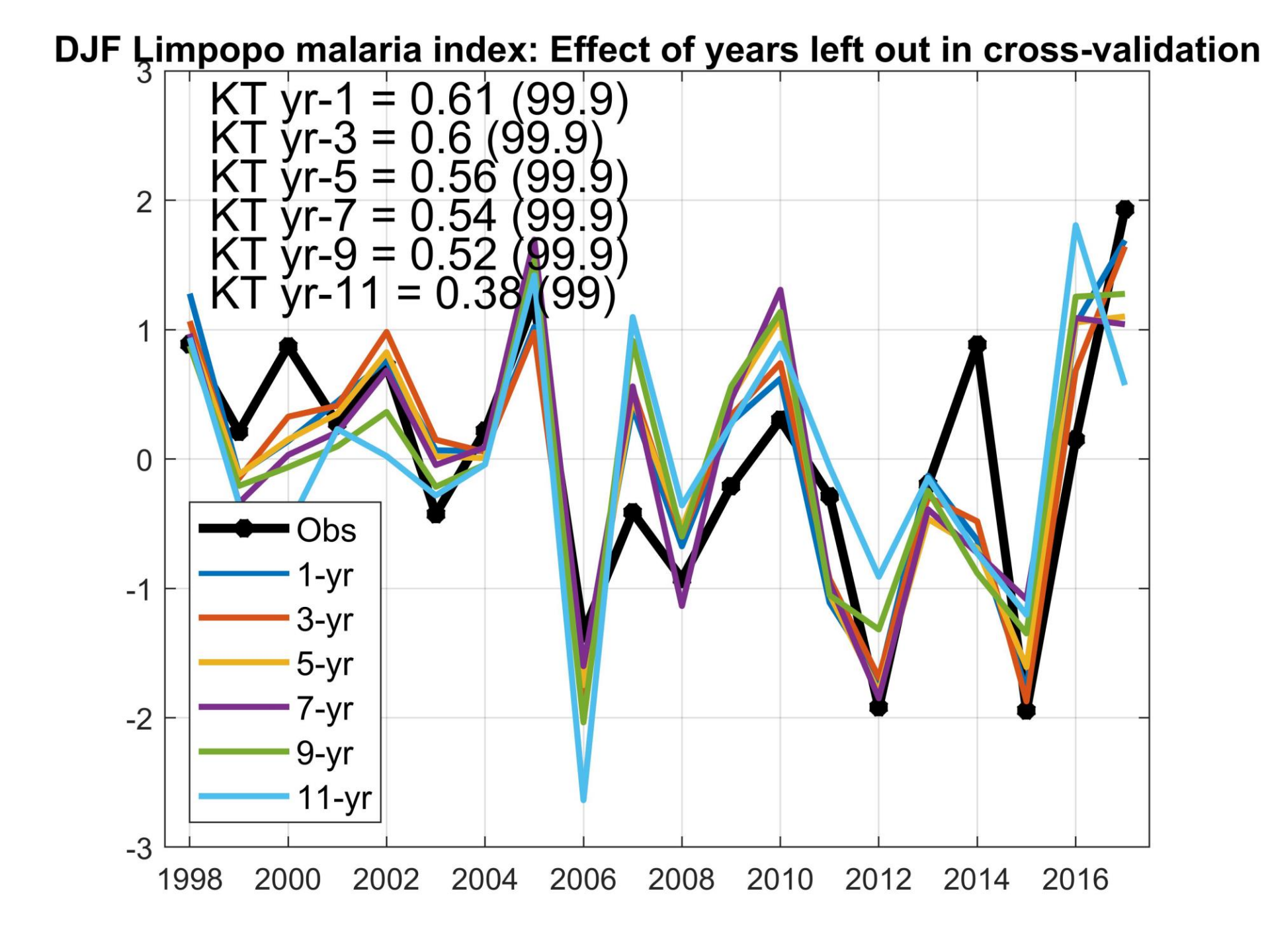


Numbers indicate areas under curve, left of -1.96, right of 1.96, and between.

Autocorrelation continued

- At Lag = 6, some high negative correlations seen
- Since only interested in positive autocorrelation, negative values can be discarded (1-tailed test)
- The significance thresholds (sloping lines) are calculated for varying sample size – critical level increases with decreasing sample size

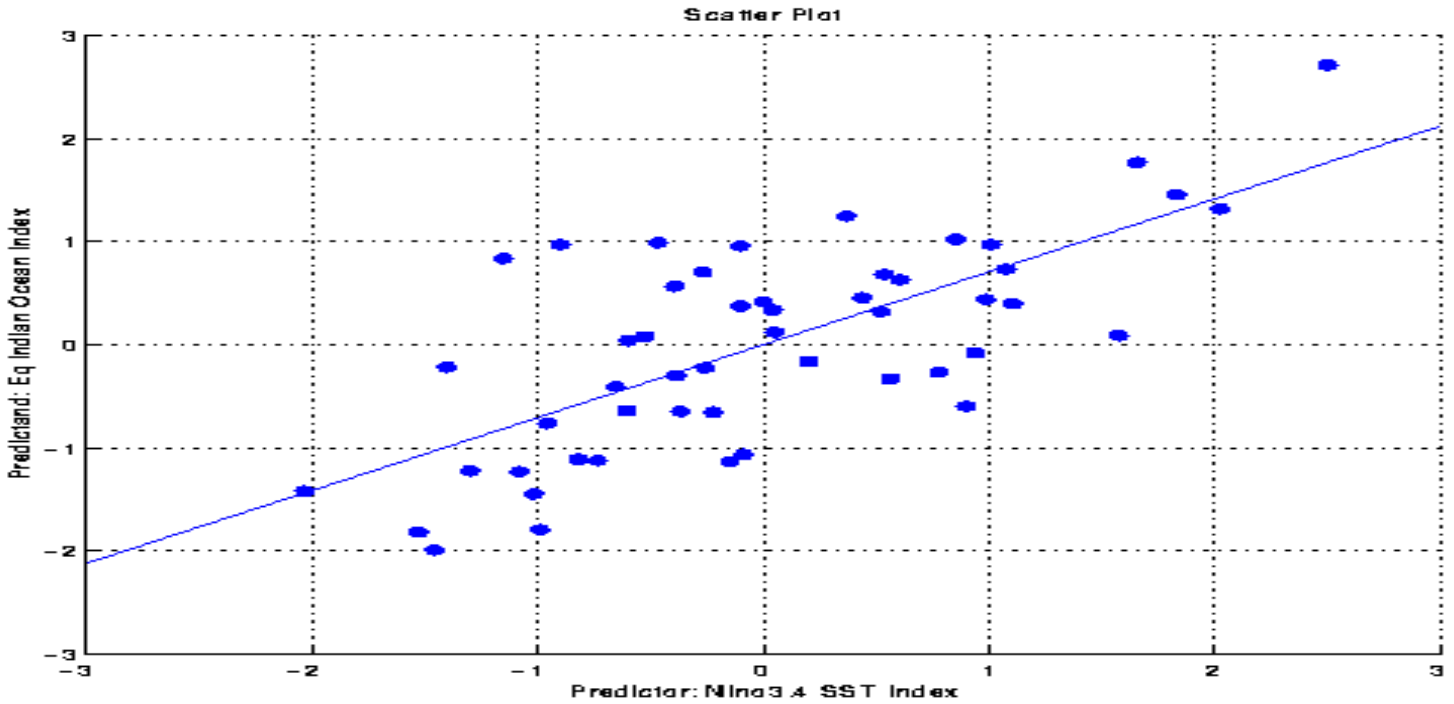



DJF Limpopo malaria index: Effect of years left out in cross-validation
KT yr-1 = 0.61 (99.9)
KT yr-3 = 0.6 (99.9)
KT yr-5 = 0.56 (99.9)
KT yr-7 = 0.54 (99.9)
KT yr-9 = 0.52 (99.9)
KT yr-11 = 0.38 (99)
Obs
1-yr
3-yr
5-yr
7-yr
9-yr
11-yr
3
2
1
0
-1
-2
-3
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016

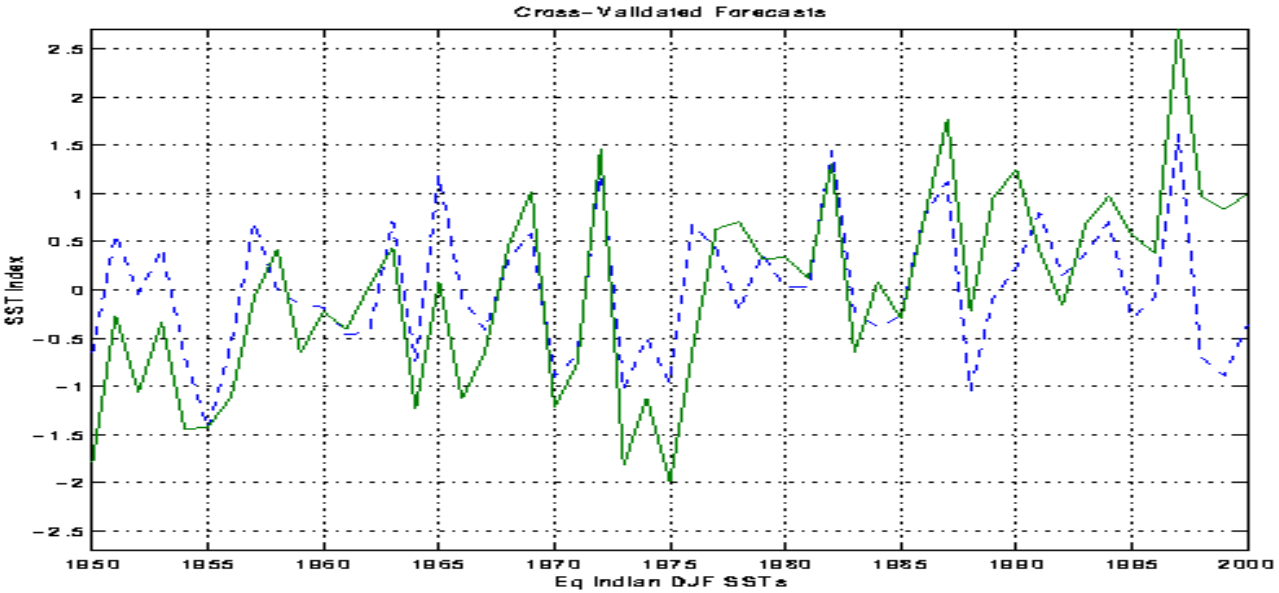
Variance Adjustment

- Least-squares statistical models minimize squared errors, not the absolute value of actual errors – damping is caused
- Observed variance (σ_o) is subsequently underestimated (perpetual near-normal forecasts may result)
- However, other regression formulas called LAD (least absolute deviation) are based on the absolute value of actual errors – damping much less severe
- For least-squares methods, one should try to raise the σ_f to σ_o
- Here, $\hat{y}_{va} = \hat{y}/CVcorr$

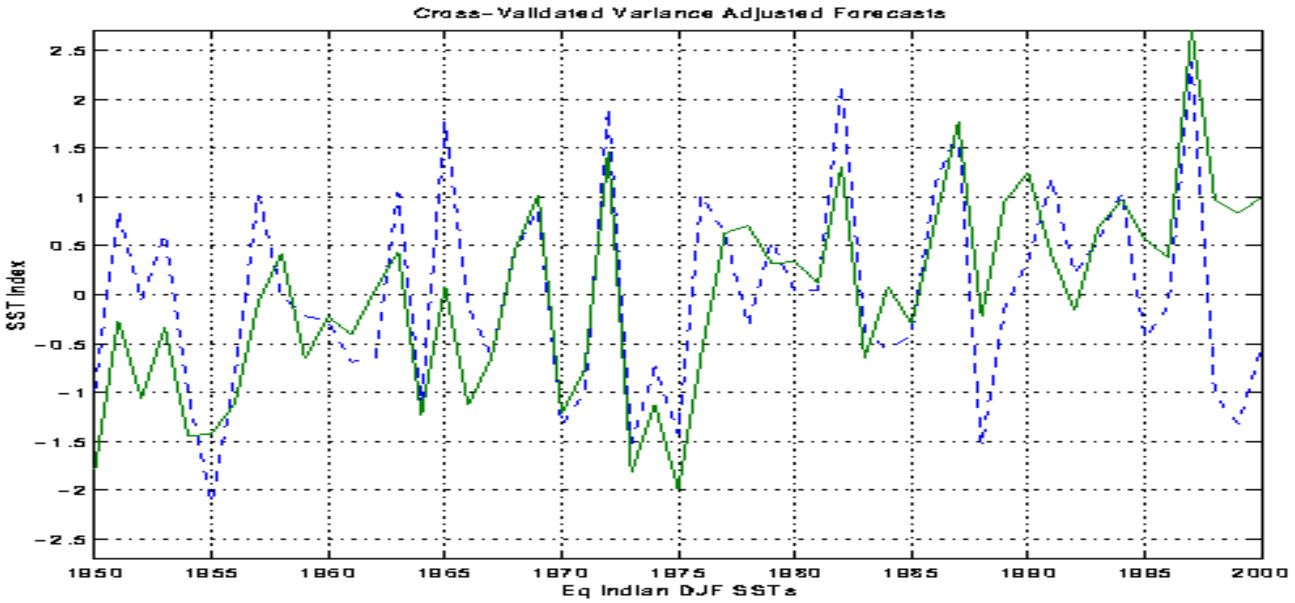
A simple Indian Ocean forecast model – scatter plot of Nino3.4 and equatorial IO:

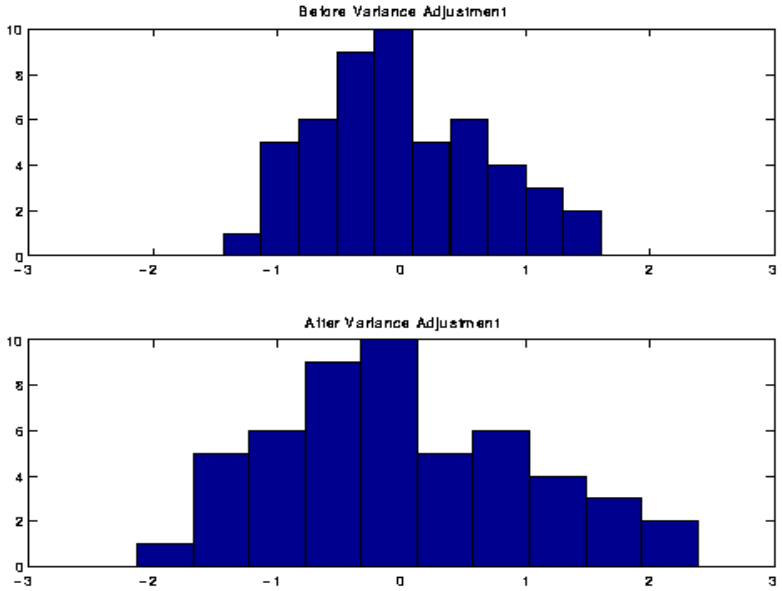


Cross-Validated Forecast of equatorial IO:

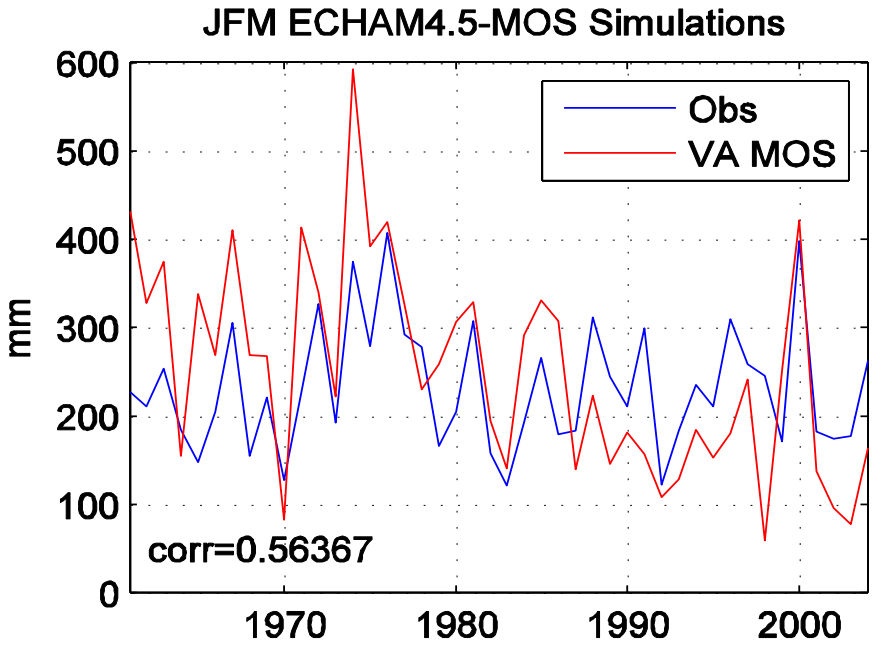


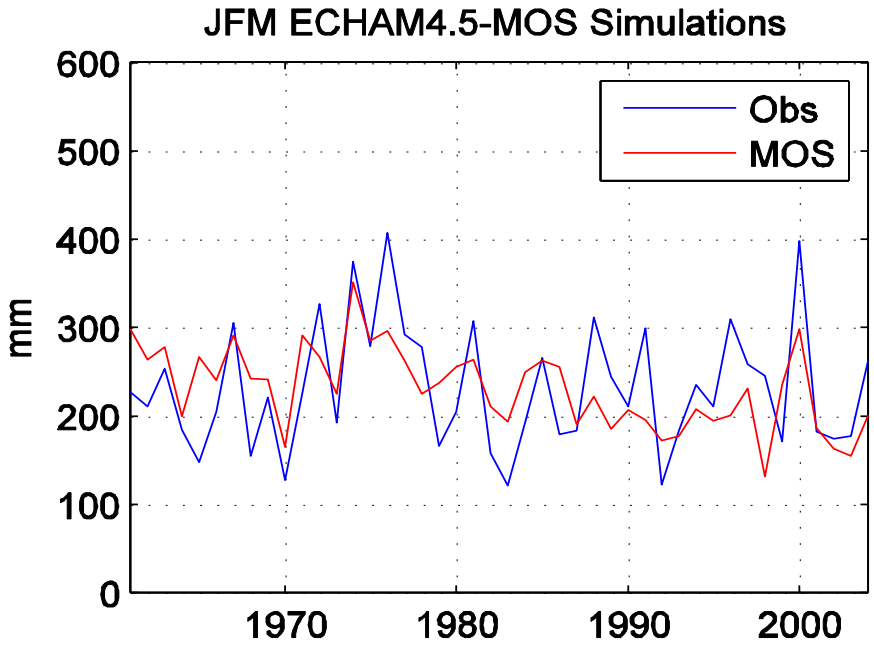
Cross-Validated and **Variance Adjusted**

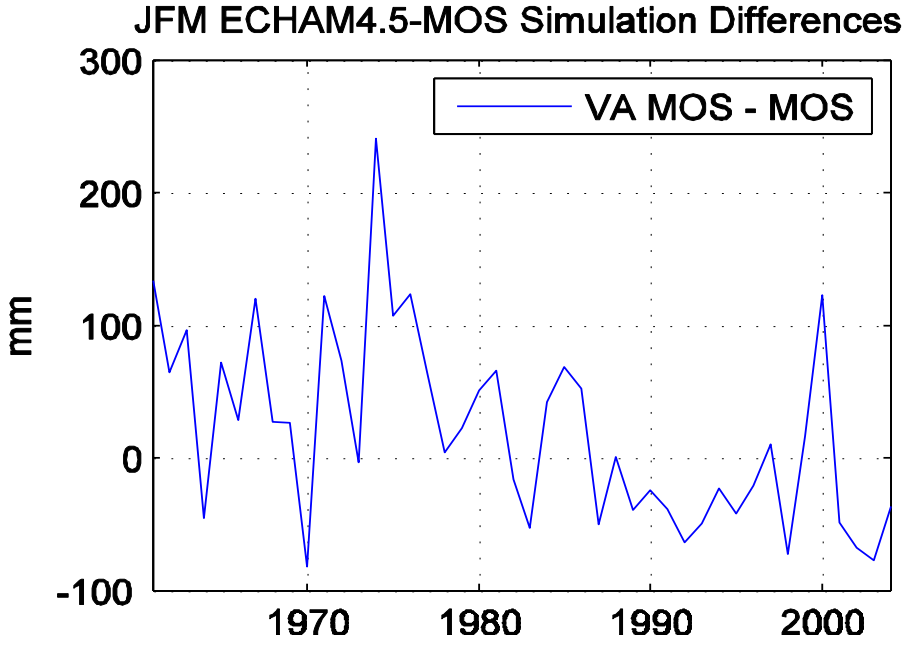


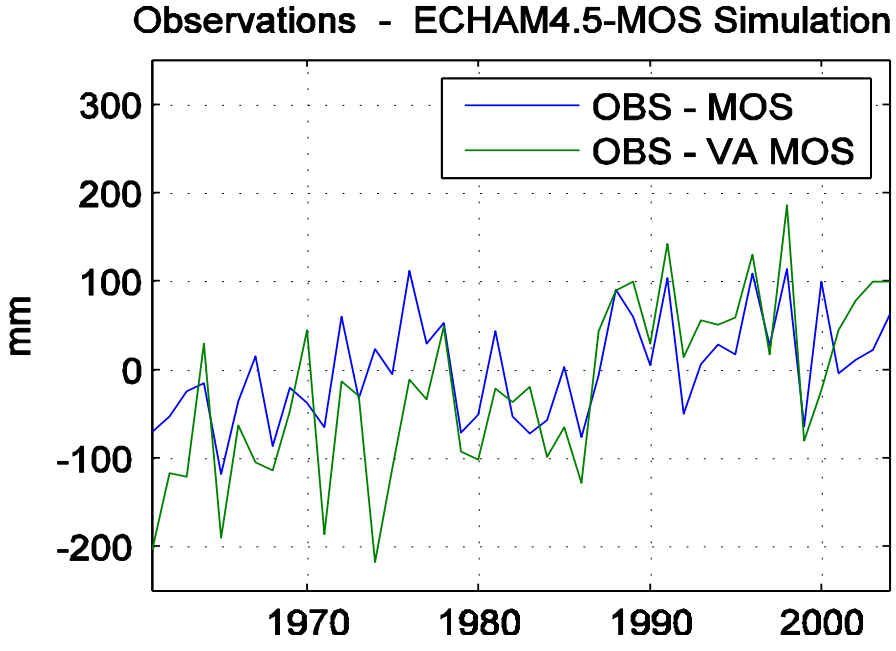


- Histograms of forecast equatorial Indian Ocean SST indices before and after variance adjustment
- The same number of bins, i.e. 10, is used
- A larger number of extremes is found after variance adjustment

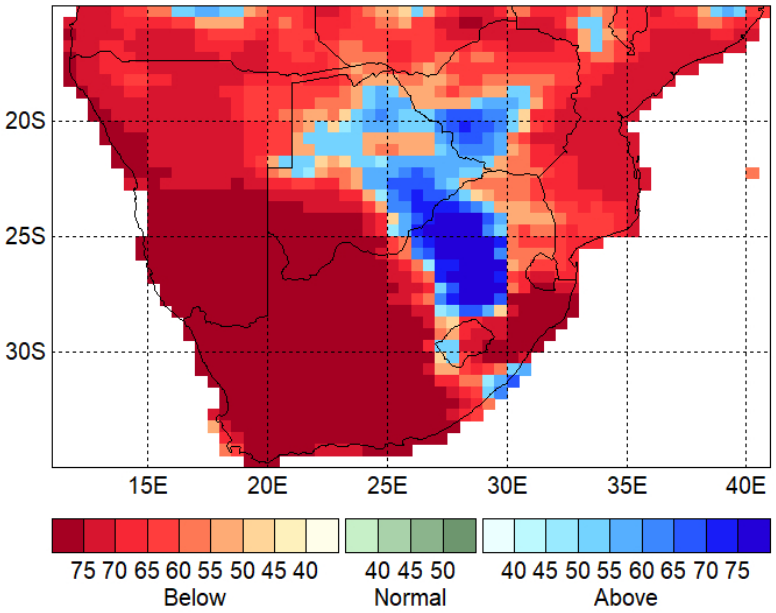

JFM ECHAM4.5-MOS Simulations
Obs
VA MOS
mm
corr=0.56367
600
500
400
300
200
100
0
1970
1980
1990
2000


JFM ECHAM4.5-MOS Simulations
Obs
MOS
mm
600
500
400
300
200
100
0
1970
1980
1990
2000

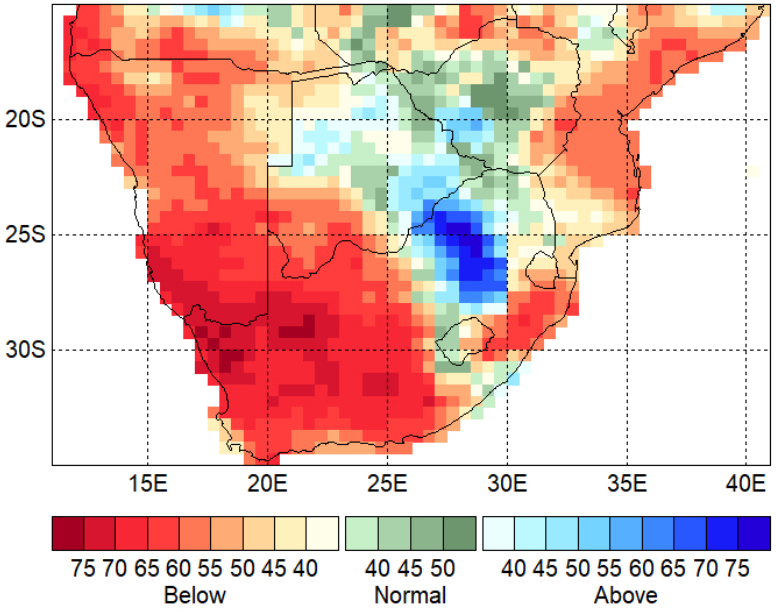

JFM ECHAM4.5-MOS Simulation Differences
VA MOS - MOS
mm
300
200
100
0
-100
1970
1980
1990
2000


Observations - ECHAM4.5-MOS Simulation
OBS - MOS
OBS - VA MOS
mm
300
200
100
0
-100
-200
1970
1980
1990
2000

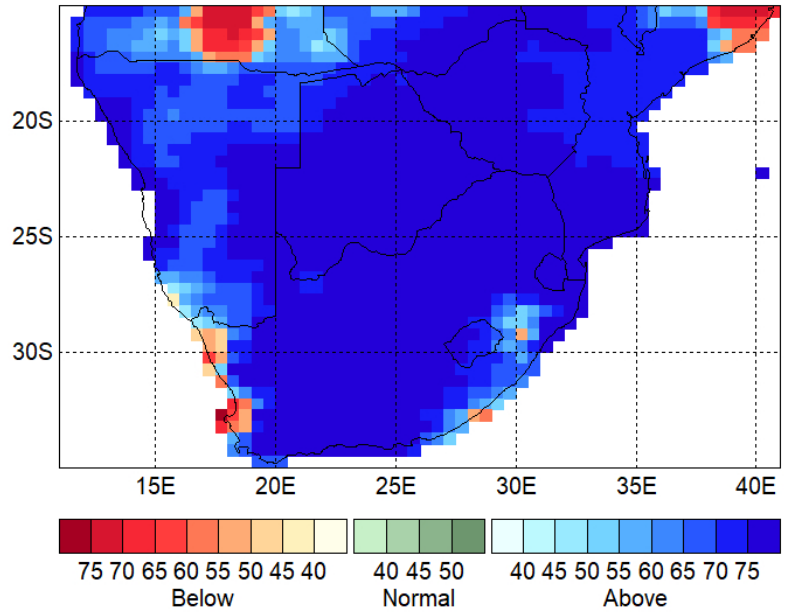
DJF 2015/16 Precip; ICs: Nov; Mean bias



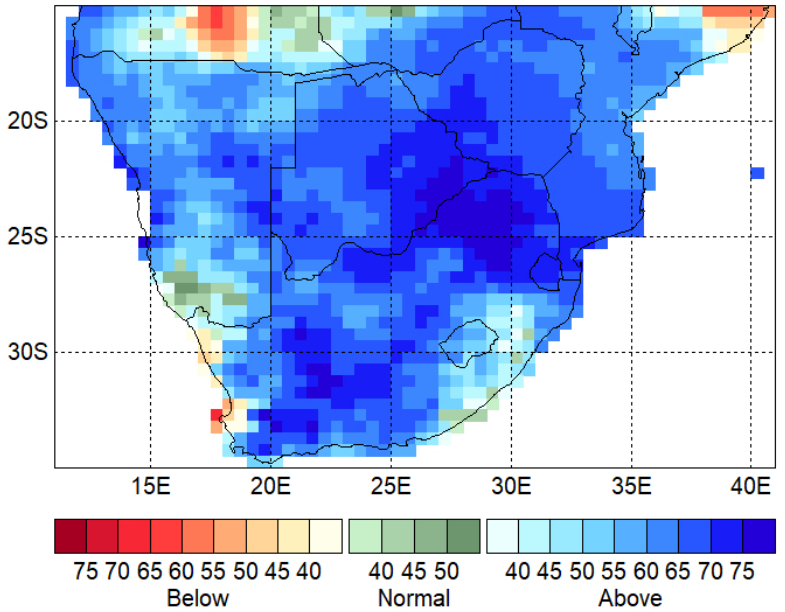
DJF 2015/16 Precip; ICs: Nov; Mean+Variance



DJF 2016/17 Precip; ICs: Nov; Mean bias



DJF 2016/17 Precip; ICs: Nov; Mean+Variance



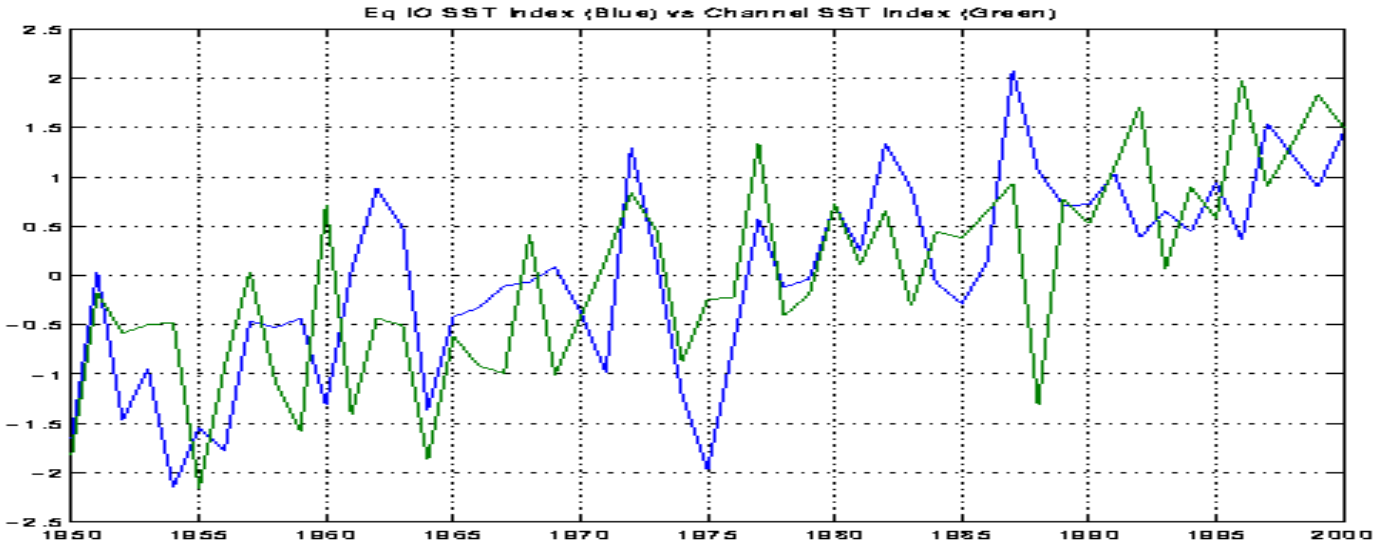
Pros and Cons of Variance Adjustment

- PROS:
  - Forecasts' variance similar to observed
  - High amplitude events are better captured if model skill is not low
- CON:
  - Large forecast discrepancies are further magnified

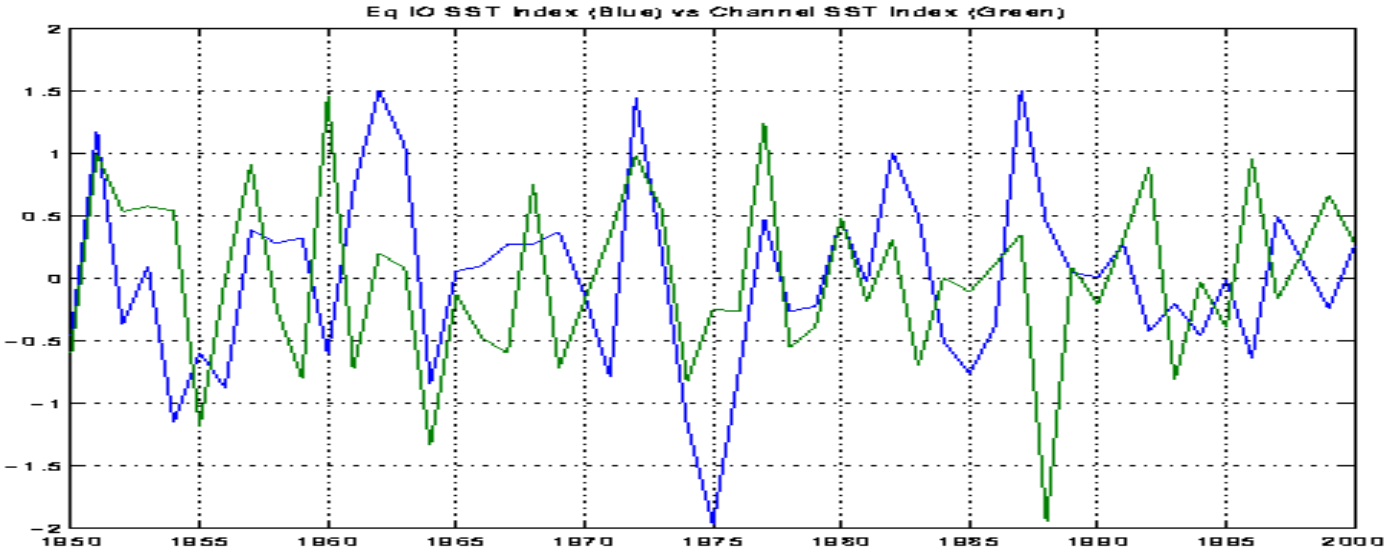
Is the linear correlation between two variables telling us everything we need to know…?

- Strong but ***nonlinear*** relationships between two variables may not be recognized
- Correlation coefficient provides no explanation for the ***physical relationship*** between two variables (MMMMW example)
- What about ***trends*** in the data?

Trends in Predictor/Predictand Correlation = 0.5937



Detrended
Correlation = 0.1665



Confirmation of scientific theories

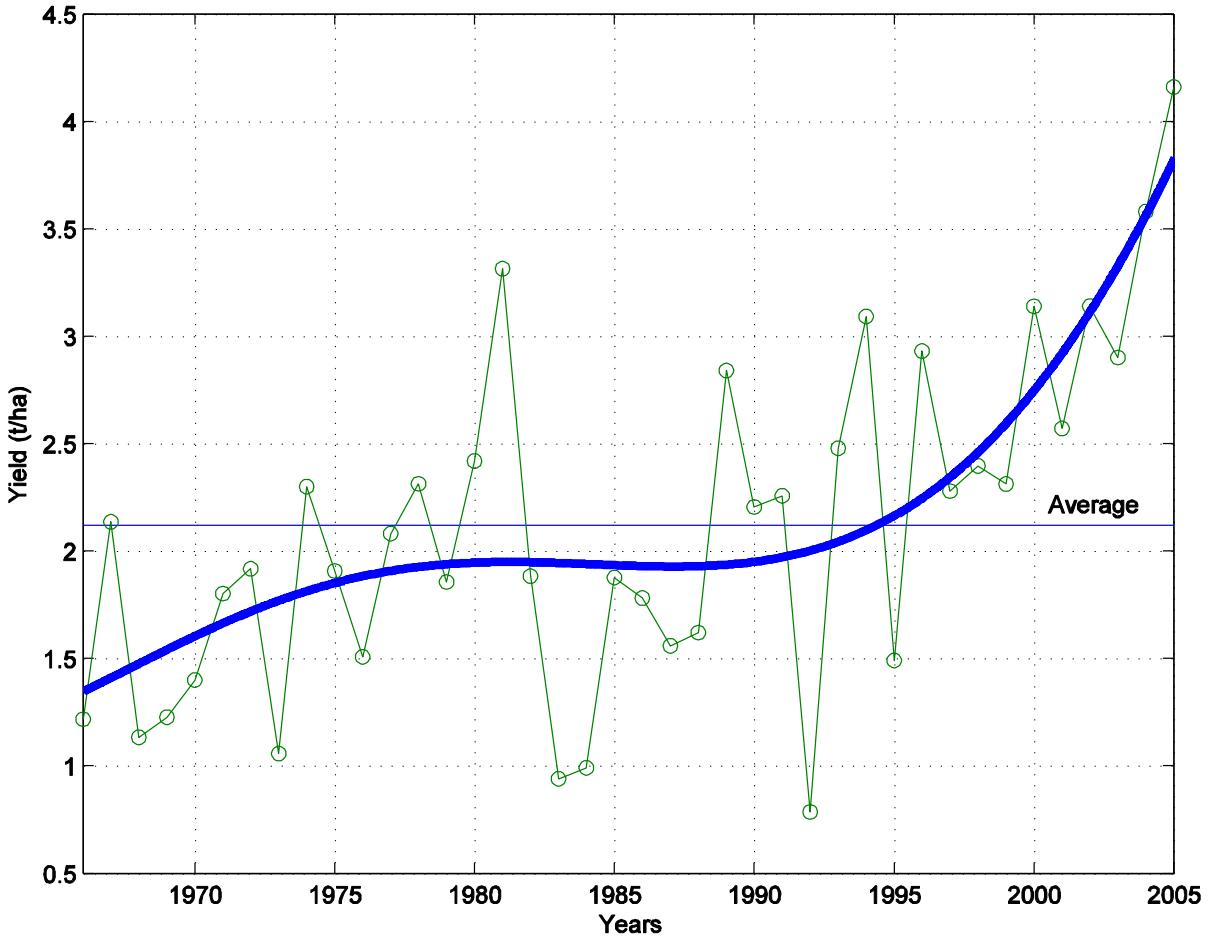
- Can we *never* have *any* grounds for supposing a scientific theory is true?

But there are grounds for supposing that certain scientific theories are true

- Assuming that scientific theories can be scientifically confirmed… under what circumstances are they *best* confirmed?

For a theory to be strongly confirmed it needs to make predictions that are both surprising and true

Is it surprising that a forecast of "above the average yield" was made for 2005?



Collinearity (1)

- Independent variables (the predictors):
  - They add more to the final prediction when they are not highly inter-correlated
  - If they are strongly correlated, then either one of them will do nearly as well as both together
  - If they are extremely highly correlated (e.g. >0.98), the regression scheme will bomb

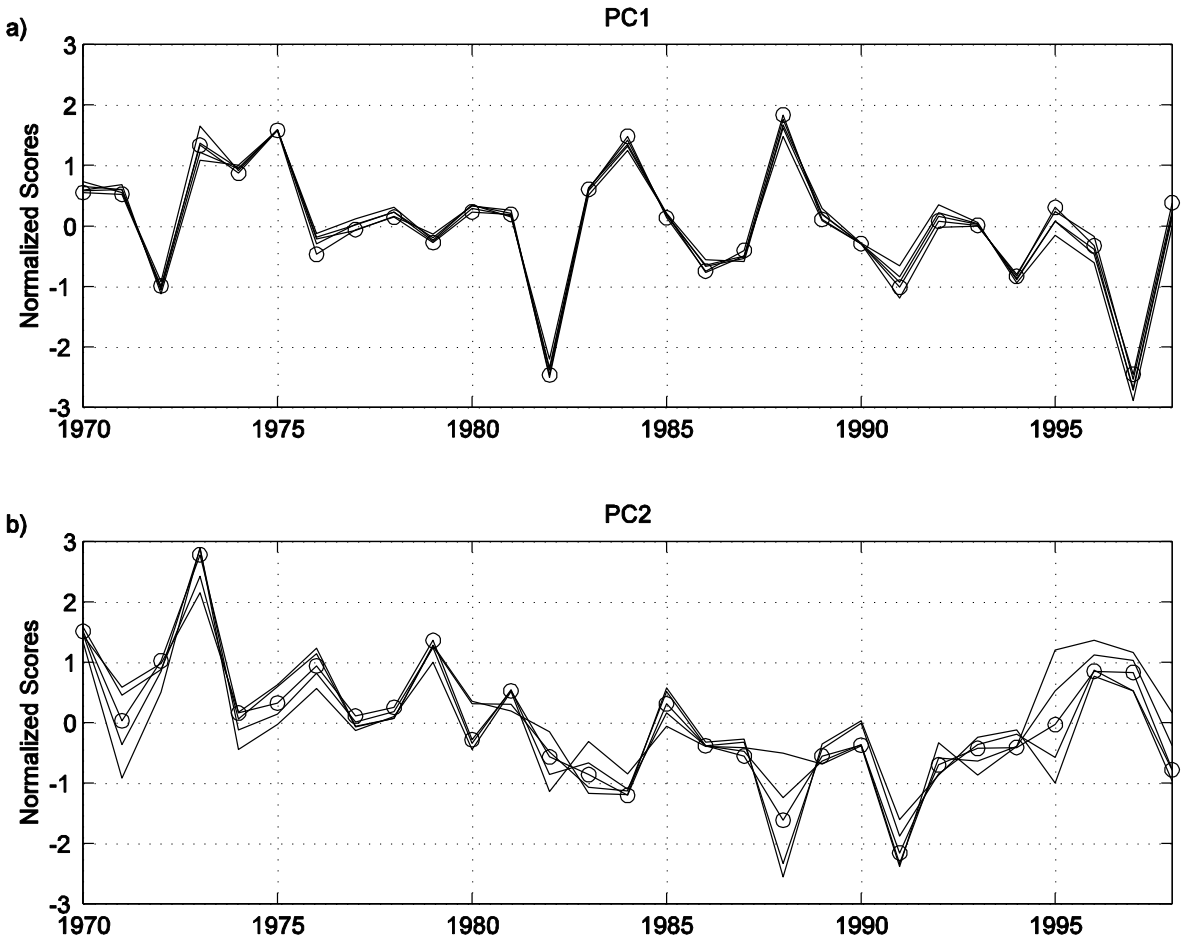
Collinearity (2)

- When the independent variables are *not correlated at all*:
  - The equation coefficients indicate the sign and strength of the relationship between the given independent variable and the predictand
- When independent variables are inter-correlated, the coefficients cannot be interpreted in a simple way - the role of each independent variable may be uncertain

Collinearity (3)

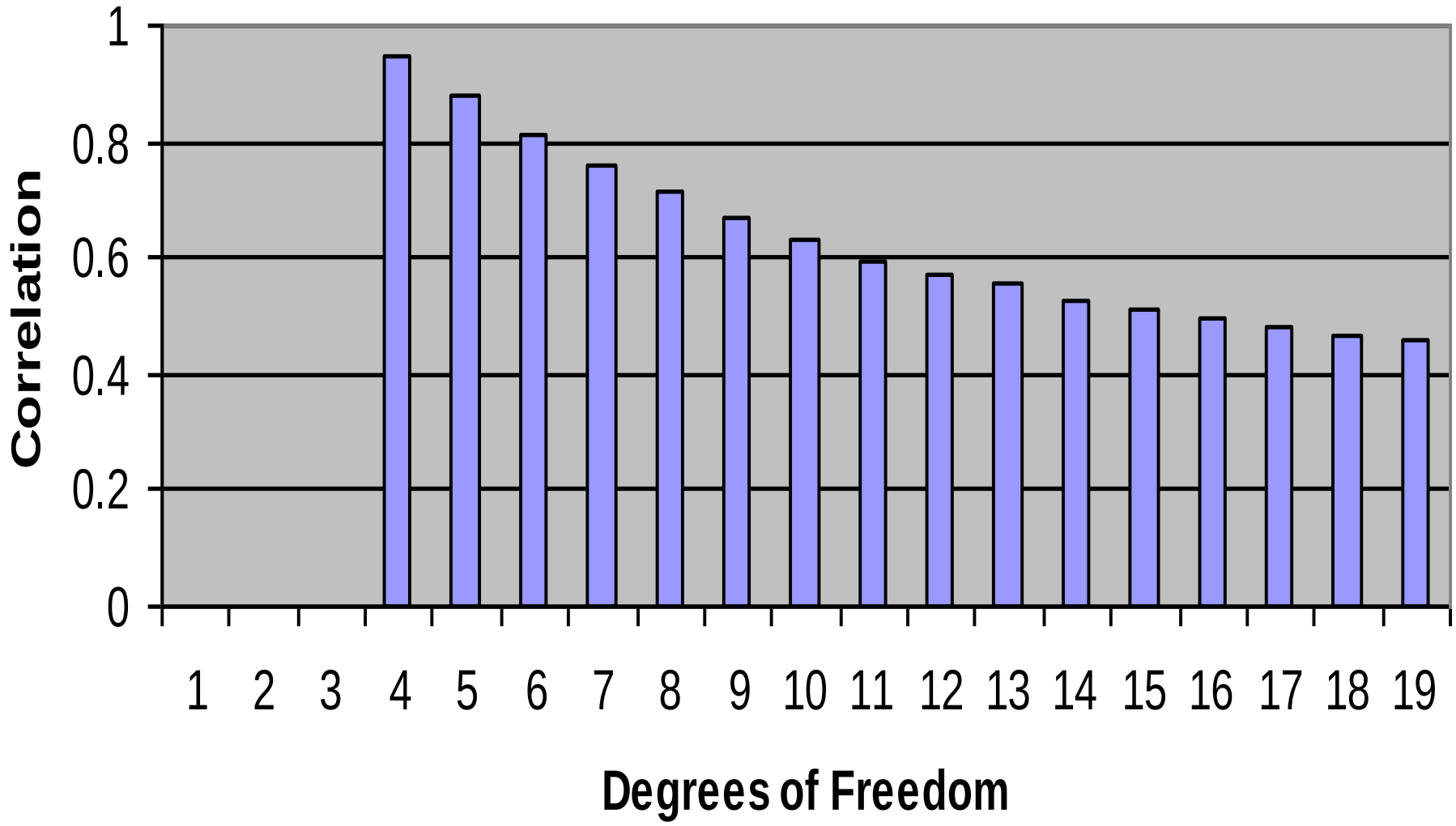
- Interpretability of coefficients:
  - Perfect interpretability is not normally a goal of multiple regression
  - If correlation among predictors exists, interpretability will lessen, but regression will still work properly (unless collinearity is extreme)
- Example:
  - Two predictors are correlated, say, correlation = 0.7, and both correlate positively with predictand individually
  - The regression equation might have strong positive coefficient and strong negative coefficient
  - Multiple regression still usable (provided stability of regression model is tested with cross-validation or retro-active forecasting)

PC time scores of gpm heights at various pressure levels – **not to be used together in one statistical model!**



Model 1: CV correlation = 0.2; Model 2: CV correlation = 0.4.Which one is the better model?

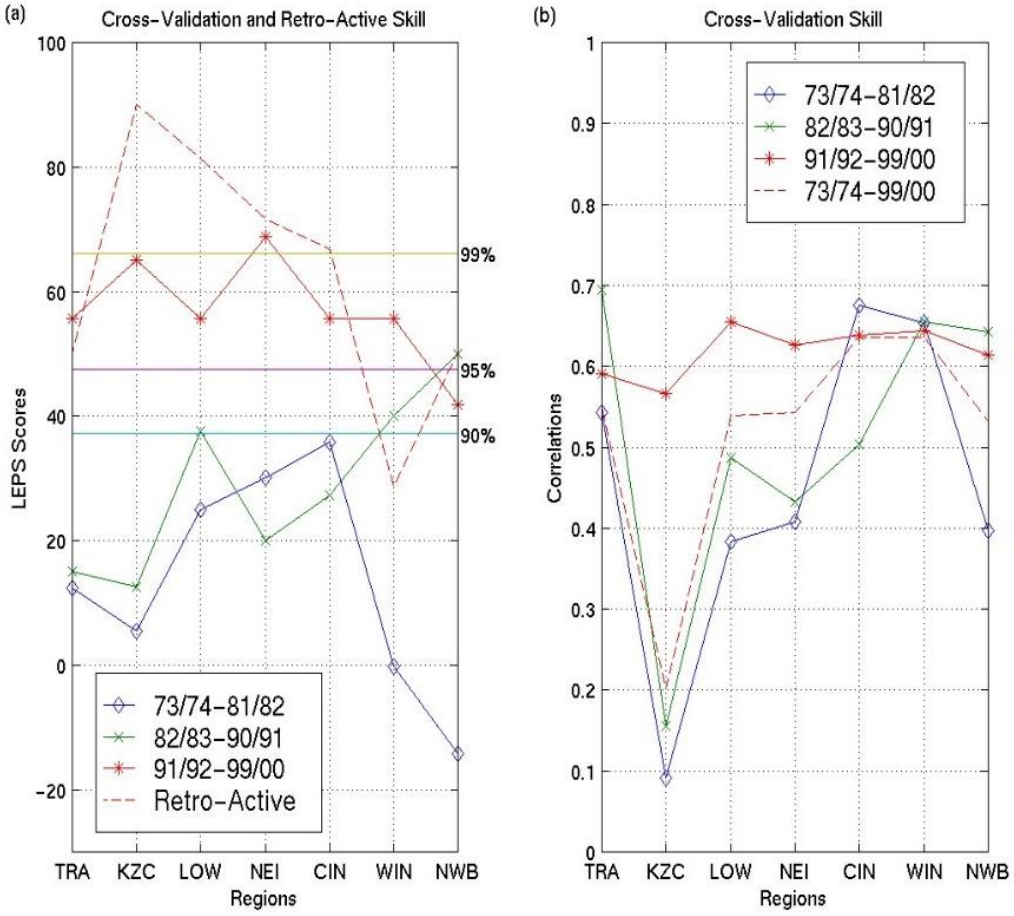
- Depends on the length of the respective model training periods
- The shorter the climate period, the higher the required correlation for statistical significance


95% Level of Significance
Correlation
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Degrees of Freedom

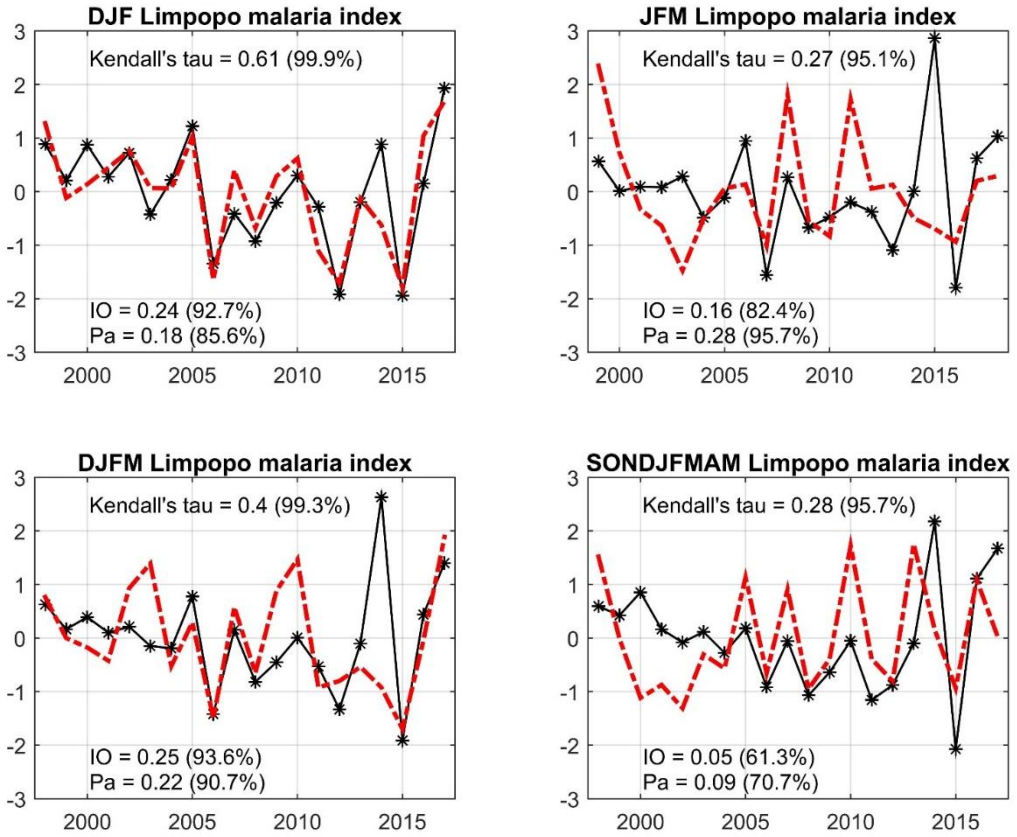
Assumptions on Stability

- Predictability remains constant
- The relationships between predictor and predictand variables remain valid under future climate conditions

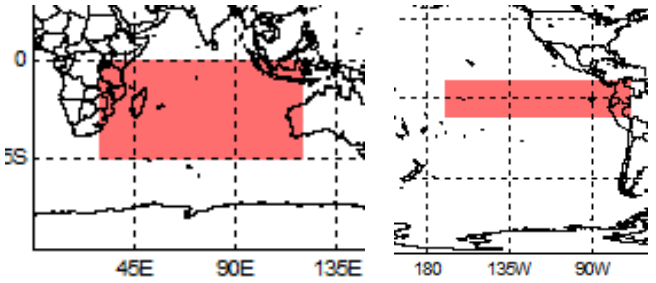
Variation in Forecast Skill

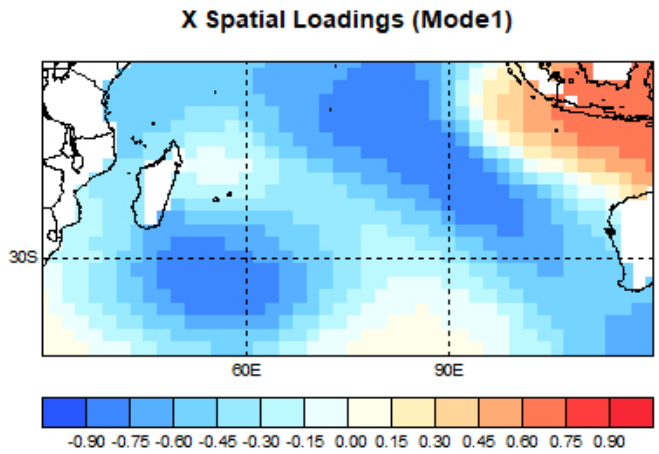


A large increase in LEPS scores is seen for the most recent of the three 9-year periods considered here. The skill is therefore seen not to be stable throughout this cross-validation period. The increase in skill may be attributable to the large number of ENSO events during the 1990s, since the main contribution in forecast skill of the model comes from the equatorial Pacific Ocean

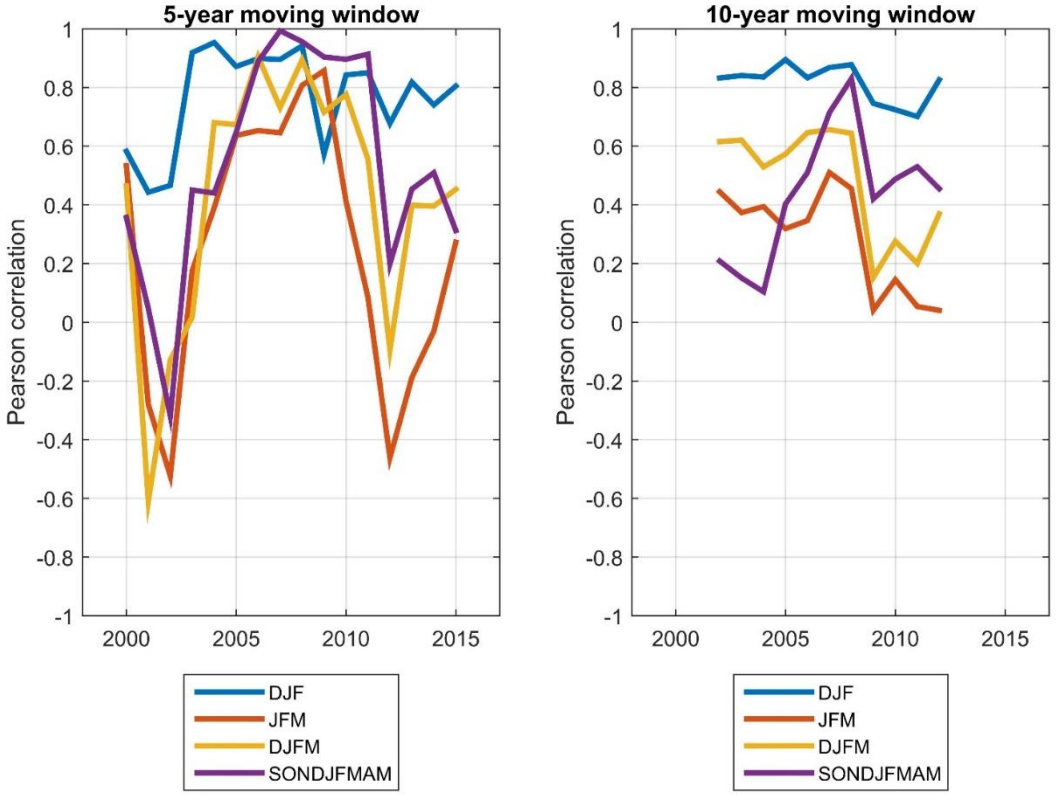


SST as predictors are used as baseline: Multiple linear regression models



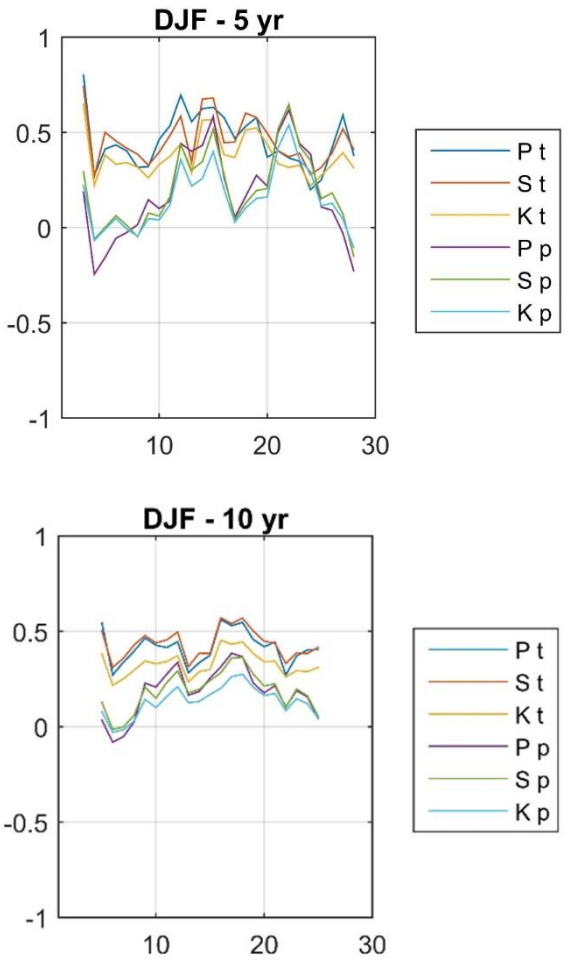


Cross-validated malaria cases hindcasts (1-month lead) for the four seasons indicated. Time series have been normalised. The top value on each panel is the Kendall's tau (Kt) correlation between hindcast (red dashed) and observed (black asterisked). IO: Kt when using Indian Ocean SST as predictor; Pa: Kt when using equatorial Pacific Ocean SST as predictor.

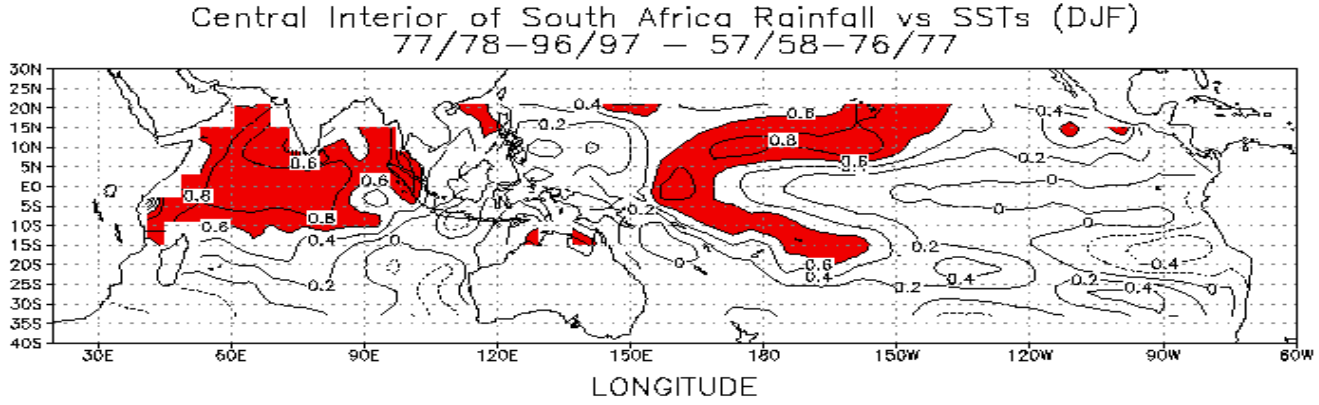


Respectively 5- and 10-year moving windows are used to calculate correlations between observed and hindcasts for the four seasons indicated. The DJF season is not only found to be associated with highest skill (correlations), but is also the season during which forecast skill remains the most consistent.

Variation in rainfall & max temp forecast skill from ECHAM4.5



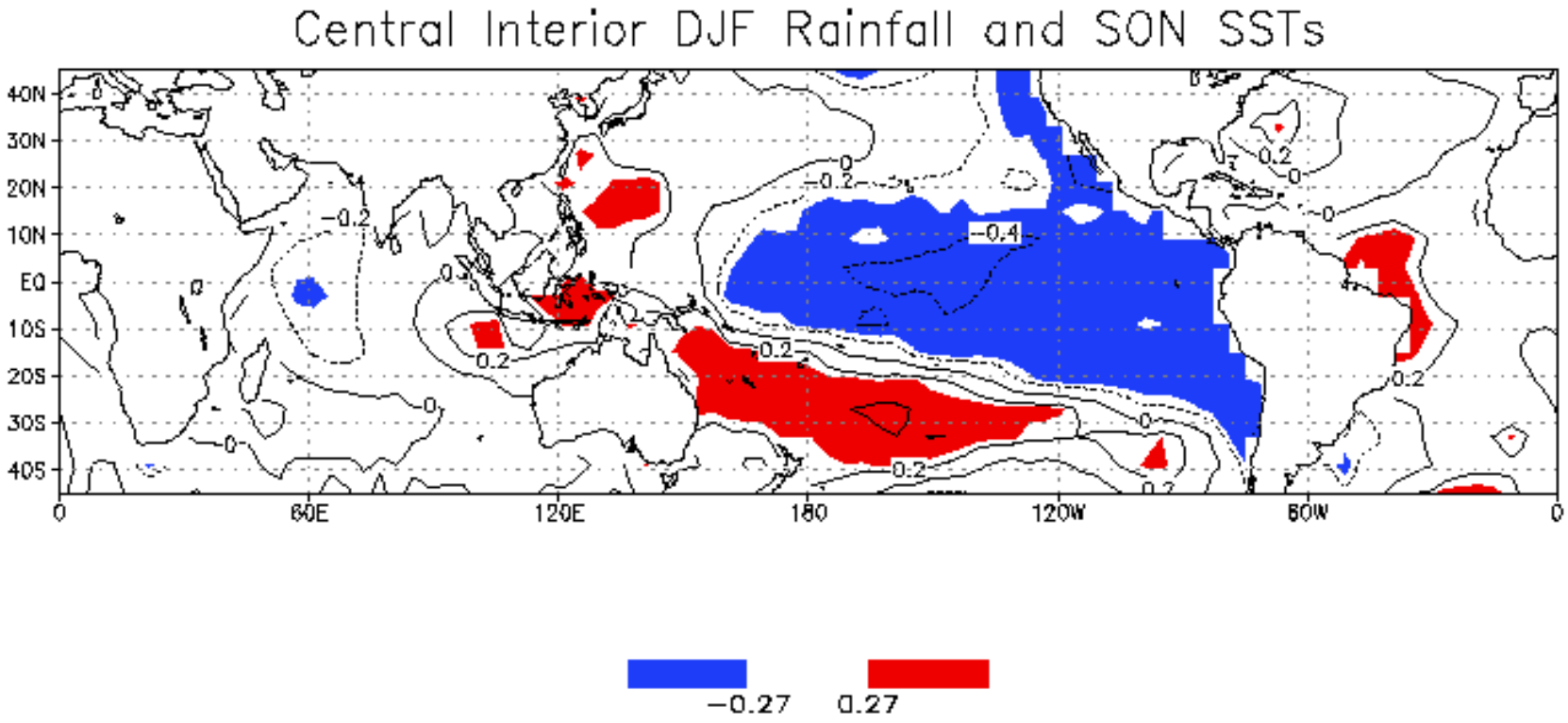
Pearson (P), Spearman (S) & Kendall's tau (K)



This figure shows where the largest changes in the association (correlation) between DJF Indo-Pacific SSTs and central south African DJF rainfall (1977/78 to 1996/97 - 1957/58 to 1976/77) are found, and indicates that the climate system is not always stable

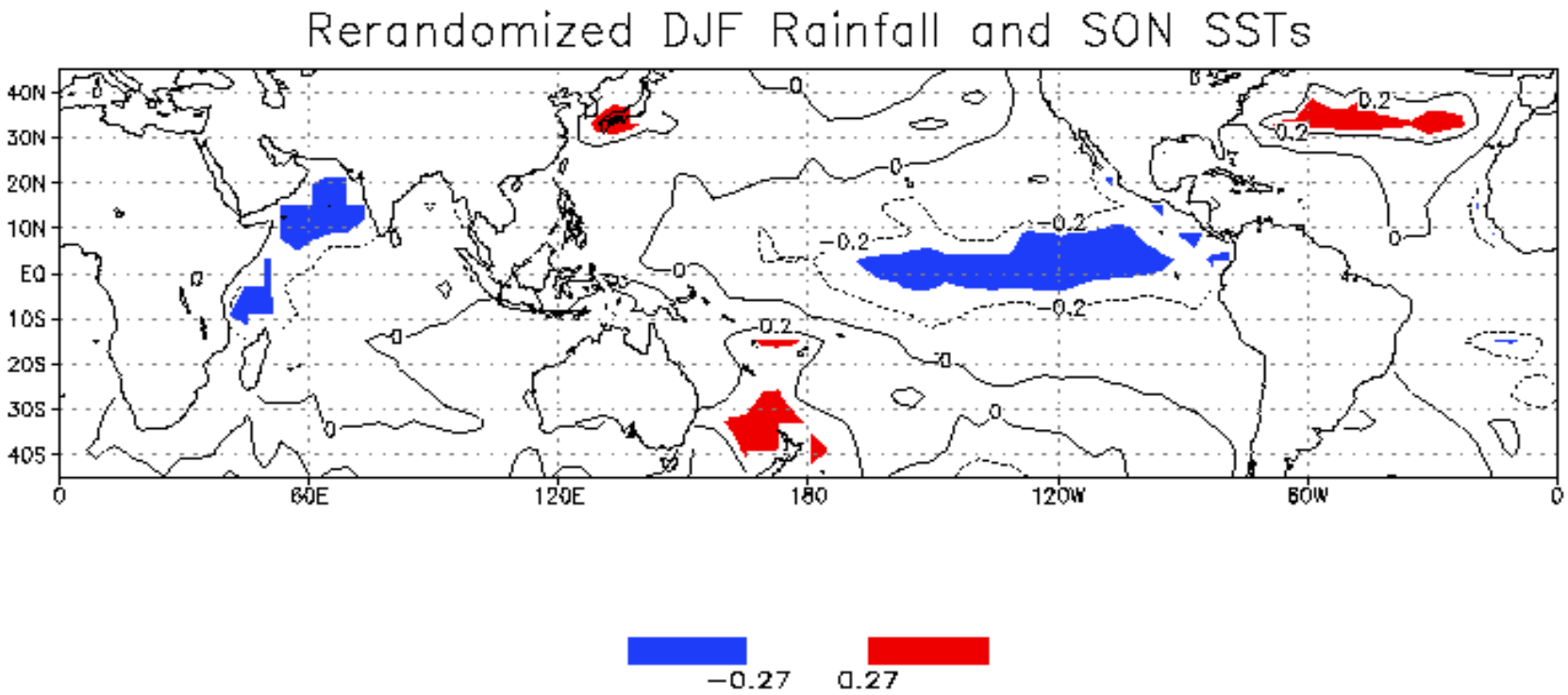
Field Significance and Multiplicity

- Special problems with statistical tests involving atmospheric fields – testing for pattern significance
- Positive grid-point-to-grid-point correlation of underlying data produces statistical dependence among local tests
- Multiplicity: the problem when the results of multiple independent significant tests are jointly evaluated


Central Interior DJF Rainfall and SON SSTs
40N
30N
20N
10N
EQ
10S
20S
30S
40S
0
60E
120E
180
120W
60W
0
−0.4
−0.2
0.2
−0.27
0.27

…and…

…after only a few rerandomization of the rainfall time series…



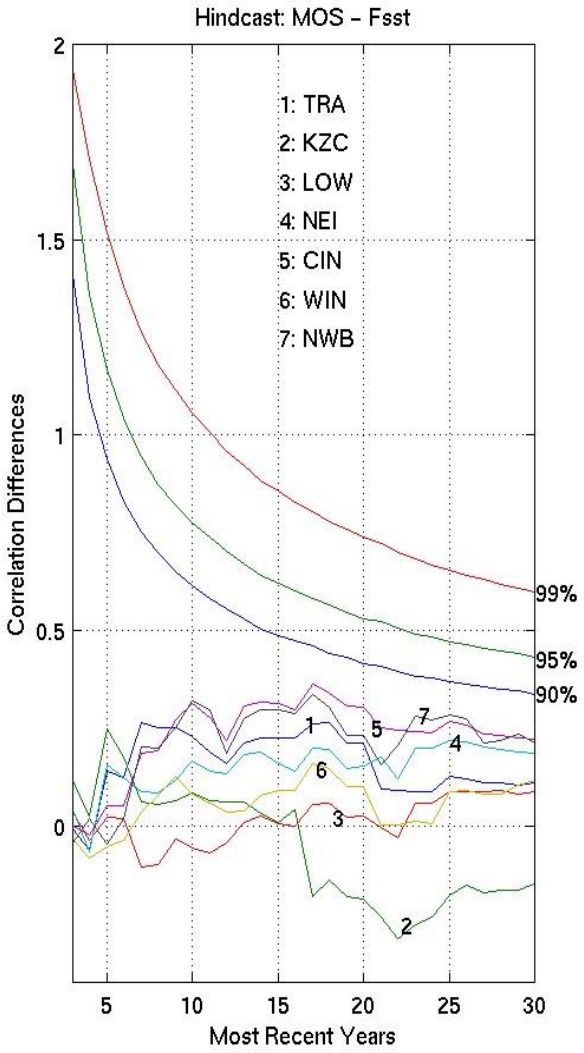
Using a Monte Carlo approach, it was possible to design a rerandomized rainfall time series that produced an El Niño type spatial pattern in the oceans. Clearly the real association between SON SSTs and the series of random numbers is zero (!!!), but the substantial grid-point-to-grid-point correlation among the SON SSTs yields spatial coherent areas of ***chance*** sample correlation that are deceptively high (due to the high spatial correlations the spatial degrees of freedom is far less than the number of grid-points).

Will empirical modelling become *obsolete*?

- No!
- Simple models can serve as a base-line against which the skill of elaborate models such as GCMs can be compared
- Empirical modelling can be applied to post-processing of dynamical model forecast output (*beware*, the same pitfalls are prevalent as discussed here for “ordinary” empirical modelling)

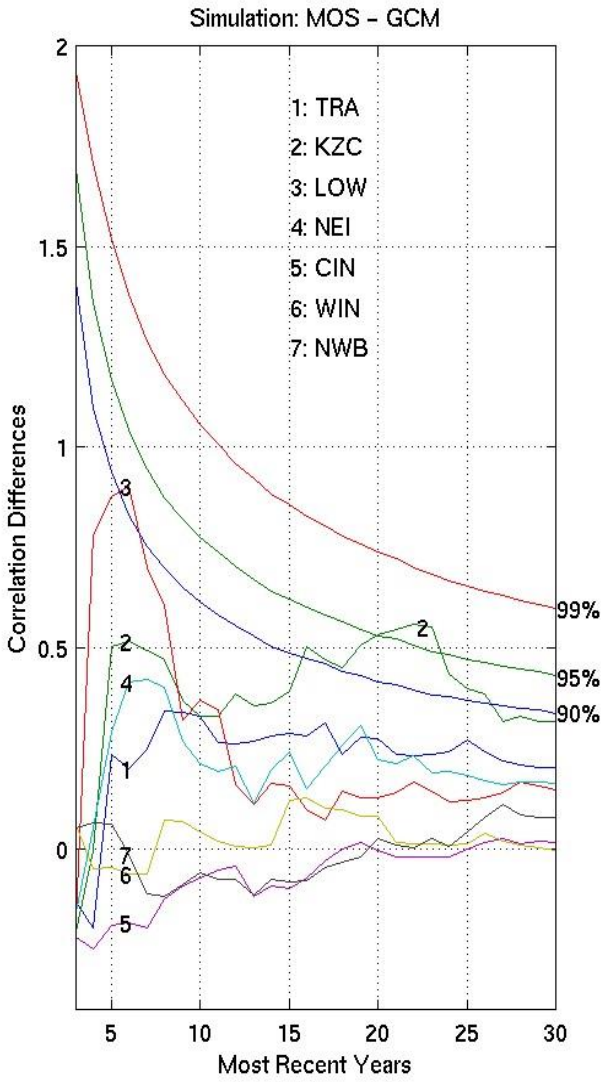
GCM-based forecast skill improvement over simple SST-Rainfall model skill

GCM-based forecasts generally outscore baseline model



Improvement over raw GCM output using Statistical Post-Processing

Post-processed GCM forecasts generally outscore raw GCM output



Do Statistical Pattern Corrections Improve Seasonal Climate Predictions in the North American Multimodel Ensemble Models?

ANTHONY G. BARNSTON

International Research Institute for Climate and Society, Columbia University, Palisades, New York

MICHAEL K. TIPPETT

Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, and Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

(Manuscript received 31 January 2017, in final form 7 July 2017)

ABSTRACT

Canonical correlation analysis (CCA)-based statistical corrections are applied to seasonal mean precipitation and temperature hindcasts of the individual models from the North American Multimodel Ensemble project to correct biases in the positions and amplitudes of the predicted large-scale anomaly patterns. Corrections are applied in 15 individual regions and then merged into globally corrected forecasts. The CCA correction dramatically improves the RMS error skill score, demonstrating that model predictions contain correctable systematic biases in mean and amplitude. However, the corrections do not materially improve the anomaly correlation skills of the individual models for most regions, seasons, and lead times, with the exception of October–December precipitation in Indonesia and eastern Africa. Models with lower uncorrected correlation skill tend to benefit more from the correction, suggesting that their lower skills may be due to correctable systematic errors. Unexpectedly, corrections for the globe as a single region tend to improve the anomaly correlation at least as much as the merged corrections to the individual regions for temperature, and more so for precipitation, perhaps due to better noise filtering. The lack of overall improvement in correlation may imply relatively mild errors in large-scale anomaly patterns. Alternatively, there may be such errors, but the period of record is too short to identify them effectively but long enough to find local biases in mean and amplitude. Therefore, statistical correction methods treating individual locations (e.g., multiple regression or principal component regression) may be recommended for today's coupled climate model forecasts. The findings highlight that the performance of statistical postprocessing can be grossly overestimated without thorough cross validation or evaluation on independent data.

1. Introduction

In principle, dynamical climate prediction models are expected to produce more accurate climate predictions than statistical models on seasonal to interannual time scales. This expectation is based on the fact that dynamical models make use of the often complex and nonlinear physical laws governing oceanic and atmospheric behavior, while statistical models use only relationships (often linear) gleaned from finite records of observational data. Operationally, however, dynamical models did not show clear superiority over statistical models in predicting monthly or seasonally averaged climate until near the turn of the twenty-first century, as more advanced data assimilation methods and computer power finally enabled them to perform closer to their potential.

While comprehensive coupled ocean–atmosphere dynamical models are now heavily relied upon for seasonal climate predictions, they still have aspects in need of further improvement. Their systematic errors, or biases, vary by model, season, lead time, and location. The presence of biases creates an opportunity for statistical models to detect and correct them, resulting in improved final forecast quality. Such methods can be

Denotes content that is immediately available upon publication as open access.

Corresponding author: Anthony G. Barnston, tonyb@iri.columbia.edu

DOI: 10.1175/JCLI-D-17-0054.1

To make empirical forecasts useful:

- Be aware that some models may only appear to be useful
- Always test forecasts on independent data
- “Fine tuning” (e.g., variance adjustment) of forecasts may have pros AND cons
- Collinearity, instability and multiplicity - modellers beware!
- Use these models correctly, since empirical models are not and will not become obsolete