Department of Applied Economics and Management

Christopher B. Barrett
Associate Professor
Co-Director, African Food Security & Natural Resources Management Program

Cornell University
315 Warren Hall
Ithaca, NY 14853-7801

Telephone: 607-255-4489
Fax: 607-255-9984
E-mail: cbb2@cornell.edu
Web: aem.cornell.edu/faculty_sites/cbb2/

January 22, 2002

Memorandum

To: BASIS CRSP project team

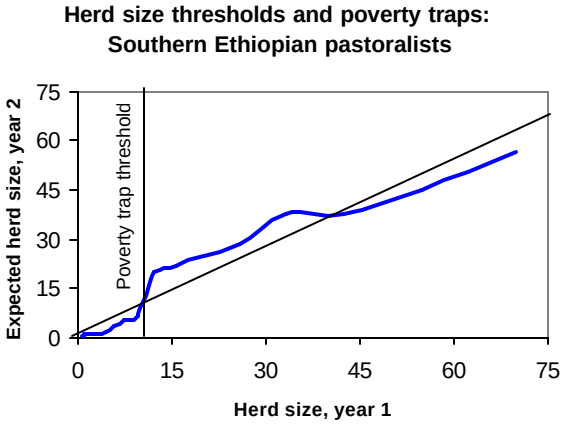
Subject: Thoughts on data collection and analysis

Hello, everyone. At long last I have finished writing up my thoughts on data collection and modeling issues in our BASIS project. I apologize profusely for having failed to produce this a couple of months ago, as intended. It unfortunately kept getting pushed to the back burner. My apologies for any inconvenience caused by my tardiness. Anyway, here goes, with a mix of conceptual issues and mundane details.

It is important that we each think about the bigger picture design of this project and what it implies for the details of data collection and subsequent data analysis. In essence, there are three steps to the research component of this project: (i) descriptive analysis of wealth and welfare dynamics, (ii) econometric analysis of the causal factors underlying poverty traps, and (iii) calibration and validation of the CLASSES bioeconomic modeling tool for particular stylized sites. [There are also training and outreach activities I'm not addressing in this memo.] So let me organize the bulk of this memo around these three steps.

Step one: Descriptive analysis of wealth and welfare dynamics

In the first stage of the project, we document wealth and welfare dynamics as a means of describing how wealth or welfare at one point in time (period t) projects to wealth or welfare at a future point in time (period t+i). Do there indeed appear to be poverty traps? If so, for what portion of the subpopulation? This is really just descriptive statistics, albeit dressed up a bit, most likely in the form of transition matrices and the sorts of recursion diagrams that I presented in Kerugoya (using the plot of herd size dynamics in



southern Ethiopia, reproduced in slightly different form at right). These convey good, basic information about how observed conditions today are expected to project into conditions tomorrow, especially how these mapping might vary with current wealth or welfare. These statistics and tools are also readily understood by lay audiences, which is important to communicating our results. The transition plots can also help us to identify thresholds, points at which dynamic behavior bifurcates, which is extremely important to modeling poverty traps.

In practice, this descriptive analysis requires us to represent the concepts of wealth and welfare with specific, measurable variables (or functions of variables). The appropriate variables will depend on (i) the site and (ii) the variables that were measured in the previous survey(s) on which we are building in this project. For example, livestock ownership is clearly the most relevant measure of wealth in the northern Kenya sites, so we're measuring the exact same livestock variables in repeated periods to capture the dynamics of wealth transitions in this particular form among this particular subpopulation. In order that we are truly comparing the same variable over time, ***it is absolutely essential that the new round of surveys retain the original wording from the previous survey round*** . If we start tinkering with definitions and wordings so as to achieve greater standardization across the survey instruments fielded in different sites, we'll lose comparability across time, which is far more important for present purposes.

As far as variables go, and having looked over almost all the original questionnaires now, let me first tackle the wealth dynamics issue. In general, we will want to capture reasonable proxies (where "reasonable" varies a bit by site) for each of the following sorts of wealth: (i) financial capital stocks (e.g., bank savings), (ii) livestock owned (by species so as to be able to aggregate into tropical livestock units), (iii) land owned, (iv) land / soil quality, (v) value of other agricultural / pastoral capital owned (e.g., tractors, trees), (vi) nonfarm productive assets (e.g., housing stock, radio / TV / major household items, transport, stores). We will likely create both wealth dynamics measures specific to particular assets – especially where that asset accounts for the vast majority of wealth, as in the case of livestock in the pastoral systems – and create aggregate wealth measures (whether money metric or indices based on factor analysis) the dynamics of which we will also study.

Most of these will be reasonably straightforward, with the notable exception of land / soil quality, where we will commonly need to rely on coarse categorical measures of quality and soil type. Based on Frank's earlier comments, discussions at Kerugoya and subsequently with Ben, I'm thinking that we'll need to rely on respondents' subjective perception as to to the current quality of their soils and as to what has happened to the condition of their soils since the last survey (perhaps using a 5-point ordinal scale: it's improved a lot, improved a

little, stayed roughly the same, deteriorated a little, deteriorated a lot). Is there any point in trying to break down the soil conditions finer (e.g., fertility, drainage, erosion, etc.)? I'm inclined to think not, but welcome others' thoughts on this. [Frank, you've had to deal with this in western Kenya in the past, and Jean Claude, you dealt with this in the 2000 survey in Madagascar, so you two in particular might have good insights.]

As asset stocks are usually well registered with individuals. So we can and should ask recall questions to reconstruct their investments in productive assets (land, livestock, trees, housing, vehicles, education, soil and water conservation structures) since the last survey. Since several participants at the Kerugoya meeting emphasized the importance of education, it will be especially important to be sure we capture accurately any further investments in education (i.e., do not assume that educational attainment remains unchanged from the previous survey for household members).

The welfare dynamics issue is a bit more tricky because welfare is a somewhat more elusive concept than wealth. In particular, we face the classic income versus expenditures versus non-money metric welfare measurement challenge. Although in principle I prefer expenditures (for the usual reasons associated with the permanent income hypothesis) and non-money metric measures such as anthropometric status, we haven't got funds to do the latter properly and we do not have baseline data for expenditures in all of the survey sites. I therefore suggest that we collect (disaggregated) income data in all sites and use that as the common basis for intertemporal welfare comparisons within a site. But we also collect (disaggregated) expenditures data in those locations where such data were collected previously.

A key concern in panel data collection and analysis is sample attrition. Because we want to understand what is likely to happen to a household in time t+1 conditional on its state at time t, we need to be careful that the households for which we have complete observations on both time t and time t+1 are not a biased subsample of the original sample at time t. For example, if all the chronically poor people migrate out of the area or die between the two survey rounds and we estimate off just the remaining subsample, we may get a highly inaccurate view of poverty as almost entirely a transitory phenomenon. [The opposite mistaken inference and lots of others are obviously possible as well.] So we need to really push enumerators to track down every household from the prior survey. If the household has left the area, if possible, we should make contact with them at their new location and establish why they left. When this is quite far from the survey site, it will probably prove infeasible. In such cases, we will need to establish what happened to those households that can no longer be found by questioning neighbors or local leaders (including teachers or postal

workers) to find out whether the household migrated (and why), died (when and how), never existed (i.e., baseline data were in error … or worse), or what have you.

Another serious concern in panel data in changing household structure. Especially where several years have elapsed since the last survey round, we will likely find many individual movements into or out of the household. The most severe form is household splitting, as when a married couple divorces and each takes some of the assets of the household. Because we need to control for household composition in welfare measures, we need to document changes in household composition carefully. There is no unambiguously preferable way to handle household splitting. My instinct is to follow the simple rule of resurveying only that part of the household now controlled by the household head identified on the previous survey round or, if s/he is no longer available, then resurvey the household associated with the next-oldest surviving member of the household from the previous round who continues to reside in the location, and use this "offspring" household as the successor round representation. I'm not at all wedded to this particular approach, however, so let's have an email (and in person) discussion about this so that we have a common protocol on this across the sites, ok?

However we handle household splitting, in the resurveyed households, we need to identify which of the previous household residents remain, which have left and why, and any individuals who have joined the household since the last survey. In panel data circles, this is sometimes known as the "same ID code" method of household roster verification. (On this and related technical details of panel data collection, see the panel data chapter by Paul Glewwe and Hanan Jacoby in Margaret Grosh and Paul Glewwe, eds., *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 years of the Living Standards Measurement Study*, Washington: World Bank, 2000. I can provide a copy of this to anyone who needs it.) In updating the household roster, we need to take care, therefore, to record all deaths of the last round's members (cause of and age at death as well), births, fostering (in or out) of children, and other in- or out-migrants.

<u>Action items for step one:</u>

1. Each site team needs to use its old questions on assets, income, and expenditures to establish wealth and welfare dynamics in that location. Some recall questions may need to be added (e.g., on soil quality) where baseline data is lacking.
2. Each site team needs to ensure it has a clear and consistent household ID code system so that new data from each household can be readily matched to old data on the same household.

3. We need to come to agreement across the survey sites on a protocol for tracking down households that no longer reside in the survey site, for the information we want on households that cannot be found, and on how to handle household splitting (i.e., which branch to track, or both).

Step two: Causal analysis of poverty traps

The second step in the project involves understanding why we find the relationships observed in step one. Our early project documents advance four hypothesized sources of poverty traps:

1. Poor market access creates significant fixed costs to market participation, and poorer producers in areas of weak market access tend to opt out of markets in favor of low-return self-sufficiency.
2. High-return production strategies (e.g., dairy) entail significant fixed costs that result in a minimum efficient scale of investment and operation commonly beyond the reach of the poorest people.
3. Poorer households lacking capital to finance productive investments may be unable to undertake lumpy investments, regardless of their expected returns.
4. Risk and subsistence constraints discourage poorer, more risk-averse households from accumulating assets and increasing productivity.

These hypotheses now drag us into the realm of cross-sectional data needs, for which we do not necessarily need repeated observations on the same variables for each household. This is quite different from the dynamic wealth and welfare data we need for step one, above. [As an aside, Larry Blume and I have been working, slowly, on putting together a conceptual / theoretical paper on poverty traps to underpin this section of the work. It should be ready in time for the June 2002 team meeting in Kakamega. This lays out the structural origins of the reduced form relationships described in the above four hypotheses. Ultimately, we're just going to be doing reduced form estimation, so I'm not bothering to include this material here.]

Hypothesis one concerns the costs of market access, i.e., distance and transport modes / costs to appropriate locations to buy / sell inputs and outputs. Fixed costs (presumably related to getting the person to market irrespective of the volume transacted) are especially important here, but so are variable costs associated with per unit transport / assembly / storage / taxation as well as search costs ("how frequently do you fail to find a buyer when you go to market?" "how much time does it take for you to find a buyer when you go to market?"). So each questionnaire will need a section that identifies the markets (if any) the respondent uses and what it cost for them to go to market the last few times they purchased / sold any inputs / outputs (i.e., we're not interested in the last time someone visited the nearest stall to buy a half kilo of sugar). In order to try to get

at whether changing market access contributes anything to changing welfare and wealth, we need also to ask whether the cost of getting to market has increased, decreased or stayed about the same since the last survey. Should we be asking this with respect to multiple markets (in space or commodity)? But which ones? This probably varies across sites.

Hypothesis two gets into the nature of the shape of production (or cost or profit) functions, especially multi-output production functions. Are there economies of scale or scope? To estimate these production function relationships, we need to know detailed information on inputs to and outputs from crop and livestock production (and nonfarm activities, especially value-added activities in agriculture such as milling or butter or yogurt making). For the bioeconomic model in step three, we will need the crop production data to be collected at plot level. Note that we do not need longitudinal data on these variables, so some (imperfect) harmonization of questions here can be helpful in so far as it yields greater comparability across sites. One question that we need to be sure to include in this section relates to how long the plot has been in continuous cultivation (and whether it rests seasonally or never) and the duration of its last fallow. By knowing something about recent soil nutrient amendments (inorganic fertilizer, lime, rock phosphate, manure), soil type, time since the plot was brought into cultivation, and duration of last fallow, we can begin to estimate soil degradation functions using the chronosequence as a set of treatments.

Hypothesis three focuses on financing constraints that impede investment. Each site's survey instruments therefore needs to include a section on access to both formal and informal (i.e., nonbank) credit, asking not just whether people borrowed but also whether they would have liked to have borrowed more at the interest rate they received on any loans they took out (in order to establish whether they were quantity rationed in the credit market) and why they didn't. In the case of respondents who borrowed nothing, we need to ask why they didn't and whether they would have liked to borrow at prevailing local interest rates. We should also ask what they would do with a new loan if they were to receive one (i.e., pay school fees for children? Buy a new radio? Buy a cow or a sack of fertilizer?).

It is important here to be sure we've also captured the nonfarm income households generate, since people commonly use labor markets to overcome financial market failures. The salary one household member earns as a teacher or factory worker can enable that family to undertake crucial investments (even small ones, such as fertilizer) that an otherwise identical household cannot make. Similarly, we need to be sure to understand the agricultural calendar people follow and input/output purchases related to farming activities in order to understand whether crop/livestock revenues are spread across the year in a way

that facilitates investment or bunched in a single season. People often refer to the steady stream of revenue as one of the big advantages of smallholder dairying for crop producers. But there's also the question of off-season cropping and whether the income from one season's harvest helps with planting season investments in the main crop. Chris Moser's work on SRI in Madagascar certainly suggests that off-season potato and barley crops help finance adoption of improved rice technology by Malagasy smallholders.

Hypothesis four concerns risk and subsistence constraints. The latter is basically the physical manifestation of the financing constraints in hypothesis three. Risk, however, can be understood in either or both of two ways. First, people may choose not to undertake investments ex ante because of riskiness. As Frank pointed out in an earlier note, this raises the question of heterogeneous time preferences and risk aversion. We can and should try to estimate those parameters using some of the (relatively cumbersome) methods available and see whether these explanations indeed seem to contribute to understanding any observed patterns of poverty traps. I don't think we want to invest a great deal in this particular area, however, given that time and risk preferences are both simply measures of curvature in unobservable utility functions and are therefore highly sensitive to model specification. Given the relatively small sample sizes involved and the nature of the data we're collecting, I frankly don't expect to find very robust estimates of households' risk or time preferences and therefore would not want to stake a great deal on those explanations for whatever dynamics we observe. But I'll be very happy to be proved wrong on this one and, again, we need to take a look at these issues.

The second sense in which risk may matter is the way in which people cope ex post with shocks to their livelihoods. If the poor are observed to suffer serious shocks more frequently, if they tend to have suffered more severe shocks, or if they have to liquidate productive assets more frequently for a given level and frequency of shocks experienced, then vulnerability and poverty become inextricably linked in a way that matters to wealth and welfare dynamics. To get at this issue, we need to include a section on shocks in each questionnaire, wherein we ask whether over the period since the last survey this household has experienced any of the following types of shocks: (i) human injury or illness necessitating hospitalization or continuous medical treatment, (ii) death of a household member, (iii) death of a family member not resident in the household, (iv) complete or near-complete crop loss due to drought, (v) complete or near-complete crop loss due to causes other than drought (specify the cause: flood, hail, locusts, etc.), (vi) complete or near-complete herd loss (specify the cause: theft, wildlife predators, drought, disease, etc.), (vii) loss of permanent employment by a household member, (viii) major cut in household income due to falling price of crop or livestock (identify relevant commodity), or (ix) household

or business loss due to fire, theft or violence (identify cause). For each of these shocks, identify (a) month(s) and year(s) when this happened, (b) what steps were taken by household members in response to this difficulty? (to which there can be multiple, coded replies, e.g., nothing, pray, sell off livestock, sell next harvest in advance at below-market price, sell household food supply, sell land, sell other household assets, use savings, borrow money using land or crop as collateral, borrow money without any collateral, receive assistance from friends or family, receive food aid, cut household expenses, take on an extra job, other (identify)), and (c) what was the approximate cost of attempts to overcome this difficulty? Questions along these lines have been fielded successfully in other studies (e.g., in Indonesia) and could go a long way toward helping explain who falls into poverty and why.

Empirical exploration of these four hypotheses will provide the foundation for our analysis of the causal factors behind wealth and welfare dynamics and their interrelationship. Here we should be able to build on the approach Michael Carter and Julian May present in their December 2001 *World Development* article on poverty dynamics in South Africa (I'm attaching the pre-publication working paper as a .pdf file). Carter and May project welfare off of wealth measures, interpreting the resulting predicted welfare values as the structural component, then comparing these predicted values to observed welfare and interpreting the difference as the transitory component to current welfare. This enables them to try to identify what subpopulation of the poor in any period appear to be structurally poor as opposed to suffering a transitory, negative deviation from long-term welfare. Since core policy concerns about poverty stem less from its existence than from its persistence, this is a very helpful first step (and one that Michael, as the BASIS CRSP Director, would obviously be keen to see us pursue in this project).

As much as we all like hard, quantitative data based on good survey techniques, it will be very important to supplement the quantitative descriptions and inferences we draw from the survey data with qualitative work designed to draw out the stories behind the numbers. We discussed this in Kerugoya and I know that the Ben, Festus, Frank and Willis discussed this further during Ben's recent trip. The question is how to proceed. Here is my suggestion, based on various discussions with individuals on the team.

One can think of a two-by-two matrix defined by above/below poverty line in each period. I propose that we do open-ended interviews to capture the oral histories of eight households in each site, two each in the $poor_t$- $poor_{t+1}$, $nonpoor_t$-$poor_{t+1}$, $poor_t$- $nonpoor_{t+1}$, $nonpoor_t$- $nonpoor_{t+1}$. How did some people climb out of poverty? Why can't others climb out? How do some fall into poverty? This necessarily means that the qualitative work has to be sequenced with the

quantitative work – once data entry and basic descriptive analysis has been done, most likely by September-October 2002 – which has the added advantage of buying us a bit of time to find the funds to add this component on to the study. We would probably want a good anthropologist or sociologist to do this work. I believe Willis and Frank have already identified a couple of people at the University of Nairobi for the western and central Kenya work. Perhaps they could work up north too, or maybe Kevin Smith or someone else could do Baringo and Marsabit? What do you think, John? Bart and Jhon, do you have ideas on this for the Madagascar sites?

One final data collection issue in step two concerns replacement and/or supplementation sampling. In the ideal world, one wants to have data that are statistically representative of the community in each survey period. Typically, the first round of the survey randomly sampled households from a household sampling frame constructed for sample villages (so-called "cluster sampling"). So the first round is statistically representative of the village at time t. At time t+1, however, the village population has almost surely changed, so simple revisiting of the period t survey households, even if we found each and every one, would no longer be statistically representative of the village in period t+1. The usual solution to this is to establish how many additional households have joined the village since the previous survey and then replace any "lost" households from the previous round and supplement appropriately with additional households by randomly sampling from among the new households using the same sampling fraction that was applied in the first round. [For example, if one of every forty households was selected into the sample in the first round and when we return for the second round, we find the village has grown by 120 households and we cannot locate two of the old survey households, then we randomly draw five households from the set of new households to supplement with three new households as well as to replace the two lost households.] This requires a bit of extra effort at the start to create the sampling frame of new households and to establish (at least an estimate of) the change in the number of households since the previous period. But this way the cross-sectional estimates from either round yield statistically representative parameter estimates of production functions, market participation functions, etc. The key, therefore, is that we must have the original household count and sampling ratio from the previous round. Does each site team have those data on the previous survey round?

<u>Action items on step two:</u>

1. While in Kenya and Madagascar, I will go over questionnaires with each of the teams. We need to come to agreement, in particular, on what market access coverage is appropriate for the site in question.
2. In the coming week, Ben will go through all the baseline survey questionnaires again to establish which one(s) have the best modules on

market access costs, production, credit, and shocks and will send these to the whole team with his thoughts as to what other bits are missing.
3. Frank and Willis are working on a proposal for funding the qualitative work in Kenya. Jhon, Bart and I will need to talk about this for Madagascar. Need to identify individuals who would carry out the qualitative work in each site (John, who for the northern Kenya sites?).
4. Each site team must identify the total sampling frame count and the sampling ratio from the previous survey round in order to do proper replacement and supplemental sampling for the new round. Please confirm to Chris that you have these figures in hand.

Step three: Predictive/prescriptive analysis based on CLASSES model

The third step in the project will be to use the descriptive evidence assembled in the preceding two steps to create a predictive bioeconomic model that can be used for policy analysis, in particular for ex ante impact assessment. This is the Crop, Livestock and Soils in Smallholder Economic Systems (CLASSES) model, the development of which Ben is spearheading. We expect to have a crude prototype ready for the October 2002 bioeconomic modeling course in Ithaca. [The June two-day short course will be on general techniques and methods, not on CLASSES specifically.]

CLASSES requires calibration, some of which will be done econometrically (e.g., using the production functions estimated above) and some of which will be done statistically using starting values from the literature and from preexisting biophysical studies of the sites. Alice Pell leads a Cornell-ICRAF-KARI group that has proposed a major NSF biocomplexity grant that would undertake detailed, rigorous study of the biophysical elements of the Baringo, Embu and Siaya/Vihiga sites if it wins funding (we should know before the June 2002 team meeting). But we are presently working on the assumption that Ben will do the rest of the biophysical calibration off of secondary data collected by agricultural scientists working in these and similar sites. Your cooperation in helping him identify and obtain the necessary reports and data will be greatly appreciated.

Action items on step three:

1. Ben identifies remaining gaps in biophysical data and requests specific information from site teams.
2. Ben, Chris and Larry develop CLASSES prototype for October 2002 bioeconomic modeling course.

Data entry, cleaning, storage and access

It is best if data entry and initial cleaning (e.g., for impossible values) can be done promptly by the site teams. On-site data entry and checking allows the opportunity to revisit households as needed to correct any irregular entries or complete any missing values. [Please be sure to use a single, distinctive missing value code that cannot be confused with a proper entry, including of a true zero valued observation.]

Once complete data have been entered and given that initial, coarse cleaning, it will be time to begin data assembly and more systematic data cleaning, then move on to more meticulous data cleaning and preliminary descriptive analysis. We therefore need to come to tentative agreement on the division of responsibilities for data cleaning and preliminary descriptive analysis (simple tables of descriptive statistics and cross-tabulations). I will have three graduate students ready to do empirical work this summer, one taking the two northern Kenya sites (Andrew Mude), one taking the western and central Kenya sites (Heidi Hogset), and one on the two Madagascar sites (Marc Bellemare). They will be tasked with assisting the site teams with data cleaning, description, and preliminary analysis. We need to discuss how best to use these students to complement your site teams' work and not duplicate efforts unnecessarily.

We will also need a central repository to which all team members will have access, and we will need to establish a clear data access policy so as to balance the need to protect the investment some are making in data collection against the need to make data available to the broader team and community for good empirical work to inform policy and science. The subcontracts FOFIFA, ICRAF and KARI have signed with Cornell explicitly give both parties to the contracts (i.e., all four of our institutions) full rights in the data collected and questionnaires used in this project. But we need to agree on exactly what that means in terms of the baseline data and sharing these data with third parties (e.g., other researchers or students within our institutions, researchers at other institutions). These topics will be on the agenda at the June 2002 team meeting in Kakamega, Kenya. I would welcome your thoughts in advance of that meeting, if possible.

Other Issues

It would help enormously if we could establish a bibliography – or, better yet, a library – of documents written from the baseline data that we'll be using in each site, as well as any related studies of the site that you think might be germane (e.g., soils studies, ethnographies, etc.). I will assign Jason Frasco, a very talented undergraduate working with me, to assemble this bibliography and making it available on the project web site as soon as possible (web site is at http://www.aem.cornell.edu/special_programs/AFSNRM/Basis/index.htm). It would help enormously if you could send me and Jason (jbf26@cornell.edu) the

listing for your site (Bart, Jhon and Jean Claude: Fianarantsoa and Antsirabe; Willis: Madzu, Frank: Shinyalu and Embu, John: Marsabit and Baringo). If possible, please also send along any electronic copies you might have of the relevant documents. Thank you!

Our first project policy brief, which merely describes our project, is finally coming off the presses this week. I will get you each copies. Through Jhon, we are getting a French translation made for distribution in Madagascar in advance of our pre-study workshops in Fianarantsoa and Antananarivo next month. Our project proposal commits us to two further briefs this fiscal year (by end-September). I'm envisioning one will explore the theory of poverty traps and the relevant implications for development policy, based on the work Larry and I are doing now, supplemented by a few brief case study boxes to illustrate the idea. We should be able to generate at least one simple descriptive brief off the new data by September.

Conclusion

I apologize for subjecting you to an extremely long memo. But it seemed there were quite a number of things that we need to discuss in a somewhat holistic fashion. I look forward to an email dialogue on many of the above points over the coming few weeks and to visiting with the Kenya-based team next week and the Madagascar-based team the week following.

You have hopefully all blocked June 9-11, 2002, out on your calendars for our second annual team meeting, this one to be held in Kakamega, Kenya. Ben Okumu, Frank Place and Martins Odendo have already made many of the logistical arrangements. The two-day bioeconomic modeling short course will be held in Nairobi, probably immediately prior to the Kakamega meeting. Those who will be attending should plan their travel accordingly.

Thank you very much for all your hard work on this project. You should all know that due to your superb efforts thus far, Michael Carter is apparently talking up our particular project as the crown jewel of the BASIS CRSP. Partly as a result, Lena Heron, the program officer for the BASIS CRSP at USAID/Washington, will be accompanying me for the last couple of days in Kenya and the week in Madagascar. Keep up the great work!

Christopher B. Barrett
Associate Professor