Creating Empirical Models

Simple Linear Regression

Many of the problems in seasonal climate prediction are concerned with determining a relationship between two sets of variables. For example, the sea-surface temperature (SST) variability of the eastern equatorial Pacific (variable one) is believed to have a relationship with rainfall variability (variable two) at certain regions of the globe, at a time scale that exceeds that of day to day variations. Knowledge of the extent to which such a relationship may be valid would have as a result that the expected rainfall could be predicted if knowledge of the state of the eastern equatorial Pacific Ocean is known.

In the example to be presented here, there is a single response variable *Y* which depends on the value of an input variable *X*. In the literature, the *Y* variable is referred to as a dependent variable, and the *X* as an independent variable. In seasonal climate prediction, the dependent variable is often referred to as the *predictand*, and the independent variable the *predictor*. For the example above, the predictand is rainfall and the predictor the SSTs of the eastern equatorial Pacific Ocean, both on a seasonal time scale of several months (mostly 3 months). In practice, however, due to random errors it is never possible to *exactly* predict the response for one or a set of predictors. Notwithstanding, one should be able to construct a simple regression equation that supposes a linear relationship between the mean response and a single independent variable:

$$Y = \alpha + \beta x$$

where x is the value of the predictor and *Y* the response. In order to see if one can use the above equation as a prediction tool, a good starting point would be to make a simple plot, called a *scatter diagram*, to determine if a linear relationship (or any other) exists between the predictor and predictand values. Consider the following 20 data pairs (x_i, y_i), i=1,2,…,20, relating *y*, a 3-month averaged rainfall index of a region of the globe (i.e. central southern Africa), to *x*, a 3-month averaged SST index of the eastern equatorial Pacific Ocean.

i	x_i	y_i	i	x_i	y_i
1	-1.3719	-0.2179	11	0.0670	0.4004
2	-0.4962	0.9133	12	0.0131	-1.0518
3	0.9218	-0.4753	13	1.7473	-1.2373
4	-1.8650	2.0254	14	0.3468	-1.0691
5	-1.0855	0.4186	15	-0.2330	-0.4616
6	-1.8062	2.1255	16	0.0106	-0.6662
7	0.0720	0.3577	17	1.0150	-0.9299
8	0.5173	-0.0106	18	1.5223	1.5044
9	0.2192	-0.8236	19	-0.7204	0.2140
10	0.7199	-0.5498	20	0.4061	-0.4662

A plot of y_i versus x_i can be seen in Figure 1. The scatter diagram seems to reflect a linear relationship between x and y. In this case the relationship is negative: when the SST index increases (decreases), the rainfall index decreases (increases). Take note of an *outlier* at approximately (1.5, 1.5).

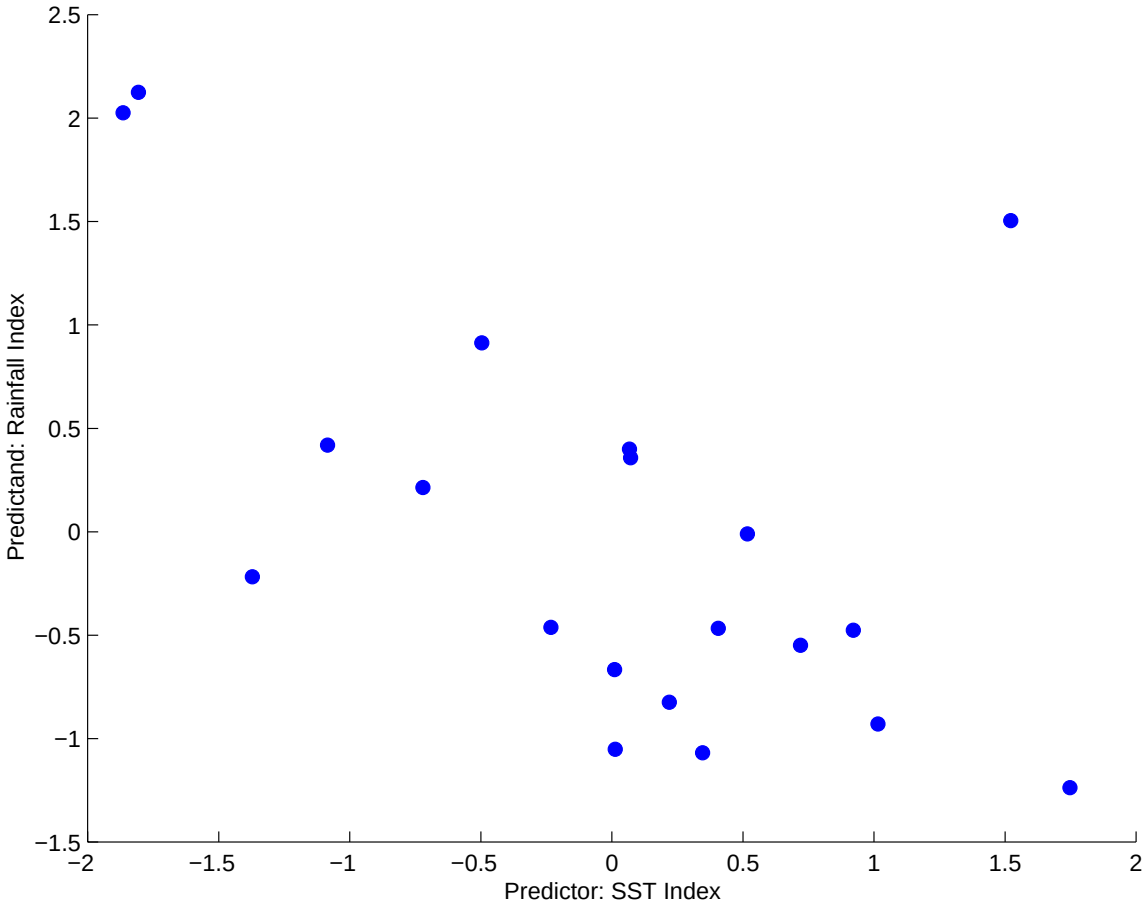

Figure 1. Scatter plot

For the example presented here, the SST index is, say, for the 3-month period September to November (SON), and the rainfall index is the 3-month period January to March (JFM) that follows the sea-surface temperature period. Thus, if knowledge of the SON SST index can obtained at the beginning of December, one should be able to make a qualitative statement about the expected JFM rainfall. For example, if the SON SST index has been observed to be positive, it is likely that the following JFM season will be associated with a negative rainfall index, i.e. drier than average conditions. However, it would be more desirable if a more quantitative statement about the rainfall of the coming JFM season can be made. Such a statement would have to include indicators of how much above or below the climatological average the rainfall index is expected to be and, very

importantly, the degree of uncertainty of the expected outcome (reflected in the random error of the relationship).

In order to determine estimators of α and β of the simple regression equation of above, the method of *least-squares* is employed. If *A* is the estimator of α, and *B* is the estimator of β corresponding to the data set x_i, Y_i, i=1,2,…,20, then the straight line $A + Bx$ is called the *estimated* regression line. Any book on statistical methods will have an elaborate explanation on how to determine these estimators and will thus not be discussed here. Furthermore, all statistical computer software packages also have the ability to easily calculate these estimators. Figure 2 shows the result of using such a software package to determine the best linear fit to the data presented above. The α estimator *A* is equal to 0, and the β estimator *B* is equal to –0.5598, which leads to an estimated regression line of $-0.5598x$.

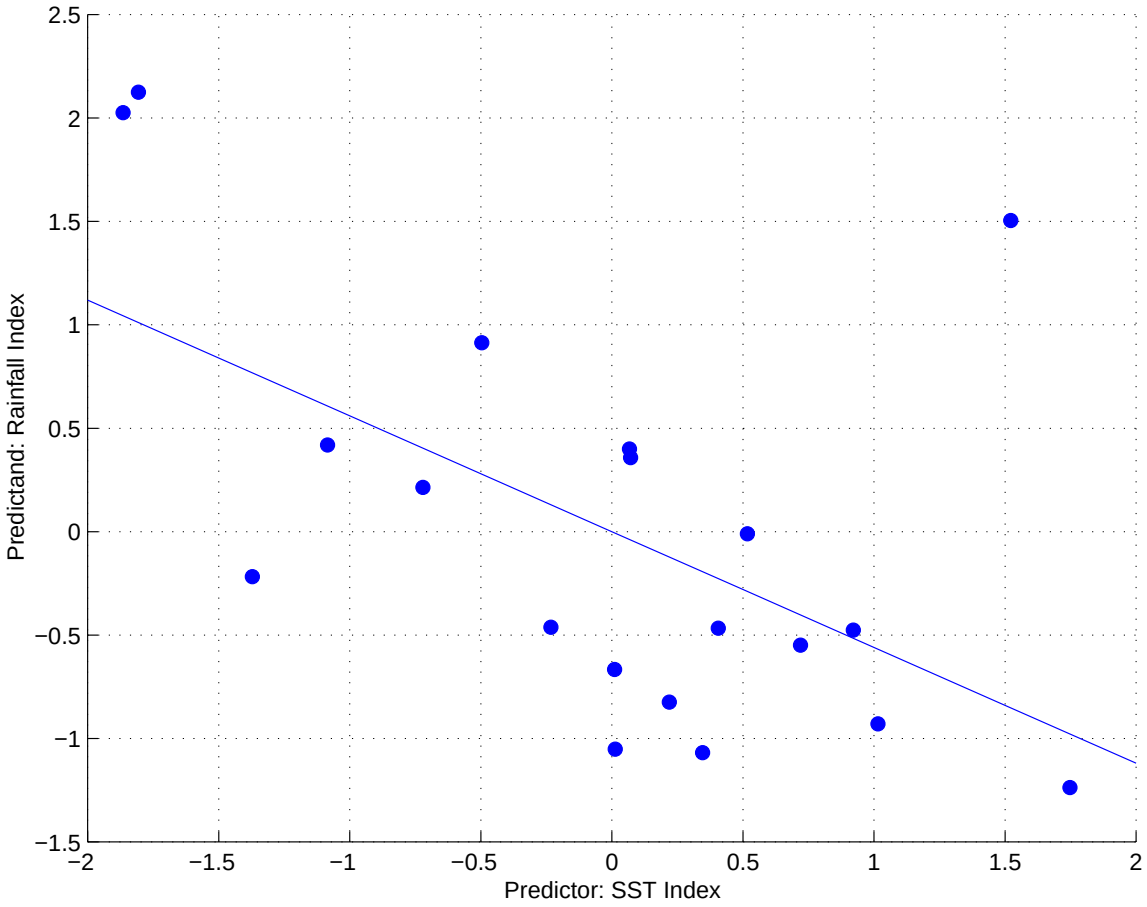

Figure 2. The best linear fit to the data.

A central part of the regression output of the statistical software packages is a summary of the foregoing information in an *ANOVA* (analysis of variance) table. Not all the information in an ANOVA table will be of interest, but it presents

related measures of the "fit" of a regression. The fit is an indication of the correspondence between the regression line and the scatter diagram of the data. From a forecasting point of view, the *mean-squared error* (MSE) is perhaps the most fundamental of the measures, since it indicates the variability of the observed predictand around the forecast regression line. In the case of a perfect linear relationship between predictor and predictand the MSE will be zero, but in the case of a poor linear relationship the MSE will be large. Another measure of the fit of a regression is the *coefficient of determination* (R^2) which is, for the case of simple linear regression, the *squared* value of the Pearson *correlation* between predictor and predictand. Qualitatively, the R^2 can be interpreted as the proportion of the variation of the predictand that is described or accounted for by the regression. For a perfect regression, the $R^2 = 1$, but for R^2 near 0 very little of the variation is explained. In the majority of applications, however, the response of a predictand can be predicted more adequately not on the basis of a single independent input variable but on a collection of such variables.

Multiple Linear Regression

Multiple linear regression is the more general and also more common case of simple linear regression. There is, as is the case with simple linear regression, still only one predictand, but there is more than one predictor variable. For example, seasonal rainfall may be influenced by the El Niño / Southern Oscillation (ENSO) AND by the SSTs in the Indian Ocean AND/OR the Atlantic Ocean.

For any K number of predictor variables, the response Y is related to them by the relation:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Simple linear regression constitutes the special case of $K = 1$. For multiple regression, each one of the K predictor variables has its own coefficient and the K+1 parameters are found by simultaneously solving K+1 equations using matrix algebra. In practice, the calculations are usually done using statistical software packages, and the calculations are again summarized in an ANOVA table. Care should however be taken when interpreting the R^2 value, because it is not the square of the Pearson correlation coefficient between the predictand and any of the predictor values in particular.

Designing a Scientifically Sound Regression Model

The statistical forecasts produced by the linear equations are objective in the sense that a particular set of predictors will *always* produce the same forecast for the predictand, once the forecast equation has been developed. However, many

subjective decisions go into the development of the forecast equations. An example of a subjective decision is the selection of the predictor set. There are almost always more potential predictors available that can be used in a regression, and finding the appropriate combination of predictors is not a straight forward procedure. The process is definitely *not* simply an addition of the members on the list of potential predictors until an apparently good relationship is achieved, and there are dangers associated with including too many predictor variables in a forecast equation, which will lead to *overfitting* the data. An overfit regression will fall apart when it is used as an operational prediction model. To counter the possible dangers of including too many predictor variables, firstly begin the development of the regression equations by choosing only physically reasonable or meaningful potential predictors. Secondly, the tentative regression equation needs to be tested on a sample of data not involved in its development. Thirdly, one needs a reasonably large developmental sample if the resulting regression is to be stable. By stability is usually understood that the regression coefficients are also applicable to independent (or future) data, and have not resulted from an overfit regression.

Suppose that only those predictor variables that have physical relevance are included in a set of potential predictors. However, it would generally not be useful to include all of the potential predictors in a final regression equation, because the predictor variables are almost always mutually correlated, so that the full set of potential predictors contains redundant information. One needs therefore, a method to choose among potential predictors, and of deciding how many of them are sufficient to produce a good prediction equation. This method of selecting a good set of predictors from a pool of potential predictors is called *screening* or *stepwise regression* of which the most commonly used procedure is known as *forward selection*. On the forward selection step, all the potential predictors are examined individually for the strength of their linear relationship to the predictand. The predictor whose linear regression is best among all candidate predictors is chosen as x_1. At the next stage, the predictor variable that is chosen as x_2, is the one that improves the model the most. Subsequent step will follow this procedure exactly. An alternative screening approach is called *backward elimination*, which is opposite to that of forward selection: the initial point is a regression containing all the potential predictors, and at each step the least important predictor variable is removed from the regression equation. Without any stopping rule, both forward selection and backward elimination will continue until all the candidate predictor variables were included (forward selection) or until all the predictor variables had been eliminated (backward elimination). The MSE could be used as such a rule: if a regression equation were being developed little would be gained by adding more predictors if the root of the MSE lies within a 95% confidence interval around the forecast value. Ideally the stopping rule would be at a point where the MSE does not decline appreciably with the addition of more predictors. It is normally good practice to have the number of predictors substantially less than the sample size (training period).

Once the stepwise regression has been used to design the regression equation which contains the most suitable predictor or a combination of predictors, one way of subsequently testing the operational performance of the model is to use a technique called *cross-validation,* where the value that is predicted is omitted from the developmental sample or training period. Cross-validation can be performed by removing a distinct but corresponding partition from the predictor and predictand data. Suppose there are *n* partitions of the original predictor and predictand fields. For some partition index η, where $1 \leq \eta \leq n$, the ηth predictor and predictand fields are removed. The remaining *n*-1 partitions are used to train the model. The model prediction is then tested against the predictand of index η, which was withheld from the training period. This procedure is repeated *n* times, sequentially omitting a single partition pair, resulting in a series of *n* forecast partitions, each of which can be directly compared with the observed partitions. The following diagram illustrates the idea of cross-validation further by showing a simple example of 6 years of data for an arbitrary predictor and predictand setup.

	Year 1	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						

Model 1 uses the information of the predictor and predictand data (Year 2 to Year 6) to predict the predictand of the first year, Model 2 for the second year, and so on until Model 6 is used predict the outcome of Year 6. Each of these models has a 5-year training period. Model 7, that uses all 6 years of the training period, does ***NOT*** estimate the model's true forecast performance, and will lead to inflated skill levels because the value that is predicted is also contained in the training period of the regression model. Although the example above shows a removal of a single year from the training set, it is good practice to first determine of any temporal autocorrelation exists in the data. If significant autocorrelation exists, it is recommended that the years omitted from the training set for each model be expanded to three years (one year on either side of the omitted years of Model 2 to 5 in the above diagram, but for the year after the omitted year in Model 1 and the year before the omitted year in Model 6). This procedure may be necessary when there are trends in the data. For long-term trends, however, a wider time window (more years) will not solve the problem, and the data *may* (but not necessarily) need detrending.

A further estimation of the performance of a regression model is a technique called *retro-active forecasting*. The process involves the evaluation of prediction compared to observations excluding any information following the target year and can be explained by considering a retro-active forecast procedure during a 6-

year period following the 6-year training period of Model 7. The regression model is first trained considering information leading up to and including Year 6. Three, say, consecutive years are predicted (Year 7, 8 and 9) using the same information. The model subsequently uses information leading up to and including data for Year 1 to 9 to predict the following three years (Year 10, 11 and 12). It should be noted that the training set of the above example of 6 years (Model 7) is inadequate, and in practice should exceed several decades.

Forecast Skill Estimation

There are many different ways to estimate the *forecast skill* (the relative accuracy of a set of forecasts) of a regression model. By accuracy is meant the correspondence between individual forecasts and the events they predict. In estimating the skill, a possible scoring method to use is a variation of the standard tercile approach. Each of the observed and predicted fields is separated into its own three equiprobable classes or terciles. These terciles are referred to as above-normal (A), near-normal (N), or below-normal (B) categories. The *categorical* forecast is compared with that of the observed to calculate the model skill. Categorical means that the forecast consists of a statement that one and only one of a set of possible events will occur.

Categorical forecast verification often involves the use of a *contingency table* of absolute counts. The following 3 X 3 contingency table is a 3-category (A, N or B) forecast verification where each variable in the nine boxes represents nine possible forecast outcomes:

	O_A	O_N	O_B
F_A	**R**	**S**	**T**
F_N	**U**	**V**	**W**
F_B	**X**	**Y**	**Z**

For example, **R** represents the number of cases from a verification sample when the observed (O) and forecast (F) categories were for above-normal values to have occurred, and **W** represents the number of cases when the observed category was below-normal and the forecast category was near-normal, and so on. The sample size of the verification data is the summation of all the entries in the table.

Only very basic accuracy measures of the multicategory forecasts will be discussed in some detail here, and they are the *hit score*, *false alarm ratio* (FAR) and the *bias*. The hit score is the number of times a correct category is forecast (**R**+**V**+**Z**). The FAR is that fraction of forecast events that failed to materialize (best possible FAR is zero and worst possible FAR is one), and the bias is the comparison of the average forecast with the average observation. A bias greater than one indicates overforecasting and a bias less than one indicates underforecasting. The bias for forecasts of above-normal is (**R**+**S**+**T**)/(**R**+**U**+**X**), of near-normal is (**U**+**V**+**W**)/(**S**+**V**+**Y**), and of below-normal is (**X**+**Y**+**Z**)/(**T**+**W**+**Z**). The FAR for forecasts of above-normal is (**S**+**T**)/(**R**+**S**+**T**), of near-normal is (**U**+**W**)/(**U**+**V**+**W**), and of below-normal is (**X**+**Y**)/(**X**+**Y**+**Z**).

An Example of Model Testing

This section deals primarily with the process of testing a regression model, assuming that only physically plausible predictors have been considered and the screening process has been adequately performed. The data used to design the model is that of the above example where an index of SSTs are related to a rainfall index. The lag-1 autocorrelations for both the predictor and predictand sets are less than 0.2, so it would be safe to test the model with a cross-validation design that omits only one year at a time (like the example presented above). The estimated regression lines for each of the 20 cross-validation models are (with the training period in parenthesis):

Model 1: $Y_1 = 0.0579 - 0.6434x$ (years 2 to 20)
Model 2: $Y_2 = -0.0339 - 0.5421x$ (year 1 and 3 to 20)
Model 3: $Y_3 = -0.0022 - 0.5620x$ (years 1 and 2 and 4 to 20)
Model 4: $Y_4 = -0.0640 - 0.4342x$ (years 1 to 3 and 5 to 20)
Model 5: $Y_5 = 0.0106 - 0.5719x$ (years 1 to 4 and 6 to 20)
Model 6: $Y_6 = -0.0716 - 0.4237x$ (years 1 to 5 and 7 to 20)
Model 7: $Y_7 = -0.0209 - 0.5614x$ (years 1 to 6 and 8 to 20)
Model 8: $Y_8 = -0.0149 - 0.5679x$ (years 1 to 7 and 9 to 20)
Model 9: $Y_9 = 0.0370 - 0.5512x$ (years 1 to 8 and 10 to 20)
Model 10: $Y_{10} = 0.0080 - 0.5537x$ (years 1 to 9 and 11 to 20)
Model 11: $Y_{11} = -0.0230 - 0.5614x$ (years 1 to 10 and 12 to 20)
Model 12: $Y_{12} = 0.0550 - 0.5590x$ (years 1 to 11 and 13 to 20)
Model 13: $Y_{13} = 0.0164 - 0.5296x$ (years 1 to 12 and 14 to 20)
Model 14: $Y_{14} = 0.0464 - 0.5428x$ (years 1 to 13 and 15 to 20)
Model 15: $Y_{15} = 0.0313 - 0.5674x$ (years 1 to 14 and 16 to 20)
Model 16: $Y_{16} = 0.0348 - 0.5594x$ (years 1 to 15 and 17 to 20)
Model 17: $Y_{17} = 0.0202 - 0.5382x$ (years 1 to 16 and 18 to 20)
Model 18: $Y_{18} = -0.1423 - 0.7878x$ (years 1 to 17 and 19 and 20)
Model 19: $Y_{19} = 0.0103 - 0.5675x$ (years 1 to 18 and 20)
Model 20: $Y_{20} = 0.0127 - 0.5543x$ (years 1 to 19)

Take note of the regression equation of Model 18. The estimators of the equation are quite different from those of the other 19 models, which is the result of an *outlier* in the predictand data (rainfall). This outlier can be seen at approximately (1.5, 1.5) in Figure 3 that shows the linear fits of the 20 models. The red line is the best linear fit of Model 18. The Pearson correlation coefficients between the predictor and predictand data for the different training periods for all the models except Model 18 range from 0.44 to 0.61, but the predictor-predictand correlation associated with Model 18 is 0.79. Model 18 is probably a more suitable regression that describes the physical association between the rainfall and the SSTs, because this extremely large rainfall anomaly was caused by a single synoptic event that lasted for a relatively short period of time (about one week) during the JFM rainfall season, and was more than likely not associated with the SST variability of the eastern equatorial Pacific (the predictor in the regression models). However, one should be careful when omitting such seasons for which the model did not make an accurate forecast simply because the exclusion of them would make the model look good. There may be other physical predictors that were not considered in designing the model that may have the potential to improve on the forecast accuracy of those seasons that were poorly predicted. Notwithstanding, any statistical model will most likely always have difficulty in forecasting rainfall during a season accurately if most of that season's rainfall is the result of a single synoptic event that causes large amounts of rainfall.

By removing the outlier from the data there are only 19 seasons left. Because it is preferred that the 3 classes of above-normal, near-normal and below-normal are equiprobable (equal chance of an event to fall into any of the 3 categories), we would like to use a data set consisting of a number of events that are divisible by 3. Therefore, we will subsequently redo the cross-validation exercise using only 18 seasons by excluding the first season (or any other) of the 19-event predictor-predictand data. Figure 4 illustrates the cross-validated time series of the 18 seasons used in the analysis. The model seems to be able to capture most of the rainfall variability accurately, reflected also by the high correlation value. However, it is necessary to obtain a clearer indication of the model's ability to predict the categories correctly. The near-normal category is shown in Figure 4 as the horizontal lines on either side of the zero line. That they are not symmetric around the zero line demonstrates the rainfall data's *skewness*. Rainfall data are usually skewed, especially over arid and semi-arid regions, and is a result of the fact that rainfall measurements are physically constrained to lie above zero mm, which has as a result that the data are *positively* skewed. Most of the statistical techniques, including the one being discussed here, make the assumption that the data used to calculate the estimators of the equation follow the familiar bell-shaped curve of the *Guassian* distribution. If the rainfall indices were perfectly Gaussian, there would not be any observed skewness. However, as the sample size (i.e. the training period) becomes large, the arithmetic mean of a set of independent observations will have a Guassian distribution. This statement is called the *central limit theorem*, and is directly applicable to atmospheric data. Thus, it is necessary to use as long as possible records of data to construct the

linear model, provided that the data have been quality controlled. As a rule of thumb, linear models predicting seasonal rainfall variability should be based on data that have record lengths exceeding 30 years.

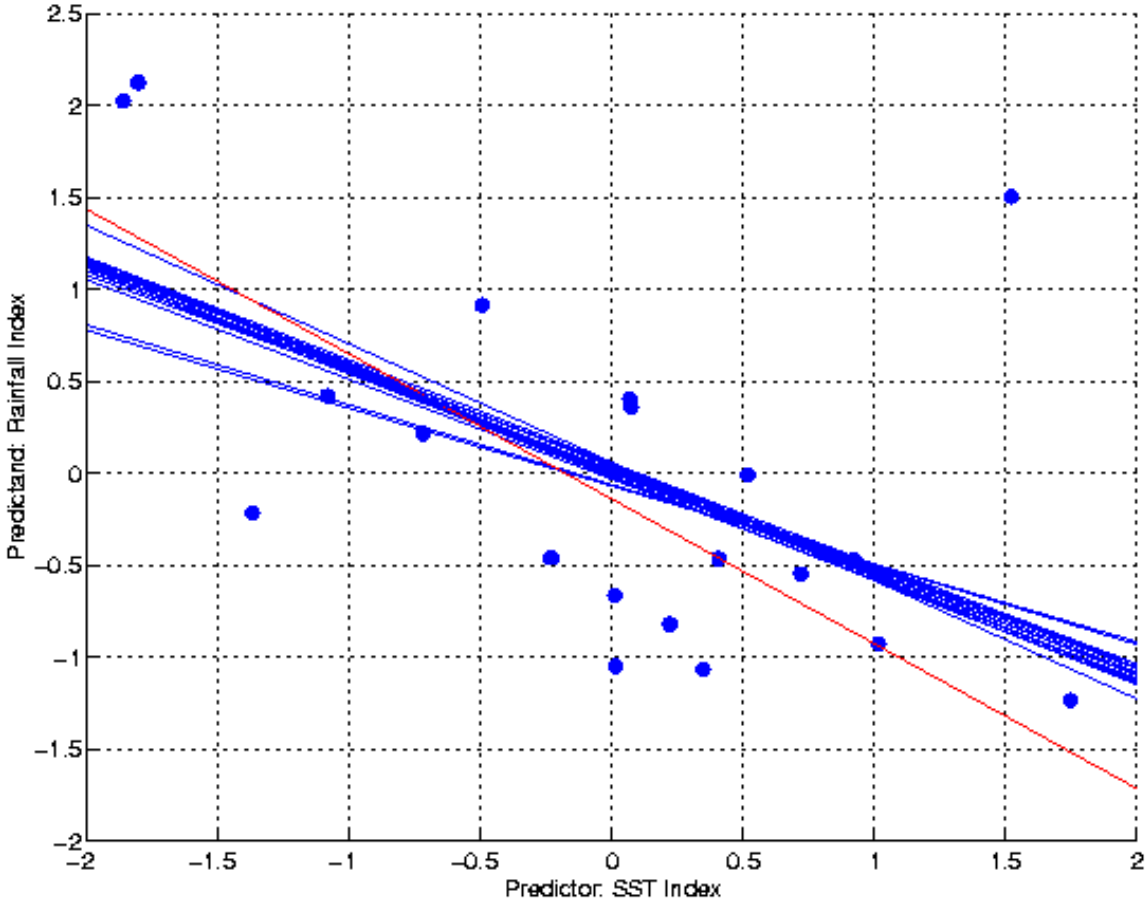

Figure 3. Linear fits of the 20 cross-validation models. The red line is the fit of Model 18.

The near-normal category was easily obtained by ranking the 18 rainfall seasons from the largest to the smallest index. The 6 largest numbers are associated with the above-normal category, the 6 smallest numbers with the below-normal category, and the remaining middle 6 values with the near-normal category.

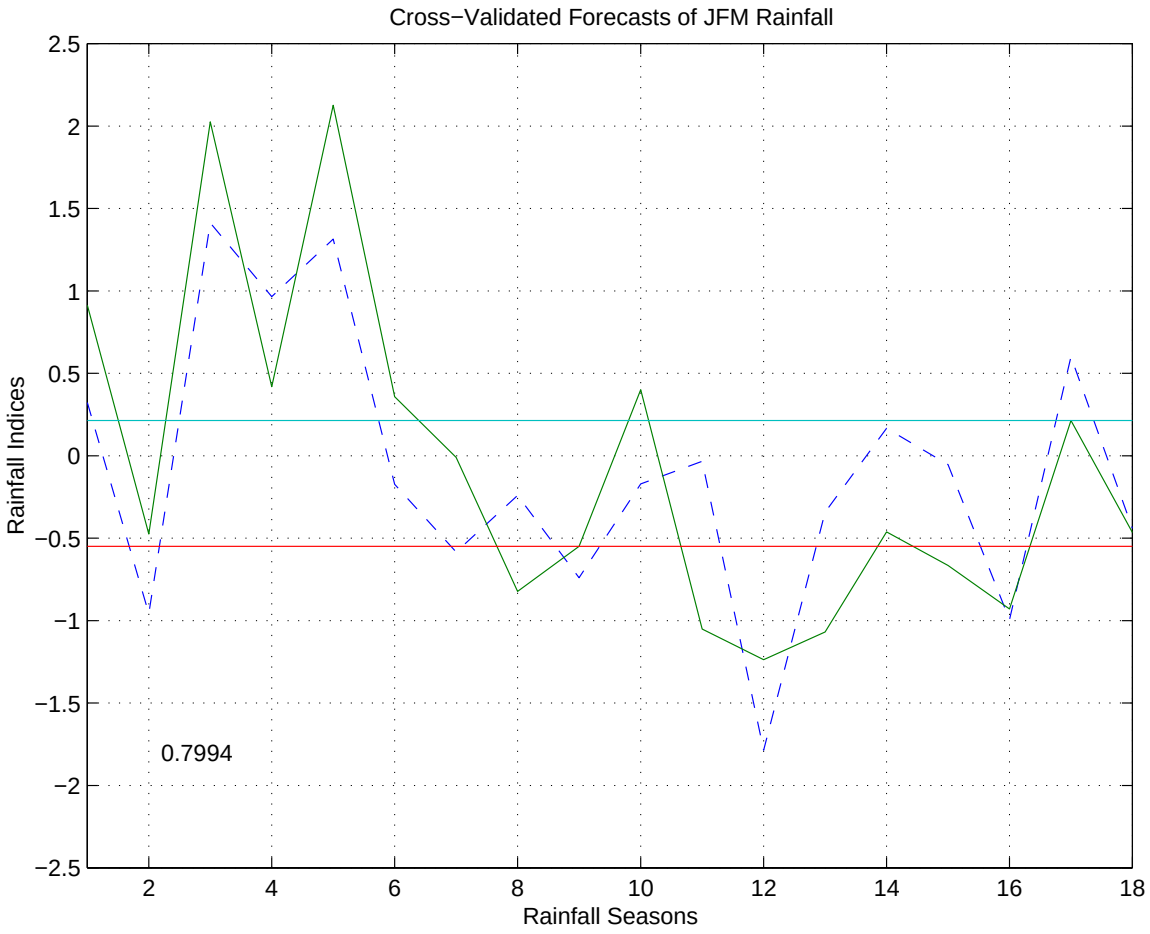

Figure 4. Cross-validated forecasts (dashed line) and observed (solid) rainfall indices using the data from 18 seasons. The horizontal lines on either side of the zero line are the upper and lower limits of the near-normal category. The correlation between the observed and predicted indices is in the bottom left corner of the graph.

Based on this description of the categories, the following 3 X 3 contingency table summarizes the forecasts shown in Figure 4 categorically:

	O_A	O_N	O_B
F_A	**4**	**1**	**0**
F_N	**2**	**2**	**4**
F_B	**0**	**3**	**2**

The hit score is the sum of the entries on the diagonal and is seen to be 8. The bias for predicting above-normal is (4+1+0)/(4+2+0) = 5/6 = 0.8333, for predicting near-normal is (2+2+4)/(1+2+3) = 8/6 = 1.3333, and for below-normal is (0+3+2)/(0+4+2) = 5/6 = 0.8333. Thus, for this model the above- and below-normal categories were being underforecast, and the near-normal category overforecast. It is not uncommon that statistical forecasts do not have the same variance as the observed data, and may therefore be variance adjusted by inflation the amplitudes of the forecasts in order to counter the near-perpetual forecasts of near-normal rainfall (or any other predictand). The adjustment could be done by dividing each of the forecasts of the training period by the Pearson correlation. However, the adjustment did not make much of a difference for the example presented here. The FAR for forecasts of above-normal rainfall is (1+0)/(4+1+0) = 1/5 = 0.2, for near-normal is (2+4)/(2+2+4) = 6/8 = 0.75, and for below-normal is (0+3)/(0+3+2) = 3/5 = 0.6. Thus, when the model predicts above-normal rainfall it should get it correct most of the times. This is less so for the other two categories where the FARs are approaching the value of one.

Other Statistical Methods

Discriminant analysis, cluster analysis, analogue methods, optimal climate normals and neural networks.