



**AN INTRODUCTION TO THE ART OF
AGROMETEOROLOGICAL CROP YIELD FORECASTING
USING MULTIPLE REGRESSION**

**René Gommès
FAO Senior Agrometeorologist, SDRN**

**Crop Monitoring and Forecasting Group
Crop Yield Forecasting and Agrometeorology Sub-Project
UTF/BGD/029, ASIRP/DAE
Dhaka**

9 June 2001

1. Introduction

Agrometeorological yield forecasting using a multiple regression always starts with a table of data containing yields and a series of agrometeorological and other variables which are thought to determine the yields

An example of such a table is given below (Table 1) with same data from for T-Aman in Rajshahi district between 1983 and 1998. We shall refer to this table as the **input matrix**.

Table 1: A typical input matrix for a multiple regression crop yield “model”

	HYV Yield Kg/Acre	July Rain mm	August Max.T deg. C	Sept. Min.T deg. C	October Sunshine hours	November Rel. hum. %	Excess Water mm	Evapotran- spiration mm
1983	810	366	31.7	26.2	6.2	76	1416	629
1984	817	294	32.3	24.9	8.1	72	806	712
1985	869	274	32.5	25.4	8.2	77	385	345
1986	864	262	33.2	24.9	7.5	82	907	631
1987	788	409	31.9	26.1	8.2	79	912	593
1988	777	301	32.4	25.6	8.5	77	445	473
1989	906	357	32.9	25.1	7.5	80	370	675
1990	925	413	33.4	24.8	5.9	79	392	643
1991	943	285	33.3	24.9	6.1	79	619	557
1992	966	238	33.1	25.2	7.5	79	132	575
1993	932	217	32.1	25.5	7	82	613	871
1994	866	171	32.2	24.8	7.7	76	375	609
1995	826	285	31.4	25.8	7.8	83	757	596
1996	952	105	32.5	26.1	8.1	80	486	633
1997	941	725	32.5	25.2	8.5	81	899	594
1998	872	404	32.8	26.2	6.6	84	638	661

A regression equation (usually linear) is derived between crop yield and one or more agrometeorological variables, for instance

$$\text{Yield} = 5 + 0.03 \text{Rain}_{\text{March}} - 0.10 T_{C, \text{June}}$$

with yield in tons Ha⁻¹, March rainfall in mm and June temperature in °C. Beyond their simplicity, their main advantages are the fact that

- calculations can be done manually,
- data requirements are limited.

The main disadvantages are:

- the ease of derivation of the equations using standard statistical packages or a spreadsheet

- their poor performance outside the range of values for which they have been calibrated, i.e. their inability to yield correct values in the event of extreme factors¹.

Because of the disadvantages listed above, multiple regression “models” sometimes² lead to nonsensical forecasts. The equation above, for instance, suggests that low March rainfall (a negative factor) could be corrected by below zero temperatures in June (frost), which obviously does not make sense. Another disadvantage is the need to derive a series of equations to be used in sequence as the cropping season develops.

Many of the disadvantages of the regression methods can be avoided when value-added variables are used instead of the raw agrometeorological variables³, as is done in the FAO method (see section below). Such a value-added variable would be, for instance, actual crop evapotranspiration, a variable known to be linked directly with the amount of solar radiation absorbed by the plant under satisfactory water supply conditions or light water stress.

Crop forecasting is as much art as science: with the same input data, some experts produce reliable and stable methods, while others come up with equations⁴ which the experienced eye can discard at the first glance.

Bad equations are produced when the blind application of statistics prevails over common sense and agronomic knowledge.

The present note tries to summarise some of the considerations which the crop forecaster should keep in mind when deriving multiple regression equations (so-called ***Yield Functions***) which will eventually be used for forecasting crop yields. The process by which the coefficients of a yield function are derived are known as ***calibration***.

Little attention is paid to regression techniques proper. The computations can be done with any statistical package, with EXCEL or with FAOMET.

An didactic example is given for the forecasting of T-Aman in Rajshahi. The diskette with the sample files is part of the present report.

Apart from the country-specific parts, the introductory sections and the lay-out, this reports draws largely from the manual on Crop Yield Weather Modelling prepared by the reporter in 1998 for WMO. The reader is referred to the mentioned manual for the bibliographic references, which were all omitted from the present note.

¹ Extreme factors are, **by definition**, factors or combinations of factors that occur rarely.

² I hesitated between “sometimes”, “often” and “mostly”

³ This only the first of several statements in these notes insisting on the fact that agrometeorological crop forecasting is an agrometeorological exercise, not a statistical one.

⁴ Strictly speaking, multiple regression “models” are no models at all. They are just equations. However, if the equations use variables (factors) which are very significant from a physiological and agronomic point of view, then an “equation” can come close to a model.

2. The FAO crop forecasting philosophy

It is suggested that the approach used by FAO and a number of developing countries for crop forecasting at the national level strikes a good compromise between input requirements and ease of validation⁵. The section thus describes the FAO crop modelling and forecasting philosophy.

The word “philosophy” is preferred to “methodology” because the position of FAO has been to propose a general framework of which the totality, or only some elements, can be adopted by the countries for their national crop forecasting methodology for food security. It is also felt that “philosophy” stresses the fact that, when operating in a field with many partners (economists, marketing experts, nutritionists, statisticians, demographers, *etc.*), the most serious problems are not technical but organisational and institutional⁶: co-ordination of the participants and integration of different sectoral approaches.

2.1 Flow of data

The flow of data is illustrated in figure 1. The left hand side of the figure (elliptic boxes) lists the sources of the data: the meteorological network, satellites, field observers (mostly agricultural extension staff) and national services dealing with soils (e.g. soil survey), crops (Ministry of Agriculture) and National Agricultural Statistics. The number of partners and the diversity of the data types creates some difficult as well as interesting problems which were described elsewhere.

Each of the sources may contribute one or more types of data (second column, rectangles). For instance, meteorological data can be provided, in addition to the *ad hoc* national network, by remotely sensed sources. Indeed, several methods are now routinely available which are used to derive or interpolate rainfall or sunshine data from satellite information.

The same applies to some crop data, for instance planting dates, which may be estimated, under adequately known conditions, from vegetation index (NDVI⁷) time series.

Based on the meteorological and agronomic data, several indices are derived which are deemed to be relevant variables in determining crop yield, for instance actual evapotranspiration, crop water satisfaction, surplus and excess moisture, average soil moisture... The indices (variables) then enter the yield function⁸ to estimate station yield. At this stage, the data are still station-based since most input are by station.

⁵ Validation, in this context, covers basically the statistical calibration of the model, as the underlying processes can hardly be verified at the considered scale.

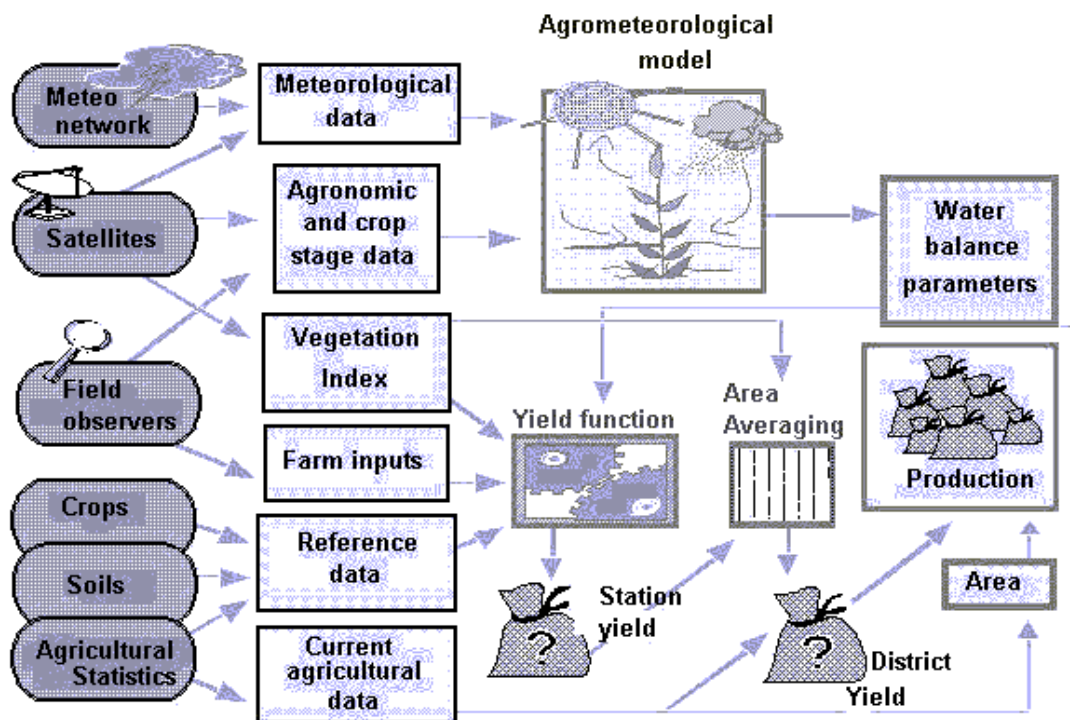
⁶ Just as an example of institutional issues, remember that four independent crop forecasts are produced in Bangladesh, by BBS, SPARRSO, FPMU and DAE. I said “independent” although this is not quite correct as the institutions mostly use the same data sources for their forecasts.

⁷ Normalised difference vegetation index. A satellite index which is roughly and not too far from linearly correlated with standing living biomass. Under normal circumstances, the condition of natural vegetation and crop condition are related.

⁸ The yield function is usually an equation, linear in most variables, which was obtained by multiple regression of a combination of time series and cross-sectional data.

It is stressed that the derivation of the indices above constitute a major difference with the process-oriented or simulation models⁹ : they constitute some of the **value-added variables** that will be used in the yield function. The word value-added is used because the variables derive from an agrometeorological analysis, which includes agronomic information and a number of soil and weather data. They are highly synthetic variables which are very superior to raw climatological variables.

Figure 2 : The flow of data in FAO-promoted crop forecasting systems for food security.



Station yields are then area-averaged using, for instance, NDVI as a background variable (see the SEDI methods below), possibly adjusted with other yield estimated provided by national statistical services, multiplied by planted area to yield a district production estimate¹⁰.

As indicated, according to countries, variants to this general scheme can be introduced at almost every step. The technical options were adopted mainly to reduce computing overhead and bypass, for the time being, some problems which are still difficult to handle in the context of developing countries. More details will be given below, but simple, even elementary solutions are sometimes preferable to complex solutions for which the necessary inputs are not available and must virtually be guessed. It is also suggested that a codified system and reproducible approach,

⁹ Of which the most popular are the Wageningen-WOFOST family, the Hawaii-CERES family (Gordon Tsuji, Gerrit Hoogenboom, Tony Hunt...) , the EPIC model of Texas University (Jim Williams), CropSyst of Washington State University (Claudio Stoeckle) and many others.

¹⁰ In practice, the situation is slightly more complicated as “station yields” have themselves been calibrated against agricultural statistics which are given by administrative units.

even if very far from perfect, is preferable to no system because it leaves relatively little room to the personal interpretations of the forecasting officers .

To illustrate the previous point: many countries estimate crop production¹¹ by calling a meeting of knowledgeable people (grain board, statistics, agrometeorological services) and, **through bargaining**, eventually reach an agreement on the current crop production estimate. No specific methodology is followed, and strong political bias - conscious or otherwise - is often a basic ingredient in the forecast.

Under such circumstances, any "system" which will avoid political bias and ensure at least a reasonable degree of consistency from year to year and from place to place is to be preferred.

2.2 Standard Technical options

The main technical options adopted in the FAO crop forecasting philosophy are the following:

- agrometeorological and remotely-sensed data are integrated at all levels whenever possible: at the level of data (rainfall, phenology) and at the level of products (area averaging of yields);
- gridding is done after modelling¹², under the assumption that there exist variables, such as NDVI, which are at least qualitatively linked to crop condition **in a given area**. If this assumption does not hold in quantitative terms over large areas is not relevant for the interpolation procedures adopted. This also assumes that such factors as soil fertility and the effect of greater soil water holding capacity is captured by NDVI;
- the time step mostly adopted is the decadal: all calculations are done at a ten-daily step, following the recommendation of WMO. 7-day periods are adopted in many countries of British meteorological tradition, although this poses some practical problems
- results are calibrated against agricultural statistics through empirical yield functions. **It is clear that the accuracy of the forecasts cannot possibly be better than the agricultural statistics used to calibrate them.** There is thus

¹¹ In most developing countries there are not many alternatives to agrometeorological crop forecasting, with or without remote sensing inputs. Some countries in the Sahel conduct rapid estimates based on interviews with farmers. **Other countries have developed biometric systems based on measured crop indices** (plant density, maize cob size). In some countries agricultural statistics are so uncertain that the agrometeorological forecasts are taken as final yield and production figures. The agrometeorological approach usually gives best results in semi-arid areas where the water deficit is the main limiting factor. It performs poorly in some mountainous areas and where (i) farming does not follow a homogeneous pattern, (ii) coverage by the weather stations is insufficient and (iii) water surplus, or pests and diseases, tend to be the main limiting factor(s). Simple statistical (trend) models perform very poorly in semi-arid countries, where the inter-annual variability of yields reaches very high values. This being said, after an initial spell of enthusiasm, the hope to use direct correlations between satellite indices and yields as a forecasting tool, was gradually abandoned. The methods worked only in few countries, if given the help of additional data collected at ground level.

¹² Gridding of actual data, for instance weather data for short time intervals, is the typical example where we feel that the available techniques have not reached the a level of reliability which would justify our transferring the methodology to national services in developing countries.

some uncertainty about the accuracy, 10% to 30% is probably a good guess. At the scale at which FAO works, e.g. districts, provinces, etc., models developed at the field level do not apply. The “agrometeorological models” mentioned in figure 1 are thus usually very simple. They aim more at assessing growing conditions through value-added “water balance parameters” than actually simulating crop-weather-soil interactions. It is, therefore, justified to use empirical yield functions which, in addition, avoid to touch on the most difficult issue of geographic scale effects;

- tools are modular, i.e. the crop forecasting system uses a number of software tools that carry the analysis from the data to the final production estimate. Depending on the local conditions, national services can choose between different tools (for instance for area averaging). Any specific tools can be changed without touching the whole structure of the system: the system remains light and easily upgradable and maintainable. This is facilitated by standardisation¹³ through common file names and structures and early reduction¹⁴ of RS images. What this means is that the users, who are responsible for carrying out the analyses and the forecasts, need not worry about the technical (remote-sensing technical) aspects of satellite inputs.

2.3 Suggestions for adapting the FAO methodology in Bangladesh

Most of the ingredients (data, tools) for the application of the FAO methodology are available in Bangladesh, and so is the expertise, even if it is spread FPMU, CMFG in DAE, SPARSSO, BARC, BS etc.

Given the relative homogeneity of Bangladesh climate, as well as the typical features of the agriculture, it is suggested that the considerations listed below are worth paying some attention to.

(1) **Radiation** is the main limiting factor for crop production for some crops (Aus, Aman), while water is more important for others (Boro). Minimum temperature is a good measure of night-time respiration loss. Variables standing for water balance, respiration loss and sunshine should, at least in the exploratory phase, be part of all crop forecasting systems;

(2) **Using the SEDI method, all model inputs (including weather) should be gridded** systematically by the CMFG in DAE, and averaged over the units for which the yield data for calibration are available. The area wide averages should be used for the calibration and the operational forecasts. This will solve once for ever the issue of the spatial distribution of meteorological stations and will not prevent the improvement of the grids when new stations become available;

(3) **The calibration should not be done with time series from one area only.**

The time series are typically short (about 20 years) compared with the number of regression variables (typically 30 to 40).

¹³ This issue was addressed by a recent meeting organised by FAO (FAO, 1995).

¹⁴ Image reduction here refers to the corrections (geometric, collocation, radiometric, etc.) which must be made on the images before they can be used for applications.

Upazilas or districts should be grouped into homogeneous areas. The addapix software and NDVI time series is the ideal tool to carry out this spatial clustering operation, in combination with the common sense of people knowing the areas under consideration. Remember: good local agronomic common-sense is more important in crop forecasting than statistical analyses or the beauty of satellite imagery.

By combining districts/Upazilas, the mix of cross-sectional and time-series data will provide calibration sets with large numbers of “observations” . This is not only necessary for better statistical significance, but also reduces the work-load as calibration has to be redone annually with the new data available from BBS.

(4) The calibration of the yield function (i.e. the computation of the coefficients using multiple regression) must be preceded by a principal components analysis. This has the multiple advantage of

- reducing the number of variables (which is good a basis for a meaningful statistical analysis) ,
- identifying how variables correlate and which are the relevant ones to be retained for the multiple regression, and
- showing which are the most significant factors affecting yields.

(5) Never do any statistical analysis without having first removed the technological trend from the yield series. Needless to say, the trend must then subsequently be added.

(6) never retain any variable which has too low a ration of Coefficient of regression / standard error of the coefficient.

3. Some methodological issues for consideration of crop forecasters

3.1 Scaling

The word “scaling” generally refers to the fact that we are working with data which belong to different spatial scales. There is also a time scaling problem (e.g. weekly observations and monthly normals), but this is not specifically dealt with here.

The most critical spatial scaling issue is related with the fact that most socio-economic data, including agricultural statistical data (yields etc) come by administrative units, while weather data are observed at stations (points). Still other information can be available by polygons, for instance soil features, planting dates and cropping patterns.

The issue of scaling is thus very relevant for crop forecasting. It is generally solved in operational context using geostatistical methods, i.e. methods that estimate the value of a parameter at co-ordinates xy based on the observations at more or less distant locations. If the estimates are done for regularly spaced points (a grid of points), the process is usually referred to as **gridding**. Based on the grid, it is then possible to compute an average value (estimate) over any spatial unit. This is known as **area averaging**. Note that a grid is also often called a “surface” by analogy with elevation grids which describe the terrain (digital terrain models).

The interested reader is also invited to meditate the links between gridding and the issue of missing data, one of the main obstacles to routine crop monitoring and forecasting. It is also suggested that users of all types of grids should be aware of the differences between grids (values at points) and pixels (average values over a larger area).

3.2 Spatial interpolation

As mentioned above, spatial interpolation of missing data consists in the estimation of the unknown values at one P point in space based on the known values at neighbouring points. Area averaging is the estimation of the spatial average, within a contour¹⁵, of a variable measured at several points. A typical example is the estimation of a district crop yield when the value has been actually sampled or computed at several points only, or the estimation of a temperature based on nearby stations.

Gridding, area averaging and the estimation of missing data from neighbouring stations are thus different facets of the same problem of spatial interpolation. Spatial interpolation has become a central issue in regional crop-weather model and yield forecasting, to the extent that many of the comprehensive crop modelling environments like DSSAT now include tools for geostatistical analysis. Gridded data are directly compatible with the Geographic Information Systems

¹⁵ A contour or a polygon. A district boundary is a typical polygon.

The available methods are many; they vary in their complexity, constraints on inputs, and computer implementation. In addition to the “historical” Thiessen and Voronoi diagrams, we can mention inverse distance weighting, kriging and co-kriging, thin-plate splines, etc. For some general and climatological references, refer to **cokrig** software and manual available from the FAO website (<http://artmet.fao.org>).

The spatial interpolation can either be purely geo-statistical, or take advantage of the additional knowledge obtained from external variables. A third approach, known as mesoscale modelling, uses the physics of the phenomenon to model its spatial behaviour.

In the first category, we mention the method known of inverse distance weighting and simple kriging. In the second, the method we can list co-kriging and Satellite Enhanced Data Interpolation (SEDI).

The purely geostatistical methods treat spatial interpolation as a statistical problem: the nature of the variables being interpolated does not matter. The second methods takes advantage of the correlation between the variable to be interpolated and some other variable (external variable), for instance the well known linear decrease of temperature with elevation. If a digital terrain model¹⁶ is available, the spatial interpolation can be significantly improved.

Crop forecasters are specifically warned against the blind use of generic interpolation software, such as **Surfer**. Such packages apply the same method, regardless of the variable being gridded. It is obvious that rainfall, temperature, soil moisture etc do not have the same spatial behaviours.

3.3 Inverse distance weighting IDW

The *inverse-distance weighting* is one of the most straightforward methods of spatial interpolation; it is simple and transparent, and provide the user good control over the process. It can be implemented with FAOMET/FAOCAST for tables of data with missing values.

IDW takes into account the distance d between the “known” and “unknown” points and their relative importance in the estimation. For instance, close-by points are assigned a higher weight than far-away ones. If the unknown value X_P at a point P has to be determined, the first step is to compute the distances between the point and all the points where the value X_i is known, subsequently discarding all the values beyond a certain distance and retaining only n neighbours. The user usually has the option to interpolate X_P only if the number of numbers is sufficient (for instance, $n > 5$). The distance between P and the n other points is d_i .

$$X_P = \frac{\sum_1^n X_i d_i^a}{\sum_1^n d_i^a}$$

¹⁶ A Digital Terrain Model is a grid of elevations.

in which the exponent a is determined empirically. The method has several advantages, including simplicity and the fact that only values in the range of X_i are determined. The main disadvantage is the lack of indication of the statistical error affecting X_p .

3.4 Satellite enhanced data interpolation (SEDI)

SEDI takes advantage of the correlation between the variable to interpolate and an environmental variable, for instance NDVI/biomass and agricultural yields. One of the ways to approach this is co-kriging, implemented in the FAO-UCL Cokrig software mentioned above.

The SEDI interpolation method originated in a Harare based FAO/SADC Regional Remote Sensing Project. It was originally developed to interpolate rainfall data collected at station level using the additional information provided by METEOSAT CCD. The methods proved powerful and versatile, and it is now regularly used to spatially interpolate other parameters as well (e.g. potential evapotranspiration, crop yields, actual crop evapotranspiration estimates, etc.).

The SEDI functions were incorporated into the WINDISP_3 software available from the FAO web site. A new, improved version was recently finalised (2001).

SEDI is a simple and straightforward method for 'assisted' interpolation. The method can be applied to any parameter of which the values are available for a number of geographical locations, as long as a 'background' field is available that has a negative or positive relation to the parameter that needs to be interpolated.

Three requirements are a prerequisite for the application of the SEDI method:

The availability of the parameter to interpolate as *point data* at different geographical locations (e.g. rainfall, potential evapotranspiration, crop yields). In the present case of statistical variables, they were assigned a co-ordinate corresponding to the centre of gravity of the administrative unit;

The availability of a background parameter in the form of a *regularly spaced grid* (or field) for the same geographical area (e.g. the above-mentioned NDVI variables, altitude).

A monotonous relation, **at least locally**, between the two parameters (*negative or positive*; Yield/NDVI is positive, temperature/altitude would be negative). A Spearman rank correlation test can reveal whether a relation exists, and how strong this relation is.

The SEDI method yields the parameter mentioned under point 1 as a field (i.e. an image covering the whole area under consideration).

Let us illustrate the method below using rainfall and CCD: it is implemented in three steps (i) extracting CCD values from the satellite image and calculating the ratio of point and image values; (ii) gridding the ratios to form a regularly spaced grid, using any method, for instance inverse distance weighting or kriging; (iii) multiplying the grid of interpolated ratios with the grid of CCD (image) pixel by pixel to obtain an estimated rainfall grid (image).

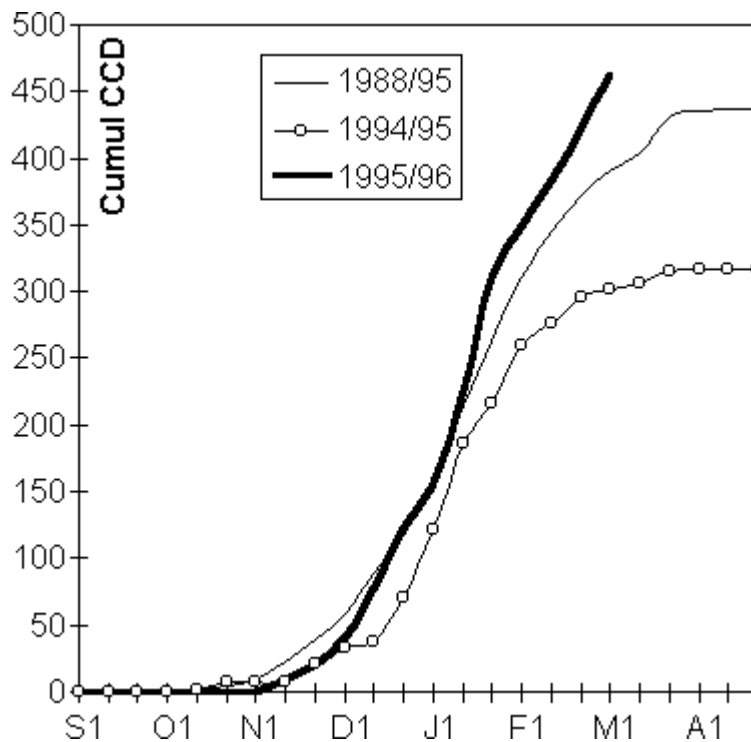
There are several variants of the interpolation method thus described, for instance the one described by Herman et al., 1997. The authors include an estimate of orographic rain based on clouds with a relatively warm top. Note that the described methods apply only to tropical conditions.

3.5 Cold Cloud Duration (CCD)

Cold Cloud duration is used in combination with SEDI to create rainfall grids. If CCD is not available, basically ant cloud index or variable can be used with the SEDI technique, provided the images are available in the IDA7WEINDISP format.

CCD is determined using the geostationary satellites of the METEOSAT or GMS types. Due to the low temperature threshold corresponding to high elevations, there is relatively very little atmospheric effects to be corrected in CCD when compared with NDVI.

Figure 1: Accumulated CCD (hours) in central Malawi during two cropping seasons (1994/95 and 1995/96) in comparison with the reference period 1988/1995. The abscissa covers the period from the first dekad of September to the end of August in the following year.



GMS-4 is a Japanese satellite in geostationary orbit over the equator at approximately 140E. The satellite is equipped with the Visible and Infrared Spin Scan Radiometer (VISSR) imaging sensor, which uses the spin motion of the satellite to scan the earth in the East-West direction. GMS begins a North-South scan every hour on the half hour, with four additional scans daily for wind estimation. At the vertical of the satellite, the visible (0.5-0.75 μm) channel has a resolution of 1.25 km and the infrared (10.5-12.5 μm) channel has a resolution of 5 km. This gives

approximately 10,000 visible and 2,500 infrared lines and samples for each full-globe image.

METEOSAT provides weather oriented imaging of the earth. Like GMS, it covers visible and infra-red wavelengths sampled at half-hourly intervals.

CCD has been used extensively in a food security context to estimate rainfall using various techniques. It is defined, for each pixel and for a given period (usually ten days), as the number of hours during the temperature was below a “cold” temperature threshold around $-40\text{ }^{\circ}\text{C}$, which corresponds to convective clouds with high vertical extension assumed to produce rainfall. The technique has been used to estimate rainfall with good results in tropical countries only.

Although the relation between the solar radiation reaching the ground and clouds is far from direct, the development of more or less empirical methods is progressing, usually with much better results than with rainfall, among others because the role of clouds in radiation interception is far more direct than in rainfall production. In addition, the methods, once they have been properly calibrated, apply in tropical and temperate countries alike. It is suggested that the CMFG in DAE experiment with the available GMS data to develop a method to estimate radiation based on ground data and the cloud imagery, using SEDI.

The original approach was to try and estimate rainfall based on CCD only, assuming a constant intensity R_i (mm hour^{-1}). Because of the large spatial and temporal variability of R_i , the method is now being replaced by more or less real-time calibration against ground data, using CCD as an auxiliary variable in the spatial interpolation of raingauge measurements.

The major methodological issues regarding rainfall estimation and CCD can be listed as:

- ☞ rains are known exactly only for a given duration only for a very limited area around raingauge. According to the period covered (hours, days, dekads, months), the radius within which a raingauge provides a representative sample varies from a couple of hundred meters to 200 km;
- ☞ CCD indices cover a METEOSAT or GMS pixel (about 50 km^2 at the equator), and correspond to a discontinuous sample in time (one observation every 30 minutes);
- ☞ a plot of rainfall as a function of CCD thus compares two rather different variables, both of which are used as proxies for a third unknown variable, the average pixel rainfall. One of the consequences is a rather poor correlation¹⁷ between CCD and rainfall, and usually not usable for rainfall estimation. Newer and significantly more efficient techniques are now available (see 5.3.3);
- ☞ for short time intervals (one day), an additional difficulty is the difference between the time covered by rainfall measurements (09 GMT to 09 GMT the next day) and the satellite images.
- ☞ It remains however that CCDs are associated with rainfall and that they provide a useful monitoring tool, as shown in figure 21. This figure covers one of several

¹⁷ There is a theoretical limit to the correlation coefficient : about 0.7, i.e. about 50% of the variance can be accounted for.

“homogeneous” rainfall units of the SADC region which are regularly published by the regional monitoring system.

3.6 Observations of microwave satellites

Radar satellites have several advantages over optical satellites, at least in theory. To start with, they “see” through clouds, which is very useful for the monsoon period in Bangladesh when cloud cover can be almost permanent. Next, they are ideal tools for seeing water, and their use in a flood prone country need not be elaborated on.

But the use of satellites for the direct estimation of surface moisture involves a number of difficulties. Active microwave (or radar) satellites operating in the centimetre range of wavelengths are relatively unhindered by clouds, and satellites such as ERS-1 and JERS-1 have been providing images of the earth since the early nineties. Radar is an active sensor in that it emits a beam of energy which is analysed after having been scattered back by the surface: it provides information about the surface, either crop canopy structure or soil surface. Regarding crops, active microwave responds well to row spacing and orientation, and even to leaf orientation. Little operational use has so far been made of the technique because of its large sensitivity to such factors as wind effects on the surface, including leaf orientation!

Much hope is placed in the technique to estimate soil surface moisture directly, and possibly crop water content for plants with a planophile or near-planophile leaf distribution. For the estimation of soil moisture, refer to Wagner et al., 1999a, 1999b. Regarding the use of radar remote sensing, often combined with visible imagery, for crop modelling and yield forecasting.

Passive microwave measure the centimetre radiation emitted by the surface. It is used to determine the brightness temperature¹⁸, or effective temperature which can be used in biomass estimations.

¹⁸ Brightness temperature is the temperature of a blackbody radiating the same amount of energy per unit area at the wavelengths under consideration as the observed body.

4. Multiple Regression crop forecasting in practice

4.1 The golden rules of multiple regression crop forecasting

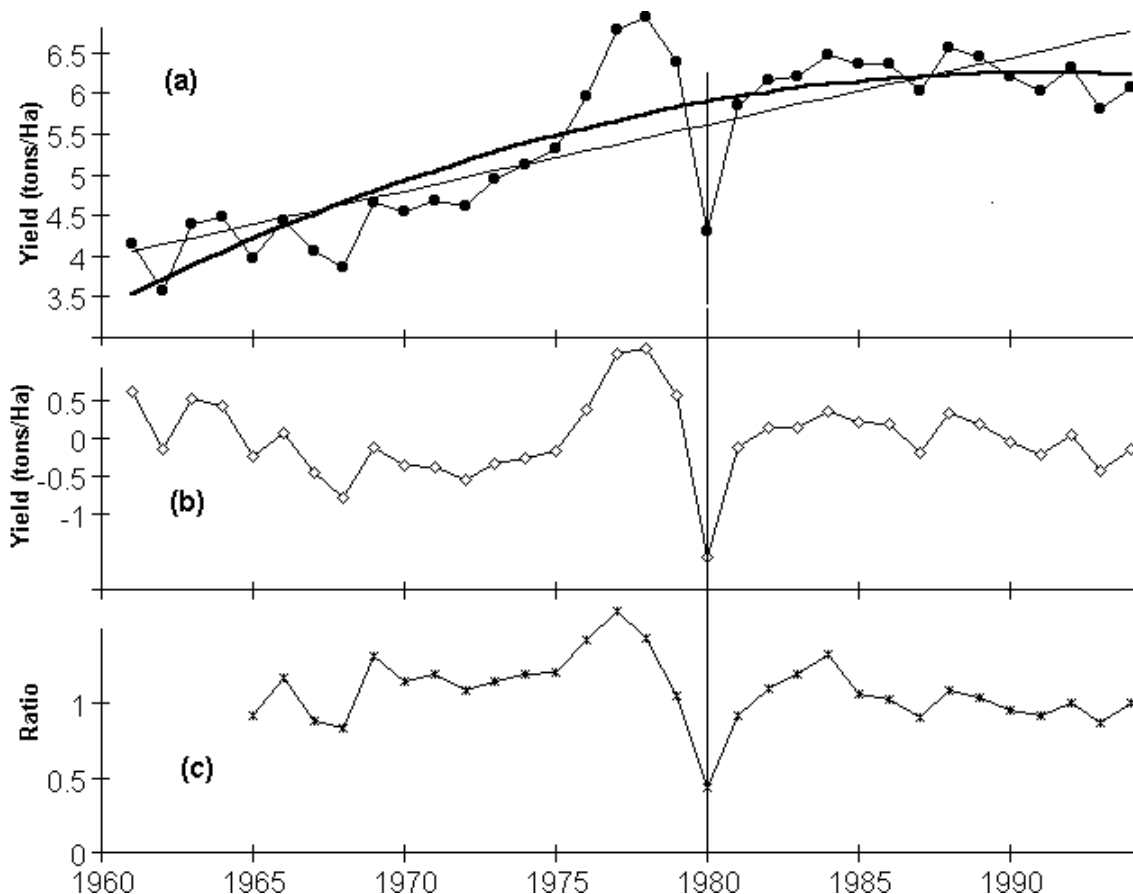
- Use only variables which are known to be meaningful for the crop under consideration. When there are good reasons to suspect that the response of crop production to a given variable is not linear, use a quadratic term in addition to the linear term.
- Retain only those variables for which the coefficients are significantly different from 0. This is to say that the regression coefficients must be significantly larger (absolute values) than their standard errors. This can be tested statistically (ratio of coefficient to its error), but common sense is usually enough.
- The sign of the coefficients must correspond to what is known about the response of the crop to the variable considered. This applies also to the quadratic terms.
- The coefficients must be spatially coherent, which is to say that they must vary smoothly over adjacent districts
- The quality of a regression equation is given, in addition to the statistics (R , R^2 , coefficients significantly different from 0...), by the average error of estimated yields
- Trends **MUST** be removed before carrying out the regression work proper. The trends need not be linear.
- Be aware of the fact that there are two types of variables: continuous-quantitative ones (e.g. minimum temperature affecting crops through night-time respiration) and qualitative ones (e.g. male sterility induced by high temperatures) .
- Always use a variable which stands for the local yield potential
- A yield function does not have to be linear. In some cases, a multiplicative function can be more appropriate.

4.2 Some good practice advice

- Compute the correlation matrix between all variables to get a better feel for the redundancy of the information...
- Plot the yield against time, to get an idea of the shape of the trend
- Run a Principal Components analysis to realise how redundant your data set actually is, and to identify the most important factors. Run the PCA twice (1) excluding the yield as a variable, to get a feel for the variable groupings and redundancy and (2) with yield to identify the variables which are associated with the yield, as well as those that are irrelevant.
- After removing the trend, plot de-trended yield against each individual variable to see the shape of the regression curve and the strength of the statistical correlation, if any

- As far as possible, ignore redundant variables or use the regression through a principal component analysis. Always prefer techniques with (annual or “automatic”) addition of variables to techniques with deletion of variables
- Use techniques to ensure the stability of the coefficients (randomly or systematically eliminating up to 50% of the observation points of the time series)
- Use jack-knifing to determine the actual accuracy of the method
- Yield functions typically “expire” after a couple of years... after which they need recalibrating. **A yield function older than 3 years is definitely worthless!**

Figure 3 : Yield of total paddy in the Korean Republic between 1960 and 1994 (based on FAO statistics). The top curve (a) indicates the actual yields with their linear and quadratic trends; the middle curve (b) is the detrended yield, i.e. the difference (residual) between actual yield and the quadratic trend; the lower curve (c) shows the ratio between the yield of year N upon the average of the 4 years from N-1 to N-4.



4.3 Detrending yields

In Bangladesh, a very large percentage (80-95%) of national cereal production is accounted for by trend, i.e. mostly the technology component, in particular HYVs. At district level, the trend component is less, but still significant (up to 50% of the inter-annual variability).

As the trend does not depend on weather, it must be removed before subjecting the data to any analysis. This is known as de-trending, and the resulting yield is said to be de-trended.

There are several techniques to resolve this problem when there is no weather trend, such as the one that occurred in the West African Sahel between 1960 or so and 1984.

One can, for instance, detrend the yield series (Figure 2). The example (Korean Republic) shows a typical upward trend due to improved technology (varieties, management, inputs) as well as the linear and quadratic trend. The coefficients of determination¹⁹ amount to 0.74 and 0.71, respectively. The coefficient achieved with the “best” trend model (a sigmoid, not shown) amounts to 0.80. Within the remaining 20%, weather accounts probably to about half.

The sharp drop in 1980 was due to severe low temperatures around the heading stage through early ripening stage. Tong-il varieties, high yielding hybrids, are very sensitive to abnormal cool temperature at that stage due to the failure in pollination. In the late 70s, the weather had been mostly favourable to rice cultivation, especially to Tong-il type (Byong-Lyol Lee, personal communication).

The middle curve shows the detrended yield (using the quadratic trend). This is the yield that will be used to calibrate a regional crop forecasting model. The lower curve shows the ratio between the yield of the current year and the average of the yields of the 4 preceding years, assuming that the trend is not significant over such short period. The advantage of this approach is that no trend has to be determined, and no hypothesis has to be made about the shape of the trend.

When there is a marked trend in the weather variables, it is probably better not to detrend the time series, but to add time as one of the variables in the calibration process, which we can write as

$$Y_y = Y_0 + f_1(y) + f_2(\text{simulation}) + e \quad (64)$$

where Y_y is the yield of year y computed from a function f_1 of time and a function f_2 of simulation model outputs. An additional reason for adopting this approach is that management is both time and weather dependent.

4.4 What are the “good” forecasting variables

The best variables are those with added agronomic value, such as actual evapotranspiration ETA, calculated soil moisture, radiation, rainfall, temperatures, fertiliser use, maximum yield. They can be used in a “raw” form, or they can be modified to come closer to their actual impact on yield. The Raw variables are usually quite sufficient for any type of exploratory work.

Maximum yield (Y_{max}) is a somewhat tricky variable. It is used to account for differences from place to place in the production potential. It is thus an absolute must when both cross-sectional and time series data are used in a calibration, for instance,

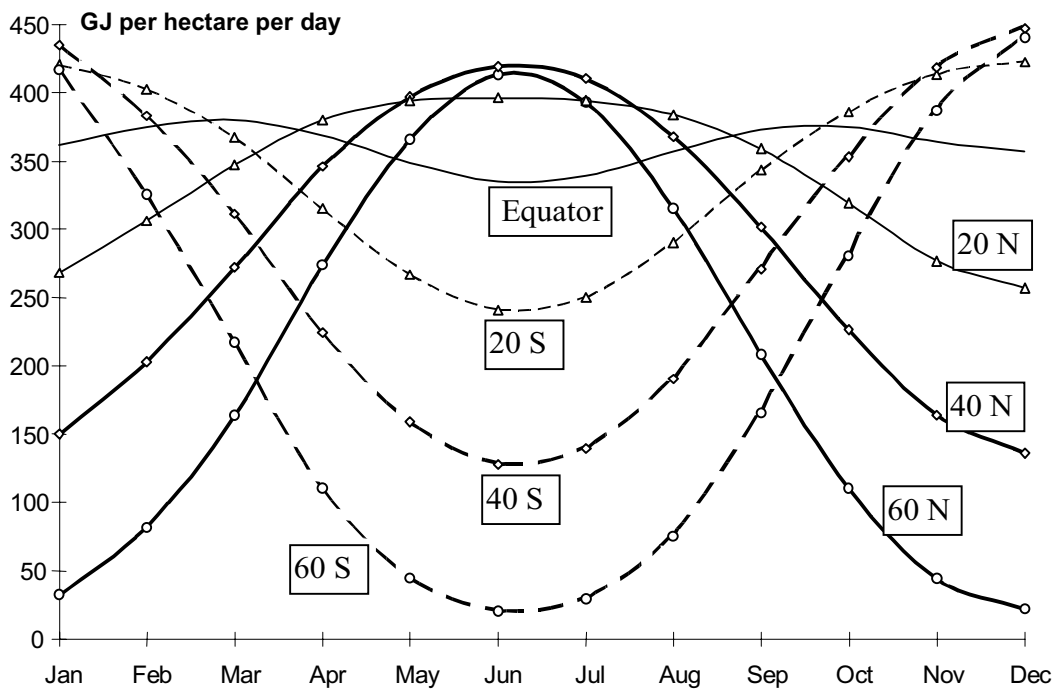
¹⁹ The coefficient of determination is the square of the coefficient of correlation. It expresses which percentage of the variance is accounted for by the trend.

when attempting to derive an equation that would be valid over large areas. To some extent, ETA can play the same role as Ymax, but when non climatic variables (soil etc) play an important part, Ymax should be used as one of the variables

4.4.1 Radiation

Radiation data are not always available; they can conveniently be replaced by sunshine hours or (better) sunshine fraction. This is because, at the latitude of Bangladesh, and within the same season, radiation and daylength can roughly be considered to be constant (Figure 4). For a more advanced regression equation, the expert could decide that the proposed approximation is not good enough, but the constancy of radiation and daylength can safely be adopted for the exploratory work

Figure 4 : Global radiation at the upper limit of the atmosphere as a function of month and latitude



If still more sophistication is required and justified²⁰ the expert could try and remember that net carbon dioxide absorption during photosynthesis follows a characteristic light response curve (Figure 5) given by de Wit as a function of absorbed photosynthetically active radiation (PAR), R_{HC} :

$$F_n = F_d + (F_m - F_d) \left(1 - \exp\left(-\frac{E_v \cdot R_{HC}}{F_m}\right) \right)$$

where the symbols have the definition and units given in table 2.

²⁰ Which I suggest is unlikely in the real world!

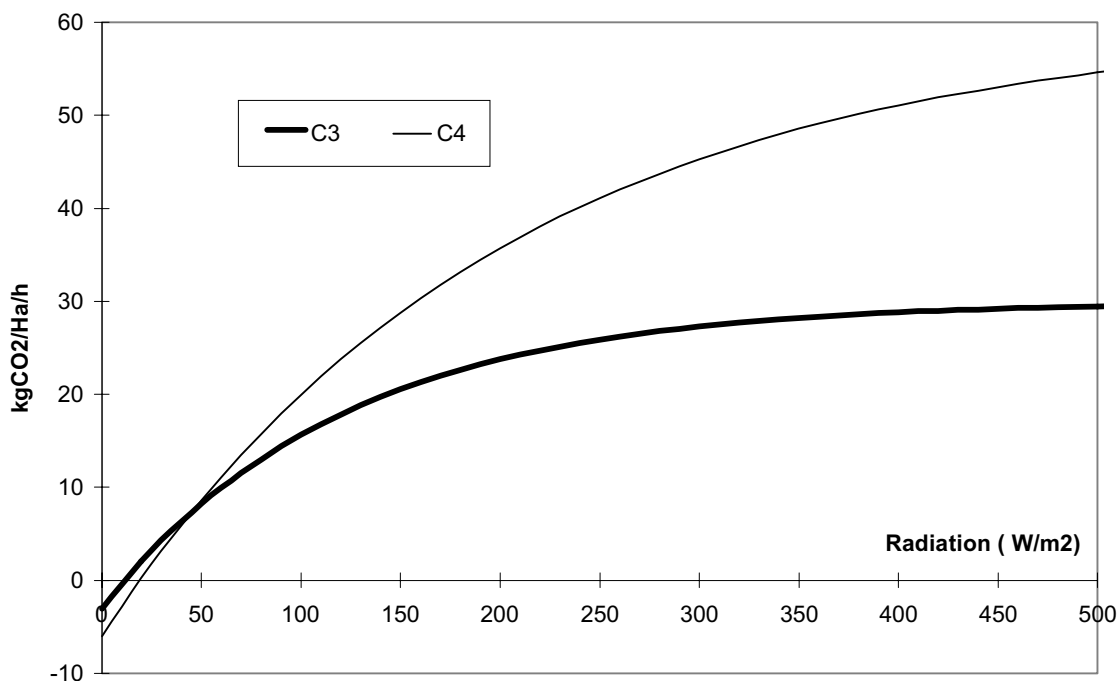
Note that equation above gives short-term net assimilation in $\text{Kg CO}_2 \text{ Ha}^{-1} (\text{leaf}) \text{ hour}^{-1}$, thus assuming a huge horizontal leaf with an area of 1 Ha. The radiant flux in the 400-700 nm range corresponds to the so-called Photosynthetically Active Radiation (PAR), roughly equivalent to 50% of global radiation at ground level²¹. The Efficiency at the light compensation point is the slope of the F_n curve at the light compensation point.

The curve below illustrates the equation. Above the light compensation point, assimilation exceeds respiration, and there is thus a net biomass accumulation.

The majority of plants fall into one of the C3 or C4 categories according to the number of carbon atoms of one of the first acceptors of CO_2 during photosynthesis. C3 plants respond well to increased atmospheric CO_2 , while C4 plants, mainly tropical grasses (maize, sorghum, sugarcane, millet) or halophytes, respond better to higher temperatures. Most agricultural plants (cereals, legumes, vegetables) are adapted to lower temperatures and sunshine; they belong to the C3 group. WUE tends to be higher in C4 plants.

C3 plants respond better at low intensities, but rapidly reach light saturation; C4 plants, which are more typical of tropical regions and include millet, sorghum, sugar cane and maize, continue increasing the assimilation over a much large range of radiation.

Figure 5 : Net assimilation of carbon dioxide per hectare of leaf in C3 and C4 plants as a function of absorbed PAR (R_{HC}). The values read from the graph for C3 plants at 400, 40 and 4 W m^{-2} are 28.82, 6.35 and $-1.92 \text{ KgCO}_2 \text{ hectare}^{-1} \text{ hour}^{-1}$, respectively. For 220 and 100 W m^{-2} , they amount to 24.72 and $15.66 \text{ KgCO}_2 \text{ hectare}^{-1} \text{ hour}^{-1}$.



²¹ Again... the sophisticated expert can decide to derive net radiation at the ground as a function of its astronomical value using angstrom's formula.

Table 2: Some characteristics of photosynthesis as driven by light, and the orders of magnitude of the ecophysiological characteristics for C3 and C4 plants.

			C3-plants	C4-plants
Net assimilation	Kg CO ₂ / Ha leaf / hour	F _n		
Maximum rate of net assimilation	Kg CO ₂ / Ha leaf / hour	F _m	30 (15 to 50)	60 (30 to 90)
Net assimilation in the dark	Kg CO ₂ / Ha leaf / hour	F _d	-3	-6
Absorbed radiant flux in the 400-700 nm range	Joule / m ² / s	R _{Hc}		
Efficiency at light comp. Point	Kg CO ₂ / Joule	E _{lc}	0.25	0.30
Temperature-dependent F_m ?			No	Yes

4.4.2 Actual evapotranspiration ETA

de Wit was among the first who recognised in the mid fifties that there is a direct link between transpiration and plant productivity. Transpiration can be limited due to short supply of water in the root zone, or by the amount of energy required to vaporise the water. It can be said that plants growth (biomass accumulation) is driven by the available energy, but that plants pay for the energy by evaporating water. This one of the basic “dogmas” of agrometeorology. Any crop forecasting system that does not incorporate ETA is unlikely to produce good results.

Maximum evapotranspiration (LE_m)²² and maximum assimilation (F_m) occur when the stomates are completely open.

We define the relative evapotranspiration as $Q = LE / LE_m$ and the relative assimilation as $R_{ass} = F / F_m$.

A plot of relative assimilation R_{ass} as a function of relative transpiration Q is given in Figure 6. It appears that, when Q values are relatively high (at least Q>0.6), and if other effects can be assumed to be constant, the relative assimilation is directly related to relative evapotranspiration.

$$\text{Daily biomass accumulation} \propto K * \text{ETA}$$

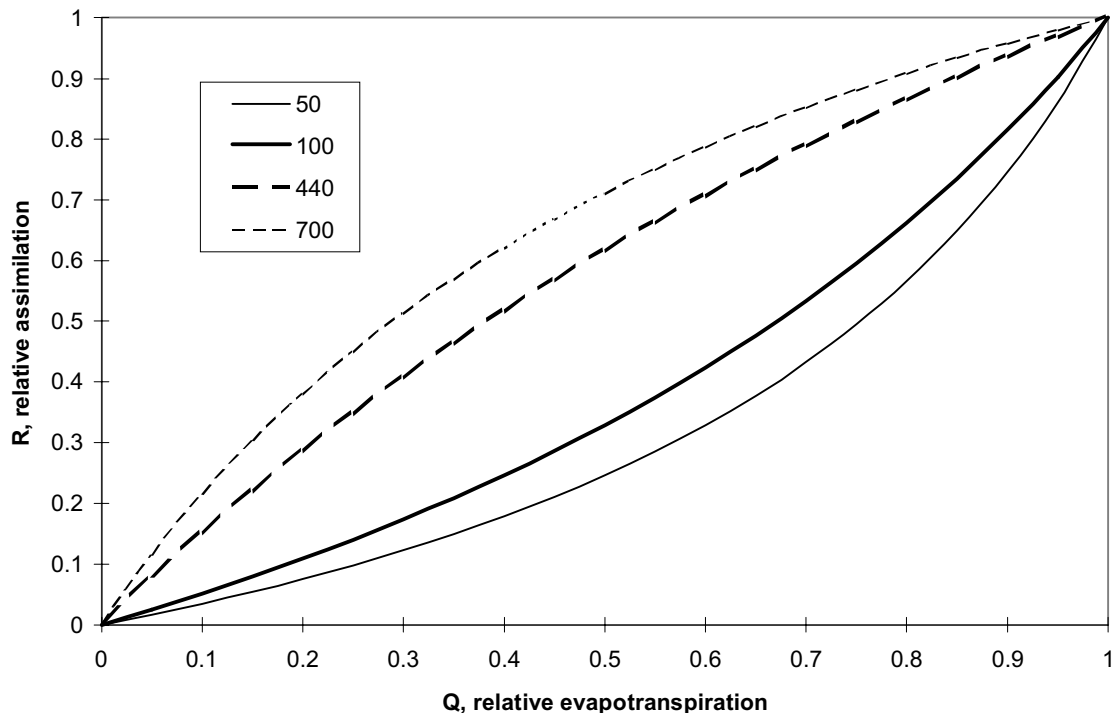
²² LE, is thus the evaporative heat loss (J m⁻² d⁻¹), the product of E, the rate of water loss from a surface (kg m⁻² d⁻¹) and L, the latent heat of vaporisation of water is 2.45 10⁶ J kg⁻¹.

ETA is one of the best forecasting variables in absolute, because, as indicated above, it is directly related to biomass production, but also because of its synthetic nature: it also includes radiation as one of its main driving forces. It is strongly recommended to include actual ET as one of the variables in any serious crop forecasting using multiple regression.

Of course, ETA must be estimated using a water balance.

Since we deal with crop forecasting, the water balance must be carried to the end of the cropping cycle, i.e. it must be based on actual (past) data as well as on future data. In fact, the main difference between crop modelling *per se* and yield forecasting is the fact that a forecast needs future data, i.e. an estimation of weather data between now, the time of the forecast, and the harvest .

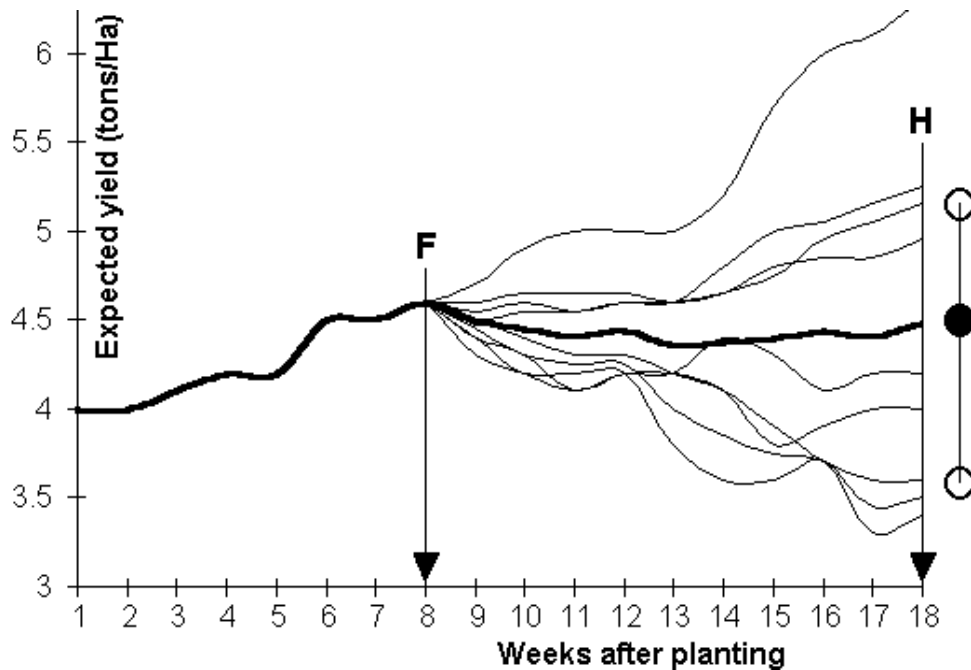
Figure 6 : The shape of the relation between relative assimilation R_{ass} and relative evapotranspiration Q depends on several variables, such as mesophyll resistance (values of 50, 100, 440 and 700). It is not useful to enter into more details here, just observe that the relation is roughly linear for non-water-stressed situations.



One of the difficulties of using a water balance model (or any deterministic model) for crop forecasting is the fact that yield has to be estimated before the end of the cropping season. This is to say that some of the data (the “future” data) are not known yet, as illustrated in figure 7.

Several techniques can be used: either one uses “normal” weather, or the historical data that have occurred between F and the time of harvest, or one uses a random weather simulator. The two last approaches are preferable in that they provide not only a yield estimate, but also a confidence interval.

Figure 7 : Yield forecast F at week 8, and harvest H at week 18. The black dot represents the estimated value, together with its confidence interval.



4.4.3 Maximum yield Ymax

The rationale for using Ymax was given above. Again, several techniques have been used. One of them consists in taking the average of the three or five highest yields ever achieved in the area. Note that this obviously does not refer to detrended yields, but to actual recent yields.

Other, more theoretical techniques are possible. For instance, the use of one of the numerous methods to derive climatically possible yield.

An interesting equation is given Uchijima and Seino, the “Chikugo” model. It involves several terms of the water balance: radiation and rainfall. It is written

$$NPP = 0.29H \cdot e^{(-0.216 RDI^2)}$$

where H is the annual net radiation (Kcal cm⁻²) and RDI is the “radiative dryness index” defined as the ratio H/(L.Prec) between annual net radiation and the product of L and Prec, L being the latent heat of evaporation (580 cal/gH₂O) and Prec annual precipitation in cm. RDI expresses how many times the available energy can evaporate the rainfall.

Converted to SI units, this becomes

$$NPP = 6.938 \cdot 10^{-7} H e^{\left(-3.6 \cdot 10^{-14} \left(\frac{H}{Prec}\right)^2\right)}$$

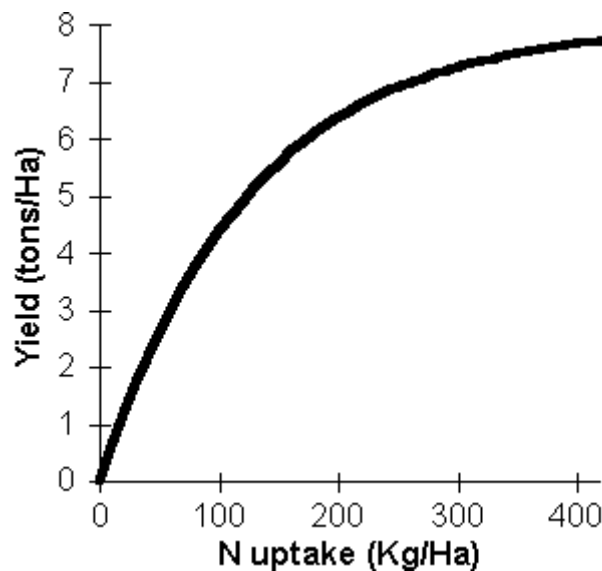
with NPP in g (DM) m⁻² year⁻¹, H in J m⁻², Prec in mm (equivalent to Kg m⁻²). Refer to the exercises for more details (7.4).

This equation would prove useful in Bangladesh where there are few temperature (frost) limitations.

4.4.4 Effect of fertiliser

There are several techniques to quantify the effect of nitrogen uptake on yields in multiple regression models. One of them is to use fertiliser amounts as a linear term. However, the response curve of yield to nitrogen is of the saturation type (figure 8).

Figure 8: Response of crop yield (Tonnes Ha⁻¹) to nitrogen uptake [Kg (N) Ha⁻¹]



It is unlikely that actual fertiliser use in Bangladesh will soon reach the plateau of the curve. Therefore, it will be sufficient to assume a linear response to fertiliser.

4.5 Uncertainty analysis: reliability, accuracy, precision and bias

Uncertainty analysis examines the sources errors in model outputs. While this is a straightforward issue with regression models, it is still worth paying attention to the points below.

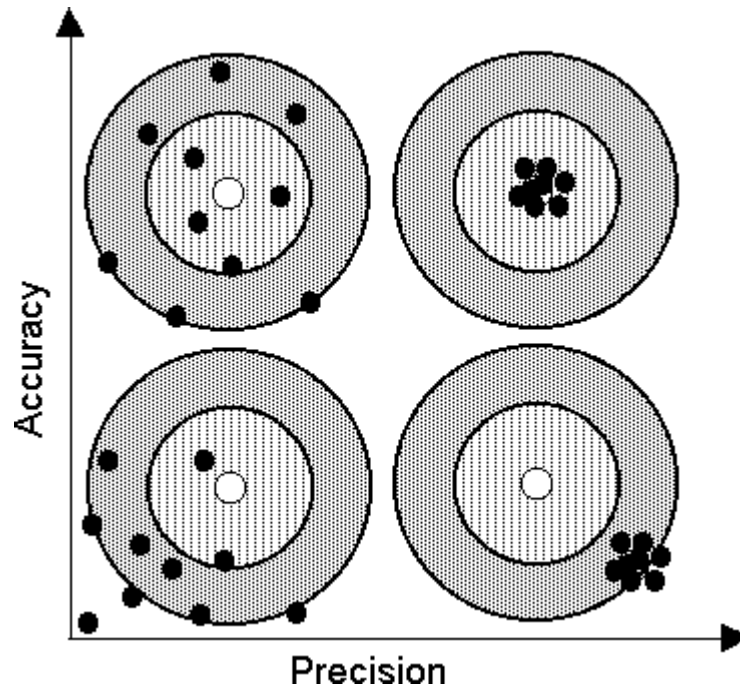
In Bangladesh BBS indicated an accuracy of about 6% for rice statistics at the national level, and 10% at the district level. If one considers that about 90% of the variability of the yields is accounted for by the technology component, this leaves only 10% to be explained by the weather... assuming that all of the residual variability is actually due to weather.

The accuracy of the agrometeorological forecast cannot possibly be better than the accuracy of the data used for calibration (the “training data”).

Uncertainty analysis thus deals mainly with model reliability, a term which encompasses both *accuracy* and *precision*. They are essential criteria in the empirical world of crop forecasting where the objective is mainly one: ensuring that the model accurately forecasts yield.

It should be remembered that however good a model, it will never be able to reproduce the total variability of the input yield data. But we should not forget that we should also value a forecasting technique by estimating the error associated with the estimates: this is not normally computed by statistical packages!

Figure 9 : A graphical representation of the differences between accuracy and precision.



A model output is accurate if in the long-term or under different conditions its average is close to the actual average; the difference between the averages is the bias. A model is precise if there is little dispersion of the outputs. This is illustrated in figure 9 using a presentation similar to that adopted by Sakamoto.

Low accuracy is probably easier to correct than low precision; the former can be corrected through the identification of some error in the model variables or parameters. Low precision is more difficult to tackle, as its source is more likely to be in the input data, including their selection, than in the model proper. In fact, the accuracy-precision discussion is useful in that it allows some identification of the source of the errors.

As mentioned above, it is a rather common observation that all models, be they deterministic or statistical, tends to underestimate the variability present in actual cropping. In other words: models tend to simulate average conditions more often than desired.

We conclude this section with a note on extreme events: as it is unlikely that a model will ever have been calibrated with very unusual²³ inputs, it is equally unlikely that it will be able to capture the impact of the extreme conditions.

²³ This is the definition of an *extreme* event.

4.6 How to assess the relevance and robustness of a yield function?

Below are some empirical rules that can be used to decide if a yield function is a good forecasting tool. We do assume that all the verifications regarding agronomic and statistical significance, as well as average error of the estimated yields have been worked out.

- Once you have carried out the calibration of the yield function, split the input matrix in two halves covering an equal number of years, for instance the first years in one matrix, and the recent years in the second. The matrices can also be split more or less randomly. Then recalculate the coefficients of regression and verify that they are similar in the two sets. If they are not, different factors are obviously at work in the two sets, which casts some doubt about the selection of the variables. Needless to say, this exercise has to be done with detrended yields!
- Verify that the coefficients for the yield functions vary smoothly over adjacent areas.

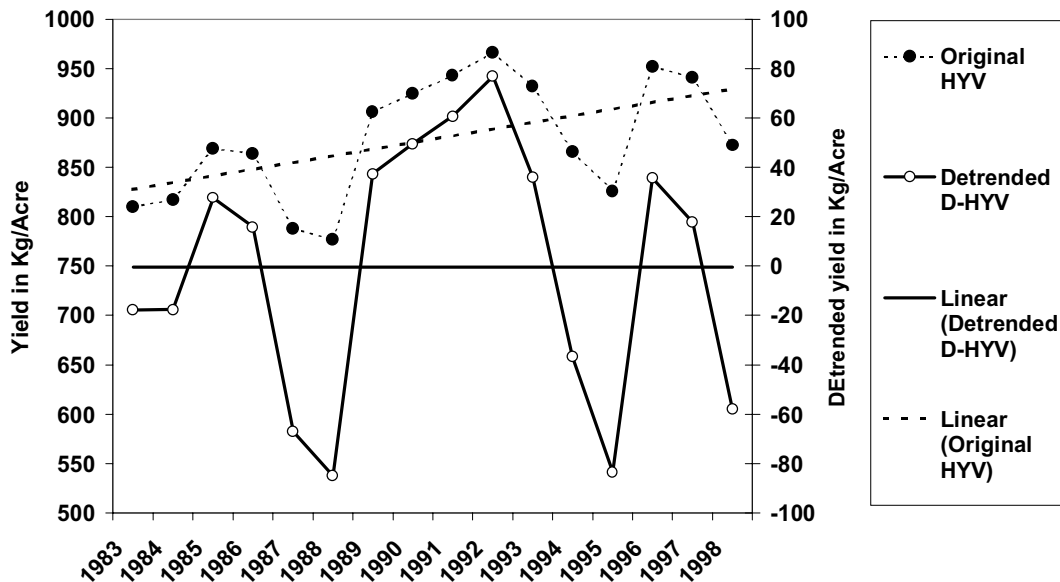
5. A case study: forecasting T-aman in Rajshahi

5.1 Input matrix and yield trends

The input matrix is given in file **Summary data 8398.xls**. The available variables include monthly averages/totals of Rainfall (**Rain**), Maximum and minimum temperature (**T_x** and **T_n**), sunshine hours (**Sh**) and relative humidity (**HR**) for the months of July through December (5 variables for 6 months, i.e. 30 variables). Together with the water balance parameters (see below), the total number of variables thus amounts to 34. They are listed in Annex I.

The trends were computed based on the input matrix. They account for 28 % of the yield variations of HYV and 12 % of the local variety (Figure 10). Refer to the spreadsheet for details.

Figure 10: T-Aman yield, trend and detrended yields, in Kg/Acre, between 1983 and 1998 in Rajshahi.



5.2 Water Balance

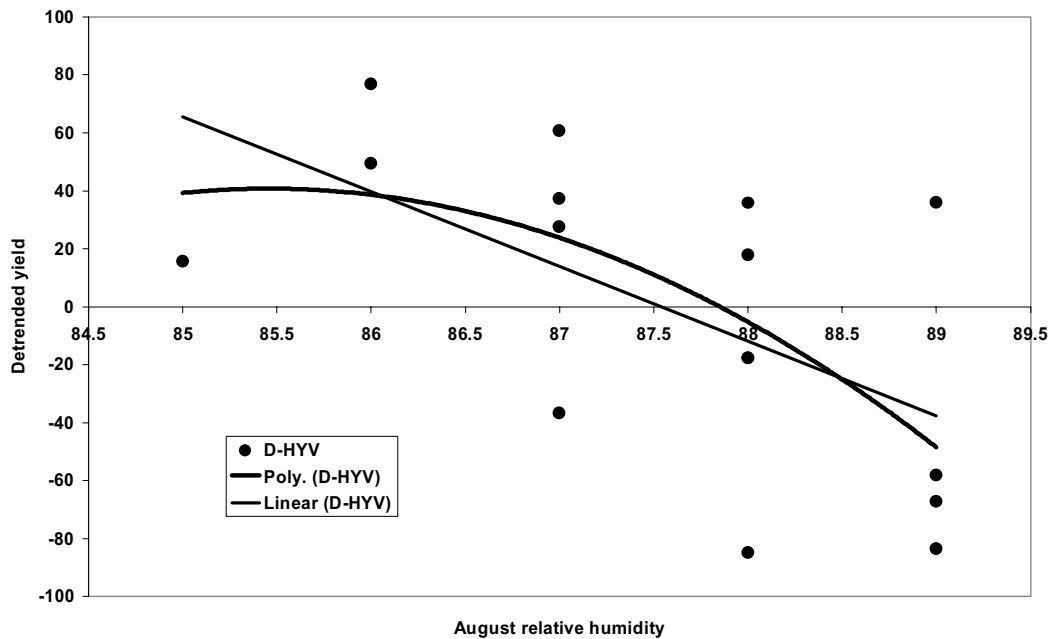
The water balance calculations were done with FAOINDEX, the parameters and the files given in subdirectory **WABAL**. The water balance parameters were extracted from file **Bdwx01xa.dat** and copied to the input matrix. They include **IndxLast**, the final value of the FAO water satisfaction index, **EXWT** total excess water in millimetres, **DEFWT** the total of dekad deficits and finally **ETA**, the actual crop evapotranspiration.

5.3 Straight multiple linear regression

The statistical work related with the standard multiple linear regression is given in file **Work stats 8398.xls**.

Before embarking on the calculations of the regression, a correlation matrix was computed between all the variables, and the variables with the highest correlation with detrended yield were identified. Yield was plotted against each of them individually to assess the shape of the relationship. An example is given for relative humidity in August (figure 11)

Figure 11: Rajshahi T-Aman yield as a function of August relative humidity, together with linear and quadratic trends.



For the 5 most correlated variables, a regression was computed yielding the result shown in table 3.

Table 3: Regression coefficients, standard errors and the t-stat (ratio between the regression coefficients and their standard errors).

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>
Intercept	-16001.25014	63876.46407	-0.250503067
Tx8	41.31125996	25.21611275	1.6382882
Tn11	-13.96437662	10.86483306	-1.285282207
Sh8	-1.660430979	21.12267193	-0.078608946
HR8	340.1707596	1471.955165	0.231101305
HR8*2	-1.977196912	8.482618744	-0.233088032
ETA	0.469277794	0.288989491	1.623857648

Note that HR8 was taken as both the linear and the quadratic term. Of the 6 variables, only 2 turn out to have acceptable coefficients of regression: August maximum temperature and ETA, possibly November minimum temperature.

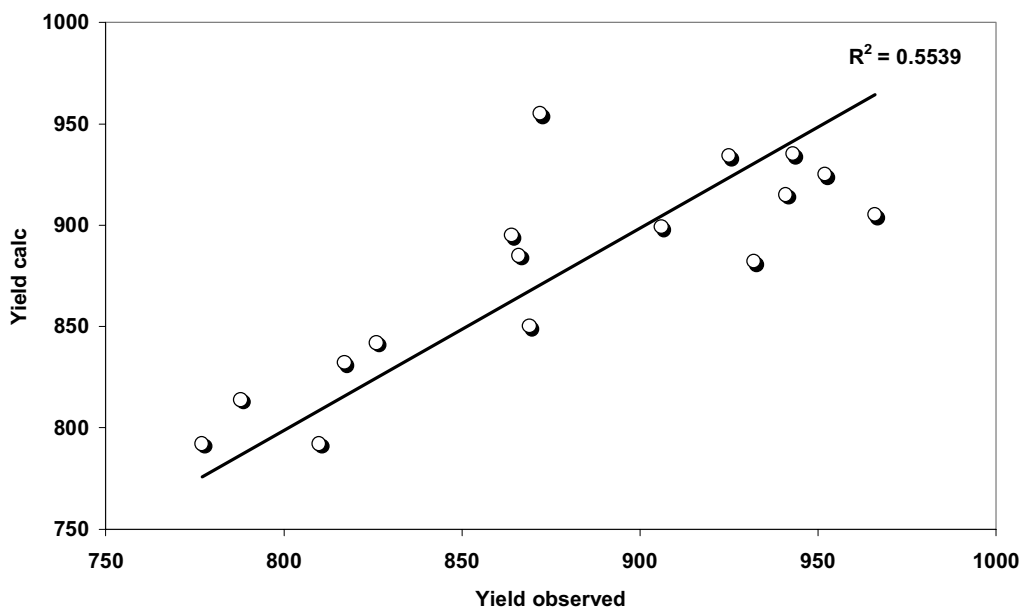
The calculations, when redone with only those two variables, yield the results given in table 4.

Table 4: Regression coefficients, standard errors and the t-stat for detrended yield as a function of August maximum temperature and ETA.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>
Intercept	-2078.073331	542.2848552	-3.832069643
Tx8	55.14678345	16.72811835	3.296651919
ETA	0.467996932	0.241381241	1.938828926

The final regression thus includes only two variables and accounts for 55.39 % of the variance, with an average²⁴ error of estimate of 3.0 %. The plot of the estimated values against the actual values is shown in figure 12. It shows an obvious lack of sensitivity of the yield functions at high yields (high yields are underestimated). This is clearly due to some errors in the water balance parameters as ETA saturates rather quickly. The suggested solution is to use actual ETA rather than normal values, as well as a better tuning of the water balance parameters.

Figure 12: Computed yield based on Tx8 and ETA versus actual yield.



²⁴ This is the average of the absolute values of the errors.

5.4 Principal component analysis (PCA) and factor analysis

PCA provides a convenient tool to reduce the number of variables used in a multiple regression, while at the same time providing a useful insight into the factors at play. IN this example, PCA was computed with FAOMET/FACAST

For instance, in the Rajshahi example, 15 components account for 100% of the variance in the input data, which is to say that there is a large amount of redundancy in the original 34 input variables (table 5).

Table 5: percent of the total variance of the input matrix absorbed by the 15 first principal components.

Component	%variance	%accum. Variance
1	18.9	18.9
2	17.2	36.1
3	15.3	51.4
4	11.1	62.5
5	8.4	70.9
6	7.7	78.5
7	6.3	84.8
8	4.0	88.9
9	2.8	91.6
10	2.2	93.8
11	2.1	95.9
12	1.7	97.6
13	1.7	99.3
14	0.4	99.7
15	0.3	100.0

Based on the correlation matrix in annex II, the first components are correlated with the original variables as shown in table 6.

Table 6: Interpretation of the first 5 principal components in terms of their correlation with the original variables.

Component	STRONG POSITIVE LINK WITH...	STRONG NEGATIVE LINK WITH...
C1	Tx8, Sunhours8	Tn8,Tn10,Tn11, HumRel8
C2	Sunhours11, Defwater	Index, ETA
C3	Rain7, HumRel12	Sunhours11, Sunhours12
C4	Sunhours7 , ExcessWater	Tx12
C5	Tx7 ,Sunhours9	(Sunhours12)

It appears immediately that Tx8 and ETA, the two variables which survived the previous analysis (5.3) are strongly correlated with the two first components.

In fact, only 6 variables account for 99% of the variance in the first component C1. They are listed in table 7.

Details are provided in spreadsheet **Explain PCA 1.xls**

Table 7: Regression coefficients, standard errors and the t-stat for the first principal component as a function of the 6 correlated variables with the highest correlation coefficients to it.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>
Intercept	101.7749147	6.986582474	14.56719578
AugHR	-0.581031224	0.11557561	-5.027282344
NovTn	-0.520396429	0.118143726	-4.404774129
OctTn	-0.726566911	0.114023337	-6.372089543
SepTn	-0.803008094	0.264855321	-3.031874503
DecTn	-0.360853231	0.090956736	-3.967306276
NovRain	-0.034098811	0.007328293	-4.653036133

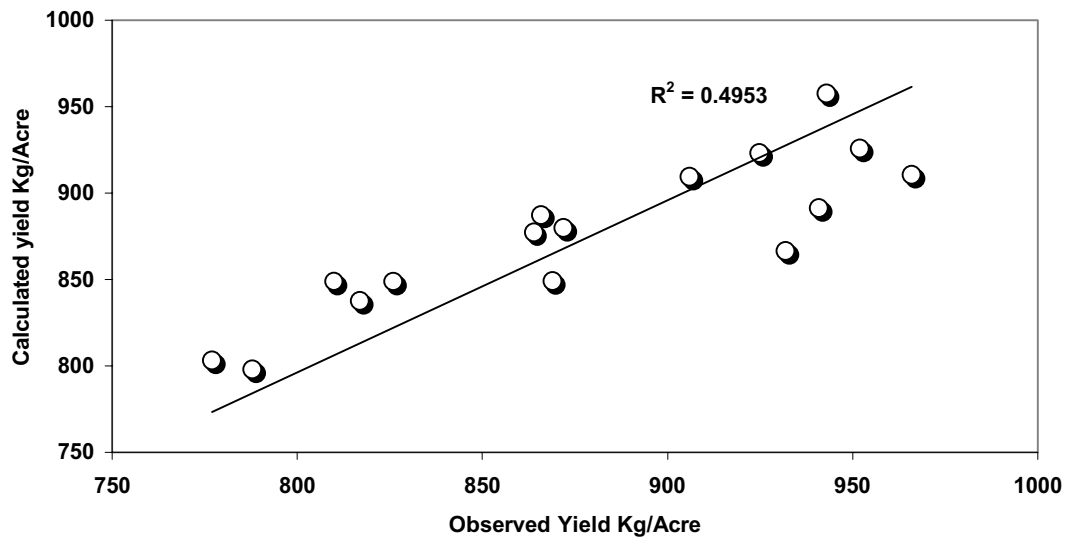
In this case, the final regression of yields as a function of the Components needs only the two first components (table 8)

Table 8: Regression coefficients, standard errors and the t-stat for yield as a function of the first 5 components.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>
Intercept	-0.414910999	16.05433607	-0.025844171
C-1	12.87817241	4.256602292	3.025458225
C-2	-10.00351436	4.466475645	-2.23968855
C-3	3.078950811	4.734296042	0.650350291
C-4	-4.669309805	5.564084196	-0.839187482
C-5	3.65805863	6.37503896	0.573809612

As shown in spreadsheet **PCA work.xls** and in figure 13, the final regression accounts for 78% of the variance with a slight improvement over the simple approach above. The average error affecting the yield drops to 2.8 %.

Figure 13: Computed yield based on the two first principal components versus actual yield.



Annex I: list of variables for Rajshahi yield function derivation

Nr	Name	Number missing	Minimum	Maximum
1	JulRain	0	105	725
2	AugRain	0	96	505
3	SepRain	0	89	511
4	OctRain	0	4	358
5	NovRain	0	0	54
6	DecRain	0	0	92
7	JulTx	0	31.3	33
8	AugTx	0	31.4	33.4
9	SepTx	0	31.3	33.5
10	OctTx	0	30.3	32.6
11	NovTx	0	28.1	30.2
12	DecTx	0	23.4	27
13	JulTn	0	25.2	27.5
14	AugTn	0	25.4	26.8
15	SepTn	0	24.8	26.2
16	OctTn	0	20.6	25
17	NovTn	0	15.2	20.3
18	DecTn	0	10.7	14.1
19	JulSh	0	2.6	5.3
20	AugSh	0	3.4	6.5
21	SepSh	0	2.6	6.6
22	OctSh	0	5.9	8.5
23	NovSh	0	6.8	9.5
24	DecSh	0	6.2	8.5
25	JulHR	0	85	91
26	AugHR	0	85	89
27	SepHR	0	82	90
28	OctHR	0	76	87
29	NovHR	0	72	84
30	DecHR	0	75	85
31	IndxLast	0	0	69
32	EXWT	0	132	1416
33	DEFWT	0	195	1295
34	ETA	0	473	629

Number of data lines: 16 / Code for missing data: -999

Annex II: Correlations between Components and original variables

	1	2	3	4	5	6	
	JulRai	AugRai	SepRai	OctRai	NovRai	DecRai	
1	-0.1821	-0.4067	-0.1177	-0.0792	-0.6313	0.2967	C-1
2	0.0507	0.2480	-0.5793	-0.6361	0.0197	-0.4565	C-2
3	0.6035	-0.0953	0.5752	-0.5746	0.4066	0.4336	C-3
4	0.2421	0.5603	-0.1072	0.3582	0.0122	0.2358	C-4
5	0.0187	-0.0752	-0.1339	-0.1429	0.2216	0.0721	C-5
6	0.5426	0.6128	-0.2359	0.0731	-0.3242	0.0720	C-6
7	0.2759	0.1695	-0.2584	0.2255	0.3232	-0.5027	C-7
8	0.1477	0.0142	0.1720	0.0108	-0.0557	-0.2354	C-8
9	-0.3161	0.0424	0.2483	0.0544	-0.0118	0.0861	C-9
10	-0.1711	-0.0669	-0.1063	-0.1103	-0.0202	-0.2201	C-10
11	-0.0476	-0.0891	-0.0468	0.1726	-0.0063	0.0317	C-11
12	-0.0099	0.0591	0.1397	-0.0609	-0.3844	0.1367	C-12
13	-0.1211	0.1122	0.1871	-0.0372	0.1471	-0.2233	C-13
14	0.0486	-0.0423	0.0843	-0.0184	0.0095	0.0224	C-14
15	-0.0062	0.0886	-0.0382	-0.0608	0.0601	0.1249	C-15
16	-0.0065	0.0002	-0.0017	0.0014	-0.0046	0.0015	C-16
17	-0.0005	0.0007	0.0006	0.0007	0.0039	-0.0007	C-17
18	0.0057	0.0011	-0.0003	-0.0011	-0.0006	-0.0045	C-18
19	0.0028	-0.0006	-0.0001	0.0009	0.0028	0.0019	C-19
20	0.0036	-0.0012	0.0044	-0.0001	-0.0028	-0.0007	C-20
21	0.0026	-0.0013	-0.0015	0.0012	0.0023	0.0014	C-21
22	0.0002	0.0013	-0.0002	-0.0018	-0.0014	0.0017	C-22
23	0.0009	-0.0018	-0.0011	0.0023	-0.0012	0.0001	C-23
24	0.0001	0.0005	0.0006	-0.0003	-0.0009	-0.0015	C-24
25	0.0000	0.0000	0.0004	0.0002	0.0000	-0.0000	C-25
26	0.0000	-0.0001	-0.0000	-0.0001	-0.0000	-0.0000	C-26
27	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-27
28	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-28
29	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-29
30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-30
31	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-31
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-32
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-33
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-34

	7 JulTx	8 AugTx	9 SepTx	10 OctTx	11 NovTx	12 DecTx	
1	-0.1132	0.6197	0.2871	-0.3978	0.0058	-0.1936	C-1
2	-0.5302	-0.2879	0.3964	0.5259	-0.1030	0.1927	C-2
3	-0.3867	0.2179	-0.3911	-0.0511	-0.2163	-0.5380	C-3
4	0.1681	-0.3463	-0.3488	-0.2637	-0.4110	-0.6223	C-4
5	0.6498	0.4035	0.5088	0.0588	0.3549	-0.0083	C-5
6	0.0629	-0.0055	0.1648	-0.2249	0.7529	-0.1949	C-6
7	-0.2250	0.2765	-0.2494	-0.4014	0.1397	0.2634	C-7
8	-0.1631	0.2028	-0.1739	0.2135	-0.0714	-0.3115	C-8
9	0.0801	0.0725	-0.0087	-0.3252	-0.1589	0.1237	C-9
10	0.0057	-0.1254	-0.0365	-0.0358	0.0628	-0.1540	C-10
11	0.0477	0.0837	-0.2970	0.3372	0.0185	0.0238	C-11
12	0.0541	0.1894	0.0544	0.0432	-0.1255	0.0332	C-12
13	0.1060	-0.1354	-0.0424	-0.0640	-0.0893	-0.0359	C-13
14	-0.0438	-0.0080	0.1106	-0.0709	-0.0357	-0.0456	C-14
15	0.0028	0.0543	-0.0274	0.0250	-0.0409	-0.0125	C-15
16	-0.0104	0.0005	-0.0032	-0.0085	0.0052	0.0016	C-16
17	0.0073	0.0048	0.0014	-0.0009	0.0093	-0.0003	C-17
18	0.0036	0.0063	0.0046	-0.0010	-0.0067	-0.0009	C-18
19	-0.0009	-0.0038	0.0041	0.0017	-0.0019	0.0042	C-19
20	0.0020	-0.0031	0.0010	-0.0003	0.0013	0.0021	C-20
21	0.0010	0.0019	-0.0018	-0.0021	-0.0019	0.0023	C-21
22	0.0010	-0.0007	-0.0014	0.0001	0.0007	0.0022	C-22
23	0.0003	-0.0002	-0.0002	-0.0006	-0.0005	-0.0008	C-23
24	0.0000	0.0002	-0.0003	-0.0000	0.0002	0.0002	C-24
25	0.0000	-0.0000	0.0000	-0.0000	-0.0000	-0.0001	C-25
26	-0.0000	-0.0000	-0.0000	-0.0000	0.0000	0.0000	C-26
27	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-27
28	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-28
29	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-29
30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-30
31	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-31
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-32
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-33
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-34
	7	8	9	10	11	12	

	13 JulTn	14 AugTn	15 SepTn	16 OctTn	17 NovTn	18 DecTn	
1	-0.4369	-0.7710	-0.6924	-0.7277	-0.7805	-0.6877	C-1
2	0.0147	-0.1388	-0.0242	-0.2522	0.0905	0.2024	C-2
3	-0.3194	0.0917	-0.2997	-0.2345	0.0401	0.2710	C-3
4	0.5689	0.1252	-0.0398	-0.0327	0.0816	-0.0198	C-4
5	0.3386	0.2740	-0.0049	-0.1730	0.2478	0.1066	C-5
6	-0.4182	-0.3539	0.3899	-0.0194	-0.0006	-0.0025	C-6
7	0.0019	0.0760	-0.2117	-0.3735	0.5398	0.1600	C-7
8	-0.1694	0.2385	-0.1497	0.1994	-0.0265	0.0994	C-8
9	-0.0864	-0.0566	0.0265	-0.1886	-0.0479	0.5151	C-9
10	-0.0683	0.1054	0.3636	0.0750	-0.0522	-0.0795	C-10
11	-0.1466	-0.0383	-0.1251	0.2752	-0.0066	0.0514	C-11
12	0.1592	-0.0789	0.2254	0.1210	0.0830	0.0740	C-12
13	0.0108	-0.2561	-0.0680	0.0084	-0.0330	-0.2807	C-13
14	0.0364	0.0573	-0.0270	0.0955	0.0844	-0.0717	C-14
15	-0.0675	0.0910	0.0671	-0.0248	0.0259	-0.0413	C-15
16	0.0091	0.0136	-0.0030	-0.0013	-0.0079	-0.0027	C-16
17	0.0045	0.0014	-0.0049	0.0072	-0.0058	0.0007	C-17
18	-0.0005	0.0030	-0.0029	0.0010	-0.0018	-0.0005	C-18
19	0.0037	-0.0023	0.0001	0.0012	0.0000	0.0028	C-19
20	-0.0024	0.0032	0.0000	-0.0032	0.0005	-0.0021	C-20
21	-0.0016	-0.0021	0.0013	0.0002	-0.0040	0.0005	C-21
22	-0.0003	0.0005	-0.0022	0.0002	0.0003	-0.0002	C-22
23	0.0000	0.0001	0.0004	-0.0000	0.0006	0.0001	C-23
24	-0.0000	0.0001	0.0002	0.0003	-0.0002	0.0011	C-24
25	0.0001	-0.0000	0.0001	-0.0001	0.0000	-0.0004	C-25
26	-0.0000	0.0000	-0.0000	-0.0000	0.0000	0.0000	C-26
27	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-27
28	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-28
29	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-29
30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-30
31	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-31
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-32
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-33
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-34
	13	14	15	16	17	18	

	19 JulSh	20 AugSh	21 SepSh	22 OctSh	23 NovSh	24 DecSh	
1	0.1851	0.6913	0.0524	-0.1237	0.1592	-0.4357	C-1
2	-0.2648	-0.3009	0.2282	0.7375	-0.2993	0.0707	C-2
3	-0.0100	-0.0486	-0.4050	0.0981	-0.6088	-0.7070	C-3
4	0.7157	0.2029	0.0536	0.0707	-0.1189	-0.2270	C-4
5	0.5514	-0.1113	0.5847	-0.2441	-0.1610	-0.3528	C-5
6	-0.0867	0.0634	0.1096	-0.0832	0.2991	0.0812	C-6
7	-0.1567	0.4104	0.0508	-0.1361	-0.1457	0.1812	C-7
8	0.1167	0.1402	0.4934	0.4573	0.4975	-0.0798	C-8
9	-0.0510	-0.1363	0.0283	0.2164	0.1590	0.0980	C-9
10	0.0294	0.3222	-0.1519	0.1852	0.0347	-0.0339	C-10
11	-0.1069	0.1028	0.0777	-0.2172	-0.1587	0.1820	C-11
12	-0.0738	0.2083	0.1433	0.0306	-0.2192	0.0681	C-12
13	-0.0258	-0.0806	0.3470	-0.0238	-0.1044	0.0396	C-13
14	0.0852	0.0091	-0.0818	-0.0398	0.0740	0.1824	C-14
15	-0.0720	-0.0231	0.0287	-0.0240	0.0035	-0.0150	C-15
16	-0.0048	0.0033	0.0070	-0.0034	-0.0022	0.0007	C-16
17	-0.0044	-0.0012	-0.0052	0.0074	-0.0016	-0.0011	C-17
18	-0.0046	0.0016	-0.0004	-0.0008	-0.0002	0.0011	C-18
19	-0.0032	0.0051	0.0019	-0.0001	0.0029	-0.0029	C-19
20	-0.0018	-0.0012	0.0008	0.0028	-0.0033	0.0025	C-20
21	0.0028	0.0017	0.0001	0.0016	-0.0009	0.0030	C-21
22	0.0004	-0.0006	0.0009	-0.0000	0.0022	0.0012	C-22
23	-0.0010	-0.0018	0.0006	0.0002	0.0003	-0.0008	C-23
24	0.0006	-0.0003	-0.0004	-0.0010	-0.0004	-0.0006	C-24
25	-0.0004	0.0001	-0.0002	0.0000	0.0000	-0.0003	C-25
26	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	C-26
27	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-27
28	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-28
29	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-29
30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-30
31	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-31
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-32
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-33
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-34
	19	20	21	22	23	24	

	25 JulHR	26 AugHR	27 SepHR	28 OctHR	29 NovHR	30 DecHR	
1	-0.2514	-0.8196	-0.2785	-0.0407	-0.4722	-0.0178	C-1
2	0.3667	0.1139	-0.6216	-0.6421	-0.2548	0.0921	C-2
3	0.4913	-0.0753	0.5724	0.1802	0.4853	0.8840	C-3
4	-0.5632	-0.0283	-0.2454	-0.5425	-0.3311	0.1022	C-4
5	-0.1291	-0.0866	-0.2262	0.1905	0.3533	0.2713	C-5
6	0.1623	0.1734	0.1793	0.3025	-0.1203	0.2559	C-6
7	0.1452	-0.4381	-0.1463	-0.1686	0.1818	-0.0628	C-7
8	0.1092	-0.0273	-0.1318	0.0576	-0.1025	0.0550	C-8
9	-0.1130	-0.1622	-0.0476	0.1715	0.0362	0.0123	C-9
10	-0.2466	-0.1411	0.0685	0.0620	0.3870	0.0284	C-10
11	-0.1786	-0.1464	-0.0788	0.1053	-0.0339	0.2027	C-11
12	0.2177	-0.0733	-0.0298	-0.1398	0.0974	-0.0144	C-12
13	0.0946	0.0086	0.0457	0.1544	0.1491	-0.0378	C-13
14	0.0730	-0.0675	-0.0498	0.0761	-0.0132	0.0179	C-14
15	-0.0268	-0.0286	-0.0865	0.0791	-0.0086	-0.0962	C-15
16	0.0016	0.0081	0.0018	0.0043	-0.0004	0.0044	C-16
17	0.0022	-0.0029	-0.0030	-0.0015	-0.0002	-0.0007	C-17
18	-0.0051	0.0039	0.0065	-0.0003	0.0002	-0.0043	C-18
19	-0.0019	-0.0005	0.0005	0.0032	-0.0009	0.0027	C-19
20	-0.0047	-0.0001	-0.0029	0.0000	-0.0010	0.0035	C-20
21	0.0008	0.0053	-0.0033	-0.0001	0.0023	0.0008	C-21
22	-0.0008	-0.0007	0.0024	-0.0025	0.0022	0.0001	C-22
23	0.0004	-0.0003	0.0001	0.0001	0.0010	-0.0002	C-23
24	-0.0003	0.0002	-0.0003	0.0007	-0.0007	-0.0003	C-24
25	0.0001	0.0000	-0.0003	-0.0002	0.0000	-0.0001	C-25
26	-0.0000	0.0000	-0.0000	0.0000	-0.0000	-0.0000	C-26
27	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-27
28	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-28
29	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-29
30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-30
31	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-31
32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-32
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-33
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	C-34
	25	26	27	28	29	30	

	31 IndxLa	32 EXWT	33 DEFWT	34 ETA	
1	-0.4108	-0.4296	0.1962	0.0187	C-1
2	-0.7005	-0.3294	0.8574	-0.8030	C-2
3	-0.2837	0.0480	0.1941	-0.0345	C-3
4	-0.3467	0.6052	0.2115	-0.0238	C-4
5	0.2651	-0.3348	0.1458	-0.3483	C-5
6	-0.0152	0.3811	0.1476	-0.2594	C-6
7	0.0276	0.0671	-0.1167	0.1605	C-7
8	0.1768	-0.0827	-0.1291	0.1160	C-8
9	0.0012	0.1978	0.1271	-0.2407	C-9
10	-0.0924	-0.0776	0.0941	-0.0293	C-10
11	-0.0307	0.0818	0.1627	-0.2321	C-11
12	-0.0274	0.0529	-0.0828	0.0612	C-12
13	-0.1439	0.1206	0.0501	0.0614	C-13
14	-0.0533	-0.0217	0.0959	-0.0193	C-14
15	-0.0592	-0.0019	0.0213	0.0685	C-15
16	-0.0026	0.0002	0.0025	-0.0026	C-16
17	-0.0053	0.0039	0.0001	0.0042	C-17
18	-0.0011	0.0012	0.0010	-0.0043	C-18
19	-0.0019	-0.0001	0.0009	0.0030	C-19
20	-0.0004	-0.0004	-0.0024	0.0015	C-20
21	0.0007	-0.0015	0.0017	0.0017	C-21
22	0.0000	-0.0008	0.0024	0.0017	C-22
23	-0.0011	-0.0004	0.0010	0.0006	C-23
24	-0.0003	-0.0007	0.0012	0.0013	C-24
25	0.0009	0.0002	0.0007	-0.0000	C-25
26	0.0000	0.0001	0.0000	0.0000	C-26
27	0.0000	0.0000	0.0000	0.0000	C-27
28	0.0000	0.0000	0.0000	0.0000	C-28
29	0.0000	0.0000	0.0000	0.0000	C-29
30	0.0000	0.0000	0.0000	0.0000	C-30
31	0.0000	0.0000	0.0000	0.0000	C-31
32	0.0000	0.0000	0.0000	0.0000	C-32
33	0.0000	0.0000	0.0000	0.0000	C-33
34	0.0000	0.0000	0.0000	0.0000	C-34
	31	32	33	34	