

TECHNICAL PLAN for a DEMONSTRATION GLOBAL FORECASTING MODEL

by Paul Williamson

11-2-03; rev. 8-3-04, 10-19-04, 10-31-04, and 11-15-05.

1. Overview

- Conflict-economics coupling,
- Economic GDP growth model,
- Other couplings, elements—make easy to insert them later.

Key considerations include:

1. keep the initial model comparatively simple
2. insure ease of later addition and substitution of
 - other mechanisms (neural networks, fancy economic models, etc.), and
 - other factors (weather, etc.)
 - variable transformations [e.g. $\log(\text{GDP})$ in place of GDP]

... so that we have an effective research tool (rather than just one static model that will be obviously wrong the minute we run it).

The values of all constants and inputs are to be supplied from files external to the computational region of the program.

The focus will be on the conflict-economics coupling + economics model. The initial version (sketched in Parts 2 through 4, below) will be the essence of simple mindedness, though still programmatically complex.

By graphical user interface (GUI) I mean the user-program interface that appears on the monitor when one goes to www.globechange.org, then accesses the appropriate link. (That is, GUI means what the final user will see upon accessing and using the program. For example, in the MS Windows operating system, one "points and clicks" various icons to direct the actions of the computer.)

Further down the list of priorities is the ability to produce graphical displays--scatter plots, over-time plots, etc.--from outputs.

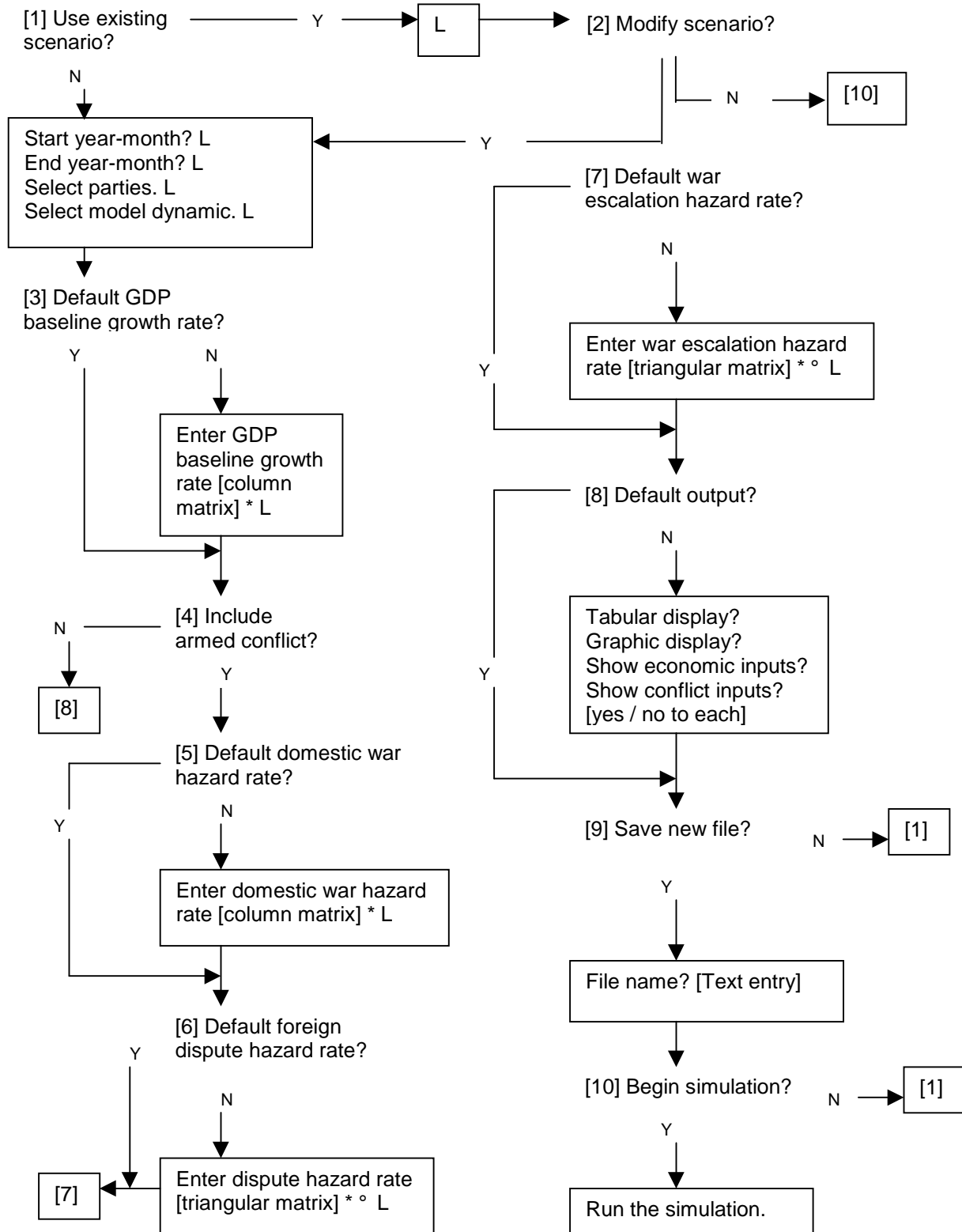
Objective is to write a program incorporating the formulae provided below, allowing the user to input the variable values (including input by selecting from data sets online) and view the resulting computations.

The "shell" means the program that gives overall direction to all the sub-programs. That raises two further questions: a. What are the sub-programs? b. What do I mean by overall direction? These are answered in part 2, below.

The desired functioning of this shell is summarized in Figure 1, on the following page:

FIGURE 1. FLOW DIAGRAM FOR DEMONSTRATION GLOBAL FORECASTING MODEL

* = alter selected entries or all entries.
 ° = named party versus each of all other parties.
 L = choose from list.



2. Technical description, general

a. Sub-programs:

The program of current interest, i.e. the program to estimate changes in GDP, is an example of what I mean by a sub-program. In what follows, I will use the letter f to denote the generic variable, of which we are seeking to estimate the changes. So the g of GDP, Parts 3 and 4 below, is a specific instance of the discussion of f that follows.

I should emphasize that the GDP sub-program is the only one we actually, for the moment, intend to build. The purpose in the structure that I am describing is simply to leave room for additional sub-programs later on.

Each sub-program will consist of an equation that computes a change, Δf , in the starting value of one dependent variable denoted by f ...

- starting at an initial referent time t ,
- over a time increment Δt to a new time period $t + \Delta t$,
- for a given set of parameter values,
- for a given set of starting values in all relevant input variables.

"Initial referent time" t means the value of time, in years, for which the starting input values are true. For example, it might be that $t = 2001$.

Δt means the period over which the change Δf is to be predicted. For example, it might be 1 year; in which case, if also $t = 2001$, then the new time period for which the prediction applies is 2002; and $\Delta t = 2002 - 2001 = 1$.

The parameter values are values of things that do not change during a particular use of the program. In the GDP sub-program, they are represented by the variously subscripted symbols a , b , c , and d mentioned below, beginning in equation (5). The starting values refer to inputs that can change in value during a single use of the model. In the GDP sub-program, they are represented by the variously subscripted symbols g , $x(k)$, and $y(k)$, where k is a positive integer, also beginning in equation (5).

In the above, the parameter values are to be inferred from some best-fit procedure based on empirical data, or conjured by some theoretical argument, or guessed. Of the variables, there are 3 source types: a) the empirically observed (historical) values seen in a particular starting year; b) estimated values for a particular year, based on a previous use of the sub-program to predict from prior values; and c) hypothetical values. For example, if the goal is to predict GDP in 2003, one might input the observed present values for 2001 (source type a), use them to predict the values for 2002, then use the latter values (source type b) to predict 2003.

Given unavoidable delays in data availability, when it comes to actual real-time predictions, lags of a year or more—for instance, predicting 2003 from 2001—are likely to be necessary.

The initial referent time, time increment, parameter values, and starting values will be input for each iteration of the sub-program; i.e. for each separate instance of computing an estimated change Δf in the dependent variable f .

As mentioned, the sub-program with which we are starting is the one for the change in GDP. The reference to "other couplings" means the other sub-programs later to be added but not in this cycle of development.

b. Overall direction:

The shell must do the following.

- Respond to user input regarding "run time" for the particular use of the program. Run time means the number of separate iterations of a sub-program required to get a particular forecast. If just one year's inputs are used to predict the next, then the run time is 1 year (or other time period). If the empirical observations or the predictions of one year are used as inputs to the next year, then the run time is however many such iterations are used. In the above example of predicting for 2003 from 2001 inputs, the run time would be 2 iterations.
- Respond to user input regarding referent initial time and time increment (for example, that referent initial time is 2001 and time increment is 1 year). Looking down the road, eventually we may want time increments as small as 1/10 year (or 1/12, to match the 12 months of a year); but for now it will likely be 1 year and, in some cases, 1 decade. (The 1/10 - or 1/12 - year increments will encounter the fact that many empirically observed data items are available at intervals no shorter than 1 year. For these, use of an interpolation method, such as LaGrange's polynomial or cubic splines, will be required. However, initially, these interpolation methods will be exogenous to the computational program.)
- Respond to user input regarding cases. Initially, the cases will be nations; later they will include other types of cases as well.
- Stage in proper order the separate uses of programs and, within uses, the separate iterations.
- For each iteration, input the applicable referent times initial, time increments, parameter values, and starting values. This involves one of 3 cases: As mentioned earlier, starting values are one of the types...
 - a) empirically observed;
 - b) saved estimates from prior iterations;
 - c) hypothetical.

In case a, the shell must read values from a previously constructed and loaded table, for the appropriate case and time period (initially, "year"). In case b, the shell must read the estimated values computed for the previous year or years. In case c, the inputs will reflect a scenario (again, read from a previously constructed and loaded table).

Later, when we insert additional sub-programs, the shell will need to be able to reach across them, using the observed and estimated starting values at time t from one sub-program for estimating outputs at time $t + \Delta t$ in another sub-program. [That requirement means that the region where empirically observed or hypothesized (type a or c starting values) and computed estimates (type b starting values) are stored must be "separate", in some sense, from the sub-program region.]

- Save all computations from each iteration of each use of each sub-program.
- Display output values, as requested by user.

3. Technical description, initial GDP change model

In what follows, first I will specify the mathematical form of the model; then I will explain what the model form means and otherwise comment. This will happen in several iterations, running into Part 4, between giving specifications, followed by comments, followed by additional specifications, etc.

Below, GDP means gross domestic product.

$$g_t \equiv \text{GDP accumulated during time period from moment } t - \Delta t \text{ to } t. \quad (1)$$

The variable t , defined below, will be called “time t ”. The default value of the time increment appearing in equation (1) will be $\Delta t = 1$ (year).

Definitions and notation:

The variable

t

denotes the referent time period. This is the most recent time period from which input data are to be drawn, on which to base the forecast of a particular future datum value.

Comments about referent time period:

1. At first, this period will be one year. (See Part 2, above.)
2. At first the referent time period will immediately precede the time period for which a value is to be forecast; that is, there will be a lead time of as little as 1 year between input and forecast data.
3. Combining the above two remarks, initially t will denote the last year from which inputs are taken, in order to forecast a datum value in the year $t + 1$. For example, if the year for which a forecast is to be made is 1990, then $t = 1989$.
4. One might ask to what moment, within the referent time period, do the data refer; e.g. if $t = 1989$, do we mean May 1st, or December 31st, or ...? First it should be noted that, as a practical matter, empirical data frequently are not very exactly dated. To take an example from another domain eventually to be incorporated, a particular estimate of military personnel strength may be vague as to exactly when, within the stated time period, it refers. This may apply also, to lesser degree, in the present instance of GDP. Often, one is interpolating, extrapolating, or guessing about such data; and there may be great inherent imprecision. To fix a standard, let us say that, where precision is appropriate, we mean the final moment within the stated period. So, for the time period of a year, we mean the final seconds of New Years Eve.

In what follows, we need to talk about the *order* of a change. First order refers to the change in a quantity over some time period. If x_t names the value of a quantity at time t , then this first order change is given by

$$\Delta x_{t+\Delta t} \equiv x_{t+\Delta t} - x_t. \text{ Second order refers to the change in the change, given by } \Delta^2 x_{t+\Delta t} \equiv \Delta x_{t+\Delta t} - \Delta x_t.$$

Third and higher order changes would be defined in a similar manner. With this terminology in mind, we identify the following first order changes:

$$\begin{aligned}
\Delta g_{t+1} &\equiv g_{t+1} - g_t, \\
\langle \Delta g \rangle_{t+1} &\equiv \text{a model estimate of } \Delta g_{t+1}, \\
r_{t+1} &\equiv \Delta g_{t+1} - \langle \Delta g \rangle_{t+1}.
\end{aligned}
\tag{2}$$

Put into words, the model estimate refers to the quantity that the program computes, as the estimated value of the change in the value of g from the year t to the year $t + 1$. The quantity Δg_{t+1} is the change in the value of g (i.e. of GDP) from time period t to time period $t + 1$.

This means, to get Δg_{t+1} put the value of g at time $t + 1$ into "slot 1", put the value of g at the prior time t into "slot 2", and subtract the value in slot 2 from that in slot 1. (Again, the default time increment is 1 year.) The quantity r_{t+1} is the error (residual) between the true and estimated values of Δg_{t+1} .

Second order changes are given by:

$$\begin{aligned}
\Delta^2 g_{t+1} &\equiv \Delta g_{t+1} - \Delta g_t, \\
\langle \Delta^2 g \rangle_{t+1} &\equiv \text{a model estimate of } \Delta^2 g_{t+1}, \\
r_{2,t+1} &\equiv \Delta^2 g_{t+1} - \langle \Delta^2 g \rangle_{t+1}.
\end{aligned}
\tag{3}$$

The quantity $\Delta^2 g_{t+1}$ is the second order GDP change (change in GDP change) from time period t to time period $t + 1$. The quantity $r_{2,t+1}$ is the error (residual) between the true and estimated values of $\Delta^2 g_{t+1}$.

In other words, the output (dependent) variable is the annual change in the change: The change in g in time period $t+1$ from time period t is computed, also the same for time period t compared with time period $t-1$; then the difference in these two changes is computed. This latter is the output variable to be replicated by a linear model, a neural network, or other computational scheme.

One uses the 2nd order ("change-of-change") values to estimate the 1st order changes themselves, by adding these estimates to the previously observed [time period t] - [time period $t-1$] change values. The science of it is then to check the empirical fit between estimate and observed values, where the discrepancy between the two of them is given by $r_{2,t+1}$.

Inputs:

$$\begin{aligned}
g_{t-n} &\equiv \text{previous value of GDP of referent party, for time period } t - n, n = 0,1,2,\dots; \\
x_{t-n}(i) &\equiv \text{domestic economic factor } i, i = 1,2,\dots, \text{ for time period } t - n, n = 0,1,2,\dots; \\
y_{t-n}(j) &\equiv 1 \text{ if conflict factor of type } j, j = 1,2,\dots, \text{ occurred during time period } t - n, n = 0,1,2,\dots; \\
&\equiv 0 \text{ otherwise.}
\end{aligned}
\tag{4}$$

The quantity $x_t(i)$ has one of two possible references: As with GDP, discussed above, eq.(1), for accumulating values such as production of some economic good, $x_t(i)$ is the accumulation for the period beginning at the moment $t - \Delta t$ and ending at the moment t ; for instantaneous values such as unemployment rate, this quantity is the value at the moment t . (Again, the default is $\Delta t = 1$.) The quantity $y_{t-n}(j)$ of conflict factor j in lag time period n , reflects the intent to consider presence or absence of conflict, in periods preceding the referent period, as a factor influencing changes in GDP.

There are two linear model forms:

$$\begin{aligned} f_{lin1,t+1} &\equiv a_0[g_t] + a_1[g_{t-1}] + a_2[g_{t-2}] + \dots \\ &+ b_{10}[x_t(1)] + b_{11}[x_{t-1}(1)] + b_{12}[x_{t-2}(1)] + \dots + b_{20}[x_t(2)] \\ &+ c_{10}[y_t(1)] + c_{11}[y_{t-1}(1)] + c_{12}[y_{t-2}(1)] + \dots + c_{20}[y_t(2)] + \dots + d \end{aligned} \quad (5)$$

(in words, this equation is saying is that the quantity being estimated, $f_{lin1,t+1}$, is a weighted sum of all the quantities in brackets []); and

$$\begin{aligned} f_{lin2,t+1} &\equiv a_{01}[\Delta g_t] + a_{11}[\Delta g_{t-1}] + a_{21}[\Delta g_{t-2}] + \dots \\ &+ a_{02}[g_t] + a_{12}[g_{t-1}] + a_{22}[g_{t-2}] + \dots \\ &+ b_{10}[x_t(1)] + b_{11}[x_{t-1}(1)] + b_{12}[x_{t-2}(1)] + \dots + b_{20}[x_t(2)] \\ &+ c_{10}[y_t(1)] + c_{11}[y_{t-1}(1)] + c_{12}[y_{t-2}(1)] + \dots + c_{20}[y_t(2)] + \dots + d \end{aligned} \quad (6)$$

where a_n , a_{n1} , a_{n2} , b_{in} , c_{jn} and d (with $i = 1,2,3,\dots,I$, $j = 1,2,3,\dots,J$, and $n = 0,1,2,3,\dots,N$), are constants to be determined empirically. (Note the ranges of variation in i , j and n match the corresponding indices in eqs. (4), above.) The quantities $f_{lin1,t+1}$ and $f_{lin2,t+1}$ correspond to forecasting approaches based on the first- and second- order change quantities defined above, equations (2) and equations (3), respectively.

Artificial neural network (non-linear) forms are:

$$\begin{aligned} f_{mer1,t+1} &\equiv f\{g_t, g_{t-1}, g_{t-2}, \dots \\ &x_t(1), x_{t-1}(1), x_{t-2}(1), \dots x_t(2), x_{t-1}(2), x_{t-2}(2), \dots \\ &y_t(1), y_{t-1}(1), y_{t-2}(1), \dots y_t(2), y_{t-1}(2), y_{t-2}(2), \dots d_1, d_2, d_3, \dots\} \end{aligned} \quad (7)$$

and

$$\begin{aligned} f_{mer2,t+1} &\equiv f\{\Delta g_t, \Delta g_{t-1}, \Delta g_{t-2}, \dots g_t, g_{t-1}, g_{t-2}, \dots \\ &x_t(1), x_{t-1}(1), x_{t-2}(1), \dots x_t(2), x_{t-1}(2), x_{t-2}(2), \dots \\ &y_t(1), y_{t-1}(1), y_{t-2}(1), \dots y_t(2), y_{t-1}(2), y_{t-2}(2), \dots d_1, d_2, d_3, \dots\} \end{aligned} \quad (8)$$

where $d_{m,q,p}$, $m = 0,1,2,3,\dots,M$ and $q, p = 1,2,3,\dots,N, P$, respectively, are a set of constants (the "node weights" of the network). In this scheme, m refers to the layer of the neural network, with $m = 0$ the output layer, $m = 1$ the first hidden layer ("first" meaning the final layer encountered before data reach the

output layer), $m = 2$ the second hidden layer, and so on. For now, the neural networks we use will have the output and just one hidden layer. Further, q refers to the q^{th} node, and p refers to the p^{th} input, of the given layer.

In parallel to equations (5) and (6), $f_{met1,t+1}$ and $f_{met2,t+1}$, respectively correspond to forecasting approaches based on the first- and second- order change quantities.

Four models may then be defined as:

model 0101-L

$$\begin{aligned} g_{t+1} &= g_t + \langle \Delta g \rangle_{t+1} + r_{t+1}, \\ \langle \Delta g \rangle_{t+1} &= f_{lin1,t+1}; \end{aligned} \tag{9}$$

model 0102-L

$$\begin{aligned} g_{t+1} &= g_t + \Delta g_t + \langle \Delta^2 g \rangle_{t+1} + r_{2,t}, \\ \langle \Delta^2 g \rangle_{t+1} &= f_{lin2,t+1} \end{aligned} \tag{10}$$

model 0103-N

$$\begin{aligned} g_{t+1} &= g_t + \langle \Delta g \rangle_{t+1} + r_{t+1}, \\ \langle \Delta g \rangle_{t+1} &= f_{met1,t+1}; \end{aligned} \tag{11}$$

and model 0104-N:

$$\begin{aligned} g_{t+1} &= g_t + \Delta g_t + \langle \Delta^2 g \rangle_{t+1} + r_{2,t}, \\ \langle \Delta^2 g \rangle_{t+1} &= f_{lnet2,t+1}. \end{aligned} \tag{12}$$

In words model 0101-L, for example, is saying that the value of g in the year $t + 1$ will equal the sum of three terms: the value of g in the previous year t , the model estimate of the change in g from year t to year $t + 1$, and the residual (error) in the estimate. The second term (the model estimate) is what the computer program will compute, the first term will come from an input table or have been computed in a previous step. The third term is uncontrollable and constitutes the part (namely error) that we are seeking to minimize. Similar remarks apply to the other three models.

4. Preliminary *.xls mockup, initial GDP change model

The Excel mockup, example_02_041016_3-nations.xls emulates model 0102-L, equations (10), the second of which, combined with eq. (6), becomes

$$\begin{aligned} \langle \Delta^2 g \rangle_{t+1} &= f_{lin2,t+1} \\ &= a_{01}[\Delta g_t] + a_{11}[\Delta g_{t-1}] + a_{21}[\Delta g_{t-2}] + \dots \\ &\quad + a_{02}[g_t] + a_{12}[g_{t-1}] + a_{22}[g_{t-2}] + \dots \\ &\quad + b_{10}[x_t(1)] + b_{11}[x_{t-1}(1)] + b_{12}[x_{t-2}(1)] + \dots + b_{20}[x_t(2)] \\ &\quad + c_{10}[y_t(1)] + c_{11}[y_{t-1}(1)] + c_{12}[y_{t-2}(1)] + \dots + c_{20}[y_t(2)] + \dots + d, \end{aligned}$$

which is the general scheme that is used in the mockup version of 0102-L, for estimating the second-order change, $\langle \Delta^2 g \rangle_{t+1}$, in the logarithm of GDP. [Note this Excel file is not online but a copy will be provided on request. Contact Paul Williamson at paulrw@globechange.org .] When this scheme is further developed by limiting it to the coefficients and input variables actually used, the above equation becomes

$$\begin{aligned} \langle \Delta^2 g \rangle_{t+1} &= f_{lin2,t+1} \\ &= a_{01}[\Delta g_t] + a_{11}[\Delta g_{t-1}] \\ &+ a_{02}[g_t] + a_{12}[g_{t-1}] + a_{22}[g_{t-2}] \\ &+ b_{10}[x_t(1)] + b_{11}[x_{t-1}(1)] + b_{12}[x_{t-2}(1)] \\ &+ b_{20}[x_t(1)] + b_{21}[x_{t-1}(1)] + b_{22}[x_{t-2}(1)] \\ &+ b_{30}[x_t(1)] + b_{31}[x_{t-1}(1)] + b_{32}[x_{t-2}(1)] \\ &+ c_{10}[y_t(1)] + c_{11}[y_{t-1}(1)] + c_{12}[y_{t-2}(1)] \\ &+ d . \end{aligned} \tag{13}$$

The corresponding referent year and the non-vanishing constants and variables are shown in Table 1:

TABLE. NOTATION FOR COEFFICIENTS AND VARIABLES IN EXCEL MOCKUP

[Note, again, this Excel file is not online, but a copy will be provided on request. Contact Paul Williamson at paulrw@globechange.org .]

Notation, this document		Notation, example_02_041016_3-nations.xls	
coeff	variable	var_name	var_symbol
[none]	t	referent year = 2000	t
a_{01}	Δg_t	dlt gdp(2000, 1999)	dlt-g_t
a_{11}	Δg_{t-1}	dlt gdp(1999, 1998)	dlt-g_t-1
a_{02}	g_t	gdp, referent year ($t = 2000$)	g_t
a_{12}	g_{t-1}	gdp, referent year - 1 (= 1999)	g_t-1
a_{22}	g_{t-2}	gdp, referent year - 2 (= 1998)	g_t-2
b_{10}	$x_t(1)$	baseline gdp growth rate, referent year	$x_t(1)$
b_{11}	$x_{t-1}(1)$	baseline gdp growth rate, referent year - 1	$x_t-1(1)$
b_{12}	$x_{t-2}(1)$	baseline gdp growth rate, referent year - 2	$x_t-2(1)$
b_{20}	$x_t(2)$	price growth rate, referent year	$x_t(2)$
b_{21}	$x_{t-1}(2)$	price growth rate, referent year - 1	$x_t-1(2)$
b_{22}	$x_{t-2}(2)$	price growth rate, referent year - 2	$x_t-2(2)$
b_{30}	$x_t(3)$	population growth rate, referent year	$x_t(3)$
b_{31}	$x_{t-1}(3)$	population growth rate, referent year - 1	$x_t-1(3)$
b_{32}	$x_{t-2}(3)$	population growth rate, referent year - 2	$x_t-2(3)$
c_{10}	$y_t(1)$	conflict factor, referent year	$y_t(1)$
c_{11}	$y_{t-1}(1)$	conflict factor, referent year - 1	$y_t-1(1)$
c_{12}	$y_{t-2}(1)$	conflict factor, referent year - 2	$y_t-2(1)$
[none]	d	dependent variable intercept term	D

Note: A piece of equation (13) [which, in turn, comes from (6) and (10)] is $f_{lin2,t+1} = a_{01}[\Delta g_t]$. For this case, in which the output $f = \text{GDP } 2^{\text{nd}} \text{ order change}$, this particular piece is explored in `tbl1002_dd_lg-v-d_lg-1.xls`. The summary result, Pearson $r \cong -0.58$ is reasonable encouragement to proceed by seeking a linear model of the remaining residuals based on (13). In fact, this is the reason why model 0102-L, equation (10), has been chosen for this initial realization of the GDP change model.