NON-PARAMETRIC CROP YIELD FORECASTING, A DIDACTIC CASE STUDY FOR ZIMBABWE

R. Gommes

Environment, Climate Change and Bioenergy Division
FAO, Rome, Italy
rene.gommes@fao.org

Commission VIII, WG VIII/10

KEY WORDS: Agriculture, Climate, Crop, Developing Countries, Impact Analysis, Prediction, Decision Support, Accuracy

ABSTRACT:

Operational crop yield forecasting is mostly achieved with empirical statistical regression equations relating regional yield with predictor variables, termed “factors”. Regional yield (the “dependent variable”) refers to average yield over districts, provinces or, more rarely, whole countries; they are provided by national statistical services. The factors can be any combination of raw environmental variables such as weather variables or indices, satellite indices such as Normalised Difference Vegetation Indices (NDVI), farm inputs (fertiliser use) or outputs from simulation models, for instance water transpired over a given phenological phase, maximum leaf area index (LAI), average soil moisture, etc. The approach above is termed “parametric” for two reasons: (1) it derives or requires a number of parameters, for instance regression coefficients and the parameters characterise crop simulation models and (2) it attempts to identify the factors that condition yields and to understand their action. The difference between “parametric” and “non-parametric” methods is not clear-cut; it is mostly operational. Parametric forecasting approaches derive a “model” (through a process known as “calibration”) based on historical yield and climatic data. The model is subsequently applied to current crops and within season data to issue a forecast of yields. A number of calculations are performed; they are basically the same in the calibration and in the forecasting phases. Non-parametric crop yield forecasting techniques attempt to establish a typology (qualitative description) of the environmental conditions that occur during the growing season, assuming that similar types of seasons lead to similar yields. Similar years are grouped in classes. During the calibration phase, the types of seasons are defined in such a way as to minimize the variability of yields within classes and maximise between-classes variance. The forecast proper is done by categorizing the current year into one of the classes, and by assigning the class yield to the current forecast. Depending on the actual method, the forecast itself may require little more than comparing some variables with reference values, e.g. a threshold. This paper offers a rough comparison of simple yet classical parametric approaches with two different non-parametric methods, applied to national maize yields in Zimbabwe. The conclusion suggests that the simple non-parametric approaches are not inferior, in terms of accuracy and ease of use to the more complex parametric models

.

1. INTRODUCTION

Operational crop yield forecasting is mostly done with crop simulation models and empirical statistical regression equations relating yield with predictor variables, usually termed “factors”. For the purpose of this paper, crop forecasting and crop yield forecasting refer to operational within-season regional yield forecasts, i.e. forecasting of average crop yield (tons of agricultural product per ha) over large areas. The areas are administrative units, as this is the scale at which most socio-economic data and crop statistics are available to decision makers.

It is stressed that crop forecasts are eventually calibrated against crop statistics, so that, strictly speaking, crop forecasts are actually forecasts of agricultural statistics; they incorporate all the errors and biases that affect statistics.

Crop forecasts are typically issued between the time of planting and the time of harvest. They use past data (data between planting or before, and the time of the forecast) and “future” data. Future data can be implicit or explicit. In the first case, the future is assumed to be “normal” whereas the second requires that numerical values be actually specified, for instance historical data or stochastic weather generator outputs (Lawless and Semenov, 2005; Hansen et al., 2006).

There is a variety of generic forecasting methods, of which most can somehow be applied to crop forecasting as well (Petr, 1991). According to Armstrong (2001b), “judgement pervades all aspects of forecasting”, which is close to a definition which the author has frequently applied to crop yield forecasting, which can be seen as “the art of identifying the factors that determine the spatial and inter-annual variability of crop yields” (Gommes, 2003). In fact, given the same set of input data, different experts frequently come up with rather different forecasts of which, however, some are demonstrably better than others, hence the use of the word “art”.

There appears to be no standard classification of forecasting methods (Makridadis et al., 1998; Armstrong, 2001a). Forecasting methods can be subdivided into various categories according to the relative share of judgement, statistics, models and data used in the process. Armstrong identifies 11 types of methods that can be roughly grouped as ***judgemental***, based on stakeholders’ intentions or on the forecaster’s or other experts’ opinions or intentions, and ***statistical***, including univariate (or extrapolation), multivariate (statistical “models”) and theory-based methods. ***Intermediate*** types include expert systems, basically a variant of extrapolation with some admixture of expert opinion, and analogies, which Armstrong places between expert opinions and extrapolation models.

In this paper, we consider "parametric models" to be those that attempt to interpret and to quantify the causality links that exist between crop yields and environmental factors – mainly weather-, farm management and technology. They include essentially crop simulation models and statistical "models" which relate crop yield with assumed impacting factors. Obviously, crop-yield-weather simulation belongs to Armstrong's Theory-based Models. Non-parametric forecasting methods are those that rely more on the qualitative description of environmental conditions and do not involve any simulation as such (Armstrong's expert systems and analogies).

There are few explicit applications of non-parametric forecasting methods to agricultural yields, among others because of differing usage of the term "non-parametric (*e.g.* in Orlandini et al., 2004). Some methods are described by Gommes et al., 2007.

The paper provides a rapid comparison of four forecasting methods in the specific case of maize in Zimbabwe (Southern Africa). After a short introduction to climate and cropping in the country (2.1), two parametric methods are illustrated first (regression, in section 2.2, and model and regression based, section 2.3). The two non-parametric methods that are described next include a threshold-oriented approach (2.4) and an approach using the statistical clustering of annual rainfall profiles (2.5). The overall performance of the four methods is summarised in table 2.

2. A SIMPLE CASE STUDY FOR MAIZE IN ZIMBABWE

2.1 General setting and removal of yield trends

To illustrate and compare some non-parametric methods, a didactic example was prepared to estimate yields in Zimbabwe (Southern Africa) covering 41 years from 1960-61 to 2001-2002 (21 years from 1982-82 to 2001-02 for the simulation approach in 2.3).

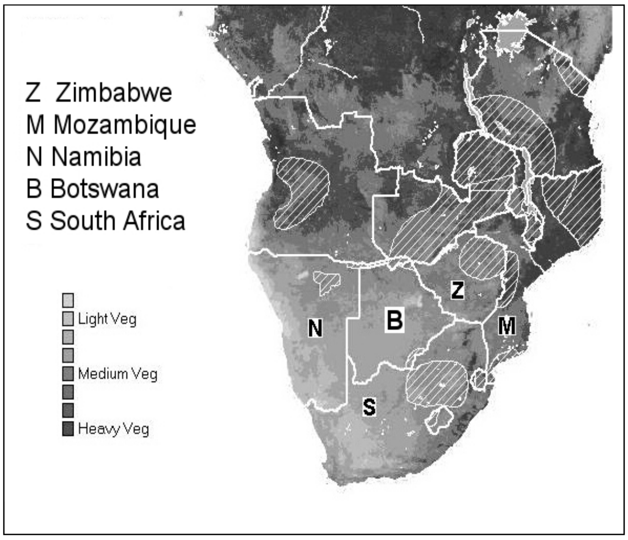

Figure 1: Map of Southern Africa. The hatched area corresponds to the main maize growing areas. The background map shows vegetation densities as estimated from satellite indices (light, medium and heavy vegetation)

Rainfall over the main maize growing area was extracted from NOAA monthly rainfall grids[*] using the WINDISP[**] software after the grids were converted to WINDISP format. All climate and crop statistics given hereafter refer to the maize growing area of NE Zimbabwe illustrated in Figure 1.

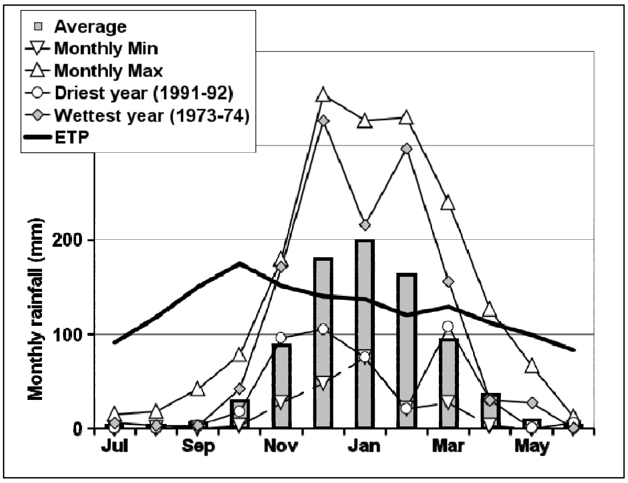

Figure 2: : Rainfall and ETP (evapotranspiration potential) patterns in Zimbabwe between 1960-61 and 2001-2002: average monthly values, maximum and minimum recorded for each month, as well as rainfall profiles of driest and wettest years.

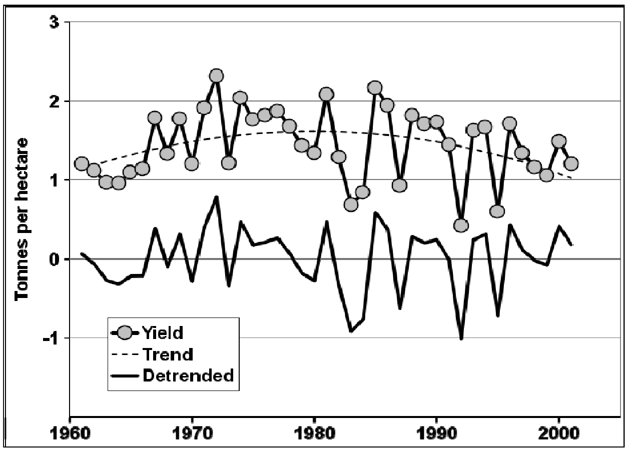

Figure 3: Maize yields in Zimbabwe, together with their trend and the and detrended value (departure of actual values from the trend).

Average rainfall amounts to 812 mm per year, but the driest year (1991-92, an El Niño year) recorded only 462 mm, while the wettest experienced 1278 mm in 1973-74. Note, incidentally, that 1973-74 corresponds to a severe drought in the West African Sahel; this "correlation" derives directly from the movements of the Inter-Tropical Convergence Zone (ITCZ). In Zimbabwe, the growing season roughly covers the period from November to March-April (Figure 2), and is also

[*] The data are available upon request from the website ftp://ftpprd.ncep.noaa.gov/pub/precip/50yr/gauge/0.5deg/

[**] WINDISP is a software developed by FAO and other agencies (USGS, USFS, FEWS, SADC) to process satellite imagery in food security projects. The latest version can be used for a number of gridded data, including rainfall. The software is available at ftp://ext-ftp.fao.org/sd/reserved/agromet/windisp

dependent on the different behaviours of the two main sectors of the Zimbabwean agriculture, i.e. large-scale commercial farms on the most suitable soils and subsistence farmers in so-called "communal lands". Part of the country being semi-arid with a marked dry season, water is the dominant factor driving the inter-annual variability of crop yield.

Since independence in 1980 but particularly after 1990, the country has been affected by a somewhat disorderly land reform aiming at redistributing part of the land under large-scale farms. The combination of land reform and changes that were made to the national agricultural statistical system, in particular the inclusion of "communal lands" in the statistics, results in figure 3. A curvilinear trend had to be fitted to the data. Clearly, the trend itself is due to a combination of factors where weather plays only a minor part. The trend must be removed before any agrometeorological analysis can be carried out. This was done for the lowermost curve in Figure 3, where yield is expressed as the difference between the observed values and the trend. The trend accounts for 13 % of the interannual yield variability, which is in line with the fact that parts of the country are semi-arid.

2.2 First parametric approach: yield-rainfall relations between 1961-62 and 2001-02

The simplest possible parametric method to estimate crop yields is to regress them against rainfall, particularly in areas where water is the dominant limiting factor to agricultural production (Palm, 1997). Figure 4 shows the roughly linear relation between yield and rainfall, with a coefficient of determination amounting to 0.4563, i.e. about 46% of the variability of detrended yields can be assigned to rainfall (Table 2).

2.3 Second parametric approach: simple simulation of maize yields in Zimbabwe (1981-82 to 2001-02)

The second parametric approach that is being illustrated uses the standard FAO methodology (Gommes et al., 1998; Gommes, 2003) and the AMS* software. A crop specific soil water balance was computed for the years 1981-82 to 2001-02 using 10-daily data from 25 meteorological stations in Zimbabwe and 245 in the surrounding countries. Actual maize crop evapotranspiration (ETA, mm) was computed for all the stations, gridded over the region and averaged for the maize growing areas.

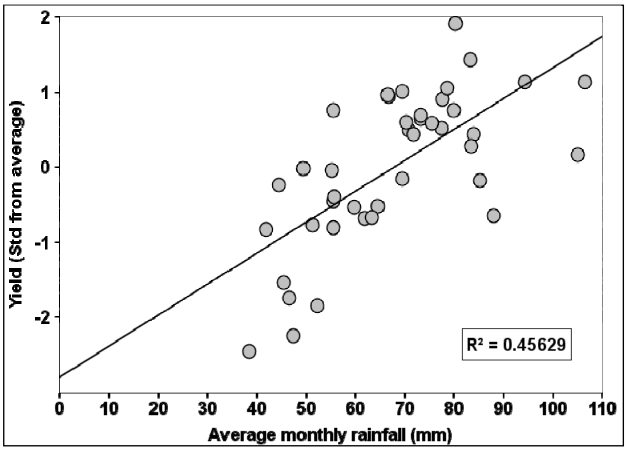

Figure 4: Relation between detrended maize yield (expressed as standard deviations from average) and average July-June rainfall in maize growing areas of Zimbabwe between the cropping seasons of 1961-62 and 2001-02.

Water balance parameters, in particular ETA, are ideal "value-added" variables to be used in crop forecasting (Gommes, 1998), and they are at the heart of the FAO crop forecasting approach. Water balance parameters include actual crop evapotranspiration, water surplus and water deficit over main crop stages (e.g. emergence, vegetative phase, flowering).

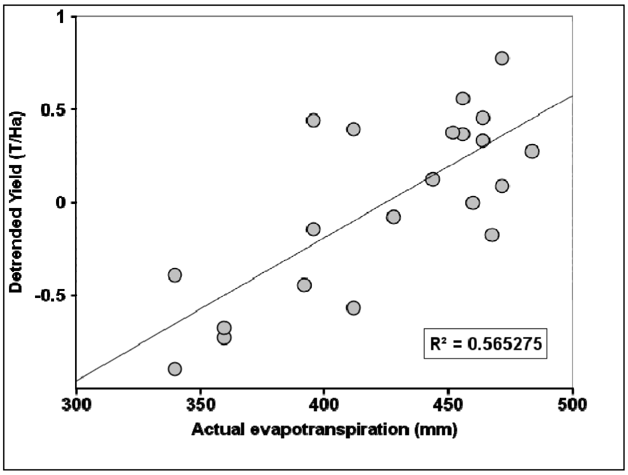

Figure 5: Relation between detrended yield and actual maize evapotranspiration in maize growing areas of Zimbabwe between 1982 and 2002.

The basic idea behind the methodology adopted by FAO is that, as de Wit was among the first to recognise in the mid fifties, there is a direct link between plant transpiration and productivity (van Keulen and van Laar, 1986). For "not too severe" water stresses, yields are rather linearly correlated with actual evapotranspiration. Interestingly, this relation holds across various spatial scales, from leave to plant to field to administrative region. The relation between ETA and maize yield is shown in Figure 5 Altogether, ETA and trend account for about 73.75 % of the interannual variability of maize yields (Table 2).

2.4 First non-parametric approach: threshold based yield forecasting (1961-62 to 2001-02)

The first non-parametric method is a simple threshold-based crop-forecasting table.

* AMS, the AgroMetShell is the standard FAO software used for crop forecasting at the national level. It can be downloaded from ftp://ext-ftp.fao.org/sd/reserved/agromet/agrometshell

In Zimbabwe, like in most of southern-central Africa, there is a tendency for rainfall distribution to be bimodal, with a dry period in January or February, as can be seen in Figure 2. In fact, when correlating yields with monthly rainfall, the coefficient turns out to be highest in January (R=0.656) and February (0.500). The next highest value corresponds to March (R=0.367)

Criteria 1 January rainfall (mm)	Yield (average and 95% confidence interval)	Criteria 2	Threshold	
75 to 155	-1.07 -1.64 to –0.50	February rainfall	< 120 mm	>120 mm
			-1.74	-0.52
			-2.35 to -1.13	-1.16 to 0.12
156 to 249	0.25 -0.05 to 0.55	February rainfall	< 170 mm	> 170 mm
			0.07	0.57
			-0.50 to 0.35	0.25 to 0.89
250 to 327	0.78 0.35 to 1.08	December rainfall	< 190 mm	> 190 mm
			0.92	0.66
			0.23 to 1.63	0.08 to 1.25

Table 1: Example of a threshold-based crop forecasting table for maize in Zimbabwe, based on yields recorded during the period 1961-62 to 2000-2001. Yields are expressed in standard deviations about the average for the period.

It was found that a good separation of yield categories could be achieved when grouping years by January and February rainfall totals as shown in Table 1.

For instance, yields fall in the range of –0. 05 to 0.55 standard deviations from average when January rainfall is between 156 and 249 mm: they are about average (0.25 standard deviations higher than the average). When we now separately examine the group of years characterized by January rainfall from 75 to 155 mm (Group 1, Table 1), 156 to 249 mm (Group 2) and 250 to 327 (Group 3) rather contrasting correlations are found between yields and monthly rainfall.

In Groups 1 and 2, the highest correlation is between yield and February rainfall while in Group 3, we find a negative highest correlation between yield and December rainfall. This results from the fact that high January rain will not have a detrimental effect on yield only if December is relatively dry.

The described method will of course forecast only the six yield values that appear in Table 1, together with their confidence interval. Yet, the strength of the correlation remains comparable with the one obtained with the less empirical simulation approach (Figure 6).

The threshold-based approach also illustrates the fact that this non-parametric method can somehow be seen as a discrete variant of 2.2: two regression equations between yield and rainfall could be developed, one using January and February rainfall during relatively dry years, and another based on January and December precipitation during wetter years.

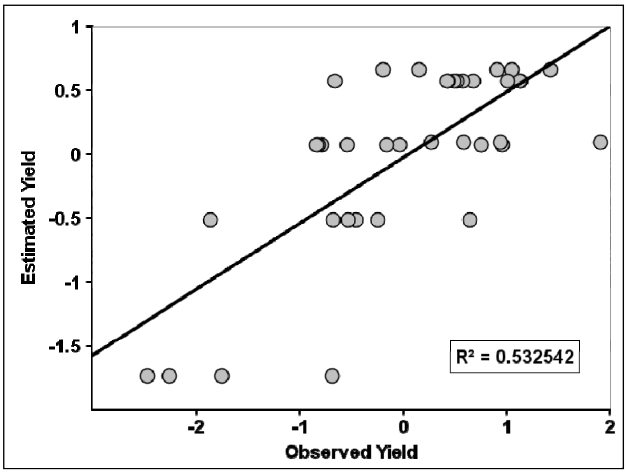

Figure 6: Comparison of estimated and observed yields in Zimbabwe between 1961-62 and 2001-02 using the threshold method described in Table 1. Yields are expressed in standard deviations from the average.

2.5 Second non-parametric method: rainfall profile clustering method (1961-62 to 2001-02)

For this second non-parametric method, the basic assumption is that similar rainfall profiles (July to June) will on average result in similar yield categories.

The profiles were obtained using the ADDATI* multivariate statistical package developed by Griguolo at the university of Venice. The number of classes to adopt is somewhat arbitrary. In this case, 12 were found to be a good compromise. Some typical rainfall profiles are shown in Figure 7.

They can all be described in terms of rainfall distributions and amounts. For instance class 1 stands for “low but well distributed rainfall”, class 3 for “abundant and well distributed rainfall with a mid-season dry spell”, etc.

Regarding the potential value of the method as a crop-forecasting tool, the coefficient of determination R^2 of 0.5692 is amazingly close to the one obtained with the crop specific soil water balance (0.5653, Figure 5).

With the classification method, the number of different yields is obviously the same as the number of classes. To use the approach for crop forecasting in operational mode, a given season is compared with the 12 classes and assigned to one of them. The yield for the year is then taken as the average yield (with confidence interval) of the class.

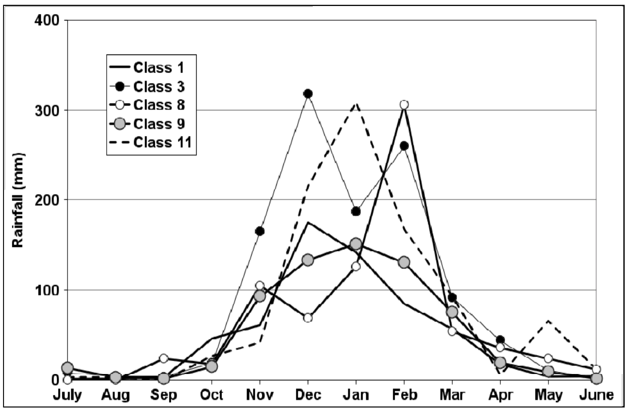

Figure 7: Some typical rainfall profiles for Zimbabwe. Each

* The latest update can be downloaded from the website given hereafter: http://cidoc.iuav.it/~silvio/addati_en.html

profile is the average of a number of different years that have been assigned to that class by the clustering programme. Rainfall is expressed in mm.

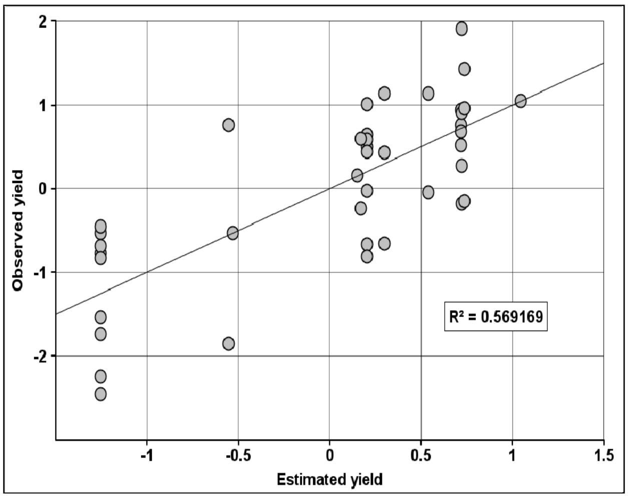

Figure 8: Comparison of estimated and observed yields in Zimbabwe between 1961-62 and 2001-02 using the rainfall profile method. Yields are expressed in standard deviations from the average.

A technique is needed to assign the current season to one of the classes. This can be done with various distance functions between the profiles, and the reference curves. It may, naturally happen that the assignment of a given year to a reference class changes as the season develops. If necessary, this can be avoided by expressing yield as a weighted average of several classes, where the weight is given by the above-mentioned distance function.

Method	R^2		
	Trend	Method	Total
Average Rainfall	0.1702 +	0.4563	0.6265
Water Balance		0.5653	0.7355
Threshold		0.5311	0.7013
Clustering		0.5692	0.7394

T*able 2: comparison of several maize yield forecasting approaches in Zimbabwe (fraction of variance accounted for by the different methods).*

3. CONCLUSIONS

The inter-comparison of the four illustrated crop-forecasting approaches is given in table 2. Remember that the number of years and the source of the weather data is different for the water balance approach compared with the other methods: the length of the time series is shorter (21 years instead of 41), but the quality of the data is better given that actual station information was used rather than grids.

It appears that, with the exception of the simple use of seasonal rainfall totals, the methods tested yield similar results.

The reason why, in spite of their conceptual simplicity, non-parametric approaches yield good results lies in the fact that weather variables are correlated, and the fact that they do not vary randomly over space and time. Therefore, one variable, especially a seasonal time-profile, can indirectly capture a number of environmental factors. It is certainly worth remembering the classical work of Cane et al (1994) on the relations between El Niño-Southern Oscillation (ENSO) indices and maize yields in southern Africa: better correlations are obtained between ENSO indices and yields than between rainfall and yield. ENSO is a more comprehensive variable that describes, albeit indirectly, the overall behaviour of weather conditions during the growing season better than a single variable.

Timeliness, cost and spatial scale are some of the criteria that are adopted when selecting crop-forecasting methods. In view of the ease of implementation of the non-parametric methods, it is certainly worth exploring their potential further.

It appears further that non-parametric methods are as accurate as the deterministic ones, and that they are comparable in terms of timeliness. Non-parametric approaches, however, are much less demanding in terms of inputs and "technology" (processing power), so that some of them can even be applied at village level (the "threshold-based" approach).

4. REFERENCES

Armstrong, J.S. (Ed), 2001a. *Principles of forecasting, a handbook for researchers and practitioners*. Kluwer Acad. Press, Boston, Doordrecht, London, 849 pp.

Armstrong, J.S. , 2001b. *Introduction*, pp. 1-12 *in* Armstrong, J.S. (Ed), 2001a.

Cane, M.A., G. Eshel and R.W. Buckland, 1994. Forecasting Zimbabwean maize yields using eastern equatorial Pacific sea-surface temperature. *Nature*, 370: 204-205.

Gommes, R. 1998. *Roving Seminar on crop-yield weather modelling; lecture notes and exercises*. WMO. Geneva, 153 pp.

Gommes, R., 2003. *The FAO crop monitoring and forecasting approach*. Proc. of JRC-FAO Workshop on Crop Monitoring in the Greater Horn off Africa. Nairobi, 28-30 Jan 2003, pp 45-48 *in* Rijks, D., F. Rembold, T. Nègre, R. Gommes and M. Cherlet, Editors. 2003. *Crops and Rangeland Monitoring in Eastern Africa for early warning and food security*. Proc. Of JRC/FAO Internat. Workshop, Nairobi 28-30 Jan. 2003. FAO/JRC, EUR 20869EN, European Commission, Official Publications of the EU, EUR 20869EN, Luxembourg, 184 pp + 1 CD-ROM.

Gommes, R., F.L. Snijders and J.Q. Rijks, 1998. *The FAO crop forecasting philosophy in national food security warning systems*, pp. 123-130, *in* Rijks, D., J.M. Terres and P. Vossen (eds), 1998. *Agrometeorological applications for regional crop monitoring and production assessment*, Official Publications of the EU, EUR 17735, Luxembourg, 516 pp.

Gommes, R., H.P. Das, L. Mariani, A. Challinor, B. Tychon, R. Balaghi and M.A.A. Dawod, 2007. *WMO Guide to Agrometeorological Practices*, Chapter 5, Agrometeorological Forecasting. 70 pp. Current version downloadable from http://www.agrometeorology.org/fileadmin/insam/repository/gamp_chapt5.pdf.

Hansen, J.W., A. Challinor, A. Ines, T. Wheeler and V. Moron, 2006. Translating climate forecasts into agricultural terms: advances and challenges. *Clim. Res.*, 33:27-41.

Lawless, C. and M.A. Semenov, 2005. Assessing lead-time for predicting wheat growth using a crop simulation model. *Agric. for. meteorol.*, 135(1-4):302-313.

Makridadis, S., S.C. Wheelwright and R.J. Hyndman, 1998. *Forecasting, Methods and Applications*. John Wiley & sons, Inc. New York, 642 pp.

Orlandini, F., D. Lanari, L. Pieroni, B. Romano and M. Fornaciari, 2004. Instrumental test to estimate a model of forecast yield: a non-parametric application to the pollen index in south Italy. *Ann. Appl. Biol.*, 145:81-90.

Palm, R., 1997. *Les modèles de prévision statistique: cas du modèle Eurostat-Agromet*. Pp. 85-108 *in* Tychon, B., and V. Tonnard, 1997. *Estimation de la production agricole à une échelle régionale*. Official Publications of the EU, EUR 17663, Luxembourg. 202 pp.

Petr, J., 1991. *Weather and yield*. Developments in crop science N. 20. Elsevier, Netherlands, and Agricultural Publishing House. 288 pp.

van Keulen, H., and H.H. van Laar, 1986. *The relation between water use and crop production*, *in* van Keulen and Wolf (1986), (eds), 1986. *Modelling of agricultural production: weather, soils and crops*. Simulation monographs, Pudoc, Wageningen, 478 pp.