General Description of the CLIGEN Model and its History

Notes by Chuck Meyer
USDA-ARS National Soil Erosion Laboratory
West Lafayette, IN

Cligen is a stochastic weather generator which produces daily estimates of precipitation, temperature, dewpoint, wind, and solar radiation for a single geographic point, using monthly parameters (means, SD's, skewness, etc.) derived from the historic measurements. Unlike other climate generators, it produces individual storm parameter estimates, including time to peak, peak intensity, and storm duration, which are required to run the WEPP and the WEPS soil erosion models. Station parameter files to run Cligen for several thousand U. S. sites are available for download from this website: also data and software to build station files for international sites. With the exception of Tmin, Tmax, and Tdew temperatures (changed in January 2004), daily estimates for each parameter are generated ***independently*** of the others. With the current random number generator, subsequent runs on the same machine made with identical inputs will produce identical results.

Users of daily simulation models should consider the impacts of Cligen's characteristics on their application. Individual parameter distributions may be expected to reproduce monthly historic distributions quite well. However, if the model in question is sensitive to the daily interactions of two or more of the parameters Cligen produces, Cligen may not be the most appropriate weather generator to use. This is because for a given day, it generates solar radiation, and maximum and minimum temperatures completely independently from precipitation. Experience and common sense tell us that these parameters are NOT independent. In practice this may not be a huge issue, since it is not uncommon for models to be sensitive to one weather parameter on a daily basis, and relatively insensitive to the others, as long as their monthly trends are preserved.

Cligen was produced by Arlin Nicks and Gene Gander at the USDA Agricultural Research Service (ARS) lab in Durant, Oklahoma. They made their last significant changes to the model in the mid 1990's before Dr. Nicks' retirement and death in July 1997. David Hall and Dayna Scheele at the USDA Forest Service lab in Moscow, Idaho, acquired code from Dr. Nicks' computer to generate from historical weather data, the monthly station parameter files required to drive Cligen. They mechanically cleaned the existing station data files, and added a large number of new stations for the United States. In the Summer of 1999 Bofu Yu of Griffith University in Australia, expressed concerns that Cligen's rainfall intensity calculations were not operating correctly due to a unit conversion error, and offered revised code to correct the problem. In the Fall of 1999, Charles R. Meyer at the USDA-ARS lab in West Lafayette, Indiana, re-coded the CLIGEN source code to make it understandable, maintainable, and extensible by people who had not written it. He incorporated Dr. Yu's corrections. When performing checks on CLIGEN's *uniform random number generator* and *standard normal generator*, Meyer discovered that they were not operating correctly. A paper by Johnson, et. al. referenced below, reports effects which appear to be the result of this problem. This has major implications

for any stochastic model like CLIGEN because all its output originates from its random number generator. To correct this problem Meyer introduced a form of "quality control" borrowed from industrial engineering.

A few range checks are included in CLIGEN. While these will slightly alter the distributions produced, that may be acceptable to gain daily values that make more sense both individually, and relative to each other. For example, daily precipitation is not allowed to be negative; and minimum temperature is forced to be less than maximum for each day. Solar radiation also has prescribed limits that it must operate within. With one exception there does not appear to be anything within the Cligen code that operates as a Markov chain, as described by Johnson, et. al. That exception is use of the monthly probability of a "wet" day following a dry day, and a "wet" day following a wet day. Anecdotal evidence from chatting with others who knew Arlin Nicks suggests that the approach was included for a short time and subsequently removed due to the difficulty of adequately implementing all the interactions between parameters.

Cligen is implemented in Fortran-77. It has been modified to operate in three modes:

- Pure Interactive Mode.
- Pure Commandline Mode.
- Mixed Mode, where the user is queried for information not entered on the commandline.

(Note that successive runs made in Interactive Mode, without re-starting CLIGEN will *not* be identical. This is because the seeds for the random number generator are not re-initialized.)

Neither the ANSI standard for F-77 nor F-90 includes intrinsic routines for parsing the command line. Although most compilers include them as extensions, the actual routines vary from one compiler to the next. Thus, having the ability to enter options on the commandline also makes the code specific to the compiler used. Cligen has been modified to include the code for several popular compilers, and the user can simply enable the line in the source code appropriate to the compiler used. For command line parsing support for a more complete set of compilers, visit the Fortran-2k (F2KCLI) project.

Available from this Website:

- A slide show explaining how the basics of CLIGEN work.
- Several thousand U. S. climate station files produced by the U. S. Forest Service. Using a station parameter file, runs of any desired length can be made.
- The ***GenstPar*** and ***FindMatch*** software make it easy to generate an international Cligen station parameter file, using a matching U. S. station parameter file as a surrogate.

- Both the Win-95/98/NT/XP/2000 **executable** (cligen.exe) and the **source code** files for:
    1. **recoded CLIGEN V-4.2** verified to give results identical to the original [folder v4.2_src (Recoded)].
    2. **CLIGEN 5.22564**, which now includes:
        - values for Tmax, Tmin, and Tdew that are correlated on a day by day basis,
        - improved K-S quality-control for uniform random distributions,
        - quality-control (CI for mean & SD) for normal random distributions,
        - corrected Ipeak, Tpeak, and storm duration code,
        - limits on precipitation skew coefficient,
        - three types of interpolations.
        - command line options
        - supports concatenated files of Forest Service station data,
        - filenames can include spaces & slashes, and can be up to 256 characters,
        - hints for configuration and for international use,
        - numerous enhancements for "option-6" runs,
        - saves command line in output file,
        - cosmetic changes,
        - minor bug fixes.

    NOTE: to run from the command line, specify "-t5" for the typical WEPP/WEPS output.

    The CLIGEN (and GenStPar & FindMatch) executable must be run in Windows from a DOS window. If you are downloading the source code, be sure to get a current copy of all the required files, including the "include" (*.inc) files. The public domain ACM code to perform Chi-square tests on the standard deviations is appended to the end of the cligen.f source code file. Using the copyrighted Numerical Recipes equivalent requires using the binary file ms_chi-sq.o or solaris_chi-sq.o, depending upon your operating system. Because of licensing restrictions these cannot legally be distributed as source code.

- The GNU g77 Fortran compiler and gcc c/c++ compiler for Win-95/98/NT by Mumit Khan that produces native code utilizing the Microsoft DLL's, is available for download free of charge here. It is a very good compiler, which can link Fortran and 'C' modules to create an executable **from a single command** like UNIX compilers. (See note in the top of CLIGEN.F source code for command line required.) When used to compile CLIGEN, it gave output identical to both the Lahey commercial compiler on the PC, and the Solaris compiler on a UNIX workstation.
- Description of the Cligen_station_parameters
- Manuscripts and references related to CLIGEN.
- Source code for the various random number generators and standard normal deviate generators which were tested for possible incorporation into CLIGEN.

- Source code for the "conf_lim" subroutine employed in statistical and mathematical analyses of CLIGEN using confidence limits on the monthly means, and source code for the entire programs using it to test the outputs and random deviates.
- Some example 30 year runs using the interpolation options.

Purposes and Results of Recoding CLIGEN:

NOTE: There now are several mathematically ingenious features incorporated in Cligen:

1. Quality control on the random number generator -- Totally necessary for models depending on a random number generator, but to date not known to be provided in any other stochastic model.
2. The generic random number generation scheme used in Cligen for the Gamma distribution can be employed for any function that can be utilized between x and f(x) max-min limits.
3. Use of monthly distributions of Tmin, Tmax, and Tdew to generate ***correlated*** daily values of these parameters, which reproduce the monthly means and SD's they were generated from. Drastically reduces the need to enforce range checks.

What has Changed:

- CLIGEN V-4.2 was **recoded** to conform to the "Water Erosion Prediction Project (WEPP) Fortran-77 Coding Convention". The work was performed by Charles R. Meyer, NSERL-USDA-ARS, from August to November 1999, to facilitate the maintenance, modification, and extension of the CLIGEN code until a suitable replacement for CLIGEN is obtained. The extensive help of David Hall, USFS, Moscow, Idaho in providing many valuable definitions is gratefully acknowledged. C. R. Meyer may be contacted at: 275 South Russell Street, Purdue University, West Lafayette, IN 47907; ph: 765/494-8695; e-mail: wepp@ecn.purdue.edu.
  - The code structure was radically simplified.
  - Inline definitions were added for virtually every variable.
  - Explanatory comments were added to the code where the daily climate outputs are generated, to facilitate understanding, evaluation, and subsequent modification.
  - To correct an arithmetic underflow function DSTG generates in long runs, the variables "xx" and "fu" are declared "double precision".
  - Code that appears wrong or suspicious was commented with the string "XXX". (Meyer is not a climate/weather expert, so it is not obvious to him what should be done -- only that the code appears to be incorrect.) The user is left to his/her own interpretation.

    - The recoded code was tested to verify that it gave results identical to the original V-4.2 code from which it was derived.
- **Corrections to peak intensity calculations** proposed by Bofu Yu, were added to the code, primarily in subroutines R5MON and ALPH (which became R5MONB and ALPHB) to make peak intensity responsive to differences in latitude as originally intended. The new code has been verified to give results identical to Bofu's.
- **Corrected and extended data for individual U. S. stations provided by the U. S. Forest Service, Moscow, Idaho.** (Many derived from Arlin Nicks' Data.) Use this in preference to the old V-4.2 data.
- Because CLIGEN uses single parameters to derive daily values for the entire month, three schemes of **interpolation/disaggregation** were coded to provide more continuous daily values between the monthly ones. These are:
    1. simple linear interpolation.
    2. Fourier series.
    3. a modified linear interpolation which preserves the mean value of the parameter for the month.

  The default is "no interpolation". When interpolation is requested, all the monthly parameters (mean, SD, skewness, ...) are interpolated for each day of the run, according to the scheme selected.

  The paper by Johnson, et. al. suggests that interpolation may be of particular interest "during transitional months when large changes are noted in short time periods," like the beginning and end of growing seasons. Because maximum solar radiation can be so well predicted using only latitude and day of the year, this may also be of interest for simulations that are sensitive to changes in solar radiation.

  Note that the mathematical resolution of the station data is "monthly". Any extrapolation of the data to a finer timescale involves assumptions. When using interpolation the user must decide which assumptions (and associated tradeoffs) result in the lowest level of discomfort.

- To facilitate use of CLIGEN by user interfaces, the following **command line options** were added:
  -S state number
  -s station number
  -i input filename
  -o output filename
  -b beginning year
  -y years duration
  -r random seed
  -I0 no interpolation (default)
  -I1 linear interpolation
  -I2 Fourier interpolation
  -I3 interpolation to preserve the monthly means
  -?, -h help

-F force overwrite of existing output file
-O observed data file (option-6)

- Using confidence interval tests on the mean, C. R. Meyer determined that the numbers generated by Cligen did not correspond satisfactorily to the numbers they were supposedly generated from: the historic means, standard deviations, and skewness. Further, Cligen imposes range checks which alter the resulting distributions. However, once those are accounted for, confidence tests on the mean show that Cligen's uniform random number generator (RANDN) and standard normal deviate generator (DSTN1) do a poorer job of producing a set of numbers approximating a standard normal distribution (mean = 0.0, SD = 1.0) than some other routines that could be substituted for them. However, no combination tested worked well enough to be satisfactory. ***In general terms, Cligen was not receiving sets of random uniform and standard normal inputs of satisfactory quality, from which to generate its weather outputs.*** Johnson, et. al. (*Stochastic Weather Simulation: Overview and Analysis of Two Commonly Used Models*; Journal of Applied Meteorology, Vol. 35, October 1996) noted that while CLIGEN did an acceptable job of reproducing the historical overall means and the extreme values of the parameters studied, it did a very poor job of reproducing the variability in the historical means from one year to the next, and in reproducing their standard deviations by month. They commented that "CLIGEN failed the F test for equality of varience of mean monthly temperatures in all 72 tests."
- Numerical Recipes in Fortran points out in Ch 7.1, p 267, that "... a reliable source of random uniform deviates, ... is an essential building block for any sort of stochastic modeling...." A few pages later they refer to Park and Miller's "anecdotal sampling of a number of inadequate generators that have come into widespread use," and conclude, "The historical record is nothing, if not appalling." CLIGEN has been altered to ensure that the random numbers its outputs are based upon, are of satisfactory quality.

The approach which proved successful, was as follows. C. R. Meyer returned to use of the original RANDN and DSTN1 routines; however, he introduced a new routine (RANSET) to initialize the random values for each parameter one month at a time. (See discussion on Random Number Generators later in this document.) RANSET includes a ***data quality inspection*** feature which subjects uniform random distributions to a Kolmogorov-Smirnov "goodness of fit" test to ensure the appropriate frequency of random numbers in each of 20 size classes. It also ensures that for normal distributions both the mean and the standard deviation of the numbers generated thus far are within the confidence limits at a specified level (threshold), for example, 50 percent. (The standard normal deviates are supposed to approximate a standard normal distribution, and any two standard normal distributions are by definition "equivalent" if they have the same mean and standard deviation (or variance).) The numbers tested are those produced for the current parameter, for the current month, since the start of the run. Note there are a couple of caveats and potential drawbacks to this approach:

- If the threshold is set too "tight" (low) Cligen may never be able to converge to a solution.
- The numbers "measured" to ensure acceptable quality are the ones going into Cligen, not the ones coming out. It is still possible that the range checks imposed, the climatic equations used, or some other factor will result in unsatisfactory outputs being generated.
- The testing previously done on Cligen to determine how well it reproduces extreme events is no longer valid for the new code. That work needs to be re-done.
- Cligen is constrained from utilizing monthly sets of standard normal deviates that exceed its threshold. As the run progresses, two offsetting forces are at work: the limits on the mean get "tighter", but the "mathematical inertia" of the system gets larger, i. e., there are many more numbers in the "pool of numbers" so the impact of any individual number becomes less significant. This suggests that a really "deviant" number might be rejected early in the run, but accepted later. Exactly how this impacts the distribution in time of events produced is anybody's guess, and needs further investigation.

Version History:

V-5.101 includes code to correctly handle three consecutive identical average monthly values -- a situation that caused arithmetic errors in V-5.1.
V-5.102 includes several minor corrections ensuring that when a month's data is rejected, the system is returned to exactly its previous initial state before new data is generated. A typographical error in the standard normal deviate generator's 10-standard-deviation range check is also corrected.
V-5.103 has some cosmetic changes which do not change the output at all. The most notable is to correct for an extra space incorrectly added at the beginning of each line printed in a DOS window under Win-95. That double-spaced the CLIGEN stations and the first half scrolled off the screen so they could not be viewed.
V-5.104 produces its outputs from the exact same inputs as the original 4.2, when run in interactive mode.
V-5.105 provides a guaranteed "break out" after 10,000 iterations and a warning message on the screen if Cligen gets caught in its quality control loop.
V-5.106 allows input and output filenames in the commandline to include blanks, slashes, etc. (They must be quoted.)
V-5.107 reads station parameter files built by concatenating Forest Service individual station files. No special flags are needed.
V-5.108 corrects problems with temperature dewpoint when executing option 6; ie, command line option "-t6".
V-5.109 permits input and output filenames of up to 256 characters, instead of the 60 previously allowed. Includes some cosmetic changes to initial screen.
V-5.110 Records any command line parameters by appending them to the fifth line of the output file. Eliminates infinite loop caused by specifying an output file that already exists.
V-5.111 Except for embedded hints for configuration, and hints for international users, all changes made in this version pretain to "option 6" which uses user-supplied temperature and precip data from

one location, in conjunction with existing station parameters for another location having a similar climate. Option-6 runs are now supported from the command line. This requires using a new switch, capital-O, followed by the name of the file containing the observed data. (See option-6 example for format.) Years in the output file may now have more than 2 digits. Length of run may be specified from the command line, although 100 years is still the default. The user may now specify the beginning year from command line. If unspecified, the first year listed in observed data file is used.
V-5.200 Includes limits on precipitation skewness coefficient to keep Cligen from generating an inordinate percentage of negative values.
V-5.211 to V-5.213 Improvements to code affecting the individual storm characteristics: Tpeak, Ipeak, and storm duration. Annotation within the code of a method to generate any target distribution with a fixed x-min, x-max, f(x)-min, and f(x)-max -- as long as the PDF f(x) can be described with a function (ie, for every value of x, there is a unique value for f(x)). This is used in CLIGEN to generate a Gamma distribution used for Ipeak.
V-5.220 & V-5.221 A much more stringent form of quality control was incorporated. This should significantly improve the distributions of randomly generated variables that are based on the uniform distribution, as well as the standard normal.
V-5.222 Corrects a bug in the storm duration calculation which was in versions 5.213 through 5.221.
V-5.223 Corrects a bug in the storm intensity algorithm that was in all prior versions, and adds quality control to the gamma distribution used to generate storm intensity.
V-5.2253 Generates Tmax, Tmin, and Tdew so they are related for every day, but also the monthly distributions of each are preserved. No changes to the Cligen inputs are required.
V-5.2254 Bug fix. Previous version entered an endless loop if it could not fit the gamma distribution in 10,000 tries.
V-5.2255 Chi-square quality-control replaced by Kolmogorov-Smirnov (K-S) which reproduces the monthly distribution accurately, in addition to the mean and standard deviation.
V-5.2256 Minor change to increase lot size of numbers generated for the Gamma distribution (rainfall intensity) from 20 at a time, to 30 at a time.
V-5.22561 Minor change to correct typographical error in subroutine RYF1 which was causing a *subscript out of range* error. Many thanks to Fred Fox and Larry Wagner of USDA-ARS-WERU for their persistence and patience in pointing out that error.
V-5.22562 Very minor change. In file command5.inc variable av_len was declared to be type integer, and command5.inc was renamed command6.inc.
V-5.22563 Very minor change. Added explicit initialization of a variable used to open a file in SR USR_OPT.
V-5.22564 Change applies only to very arid climates. Calculations for locations with less than 1/2 day of precipitation per month were "blowing up". Code was changed to make them behave the same as the other months with less than 2.218 raindays per month.

Running Cligen outside the USA:

Cligen and its station parameter files were developed at the expense of the U. S. taxpayers for use by government agencies. Consequently "ready to use" station files developed are for primarily U. S. stations. However, people outside the U. S. are welcome and encouraged to use and adapt the station data (and

Cligen itself) to their use. (The Cligen source code and the free compiler used to produce the MS-Windows executable available here, are provided on this website for download. Note that I will only provide support for changed code if I incorporate it into the source code "base" available on this site.)

The U. S. has a very diverse climate, so there should be something in our several thousand station files, that is reasonably similar to just about any site on earth. These are plain-text (ASCII) files which can be modified with a text editor. They allow the user to run Cligen for any number of years desired, producing a series of daily weather values with the same statistical distribution as those observed at the site. *Survey of Climatology* by Griffiths and Driscoll (1982, Merrill Pub.) categorizes climates of the world. If the user locates their climate on the Survey map, and then finds a matching U. S. station to use as a surrogate (substitute to supply the missing data values), it's climate file can be adapted to the monthly values (Mean, SD, skew) of daily precipitation, temperature, etc. observed at the new site for its period of record. The equations used for SD and skew appear below on this webpage. If you use daily data, a spreadsheet like Excel can be employed to calculate the means, standard deviations, and skew coefficients, using its built-in functions (AVERAGE, STDEV (or STDEVP), and SKEW).

NOTE: As of September 19, 2003, software is available to **mechanically create a Cligen station file** from the online GDS data (international daily temperature and precipitation data). It employs two programs: "GenStPar" which builds the top of a station (.par) file and writes it to a (.top) file; and "FindMatch" which searches the several thousand USFS station files, lists the ten best matches, and builds a Cligen station (.par) file using the best match, supplying the missing values from the surrogate (substitute) U. S. station. The current matching algorithm uses a least squares statistic which weights monthly P(W|W) at 49%, monthly P(W|D) at 49%, elevation at 1%, and latitude at 1%. Source code is available for download and modification. (Note that the same rules relative to support apply.) In MS-Windows GenStPar and FindMatch must be run from the command line within a DOS window.

Using your daily data to run Cligen:

Here are some hints for using your own daily data to build a Cligen station (.par) file.

The format of the GDS data file is:

First line (station identifiers):
(i2,i3,a46,i3,i2,2x,i3,i2,i6)
Variables in each field:
Station ID (2 digit integer)
Country ID (3 digit integer)
Station Name (46 ASCII characters)
Degrees Latitude (3 digit integer)
Minutes Latitude (2 digit integer)
skip 2 spaces
Degrees Longitude (3 digit integer)
Minutes Longitude (2 digit integer)
Elevation (m) (6 digit integer)

Following Lines (daily data):
(3i2,2(f5.1,2x),f5.1)
Year (2 digit integer)
Month (2 digit integer)
Day (2 digit integer)
Maximum Temperature (0.1 degree C) (5 digit floating point)
skip 2 spaces
Minimum Temperature (0.1 degree C) (5 digit floating point)
skip 2 spaces
Precipitation (0.1 mm) (5 digit floating point)

Note:
- The value -999 denotes missing data.
- When using a format like 'f5.1' to read a number, a field of 5 spaces is read. If a decimal exists in the input, it is used.
However, if none is present the format inserts one, placing one digit to the right of the decimal. For example, '1.2345' is read as '1.2345' but '12345' is read as '1234.5' Blanks within the 5-space field are read as zeroes; ie, ' 5 ' is read as '0050.0'.

Your data will have to be in the GDS format above, before the program GenStPar can process it. Decimals may be inserted where appropriate; otherwise, GenStPar will assume they exist as indicated by the input format listed above. Decimals must not appear in integer fields, like dates. It will probably be easiest to delay building the first line until the others are "built". Your data needs to be listed with one day per record (line).

You can manipulate data into the GDS format above using a spreadsheet like Excel. Read the data into the spreadsheet dividing the columns so that that minimum temperature is in one column, maximum temperature is in another, precipitation in another, etc. Now move the columns into the proper order, as indicated by the GDS format. You can insert blank columns and copy data columns into them, then delete the un-needed original column of data. You will want to output your file in ASCII (plain text) format. I used a "space delimited" ".prn" option under "File>Save As". Before you actually save the file, highlight each column and go into "Format>Column>Width". Set each column as "5" or whatever is needed in the GDS format. (To skip 2 spaces, make the column to the right, two spaces larger.) To align the numbers with the right margin, select the column header; click on "Format>Cells"; click on the "Alignment" tab; and select "Right" from the "Horizontal" pull-down. After saving the the data file, open it using a text editor, and add the first line (station (site) information).

If you develop station parameter files for locations outside the U. S. please consider sending them to us so that they may posted on the website to share with others who may be interested. The file should include the correct name of the station and its elevation, latitude, and longitude.

Approach used in Recoding CLIGEN, and Results:

The McCabe Tools were used to visualize the logic, and to measure improvements made in the code structure. Code complexity was reduced by isolating sections of code into eight additional modules. Common blocks were moved to "include" files, resulting in a single copy of each. Unused variables were removed -- 66 from the common blocks alone. The extremely unstructured original code was greatly improved, largely through the addition of two variables: MOVETO a global variable, and NDLOOP a local variable -- and though careful selection of code to be isolated as a new module, based on its structure. This permitted removing the large number of GOTO's that had made spaghetti of the original code. Where a GOTO had seemed necessary, appropriate choice of module boundaries and use of the MOVETO variable was substituted.

The software metrics supplied by the McCabe software analysis tool indicate the number of unstructured constructs was reduced from 46 to 2. (The remaining 2 are in the main module.) The complexity of the most complex module was reduced from 89 linearly independent paths, to 13.

What CLIGEN does, and How:

Cligen generates nine daily weather outputs from long-term observed data:

- Maximum Temperature
- Minimum Temperature
- Dew Point Temperature (TDP)
- Probability of Precipitation (today)
- Amount of Precipitation
- Time to Peak
- Radiation
- Wind Direction
- Wind Velocity

For each of these, a pseudo random number generator (RANDN) produces a uniform distribution of random numbers with values evenly distributed between zero and one. A different set of seeds is fed to RANDN for each output parameter, providing a separate instance (copy) of the random number generator for each parameter. To generate normal distributions these random numbers are fed into a routine (DSTN1) which produces a distribution of standard normal "deviates" for each distribution; i. e., a set of numbers with mean=0 and SD=1. With the exception of Tmin, Tmax, and Tdew, each of the nine outputs is produced independently of the others. However, for normal distributions, because DSTN1 uses two random number inputs, and because the second input for today becomes the first for tomorrow, tomorrow's value is not independent of today's.

To generate its outputs, Cligen uses the monthly averages and standard deviations from the observed period of record for the site.

- The current day's ***maximum temperature, minimum temperatures,*** and ***dew point temperature*** are assumed to be independent normal

distributions. (Common sense tells us they are not independent, but assuming that they are greatly simplifies the mathematics.) When the SD of Tmin is less than that of Tmax, Tmin's daily value is generated by scaling (multiplying) the SN deviate (positive or negative) by the parameter's standard deviation for the current month, and adding its monthly mean. Tmax and Tdew are generated from a "difference" distribution. When SD of Tmax is less than that of Tmin, the calculations for Tdew and Tmin are similarly based on Tmax.

- For ***precipitation***, things get a bit complicated. Precip is calculated using a joint distribution of the probability that precip occurs today, given precip or no precip yesterday. I.e., Cligen uses the probability of a wet day following a dry day, and a wet day following a wet day. A random number is compared to the probability of an event at the site today, given that yesterday was wet, or that it was dry. If this indicates precip should occur, the amount is generated using a Pearson Type-III equation and the monthly skewness, in addition to monthly mean and SD. Precipitation is assumed to follow a skewed normal distribution.
- To generate ***solar radiation***, a random standard normal deviate is produced and multiplied by the monthly std. dev. (as with max. & min. temp.). This value is then multiplied by 1/4 the difference between the monthly mean and its theoretical maximum, and added to the monthly mean.
- ***Wind speed*** is generated using skewness in addition to mean and SD, from 16 different directions, and the percent time that it comes from each of those directions.

(Most of the equations that perform this work are clearly identified in subroutine CLGEN (CLIGEN5106.F, starting in line 932).)

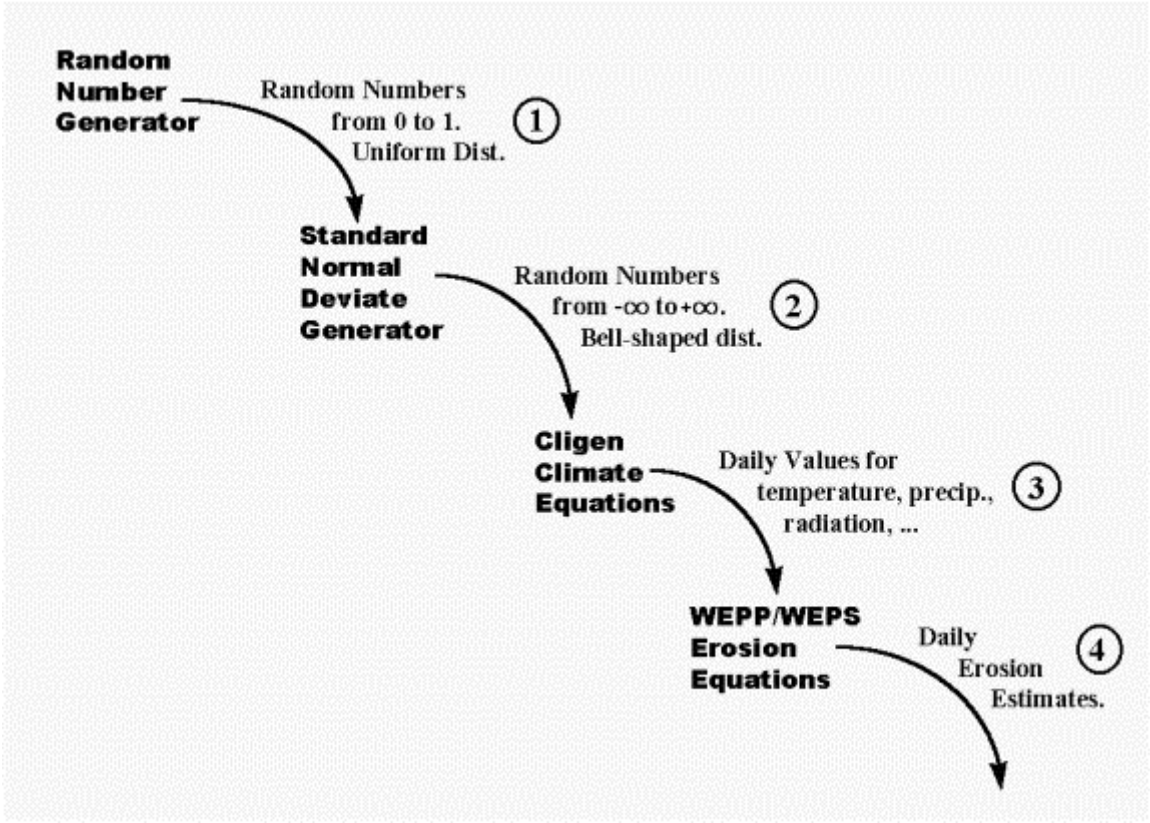
A simple range check forces daily minimum temperature to be less than maximum temperature. This alters the distribution of minimum temps, depressing their long-term mean. (While one can argue that this is not strictly "correct", it only have a significant effect on the output of a model if it is sensitive to minimum temperature. Preliminary tests with WEPP indicate that it is not.) There is also a range check for Solar Radiation which can alter its distribution. With radiation there is a "theoretical maximum" (RMX) for the site. If RMX is less than the monthly observed value, it is set to the observed value. If the generated value is more than RMX, it is set to 90 percent of RMX. If it less than zero, it is set to 5 percent of RMX. (Note that these checks alter the resulting distribution so that the output can not be guaranteed to match the input numbers from which it was generated.)

General Design and Purpose:

The user should note that CLIGEN can be expected to do a very good job of reproducing a long-term distribution, like 30 years, for its parameters. However, it clearly was not designed to produce outputs that make sense when examined on an the basis of an individual day. For example, for any given day, temperature, precipitation, and solar radiation are generated independently of each other.

Testing CLIGEN's ability to Reproduce its Input Parameters without Quality Control on the Random Number Generator:

Overview of the Stochastic Process



Testing the weather values output from CLIGEN:

It is desirable to be able to demonstrate that the CLIGEN outputs (step 3 in figure above) are reproducing the historical means from which they are generated. Not all the distributions are normal distributions; however, the Central Limit Theorem of statistics asserts that the means of samples taken from *ANY* type of distribution, approach a normal distribution as the number of means gets large -- even if the original distribution is NOT a normal distribution. This very conveniently makes "confidence intervals on the mean" an appropriate statistical test, without requiring knowledge about the type of underlying distribution. Confidence interval software was written to test the monthly means of the nine CLIGEN parameters generated from random numbers, against the monthly means and standard deviations used to generate them. Ideally, the CLIGEN outputs should converge back to the original values from which they were generated. Generally, without imposing quality control on the random number generator, they did not.

Testing the standard normal values:

Note that in contrast to the weather outputs from Cligen, the random standard normal deviate inputs to the CLIGEN climate equations (step 2 in figure above) are supposed to form a standard normal distribution,

so their means and variances may both be legitimately tested assuming the underlying distribution is standard normal. Cligen was modified accordingly to output the standard normal deviates to a file for inspection and testing. The monthly means and variances (SD's) of these values were tested against the population they are supposed to approximate, i. e., with mean=0 and SD=1. (Interval tests on the means of generated daily climate values may be expected to give very similar results; however, they are not always identical because of the need to use the calculated SD from the observed data.) The versions of CLIGEN without quality control did not reproduce standard normal distributions.
Part of the confidence interval software written was the "conf_lim" subroutine. It displays the probability that the mean of the generated numbers for a given month is different from the mean of the population from which they were derived, with the given monthly standard deviation. One can examine whether the number of values at or above 90 percent exceeds the expected value (10 percent of the number of means tested). If so, the result obviously is unacceptable. It is much more difficult to specify what constitutes acceptable results.

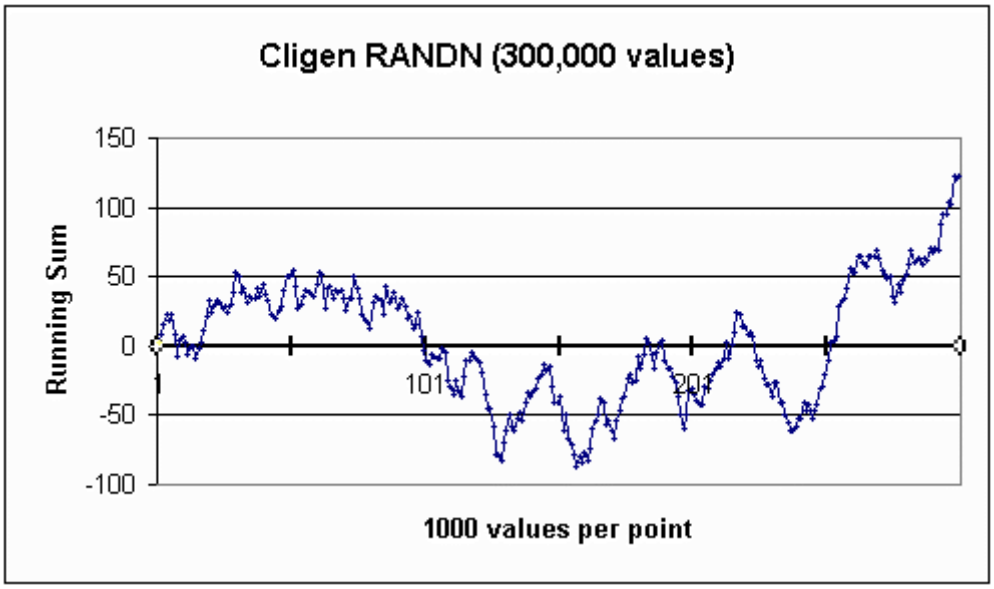
Testing the most elementary random values:
The RNG from which all CLIGEN's stochastic numbers flow (step 1 in the figure above) is supposed to produce a ***uniform distribution*** such that all values between zero and one are equally likely to occur. It is actually fairly simple to demonstrate that the RNG is not working as desired and expected. The uniform distribution has an expected value of 0.5 . If $E(x_i) = 0.5$, then for any X_i that might be generated, $E(x_i - 0.5) = 0$. As can be deduced from the three statements below, the running sum of $x_i - 0.5$ should converge to zero.

$$E\sum_{i=1}^{N}(x_i - 0.5) = \sum_{i=1}^{N} E(x_i - 0.5)$$

$$E(x_i - 0.5) = 0$$

$$E\sum_{i=1}^{N}(x_i - 0.5) = 0$$

However, in our tests taking 300,000 values from a stream of 3.65 million iterations of CLIGEN'S RNG, **it did not!**



From a visual inspection of the graph above, it seems clear that CLIGEN's RNG is not producing a uniform distribution as advertised. If it were, the graph should converge to zero. Unfortunately, we observed similar results from a couple other popular RNG's. This suggests that a number of stochastic models may be producing results below the quality that would otherwise be expected, solely due to the poor quality of the distributions produced by their random number generators. For CLIGEN, this indicates the need for some sort of quality assurance for the distributions coming out of the RNG.

Comparisons:

CLIGEN was run for four climatically diverse sites in the U. S. (College Station, TX; Indianapolis, IN; Moscow, ID; and Tucson, AZ. Summary table from 30 year run below.), for periods of 1 to 1000 years. The monthly means of min. temp., max. temp., precip., and radiation outputs were subjected to confidence interval tests. Lessons learned are summarized immediately below.

- Prior to the introduction of RANSET, CLIGEN was tested both with, and without linear interpolation (applied to means, std. dev. and skewness). In all four climates it converged on its final mean monthly value for all four parameters within 28 years. (This seems to be due to the limits of the single precision calculations. The standard normal deviates, which were numbers of smaller magnitude and were stored as double precision, did not converge in 28 years.) Weather experts consider 30 years to be the time required to achieve a "representative sample" of weather data. It is also considered to be the length of run needed to establish "representative weather" from a climate generator. These observations coupled with CLIGEN's numeric convergence in 28 years imply that the user should run CLIGEN for 30 years to generate a weather file representative of the site. (Telephone conversation with Clarence Richardson.)
- Using the original version of Cligen, mean monthly values out of range at the 90 percent level of probability or higher were still

produced at all sites when the system reached steady state convergence at 28 years. It was common for the number to exceed the expected value of 10 percent of the number sets tested. Predictably, the generated outputs followed the same pattern as the generated random standard normal deviates produced by DSTN1. (There is some possibility of differences because of the use of observed standard deviations for the generated outputs.)

- Replacing the ***random number generator*** RANDN with the UNIX BSD-4.2 C-language RANDOM function significantly reduced the number of monthly means differing at the 90 percent level of confidence or above, regardless of the length of run. It also improved (reduced) the average level at which one is confident of a difference between the outputs and the inputs which generated them. This was especially evident for short runs. A random number generator (in FORTRAN) from MIT; another from ACM; the GNU g77 RAND routine, and a one-line random number generator (in FORTRAN) were all tested. (They are available from this site.) Most of these are "linear congruential" random number generators: the BSD "random" function is not. Even though alternative random number generators offered obvious improvement over the original CLIGEN RANDN, the results still were often not satisfactory, because the number of "outliers" exceeded the expected percentage.
- Replacing DSTN1 with another function adapted from an MIT author, to generate ***standard normal deviates*** also improved the results. Replacing it with the analogous routine from ACM improved the results further. However, as with the random number generator, the overall results still proved unsatisfactory.
- Random number generators produce a "stream" of numbers which follow specified characteristics. However, those characteristics can only be ensured if consecutive numbers are collected from the stream. You cannot simply sample every 9^{th} number from a routine that generates a standard normal distribution, and expect the result to be a standard normal distribution. For this reason subroutine RANSET was coded to generate all the values for a single parameter for a month, before moving on to the next parameter. In the original CLIGEN code, the random numbers were generated for each parameter in sequence for the current day, and then for the next day, and so on. This worked well because the original CLIGEN uses a separate instance of the random number generator for each parameter generated. Attempts to use alternate single-instance generators only proved successful when streams of numbers were generated for a month at a time. (CLIGEN uses the random numbers in the original order.) Additionally, DSTN1 uses two random numbers to generate each value, and a new number is rotated into one of these positions each day. This offers the added benefit of ensuring that each day's value for a parameter is not independent of its neighbor a day earlier or a day later. For these reasons, there was strong incentive to utilize the original RANDN and DSTN1 code. This was realized through use of the feedback loop for "quality control", which was coded into subroutine RANSET.

Summary of 30 year WEPP Runs:

		Avg Ann Precip mm	Avg Ann Runoff mm	Avg Ann Detach kg/m2	Avg Ann Depos kg/m2	Avg Ann Sed. Loss kg/m Wid.
IND	Orig	1034.93	144.54	10.117	29.431	65.573
(1013.1)	Orig+Yu	1034.93	148.59	10.799	35.490	66.334
	Rand	1011.87	130.65	9.702	30.517	60.823
	Rand+Yu	1011.87	138.13	10.730	36.456	64.829
COS	Orig	1040.39	307.61	33.614	67.413	214.908
(959.5)	Orig+Yu	1040.39	322.53	36.169	73.072	219.673
	Rand	987.65	265.40	28.093	60.966	174.056
	Rand+Yu	987.65	271.30	30.210	71.948	169.293
MOS	Orig	643.03	21.71	1.920	7.995	11.264
(621.5)	Orig+Yu	643.03	8.08	0.784	5.112	3.121
	Rand	644.40	21.41	1.475	6.453	8.407
	Rand+Yu	644.40	4.76	0.317	1.669	1.776
TUC	Orig	310.07	25.45	4.213	11.327	21.925
(293.2)	Orig+Yu	310.07	21.59	3.694	11.149	16.162
	Rand	291.06	21.20	3.893	10.022	19.446
	Rand+Yu	291.06	13.41	2.488	7.577	9.782

COS - College Station, TX

MOS - Moscow, ID

TUC - Tucson, AZ

Orig - Recoded/Original Version 4.2 CLIGEN

Orig+Yu - Recoded/Original Version 4.2 CLIGEN with storm intensity corrections by Bofu Yu.

Rand - CLIGEN 4.2 with Random Number checks for both Mean and SD.

Rand+Yu - CLIGEN 4.2 with Random Number checks and storm intensity corrections by Bofu Yu.

In a test of 100-year Cligen runs at 95 stations in the 48 states of the continental U. S., Cligen V-5.x (with RNG quality control) more closely matched actual average annual precipitation values than V-4.x (without RNG QC) at 80 percent of the stations. V-4.x showed a bias to over predict by an average of 16.9 mm (2.27%), and V-5.x underpredicted by an average of 1.57 mm (0.16%). Subsequent confidence interval tests on the monthly values showed that high monthly skew values were sometimes associated with generated monthly mean values below 40% of the observed values for both V-5.x and V-4.x. Consequently extreme skew values received further investigation.

Skew Coefficient for Precipitation:

When running confidence interval tests on the monthly precipitation values from 100-year runs of Cligen, Meyer discovered that skew coefficients above 4.5 seem to be associated with the generation of values that matched the "target" values very poorly. This results from the generation of a high percentage of precipitation events with "negative" amounts, which were subsequently converted to the default value of 0.01 inch (0.3 mm). It is assumed that the Pearson Type III equation used to estimate precipitation in Cligen, is not sufficiently robust to be used with high skew values. The target values were calculated by computing the probability of precipitation P(W) from the conditional probabilities P(W|W) and P(W|D) (equation below). P(W) is then multiplied by the number of days in the month, and the mean precipitation per event, to give the long term mean monthly precipitation.

The probability of precipitation

$$P(W) = \frac{P(W \mid D)}{1 - P(W \mid W) + P(W \mid D)}$$

The skew coefficients calculated by Microsoft Excel from the historical data, closely match those in the corresponding station parameter files. There are many ways to calculate a skew coefficient! The equations used by Excel are listed below.

In Excel the skew coefficient is

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s}\right)^3 \qquad \text{where} \quad s=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n-1}}$$

Some Practical Considerations -- Mathematics VS. the Real World:

- The confidence interval tests show the highest level at which we can be sure of a difference. When we say we are 90 percent certain two sets of numbers are not from the same population, we are implying there is a 10 percent probability that they are from the same population; i.e., that we are wrong. However, we generally accept the risk, because we would rather have a 90 percent chance of being right, than only a 10 percent chance. Because we will be wrong 10 percent of the time, it is completely reasonable that 10 percent of our measurements may fall outside our 90 percent confidence limits! (Or if 5 percent fall out side the 95 percent limits, etc.) So, in 12 monthly means for 4 different parameters, 4.8 (5) of them could reasonably be outside the 90 percent confidence limits. A higher number would not reasonable. So, we can reject as "unacceptable" runs of CLIGEN with differences above prescribed levels. However, there is no analogous guideline for automatically prescribing a run as "acceptable".
- Some anomalies can occur when looking at the output parameter values themselves. One should be wary of differences that are "statistically significant", but trivial in the real world. For example, after a 28-year run, the upper and lower 95 percent confidence limits for radiation differed by only 3 Langleys; however, review of the instrument calibration guidelines revealed that an instrument was only thought to require calibration if it was more than 50 Langleys off -- otherwise its error in measurement was deemed acceptable. So, our 95 percent confidence interval (1.5 units from the mean) was 33 times as tight as the calibration guidelines for the instrument collecting the data! Similar observations were made about the precipitation data.
- It is mathematically sensible to analyze the ***standard normal deviates*** from which parameters are generated. They are purely mathematical artifacts. Since they are not measured, they have no associated level of precision.
- Because they are the raw materials from which a weather generator's time series are produced, standard normal deviates should be tested both for appropriate ***mean*** and ***standard deviation*** values. This can be implemented in the model code as it is in CLIGEN with a standard confidence interval test.

Known Caveats, Bugs, and Problems:

- Coming from a mathematical interpretation Cligen seems well suited to reproduce the trends of its parameters over the long term; however, potential users should realize at the onset that

the generation of daily values for temperature, radiation, and precipitation that are independent of each other, implies that values taken one day at a time will not sensibly mimic the real world. That is because these values are related in the real world, but are not generated that way in CLIGEN.

- Cligen generates its value for the first day of the month from the same values as those for the last day of that month (a month later). However, the value the last day of the previous month (one day earlier) is generated from different values - those for the previous month. This troubles some people. However, the standard deviation for a month is generally several times the difference between the means of adjacent months. So, in practice, it may not make as much difference as one would guess. (Three types of interpolation have been coded, and preliminary results show little difference in WEPP erosion estimates.) As indicated above, one should consider the application which is targeted to use CLIGEN's output.
- Except as noted for Min Temp, the values for Max Temp, Min Temp, Solar Radiation, and Precipitation are all generated as independent distributions in CLIGEN. Some people note that these parameters are not independent in the physical world. For example, the temperature and radiation go down on a rainy day. However, this simplifying assumption in CLIGEN permits the generation of a large number of long term time series (including rainfall intensity and time to peak) that mimic very well what occurs in the real world. This is much more difficult to accomplish in models that require knowledge of the inter-relationships between these parameters at the site in question. The "best" approach depends upon which tradeoffs the user feels most comfortable making.
- In choosing a random number generator there are several considerations. First the period should be as long as possible. This refers to how many times the generator is executed before the seeds are used again. Second, the outputs should be uniform, and unbiased. They should not be "clustered" by value, but evenly and equally distributed over the output range. Lack of this property is evident when a discernible pattern emerges. Confidence interval tests help make this flaw visible. Another more subtle flaw occurs in random number generators based on the system clock, as many PC RNG's are. The outputs may increase in value until they reach a point where they start over again. Each individual output should be independent of those adjacent to it in time. Finally there is convergence. Does the mean value produced converge rapidly - preferably faster than the confidence limits - and does it stay at that value. (Some random numbers seemed to go "unstable" as the length of the run increased toward 100 years.) Again, confidence interval tests make this fairly obvious.
- In using a random number generator, it is desirable to capture a sequence of consecutive outputs to generate consecutive days' values for a single parameter. Ideally it should make no practical difference if we grab every 7^{th} or 15^{th} random value to generate precip. In practice, it seems to, and some RNG's are more sensitive to this effect than others. For further information on this effect, see comments on "serial correlations" in Ch 7.1 of Numerical Recipes in Fortran.