PCA and Regression Transformation

Description and usage

Principal components analysis (PCA) is used when a data set contains many time series (dimensions), and the dimensions need to be reduced, while retaining the most significant relationships between the time series. Reducing the number of dimensions reduces the data set size and removes unrelated variability. The implementation in FEWS allows you to derive a relation between independent timeseries (e.g. observations) and a dependent timeseries (e. g. a simulated timeseries).

Using historical timeseries, the PCA regression transformation produces a linear regression equation and a root mean square error (RMSE). This equation is applied with the current observations to compute an estimate of the (simulated) parameter, as well as its associated RMSE.

The PCA regression transformation was developed to update basin snow models when they drift away from realistic output. Snow updating uses historic and current snow water equivalence (SWE). Historic and current data come from monitoring stations within or near the basin, and from simulations of SWE in the basins. Historic observed data are used for PCA for a basin, and can potentially include time series from many monitoring stations. Current data are current daily SWE values, and are also either simulated (modelled basin) or observed (from monitoring stations within or near the basin). PCA finds the strongest underlying relationships between the historic observed station time series, and produces a linear equation. Current SWE values can then be input into the equation, and a PCA estimate of current basin SWE is produced.

Input/output timeseries

In this function four nonequidistant input time series must be identified:
1. historicalObserved
2. historicalSimulated
3. currentObserved
4. currentSimulated

In the snow updating use of the PCA regression transformation, these time series are subsamples of a daily time series to produce one data point per month. See the dayMonth sample page for more details.

In this function two output time series must be identified:
1. A time series with the PCA-estimated parameter value calculated by the algorithm
2. A time series with the associated RMSE calculated by the algorithm

Each time series is assigned a variable ID which is used in the actual expression.

PCA and regression transformation

a) Handling of time series gaps and irregular lengths
In order to obtain the longest possible common period of record among the input time series, the gap filling behavior has been changed from the default FEWS behavior

On the left hand side of Table 1, a dataset is shown that consists of one basin and three stations with varying start and end times. Station 3 has a gap.

In the middle of the table, the resulting time series lengths are highlighted in color for various combinations of station and basin pairings.

On the right hand side of the table (light blue highlighting), the default FEWS pairing behavior is shown.

The BPA FEWS gap handling technique uses all of the available data, resulting in a longer dataset (gray highlighting).

Table 1: A graphical demonstration of the BPA FEWS gap handling technique

					Station Pairing Behavior in BPA-FEWS							FEWS regular behavior
Date	Basin A SWE	Station 1 SNWE	Station 2 SNWE	Station 3 SNWE	Basin A + Station 1	Basin A + Sation 2	Basin A + Station 3	Basin A + Station 1 + 2	Basin A + Station 1+ 3	Basin A + Station 2 +3	Basin A +Station 1 + 2 + 3	Station A + 1 + 2 +3
04/61	39.8	47.6										
04/62	38.8	53.8										
04/63	27.2	40.3										
04/64	50.4	55.8										
04/65	54.4	58.7										
04/66	39	43.4										
04/67	43.6	54.7		15.5								
04/68	35	45.3		27.2								
04/69	37.5	47.1		27.5								
04/70	40.2	48.4	33.4	25.1								
04/71	54.8	59.8	39.9	17.8								
04/72	62	77.1	56	34.9								
04/73	28.1	36.3	22.3	23.3								
04/74	50.9	59.5	45.4									
04/75	44.4	55.9	38									
04/76	51.3	54.6	42.1									
04/77	21.3	30.2	18									
04/78	38.5	46	31									
04/79	40.2	45.2	35.7	24								
04/80	36	41.3	29	28.1								
04/81	21.8	25.7	15.7	24.3								
04/82	45.9	54.1	37.7	26.7								
04/83	32.6	39.9	24.3	18.5								
04/84	31.5	41.9		19.3								
04/85	35.5	45.1		17								
04/86	32.1	34.7		24.4								
04/87	29.6	32.7										
04/88	37.1	39.2										
04/89	38	49.5										
04/90	33.8	42.9										
04/91		48.8										
04/92		34.3										
04/93		34.2										

b) PCA transformation

i) Data Preprocessing

Before PCA calculations take place, the data set may need to be normalized and standardized (ex. where two datasets have very different means, standard deviations, or are not normally distributed) . However, no one type of preprocessing is appropriate for all time series. Therefore, to automatically assess which preprocessing type produces the 'best' results, the FEWS PCA algorithm performs a variety of preprocessing techniques. The user is presented with the result with the lowest RMSE.

Preprocessing techniques include the following attempts to normalize the dataset:
Square-root
Cube-root
Log10
No pre-processing

Preprocessing can also standardize the dataset by subtracting the time series mean and dividing by the standard deviation.

Therefore there are eight possible preprocessing types for PCA: square-root and standardizing, square-root and not standardizing, cube-root and standardizing, cube-root and not standardizing...

ii) Derivation of the PCA equation

Details below describe how a linear equation is produced from an eigenvector in FEWS.
Where a PCA equation is constructed from two historical SWE time series, and one historical modeled time series, the eigenvector matrix is:

$$\begin{bmatrix} m \\ x \\ y \end{bmatrix}$$

and the linear equation is:

$$z = (a \times x) + (b \times y) + c$$

where:

$$a = m / (x(n-1))$$

$$b = m / (y(n-1))$$

where:

'm', 'x', and 'y' are eigenvalues from basin 'm', and stations 'x' and 'y' respectively.
'z' is the PCA derived equation
'a' and 'b' are coefficients
'm' is the dimension of the matrix
'c' is a constant offset, derived by determining the mean of the historical SWE time series

c) Linear regression and multiple linear regression:
In addition to PCA analysis, the snow analysis module also attempts to minimize the RMSE by modeling using linear or multiple linear regression.

i) Data Preprocessing
Regression preprocessing and iteration procedures are identical to PCA (square-root, cube-root, log10, or no preprocessing). Data can also be either normalized or not. Therefore, there are eight regression preprocessing types: square-root and standardizing, square-root and not standardizing, cube-root and standardizing, cube-root and not standardizing...

ii) Derivation of the regression equation

The user is informed in the FEWS statistics window if regression produces the lowest RMSE, and has been chosen.

Configuration

Config example: PCA_RMSE_SnowAnalysis.xml

```
<!--this is an example of a PCA transformation module configuration-->

<?xml version="1.0" encoding="UTF-8"?>
<transformationModule xmlns="http://www.wldelft.nl/fews" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.wldelft.nl/fews http://fews.wldelft.nl/schemas/version1.0/transformationModule.
xsd" version="1.0">

<!--declare your input time series-->
        <variable>
                <variableId>Obs_hist</variableId>
                <timeSeriesSet>
                        <moduleInstanceId>DayMonthSampleSNWE_SnowAnalysis</moduleInstanceId>
                        <valueType>scalar</valueType>
                        <parameterId>SNWE</parameterId>
                        <locationId>2A16</locationId>
                        <locationId>2A18</locationId>
                        <locationId>2A21</locationId>
                        <locationId>2A22</locationId>
                        <locationId>2A23</locationId>
                        <locationId>2A25</locationId>
                        <timeSeriesType>simulated forecasting</timeSeriesType>
                        <timeStep unit="nonequidistant"/>
                        <relativeViewPeriod unit="hour" start="-240" startOverrulable="true" end="0"/>
                        <readWriteMode>read only</readWriteMode>
                </timeSeriesSet>
        </variable>
        <variable>
                <variableId>Sim_hist</variableId>
                <timeSeriesSet>
                        <moduleInstanceId>DayMonthSampleSWE_SnowAnalysis</moduleInstanceId>
                        <valueType>scalar</valueType>
                        <parameterId>SWE</parameterId>
                        <locationId>MCDQ2IL</locationId>
                        <timeSeriesType>simulated forecasting</timeSeriesType>
                        <timeStep unit="nonequidistant"/>
                        <relativeViewPeriod unit="hour" start="-240" startOverrulable="true" end="0"/>
                        <readWriteMode>read only</readWriteMode>
                </timeSeriesSet>
        </variable>
        <variable>
                <variableId>Obs_current</variableId>
                <timeSeriesSet>
                        <moduleInstanceId>DayMonthSampleSNWE_SnowAnalysis</moduleInstanceId>
                        <valueType>scalar</valueType>
                        <parameterId>SNWE</parameterId>
                        <locationId>2A16</locationId>
                        <locationId>2A18</locationId>
                        <locationId>2A21</locationId>
                        <locationId>2A22</locationId>
                        <locationId>2A23</locationId>
```

```
                        <locationId>2A25</locationId>
                        <timeSeriesType>simulated forecasting</timeSeriesType>
                        <timeStep unit="nonequidistant"/>
                        <relativeViewPeriod unit="day" start="-2" end="2"/>
                        <readWriteMode>read only</readWriteMode>
                </timeSeriesSet>
        </variable>
        <variable>
                <variableId>Current_sim</variableId>
                <timeSeriesSet>
                        <moduleInstanceId>DayMonthSampleSWE_SnowAnalysis</moduleInstanceId>
                        <valueType>scalar</valueType>
                        <parameterId>SWE</parameterId>
                        <locationId>MCDQ2IL</locationId>
                        <timeSeriesType>simulated forecasting</timeSeriesType>
                        <timeStep unit="nonequidistant"/>
                        <relativeViewPeriod unit="day" start="-2" end="2"/>
                        <readWriteMode>read only</readWriteMode>
                </timeSeriesSet>
        </variable>

<!--declare your output time series-->
                <variable>
                <variableId>PCA_swe</variableId>
                <timeSeriesSet>
                        <moduleInstanceId>PCA_MCDQ2IL_RMSE_SnowAnalysis</moduleInstanceId>
                        <valueType>scalar</valueType>
                        <parameterId>SWE</parameterId>
                        <qualifierId>pca</qualifierId>
                        <locationId>MCDQ2IL</locationId>
                        <timeSeriesType>simulated forecasting</timeSeriesType>
                        <timeStep unit="day" multiplier="1"/>
                        <relativeViewPeriod unit="day" start="-2" end="2"/>
                        <readWriteMode>add originals</readWriteMode>
                        <ensembleId>main</ensembleId>
                </timeSeriesSet>
        </variable>
        <variable>
                <variableId>PCA_rmse</variableId>
                <timeSeriesSet>
                        <moduleInstanceId>PCA_MCDQ2IL_RMSE_SnowAnalysis</moduleInstanceId>
                        <valueType>scalar</valueType>
                        <parameterId>SWE</parameterId>
                        <qualifierId>rmse</qualifierId>
                        <locationId>MCDQ2IL</locationId>
                        <timeSeriesType>simulated forecasting</timeSeriesType>
                        <timeStep unit="day" multiplier="1"/>
                        <relativeViewPeriod unit="day" start="-2" end="2"/>
                        <readWriteMode>add originals</readWriteMode>
                        <ensembleId>main</ensembleId>
                </timeSeriesSet>
        </variable>
        <!--perform the PCA calculation-->
        <transformation id="PCA">

                <regression>
                        <principalComponentAnalysis>
                                <historicalObserved>
                                        <variableId>Obs_hist</variableId>
                                </historicalObserved>
                                <historicalSimulated>
                                        <variableId>Sim_hist</variableId>
                                </historicalSimulated>
                                <currentObserved>
                                        <variableId>Obs_current</variableId>
                                </currentObserved>
                                <currentSimulated>
                                        <variableId>Current_sim</variableId>
                                </currentSimulated>
                                <enableCombinationAnalysis>true</enableCombinationAnalysis>
                                <estimatedCurrentSimulated>
```

```
                                        <variableId>PCA_swe</variableId>
                                </estimatedCurrentSimulated>
                                <errorStatistics>
                                        <variableId>PCA_rmse</variableId>
                                </errorStatistics>
                        </principalComponentAnalysis>
                </regression>
        </transformation>

</transformationModule>
```

Config example: TimeSeriesDisplayConfig.xml
The following sample describes how the PCA snow updataing display is configured.
Note the use of the dayMonthSample function (DayMonth Sample)

```
<sampleFunctions>
                <dayMonthSampleFunction/>
</sampleFunctions>
<statisticalFunctions>
        <statisticalFunction function="principalcomponentanalysis">
                <observedParameterId>SNWE</observedParameterId>
                <simulatedParameterId>SWE</simulatedParameterId>
        </statisticalFunction>
</statisticalFunctions>
```

Produces the following display:

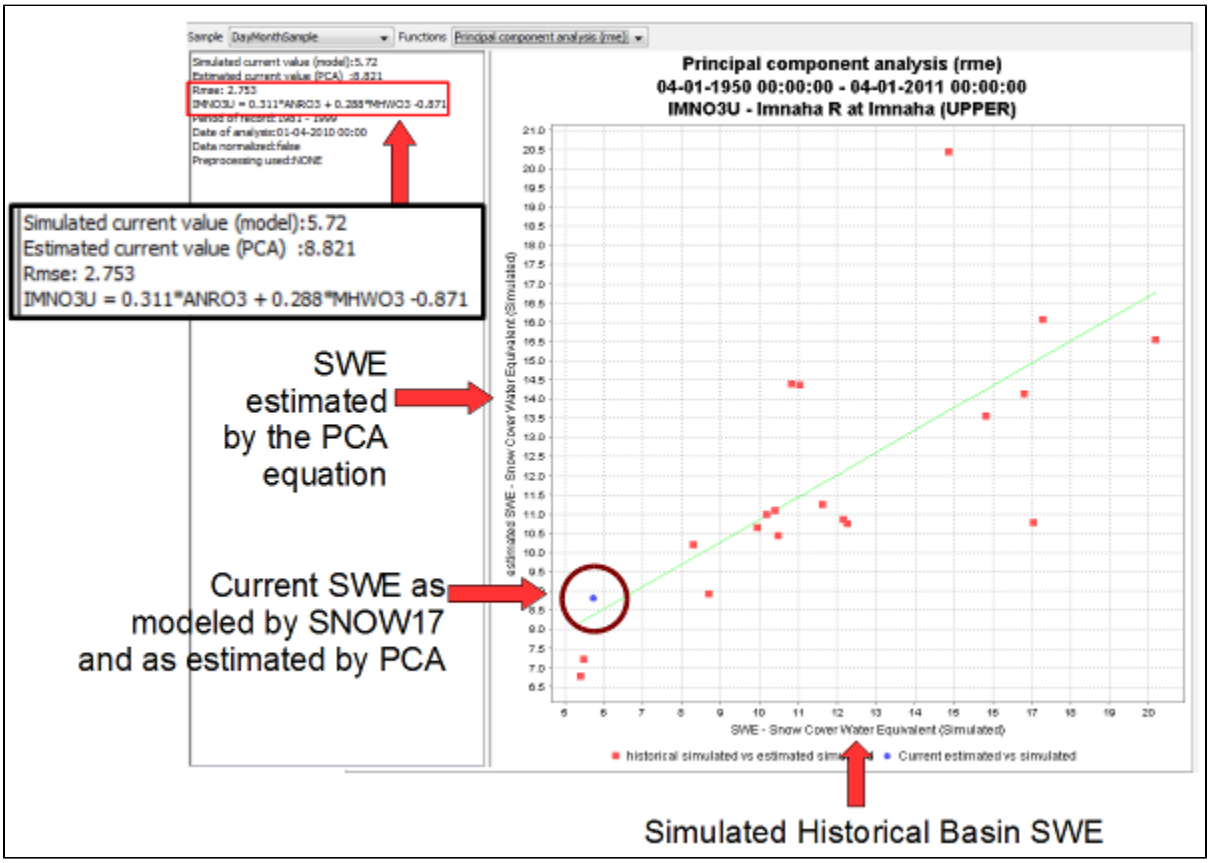

Figure 1. The snow updating GUI