Command Reference: FillMixedStation()

Fill missing data in dependent time series using the best fit from 1+ independent time series, using OLS regression or MOVE2, data transforms, one/monthly equations

Version 10.20.00, 2013-04-21

This command is under development. It is envisioned that the FillRegression() command enhancements will be completed first. Then the FillMixedStation() analysis command will utilize much of the same logic, using the output statistics table to eliminate candidate relationships and use the remaining relationships to calculate estimated values to check against standard error of prediction, etc.

The `FillMixedStation()` command fills missing data in a time series where one or more independent time series is used to sequentially fill missing data. This approach has been developed to automate analysis of regression filling (see **Mixed Station Analysis Tool** below) 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).

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 generate a report on relationships, so that the user can configure individual `FillRegression()` and `FillMOVE2()` commands in TSTool. This may be appropriate when using `FillMixedStation()` on a list of time series is inappropriate.

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.

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) and MOVE2 (see the `FillMOVE2()` command for details).
- One equation or monthly equations can be used. However, both options cannot be evaluated together due to the complexity of ranking and reporting results.
- 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:

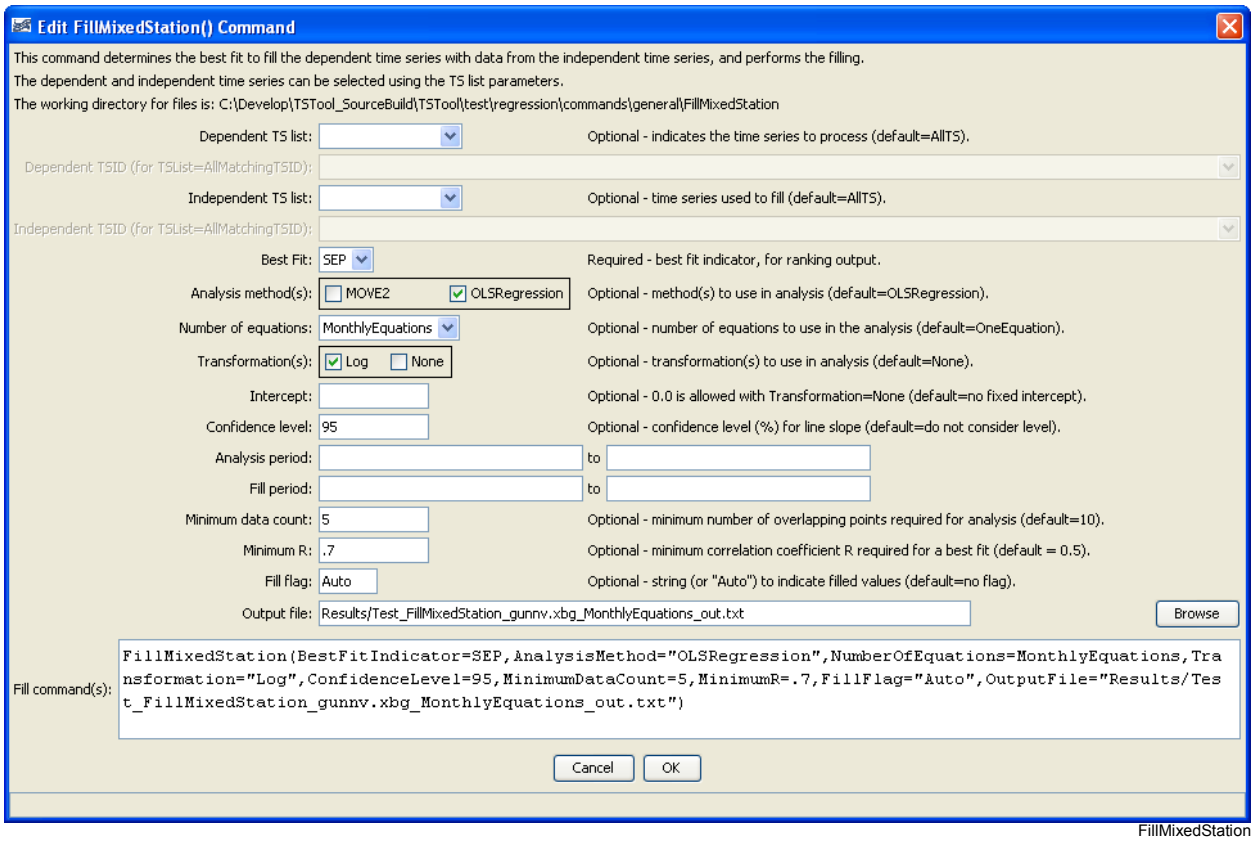
1. For each dependent time series, perform a regression analysis using a unique combination of parameters (e.g., use an independent time series, OLS regression with one equation, no data transform). This results in 1+ regression results for each dependent time series.
2. Qualifying results (those that 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 he 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 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
- Monthly relationships
- Rank on SEP
- Ordinary lease squares regression
- Minimum concurrent values = 5
- Confidence level = 95%
- Fill all time series in data set

The following dialog is used to edit the `FillMixedStation()` command and illustrates the syntax of the command. **Note that this interface will be updated to be similar to that of the `FillRegression()` command**.



FillMixedStation

FillMixedStation() Command Editor

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.	None – must be specified.

Parameter	Description	Default
	• 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.	
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 – 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.	None – must be specified.
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.

Parameter	Description	Default
`BestFitIndicator`	Specifies the indicator to use when determining the best fit, one of: ▪ `R` – correlation coefficient ▪ `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. ▪ `SEPTransformed`poin – same as SEP; however the data values have first been transformed as per the Transformation parameter. ▪ `SEPTotal`, when used with one equation, it is the same as `SEP`. When used with monthly equations, it is the `SEP` considering all months. ▪ `SEPTransformedTotal`, when used with one equation, it is the same as `SEPTransformed`. When used with monthly equations, it is the `SEPTransformed` considering all months.	`SEP`
`AnalysisMethod`	Specify the method(s) to analyze the data, in order to determine the best fit, including `OLSRegression` and/or `MOVE2`. If multiple methods are specified, separate with commas and surround with double quotes.	`OLSRegression`
`NumberOfEquations`	The number of equations to use for the analysis: `OneEquation` or `MonthlyEquations`. Only one may be chosen. If necessary, use more than one command to use different parameter combinations for different groups of time series.	None – must be specified.
`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)
`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.
`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.

Parameter	Description	Default
AnalysisStart	The date/time to start the analysis, to focus on a period appropriate for analysis. For example, specify the unregulated period for streamflow.	If blank, analyze the full period.
AnalysisEnd	The date/time to end the analysis.	If blank, analyze the full period.
FillStart	The date/time to start filling, if other than the full time series period.	If blank, fill the full period.
FillEnd	The date/time to end filling, if other than the full time series period.	If blank, fill the full period.
MinimumDataCount	The minimum number of overlapping data points that are required for a valid analysis (N1 in FillRegression() and FillMOVE2() 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.	0.5
OutputFile	Output file for the results, either as a file name to be written to the working directory, or a full path.	If not specified, partial results of the analysis may be available in the log file.

The following example command file fills natural flow time series from a StaeMod file using one equation (not monthly):

```
# Test filling the gunnison monthly baseflow time series with
# Mixed Station Analysis (all combinations for one equation)
StartLog(LogFile="fill-baseflow.log")
ReadStateMod(InputFile="gunnv.xbg")
FillMixedStation(BestFitIndicator=SEP,AnalysisMethod="MOVE2,OLSRegression",
NumberOfEquations=OneEquation,
Transformation="Log,None",OutputFile="Results.txt")
# Check for missing data - all should be filled
CheckTimeSeries(CheckCriteria="Missing",MaxWarnings=10)
# Check for negative flows - should not be any
CheckTimeSeries(CheckCriteria="<",Value1=0,MaxWarnings=10)
```