Command Reference: FillMixedStation()

Fill missing data in dependent time series using the best fit from 1+ independent time series, using OLS Regression, data transforms, one/monthly equations

Version 10.23.00, 2013-09-12

The `FillMixedStation()` command fills missing data in a time series where one or more independent time series are used to sequentially fill missing data. This approach has been developed to automate analysis of regression filling and to facilitate batch filling of many related time series. This implementation is based on the Mixed Station Model implemented for Colorado's Decision Support Systems (Ayres Associates, 2000), which was based on the similarly named approach implemented by the USGS (Alley and Burns, 1981). However, due to performing calculations in double precision in TSTool, the results are not always identical.

The time series involved in the analysis are typically related, such as being from nearby locations in a region. The main uses of the command are

1. To automatically fill every time series in a data set, using other time series in the data set. For example, for hydrologic modeling natural flow time series may have been estimated by processing measured streamflow, diversion, and reservoir time series. The natural flow time series can be filled for use in modeling.
2. To perform an analysis without filling, to guide application of individual `FillRegression()` and other commands.

Important: TSTool does not automatically exclude time series that have been filled in previous steps. Consequently, care must be taken when specifying the list of independent time series to NOT use time series that were filled in a previous step. In the future, features may be enabled to examine the data flags to determine if data have been filled.

For each dependent time series being filled, the Mixed Station Analysis (MSA) selects the independent time series and parameters that result in the best filling results, considering combinations of the following:

- The list of independent time series being considered can be constrained to a subset of available time series.
- Filling methods include ordinary least squares (OLS) regression (see the `FillRegression()` command for details). Support for MOVE2 may be added in the future (see the `FillMOVE2()` command for details).
- One equation and/or monthly equations can be used.
- The data can be transformed using $\log_{10}$, or no transformation can be applied.
- A minimum number of overlapping data points (sample size N1) can be specified to indicate a valid relationship.
- A minimum correlation coefficient *r* can be specified to indicate a valid relationship.
- A minimum confidence level for the slope of the regression line can be specified (see T-Test discussion below).
- The best fit indicator can be the correlation coefficient (*R*), or the standard error of prediction (SEP, described below).

Because extensive analysis may be necessary to evaluate all the combinations of parameters, the `FillMixedStation()` command will be slower than other commands that specifically indicate how

to perform the filling. The number of combinations can also be limited by reducing the number of parameter options and using stricter limitations on the number of overlapping points and correlation coefficients that are required for a good regression result.

The full MSA process is as follows:

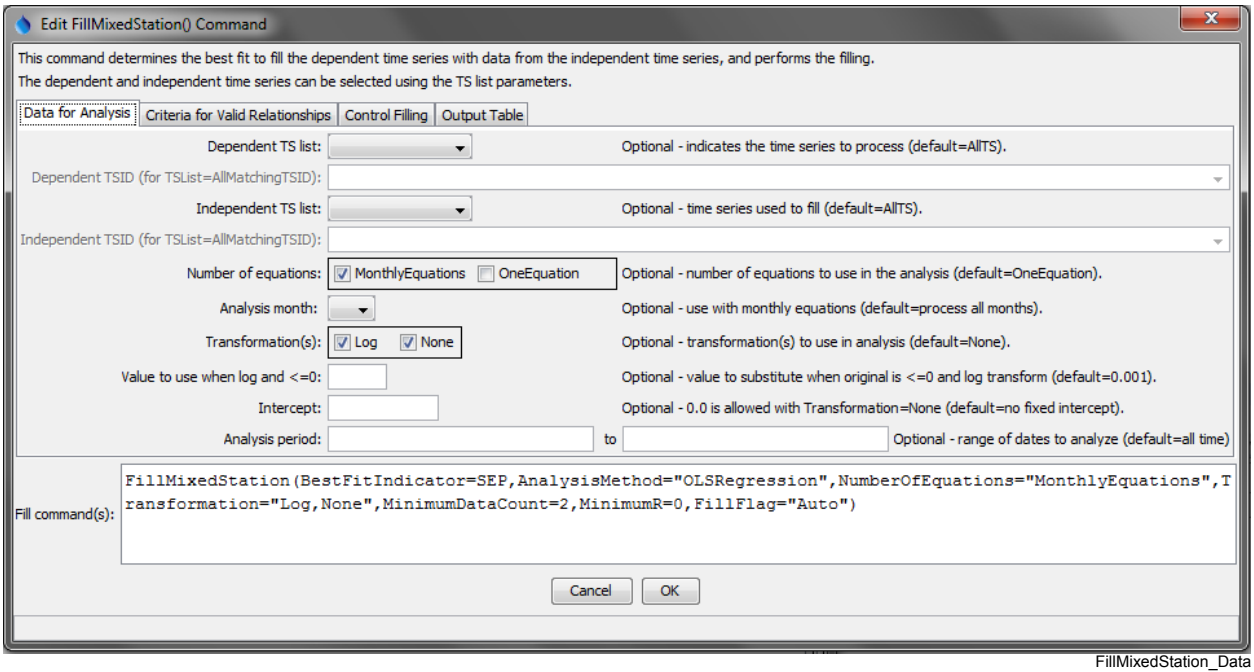
1. For each dependent time series, perform a regression analysis using a unique combination of parameters (e.g., use an independent time series, OLS regression with one equation, no data transform). This results in 1+ regression results for each dependent time series.
2. Qualifying results (those that meet the requirements of minimum number of overlapping points and correlation coefficient) are retained in a list for the dependent time series, for processing in the next step.
3. The qualifying results are used to estimate each missing value. Typically, the SEP is used to select the relationship to use (the one that has lowest SEP).
4. Missing data in the dependent time series are filled using the regression results for the selected relationship. If missing values remain, the next highest ranking regression result is used until all missing values are filled (or no additional qualifying regression results are available). Monthly filling occurs on each of the 12 months. This approach may use different stations for each filled value because of the goodness of fit of the relationship and because different stations may or may not have data that overlap the period to be filled.

Implementation in Colorado's Decision Support Systems

The Mixed Station Model implemented for the State of Colorado typically used the following input:

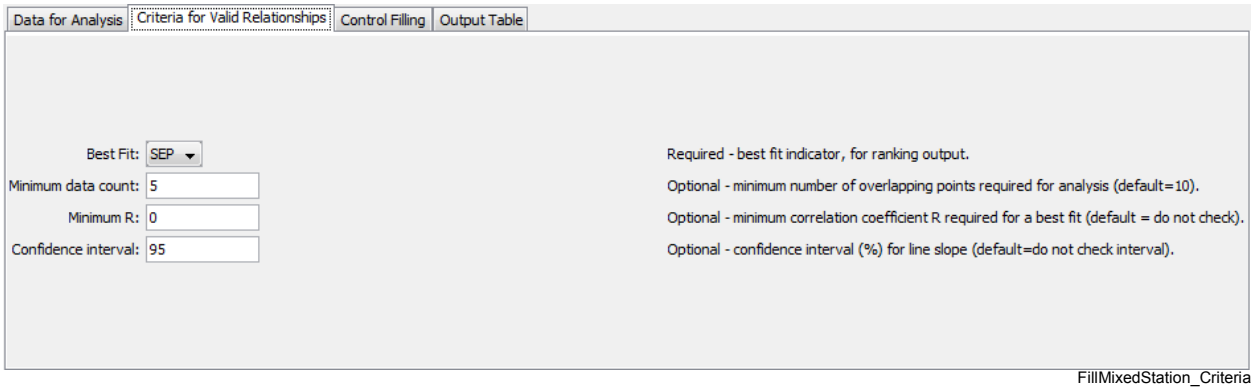
- Log transform (`Transform=Log10`)
- One and monthly relationships (`NumberOfEquations=MonthlyEquations,OneEquation`)
- Rank on SEP (`BestFitIndicator=SEP`)
- Minimum concurrent values = 5 (`MinimumDataCount=5`)
- Confidence level = 95% (`ConfidenceInterval=95`)
- Fill all time series in data set (nothing selected in filling)

The following dialog is used to edit the `FillMixedStation()` command and illustrates the syntax of the command.



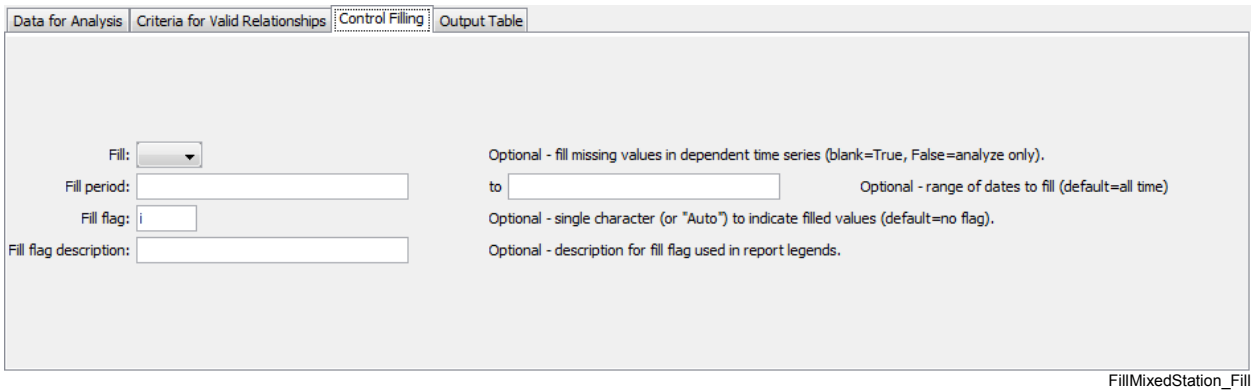
FillMixedStation_Data

FillMixedStation() Command Editor - "Data For Analysis" Parameters



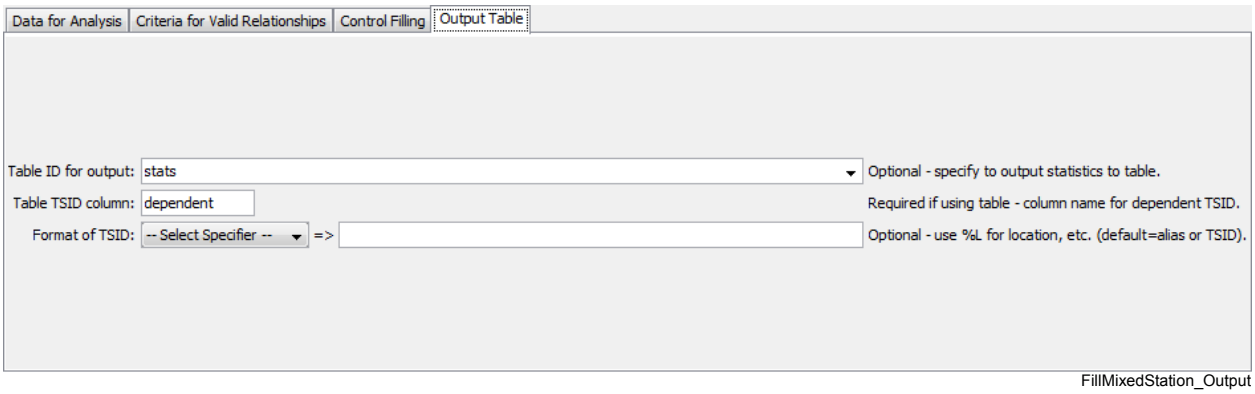
FillMixedStation_Criteria

FillMixedStation() Command Editor - "Criteria for Valid Relationships" Parameters



FillMixedStation_Fill

FillMixedStation() CommandEditor - "Control Filling" Parameters



FillMixedStation_Output

FillMixedStation() CommandEditor - "Output Table" Parameters

The command syntax is as follows:

```
FillMixedStation(Parameter=value,…)
```

Command Parameters

Parameter	Description	Default
`DependentTSList`	Indicates the list of independent time series to be processed, one of: • `AllMatchingTSID` – all time series that match the `TSID` (single TSID or TSID with wildcards) will be processed. • `AllTS` – all time series before the command will be processed. • `EnsembleID` – all time series in the ensemble will be processed. • `FirstMatchingTSID` – the first time series that matches the `TSID` (single TSID or TSID with wildcards) will be processed. • `LastMatchingTSID` – the last time series that matches the `TSID` (single TSID or TSID with wildcards) will be processed. ▪ `SelectedTS` – the time series selected with the `SelectTimeSeries()` command will be processed.	`AllTS`
`DependentTSID`	The time series identifier or alias for the dependent time series to be processed, using the `*` wildcard character to match multiple time series.	Required if `DependentTSList=*TSID`.
`IndependentTSList`	Indicates the list of independent time series to be considered for each dependent time series, one of: • `AllMatchingTSID` – all time series that match the `TSID` (single TSID or TSID with wildcards) will be processed.	`AllTS`

Parameter	Description	Default
	• `AllTS` – all time series before the command will be processed. • `EnsembleID` – all time series in the ensemble will be processed. • `FirstMatchingTSID` – the first time series that matches the `TSID` (single TSID or TSID with wildcards) will be processed. • `LastMatchingTSID` – the last time series that matches the `TSID` (single TSID or TSID with wildcards) will be processed. • `SelectedTS` – the time series selected with the `SelectTimeSeries()` command will be processed.	
`IndependentTSID`	The time series identifier or alias for the independent time series to be compared, using the `*` wildcard character to match multiple time series.	Required if `IndependentTSList=*TSID`.
`NumberOfEquations`	The number of equations to use for the analysis: `OneEquation` and/or `MonthlyEquations`.	`OneEquation`
`AnalysisMonth`	The month that data should be considered for.	All months
`Transformation`	Indicates how to transform the data before analyzing. Specify as `None` (no transformation) or `Log` (for Log_{10}). If the `Log` option is used, zero and negative values in data are set to `.001`. Missing data are ignored. If multiple values are selected, separate with a comma and surround with double quotes.	`None` (no transformation)
`LEZeroLogValue`	Value to use for data values less than or equal to zero when using a log transformation. The Log_{10} of this value will be used in calculations. Use `Missing` to ignore those values entirely. Caution: this will set 0 as a missing value in the time series.	`.0010`
`Intercept`	Specify as `0` to force the intercept of the best-fit line through the origin. This is made available only for OLS regression analysis on untransformed data, to be consistent with the `FillRegression()` command.	Do not force the intercept through zero.
`AnalysisStart`	The date/time to start the analysis, to focus on a period appropriate for analysis. For example, specify the unregulated period for streamflow.	Analyze the full period.
`AnalysisEnd`	The date/time to end the analysis.	Analyze the full period.

Parameter	Description	Default
BestFitIndicator	Specifies the indicator to use when determining the best fit, one of: ▪ R – correlation coefficient (attempts to maximize) ▪ SEP – Standard Error of Prediction, defined as the square root of the sum of differences between the known dependent value, and the value determined from the equation of best fit at the same point (attempts to minimize)	SEP
MinimumDataCount	The minimum number of overlapping data points that are required for a valid analysis (N1 in FillRegression() documentation). If the minimum count is not met, then the independent time series is ignored for the specific combination of parameters. For example, if monthly equations are used, the independent time series may be ignored for the specific month; however, it may still be analyzed for other months.	10
MinimumR	The minimum correlation coefficient required for a best fit. If the minimum is not met, then the results are not considered in the best fit ranking or filling.	
ConfidenceLevel	Required confidence level for the T-Test on the regression slope. Relationships not passing the test are not allowed for filling.	No limit on confidence level.
Fill	Indicates whether filling should occur (True) or just analyze to compute statistics (False). The latter is useful for testing combinations of statistics prior to actually performing filling. For example, use this command to analyze relationships, create an output table, and then use individual FillRegression() commands for filling specific time series.	True
FillStart	The date/time to start filling, if other than the full time series period.Fill the full period.	Fill the entire period.
FillEnd	The date/time to end filling, if other than the full time series period.Fill the full period.	Fill the entire period.
FillFlag	A single character that will be used to flag filled data. “Auto” will show whether a monthly or single equation was used to fill, as well as what rank the equation used to fill was. “I” will show the location of the equation used to fill.	Filled values will not be flagged.
FillFlagDesc	Description for the fill flag, used in reports.	Automatically generated.
MinimumR	The minimum correlation coefficient required for a best fit. If the minimum is not met, then the results are not considered in the best fit ranking or filling.	No limit on R.

Parameter	Description	Default
`TableID`	A table identifier for a table to receive output of the analysis. Note that creating the table requires a significant amount of memory, making it impractical for very large data sets.	Statistics are not written to the table. Refer to the log file for information.
`TableTSIDColumn`	The name of the column in the table that contains time series identifier information. This is used to match the table with time series being analyzed so that statistics can be written to the correct row.	Required if `TableID` is specified.
`TableTSIDFormat`	The specifier used to format the time series identifier in the TableTSIDColumn. The location part of the TSID, or the time series alias is typically used.	The alias will be used if available, or otherwise the full TSID will be used.

The FillFlag parameter shows more information about how the time series were filled:

06619400, , ACFT	06619450, , ACFT
382.9^06616000	1646.3^06620000
280.2^06616000	508.2^06620000
357.1^06616000	1033.3^06620000
1386.2^06616000	3777.0^Apr06620000
5527.1^May06616000	8576.9^06620000
7664.5^Jun06616000	13467.1^Jun06616000
1022.8^Jul06616000	941.3^Jul06616000
952.1^06616000	858.7^06620000

FillMixedStation_FlagI

Example of using the "I" fill flag

09137800, , af	720758, , af
327.4^Oct1	115.0
	0.0
	0.0
	0.0
	0.0
	0.0
	0.0
1191.2^May1	9.0

FillMixedStation_FlagAuto

Example of using the "Auto" fill flag

Various statistics for each combination of time series, such as count, mean, and standard deviation, can be output in a table:

dependent	dependent_Independ...	AnalysisMethod
09110500...MONTH	09111500...MONTH	OLSRegression
09110500...MONTH	09112000...MONTH	OLSRegression
09110500...MONTH	09112500...MONTH	OLSRegression
09110500...MONTH	09109000...MONTH	OLSRegression
09110500...MONTH	09110000...MONTH	OLSRegression
09110500...MONTH	09113500...MONTH	OLSRegression
09110500...MONTH	09114500...MONTH	OLSRegression
09110500...MONTH	09115500...MONTH	OLSRegression

More statistics...

N1	MeanX1	MeanX1_trans	SX1	SX1_trans	N2	MeanX2
138	8457.50000000	8457.50000000	12783.73216882	12783.73216882	84	9134.02380952
138	2002.46376812	2002.46376812	2293.16167221	2293.16167221	39	2669.79487179
143	20320.32167832	20320.32167832	27760.95881609	27760.95881609	793	21813.34174023
143	11589.92307692	11589.92307692	12523.90673257	12523.90673257	601	12493.64392679
143	18320.02097902	18320.02097902	18964.08889820	18964.08889820	939	20986.14483493
126	7057.21428571	7057.21428571	9174.73312319	9174.73312319	180	5900.52222222
84	45530.73809524	45530.73809524	56484.98282380	56484.98282380	804	50032.50995025
143	3932.79720280	3932.79720280	4777.07374859	4777.07374859	445	4121.51460674

FillMixedStation_Table

Example of FillMixedStation() output table

More information on the statistics in this table can be found in the FillRegression() command documentation.

The following example command file fills natural flow time series from a StateMod model file using the traditional CDSS parameters:

```
ReadStateMod(InputFile="np2008_BF.xbf")
FillMixedStation(BestFitIndicator=SEP,NumberOfEquations="MonthlyEquations,One
Equation",Transformation="Log",ConfidenceInterval=95,MinimumDataCount=5,Minim
umR=0,FillFlag="i",TableID="stats",TableTSIDColumn="dependent")
```

This page is intentionally blank.