1.6 STATISTICAL ANALYSIS MODULE

The AWRD Statistical Analysis Module, accessible via the AWRD Modules menu or through the buttons in the AWRD Interface, provides users with three different toolsets or calculators for deriving descriptive data based on either statistical summaries or probability distributions (the "Field Summary Statistics", the "Probability Distribution Calculator" and the "Summarize Theme") together with a classification and ranking tool ("Classify Theme by Multiple Criteria") and a regression tool ("Simple Linear Regression") (Figure 1.82 and Table 1.28).

FIGURE 1.82
The Statistical Analysis Module

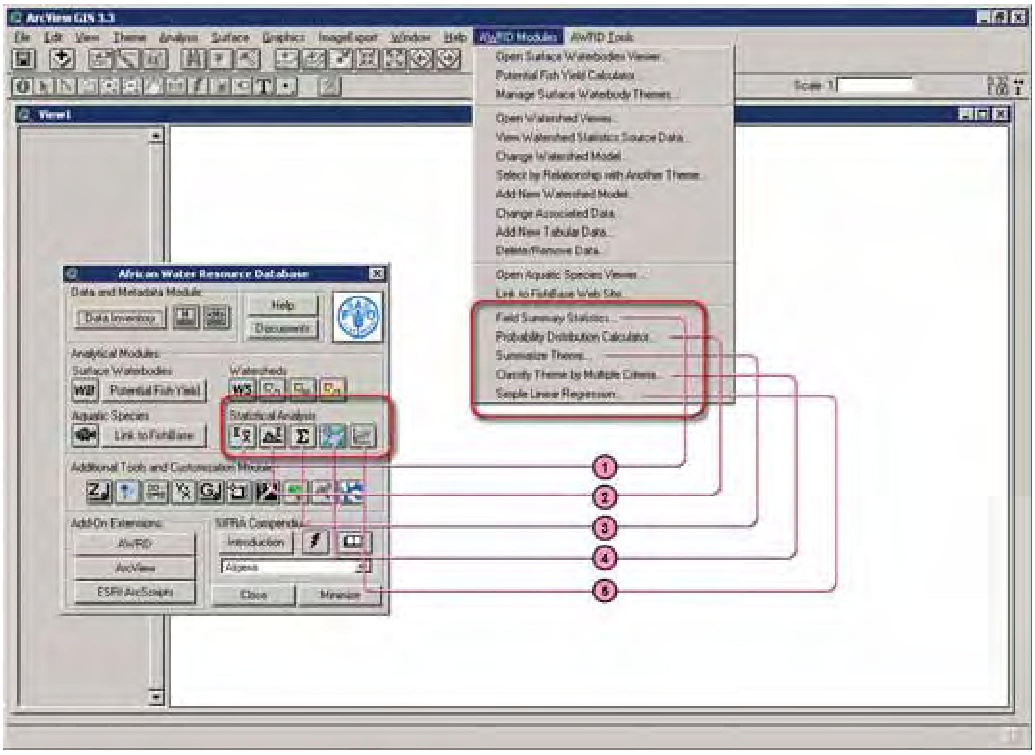

ArcView GIS 3.3
File
Edit
View
Theme
Analysis
Surface
Graphics
ImageExport
Window
Help
AWRD Modules
AWRD Tools
Open Surface Waterbodies Viewer...
Potential Fish Yield Calculator...
Manage Surface Waterbody Themes...
Open Watershed Views...
View Watershed Statistics Source Data...
Change Watershed Model...
Select by Relationship with Another Theme...
Add New Watershed Model...
Change Associated Data...
Add New Tabular Data...
Delete/Remove Data...
Open Aquatic Species Viewer...
Link to FishBase Web Site...
Field Summary Statistics...
Probability Distribution Calculator...
Summarize Theme...
Classify Theme by Multiple Criteria...
Simple Linear Regression...
Scale 1:
View1
African Water Resource Database
Data and Metadata Module
Data Inventory
Help
Documents
Analytical Modules
Surface Waterbodies
WB
Potential Fish Yield
Watersheds
WS
Aquatic Species
Link to FishBase
Statistical Analysis
Additional Tools and Customization Module
Add-On Extensions
AWRD
ArcView
ESRI ArcScripts
SIFRA Compendium
Introduction
Algeria
Close
Minimize
1
2
3
4
5

TABLE 1.28
Statistical Analysis Module buttons and menu items

Label (Fig.1.82)	AWRD Interface button	AWRD Modules menu option	Action executed
1	Σx̄	"Field Summary Statistics…"	*Field Summary Statistics*: this function allows you to generate a wide set of descriptive statistics on a theme.
2	Σ	"Probability Distribution Calculator…"	*Probability Distribution Calculator*: this function allows you to test critical values for probability levels based on several commonly-used statistical distributions. It calculates the probability, the cumulative probability and the inverse probability.
3	Σ	"Summarize Theme…"	*Summarize Theme*: this function allows you to group features in a theme based on common attribute values, and then generate several descriptive statistics for each group.
4		"Classify Theme by Multiple Criteria…"	*Classify Theme by Multiple Criteria*: this function allows you to classify features in a theme based on complex criteria, and to save the criteria sets for use with other datasets.
5		"Simple Linear Regression…"	*Simple Linear Regression*: this function allows you to conduct a simple linear regression on pairs of fields in a theme attribute table, including generating an ANOVA table, P-values, confidence bands and tests of the regression line slope.

Statistics and probability tools

The statistics and probability tools contained within the AWRD (Figure 1.82) provide users the ability to classify data, calculate a wide range of summary statistical data from themes and tables, and to easily generate both probability values and critical values from several statistical distributions.

Field summary statistics

This tool provides functions similar to those available in the basic ArcView "*Statistics…*" options under the standard "Field" menu in the Table menu bar, with the exception that there are both more options and statistics are reported with a higher level of precision. The tool may be used to generate statistics on either a theme in a view or a field in a table.

Summary Statistics on a Theme

The Σx̄ button in the Statistical Analysis Module section of the AWRD Interface will only be enabled if the user has a View open and active. When the user clicks the button, they will be prompted to identify the theme and fields they wish to calculate statistics on. If the button does not respond, click on a view to enable the button. This statistics tool is also available as a menu option in the AWRD Modules menu (i.e. "*Field Summary Statistics…*"), or as a button in the Select by Theme tool dialog and Query Builder tool dialog .

The user can also choose to calculate statistics on either all the features in the theme or only the selected features. If no features are selected, then this tool will use all the features regardless of which option is chosen. The user can also choose to calculate statistics on multiple fields at one time.

The user can also specify certain values that they do not want included in the analysis. For example, it is common practice to designate some number to mean "No Data", or to identify values that should not be involved in any analysis. Researchers often use -9999 or -99999 for this purpose, especially with datasets in which such a value would be impossible (i.e. elevation, population, area, etc.). The user can designate as many of these values as they want by entering them into the "Ignore Values" section and checking the "Ignore" box. The user can also choose between either the Simple or Advanced output.

The Simple output includes the *Sum, Number of Features, Mean, Minimum* and *Maximum*, and is reported in a text box.

1. Click on the "Add Basemap Image to View" tool 🌍 to load one of the image backgrounds (e.g. "Etopo2_2-5d.jp2") from the image database component folder. This background image is not necessary for proper functioning of these tools, but it makes it easier to locate your area of interest in the view.

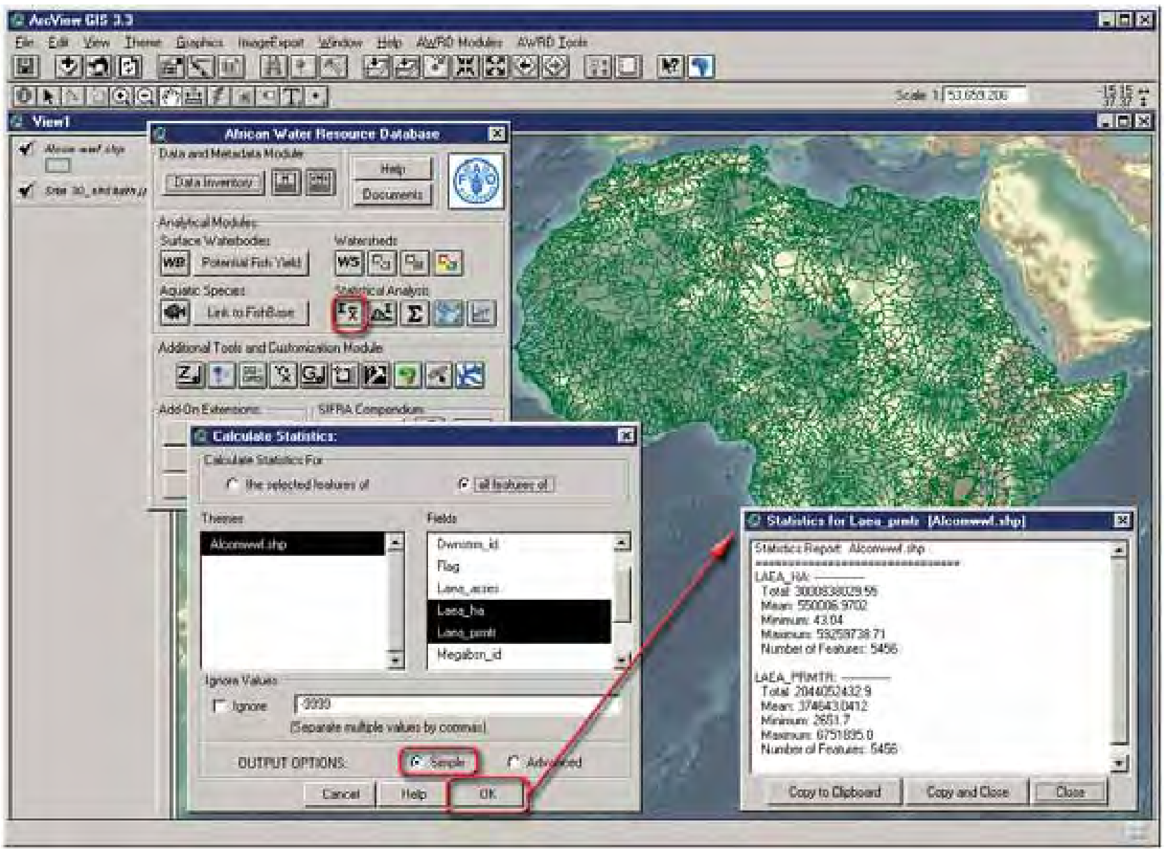
2. Click on the "Data Inventory" button Data Inventory to load one of the watershed model themes (e.g. "alcomwwf.shp") from the Watersheds database component.

3. Click on the Σx̄ button on the AWRD Interface, or the "Field Summary Statistics..." menu option in the AWRD Modules menu.

4. Select hectares (Laea_ha) and perimeter (Laea_prmtr) from the list by holding down the [SHIFT] key in the keyboard and click "OK" (Figure 1.83a).

Note The fieldname prefix "Laea", means that these values are based on the Lambert Equal Area Azimuthal projection.

FIGURE 1.83A
The statistics report (Simple output) for the FAO watershed model



Calculate Statistics:

Calculate Statistics For
○ the selected features of ● all features of

Themes: Alcomwwl.shp

Fields: Dwnstrm_id, Flag, Laea_acres, Laea_ha, Laea_prmtr, Megabsn_id

Ignore Values
□ Ignore -9999
(Separate multiple values by commas)

OUTPUT OPTIONS: ● Simple ○ Advanced

Cancel Help OK

Statistics for Laea_prmtr [Alcomwwl.shp]

```
Statistics Report: Alcomwwl.shp
*************************************
LAEA_HA: ----------
  Total: 3000838029.55
  Mean: 550006.9702
  Minimum: 43.04
  Maximum: 59259738.71
  Number of Features: 5456

LAEA_PRMTR: ----------
  Total: 2044052432.9
  Mean: 374643.0412
  Minimum: 2651.7
  Maximum: 6751895.0
  Number of Features: 5456
```

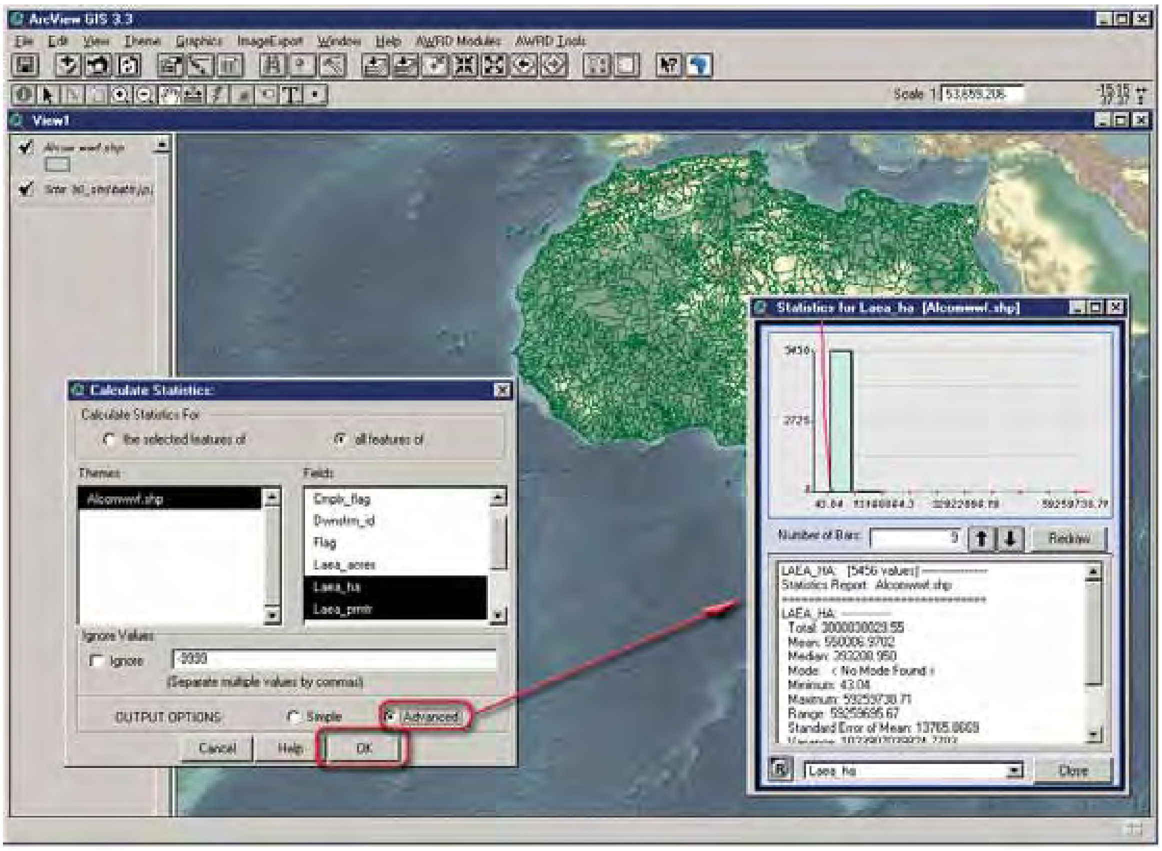
Copy to Clipboard Copy and Close Close

The Advanced output (following the same first steps listed above for the simple option) includes the *Sum, Mean, Median, Mode(s), Minimum, Maximum, Range, Standard Error of Mean, Variance, Standard Deviation, Number of Features,* and *Number of Null Values*, and is reported in a histogram (1.83b).

There will actually be only one histogram dialog open but the user can choose which set of statistics to view by choosing the field from the drop-down box at the bottom of the dialog. The user can also decide how many histogram bars to show by clicking the up/down arrows ⬆ and ⬇, and then the button Redraw. The red line behind the histogram bars shows how the bars would be arranged if the data followed a perfectly normal distribution. The R button in the bottom left of the dialog is the "Refresh" button, and is used if the image gets corrupted somehow. Click this button and the image will redraw itself.

FIGURE 1.83B

The statistics report (Advanced output) for the FAO watershed model


ArcView GIS 3.3
File Edit View Theme Graphics ImageExport Window Help AWRD Modules AWRD Tools
Scale 1: 53,659,206
View1
Statistics for Laea_ha [Alcomwwf.shp]
5456
2728
0
43.04
11860844.3
32922898.79
59259738.79
Redraw
Close

Calculate Statistics:

Calculate Statistics For: ○ the selected features of ● all features of

Themes: Alcomwwf.shp

Fields: Cmplx_flag, Dwnstrm_id, Flag, Laea_acres, Laea_ha, Laea_prmtr

Ignore Values: □ Ignore -9999 (Separate multiple values by commas)

OUTPUT OPTIONS: ○ Simple ● Advanced

Cancel Help OK

Number of Bars: 5

```
LAEA_HA: [5456 values] -----------
Statistics Report: Alcomwwf.shp
====================================
LAEA_HA: --------
  Total: 3000030023.55
  Mean: 550006.9702
  Median: 282208.950
  Mode:   < No Mode Found >
  Minimum: 43.04
  Maximum: 59259738.71
  Range: 59259695.67
  Standard Error of Mean: 13765.8669
```

Laea_ha

Summary statistics on a Field in a Table

The Σx button in the Table button bar will only be enabled if a numeric field has been selected. This tool will allow the user to generate a large number of statistics on the values in that field. The user may choose from: *Mean*; *Standard Error of the Mean*; *Confidence Intervals*; *Minimum*; *1st Quartile*; *Median*; *3rd Quartile*; *Maximum*; *Range*; *Variance*; *Standard Deviation*; *Average Absolute Deviation*; *Skewness*, both normal and Fisher's G1; *Kurtosis*, both normal and Fisher's G2; *Number of Records*; *Number of Null Values*; *Mode*; and lastly, *Total Sum* for any attribute field(s) within a set of selected records. If any records in the table are selected, then this will only operate on the selected records.

1. Click on the "Data Inventory" button Data Inventory to load one of the surface waterbody themes (e.g. "wria_swb.shp") from the Surface Waterbodies database component.

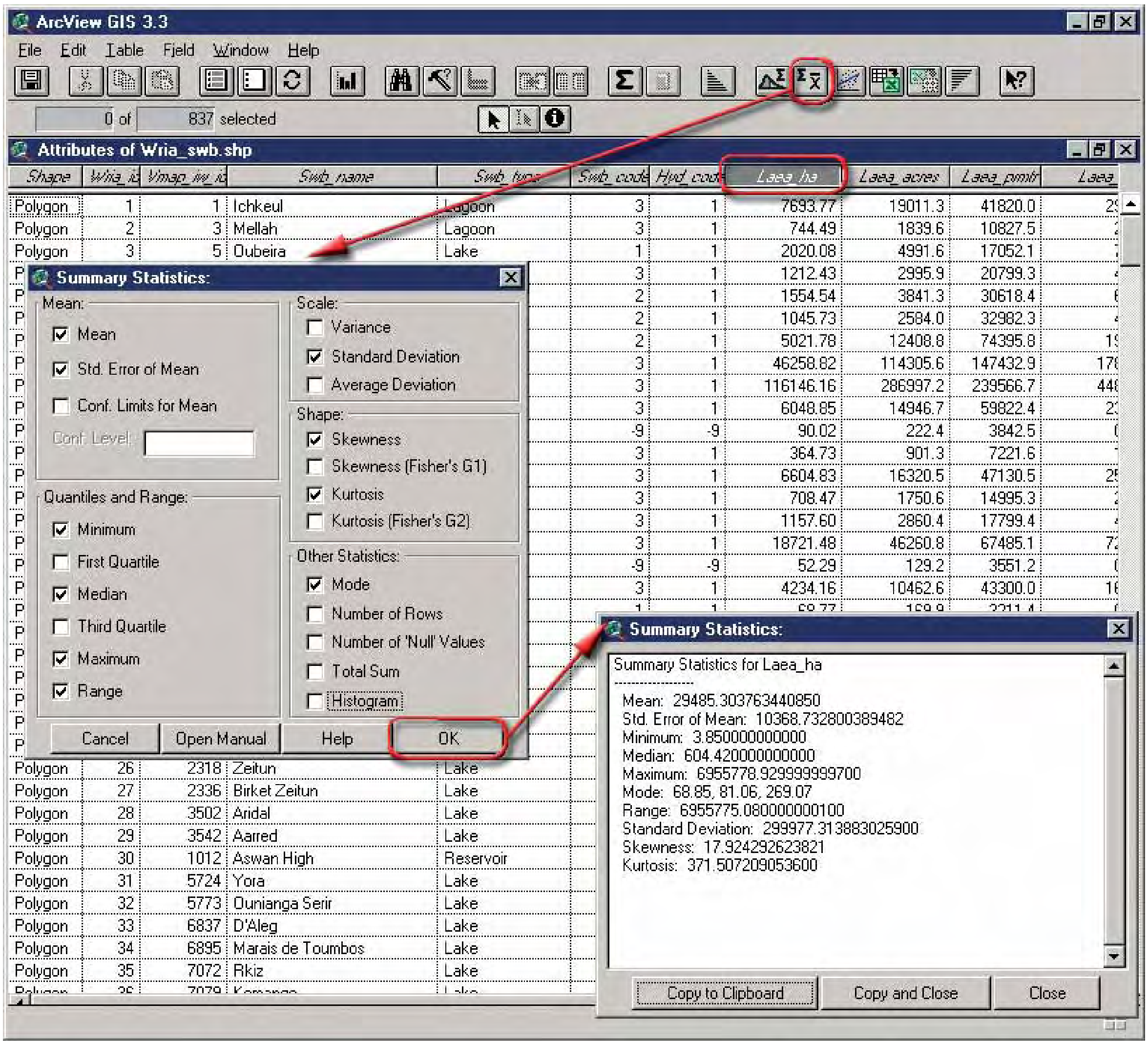
2. Open the attribute table for "wria_swb.shp" by selecting the theme in the table of contents, then clicking the standard ArcView "Open Theme Table" button ▦ in the View button bar.

3. With the table open, select the field you wish to analyse and then click the Σx button in the Table button bar.

4. Choose the desired statistics (e.g. *Mean, Standard Deviation, etc.*) and then click "OK". If you have selected the Histogram option, then the output will appear

in a histogram as illustrated in Figure 1.83b. If the Histogram option is not selected, then the output will appear in a report window (Figure 1.84).

FIGURE 1.84
Summary Statistics Calculator on a Field in a Table



Shape	Wria_i	Vmap_iw_i	Swb_name	Swb_ty	Swb_code	Hyd_cod	Laea_ha	Laea_acres	Laea_pmtr
Polygon	1	1	Ichkeul	Lagoon	3	1	7693.77	19011.3	41820.0
Polygon	2	3	Mellah	Lagoon	3	1	744.49	1839.6	10827.5
Polygon	3	5	Oubeira	Lake	1	1	2020.08	4991.6	17052.1
P					3	1	1212.43	2995.9	20799.3
P					2	1	1554.54	3841.3	30618.4
P					2	1	1045.73	2584.0	32982.3
P					2	1	5021.78	12408.8	74395.8
P					3	1	46258.82	114305.6	147432.9
P					3	1	116146.16	286997.2	239566.7
P					3	1	6048.85	14946.7	59822.4
P					-9	-9	90.02	222.4	3842.5
P					3	1	364.73	901.3	7221.6
P					3	1	6604.83	16320.5	47130.5
P					3	1	708.47	1750.6	14995.3
P					3	1	1157.60	2860.4	17799.4
P					3	1	18721.48	46260.8	67485.1
P					-9	-9	52.29	129.2	3551.2
P					3	1	4234.16	10462.6	43300.0
Polygon	26	2318	Zeitun	Lake					
Polygon	27	2336	Birket Zeitun	Lake					
Polygon	28	3502	Aridal	Lake					
Polygon	29	3542	Aarred	Lake					
Polygon	30	1012	Aswan High	Reservoir					
Polygon	31	5724	Yora	Lake					
Polygon	32	5773	Ounianga Serir	Lake					
Polygon	33	6837	D'Aleg	Lake					
Polygon	34	6895	Marais de Toumbos	Lake					
Polygon	35	7072	Rkiz	Lake					

Summary Statistics:

Mean:
- ☑ Mean
- ☑ Std. Error of Mean
- ☐ Conf. Limits for Mean
- Conf. Level:

Quantiles and Range:
- ☑ Minimum
- ☐ First Quartile
- ☑ Median
- ☐ Third Quartile
- ☑ Maximum
- ☑ Range

Scale:
- ☐ Variance
- ☑ Standard Deviation
- ☐ Average Deviation

Shape:
- ☑ Skewness
- ☐ Skewness (Fisher's G1)
- ☑ Kurtosis
- ☐ Kurtosis (Fisher's G2)

Other Statistics:
- ☑ Mode
- ☐ Number of Rows
- ☐ Number of 'Null' Values
- ☐ Total Sum
- ☐ Histogram

Summary Statistics:

```
Summary Statistics for Laea_ha
------------------
 Mean:  29485.303763440850
 Std. Error of Mean:  10368.732800389482
 Minimum:  3.850000000000
 Median:  604.420000000000
 Maximum:  6955778.929999999700
 Mode:  68.85, 81.06, 269.07
 Range:  6955775.080000000100
 Standard Deviation:  299977.313883025900
 Skewness:  17.924292623821
 Kurtosis:  371.507209053600
```

This tool can also be accessed with Avenue code, enabling more advanced users to pass these statistics to variables and then use the calculated values in other places. For details on using these functions with Avenue, please refer to the document "Notes on AWRD Statistical Tools.pdf" available by clicking the "Documents" button on the AWRD Interface dialog.

Probability Distribution Calculator

This extension includes two versions of a Probability Distribution Calculator, each of which calculates distribution data based on a variety of distributions and parameters. The first of these calculators, the "Probability Distribution Calculator" is designed to work from a view, while the second, the "Table Probability Distribution Calculator", is designed to be run from a feature attribute table or data table. Each tool allows the user to calculate distribution values from the *Beta*, *Binomial*, *Cauchy*, *Chi-Square*, *Exponential*, *F*, *Logistic*, *LogNormal*, *Normal*, *Poisson*, *Student's T*, and *Weibull* distributions. Both tools require some basic knowledge of statistical analysis and statistical probability distributions.

For example, suppose we are observing some phenomenon and we are wondering if that phenomenon is a normal event or if it represents something new. Perhaps we know that "normal" events are normally distributed (i.e. follow the Normal probability distribution), have a mean value of 10 units and a standard deviation of 2 units. If our observed phenomenon has a value of 15 units, then what is the probability that our observed event is "normal"?

Using basic statistical analysis methods, we would calculate the probability that our observed event is "normal" as being equal to the area under the normal curve that lies to the right of the observed value (where the observed value = 15 in this example). The probability Distribution Calculator allows you to do this:

1. Click on the button on the AWRD Interface or by selecting the AWRD Modules menu option "*Probability Distribution Calculator…*"

2. Specify the type of values to be calculated, i.e. Probability, Cumulative Probability, or Quantile values. In this example, we want the "Cumulative Probability", which will give us the area to the **Left** of the observed value.

3. Select a distribution type (Normal, in this example).

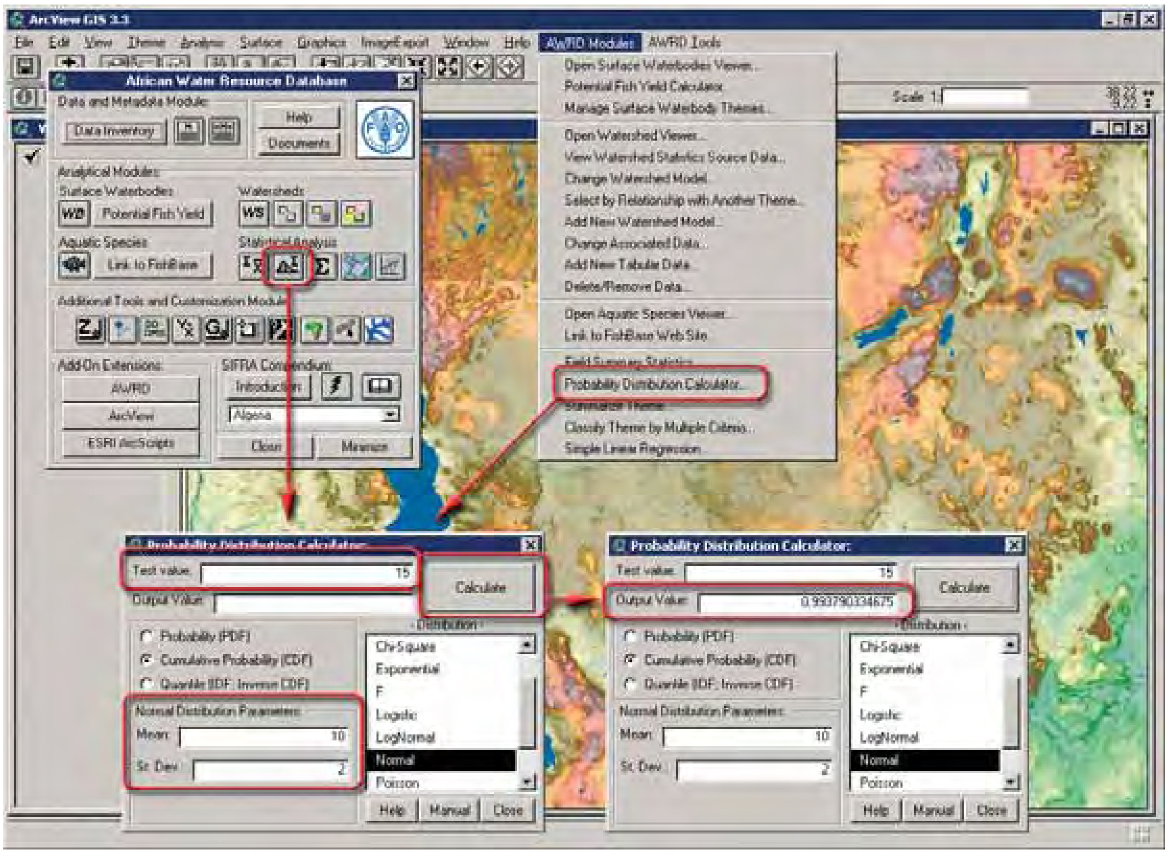
4. Enter the input and parameter values (e.g. Test value 15, Mean 10, and St.Dev. 2).

5. Click the "Calculate" button. The results will appear in the "Output Value" box of the calculator, located immediately below the "Test value" field box. The calculator will stay open until it is dismissed by clicking the "Close" button, so that a user can do a series of calculations and leave it open while other tasks are performed (Figure 1.85).

6. In this case, the cumulative area under the curve (i.e. the area to the left of the observed value) is approximately 0.9938. We are interested in the area to the right of the observed value, which is equal to (1-0.9938) = 0.0062.

7. Therefore we can conclude that there is approximately an 0.6% chance (slightly more than 6 chances in a thousand) that our observed event is a random occurance of a normal event. The probability is therefore high that our observed event represents something new.

FIGURE 1.85
Probability Distribution Calculator


ArcView GIS 3.3
File Edit View Theme Analysis Surface Graphics ImageExport Window Help AWRD Modules AWRD Tools
African Water Resource Database
Data and Metadata Module:
Data Inventory
Help
Documents
Analytical Modules:
Surface Waterbodies:
WB
Potential Fish Yield
Watersheds:
WS
Aquatic Species:
Link to FishBase
Statistical Analysis:
Additional Tools and Customization Module:
Add-On Extensions:
AWRD
ArcView
ESRI ArcScripts
SIFFIA Compendium:
Introduction
Algeria
Close
Minimize
Open Surface Waterbodies Viewer...
Potential Fish Yield Calculator
Manage Surface Waterbody Themes...
Open Watershed Viewer...
View Watershed Statistics Source Data...
Change Watershed Model
Select by Relationship with Another Theme...
Add New Watershed Model...
Change Associated Data...
Add New Tabular Data...
Delete/Remove Data...
Open Aquatic Species Viewer...
Link to FishBase Web Site
Field Summary Statistics
Probability Distribution Calculator...
Standardize Theme
Classify Theme by Multiple Criteria...
Simple Linear Regression...
Probability Distribution Calculator:
Calculate
Distribution
Chi-Square
Exponential
F
Logistic
LogNormal
Normal
Poisson
Help
Manual
Close
Probability Distribution Calculator:
Calculate
Distribution
Chi-Square
Exponential
F
Logistic
LogNormal
Normal
Poisson
Help
Manual
Close

Probability Distribution Calculator:

Test value: 15

Output Value:

○ Probability (PDF)
● Cumulative Probability (CDF)
○ Quantile (IDF; Inverse CDF)

Normal Distribution Parameters:
Mean: 10
St. Dev: 2

Distribution: Normal

Probability Distribution Calculator:

Test value: 15

Output Value: 0.9937903346675

○ Probability (PDF)
● Cumulative Probability (CDF)
○ Quantile (IDF; Inverse CDF)

Normal Distribution Parameters:
Mean: 10
St. Dev: 2

Distribution: Normal

The **Table Probability Distribution Calculator** is designed to work on either all or a selected set of records in a table, applying the distribution parameters to each value and saving the results to a field in that table. For example, suppose we are examining mercury concentrations in fish harvested from a set of surface waterbodies to see if any are unusually high, which might indicate mercury contamination in that waterbody. If we assume that natural mercury concentrations cause fish to have mercury levels of 0.4 ppm (parts per million), and that these mercury levels are normally distributed with a standard deviation of 0.2 ppm, then we can take the observed mercury concentration from each waterbody and calculate the probability that the observed mercury levels are normal. **Note:** This example is purely hypothetical, intended only to illustrate the use of this tool. The authors of the AWRD do not intend to make any claims about the actual mercury levels of surface waterbodies in Africa.

This calculator is opened from within a table by clicking on the ΔΣ button in the Table toolbar. The operation of this calculator is slightly more complicated than that of the view-based calculator and requires a user to:

1. Open the table of mercury concentration levels per surface waterbody.
2. Click on the ΔΣ button in the Table toolbar.
3. Select the field containing the "Input" values (e.g. "Mercury Concentration").
4. Choose whether to "Make a new field" or to select an existing field to save the "Output" values into.
5. Select a distribution type ("Normal", in this example) and enter the necessary parameter values (e.g Mean 0.4 and St.Dev. 0.2).

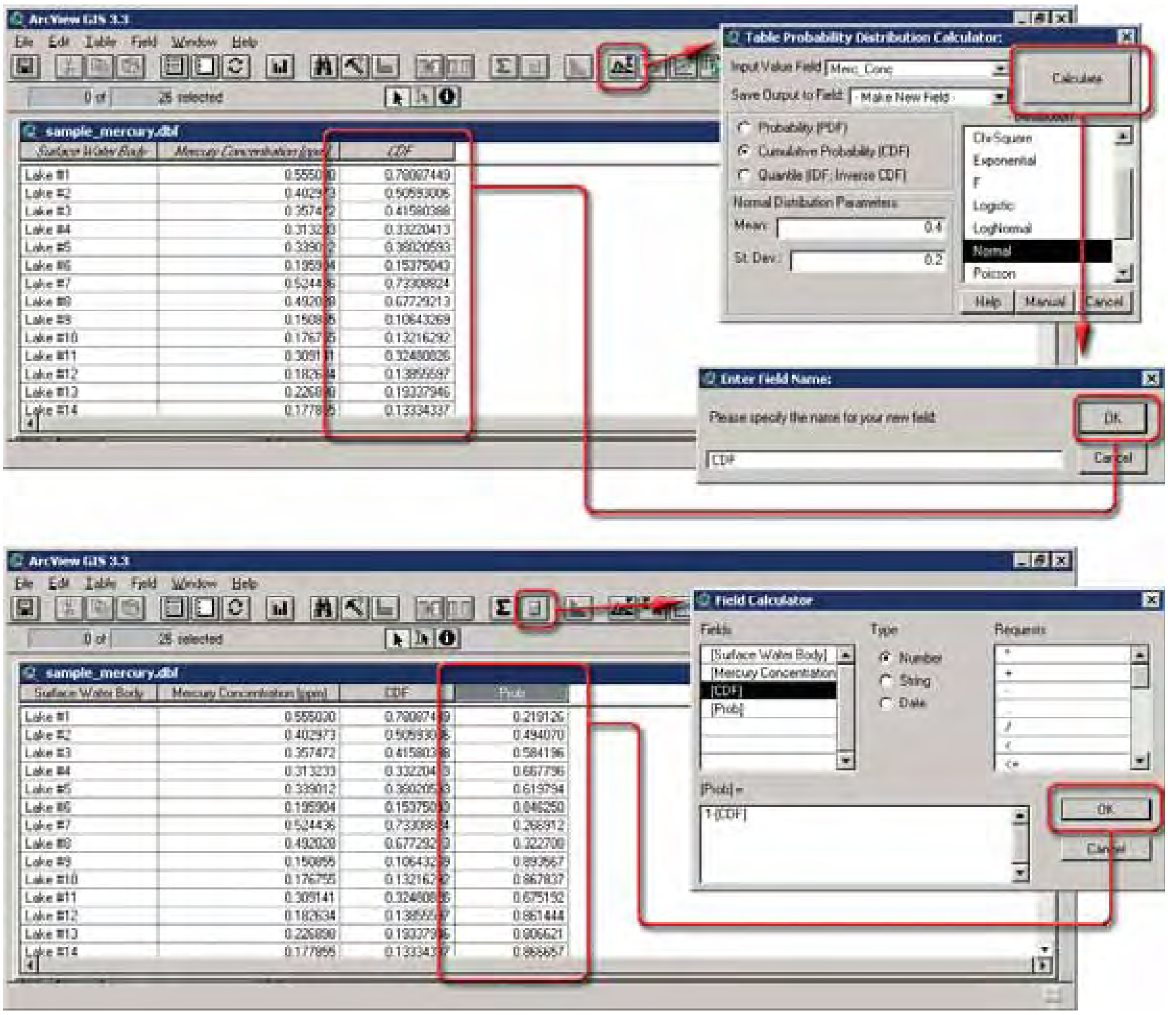
6. Specify the type of values to be calculated, i.e. Probability, Cumulative Probability, or Quantile (Cumulative Probability in this example).

7. Click the "Calculate" button to generate distribution values for all selected records. The dialog stays open until either the "Calculate" or the "Cancel" button is clicked (Figure 1.86).

8. As with the view-based calculator, this function calculates the proportion of the normal distribution curve that lies to the left of the observed value. The probability value we are interested in is equal to (1-proportion), so we now need to add a new field and manually calculate it to be equal to (1 – [CDF]).

9. The new field contains the probabilities that each lake contains a normal level of mercury.

FIGURE 1.86
Table Probability Distribution Calculator



sample_mercury.dbf

Surface Water Body	Mercury Concentration (ppm)	CDF
Lake #1	0.555030	0.78087449
Lake #2	0.402973	0.50593006
Lake #3	0.357472	0.41580388
Lake #4	0.313233	0.33220413
Lake #5	0.339012	0.38020593
Lake #6	0.195904	0.15375043
Lake #7	0.524436	0.73308824
Lake #8	0.492028	0.67729213
Lake #9	0.150855	0.10643269
Lake #10	0.176755	0.13216292
Lake #11	0.309141	0.32480826
Lake #12	0.182634	0.13855597
Lake #13	0.226890	0.19337946
Lake #14	0.177855	0.13334337

Input Value Field: Merc_Conc
Save Output to Field: - Make New Field -
○ Probability (PDF)
● Cumulative Probability (CDF)
○ Quantile (IDF; Inverse CDF)
Normal Distribution Parameters
Mean: 0.4
St. Dev: 0.2

Please specify the name for your new field
CDF

sample_mercury.dbf

Surface Water Body	Mercury Concentration (ppm)	CDF	Prob
Lake #1	0.555030	0.78087449	0.219126
Lake #2	0.402973	0.50593006	0.494070
Lake #3	0.357472	0.41580388	0.584196
Lake #4	0.313233	0.33220413	0.667796
Lake #5	0.339012	0.38020593	0.619794
Lake #6	0.195904	0.15375043	0.846250
Lake #7	0.524436	0.73308824	0.266912
Lake #8	0.492028	0.67729213	0.322708
Lake #9	0.150855	0.10643269	0.893567
Lake #10	0.176755	0.13216292	0.867837
Lake #11	0.309141	0.32480826	0.675192
Lake #12	0.182634	0.13855597	0.861444
Lake #13	0.226890	0.19337946	0.806621
Lake #14	0.177855	0.13334337	0.866657

Fields: [Surface Water Body] [Mercury Concentration] [CDF] [Prob]
Type: ● Number ○ String ○ Date
[Prob] =
1-[CDF]

The Distribution functions included with either the view-based calculator or the table calculator may be grouped into 3 categories:

- In general, the *Probability Density Functions* return the probability that the Test Value = X given that particular distribution.
- The *Cumulative Distribution Functions* return the probability that the Test Value ≤ X, given that particular distribution.

- The *Quantile Functions* (sometimes referred to as *Inverse Density Functions* or *Percent Point Functions*) return the Value *X* at which *P*(*X*) ≤ [specified probability], given that particular distribution.

Interested persons can review the references to find source code and computational methods of calculating these functions. Especially recommended are Croarkin and Tobias (date unknown) and McLaughlin (2001) for illustrations of the various distributions, and Press *et al.* (2002) and Burkardt (2001) for computational methods. For details on using Avenue code to access these functions, please refer to the document “Notes on AWRD Statistical Tools.pdf” available by clicking the “Documents” button on the AWRD Interface dialog.

Summarize Theme Tool

This tool provides functions similar to those available using the basic ArcView “Summarize” button in the Table button bar, except that this function offers more power and options and is a little more intuitive to use than the standard ArcView Summarize function. The tool is used to divide a dataset up into smaller datasets based on some attribute value, and then potentially to calculate statistics on each subset of data. The tool can also be used to combine all the features in each subset into a single feature and then to export those features into a new theme.

The user will be prompted to identify the theme and fields to use to divide the data, and optionally a set of statistics to calculate for each subset. For example, suppose a user is interested in how many surface waterbody points from the “Sadc_swb.shp” theme lie in each district. Each point has a District value in the database, so the user could use the Summarize tool to divide the data into subsets based on the District values. To do this:

1. Click on the “Add Basemap Image to View” tool 🌍 to load one of the image backgrounds (e.g. “Vrtl_map.sid”) from the image database component folder. This background image is not necessary for proper functioning of these tools, but it makes it easier to locate your area of interest in the view.

2. Click on the “Data Inventory” button Data Inventory to load one of the surface waterbody themes (e.g. SADC Surface Waterbody Database, or “sadc_swb.shp”) from the Surface Waterbodies database component.

3. Click the Σ button in the Statistical Analysis Module section of the AWRD Interface or select the AWRD Modules menu item “*Summarize Theme...*” to start the process. This button will only be enabled if the user has a View open and active.

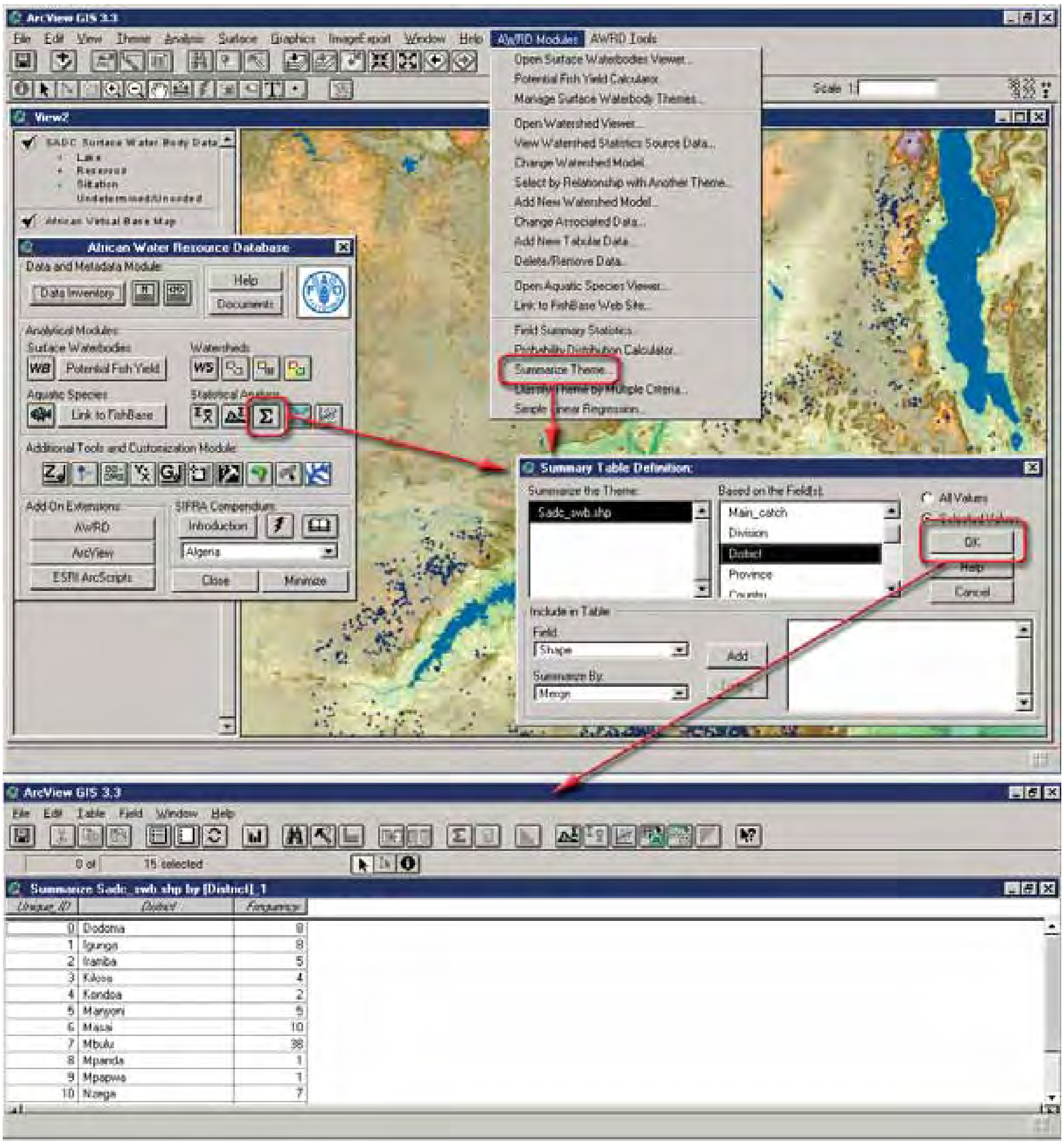
4. Identify the fields to use to separate the data into subsets (e.g. Districts). If the button does not respond, click on the view to enable the button.

5. Click “OK”.

In the illustration below (Figure 1.87a), the user is only using the currently selected values and is not calculating any statistics for each subset of data. In this case, the summary table will calculate only the number of “Sadc_swb.shp” points in each district.

FIGURE 1.87A

Using the Summarize Theme tool to divide the data into subsets based on the District values

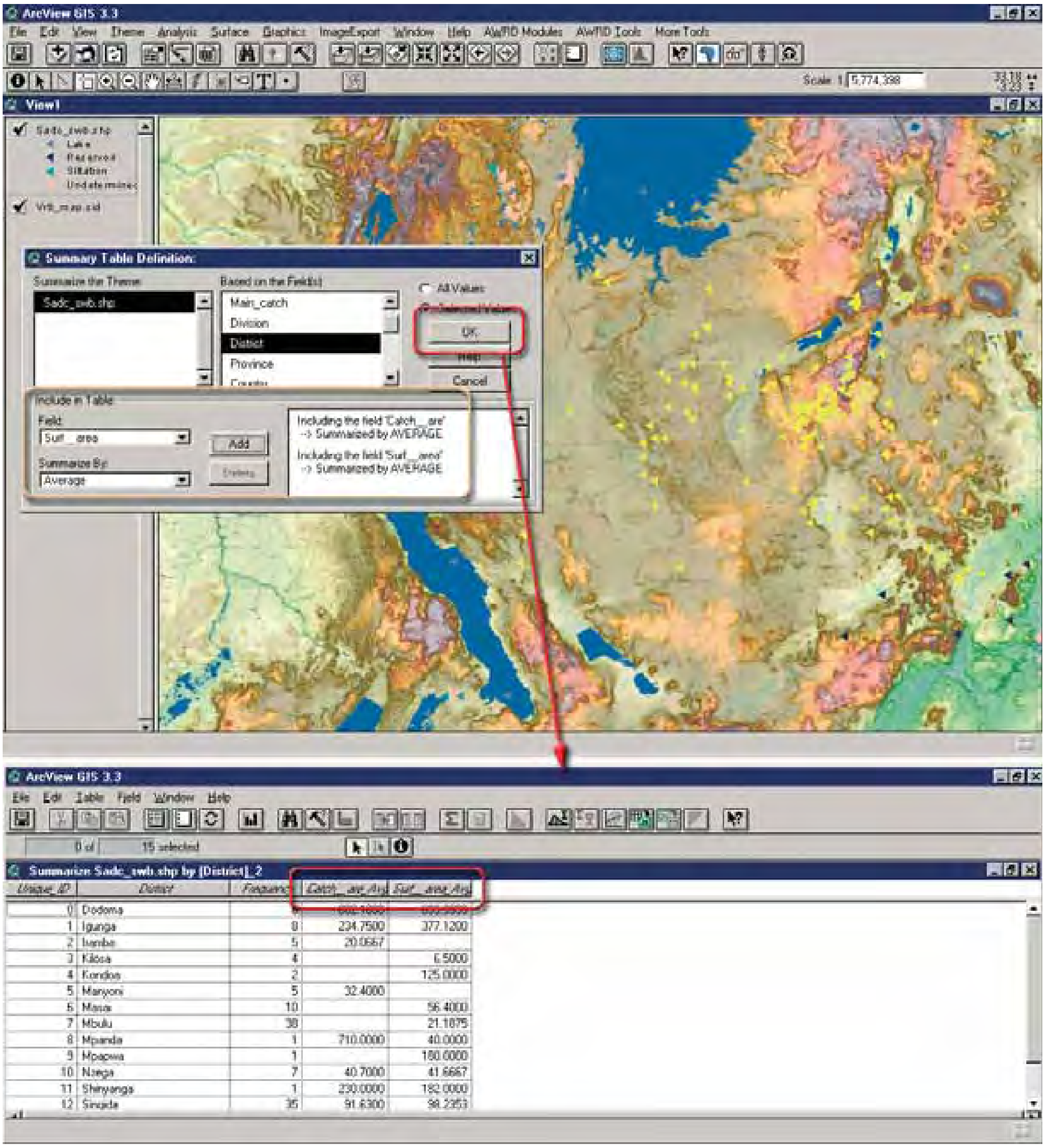


Unique_ID	District	Frequency
0	Dodoma	8
1	Igunga	8
2	Iramba	5
3	Kilosa	4
4	Kondoa	2
5	Manyoni	5
6	Masai	10
7	Mbulu	38
8	Mpanda	1
9	Mpapwa	1
10	Nzega	7

Alternatively, the user could have chosen to calculate some statistics for each subset of data. For example, if the user was interested in the average surface area and the average catchment area of these sets of data, they could have added those statistics to the analysis.

1. Identify the fields to use to separate the data into subsets (e.g. Districts).
2. Select “Surf__area” from the “Field” drop-down list, and then select “Average” from the “Summarize By” drop-down list.
3. Click “Add”
4. Select “Catch__are” from the “Field” drop-down list, and then select “Average” from the “Summarize By” drop-down list.
5. Click “Add”and then OK” (Figure 1.87b).

FIGURE 1.87B
Calculating the surface waterbodies average by Districts



ArcView GIS 3.3
File Edit View Theme Analysis Surface Graphics ImageExport Window Help AWRD Modules AWRD Tools More Tools
Scale 1: 5,774,398
View1
Sadc_swb.shp
Lake
Reservoir
Siltation
Undetermined
Vrlt_map.sid

Summary Table Definition:

Summarize the Theme: Sadc_swb.shp

Based on the Field(s): Main_catch, Division, District, Province, Country

○ All Values
● Selected Values

OK
Help
Cancel

Include in Table
Field: Surf_ area
Summarize By: Average
Add
Delete

Including the field 'Catch_ are'
--> Summarized by AVERAGE
Including the field 'Surf_ area'
--> Summarized by AVERAGE

ArcView GIS 3.3
File Edit Table Field Window Help

0 of 15 selected

Summarize Sadc_swb.shp by [District]_2

Unique_ID	District	Frequency	Catch_ are Avg	Surf_ area Avg
0	Dodoma	[illegible]	[illegible]	[illegible]
1	Iganga	8	234.7500	377.1200
2	Iramba	5	20.0667	
3	Kilosa	4		6.5000
4	Kondoa	2		125.0000
5	Manyoni	5	32.4000	
6	Masai	10		56.4000
7	Mbulu	38		21.1875
8	Mpanda	1	710.0000	40.0000
9	Mpapwa	1		180.0000
10	Nzega	7	40.7000	41.6667
11	Shinyanga	1	230.0000	182.0000
12	Singida	35	91.6300	98.2353

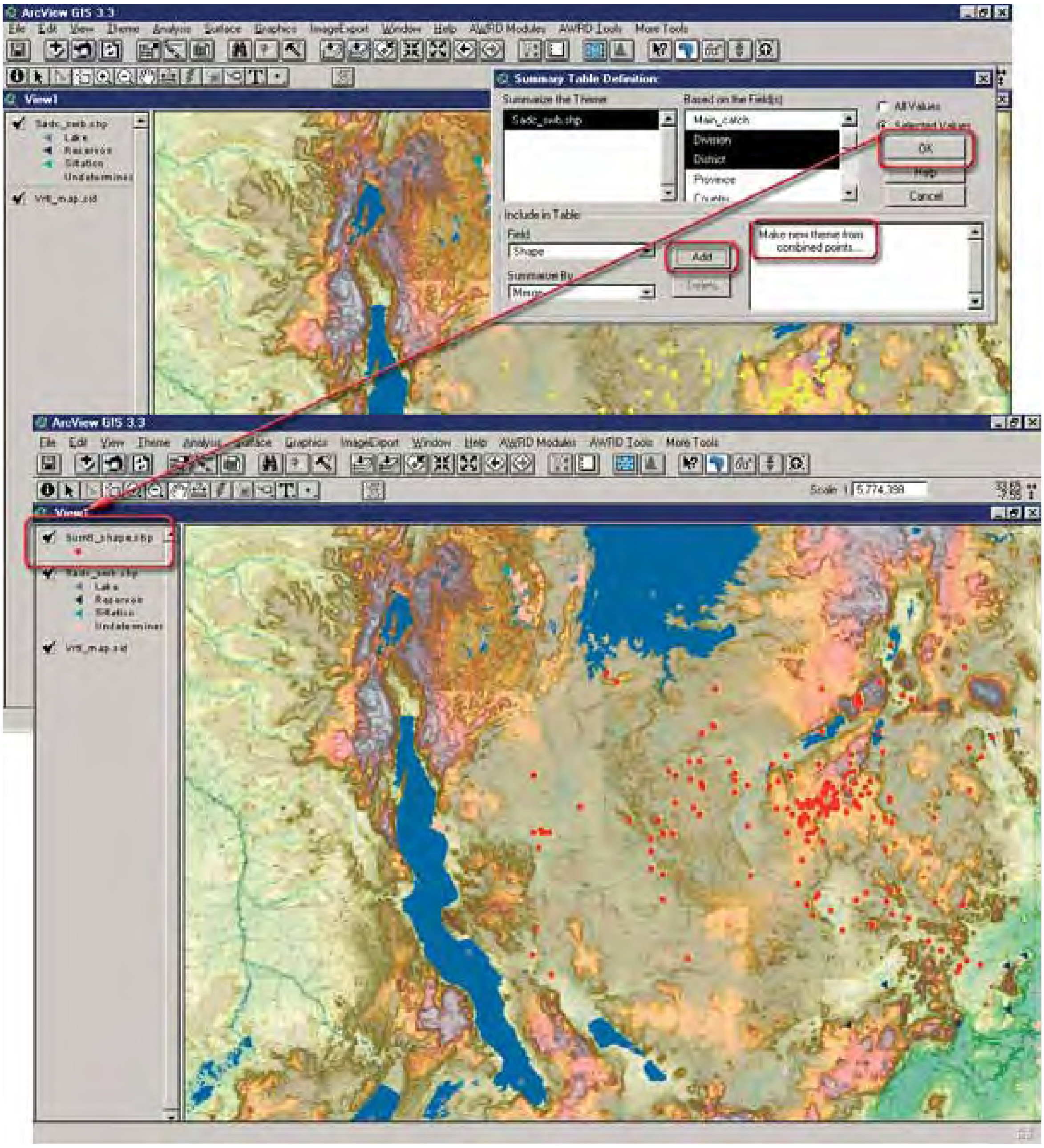
Note In the example above, the empty cell values are due to the fact that none of the records in that particular subset of data had values for Surface Area or Catchment Area.

This tool can be used to subset the data based on more than one field. In the examples above, the data was divided into subsets based on unique District values. If the user wished to divide the data into unique combinations of District and Division, for example, they could simply select both fields.

Furthermore, the user could choose to combine all features from each subset into a single feature and save them into a new theme. Do this by going to the drop-down boxes in the "Include in Table:" section, choosing "Shape" from the "Field:" drop-down box, and "Merge" from the "Summarize By:" box, and clicking the "Add" button.
"

In this case the tool will add a new theme directly to the view, and the attribute table for that theme will contain all the same information as the Summary table does (Figure 1.87c).

FIGURE 1.87C
Saving the combined features into a new theme


ArcView GIS 3.3
File Edit View Theme Analysis Surface Graphics ImageExport Window Help AWRD Modules AWRD Tools More Tools
View1
Summary Table Definition
OK
Help
Cancel
Add
Delete
Make new theme from combined points...
Scale 1: 5,774,398

Summarize the Theme: Sadc_swb.shp

Based on the Field(s): Main_catch, Division, District, Province, County

○ All Values
● Selected Values

Include in Table

Field: Shape

Summarize By: Merge

Sadc_swb.shp
Lake
Reservoir
Siltation
Undetermined
Vril_map.sid

Sum8_shape.shp
Sadc_swb.shp
Lake
Reservoir
Siltation
Undetermined
Vril_map.sid

Classification and ranking tool

The classification and ranking tools provide users with the means to classify features according to a wide variety of simple and complex functions. With these tools, users can rank features based on either single or multiple criteria, as well as identify features that do not meet any selection criteria at all. In addition, users can save specific sets of criteria so that a particular classification scheme can be replicated and modified, and various scenarios matching different criteria can be tested.

This tool is opened by clicking the button on either the AWRD Interface, the Watersheds Module, or by selecting the menu option "*Classify Theme by Multiple Criteria…*" in the View AWRD Modules menu.

When opened from the AWRD Interface or from the AWRD Modules menu, the user will be prompted to identify the theme to be classified and a unique ID field in that theme.

As with a number of the other tools within the AWRD, there are both a "simple" and an "advanced" version of this tool.

Simple version

There are many ways to employ this tool, and the simplest is to use the tool in the default or "Simple" mode by just setting some minimum and maximum value constraints for fields in the attribute table.

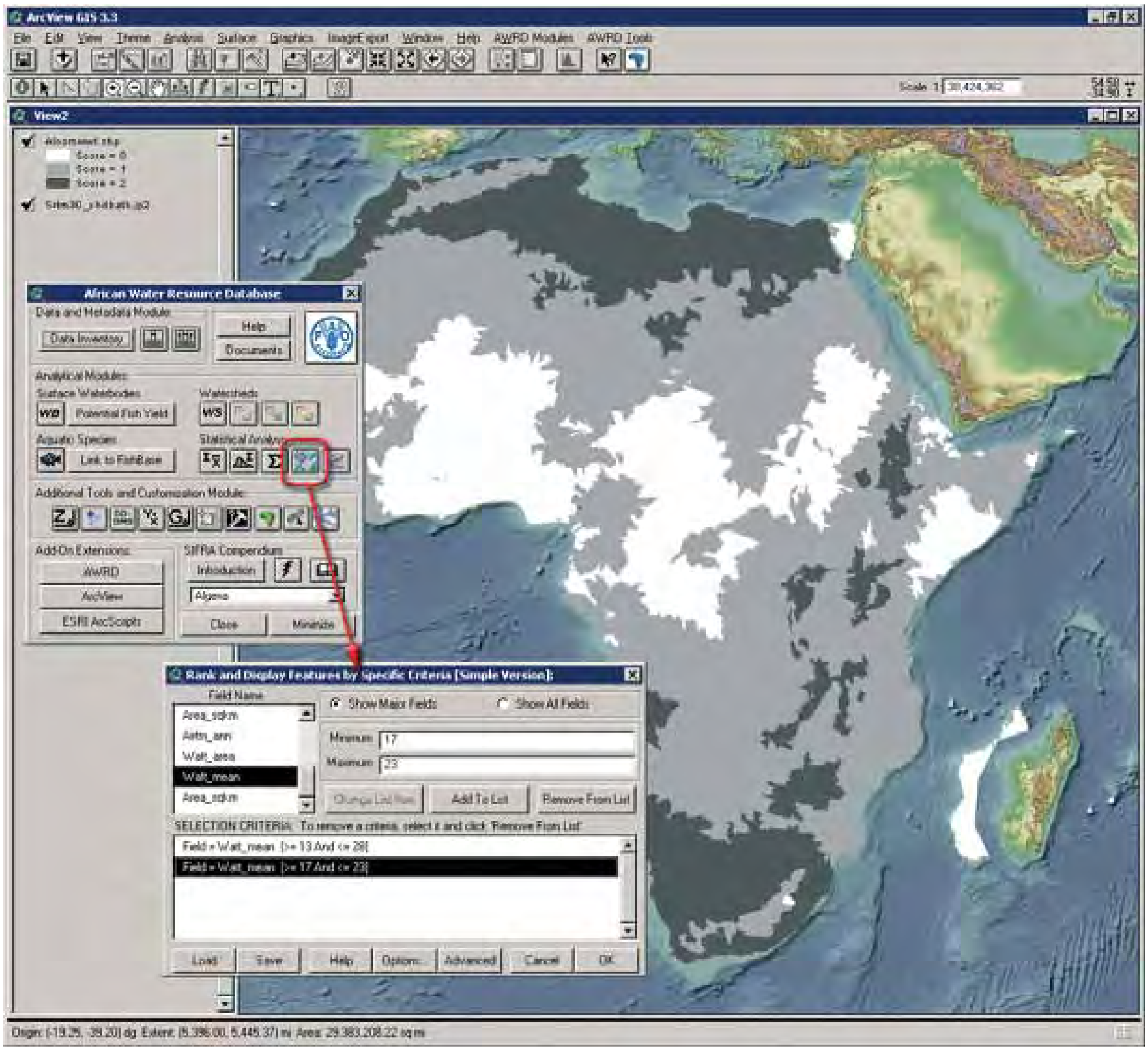
1. Click on the "Add Basemap Image to View" tool to load one of the image backgrounds (e.g. "SRTM30_shdbath.jp2") from the image database component folder. This background image is not necessary for proper functioning of this tool, but it makes it easier to locate your area of interest in the view.
2. For this example, set the default watershed model to "alcomwwf.shp" and make sure that "Mean Annual Water Temperature" is included with the "associated data" (see section 1.4 in this manual).
3. Click the "Open Watershed Viewer" button WS on the main AWRD Interface. This button makes sure that the default watershed model (alcomwwf.shp) is added to the view and that all the associated data tables are joined to the watershed model attribute table. After the watershed model is added to the view, the watershed statistics viewer may be closed.
4. Click the "Classify Theme by Multiple Criteria" button , select "Alcomwwf.shp"as the Theme and "Ws_id"as the ID Field, then click "OK".
5. Select the "Simple" option if is not set as the default option.
6. Choose one of the Field Names (e.g. Watt_mean) and set the Minium and Maximum raking criteria (e.g. 13 °C and 28 °C, representing the lowest and highest temperatures found for carp reproduction to take place in the wild) then Click "Add to list".
7. Enter a second criteria (e.g. 17 °C and 23 °C, representing the range of optimal temperatures found for carp reproduction to take place in the wild) then Click "Add to list".
8. At this point the user may choose to test their existing selection criteria by immediately clicking the "OK" button or they may set a different range for the already selected field, simply by entering the new values and clicking "Add To List" again. This process can be continued until all of the relevant test criteria have been entered and the "OK" button is clicked.

After the "OK" button is clicked, the tool will check all the features in the theme to test if any meet the ranking criteria. If a particular feature has a value within the specified range, it gets a score of 1. If it meets multiple criteria, it gets an additional point for each criteria it meets. After the analysis, all watersheds will have a score representing the number of criteria they met and the tool will display the watersheds coloured according to their score.

In the example below, the tool is used to identify all watersheds with a temperature range between 13 ≤ t ≤ 28 representing the lowest and highest temperatures found for reproduction to take place in the wild.

Based on this criterion, the tool ranked all watersheds according to how many criteria were met. Those that met no criteria were shaded white while those that met all the criteria were shaded a dark grey (Figure 1.88a).

FIGURE 1.88A
Application of the classification and ranking tool (Simple version)



Scale 1: 30,424,362

✓ Alcamanet.shp
□ Score = 0
▨ Score = 1
■ Score = 2
✓ Srtm30_ybathy.jp2

Algeria

● Show Major Fields ○ Show All Fields

Minimum 17
Maximum 23

SELECTION CRITERIA: To remove a criteria, select it and click 'Remove From List'

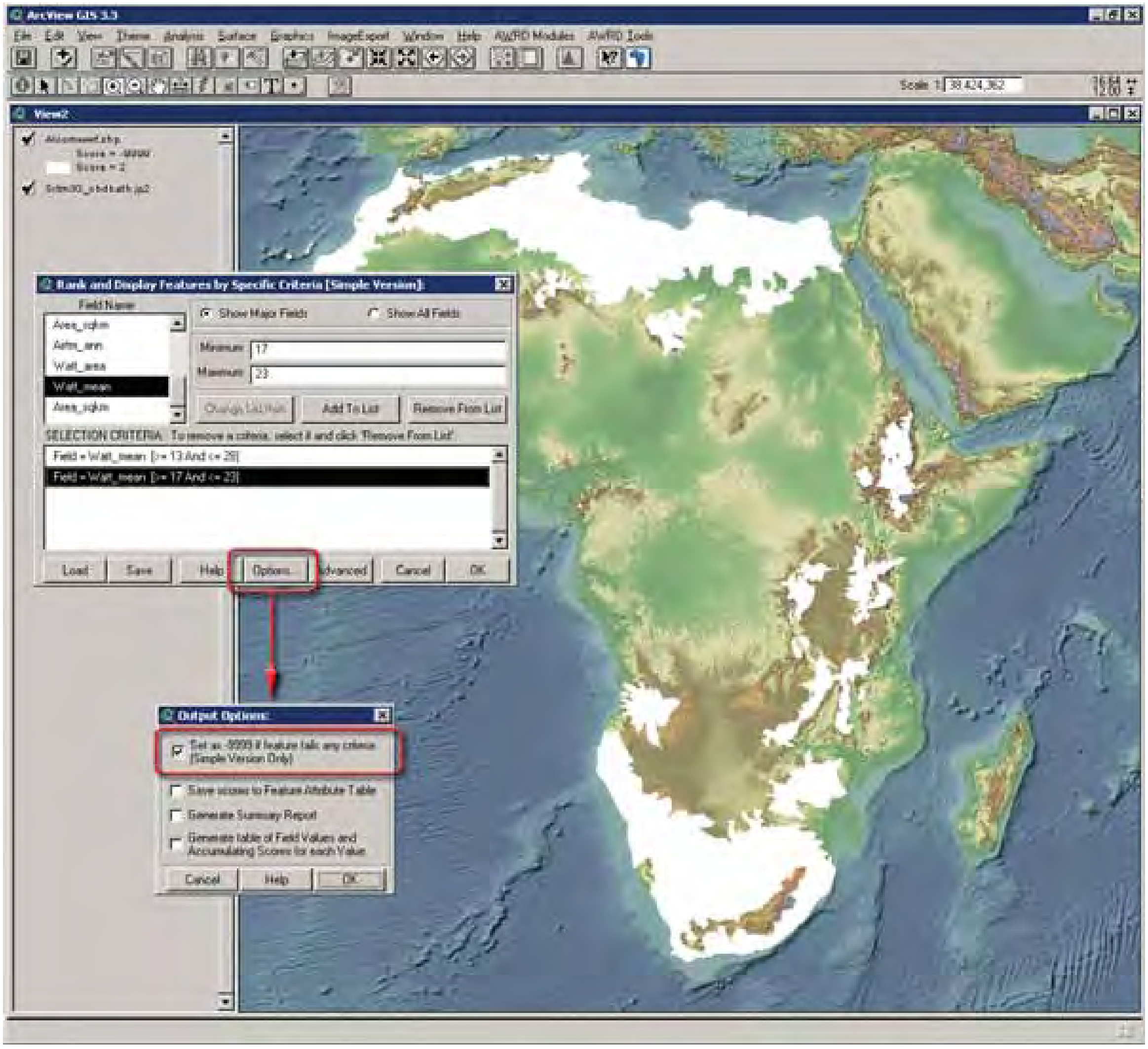
Field = Watt_mean [>= 13 And <= 28]
Field = Watt_mean [>= 17 And <= 23]

Origin: (-19.25, -39.20) dg Extent: (5,396.00, 5,445.37) mi Area: 29,383,208.22 sq mi

Alternatively, the user could have run the tool so that the watershed would be assigned a value of -9999 if any of the criteria were not met. In this case, the tool would still add up the scores, but the final classification would only differentiate between those that met all criteria (with a score of 2 in this case, coloured white), or those that failed at least one criteria (with a score of -9999, coloured transparent) (Figure 1.88b).

FIGURE 1.88B

Setting different output options with the simple classification and ranking tool



● Show Major Fields ○ Show All Fields

Minimum: 17

Maximum: 23

SELECTION CRITERIA: To remove a criteria, select it and click 'Remove From List'

Field = Watt_mean: [>= 13 And <= 29]

Field = Watt_mean: [>= 17 And <= 23]

☑ Set as -9999 if feature fails any criteria (Simple Version Only)

□ Save scores to Feature Attribute Table

□ Generate Summary Report

□ Generate table of Field Values and Accumulating Scores for each Value

Advanced version

With the advanced version, users can construct more sophisticated criteria sets which may then be used to rank different features.

This advanced version offers much more power in terms of building cumulative scores. The Criteria may use the full set of comparison operators (<, ≤, =, ≥, > and <>), and users are not limited to specifying a minimum and maximum range but may choose to skip the second criteria.

In addition, the criteria may be applied to both numbers and strings, and can include attribute fields containing words, names, or alphanumeric codes. Perhaps more importantly however, users can also choose to set weights for criteria, critical knock-out constraints, and exclusion logic.

- Weight: you may assign a particular criterion a specific weight so that it scores higher if it is more important than other criteria. You may also enter negative values for the weight if a criterion detracts from the overall ranking. The "Weight" value corresponds to the score that will be added or subtracted from the final score based on whether it meets the criteria or not. The *Simple* version of this tool applies a constant weight of 1 to all criteria.

- Critical: a “Critical” criteria is one that must be met in order for the feature to get any score at all. Here users have the option to define any criteria to be critical to the classification, such that this criterion must be met or else the final classification will automatically be set to -9999. For example, if one particular criterion was that the *mean elevation* should be greater than 1000 meters, and a particular watershed had a mean elevation of 500 meters, then a *Critical* setting would mean that this watershed would get a score of -9999 no matter how well it met any of the other criteria. A *Non-critical* setting would mean that this watershed’s score would simply not be changed based on mean elevation. The *Simple* version of this tool treats all criteria as either critical or non-critical, depending on the user’s choice in the “Options” dialog.
- Include vs. Exclude: here users have the option to decide whether to apply the score if the feature value is within the specified range (“Include”) or if it is outside the specified range (“Exclude”). For example, if the first criterion was that *mean elevation > 1 000* and the second criteria was that *mean elevation < 2 000*, then the “Include” option would cause the score to be applied if the feature elevation was between 1 000 and 2 000 meters. The “Exclude” option would cause the score to be applied if the feature elevation was either < 1 000 meters or > 2 000 meters. The *Simple* version of this tool treats all criteria as using the “Include” option.

The following example illustrates an attempt to identify potential fish farming areas for common carp using annual water temperature and water requirements for pond construction. These pond construction requirements are based on annual precipitation, potential evapotranspiration and an adjustment factor for ground seepage. An overall water requirement index was calculated as:

Water requirement = (Precip. [mm] x 1.1) - (Pot. Evap [mm] x 1.3) – Seepage ([80 mm/mo])

Based on criteria developed by FAO in *A Strategic Reassessment of Fish Farming Potential in Africa* (Aguilar-Manjarrez and Nath, 1998) we classified the landscape into 4 categories based on these water requirement values. These categories are defined in Table 1.29.

TABLE 1.29
Water Requirement Submodel

Category	Water Requirement	Interpretation
1	< -3 500 mm	Unsuitable – Many problems
2	-3 500 to -2 000 mm	Very likely to encounter water availability problems.
3	-2 000 to 0 mm	Moderately suitable for ponds
4	> 0 mm	Very suitable as a water source for ponds.

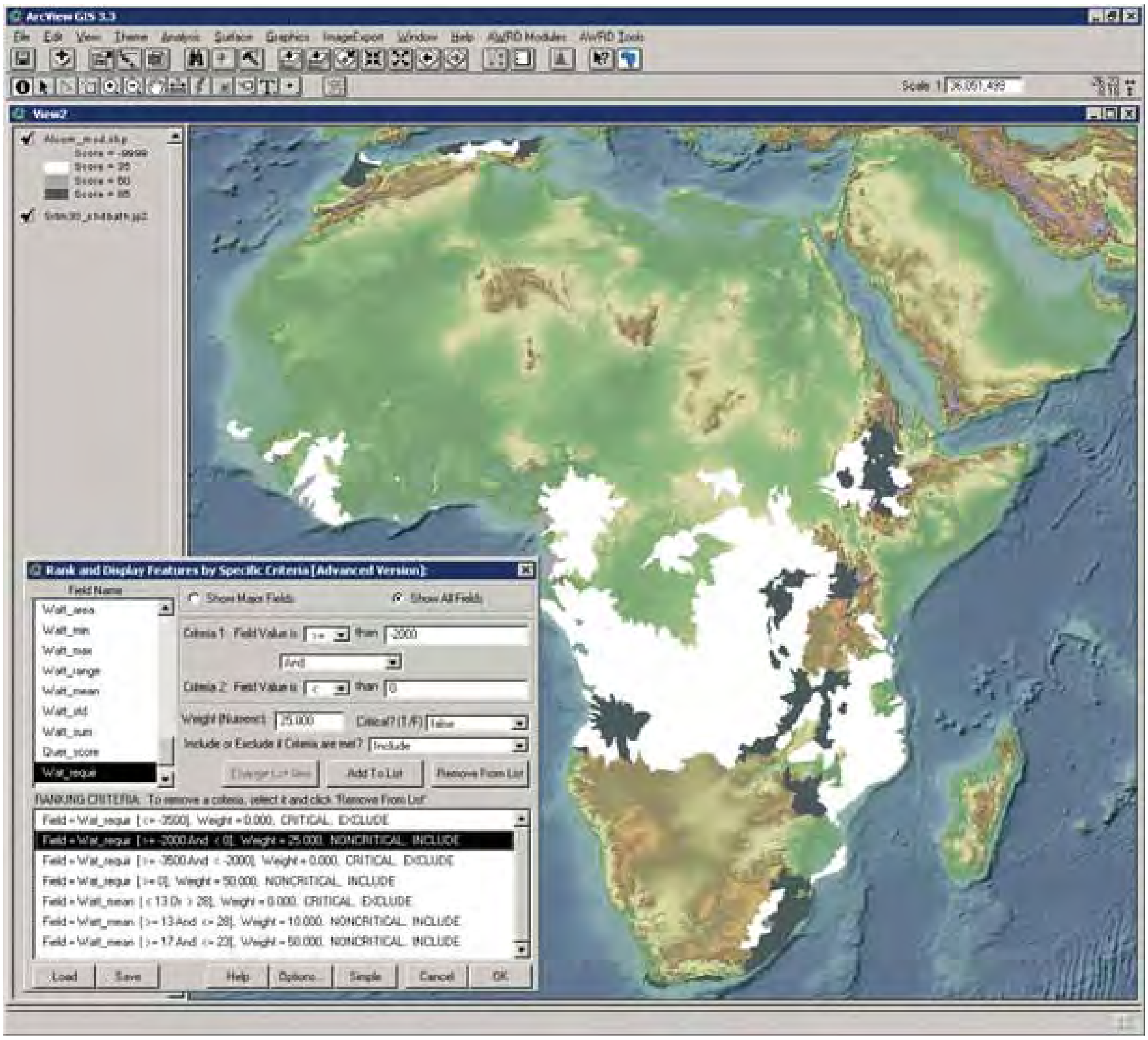
Source: Adapted from Aguilar-Manjarrez and Nath (1998)

Based on these guidelines weighting scores are assigned to the watersheds according to the following criteria to find the optimal locations for carp farming:

Mean annual water temperature (°C): < 13 and > 28, exclude critical
Mean annual water temperature (°C): 13 =< t =< 28, add 10
Mean annual water temperature (°C): 17 =< t =< 23; add 40
Water requirement category 1; exclude critical
Water requirement category 2; exclude critical
Water requirement category 3; add 25
Water requirement category 4; add 50

The watersheds that met those two criteria received scores based on how well they met the other criteria, with potential final scores ranging from 0 to 100 (50 points from the mean annual water temperature values and 50 points from the water requirement categories). Also, notice that the tool is resizable so that a user can read the entire criteria description by simply clicking on one of the corners and dragging it to a new size (Figure 1.89).

FIGURE 1.89
Application of the classification and ranking tool (Advanced option)



○ Show Major Fields ● Show All Fields

Criteria 1: Field Value is >= than -2000

And

Criteria 2: Field Value is < than 0

Weight (Numeric): 25.000 Critical? (T/F) False

Include or Exclude if Criteria are met? Include

RANKING CRITERIA: To remove a criteria, select it and click 'Remove From List'

Field = Wat_requi [<= -3500], Weight = 0.000, CRITICAL, EXCLUDE
Field = Wat_requi [>= -2000 And < 0], Weight = 25.000, NONCRITICAL, INCLUDE
Field = Wat_requi [>= -3500 And < -2000], Weight = 0.000, CRITICAL, EXCLUDE
Field = Wat_requi [>= 0], Weight = 50.000, NONCRITICAL, INCLUDE
Field = Watt_mean [< 13 Or > 28], Weight = 0.000, CRITICAL, EXCLUDE
Field = Watt_mean [>= 13 And <= 28], Weight = 10.000, NONCRITICAL, INCLUDE
Field = Watt_mean [>= 17 And <= 23], Weight = 50.000, NONCRITICAL, INCLUDE

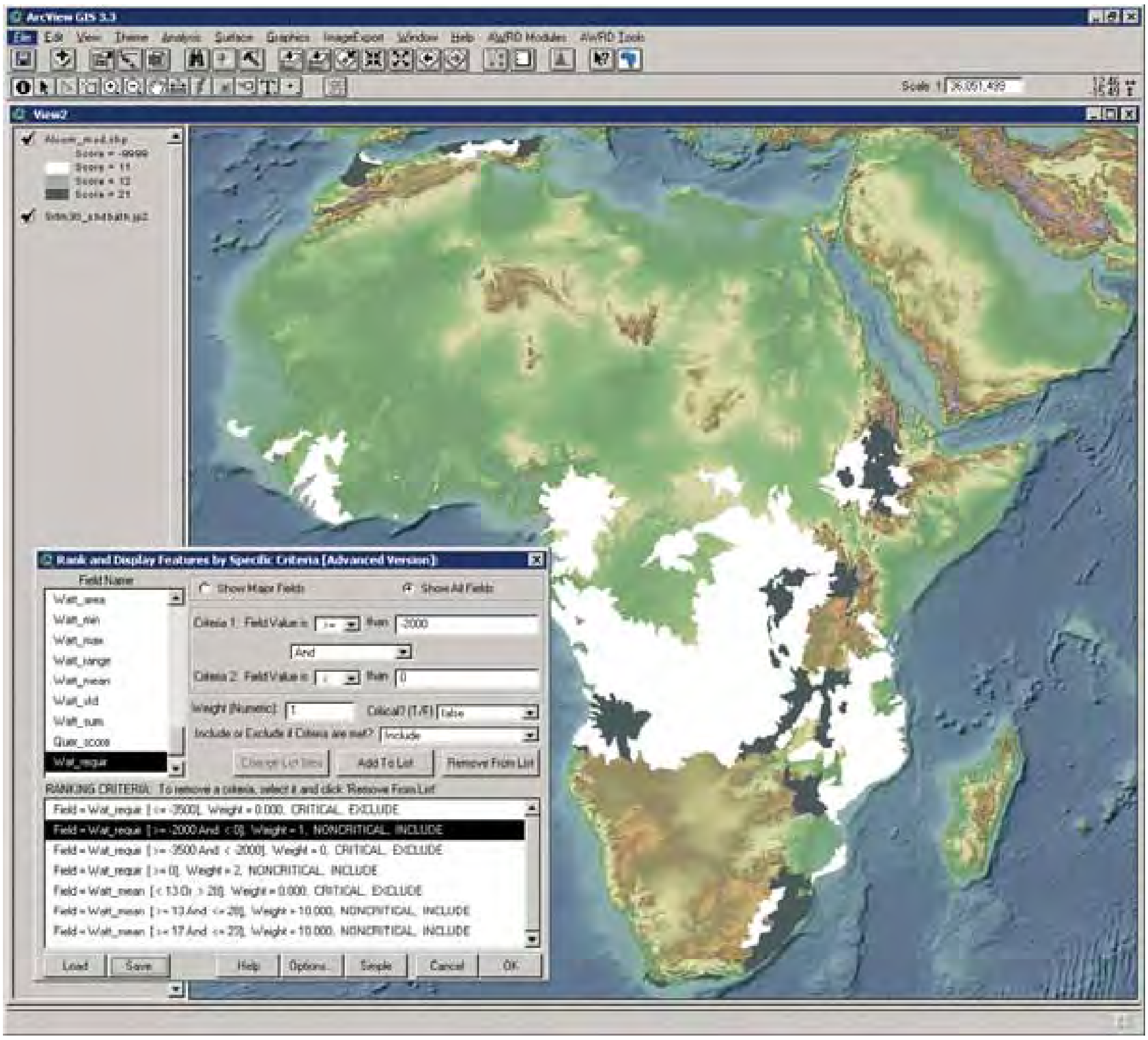
Classification and weighting strategies

Ranking strategy: This discussion naturally raises the issue of how to set the weighting scores, and there are several strategies that might be appropriate in different situations. In the example above, the maximum attainable score was set at 100. Both *water requirements* and *mean annual water temperature* each contributed up to a maximum of 50 points, implying that both categories are equally important. These weights are a matter of expert opinion and this tool gives users a way to use their own expert opinion. It also gives users the opportunity to quickly identify potentially important areas based on different classification schemes and to easily try a variety of weighting options.

Pseudo-classification strategy: another possible classification strategy would be to use the scores to produce a kind of ranked pseudo-classification of watersheds, by assigning scores with different numbers of digits to different variables. In this example (Figure 1.90), watersheds are classified according to *Water Requirements* and *Mean Annual Water Temperature* as was done in the previous example, but this time the weighting factors are set as follows:

Mean annual water temperature (°C): < 13 and > 28, exclude critical
Mean annual water temperature (°C): 13 =< t =< 28, add 10
Mean annual water temperature (°C): 17 =< t =< 23; add another 10
Water requirement category 1; exclude critical
Water requirement category 2; exclude critical
Water requirement category 3; add 1
Water requirement category 4; add 2

FIGURE 1.90
Application of the advanced classification and ranking tool to determine potential fish farming areas for common carp



Rank and Display Features by Specific Criteria [Advanced Version]

Field Name: Wat_area, Wat_min, Wat_max, Wat_range, Wat_mean, Wat_std, Wat_sum, Quer_score, Wat_requi

○ Show Major Fields ● Show All Fields

Criteria 1: Field Value is >= than -2000

And

Criteria 2: Field Value is < than 0

Weight (Numeric): 1 Critical? (T/F) false

Include or Exclude if Criteria are met? Include

RANKING CRITERIA: To remove a criteria, select it and click 'Remove From List'

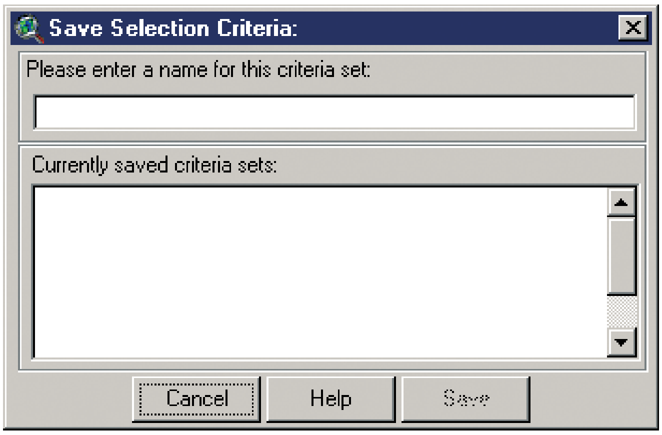
```
Field = Wat_requi [ <= -3500], Weight = 0.000, CRITICAL, EXCLUDE
Field = Wat_requi [ >= -2000 And < 0], Weight = 1, NONCRITICAL, INCLUDE
Field = Wat_requi [ >= -3500 And < -2000], Weight = 0, CRITICAL, EXCLUDE
Field = Wat_requi [ >= 0], Weight = 2, NONCRITICAL, INCLUDE
Field = Watt_mean [ < 13 Or > 28], Weight = 0.000, CRITICAL, EXCLUDE
Field = Watt_mean [ >= 13 And <= 28], Weight = 10.000, NONCRITICAL, INCLUDE
Field = Watt_mean [ >= 17 And <= 23], Weight = 10.000, NONCRITICAL, INCLUDE
```

Using this pseudo-classification strategy, all watershed scores end up having two digits. The first digit ranks the watersheds by *mean annual water temperature* and the second digit ranks the watersheds by *water requirement*.

Saving and loading criteria

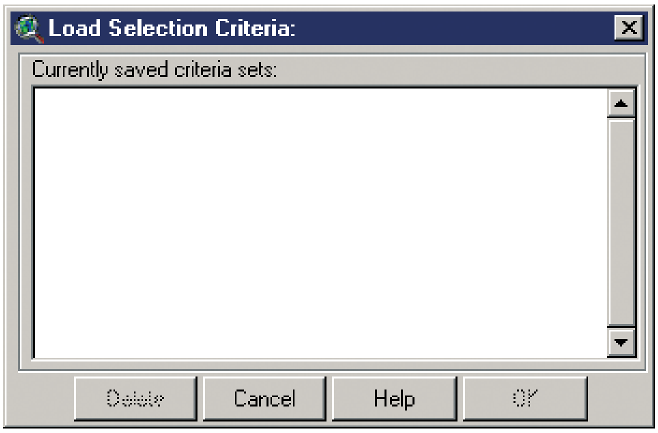
As was mentioned earlier, users of the ranking and classification tools can also save and restore lists of ranking criteria established during previous sessions, making it much easier to run analyses multiple times with small variations, run the same set of criteria on multiple data sets, and, perhaps more importantly, to share their criteria amongst colleagues in different offices or locations. After building a list of selection criteria, simply click on the Save button from the "Rank and Display Features by Specific Criteria" dialog, specify a name to identify the classification criteria set, and click "OK" (Figure 1.91).

FIGURE 1.91
Saving list of ranking criteria ("Save" button)


Save Selection Criteria:
Please enter a name for this criteria set:
Currently saved criteria sets:
Cancel
Help
Save

Users can also choose to save their "current" criteria set using the same name as a previous-saved set, in which case the newer set will replace the older set. Saved selection criteria sets may be restored by clicking on the Load button (Figure 1.92).

FIGURE 1.92
Loading saved selection criteria sets ("Load" button)



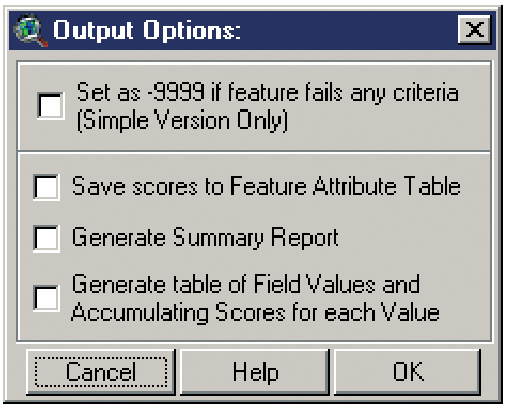
The "Load Selection Criteria" dialog also provides the option to delete any of the saved selection sets.

Classification options

Users are also provided with multiple options concerning how to save their classification output. Clicking on the Options... button brings up the following dialog (Figure 1.93).

FIGURE 1.93

Multiple Output Options for saving selection criteria sets



Output Options:

□ Set as -9999 if feature fails any criteria (Simple Version Only)

□ Save scores to Feature Attribute Table

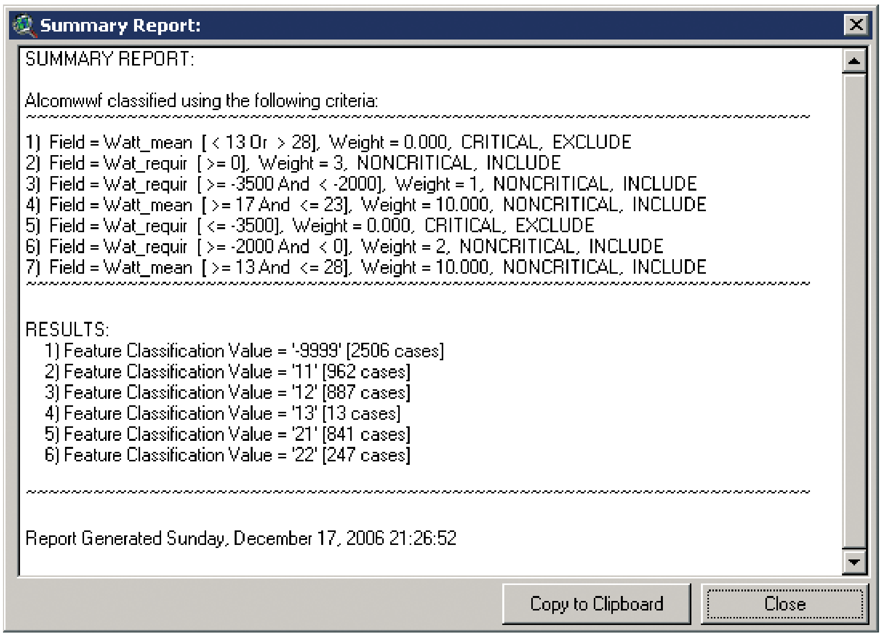
□ Generate Summary Report

□ Generate table of Field Values and Accumulating Scores for each Value

- Set as -9999 if feature fails any criteria (Simple Version Only): this option only applies to the Simple version of the dialog and will not be available if the user is using the advanced version. If this option is set, then any feature that failed any of the criteria will be assigned a value of -9999 and set transparent in the view. If this option is not set, then the features will be assigned a value equal the number of criteria that were met and coloured accordingly.
- Save Scores to Feature Attribute Table: this option adds a new field to the theme's feature attribute table and saves any scores to that field. This is a useful option if it is desirable to run a series of classification schemes against a particular theme and compare the various classifications afterwards. If this option is not selected, the tool makes no changes to the original feature attribute table. Rather, it makes a new temporary table containing the theme's ID values and the classification scores, and then joins that temporary table to the theme feature attribute table. This temporary table is destroyed and recreated every time the tool is run.
- Generate Summary Report: this option gives a breakdown of how the features were classified. Using the Pseudo-Classification example above, the Summary Report looks like the one in Figure 1.94.

FIGURE 1.94

Generating a Summary Report about the classification criteria



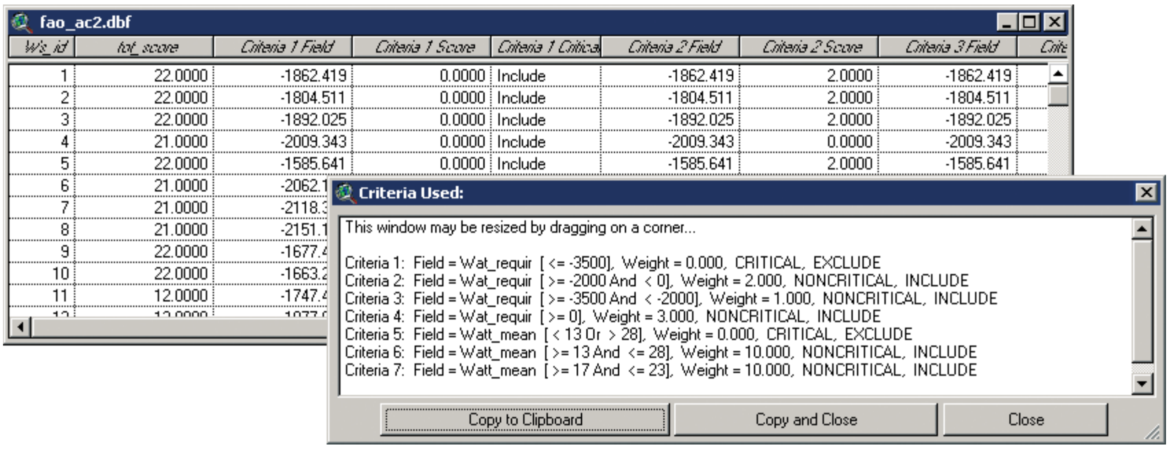
Summary Report:

```
SUMMARY REPORT:

Alcomwwf classified using the following criteria:
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
1) Field = Watt_mean  [ < 13 Or  > 28],  Weight = 0.000,  CRITICAL,  EXCLUDE
2) Field = Wat_requir  [ >= 0],  Weight = 3,  NONCRITICAL,  INCLUDE
3) Field = Wat_requir  [ >= -3500 And  < -2000],  Weight = 1,  NONCRITICAL,  INCLUDE
4) Field = Watt_mean  [ >= 17 And  <= 23],  Weight = 10.000,  NONCRITICAL,  INCLUDE
5) Field = Wat_requir  [ <= -3500],  Weight = 0.000,  CRITICAL,  EXCLUDE
6) Field = Wat_requir  [ >= -2000 And  < 0],  Weight = 2,  NONCRITICAL,  INCLUDE
7) Field = Watt_mean  [ >= 13 And  <= 28],  Weight = 10.000,  NONCRITICAL,  INCLUDE
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

RESULTS:
   1) Feature Classification Value = '-9999' [2506 cases]
   2) Feature Classification Value = '11' [962 cases]
   3) Feature Classification Value = '12' [887 cases]
   4) Feature Classification Value = '13' [13 cases]
   5) Feature Classification Value = '21' [841 cases]
   6) Feature Classification Value = '22' [247 cases]

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Report Generated Sunday, December 17, 2006 21:26:52
```

- Generate Table of Field Values and Accumulating Scores for each Value: this option creates a new table in active view that illustrates exactly how each score was established for each record (Figure 1.95).

FIGURE 1.95

Creating the new table illustrating how each score was established



fao_ac2.dbf

Ws_id	tot_score	Criteria 1 Field	Criteria 1 Score	Criteria 1 Critica	Criteria 2 Field	Criteria 2 Score	Criteria 3 Field	Crite
1	22.0000	-1862.419	0.0000	Include	-1862.419	2.0000	-1862.419	
2	22.0000	-1804.511	0.0000	Include	-1804.511	2.0000	-1804.511	
3	22.0000	-1892.025	0.0000	Include	-1892.025	2.0000	-1892.025	
4	21.0000	-2009.343	0.0000	Include	-2009.343	0.0000	-2009.343	
5	22.0000	-1585.641	0.0000	Include	-1585.641	2.0000	-1585.641	
6	21.0000	-2062.1						
7	21.0000	-2118.3						
8	21.0000	-2151.1						
9	22.0000	-1677.4						
10	22.0000	-1663.2						
11	12.0000	-1747.4						

Criteria Used:

```
This window may be resized by dragging on a corner...

Criteria 1:  Field = Wat_requir  [ <= -3500],  Weight = 0.000,  CRITICAL,  EXCLUDE
Criteria 2:  Field = Wat_requir  [ >= -2000 And  < 0],  Weight = 2.000,  NONCRITICAL,  INCLUDE
Criteria 3:  Field = Wat_requir  [ >= -3500 And  < -2000],  Weight = 1.000,  NONCRITICAL,  INCLUDE
Criteria 4:  Field = Wat_requir  [ >= 0],  Weight = 3.000,  NONCRITICAL,  INCLUDE
Criteria 5:  Field = Watt_mean  [ < 13 Or  > 28],  Weight = 0.000,  CRITICAL,  EXCLUDE
Criteria 6:  Field = Watt_mean  [ >= 13 And  <= 28],  Weight = 10.000,  NONCRITICAL,  INCLUDE
Criteria 7:  Field = Watt_mean  [ >= 17 And  <= 23],  Weight = 10.000,  NONCRITICAL,  INCLUDE
```

Simple linear regression tool

The Simple Linear Regression tool of the AWRD provides users with a powerful method for analysing relationships between data. The tool has been specifically designed to allow a user to conduct simple linear regression analyses between pairs of variables, where the first variable is considered the “Independent” or “Predictor” variable, and the second variable is considered the “Dependent” or “Response” variable. This type of regression lets the user identify whether a dependent variable varies in a predictable way over different levels of the independent variable. In addition, the analysis also provides users with a probability that any relationship established is due solely to chance.

The Simple Linear Regression tool is opened by clicking on the button on either the AWRD Interface, the Table button bar or on the Watersheds Module, or by clicking the “*Simple Linear Regression…*” menu option in the AWRD Modules menu.

Regression options

The regression options dialog allows users to conduct simple linear regression analyses between two numeric fields in a table, and examine the values in these fields for any correlation. Similar to the other statistical tools available through the AWRD, the simple linear regression tool will either analyse all of the records in a table or only those records that lie within the currently selected set. Output options associated with this tool are extensive and include: the basic R^2; an ANOVA table, where both the F-values and P-values are presented; Residuals, including both basic and standardized; a large variety of descriptive statistics; the slope variability; a description of the confidence level; predicted values; and lastly, a scatterplot.

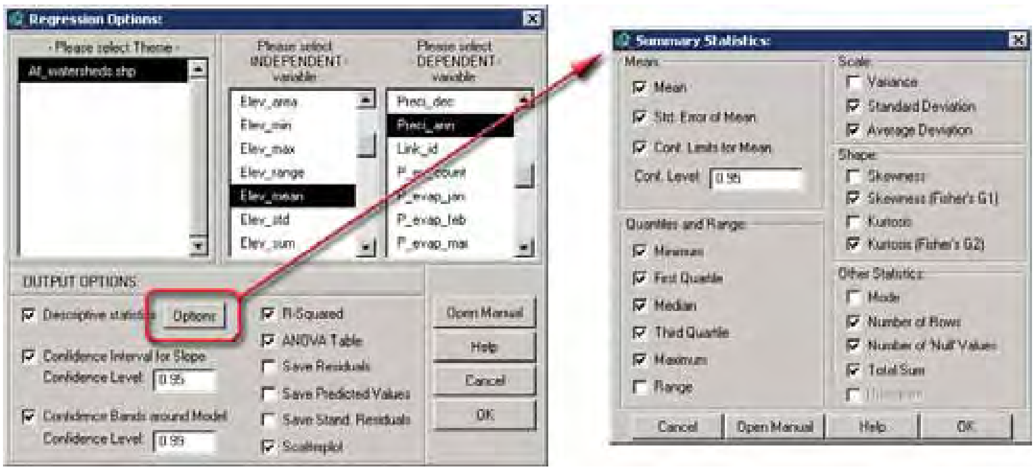
Despite the variable names (“Independent” and “Dependent”), regression analysis is not intended to demonstrate causative relationships between the dependent and predictor variables. Just because the dependent variable varies in a predictable way over different levels of the independent variable does not mean the predictor variable necessarily causes the variation. It is possible that both variables may actually be influenced by some other variable or variables, and therefore both the dependent and independent variables fluctuate in response to those other factors. True causal relationships can only be established through controlled experiments where the causal relationship is being specifically tested and measured. However, this does not diminish the value of the correlational relationships that can be identified using regression.

Output options

Descriptive statistics

The general descriptive statistics are available by selecting the “Descriptive Statistics” box and then clicking the “Options” button to open the “Summary Statistics:” dialog. The summary statistics calculator dialog (Figure 1.96) provides a wide range of statistical output associated with simple linear regression, as well as general descriptive statistics on the dependent and independent variables. The “Histogram” function is disabled for regression analyses.

FIGURE 1.96
The Summary Statistic Calculator



Theme: Al_watersheds.shp
INDEPENDENT variable: Elev_mean
DEPENDENT variable: Preci_ann

☑ Descriptive statistics
☑ Confidence Interval for Slope
Confidence Level: 0.95
☑ Confidence Bands around Model
Confidence Level: 0.95
☑ R-Squared
☑ ANOVA Table
□ Save Residuals
□ Save Predicted Values
□ Save Stand. Residuals
☑ Scatterplot

☑ Mean
☑ Std. Error of Mean
☑ Conf. Limits for Mean
Conf. Level: 0.95
☑ Minimum
☑ First Quartile
☑ Median
☑ Third Quartile
☑ Maximum
□ Range
□ Variance
☑ Standard Deviation
☑ Average Deviation
□ Skewness
☑ Skewness (Fisher's G1)
□ Kurtosis
☑ Kurtosis (Fisher's G2)
□ Mode
☑ Number of Rows
☑ Number of 'Null' Values
☑ Total Sum
□ Histogram

Other output options

In all cases, the simple linear regression tool will produce a regression report detailing all the output options that were selected in the "Summary Statistics:" dialog. This report will automatically be saved to the hard drive and opened in a text report for the user to review.

Optionally, a user may also choose to generate a scatterplot illustrating the regression relationship. If the user elects to generate confidence bands, predicted values, residuals or standardized residuals, then the tool will also produce a new theme in the active view which will be identical to the input theme except that it will also contain fields for these additional values. The name of this new theme will be the same as the input theme, appended by "_regress". The input theme will never be altered by using this tool, and any predicted values, residuals or confidence levels for the selected set will be added to this new "regression" theme.

The various output options of this tool can be illustrated by running a sample regression analysis investigating a potential relationship between *Elevation* [Elev_mean] and *Mean Annual Precipitation* [Aprec_mean] for all watersheds in Africa. A scatterplot will also be generated, as will the full range of output options by setting up the simple linear regression tool as follows (Figure 1.97b).

1. Click on the "Add Basemap Image to View" tool to load one of the image backgrounds (e.g. "SRTM30_shdbath.jp2") from the image database component folder. This background image is not necessary for proper functioning of this tool, but it makes it easier to locate your area of interest in the view.

2. For this example, set the default watershed model to "H1k_lev6.shp" and make sure that "Elevation" and "Mean Annual Precipitation" are included with the "associated data" (see section 1.4 in this manual).

3. Click the "Open Watershed Viewer" button WS on the main AWRD Interface. This button makes sure that the default watershed model (H1k_lev6.shp) is added to the view and that all the associated data tables are joined to the watershed model attribute table. After the watershed model is added to the view, the watershed statistics viewer may be closed.

4. Open the Simple Linear Regression tool by clicking on the button on the AWRD Interface.

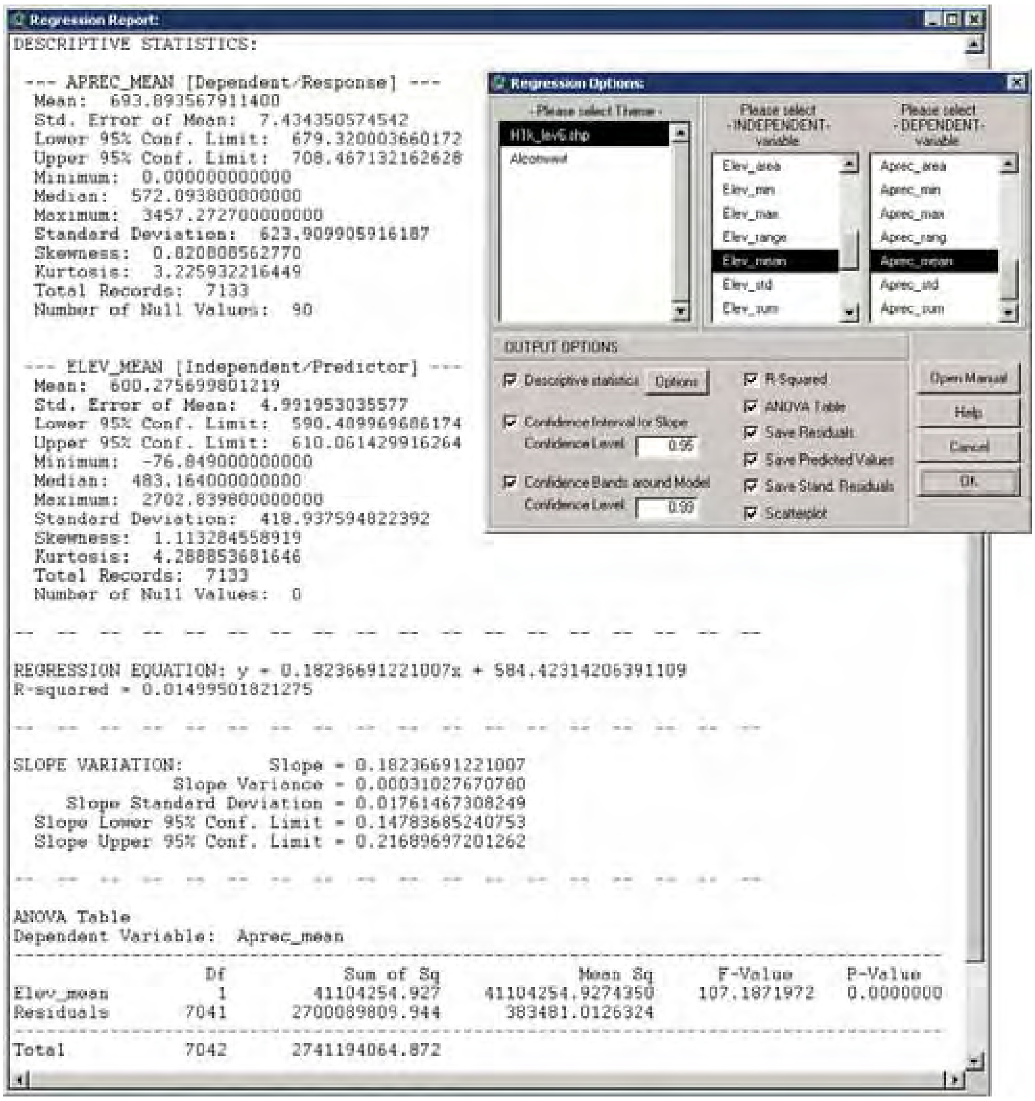
5. Select elevation (i.e. Elev_mean) as the independent variable and Annual precipitation (i.e. Aprec_mean) as the dependent variable.

6. Tick all the regression options. Write 0.95 as the confidence interval for slope and 0.99 confidence bands around Model. Then click "OK".

As soon as the tool finishes processing the data requested, the new "regression" theme will be added to the view, a scatterplot will open, and a regression report will appear (Figure 1.97a).

FIGURE 1.97A

Regression report of rainfall and elevation for all watersheds in Africa



Regression Report:

```
DESCRIPTIVE STATISTICS:

 --- APREC_MEAN [Dependent/Response] ---
  Mean:  693.893567911400
  Std. Error of Mean:  7.434350574542
  Lower 95% Conf. Limit:  679.320003660172
  Upper 95% Conf. Limit:  708.467132162628
  Minimum:  0.000000000000
  Median:  572.093800000000
  Maximum:  3457.272700000000
  Standard Deviation:  623.909905916187
  Skewness:  0.820808562770
  Kurtosis:  3.225932216449
  Total Records:  7133
  Number of Null Values:  90


 --- ELEV_MEAN [Independent/Predictor] ---
  Mean:  600.275699801219
  Std. Error of Mean:  4.991953035577
  Lower 95% Conf. Limit:  590.489969686174
  Upper 95% Conf. Limit:  610.061429916264
  Minimum:  -76.849000000000
  Median:  483.164000000000
  Maximum:  2702.839800000000
  Standard Deviation:  418.937594822392
  Skewness:  1.113284558919
  Kurtosis:  4.288853681646
  Total Records:  7133
  Number of Null Values:  0

--  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --

REGRESSION EQUATION: y = 0.18236691221007x + 584.42314206391109
R-squared = 0.01499501821275

--  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --

SLOPE VARIATION:         Slope = 0.18236691221007
               Slope Variance = 0.00031027670780
     Slope Standard Deviation = 0.01761467308249
  Slope Lower 95% Conf. Limit = 0.14783685240753
  Slope Upper 95% Conf. Limit = 0.21689697201262

--  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --  --

ANOVA Table
Dependent Variable:  Aprec_mean
------------------------------------------------------------------------------------------
                   Df          Sum of Sq               Mean Sq        F-Value      P-Value
Elev_mean           1       41104254.927      41104254.9274350    107.1871972    0.0000000
Residuals        7041     2700089809.944        383481.0126324
------------------------------------------------------------------------------------------
Total            7042     2741194064.872
```

Regression Options:

- Please select Theme -: H1k_lev6.shp

Please select - INDEPENDENT - variable: Elev_mean

Please select - DEPENDENT - variable: Aprec_mean

OUTPUT OPTIONS

☑ Descriptive statistics

☑ Confidence Interval for Slope — Confidence Level: 0.95

☑ Confidence Bands around Model — Confidence Level: 0.99

☑ R-Squared

☑ ANOVA Table

☑ Save Residuals

☑ Save Predicted Values

☑ Save Stand. Residuals

☑ Scatterplot

In all cases the regression report will include the regression equation. Using the basic equation of a line; [y = mx + b], "y" is the dependent variable, "x" is the independent variable and "m" is the slope of the regression line. This example may be interpreted to read that the dependent variable tends to increase by 0.182 units for every single unit increase in the independent variable. The entire regression equation can be used to predict new values of *y* based on values of *x*. For example, a user might be interested in predicting what the annual precipitation might be if an area was located at an elevation of 400 m. Based on this equation, the annual precipitation could be estimated as:

- Confidence Interval for Slope: this option gives a variety of statistics regarding the slope of the regression line, including several measures of the variability and uncertainty of the slope estimate. Selecting this option will provide the user with values for the slope, the slope variance, the slope standard deviation, and the upper and lower confidence limits for the confidence level specified.

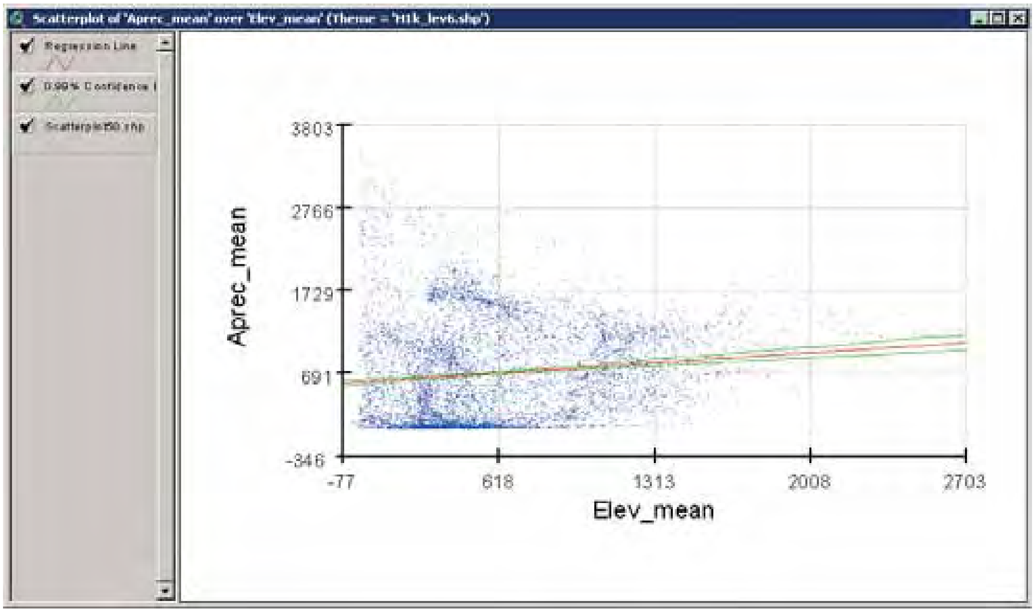
  One important use of these statistics is to confirm whether there really is a relationship between the independent and dependent variables. If the slope were equal to 0 (i.e. perfectly flat), then the dependent variable would not change at all as the independent variable changed and therefore there would be no relationship between them.

  This slope is considered an "estimate" because it is generated from a sample of precipitation values rather than the full population of all possible precipitation values. We would like to know the true population regression slope but it is rarely possible to measure the entire population, and therefore we have to accept an estimate of the slope based on a sample of the data. This is one of the fundamental foundations of statistics.

  The Confidence Limits of the slope tell us how confident we are about our estimate of the slope. Our 95 percent confidence limits should be interpreted to mean that, if we took an infinite number of samples of elevation and annual precipitation, then the true population regression slope (the one that we are really interested in) would lie between the upper and lower confidence limits 95 percent of the time. In this case, we can take this to mean that there is approximately a 95 percent chance that the true population slope lies between 0.148 and 0.216, and our best estimate of it is 0.182. If the 0-slope lay within our confidence limits, then this would be evidence that there may not be a relationship between the variables.
- R-squared: also called the *Coefficient of Determination*, this value is a measure of how much of the variability in the dependent variable can be explained by the variability in the independent variable. In the above example, an R^2 value of 0.014 indicates that only 1.4% of the variability in the dependent variable *Annual Precipitation* [Aprec_mean] can be explained by variation in the independent variable *Elevation* [Elev_mean]. This tells a user that they should be very hesitant about predicting annual precipitation in this particular area based only on the mean elevation, because *Elevation* appears to have little influence over *Annual Precipitation* (at least when analyzed over the entire continent, as we have done here).
- ANOVA Table: also called an *Analysis of Variance* table, this table provides a breakdown of the various components of the regression relationship as well as an estimate of the confidence that a true linear relationship exists between the two variables. The *P-Value* reflects the probability that the relationship examined is not linear at all, but that it is rather simply an artifact of random chance. In this case, the *P-Value* of <0.000001 indicates that the chances are extremely remote that the relationship is due to chance, and therefore it can be concluded that there is indeed strong evidence of a linear relationship between the two variables.

  It is good statistical technique to analyse the usefulness of a regression analysis using all these factors of the relationship, as well as to review a plot of the data distributions and regression line. For example, in this case there is extremely strong evidence of a linear relationship given the very small *P-Value*, but the R^2 value shows that the linear relationship really does not explain the behaviour of the dependent variable very well. Looking at the scatterplot of the output in the figure below illustrates the fact that the relationship between elevation and mean annual precipitation is not strong (Figure 1.97b).

FIGURE 1.97B



The scatterplot

In cases where a user may wish to spatially identify where a particular feature of the selected set lay in relation to the Regression line and confidence bands depicted on the figure above, the user can use the Identify tool ⓘ to click on any of the points and identify exactly which feature it represents, and then use the ID value to locate the feature back in their main view.

Confidence bands, residuals and predicted values

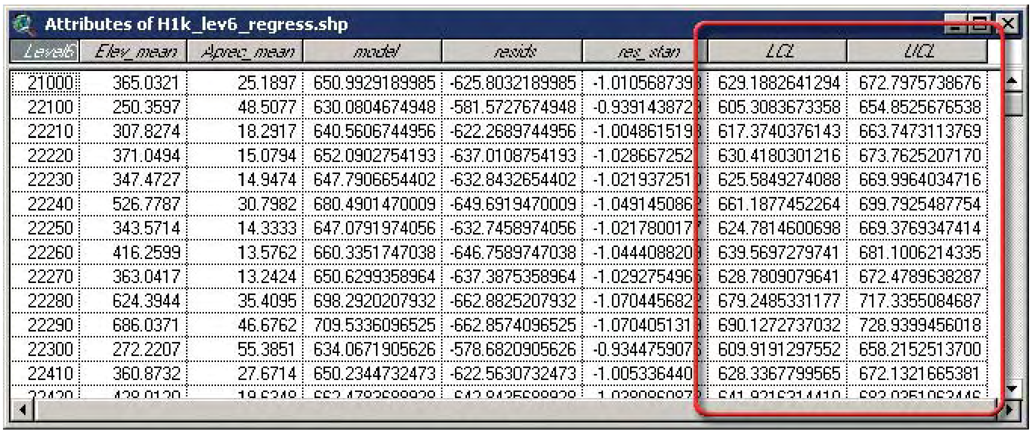
Confidence Bands: it is often wise to include some measure of the uncertainty of any statistical output and this applies to regression as well as to most other statistical analyses. Although the plotted regression line is the best estimate of the relationship, there will always be some uncertainty unless every possible combination of elevation and annual precipitation is sampled for all locations for all time.

The Confidence Bands in this case reflect the upper and lower confidence levels for the regression line over different levels of the independent variable. Since a confidence level of 95 percent was used, the results should be interpreted as "If identical regression relationships were developed an infinite number of times, based on an infinite number of random samples of the respective variable populations, then the true regression line will lie within these confidence bands 95 percent of the time."

It can also be observed from the scatterplot above that the bands tend to diverge from the regression line at higher levels of annual precipitation. This is because the regression relationship is strongest when the sample points are close to the means of the input variables. The confidence bands will always be closest to the regression line at the mean value of the Independent Variable, and diverge as one moves away from that mean.

Confidence Band values will also be added to the new "regression" ArcView theme attribute table ▦, in fields labeled "LCL" (for Lower Confidence Limit) and "UCL" (for Upper Confidence Limit) (Figure 1.97c).

FIGURE 1.97C
The new "regression" theme attribute table



Attributes of H1k_lev6_regress.shp

Level6	Elev_mean	Aprec_mean	model	resids	res_stan	LCL	UCL
21000	365.0321	25.1897	650.9929189985	-625.8032189985	-1.010568739	629.1882641294	672.7975738676
22100	250.3597	48.5077	630.0804674948	-581.5727674948	-0.939143872	605.3083673358	654.8525676538
22210	307.8274	18.2917	640.5606744956	-622.2689744956	-1.004861519	617.3740376143	663.7473113769
22220	371.0494	15.0794	652.0902754193	-637.0108754193	-1.028667252	630.4180301216	673.7625207170
22230	347.4727	14.9474	647.7906654402	-632.8432654402	-1.021937251	625.5849274088	669.9964034716
22240	526.7787	30.7982	680.4901470009	-649.6919470009	-1.049145086	661.1877452264	699.7925487754
22250	343.5714	14.3333	647.0791974056	-632.7458974056	-1.021780017	624.7814600698	669.3769347414
22260	416.2599	13.5762	660.3351747038	-646.7589747038	-1.044408820	639.5697279741	681.1006214335
22270	363.0417	13.2424	650.6299358964	-637.3875358964	-1.029275496	628.7809079641	672.4789638287
22280	624.3944	35.4095	698.2920207932	-662.8825207932	-1.070445682	679.2485331177	717.3355084687
22290	686.0371	46.6762	709.5336096525	-662.8574096525	-1.070405131	690.1272737032	728.9399456018
22300	272.2207	55.3851	634.0671905626	-578.6820905626	-0.934475907	609.9191297552	658.2152513700
22410	360.8732	27.6714	650.2344732473	-622.5630732473	-1.005336440	628.3367799565	672.1321665381

Residuals and Predicted Values: along with confidence bands, the new "regression" theme attribute table also contains fields for the predicted values, residuals and standardized residuals. The "model" field holds values for the predicted value of *Mean Annual Precipitation* based on the regression equation and that record's value for *Mean Elevation.* The *Residuals* field [resids] holds values reflecting how much the measured *Mean Annual Precipitation* deviated from the predicted model value. The *Standardized Residuals* field [res_stand] standardizes these residuals values by converting them to Z-scores, making it easy to identify extreme outliers.

Example of performing analyses on different subsets of data

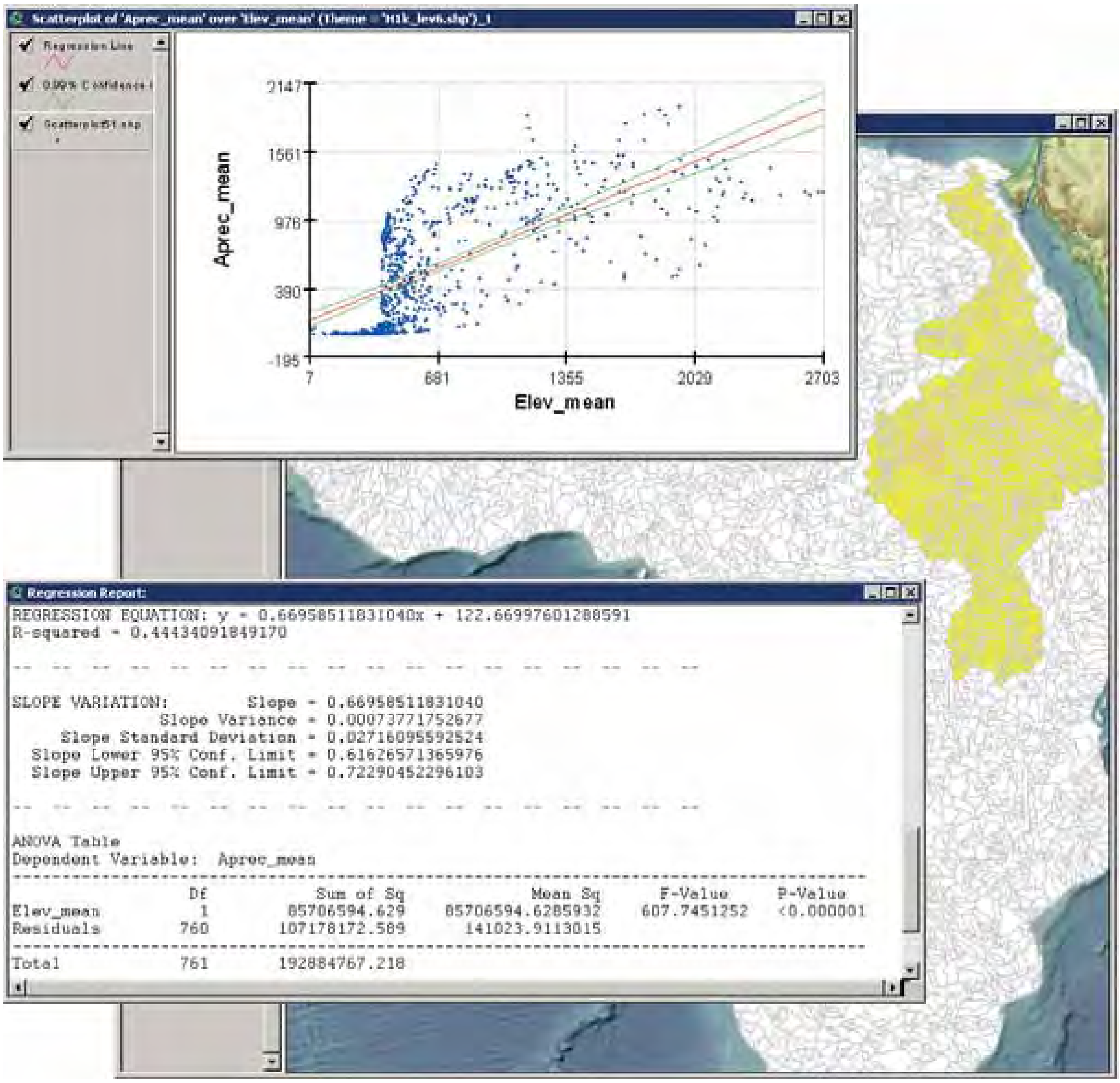
One of the strong points of the Simple Linear Regression tool is that a user can restrict the analysis to a subset of features by selecting those features prior to analysis. If any features are selected, then the tool will only operate on those selected features. If no features are selected, then the tool will operate on all features in the theme.

If, for example, a user was really only interested in phenomena occurring within the Nile River megabasin:

1. Using the HYDRO-1K watershed model from the previous example, click the Select Upstream and Downstream Watersheds" tool of the Watersheds Module. Choose the "Include Entire Basin" option and then click anywhere in Lake Victoria. This will select all watersheds in the Nile River megabasin.
2. Open the Simple Linear Regression tool by clicking on the button on the AWRD Interface.
3. Select elevation (i.e. Elev_mean) as the independent variable and Annual precipitation (i.e. Aprec_mean) as the dependent variable.
4. Tick all the regression options.
5. Write 0.95 as the confidence level for slope, and 0.99 as the confidence level for the model. Then click "OK".

The regression report and scatterplot are illustrated in Figure 1.98.

FIGURE 1.98
Regression analysis of rainfall and elevation for the Nile River megabasin



```
REGRESSION EQUATION: y = 0.66958511831040x + 122.66997601280591
R-squared = 0.44434091849170

-- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --

SLOPE VARIATION:          Slope = 0.66958511831040
                Slope Variance = 0.00073771752677
     Slope Standard Deviation = 0.02716095592524
   Slope Lower 95% Conf. Limit = 0.61626571365976
   Slope Upper 95% Conf. Limit = 0.72290452296103

-- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --

ANOVA Table
Dependent Variable:  Aprec_mean
-----------------------------------------------------------------------------------------
                    Df          Sum of Sq           Mean Sq       F-Value     P-Value
Elev_mean            1       85706594.629    85706594.6285932   607.7451252  <0.000001
Residuals          760      107178172.589      141023.9113015
-----------------------------------------------------------------------------------------
Total              761      192884767.218
```

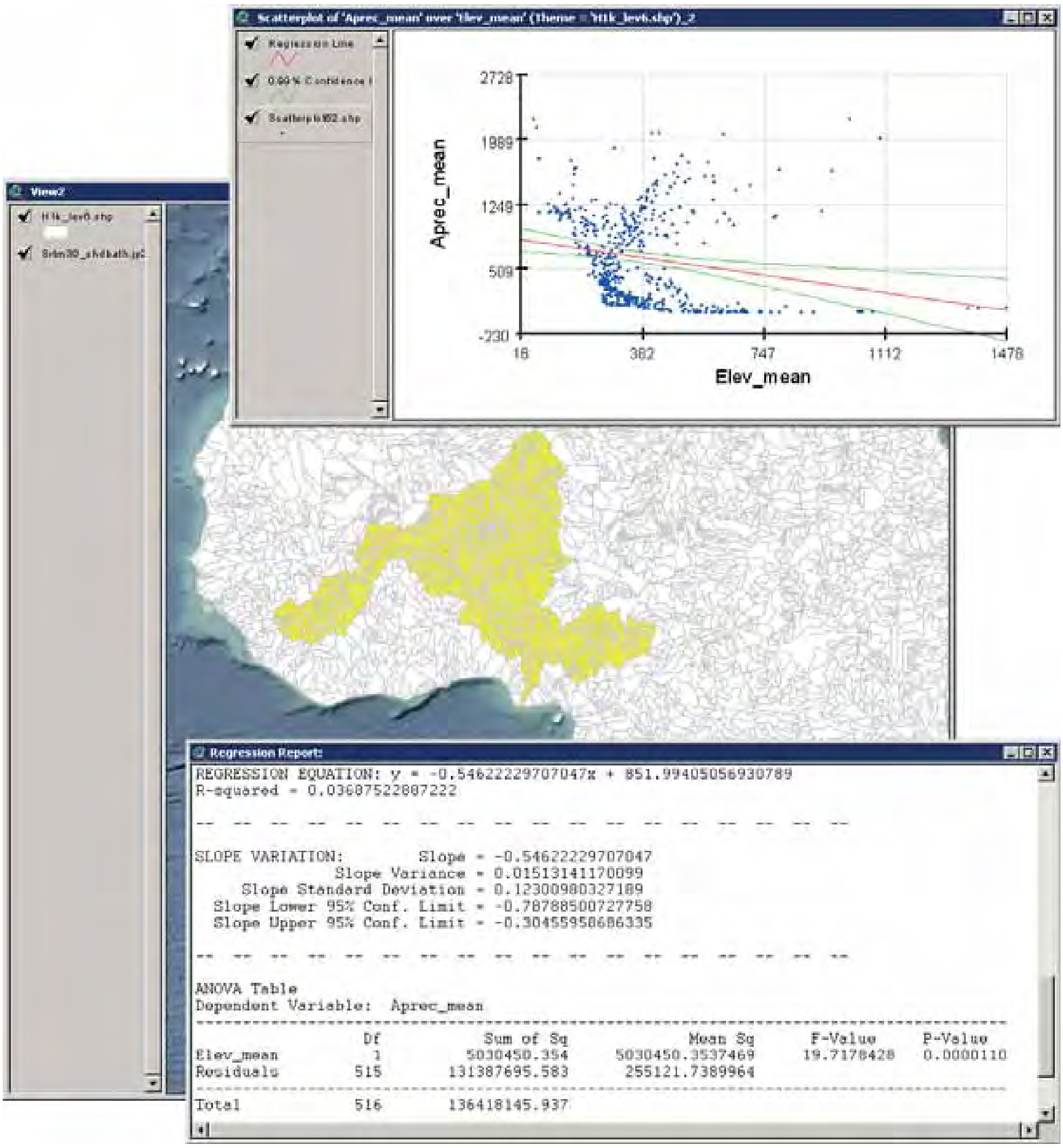
Interestingly, in this example there is actually a much stronger regression relationship where the annual precipitation appears to be more influenced by elevation (Figure 1.99).

The evidence for a linear relationship is still very strong with a *P-Value* of <0.000001 The R^2-value is far higher in this example ($R^2 = 0.44$), meaning that annual precipitation in the Nile river basin is more correlated with elevation than it is in Africa as a whole, and that higher elevations tend to get more precipitation than lower elevations.

Visual examination of the scatterplot shows that the relationship does not appear to be linear over the full range of elevation values. Lower elevations tend to have very low precipitation levels, and the precipitation becomes much more variable at approximately 400 metres. This suggests that it might be interesting to run the regression twice; once for watersheds < 400 metres and again for watersheds > 400 metres.

By clicking on the Niger megabasin, the analysis can quickly be re-run on a different region (Figure 1.99).

FIGURE 1.99
Regression analysis of rainfall and elevation for the Niger megabasin



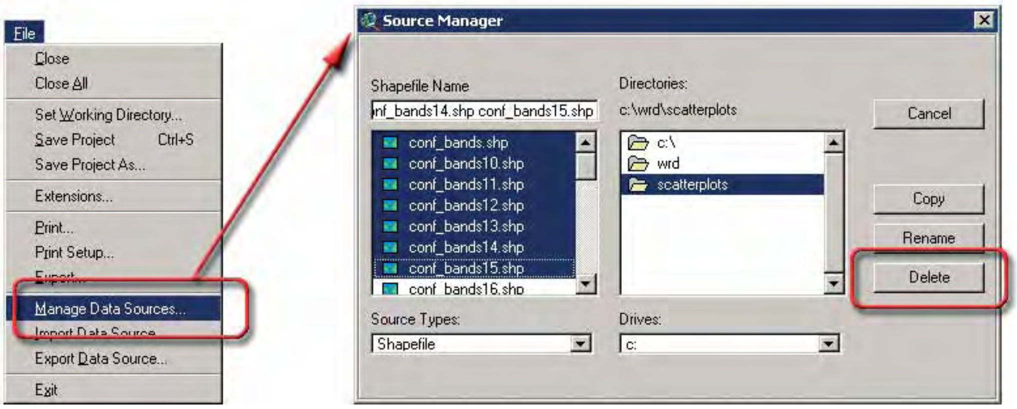
```
REGRESSION EQUATION: y = -0.54622229707047x + 851.99405056930789
R-squared = 0.03687522887222

-- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --

SLOPE VARIATION:        Slope = -0.54622229707047
               Slope Variance = 0.01513141170099
     Slope Standard Deviation = 0.12300980327189
  Slope Lower 95% Conf. Limit = -0.78788500727758
  Slope Upper 95% Conf. Limit = -0.30455958686335

-- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --

ANOVA Table
Dependent Variable:  Aprec_mean
-------------------------------------------------------------------------------
                  Df          Sum of Sq             Mean Sq      F-Value    P-Value
Elev_mean          1        5030450.354     5030450.3537469   19.7178428  0.0000110
Residuals        515      131387695.583      255121.7389964
-------------------------------------------------------------------------------
Total            516      136418145.937
```

The results of this analysis are also very interesting in that they differ from the relationship established for the continent as a whole. The evidence for a linear relationship is still very strong with a *P-Value* of 0.0000110, but the R^2-value is very low at 0.08, meaning that elevation has little influence on annual precipitation in this region. Interestingly, the regression relationship assumes a different direction than was seen in the Nile megabasin, in that here the mean annual precipitation decreases as elevation increases.

There are several excellent texts available that discuss regression in exhaustive detail. For those who are interested, we recommend. Draper and Smith (1998) and Neter *et al.* (1996).

Note This process generates three shapefiles every time a regression analysis is done. These shapefiles are stored in the "Scatterplots" sub-folder in the AWRD folder. Over time, this folder can accumulate a lot of files. We recommend that users periodically review the files in that folder and delete the ones that are not being used anymore. The easiest way to delete shapefiles is to use the "*Manage Data Sources…*" menu option in the "File" menu of any View. Click that menu option, select the shapefiles you would like to delete, and click the "Delete" button. This function will automatically delete all the multiple files that make up each shapefile, and it will check to see if the file is currently in use in your project before deleting it (Figure 1.100).

FIGURE 1.100
The Manage Data Sources



Shapefile Name: nf_bands14.shp conf_bands15.shp

Directories: c:\wrd\scatterplots

Source Types: Shapefile

Drives: c:

Note Users who wish to conduct more sophisticated regression analyses, including both simple and multiple linear regression on both tabular and grid data, may find an extension named "Grid and Theme Regression" in the set of AWRD Add-On Extensions (see Section 1.8 of this manual). This add-on extension requires that the user have ESRI's Spatial Analyst extension installed.