ORIGINAL PAPER

Extraction of information content from stochastic disaggregation and bias corrected downscaled precipitation variables for crop simulation

Ashok K. Mishra · Amor V. M. Ines ·
Vijay P. Singh · James W. Hansen

Abstract We applied a simple statistical downscaling procedure for transforming daily global climate model (GCM) rainfall to the scale of an agricultural experimental station in Katumani, Kenya. The transformation made was two-fold. First, we corrected the rainfall frequency bias of the climate model by truncating its daily rainfall cumulative distribution into the station's distribution based on a prescribed observed wet-day threshold. Then, we corrected the climate model rainfall intensity bias by mapping its truncated rainfall distribution into the station's truncated distribution. Further improvements were made to the bias corrected GCM rainfall by linking it with a stochastic disaggregation scheme to correct the time structure problem inherent with daily GCM rainfall. Results of the simple and hybridized GCM downscaled precipitation variables (total, probability of occurrence, intensity and dry spell length) were linked with a crop model for a more objective evaluation of their performance using a non-linear measure based on mutual information based on entropy. This study is useful for the identification of both suitable downscaling technique as well as the effective precipitation variables for forecasting crop yields using GCM's outputs which can be useful for addressing food security problems beforehand in critical basins around the world.

Keywords Downscaling · GCM · Mutual information · Crop model · Food security

A. K. Mishra (✉)
Pacific Northwest National Laboratory, P.O. Box 999, Richland,
WA 99352, USA
e-mail: akm.pce@gmail.com

A. V. M. Ines · J. W. Hansen
International Research Institute for Climate and Society,
The Earth Institute, Columbia University, Palisades,
NY 10964, USA
e-mail: ines@iri.columbia.edu

J. W. Hansen
e-mail: jhansen@iri.columbia.edu

V. P. Singh
Department of Biological and Agricultural Engineering,
Texas A&M University, 2117 TAMU, College Station,
TX 77843, USA
e-mail: vsingh@tamu.edu

V. P. Singh
Department of Civil and Environmental Engineering,
Texas A&M University, 2117 TAMU, College Station,
TX 77843, USA

1 Introduction

Vital to the role of crop yield models as early warning and assessment tools for food security applications is the prediction of seasonal climate (Hansen et al. 2010; Robertson et al. 2007). Many studies have been conducted to explore this linkage, and several approaches have been proposed to reconcile the mismatch between the scale of climate data and the requirements of crop models (Hansen and Indeje 2004). Crop models require daily weather inputs, while climate forecasts come in seasonal formats not compatible with what the crop models need. One of the promising approaches is the use of GCM outputs to drive crop simulation models as they can predict climate in advance before the growing season and they can provide daily weather inputs. The GCM outputs have been used for water resources planning as well as addressing other socio-economic challenges (Sivakumar 2011). Several methods are used to downscale GCM outputs at the river basin scale which include multiple linear regression, robust regression,

ridge regression, artificial neural networks, and Bayesian neural networks to identify an appropriate transfer function in statistical downscaling models (Mishra and Singh 2009; Sivakumar 2011; Jeong et al. 2012).

Using raw daily GCM outputs directly as inputs to the crop model, however, is not advisable due to the biases in the GCM data (Baron et al. 2005). Before applying the GCM outputs for required applications, the bias needs to be removed for better results (Ines and Hansen 2006; Wood et al. 2004; Maurer 2007; Mishra and Singh 2009; Kyoung et al. 2010). Ines and Hansen (2006) applied a simple bias correction method and successfully removed the biases in rainfall frequency and intensity of daily GCM rainfall. When linked with a crop model, the two-step bias correction procedure they used improved simulated yields, although the yields were still under-predicted, which they attributed to the inability of the bias correction method to correct the time structure in daily GCM rainfall. Baigorria et al. (2008) adapted the method in regional climate model (RCM) outputs, including daily temperature and solar radiation, and showed that the improvements made to the RCM outputs corrected biases in the predicted yields as compared to using raw RCM outputs directly. Recently, Ines et al. (2011) showed that by linking bias corrected GCM rainfall with a stochastic disaggregation method to redistribute the corrected wet days properly in the time series could somehow correct the time structure mismatch in the bias corrected GCM rainfall, and further improve the prediction of crop yields.

Lacking in the above studies is the in-depth analyses of where these improvements in yields are coming from. Although it is generally accepted that the improvements made in the biases of the GCM rainfall frequency, intensity and dry spell lengths resulted in the improvements of predicted yields, nothing has been done to quantify the causal relationships. Delineating this information is crucial for the development of better crop yield prediction models, as this will provide insights into which climate information from the GCM is best to produce a more robust and reliable crop yield prediction system. This paper aims to present a hierarchy of crop prediction models based on dynamic simulation of crop growth using a crop model and combinations of downscaling schemes of GCM rainfall, and to extract information content from downscaled GCM rainfall variables using information theory for identifying both the downscaling technique as well as suitable precipitation variables. Since it is not possible to present the entire crop yield simulation results, our objective is only focused on Maize yield. We conduct our study in Katumani, Machakos, Kenya during the short October–December rainy season, using maize (*Zea mays*)—a staple food in the region, as a case study.

2 Methodology

2.1 Bias correction (BC), stochastic disaggregation (DisAg) and combination (BC-DisAg)

We follow the procedure of Ines and Hansen (2006) and Ines et al. (2011) for the bias correction of daily GCM rainfall. Basically, the GCM bias correction is done by:

(a) The number of wet days in the GCM for month I is corrected by fitting a GCM threshold ($x_{I,GCM}$) derived from the observed historical rainfall distribution:

$$x_{I,GCM} = F_{I,GCM}^{-1}\left[F_{I,obs}(x_{I,obs})\right] \quad (1)$$

where $x_{I,obs}$ is the threshold for a wet day from observations, and $F_{I,GCM}$ and $F_{I,obs}$ are empirical cumulative distributions of GCM and observed rainfall, respectively.

(b) Rainfall intensity of GCM is corrected by fitting the truncated rainfall distribution of the climate model into a two-parameter gamma ($G_{I,GCM}$) and then mapping it with the truncated, gamma-fitted observed rainfall distribution ($G_{I,obs}$). Correcting rainfall (x_i) for a day is done as

$$x_i' = \begin{cases} G_{I,obs}^{-1}\left[G_{I,GCM}(x_i)\right] & x_i > x_{I,GCM} \\ 0 \end{cases} \quad (2)$$

where x_i' is the corrected GCM rainfall for that day. Shape and scale parameters of GCM and observed rainfall gamma distributions are calculated using the maximum likelihood method.

The method outlined above is called BC1 in this paper.

The general notion of daily GCM rainfall is that it always over-predicts rainfall frequency, but this is not the case always. An update of BC1 called BC2 was designed to handle cases when GCM rainfall frequency is less than or more than the observed. When GCM rainfall frequency is less than the observed, we append a number of wet-day events with minimum rainfall amounts (i.e., $x_{1,obs} + 0.1$ mm) in the GCM rainfall CDF to match the observed rainfall frequency. The number of appended wet-days depends on the discrepancy between the number of wet-days in the GCM rainfall CDF above the calibrated threshold and observations. These wet days are added and distributed evenly (stochastically) in the GCM rainfall daily time series. BC2 also handles if the left-side of the empirical distribution of the GCM rainfall contains more values equivalent to $x_{1,GCM}$ above the truncation point (Ines et al. 2011) by preserving those values instead of eliminating them ensuring that rainfall frequency is preserved.

Linking the bias corrected GCM information with stochastic disaggregation (DisAg), i.e., through a conditional

stochastic weather generator (Hansen and Ines 2005), is done as follows:

We used monthly rainfall frequencies of the bias-corrected daily GCM rainfall to adjust the first- and second-order transition probabilities of a high-breed Markov chain rainfall occurrence model to correct the time distribution of wet and dry days. Formalism for adjusting the transition probabilities of a high-breed Markov chain rainfall occurrence model within the stochastic disaggregation can be found elsewhere (e.g., Katz and Parlange (1998) and Hansen and Ines (2005)) and is summarized below (see Ines et al. 2011).

The rainfall occurrence model of the stochastic weather generator used in the stochastic disaggregation is a two-state, hybrid second-order Markov chain able to simulate rainfall occurrence with a first-order chain if the previous day was wet, or a second-order chain if the previous day was dry (Hansen and Ines 2005). If the Markov model simulates the occurrence of rainfall on a given day, the rainfall amount is sampled from a mixture of two exponential distributions. Hansen and Mavromatis (2001) described details of the temperature and solar radiation sub-models.

The rainfall frequency of bias-corrected daily GCM rainfall was used to adjust the first- and second-order transition probabilities of the rainfall occurrence model to simulate time series of wet and dry days. Transition probabilities (i.e., first-order probabilities, e.g., wet day following a dry day (p_{01}), wet to wet (p_{11}), and second-order probabilities, p_{101} and p_{001}) to match a target rainfall frequency (e.g., from bias-corrected GCM rainfall) are related directly to the unconditional rainfall occurrence probability, π (Eq. 3) and persistence of dry days, ρ_1 (Eq. 4) (Katz and Parlange 1998):

$$\pi = \frac{p_{01}}{1 + p_{01} - p_{11}} \tag{3}$$

$$\rho_1 = p_{11} - p_{01} \tag{4}$$

With the assumption that persistence of dry days remains constant when a target rainfall occurrence probability changes (e.g., from bias corrected GCM rainfall), π', the first-order adjusted transition probabilities (with apostrophes) are determined by solving Eqs. (3) and (4) simultaneously, thus,

$$p'_{01} = \pi'(1 - \rho_1) \tag{5}$$

$$p'_{11} = \rho_1 + p'_{01} \tag{6}$$

Equations (5) and (6) are used to determine a wet day if the previous day was wet. If the previous day was dry, the second-order Markov chain is used to determine if the current day will be wet or dry through the adjusted second-order transition probabilities. The transition probabilities for the hybrid second-order Markov chain rainfall occurrence model can be adjusted for a given rainfall frequency by keeping the first- and second-order persistence of dry days (ρ_1 (Eq. 4) and ρ_2 (Eq. 7)) constant:

$$\rho_2 = p_{101} - p_{001} \tag{7}$$

The adjusted transition probabilities are given as follows (Eqs. 8–10) (Hansen and Mavromatis 2001; Katz and Parlange 1998):

$$p'_{11} = \pi'(1 - \rho_1) + \rho_1 \tag{8}$$

$$p'_{001} = \pi' \frac{(1 - \rho_1)(1 - \pi') - \rho_2(1 - p'_{11})}{1 - \pi'} \tag{9}$$

$$p'_{101} = p'_{001} + \rho_2 \tag{10}$$

If rainfall frequency and total are used at the same time to condition the stochastic weather generator, (i) transition probabilities (first- and second-order) of the rainfall occurrence model are adjusted based on bias-corrected monthly GCM rainfall frequency, (ii) daily rainfall realizations are generated iteratively until generated monthly rainfall total matches 95 % of the target value (e.g., bias-corrected monthly GCM rainfall), and (iii) generated daily values are re-scaled by a ratio of monthly target (R_m) and generated rainfall totals ($\bar{R}_m$) (i.e., $\frac{R_m}{\bar{R}_m}$) such that the monthly rainfall total generated matches the target value. These steps are repeated for each calendar month in a year for all the considered years (Hansen and Ines 2005). Based on BC1, BC2, combination of BC and DisAg (BC-DisAg), different downscaling methods were used in the study using different sources of rainfall information (Table 1).

2.2 Maize simulations

We used CERES-Maize (Ritchie et al. 1998) to evaluate the performance of the bias-corrected, and bias-corrected—disaggregated GCM rainfall in predicting maize yields. Data on soil properties (sandy clay loam), crop cultivar and management practices were based on a previous study at the study site (Keating et al. 1992). Every year, the water balance was re-initialized on 17 October with soil water at 20 % of the soil capacity. The date of sowing was decided when the soil water content exceeded 40 % of the soil capacity over the top 15 cm depth, or forced on 1 November, otherwise. The plant density was set at 4.4 plants m^{-2} and 20 kg N ha^{-1} as ammonium nitrate was applied at planting (low inputs system) (Ines and Hansen 2006). First, the crop model was run with observed daily rainfall, then with the rainfall outputs from the cases enumerated in Table 1. In the crop modeling, daily minimum (T_{min}) and maximum temperature (T_{max}) and solar radiation (SRAD) were all generated from monthly mean values conditioned on the occurrence or non-occurrence of rainfall.

2.3 Mutual information using nonparametric method

The relationship between rainfall variables and crop yield do not share a linear relationship as studied in many previous articles due the fact that rainfall variables

Table 1 Downscaling schemes conducted in the study

Serial no.	Methodology	Description of methods
1	Uncorrected	No bias correction
2	BC 1	Corrects under-prediction of rainfall frequency by truncating GCM rainfall distribution given a threshold. Corrects rainfall intensity using gamma–gamma transformation (Ines and Hansen 2006)
3	BC 2	Corrects rainfall frequency (under-/over-prediction) by truncating the GCM rainfall distribution given a threshold, corrects "nugget effect" truncating empirical. Corrects rainfall intensity using gamma–gamma transformation
4	BC 1-DisAg 2-Freqn	Stochastic disaggregation of BC1 corrected GCM rainfall frequency (averaged across 24 members)
5	BC 1-DisAg 2-Freqn-Total	Stochastic disaggregation of BC1 corrected GCM rainfall frequency (averaged across 24 members) + totals (averaged across 24 members)
6	BC 1-DisAg 2-Total	Stochastic disaggregation of BC1 corrected GCM rainfall totals (averaged across 24 members)
7	BC 2-DisAg 2-Freqn	Stochastic disaggregation of BC2 corrected GCM rainfall frequency (averaged across 24 members)
8	BC 2-DisAg 2-Freqn-Total	Stochastic disaggregation of BC2 corrected GCM rainfall frequency (averaged across 24 members) + totals (averaged across 24 members)
9	BC 2-DisAg 2-Total	Stochastic disaggregation of BC2 corrected GCM rainfall totals (averaged across 24 members)

(i.e., amount, intensity and frequency) are highly stochastic in nature. Therefore it is important to measure the information content between precipitation variables and crop yield, and entropy seems to be quite useful for measuring the information between two variables. The mutual information (MI) using entropy is used as a measure of statistical dependence among random variables which captures the full dependence structure, both linear and nonlinear and several applications made in the last decades (Moon et al. 1995; Sharma 2000; Mishra and Coulibaly 2010).

A high value of MI score would indicate a strong dependence between two variables. The mutual information between two random variables S and Q is defined as (Fraser and Swinney 1986):

$$MI(S,Q) = \int P_{sq}(s,q) \log\left[P_{sq(s,q)}/P_s(s)P_q(q)\right] ds dq \tag{11}$$

where $P_s(s)$ and $P_q(q)$ are the marginal probability density functions (pdf's) of S and Q respectively, and $P_{sq}(s,q)$ is the joint pdf of S and Q. In our study, S denotes maize yield, whereas Q represents individual precipitation variables (i.e., amount, intensity and frequency).

For any given bivariate sample, the MI score in Eq. (11) can be estimated as:

$$MI(S,Q) = \frac{1}{n}\sum_{i=1}^{n} \log_e\left[P_{sq}(s_i,q_i)/P_s(s_i)P_q(q_i)\right] \tag{12}$$

where s_i and q_i are the ith bivariate sample data pair in a bivariate sample of size n.

The multivariate kernel density estimator (Scott 1992; Wand and Jones 1995) using a Gaussian kernel function for estimating joint and marginal pdfs can be defined as:

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{d/2}\lambda^d \det(Sc)^{1/2}}$$
$$\times \exp\left(-\frac{(x-x_i)^T Sc^{-1}(x-x_i)}{2\lambda^2}\right) \tag{13}$$

where the mutivariate kernel density estimate is denoted by $\hat{f}_X(x)$, the sample covariance of the variable set X is denoted by Sc; and λ is known as the bandwidth of the kernel density estimate. In this study λ is chosen as the optimal Gaussian bandwidth for a normal kernel given as (Scott 1992; Silverman 1986):

$$\lambda = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{(-1/(d+4))} \tag{14}$$

where n and d refer to the sample size and dimension of the multivariate variable set, respectively.

3 Data and study area

The analyses are based on data from the Katumani Dryland Research Center ($1°35'S$, $37°14'E$, 1601 a.m.s.l) in the Machakos District of Eastern Kenya, a main maize growing region. Rainfall is bimodal in distribution and the climate is marginal for maize in both seasons. Because of strong food preferences, maize is the staple crop. The October–December short rainy season is an important maize growing season, and is fairly predictable at a seasonal lead-time using statistical (Indeje et al. 2000) and dynamic (Hansen and Indeje 2004) forecast models, making it interesting to test the utility of daily GCM rainfall from crop yield prediction.

For analysis, we used ECHAM4.5 daily rainfall outputs (Roeckner et al. 1996) forced by observed sea surface temperature (SST) (http://iridl.ldeo.columbia.edu). The climate model grid cell encompassing the Katumani Dryland Research Center in the Machakos district of eastern Kenya was selected. Daily rainfall outputs (1970–1995) from all 24 ensemble members in this grid cell were extracted. Figure 1 shows how skillful (R) is the uncorrected GCM rainfall in the study area. Daily rainfall observations from the research center were used for the GCM rainfall bias correction. Other input data needed for the quantitative evaluation (crop modeling) of the bias-corrected GCM daily rainfall were all collected from the research station (Hansen and Indeje 2004).

All 24 members were bias-corrected using the threshold, >0 mm, delineating a wet day. The bias corrections were conducted using the simple bias correction of Ines and Hansen (2006) (BC1) and that of Ines et al. (2011) (BC2) that accounts for both under and over predictions of rainfall frequency, and corrects the "nugget effect" in truncating empirical distributions. Monthly rainfall statistics for 26 years (1970–1995) were then extracted and used in the stochastic disaggregation (rainfall frequency, rainfall frequency + totals, totals).
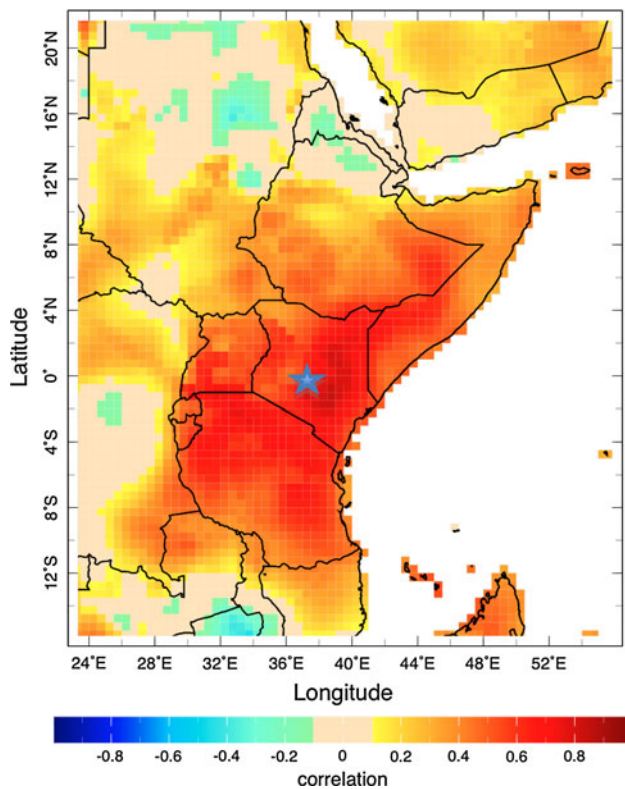

Fig. 1 Correlation of uncorrected ECHAM4.5 rainfall for October–November–December (OND) season (1970–1995) against University of East Anglia's gridded rainfall data. *Star mark* shows the location of Katumani Dryland Research Center

4 Results and discussion

4.1 Precipitation downscaling

The advantages of bias correction and the linkage of bias corrected rainfall information with stochastic disaggregation to the improvements of downscaled GCM rainfall is shown in Table 2 based on different goodness-of-fit criteria: correlation coefficient, denoted by R (Pearson product-moment correlation coefficient); root mean square error denoted by RMSE; and mean bias error denoted by MBE. Considering all three goodness-of-fit criteria, it can be said that for the month of October the BC2-DisAG2-Freqn (see Table 1 for definitions of methods) methodology performed better for total monthly precipitation. When the downscaled precipitation was not considered for bias correction, higher RMSE, MBE and lower R were observed. Interestingly, when the probability of occurrence was observed, the BC1-DisAg2-Freqn-Total methodology performed better. However, the performance of rainfall intensity for the month of October was not properly captured in most of the methodology except for the BC2-DisAG2-Freqn method which showed an improvement based on R, even though the relative performance of RMSE and MBE did not improve.

Based on the all three goodness-of-fit measures, the downscaled capability for the November total rainfall was found to be better for the BC2, BC2-DisAg2-Freqn-Total and BC2-DisAg2-Total methodologies. The performance of uncorrected downscaled rainfall could not be interpreted well using R, however based on RMSE and MBE its performance worsened. Therefore, it is important to consider all three performance measures in the selection of methodology to be considered. A similar observation was also made for uncorrected rainfall intensity for the November rainfall. For the month of December the total rainfall as well as the probability of occurrence, the overall BC2-DisAg2-Freqn methodology was found to be performing better for two statistical performance criteria, even though MBE was the lowest for other methodologies. However, the methodologies could not improve rainfall intensity in comparison to the total rainfall and probability of occurrence. Except for dry spell lengths, a majority of the biases in rainfall frequency, intensity and totals in the GCM rainfall were corrected by the deterministic transformation of the GCM rainfall. The corrections of under/over-prediction of rainfall frequency and the "nugget effect" of truncating empirical distributions (BC2) improved the corrections of rainfall frequency in the month of November, where rainfall occurrence was under-predicted by the GCM (not shown). Only when we linked the bias corrected GCM rainfall information (BC1 and BC2) with a stochastic weather generator that the majority of the dry spell length biases were reduced.

Table 2 Seasonal rainfall statistics from different downscaling schemes

Individual months	Precipitation variables	Goodness of fit	Methodology[a]								
			1	2	3	4	5	6	7	8	9
October	Total	R	0.33	0.37	0.38	0.42[b]	0.37	0.37	0.47[b]	0.38	0.38
		RMSE	1.41	1.28	1.28	1.26	1.28	1.28	1.23	1.28	1.28
		MBE	−0.54	−0.06	−0.02	−0.08	−0.06	−0.06	−0.04	−0.02	−0.02
	Probability of occurrences	R	0.26	0.34	0.32	0.34	0.51[b]	0.41[b]	0.34	0.47[b]	0.38
		RMSE	0.41	0.13	0.13	0.13	0.12	0.12	0.13	0.12	0.13
		MBE	0.39	−0.01	0.00	−0.01	0.01	0.00	−0.01	0.02	0.01
	Intensity	R	0.24	0.02	0.05	0.32	0.22	0.22	0.45*	−0.10	0.21
		RMSE	6.98	5.26	5.20	4.96	5.14	5.14	4.88	5.43	5.20
		MBE	−4.90	−0.14	−0.15	1.15	1.48	1.48	1.29	1.26	1.45
November	Total	R	0.74[b]	0.75[b]	0.75[b]	0.66[b]	0.75[b]	0.75[b]	0.63[b]	0.75[b]	0.75[b]
		RMSE	3.94	1.76	1.69	2.15	1.76	1.76	2.08	1.69	1.69
		MBE	−3.34	−0.56	0.07	−0.51	−0.56	−0.56	−0.18	0.07	0.07
	Probability of occurrences	R	0.56[b]	0.53[b]	0.55[b]	0.59[b]	0.51[b]	0.52[b]	0.43[b]	0.50[b]	0.56[b]
		RMSE	0.14	0.14	0.12	0.13	0.15	0.13	0.14	0.14	0.12
		MBE	0.06	−0.07	−0.01	−0.06	−0.05	−0.03	−0.02	0.00	0.00
	Intensity	R	0.63[b]	0.61[b]	0.62[b]	−0.30	0.58[b]	0.58[b]	0.29	0.58[b]	0.55[b]
		RMSE	7.56	2.95	2.83	3.74	2.97	2.97	3.35	2.94	2.92
		MBE	−6.90	−1.07	−0.74	0.27	0.25	0.25	0.18	0.37	0.28
December	Total	R	0.42[b]	0.43[b]	0.43[b]	0.39[b]	0.43[b]	0.43[b]	0.47[b]	0.43[b]	0.43[b]
		RMSE	2.05	1.66	1.69	1.43	1.65	1.65	1.38	1.68	1.68
		MBE	−1.48	−0.18	0.00	−0.11	−0.15	−0.15	0.19	0.05	0.04
	Probability of occurrences	R	0.53[b]	0.53[b]	0.52[b]	0.46[b]	0.50[b]	0.52[b]	0.55[b]	0.47[b]	0.47[b]
		RMSE	0.25	0.13	0.13	0.13	0.15	0.13	0.13	0.16	0.13
		MBE	0.22	−0.02	0.02	−0.01	−0.01	−0.01	0.02	0.02	0.00
	Intensity	R	0.18	0.17	0.18	0.25	0.15	0.15	0.25	0.21	0.19
		RMSE	5.74	3.23	3.30	2.51	2.66	2.66	2.53	2.61	3.01
		MBE	−5.14	−1.40	−1.54	0.18	0.34	0.34	0.40	0.16	0.33
Dry spell length		R	0.71[b]	0.69[b]	0.69[b]	0.68[b]	0.71[b]	0.70[b]	0.58[b]	0.75[b]	0.66[b]
		RMSE	7.55	4.26	4.26	4.63	4.51	4.42	4.81	3.89	4.47
		MBE	6.34	0.02	0.02	1.58	1.72	1.43	0.55	0.18	0.75

[a] The serial number of methodology used is based on Table 1

[b] Indicates the 'R' values which are statistically significant at 95 % two-tailed significance levels

4.2 Mutual information (MI) between downscaled precipitation variables and crop yield

Crop yield depends on multiple precipitation variables which include amount, probability of occurrence, intensity and dry spells. These variables may be on a seasonal time scale as well as for individual months. Different downscaling methods performed differently for the evaluation of variables calculated using downscaled precipitation and exploring their relationship with crop yield prediction. Therefore, it is important to identify suitable downscaling techniques as well as important precipitation variables, which can be useful for seasonal crop yield prediction based on the GCM output.

The mutual information values were calculated between seasonal precipitation amounts along with constituent months obtained from different downscaling methods with respect to crop yield, as shown in Fig. 2a. The MI was calculated between maize yield and downscaled precipitation variables. A high value of MI score would indicate a strong dependence between two variables. It was observed that the downscaled seasonal precipitation amount obtained from BC2-DisAg2 frequency shared more information with the crop yield. Based on overall observation, seasonal

precipitation amount shared more information, whereas the October precipitation amount shared the least information. However, the November and December month precipitation provided nearly similar type of information with crop yield. This is logical, because the month of October comprises most of the vegetative stage of the crop, and November and December are when flowering and milking stage occur, which are very sensitive to the water availability. Timing and severity of water stress during this period would drastically decrease crop yields. However, it can be said that increase in seasonal precipitation as well as

the November and December month rainfall increases crop yields (Fig. 2). The next downscaling method which shared maximum information is BC1-DisAg2 frequency, which is again based on the correction of frequency but with the daily GCM rainfall corrected by the older BC (Table 1).

The mutual information between precipitation variables and crop yield based on the probability of occurrence differed from the precipitation amount on the seasonal scale (Fig. 2b). However, the downscaled probability of occurrence using BC2-DisAg2 frequency for the month of December shared maximum information with crop yield.
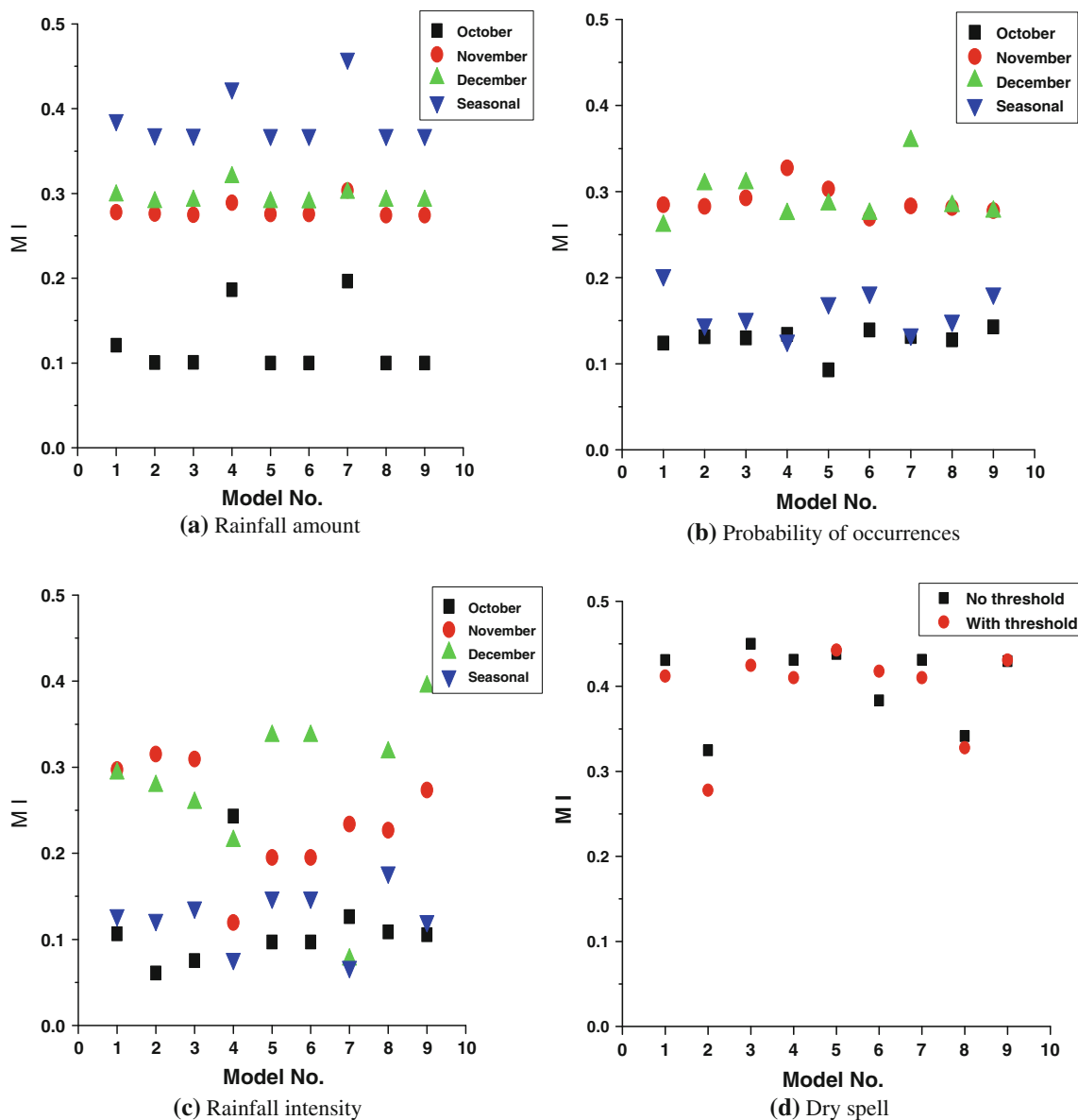

Fig. 2 The mutual information content between seasonal and individual month's **a** rainfall amount, **b** probability of occurrences, **c** rainfall intensity, and **d** dry spell for different models with the crop yield

Overall December and November month's probability of occurrence provided higher information in comparison to the seasonal and October time scales, suggesting that rainfall frequency information from seasonal time scale, as well as those from the less-sensitive period of crop growth, were not better predictors of crop yields in the study location as compared to the November and December rainfall frequencies. This is a critical insight, because in current climate prediction system, it is the seasonal values that are predicted. If a climate model is skillful to provide rainfall frequency information during the expected critical periods of crop growth, one may be able to predict better crop yields before the end of the growing season.

Similarly, the information content between precipitation intensity and crop yield differed from that of probability of occurrence as well as precipitation amount (Fig. 2c). Higher information content was observed between precipitation intensity of December downscaled based on the BC2-DisAg2-Total and interestingly it was least for December based on the BC2-DisAg2 frequency. Seasonal as well as the October precipitation intensity did not share much information with crop yield in comparison with months of December and November. The reason for rainfall intensities generated by stochastic disaggregation of rainfall frequency not giving better information could be due to the rainfall intensity model parameters not being constrained/adjusted on those months of interest. In other words, the rainfall intensities generated did not conform to the target rainfall intensities, but were likely representative of climatology (Hansen and Ines 2005). On the other hand, when we combined monthly rainfall total with rainfall frequency or using it alone in the stochastic disaggregation, we generated more appropriate rainfall intensities representing those months of interest, as shown by the increased MI in the months of November and December (Fig. 2c). It should be noted however that we did not use rainfall intensity as a predictor in the stochastic disaggregation. Usually, rainfall intensity was less (or not) predictable compared with other rainfall variables (Moron et al. 2007). Ines et al. (2011) found that the GCM model had lesser skill for predicting the October–November–December rainfall intensity ($R = 0.37$) than rainfall frequency and totals ($R \approx 0.70$) in the study area.

Dry spell lengths during the anthesis stage (November 15–December 31), on the other hand, were found to give the highest information with respect to the predictability of crop yields (Fig. 2d). For all the downscaling schemes tested, mutual information was greater than 0.30, except for BC1 (see Table 1) when a threshold (continuous >3 days no rain) was imposed to define a dry spell. Even for the case of no bias correction, MI measured a strong relationship between yield variability and dry spell during this period. It should be noted however that the bias in the

predicted yield was not measured by entropy. The yield bias diminished as we applied the hierarchy of downscaling schemes especially when we stochastically corrected the timing of wet/dry days within the time series.

5 Conclusions

The mutual information content between precipitation variables and crop yield plays an important role for seasonal crop prediction using the GCM output. This study derives a set of downscaling techniques to explore the mutual information between precipitation variables with crop yield. The following conclusions are drawn from this study:

(i)   The bias correction and stochastic disaggregation and their combination help improve the performance of downscaled GCM rainfall. It is worth noting that the performance of rainfall intensity could not be improved significantly; however the bias is greatly improved in comparison to the uncorrected precipitation. The performance of different methodologies varies among precipitation variables. Even though the performance measure based on the coefficient of correlation does not change much for the uncorrected rainfall, however, significant changes are observed in removing the bias.

(ii) The total seasonal rainfall amount obtained from the BC2-DisAg2-Freqn methodology plays an important role for crop yield in comparison to the October rainfall.

(iii) The probability of occurrence of rainfall during November and December have more importance than the October and seasonal time period in the study area. Overall the probability of occurrence obtained in December after stochastic disaggregation of BC2 corrected GCM rainfall frequency performs better.

(iv) The rainfall intensity observed during the month of December has more contribution to the better crop yield prediction than other monthly and seasonal time periods. Based on all the methodology, the BC2-DisAg2-Total performs better for rainfall intensity of the December month.

(v)   The performance of mutual information based on the dry spell shows a larger variation among different models as well as based on threshold and no threshold conditions. Overall the BC 2 method performs better with no threshold conditions. When all rainfall variables are compared it can be observed that the performance of a dry spell is an important contributing factor for crop yield as higher information is

shared based on the mutual information. Therefore, the challenge is to predict the dry spells on a seasonal lead time for better crop yield simulation.

References

Baigorria GA, Jones JW, O'Brien JJ (2008) Potential predictability of crop yield using an ensemble climate forecast by a regional circulation model. Agric For Meteorol 148:1353–1361

Baron C, Sultan B, Balme M, Sarr B, Traore S, Lebel T, Janicot S, Dingkuhn M (2005) From GCM grid cell to agricultural plot: scale issues affecting modelling of climate impact. Philos Trans R Soc B 360:2095–2108

Fraser AM, Swinney HL (1986) Independent coordinates for strange attractors from mutual information. Phys Rev A 33(2):1134–1140

Hansen JW, Indeje M (2004) Linking dynamic seasonal climate forecasts with crop simulation for maize yield prediction in semi-arid Kenya. Agric For Meteorol 125:143–157

Hansen JW, Ines AVM (2005) Stochastic disaggregation of monthly rainfall data for crop simulation studies. Agric For Meteorol 131:233–246

Hansen JW, Mavromatis T (2001) Correcting low-frequency variability bias in stochastic weather generators. Agric For Meteorol 109:297–310

Hansen JW, Tippett M, Bell M, Ines AVM (2010) Linking seasonal forecasts into riskview to enhance food security contingency planning. TR10-12. IRI Technical Report, New York

Indeje M, Semazzi FHM, Ogallo LJ (2000) ENSO signals in East African rainfall and their prediction potentials. Int J Climatol 20:19–46

Ines AVM, Hansen JW (2006) Bias correction of daily GCM rainfall for crop simulation studies. Agric For Meteorol 138:44–53

Ines AVM, Hansen JW, Robertson AW (2011) Enhancing the utility of daily GCM rainfall for crop yield prediction. Int J Climatol 31:2168–2182

Jeong DI, St-Hilaire A, Ouarda TBMJ, Gachon P (2012) Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada. Stoch Environ Res Risk Assess 26:633–653

Katz RW, Parlange MB (1998) Overdispersion phenomenon in stochastic modeling of precipitation. J Clim 11:591–601

Keating BA, Wafula BM, Watiki JM (1992) Exploring strategies for increased productivity—the case for maize in semi-arid Eastern Kenya. In: A search for strategies for sustainable dryland cropping in Semi-arid Eastern Kenya, ACIAR proceedings, no. 41. Australian Centre for International Agricultural Research, Canberra, pp 90–101

Kyoung MS, Kim HS, Sivakumar B, Singh VP, Ahn KS (2010) Dynamic characteristics of monthly rainfall in the Korean Peninsula under climate change. Stoch Environ Res Risk Assess 25:613–625

Maurer EP (2007) Uncertainty in hydrologic impacts of climate change in the Sierra Nevada, California under two emissions scenarios. Clim Chang 82(3–4):309–325

Mishra AK, Coulibaly P (2010) Hydrometric network evaluation for Canadian watersheds. J Hydrol 380(2010):420–437

Mishra AK, Singh VP (2009) Analysis of drought severity–area–frequency curves using a general circulation model and scenario uncertainty. J Geophys Res 114:D06120. doi:10.1029/2008JD010986

Moon Y, Rajagopalan B, Lall U (1995) Estimation of mutual information using kernel density estimators. Phys Rev E 52(3):2318–2321

Moron V, Robertson AW, Ward MN, Camberlin P (2007) Spatial coherence of tropical rainfall at the regional scale. J Clim 20:5244–5263

Ritchie JT, Singh U, Godwin DC, Bowen WT (1998) Cereal growth, development and yield. In: Tsuji GY, Hoogenboom G, Thornton PK (eds) Understanding options for agricultural production. Kluwer Academic Publishers, Dordrecht, pp 79–98

Robertson AW, Ines AVM, Hansen JW (2007) Downscaling of seasonal precipitation for crop simulation. J Appl Meteorol Climatol 46:677–693

Roeckner E, Arpe K, Bengtsson L, Claussen CM, Dümenil L, Esch M, Giorgetta M, Schiese U, Schulzweida U (1996) The atmospheric general circulation model ECHAM-4: model description and simulation of present-day climate, report no. 218. Max Planck Institute for Meteorology, Hamburg

Scott DW (1992) Multivariate density estimation: theory, practice and visualisation. In: Probability and mathematical statistics. Wiley, New York, p 317

Sharma A (2000) Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: part 1: a strategy for system predictor identification. J Hydrol 239:232–239

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, New York

Sivakumar B (2011) Global climate change and its impacts on water resources planning and management: assessment and challenges. Stoch Environ Res Risk Assess 25:583–600

Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall, London

Wood AW, Leung LR, Sridhar V, Lettenmaier DP (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Clim Chang 62:189–216