Laboratory of Hydraulics,
Hydrology and Glaciology
of the Swiss Federal Institute
of Technology Zurich

DamBASE

Dam Behaviour Analysis
Software Environment

User Manual
Version 1.0

Commissioned by
Swiss Federal Office of Energy SFOE

VAW 0849

Zurich, May 2015

Contents

Acknowledgements

- The development of the software is financially supported by the Swiss Federal Office of Energy (SFOE). The software was written at VAW ETH Zurich by Marco Gerber in collaboration with Marius Bühlmann and David Vetsch. The software is a reimplementation and extension of the Software DamReg (also commissioned by the Swiss Federal Office of Energy (SFOE)) written by Benedikt Weber.
- The basic methods and ideas used in the software originate from the reference report by the former working group on numerical methods of the Swiss committee on Dams headed by Dr. Georges R. Darbre [Swiss Committee on Dams, 2010]

Citation Advice

- For User Manual: Gerber M., Bühlmann M., Vetsch D.F. (2015). DamBASE User Manual - Version 1.0. Laboratory of Hydraulics, Hydrology and Glaciology (VAW). ETH Zürich. Available from http://www.dambase.ethz.ch. [27.May 2015]
- For Website: DamBASE - Dam Behaviour Analysis Software Environment, 2015.
http://www.dambase.ethz.ch
- For Software: DamBASE - Dam Behaviour Analysis Software Environment. Version 1.0. © VAW, ETH Zurich, Gerber M., Buehlmann M., Vetsch D.F., 2015

Part I

Introduction

Chapter 1

Introduction and concepts

1.1 Purpose of the software

The main goal of DamBASE is to perform statistical regression analysis for dam observation. Besides this, further concepts to simplify the modelling process are also provided such as a simplified user mode as well as a high flexibility in terms of extending and customizing the software in order to deal with problems related to the individual behaviour of a dam.

As the entire observation process supported by this software makes heavy use of regression analysis, it is assumed that a user is familiar with the basic concepts of this topic. An introduction in German can be found in [Weber, B., 2002].

1.2 Structure of the manual

Chapter *Introduction and concepts* is intended to give you an overview of the basic concepts as well as introduces the main work flow. As DamBASE supports two different operation modes, you also learn how to distinguish them and what the core idea of this separation is. Chapter *Quick go-through* on page 11 finishes the first part with an example covering all steps necessary for setting up a regression model. This part also introduces all screens necessary for understanding what features the application provides. Therefore, if you are not familiar with the software yet, it is recommended to read this part first before going on to proceeding parts.

Part *Data model setup* on page 24 explains in more detail how a data model is set up and what analysis tools are provided to get a better understanding of the data.

In part *Regression model set up for calibration, validation and prediction* on page 29 chapters *Main work flow*, *Regression model setup for calibration and validation* and *Regression model setup for prediction* cover the main work flow and discuss the separation of calibration, validation and prediction more detailed.

How to generate a report and export data for post processing as well as saving and loading a project are explained in the two final parts *Save / Load project* on page 59 and *Report generation and export of regression results* on page 56.

1.3 Basic concepts

Normally, the software is intended to be used over a longer time while new data for observation is provided frequently. During this time the construction of a dam may have changed or new approaches for observation have been introduced. Further, the proper work flow requires to have multiple regression models available for comparison and optimization. All these requirements make it necessary to have multiple regression models kept together and available at any time.

Within the software, a *project* is meant to keep all this data and regression models together (figure 1.1). In order to distinguish different projects a name, a description as well as a dam operator can be assigned.

On the other hand, the final *regression* is done by applying a *regression formula* (a polynomial consisting of a set of *regressors*) on a set of data. All these regression related elements, except for the data set, are covered within a *regression model*.

The data set used by the regression is further covered within a *data model*. As the software also supports assignment of single columns (such as pendulum, water level and temperature) to their corresponding semantic meaning (behaviour indicators and influence factors) this assignment is also part of the data model (more on this can be found in chapter *3.2 on page 25*).

Due to the nature of the work flow, many of such regression models can be set up on the same data model. In order to keep this as a unit (one data model and all regression models depending on it), all these elements are finally covered within an *analysis* *object*. Thus, many of these analysis objects finally make up the project.

Furthermore, any kind of analysis (data analysis as well as regression analysis) can be done by using *analysis tools*. The aim of an analysis tool is to do one distinct analysis on either the data model or regression model. As the kind of analysis a user is interested in at a certain time depends on personal preferences as well as the context, these analysis tools can be chosen freely by the user (more on this is explained in section *User interface concept* on page 9).

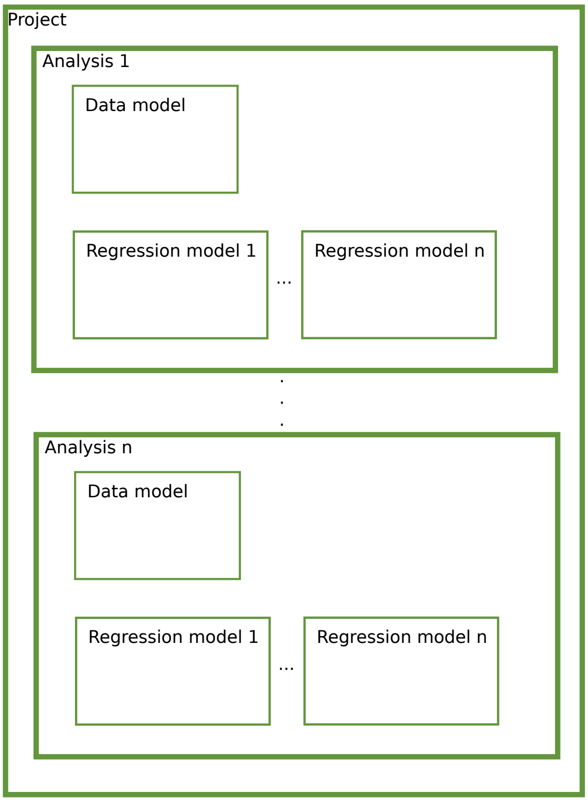


Fig. 1.1 *Basic concepts - relation*

1.4 Main work flow

With the current version of DamBASE a certain period *t* of the data is used for the setup of the regression model. Based on this model a prediction can be carried out and all related indicators are referred to as 'prediction'. Nevertheless, a new approach is proposed here, where the main work flow is split up into three chronologically ordered steps called calibration, validation and prediction (see figure 1.2).

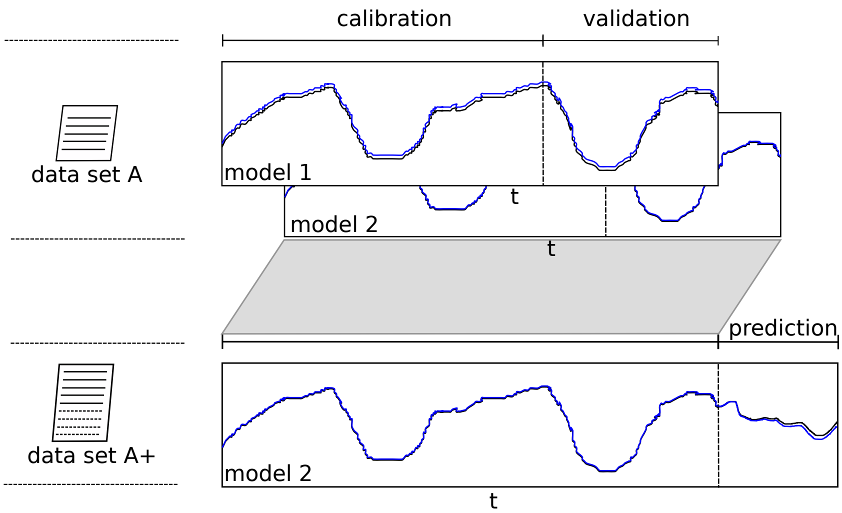


Fig. 1.2 *Separation into calibration, validation and prediction*

1.4.1 Calibration and validation

The goal of calibrating a regression model is to define a regression formula which is able to describe the behaviour of a dam as good as possible whereas validation makes sure that the model can describe values not part of the calibration period. Both steps are done with a data set for which it is assumed that the dam works correctly (*data set A* on figure 1.2 on the facing page).

As finding an appropriate model may be an iterative process, multiple models (*model 1* and *2*) can be set up and compared based on statistical and physical indicators.

1.4.2 Prediction

In step *prediction* new measurements (*data set A+*) are compared to values predicted by the regression model from the calibration and validation period (*model 2*). If indicators compared to the calibration and validation period differ too much it means that either the regression model is incomplete or the dam is behaving abnormal.

1.5 Operation modes

DamBASE supports two operation modes for setting up regression models and doing regression analysis. The core idea is to allow a user not being familiar with the details of statistical approaches, to set up a regression model that can be used for dam behaviour analysis. Please note that the basic user mode is the default mode. See *User interface concept* on the current page for how to switch between the two modes.

1.5.1 Basic user mode

The basic user mode supports regression model set up with a reduced set of regression model templates. These templates cover best practice models and abandon special case regressors. Further, the regression can be analysed exclusively with analysis tools related to a physical view of the model. This mode is the default mode when you start up the software.

1.5.2 Scientific user mode

The scientific user mode provides access to the full feature set of the software. Thus it covers all functionality of the basic user mode including special model templates with extended regressors as well as statistics related analysis tools.

1.6 User interface concept

The user interface is split up into a *navigation area* on the left and a *working area* on the right. These two areas are separated by a movable handle allowing to resize the areas, i.e. depending on user preferences (figure 1.3).

A working area is always linked to the currently selected object in the navigation area. The working areas for data and regression model are split up into three parts: a configuration area on the left upper side and two analysis areas (one below and a second on the right of the configuration area) also separated with a movable handle.

On this two working areas, a user can load analysis tools into any tab. It is possible to load an analysis tool multiple times with different configurations. Especially in cases where different aspects need to be analysed and compared it can be helpful to place two analysis tools of the same kind below or opposite each other (see figure 3.3 in section *AT* Time series *overview - Analyse values as time series* on page 26 for example). Every analysis tool has a tool bar for moving up and down as well as for removal.

It is possible to enable or disable certain graphs on plots having more than one graph by right-clicking on the corresponding graph within the legend.

After setting up either a data or regression model by using the wizard, some predefined and operation mode related analysis tools are loaded automatically on the corresponding working area. It is important to mention that additional analysis tools can be loaded individually.

Fig. 1.3 User interface concept

Chapter 2

Quick go-through

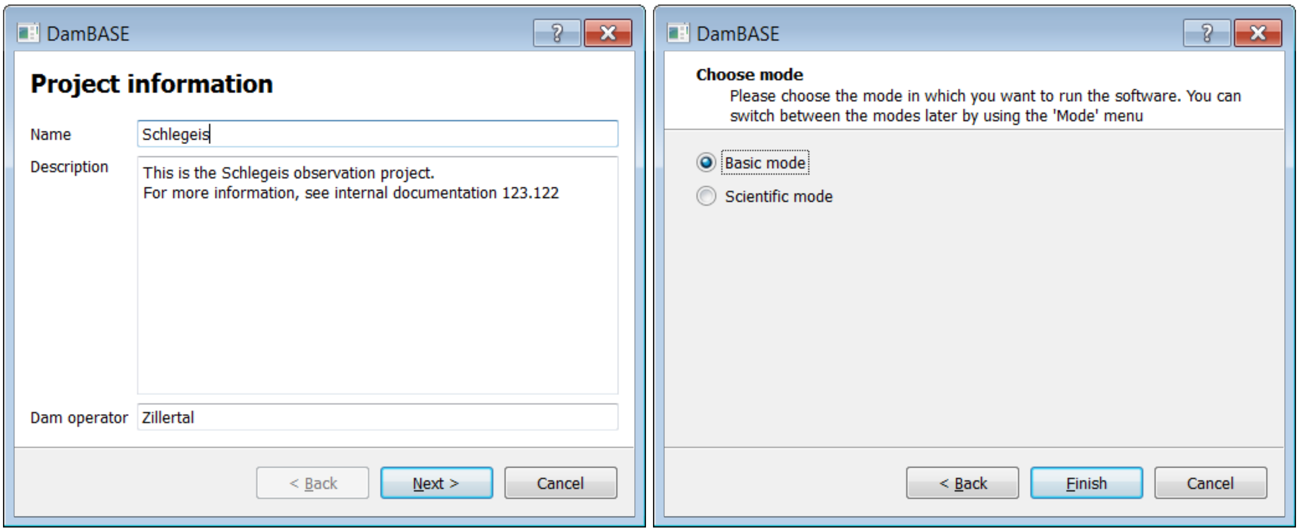
2.1 Goal

The objective of this chapter is to give the user an idea of how the main work flow looks like by covering all steps necessary to end up with a regression model which can be used for prediction/observation. The data set used within this chapter is from the Schlegeis dam in Austria as published at [ICOLD benchmark, 2001] and provided as an attachment to the software package.

2.2 Project setup

A new project can be created by clicking on the menu item "Project"->"Create project". Figure 2.1 shows the project set up wizard for project information and the preferred operation mode (the default basic user mode is used for this quick go-through).

Finally, we end up with a project overview as shown in figure 2.2 on the following page.



(a) Project information **(b)** Operation model selection

Fig. 2.1 *Project information and operation mode wizard page*

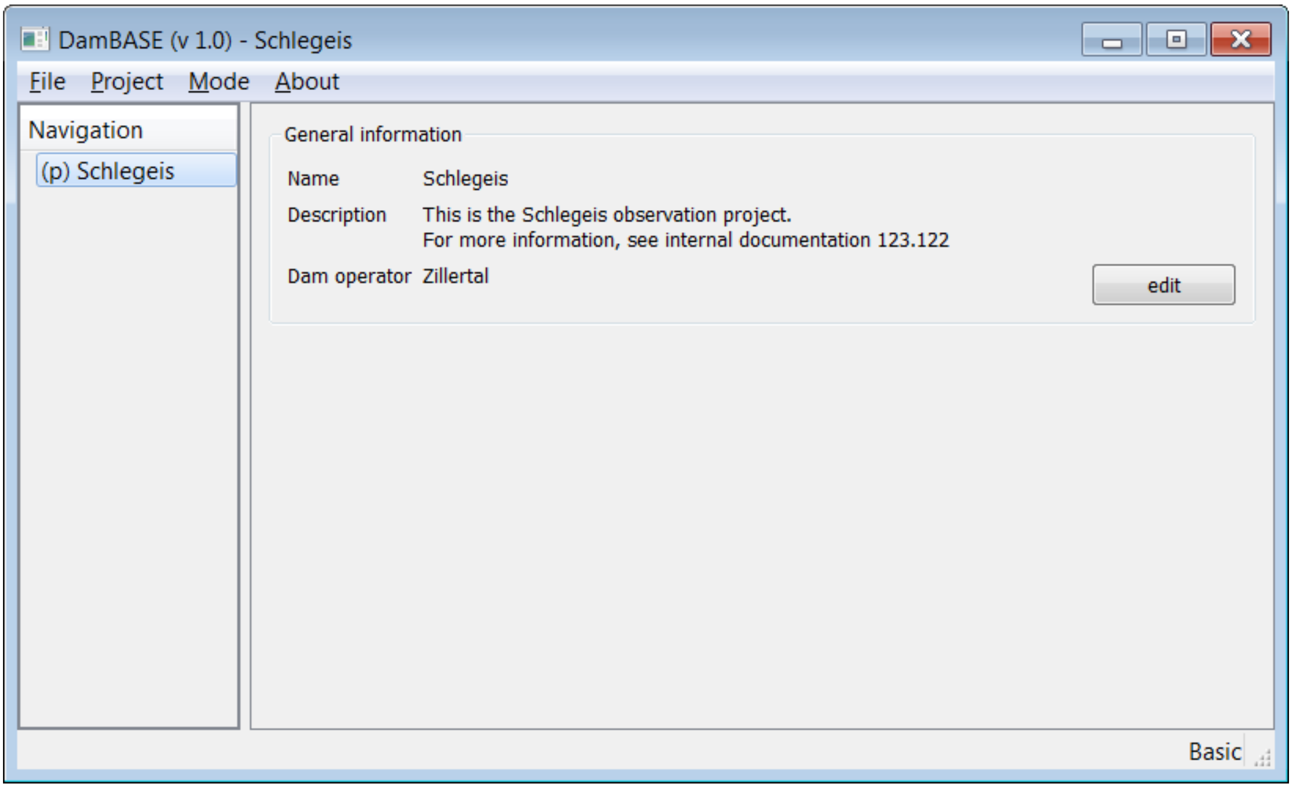


Fig. 2.2 Project overview

2.3 Set up analysis for calibration and validation

As mentioned in *Regression model set up for calibration, validation and prediction* on page 29 the main goal of calibrating a regression model is to set up a model describing the behaviour of the dam as good as possible whereas validating the model guarantees that the behaviour can be assessed by the model based on subsequent measurements not part of the calibration period.

Analysis setup

An analysis has to be set up for calibration and validation and can be created with a name and a description (figure 2.3) by right clicking on the project at the navigation tree. Finishing this wizard will lead to the analysis overview as shown in figure 2.4. From here on it is possible to define any *Water level scaling strategy* (figure 2.5) by clicking on the corresponding *Edit* button which allows to set the minimum operating level as well as the full supply level if any scaling of the water level is later required when setting up the regression model. In the case of this quick go-through, it will not be used and thus is not enabled. More about the water level scaling strategy can be found in section 4.2.1 on page 31.

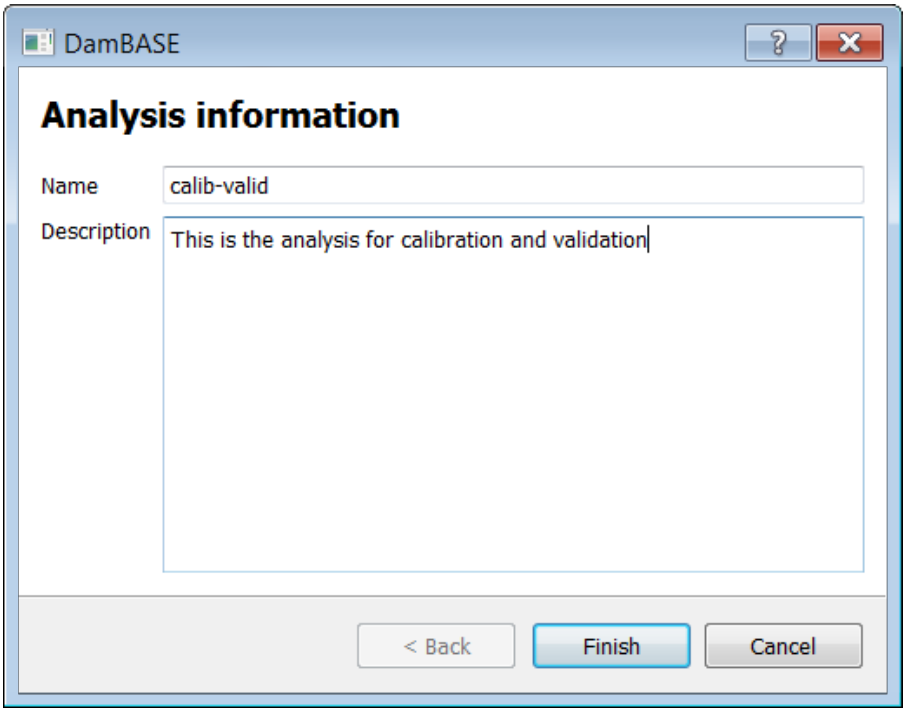


Fig. 2.3 *Create analysis*

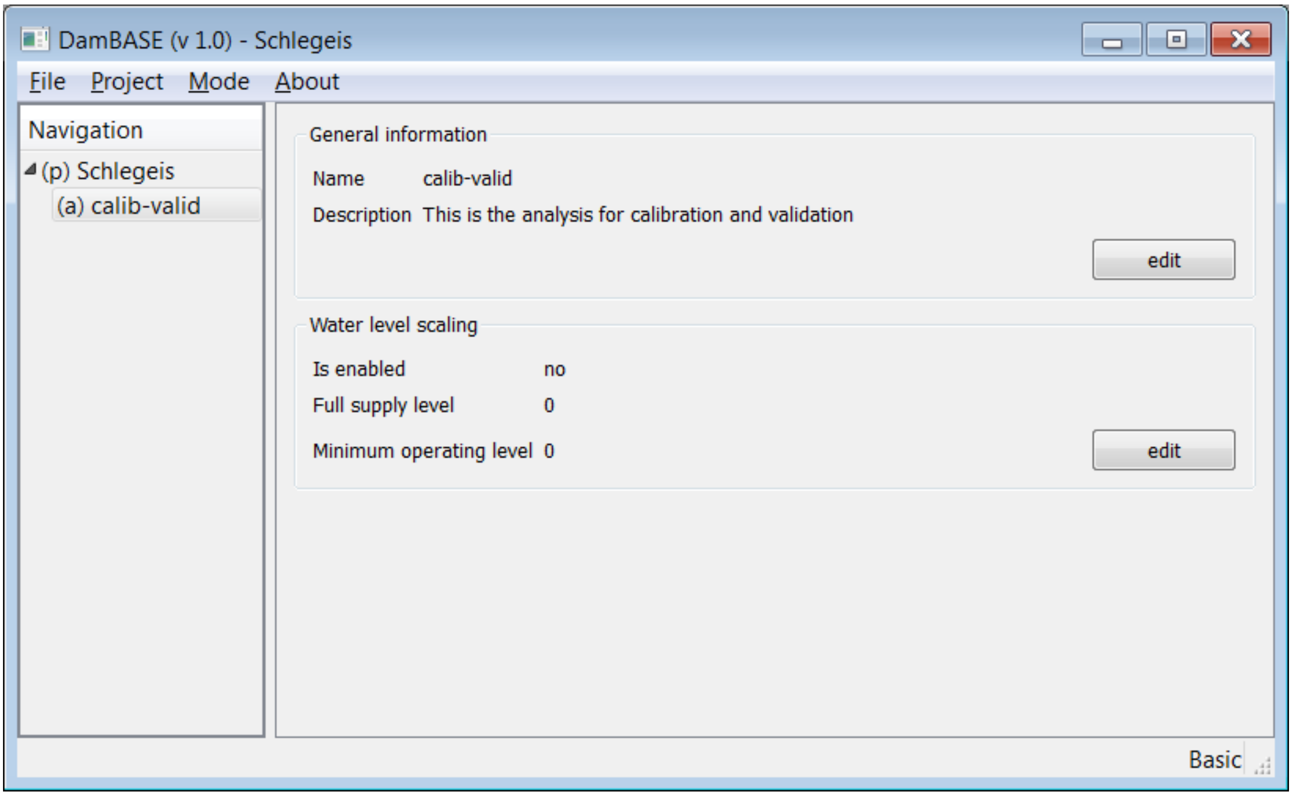


Fig. 2.4 *Analysis overview*

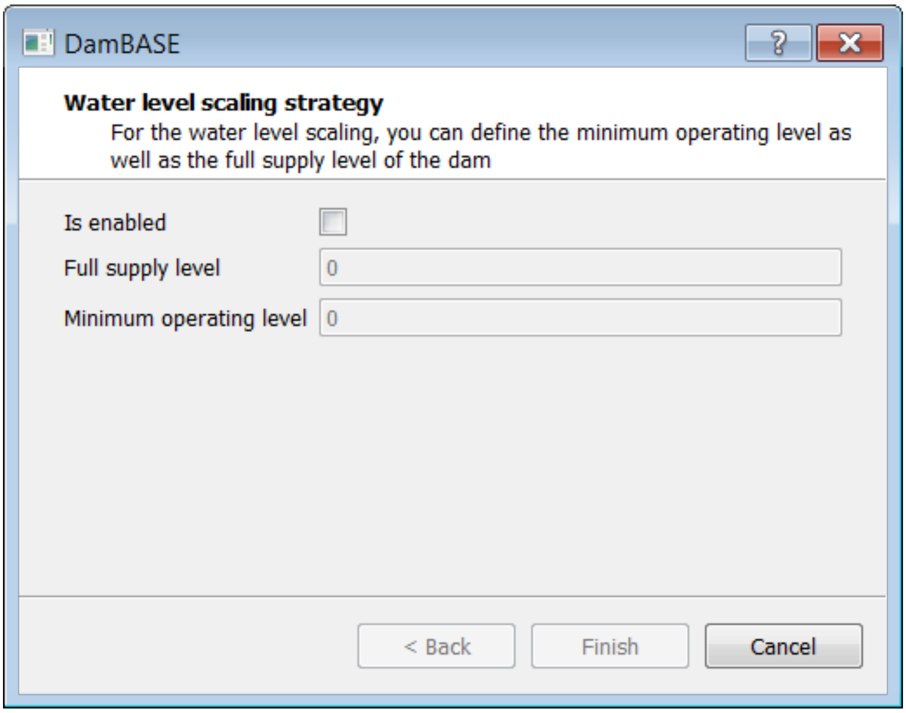


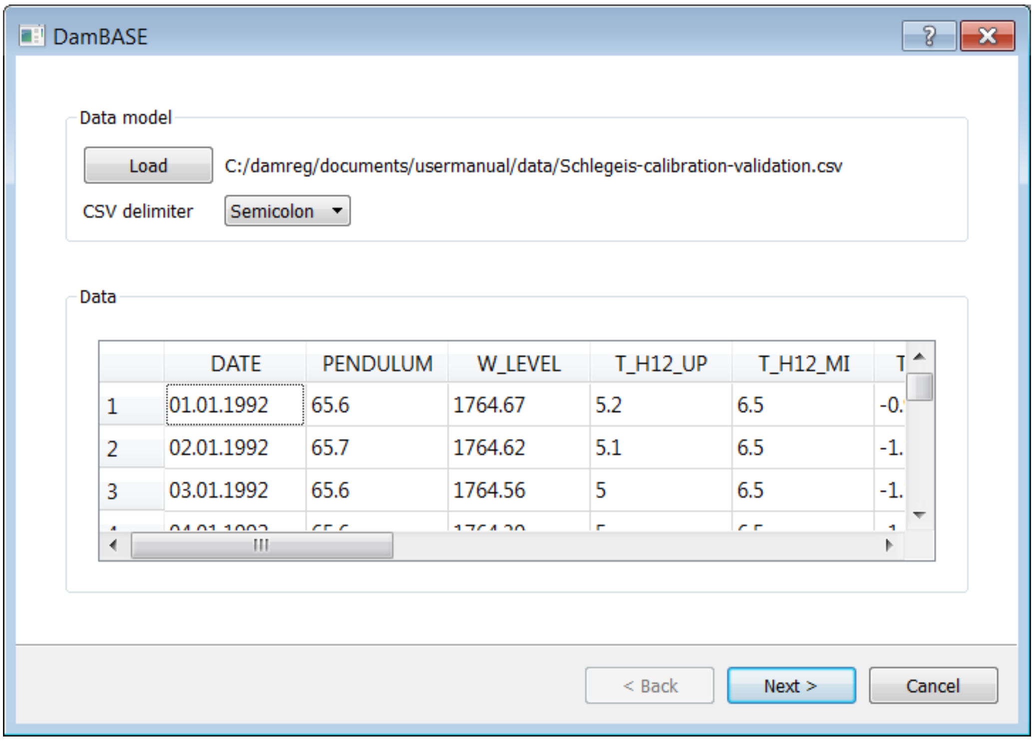
Fig. 2.5 *Water level scaling strategy*

Data model setup

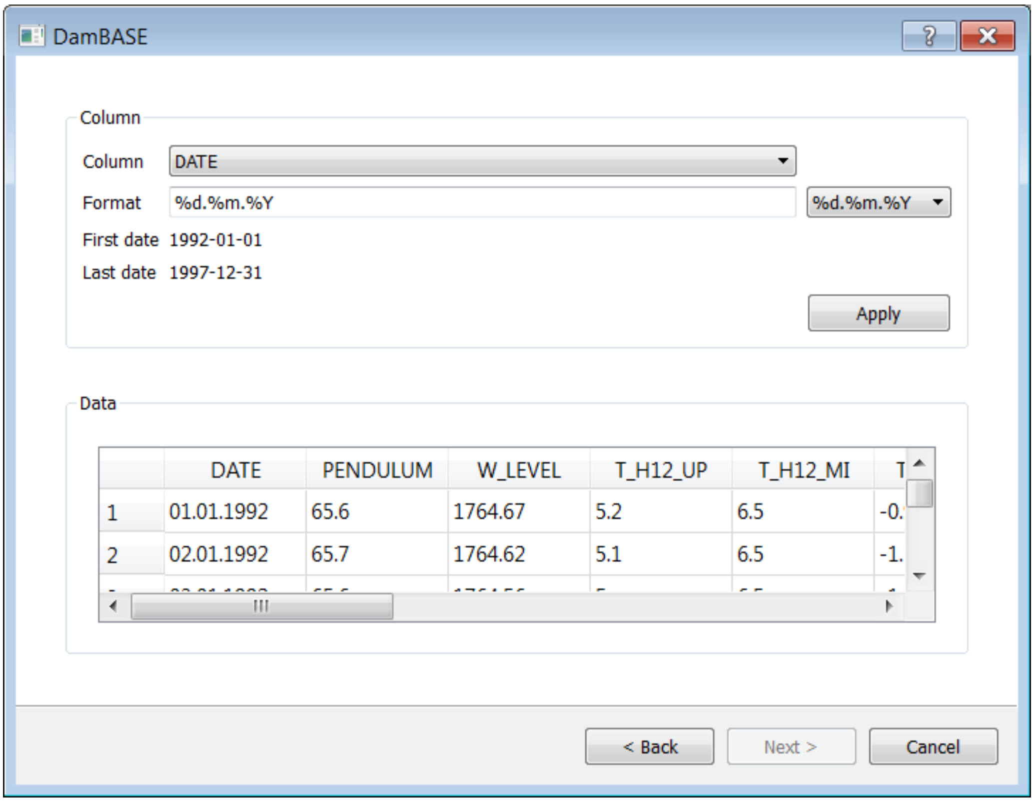
Next, the data model needs to be created by right clicking on the analysis object created in the previous step.

In the dialog, the data set file can be selected as well as the CSV delimiter (figure 2.6 on the next page). On the next wizard page, the date column has to be chosen and a date format has to be specified (standard format patterns are possible). To continue, the chosen format has to be applied to the dataset and the date values in the table view will be updated.

In the next step, the columns that represent any behaviour indicators (mostly the pendulum) and the columns which represent influence factors (such as water level and temperature) have to be assigned. The influence factors are grouped into assignment categories, namely *Water Level* and *Temperature*. It is important to note that after the regression model has been set up, the assignment done here cannot be changed any more. However, any kind of analysis on the data is still possible.



(a) Load data set dialog



(b) Apply date format dialog

Fig. 2.6 *Data model - Loading data and apply format*

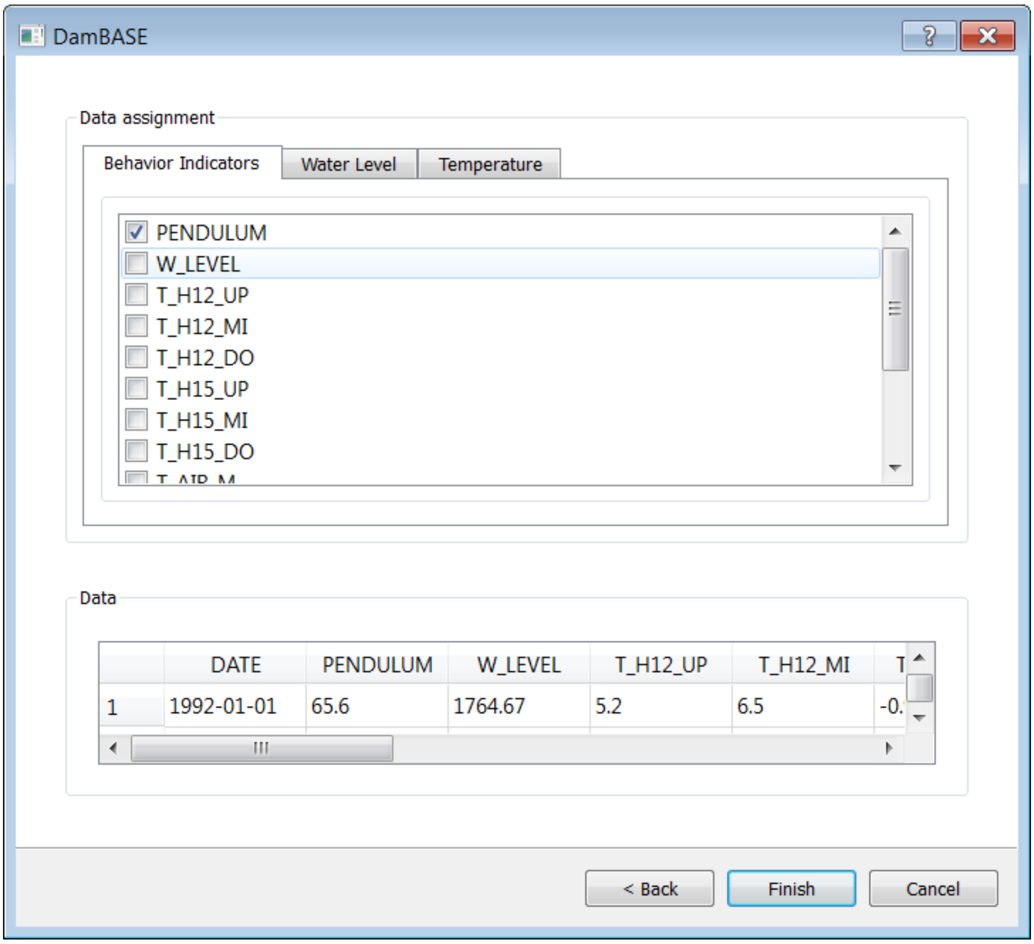


Fig. 2.7 Column assignment dialog

Data model analysis

Some predefined analysis tools related to the data model have been loaded on the working area (figure 2.8). As it is key to know how the data looks like for succeeding regression model set up, three analysis tools allow to get more insight into the data.

Therefore, the analysis tool *Time series overview* can be used to make sure that, for instance, the temperatures have been assigned correctly (value range of the y-axis as well as the periodic seasonal pattern over time). Also outliers (e.g. caused by incorrect working sensors) may be detected with this plot (see figure 3.3 in section *AT* Time series *overview - Analyse values as time series* on page 26).

On the other hand, values depending on the operation of the dam, such as the water level, can be analysed better on a scatter plot as provided by the *Influence factors* analysis tool. This graph tells a user how a dam (pendulum) behaves physically related to a change in water level.

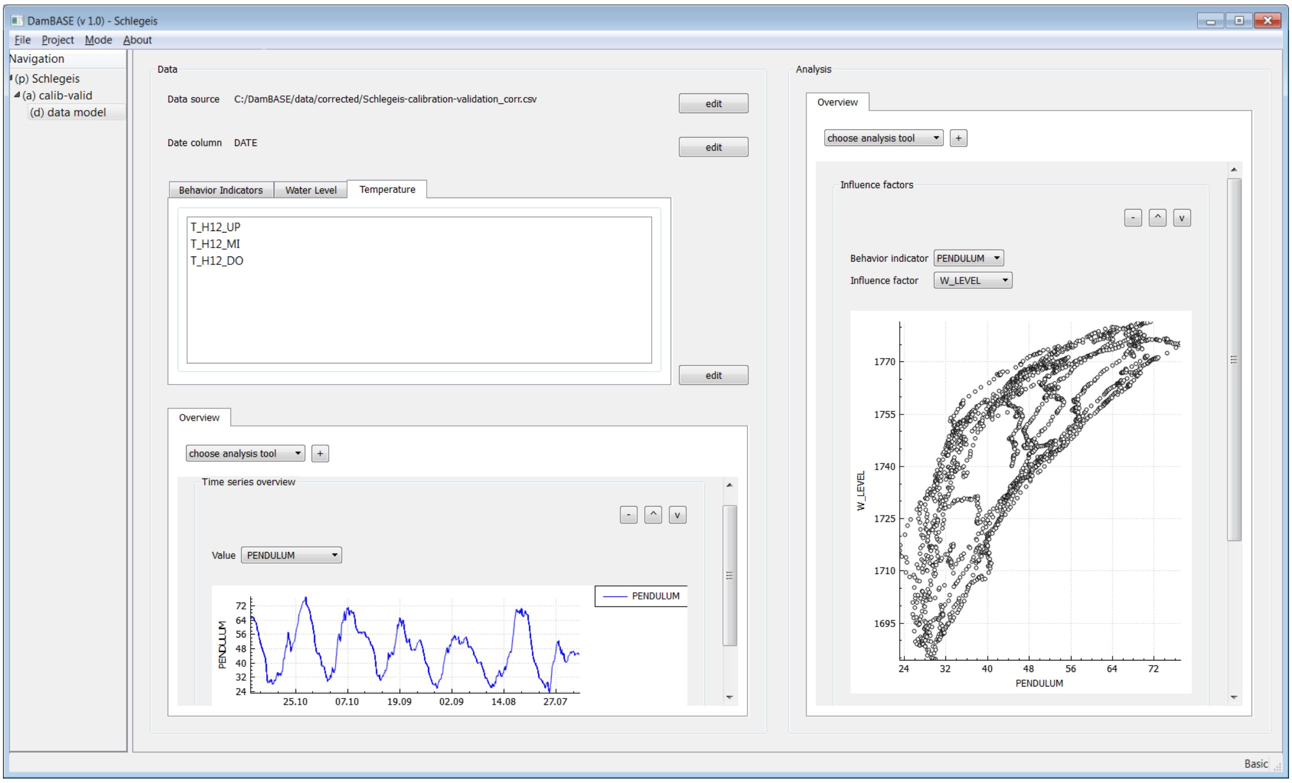


Fig. 2.8 Data model - overview with analysis tools loaded

Regression model setup

A new regression model can be created by right clicking on the analysis and choose *Create regression model* on the context menu. Next, a name and a description can be chosen (figure 2.9). It is recommended to provide this information to distinguish between multiple regression models are set up and compared to each other. Note that not only the name can be changed later time but the description as well.

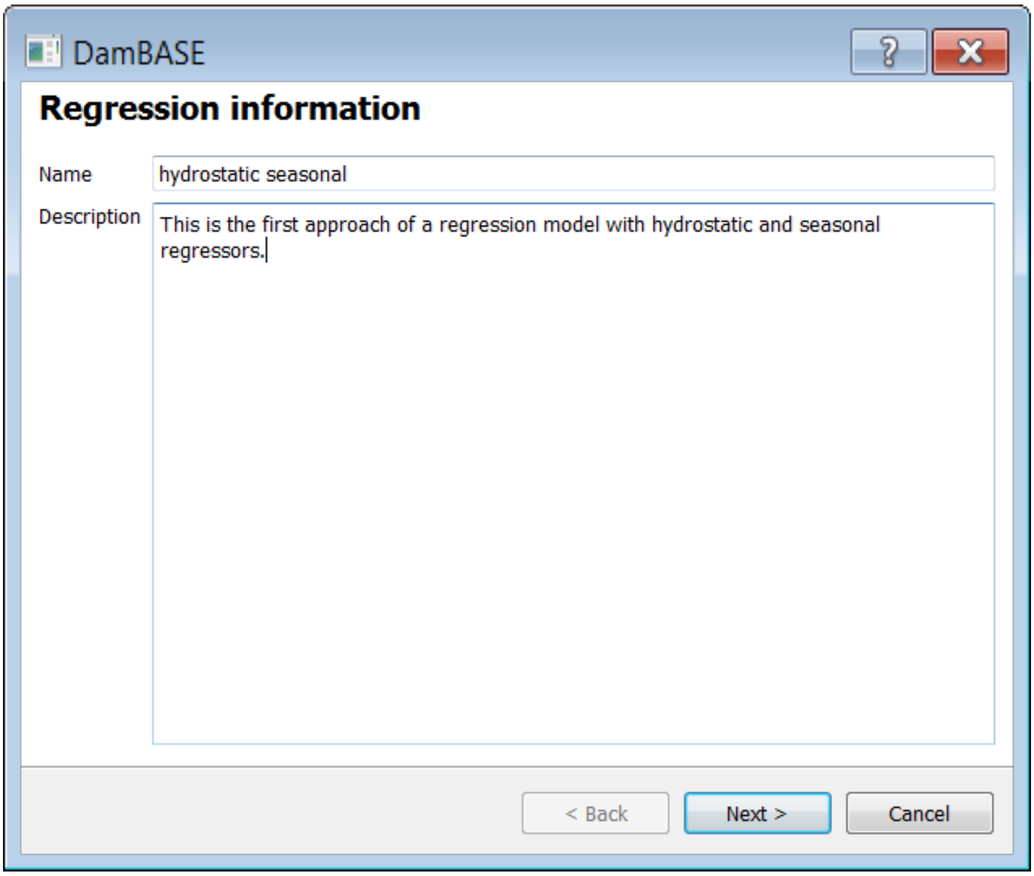


Fig. 2.9 *Regression model set up*

Regression model time range selection

On the next wizard page, a time range for the regression period can be chosen (figure 2.10). As this is the regression model set up for calibration, it is recommended to choose a time range which allows to have some validation data available after the regression period. In this example 5 years are available which are split up into 4 years for calibration and 1 year for validation (remaining data set).

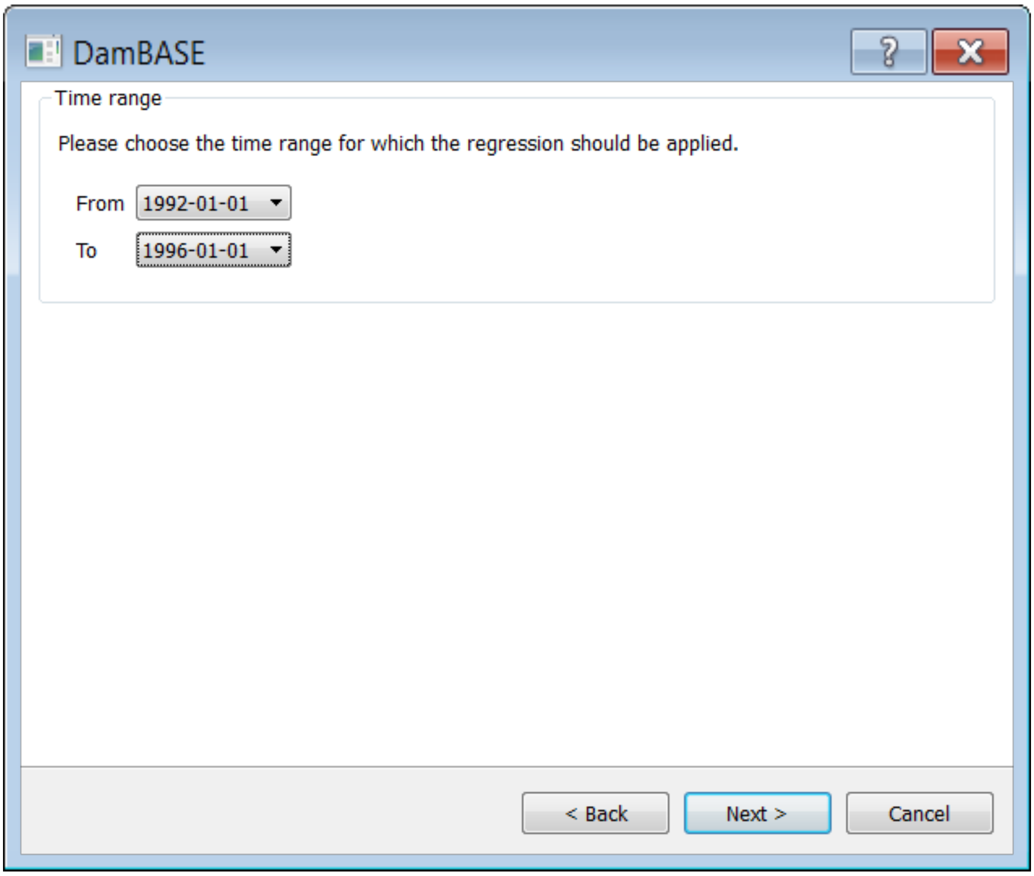


Fig. 2.10 *Regression model time range selection*

Regression model behaviour indicator selection

The behaviour indicator for this regression can be selected on the successive wizard page (figure 2.11). See chapter *Regression model setup for calibration and validation* on page 30 for more about this topic.

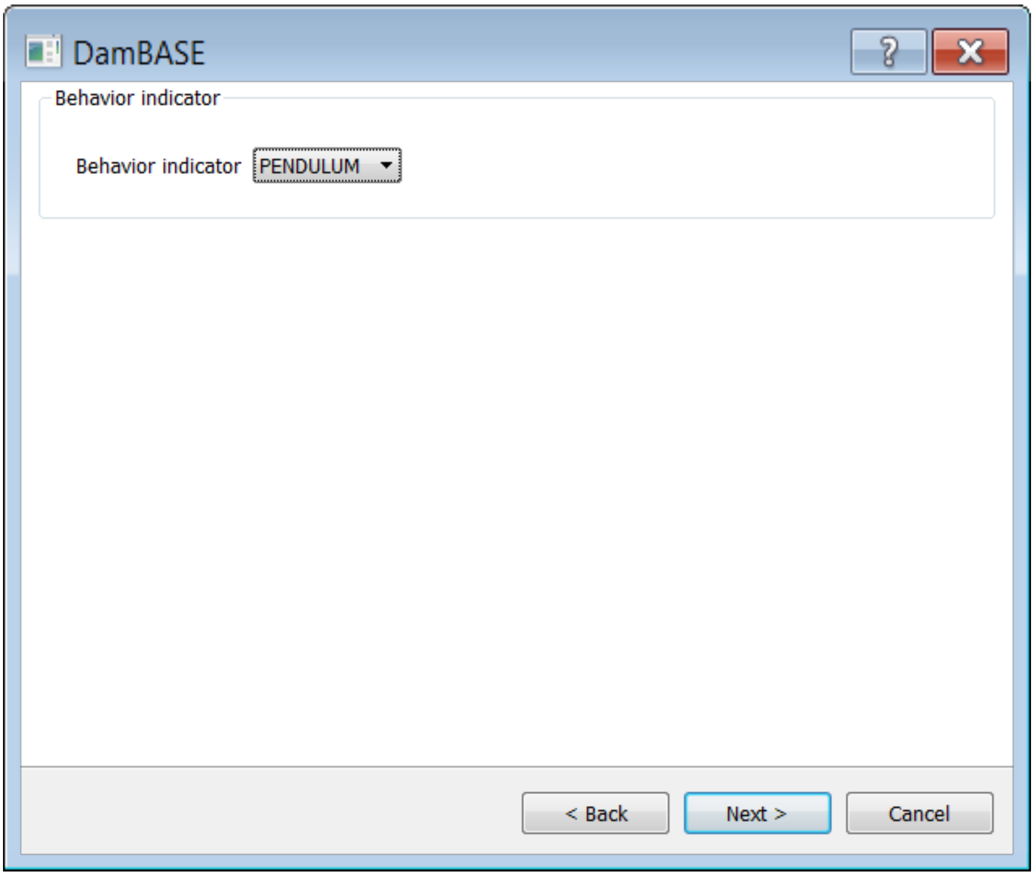


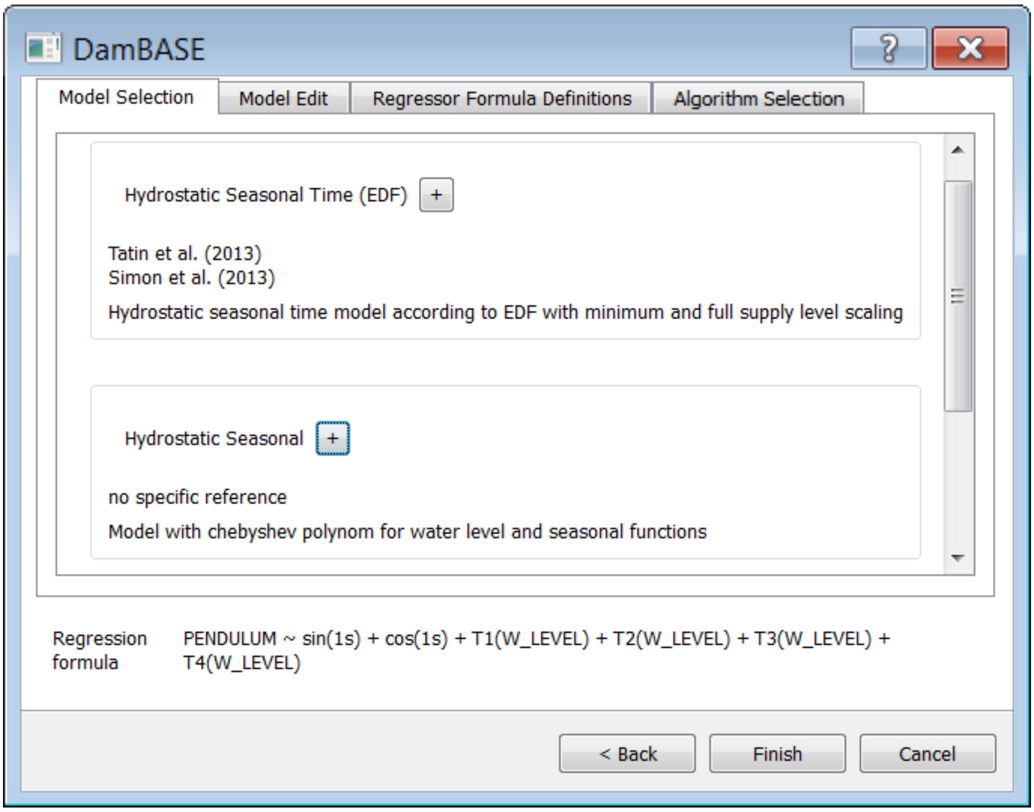
Fig. 2.11 *Regression model behaviour indicator selection*

Regression model template selection

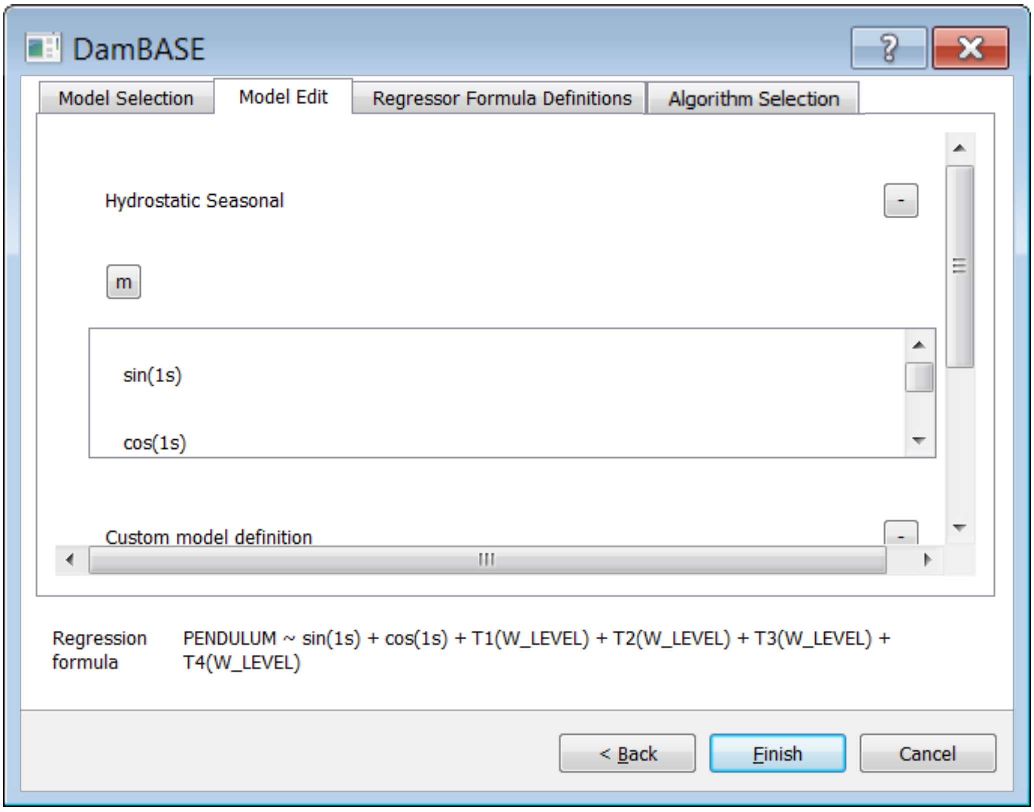
Finally, the regression formula can be set up on the next wizard page.

On the tab panel *Model Selection*, an appropriate model template, i.e. a formula, can be chosen (figure 2.12a). In this quick go-through we choose the model template *Hydrostatic Seasonal* to get a useful starting point for the first regression model. This model contains regressors (sin, cos)describing a seasonal effect as well as Chebyshev regressors describing the water level. Further, on the tab *Model Edit*, the formula can be changed in terms of adding or removing single regressors (figure 2.12b)

By clicking *Finish*, the regression is carried out (i.e. computed) and an overview with several analysis tools will be loaded onto the working area (figure 2.13).



(**a**) Model selection



(**b**) Model Edit

Fig. 2.12 *Regression model selection*

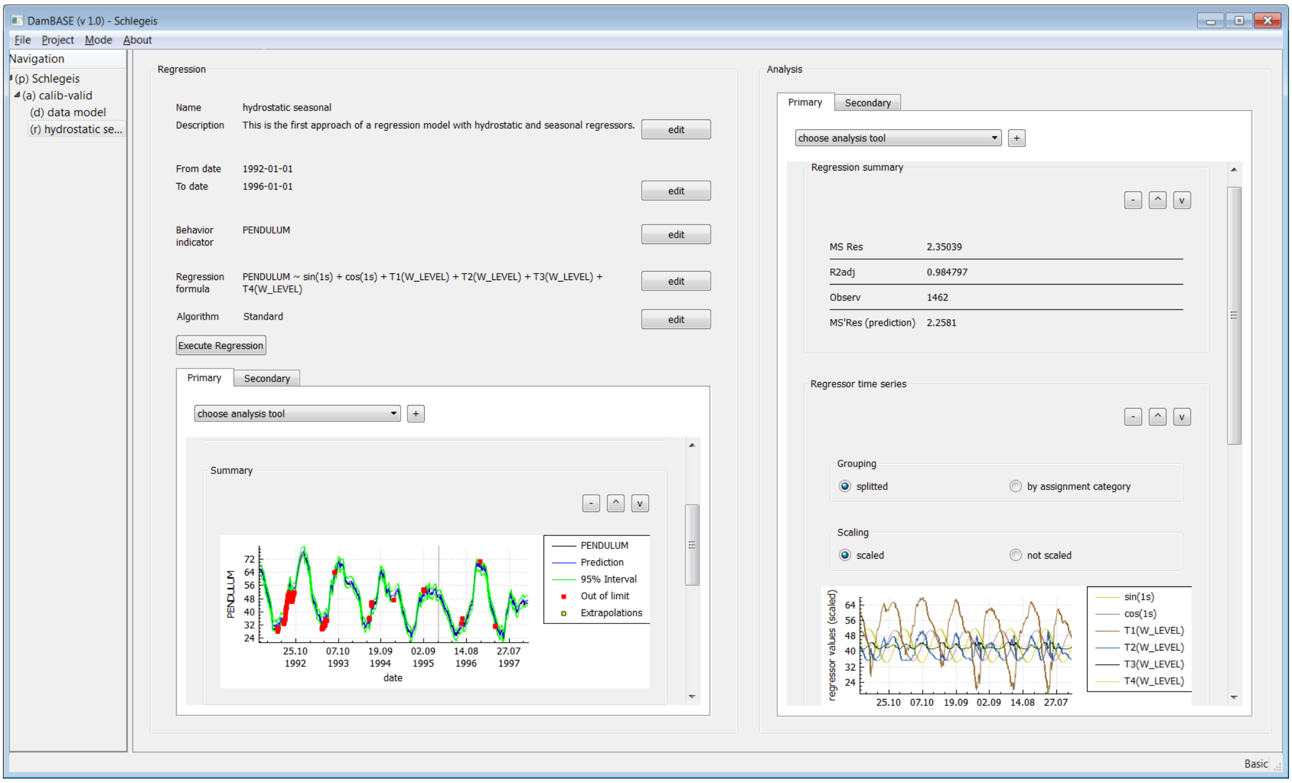


Fig. 2.13 *Regression model overview*

Regression model analysis

By setting up a data model, a set of analysis tools (AT) become available for the analysis of the regression results (figure 2.13). The area is split up into a configuration overview part including an analysis part on the left and an additional analysis part on the right. Depending on the current operation mode, different analysis tools are available by default. In basic user mode, they are related to physical analysis only.

The AT *Summary* shows the result of the regression as a time series plot split up into regression (left) and validation period (right) separated by a gray vertical line. To get an idea of how well the model describes the measurements, *MS Res* on AT *Regression summary* represents the mean square error (in the unit of the behaviour indicator) between the measured values and the calculated ones. The same is provided for values of the validation period where it is called *MS Res (prediction)*. It has to be noted that if the mean error of the validation period is different from that of the calibration, it means that the prediction may not be of sufficient validity.

If this is the case, it might be because the chosen model cannot entirely describe the real physical behaviour of the dam. To make sure this is not the case, AT *Physical behaviour* displays graphs showing how regressors related to a certain assignment category (such as water level) describing the displacement of the dam grouped by influence factors which are related to this assignment category. In the model of this quick go-through, only the water level is used as an influence factor and thus, the applied parts of the formula describing the water level (Chebyshev polynomials) are plotted against the measured water level values.

In order to get an idea of how each regressor describes the model, value plots either grouped by assignment category or split into single values are provided by AT *Regressor time series*.

Creating a model for comparison

To optimize a model additional models may be set upthat are used for comparison. multiple models can be created within the same analysis.

2.4 Set up analysis for prediction

After a regression model has been validated, the model can be used to predict new measurements. This can be done by creating a new analysis object with an extended data set (containing measurements of the calibration and validation period as well as new measurements for prediction) and a regression model setup with the formula of the before validated model. The time range in this case now covers the entire time range of the validated model (calibration and validation, see figure 1.2 on page 8) plus the period to predict. It is now possible to use the same analysis tools as above but this time for the analysis of the more recent behaviour of the dam.

Part II

Data model setup

Chapter 3

Data model setup

3.1 Introduction

Before the setup of the regression model it is important to get an idea of how the underlying data set looks like. Besides this, knowing what column represents which values facilitates the setup of a regression model. Consequently, model templates can directly be used provided that it is known which column contains water level measurements and which columns contain the temperatures values for instance. Thus a certain degree of automation can be achieved.

3.2 Data model setup and column assignment

3.2.1 Behaviour indicator and influence factors

For any dam under observation, some measurements are collected such as pendulum, water level as well as temperatures. Some of the measurements represent the behaviour of the dam (such as pendulum) whereas others have an influence on its behaviour.

The measurements that represent the behaviour are called *behaviour indicators* whereas those representing the influences are called *influence factors*. With regard to the regression model setup, this separation is key and therefore defined by the data model. Further, some data model related analysis tools such as AT *Influence factors* require to have such a separation.

It is important to note that only selected measurement data (selected by column) of assigned columns is used for further steps. For example , i.e. any kind of analysis done with analysis tools as well as for regression model setup and calculation.

3.2.2 Best practise for data model setup

After a regression model has been set up, the data model cannot be changed any more but still analysed.

Therefore, it is necessary to think about the purpose of the data model which can either be used for calibration and validation or for prediction/observation at the time of data selection. In the first case, it might be useful to assign all columns to either behaviour indicators or influence factors whereas a data model used for prediction/observation may have only those columns assigned which are used by the regression model.

3.3 Analysis tools (AT)

3.3.1 AT *Summary* - Overview of number of measurements

AT *Summary* provides an overview of how many measurements are contained within the data model as well as how many values are missing.

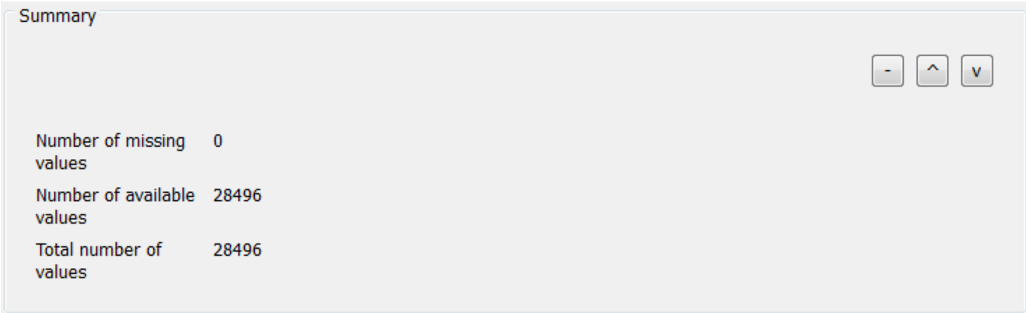


Fig. 3.1 *AT* Summary

3.3.2 AT *Time series* overview - Analyse values as time series

AT *Time series overview* plots any assigned data against the date column. This helps not only to detect any outliers (e.g. misbehaving sensors as shown in figure 3.2) but also makes sure that columns with a certain value range are assigned correctly. For example, the graph of a temperature on the air side is expected to have a certain value range (e.g. from -10°C to +25°C) and to have a seasonal trend as well. If this is not the case, the assignment might be wrong and needs to be changed.

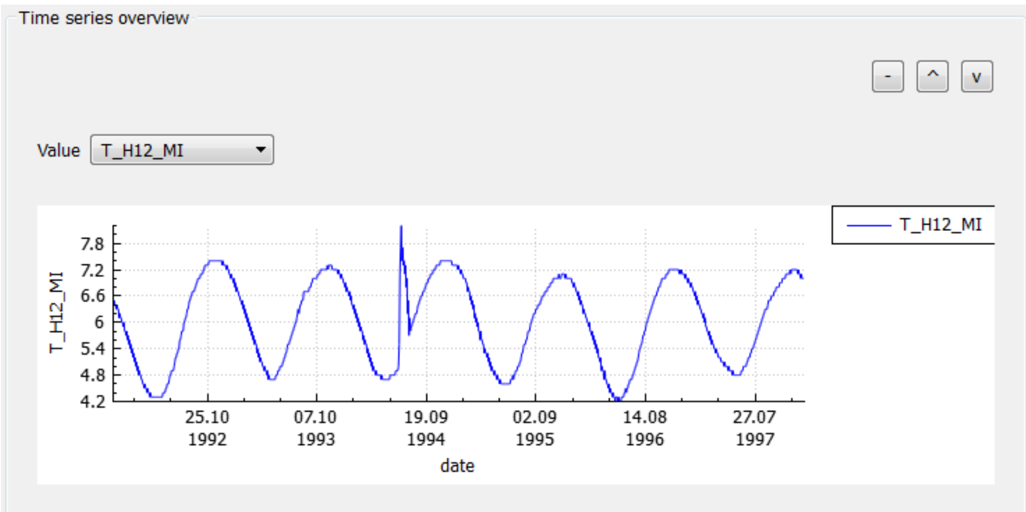


Fig. 3.2 *Outlier detected with AT* Time series overview

In case of having any unexplained measurements it might be helpful to load more than one AT below each other (see figure 3.3).

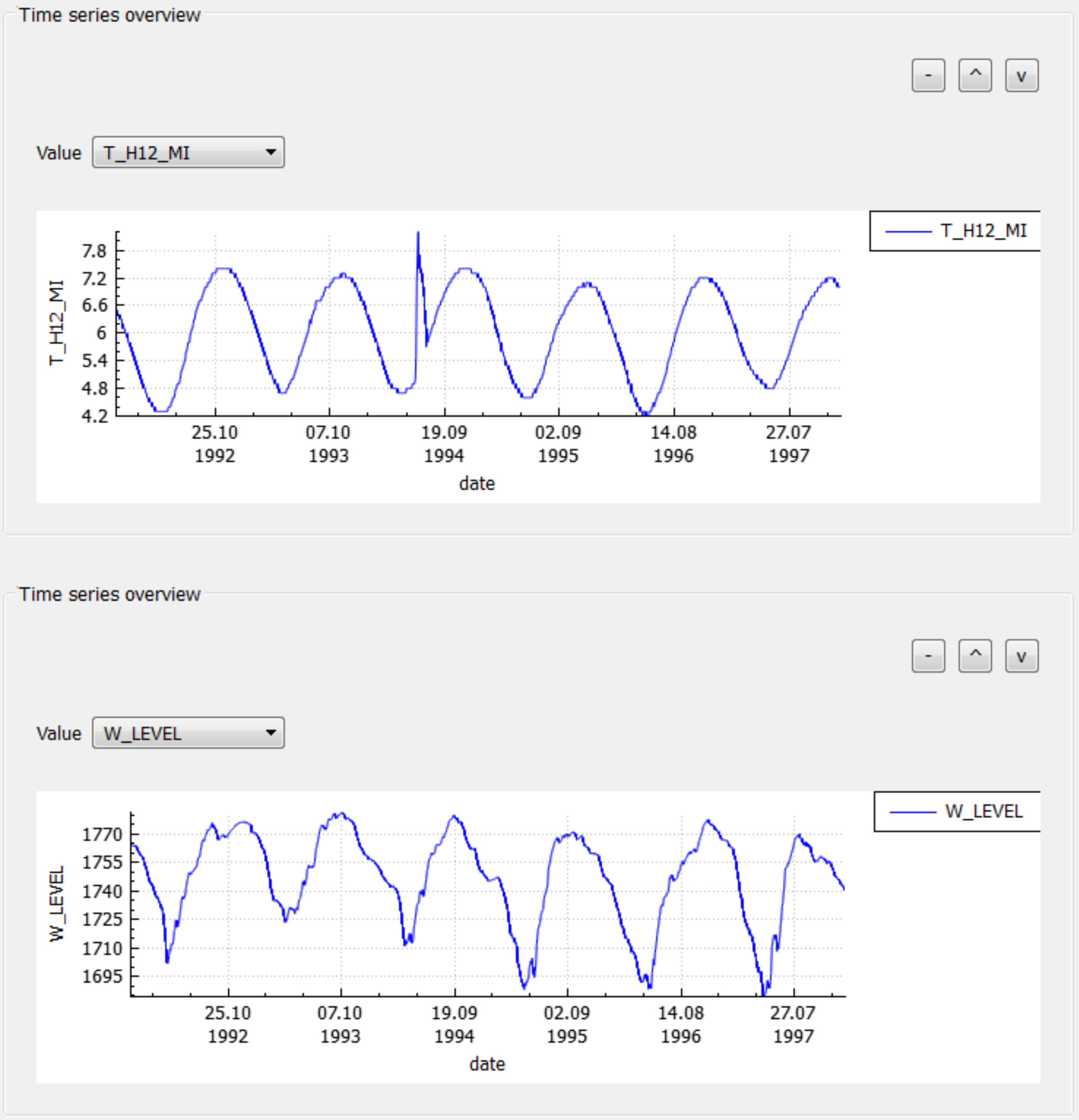


Fig. 3.3 *Multiple AT* Time series overview *open below each other*

3.3.3 AT *Influence factors* - Analyse values in relation to a behaviour indicator

Influence factors can be analysed by plotting them against the behaviour indicator (e,g., plot the water level which depends on the dam operation against the pendulum, to get an idea of how they are physically related).

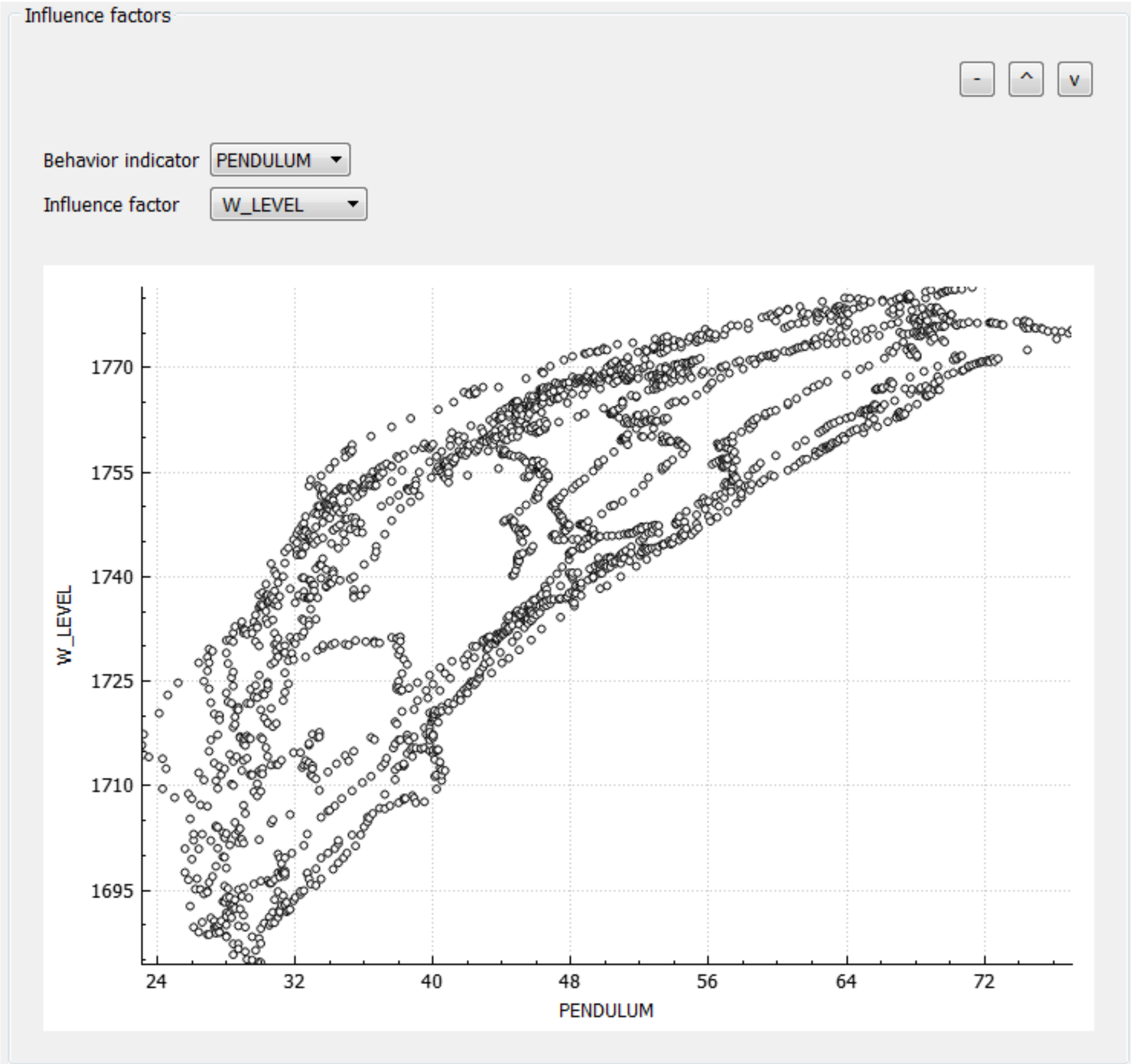


***Fig. 3.4** AT* Influence factors *with water level plotted against pendulum (displacement of the dam)*

Part III

Regression model set up for calibration, validation and prediction

Chapter 4

Regression model setup for calibration and validation

4.1 Introduction

As described in section *Main work flow* on page 8 the main goal of calibrating and validating is to create a regression model describing the behaviour of a dam as good as possible.

However, the quality of a model cannot be measured absolutely but in relation to certain properties. One of these properties is *MS Res* representing the mean square error between measured and predicted values. The same property for the validation period is called *MS Res (prediction)*.

It is important to note that in the cases where *MS Res* and *MS Res (prediction)* differ too much, it means that there are either some influences or some process related effects within the data not considered within the defined regression formula (figure 4.1).

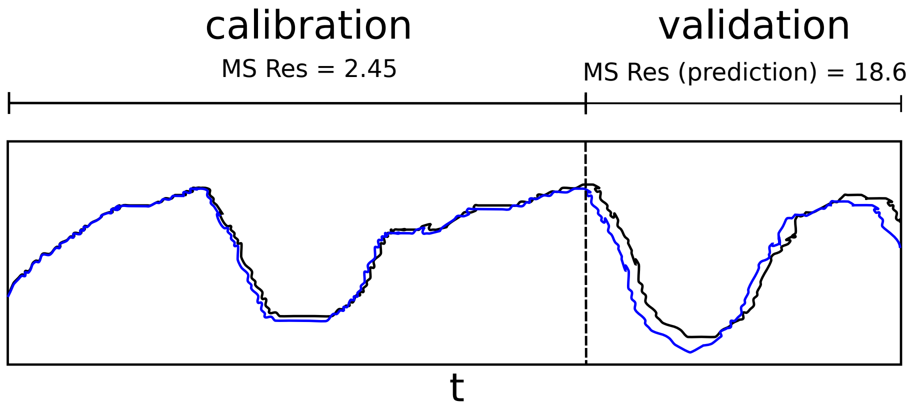


Fig. 4.1 *Differences between* MS Res *of calibration and validation period*

4.2 Analysis setup and data model selection

For calibration and validation , it is recommended to use a data set which can be split up into a period for calibrration, i.e. for the setup of the regression model and a (successive) period with enough values for validation..

Due to the ability of DamBASE to keep multiple models within a single analysis object, it may be useful to set up more than one regression model and compare them by switching between the corresponding working areas.

4.2.1 Water level scaling strategy

DamBASE allows for scaling of the water level to minimum operating and to full supply level. This feature can be enabled or disabled.

Because these values are related to a certain dam, they can be defined on the corresponding analysis (figure 4.2).

Note that scaling with this strategy is only used by polynomial water level regressors with *scaled...{min/max}* identifier pattern.

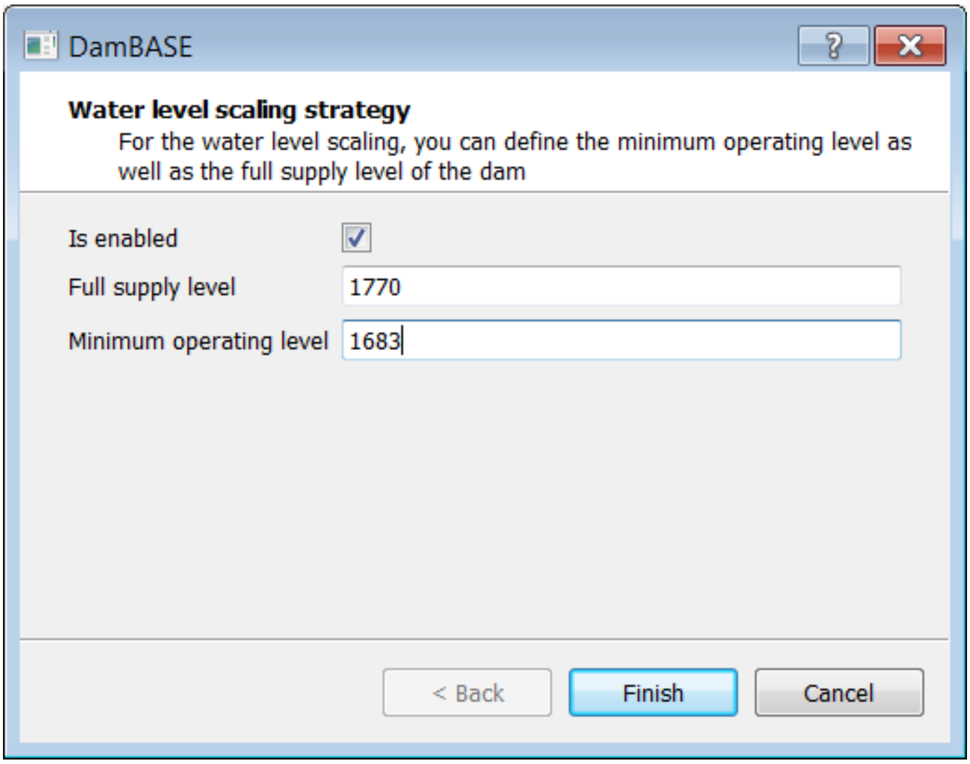


Fig. 4.2 *Analysis wizard page for water level scaling definition*

Choosing appropriate template

In order to set up a regression model, a user can choose from a list of predefined model templates. Due to the fact that every dam has an individual behaviour, a regression formula setup out of such a template is only meant to be a starting point.

As some templates assume that certain preconditions are fulfilled (e.g. temperature related influence factors are assigned on the data model), it is important to choose a template which works with the corresponding setup. If any preconditions are not fulfilled, DamBASE prevents a user from using that template.

4.2.2 Operation mode related templates

Some regressors are a special case of others (e.g. polynomial v.s. Chebyshev). If such special regressors are used within a model template, these templates are available in scientific user mode only.

4.2.3 Available templates

4.2.3.1 Hydrostatic Seasonal Time (EDF)

Model equation

$Y_0 = a_1 + a_2 \cdot t + a_3 \cdot t^2 + a_4 \cdot t^3 + a_5 \cdot t^4 + a_6 \cdot e^{-\frac{t}{\tau}} + a_7 \cdot Z + a_8 \cdot Z^2 + a_9 \cdot Z^3 + a_{10} \cdot Z^4 + a_{11} \cdot \cos(s) + a_{12} \cdot \sin(s) + a_{13} \cdot \cos(2s) + a_{14} \cdot \sin(2s)$

Scaling

Hydrostatic:	$Z = \frac{RN-H}{RN-Rvide}$
	RN: full supply level, normal water level
	R_{vide}: minimum operation level
	H: water level of measurement
Seasonal:	s: $2\pi \cdot \frac{J}{365.25}$
	J: number of day in year

Notes

Generally the linear term is used for the irreversible effects. Also note, that it is recommended to use only a subset of the terms describing time-effects.

References [Simon *et al.*, 2013][Tatin *et al.*, 2013]

4.2.3.2 Hydrostatic Seasonal

Model equation

$Y_0 = a_1 \cdot \sin(s) + a_2 \cdot \cos(s) + a_3 \cdot T_1(H) + a_4 \cdot T_2(H) + a_5 \cdot T_3(H) + a_6 \cdot T_4(H)$

Scaling

Hydrostatic:	T_n :Chebyshev polynomial of degree n
	H: water level of measurement
Seasonal:	s: $2\pi \cdot \frac{J}{365.25}$
	J: number of day in year

4.2.3.3 Hydrostatic Thermal

Model equation

$Y_0 = a_1 \cdot T_1(H) + a_2 \cdot T_2(H) + a_3 \cdot T_3(H) + a_4 \cdot T_4(H) + \sum_{i=0}^{i=n} a_{5+i} \cdot \theta_i$

Scaling

Hydrostatic:	T_n : Chebyshev polynomial of degree n
	H: water level of measurement
Temperature:	θ_i: temperature of measurement i

4.2.3.4 Water level polynom 3rd degree

Model equation

$Y_0 = a_1 \cdot H + a_2 \cdot H^2 + a_3 \cdot H^3$

Hydrostatic: H: water level of measurement

4.2.3.5 Water level polynom 4th degree

Model equation

$Y_0 = a_1 \cdot H + a_2 \cdot H^2 + a_3 \cdot H^3 + a_4 \cdot H^4$

Hydrostatic: H: water level of measurement

4.2.3.6 Seasonal with scaling 2

Model equation

$Y_0 = a_1 \cdot \sin(s) + a_2 \cdot \cos(s) + a_3 \cdot \sin(2s) + a_4 \cdot \cos(2s)$

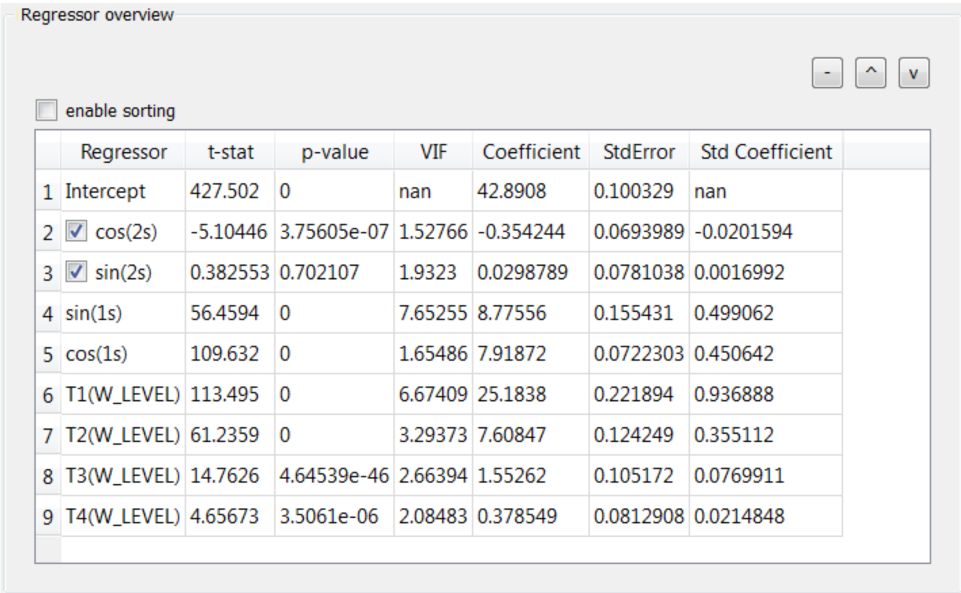
Scaling

Seasonal:	s: $2\pi \cdot \frac{J}{365.25}$
	J: number of day in year

4.3 Enabling / disabling regressors

Regressors can be enabled or disabled on the *Regressor overview* analysis tool (figure 4.3). After enabling or disabling, a new regression can be calculated by clicking on the *Execute Regression* button right below the configuration area.

However, regressors related to a model template cannot be modified and therefore neither enabled nor disabled. In such cases, the related model definition needs to be enabled for modification which can be done by clicking on the “m” button of the corresponding model definition (figure 4.4). As a consequence, the name of the model definition changes and marks the fact that it is no longer a model of a certain template with a preceding * (figure 4.5). As a result, all regressors can now be enabled or disabled (4.6)



	Regressor	t-stat	p-value	VIF	Coefficient	StdError	Std Coefficient
1	Intercept	427.502	0	nan	42.8908	0.100329	nan
2	☑ cos(2s)	-5.10446	3.75605e-07	1.52766	-0.354244	0.0693989	-0.0201594
3	☑ sin(2s)	0.382553	0.702107	1.9323	0.0298789	0.0781038	0.0016992
4	sin(1s)	56.4594	0	7.65255	8.77556	0.155431	0.499062
5	cos(1s)	109.632	0	1.65486	7.91872	0.0722303	0.450642
6	T1(W_LEVEL)	113.495	0	6.67409	25.1838	0.221894	0.936888
7	T2(W_LEVEL)	61.2359	0	3.29373	7.60847	0.124249	0.355112
8	T3(W_LEVEL)	14.7626	4.64539e-46	2.66394	1.55262	0.105172	0.0769911
9	T4(W_LEVEL)	4.65673	3.5061e-06	2.08483	0.378549	0.0812908	0.0214848

Fig. 4.3 *AT* Regressor overview *with some editable regressors*

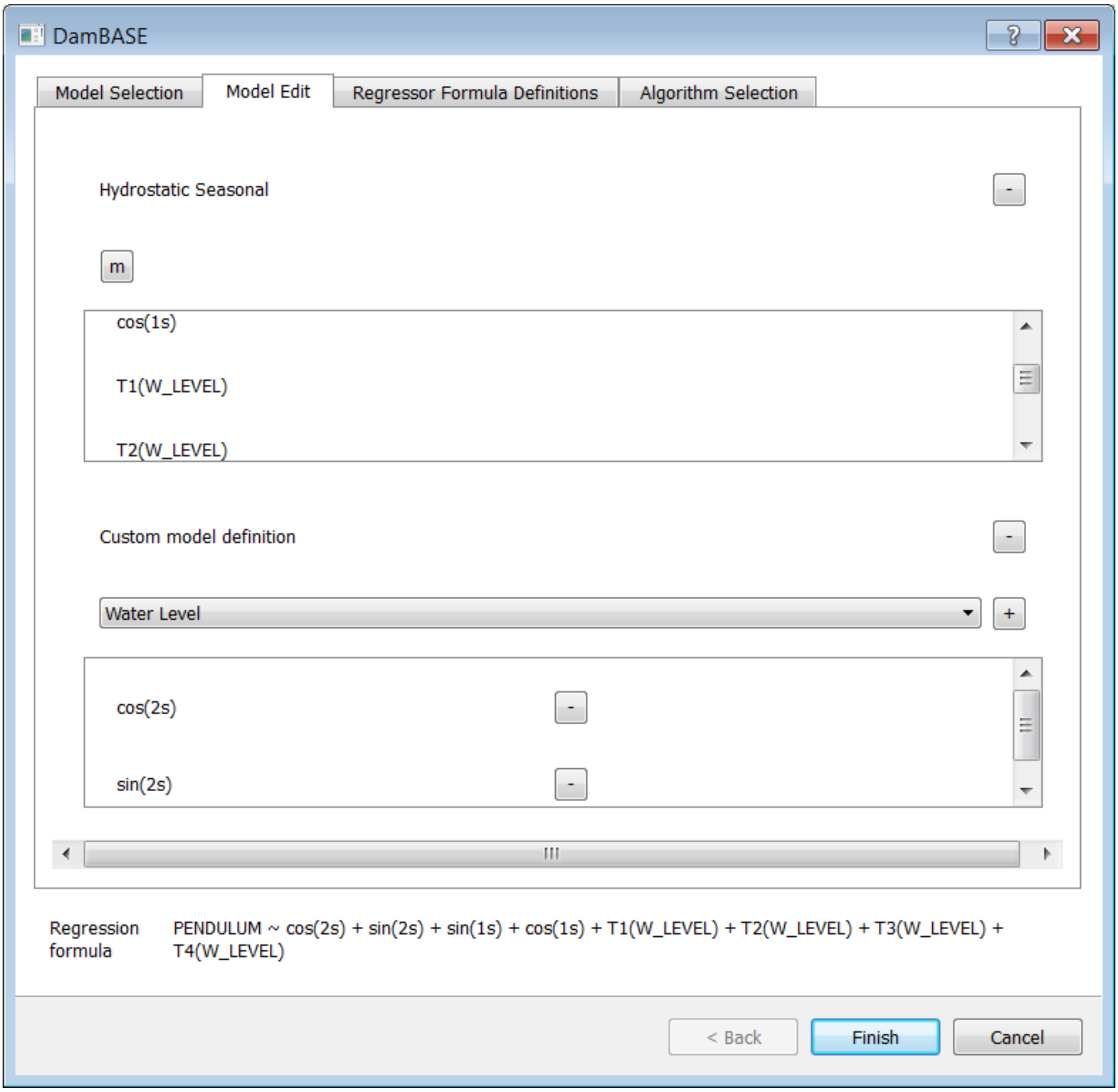


Fig. 4.4 *Before editing of the model (clicking on the 'm' button allows to derive the model definition for modification)*

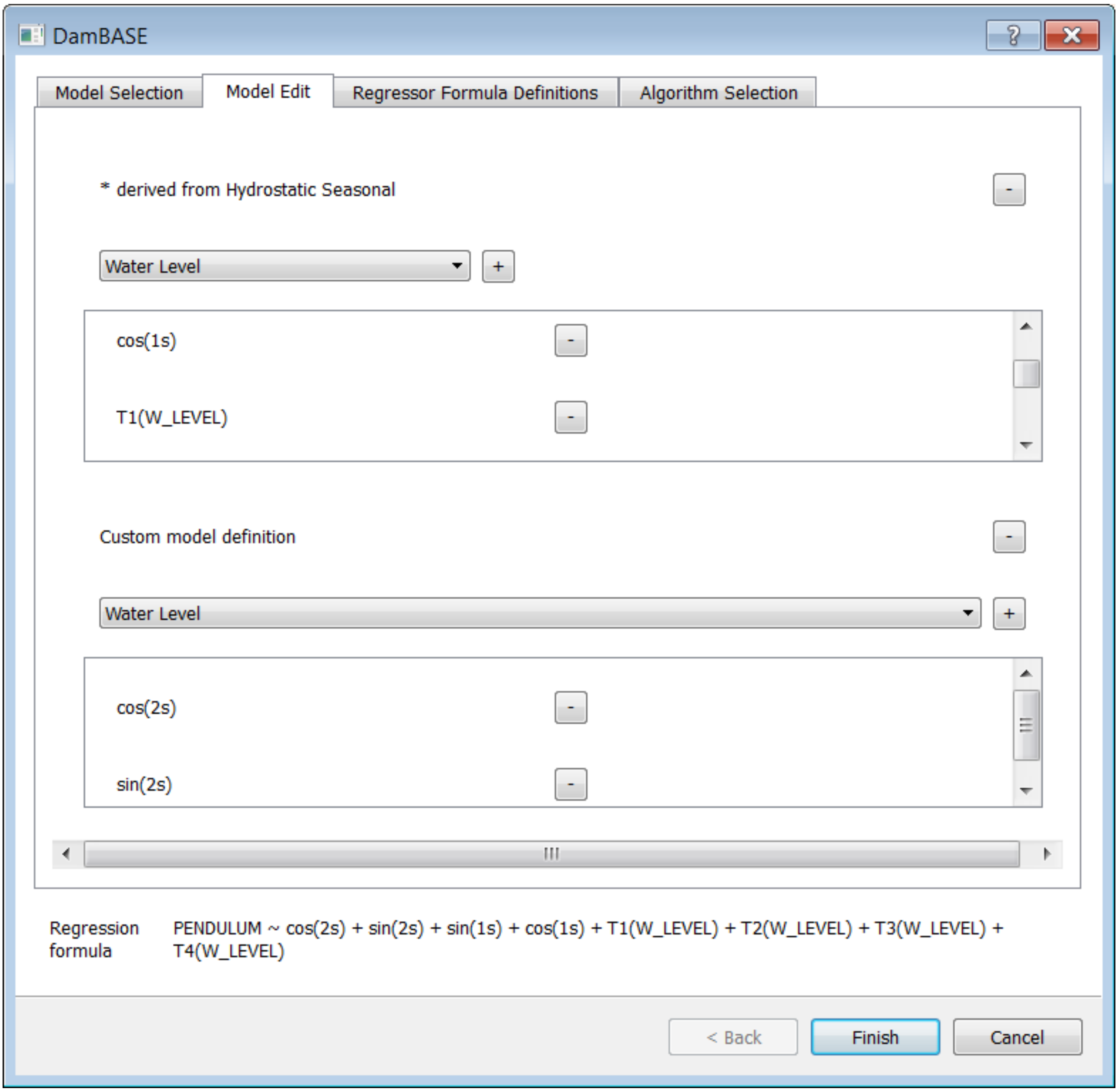
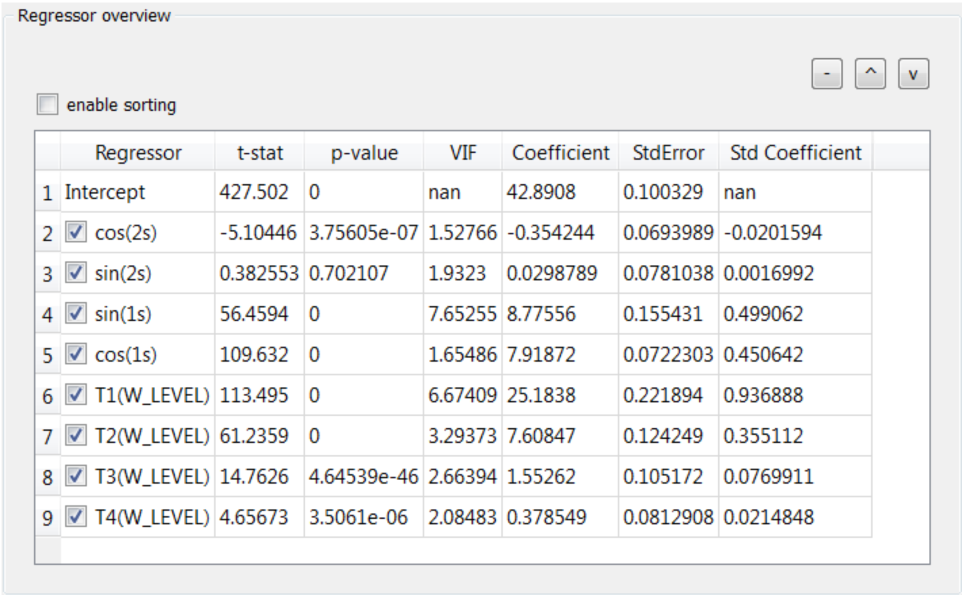


Fig. 4.5 *Derived model definition ready for editing*



	Regressor	t-stat	p-value	VIF	Coefficient	StdError	Std Coefficient
1	Intercept	427.502	0	nan	42.8908	0.100329	nan
2	cos(2s)	-5.10446	3.75605e-07	1.52766	-0.354244	0.0693989	-0.0201594
3	sin(2s)	0.382553	0.702107	1.9323	0.0298789	0.0781038	0.0016992
4	sin(1s)	56.4594	0	7.65255	8.77556	0.155431	0.499062
5	cos(1s)	109.632	0	1.65486	7.91872	0.0722303	0.450642
6	T1(W_LEVEL)	113.495	0	6.67409	25.1838	0.221894	0.936888
7	T2(W_LEVEL)	61.2359	0	3.29373	7.60847	0.124249	0.355112
8	T3(W_LEVEL)	14.7626	4.64539e-46	2.66394	1.55262	0.105172	0.0769911
9	T4(W_LEVEL)	4.65673	3.5061e-06	2.08483	0.378549	0.0812908	0.0214848

Fig. 4.6 *All regressors are selectable*

4.4 Available regressors

This section describes all regressors available in DamBASE. If any regressor not listed here is required, a new one can be derived from an existing one and modified to any new requirement (see *Customizing regressors* on page 44 and section *Creating new regressors* on page 45 for how to do this).

Because some assignment categories require to have an influence factor of a certain assignment, any *direct_x* regressors need such an influence factor on instantiation. This is important if you intend to use a temperature influence factor directly then a *direct_temp* regressor needs to be used rather than *direct_avg* because in this case, only temperature influence factors are available when using such a regressor within the formula. Also, the software prevents to use a regressor with a certain influence factor required if no such influence factor is assigned on the data model.

All time-related regressors use time steps for their calculations. This is important to note because depending on the distribution and frequency of the data, some approaches will not work appropriate if the data properties do not fulfil certain conditions.

Water level regressors

Tab. 4.1 *Water level regressors*

Identifier	Formula	Description
Chebyshev polynom of degree n	$T_n(\$)$	Chebyshev polynom up to degree n related to influence factor $\$$
Polynomial of degree n	$h(\$)^n$	Polynomial up to degree n
Scaled polynomial of degree n (min oplevel and full supply level)	$h(\$)^n\{min/max\}$	Polynomial up to degree n. Influence factor used will be scaled to minimum operation and full supply level (see *Water level scaling strategy* on page 31 for more on how to define levels)

Seasonal regressors

Note that all seasonal regressors are intended to be used for temperature approximation in cases where no temperature data is available. Therefore, they are synchronized to the day of the year of the regression start date. If other seasonal approaches are required, a new regressor needs to be defined (section 4.7 on page 45). Also note that *sin* as well as *cos* regressors are intended to be used as pairs with the same scaling factor.

Tab. 4.2 *Seasonal regressors*

Identifier	Formula	Description
sin(ns)	$sin(ns)$	s: $2\pi \cdot \frac{J}{365.25}$ n: scaling factor J: number of day in year
cos(ns)	$cos(ns)$	s: $2\pi \cdot \frac{J}{365.25}$ n: scaling factor J: number of day in year

Drift regressors

Drift regressors support modelling of long time behaviour (e.g. creep or inelastic deformation of the foundation).

Tab. 4.3 *Drift regressors*

Identifier	Formula	Description
Exponential	$exp(-t/n)$	Exponential drift with divisor n
Square root	$sqrt(t)$	Square root
Direct	t	Direct usage of t
Ln	$ln(1+t)$	Logarithm naturalis of 1 + t
t^n	t^n	exponential t

Temperature regressors

Temperature regressors like 1D heat equation can be used to calculate temperature on a certain depth out of surface values. They are implemented in semi-infinit domain with a convolution integral.

Because small depths near the surface respond to high frequency variations, it is recommended that the following depth values are chosen depending on the frequency of the data:

Tab. 4.4 *Frequency of the data for a given depth from the surface of the wall*

frequency	depth
<= 1 week	2m
<= 1 month	4m

Tab. 4.5 *Temperature regressors*

Identifier	Formula	Description
direct temperature	$direct\_temp(\$)$	Direct use of temperature assigned influence factor *$*
1D heat equation on depth *n*	$temp1d(\$-n)$	One dimensional heat equation for a depth of n

Time-averaging regressors

Regressors for averaging time.

Tab. 4.6 *Average regressors*

Identifier	Formula	Description
Average with *n* time steps	$avg(\$,n)$	Average with n time steps for influence factor *$*
direct average	$direct\_avg(\$)$	Direct use of influence factor *$*

Delay regressor

Regressor for time shifting of n steps.

Tab. 4.7 *Delay regressor*

Identifier	Formula	Description
Delay of *n* time steps	$delay(\$,n)$	Apply delay of *n* time steps to influence factor *$*

Extra regressors

Regressors not related to any other category. See description in table 4.8.

Tab. 4.8 *Extra regressors*

Identifier	Formula	Description
$*steps($-100)	$\$ * steps(\$ - 100)$	$step(x) \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$
sqrt(ramp($-100))	$sqrt(ramp(\$ - 100))$	$ramp(x) \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$
spline	$spline(\$)$	Spline function. X and Y coordinates can be provided by the corresponding user parameters as R vectors. E.g. : c(1, 2, 3) Please mind the spaces necessary between the coordinates to work properly!

4.5 Structure of a regressor

A regressor in DamBASE consists of the following elements:

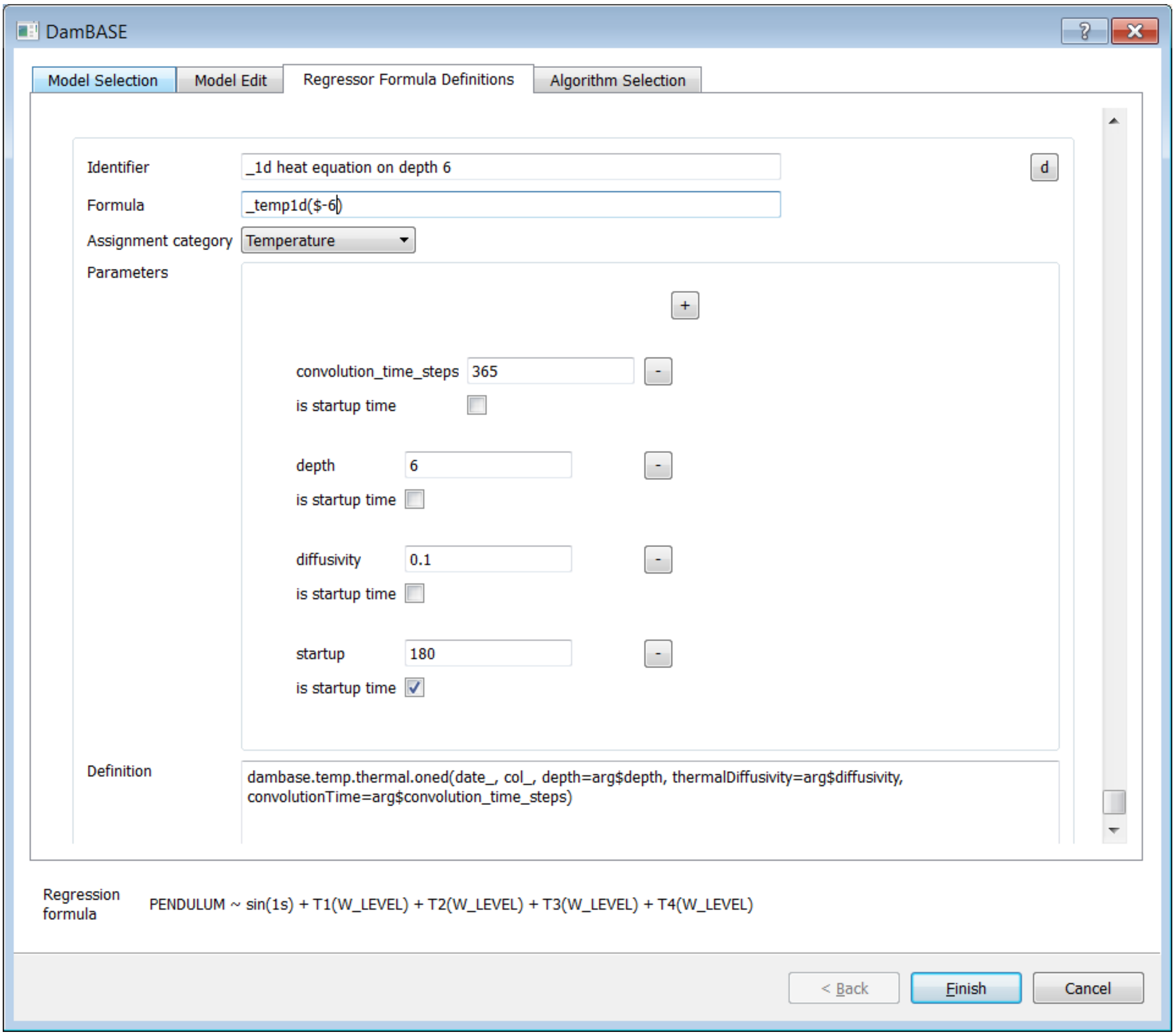


Fig. 4.7 *Regressor Formula Definition*

Identifier

Identifies a regressor with a symbolic name. This identifier will appear in the drop down menu on the model definitions area of the *Model Edit* tab (see figure 4.4 for an example).

Formula

This property assigns a symbolic formula to a regressor. However, the final formula used by the regressor is defined by property *Definition*. Compared to property *Identifier*, this property is meant to be used all over the software to identify the regressor from a mathematical point of view.

In cases where a regressor requires to work on an influence factor, a *$* can be used to highlight the influence factor within the *Formula*. For instance, if this property looks like $T5(\$)$ and later this regressor gets used within a regression formula and is assigned with water level column called *W_LEVEL* the final formula extends to $T5(W\_LEVEL)$.

Assignment category

This property is used to identify the assignment category a regressor belongs to. Some assignment categories require to have an influence factor of a certain assignment of the data model (water level or temperature) and some do not (see *Data model setup and column assignment* on page 25).

Tab. 4.9 *Influence factors by assignment category*

Assignment category	no influence factor / date column is used	water level assigned influence factor	temperature assigned influence factor	any influence factor
Water Level		x		
Seasonal	x			
Drift	x			
Temperature			x	
Average				x
Delay				x
Any influence factor				x

Parameters and Definition

Property *Definition* contains an R script being the implementation of the *Formula*. To parametrize this formula, user defined *Parameters* can be declared. For example, if the script requires a value as an argument, such as the degree of a polynomial regressor, this value can be passed to the script by adding a parameter called *degree*. Within the script, this parameter can be accessed by a preceding *arg$* followed by the name of the parameter. The final part of the script in this case would be *arg$degree*.

This concept allows customizing a regressor without having to change the script which might be error-prone.

Start-up time

Some regressors might require to have a start-up time before they are valid (e.g. 1D heat-equation regressors). In such cases, a flag called *is start up time* can be selected next to the parameter property to define a parameter as being a start-up value in time steps. Note that in these cases, the minimum start date of the regression will be changed automatically to a later date, i.e. original date + start up time.

Available properties in Definition

As mentioned above, a *Definition* consists of a full working R script and can be parametrized by user defined parameters. Besides this, other properties are available such as the date column and in cases of having an influence factor assigned, the entire column of this influence factor can be accessed as well (e.g. Assignment category is set to Water level, the water level column will be available within Definition through expression *col_*).

Tab. 4.10 *Properties available in Definition*

property	usage	always available	only available if an influence factor is required (see table 4.9)
user defined parameter xyz	*arg$xyz*	yes	
date column	*date_*	yes	
influence factor	*col_*		yes

4.6 Customizing regressors

Regressors can be customized on the corresponding *Regressor Formula Definitions* tab of the model set up wizard page.

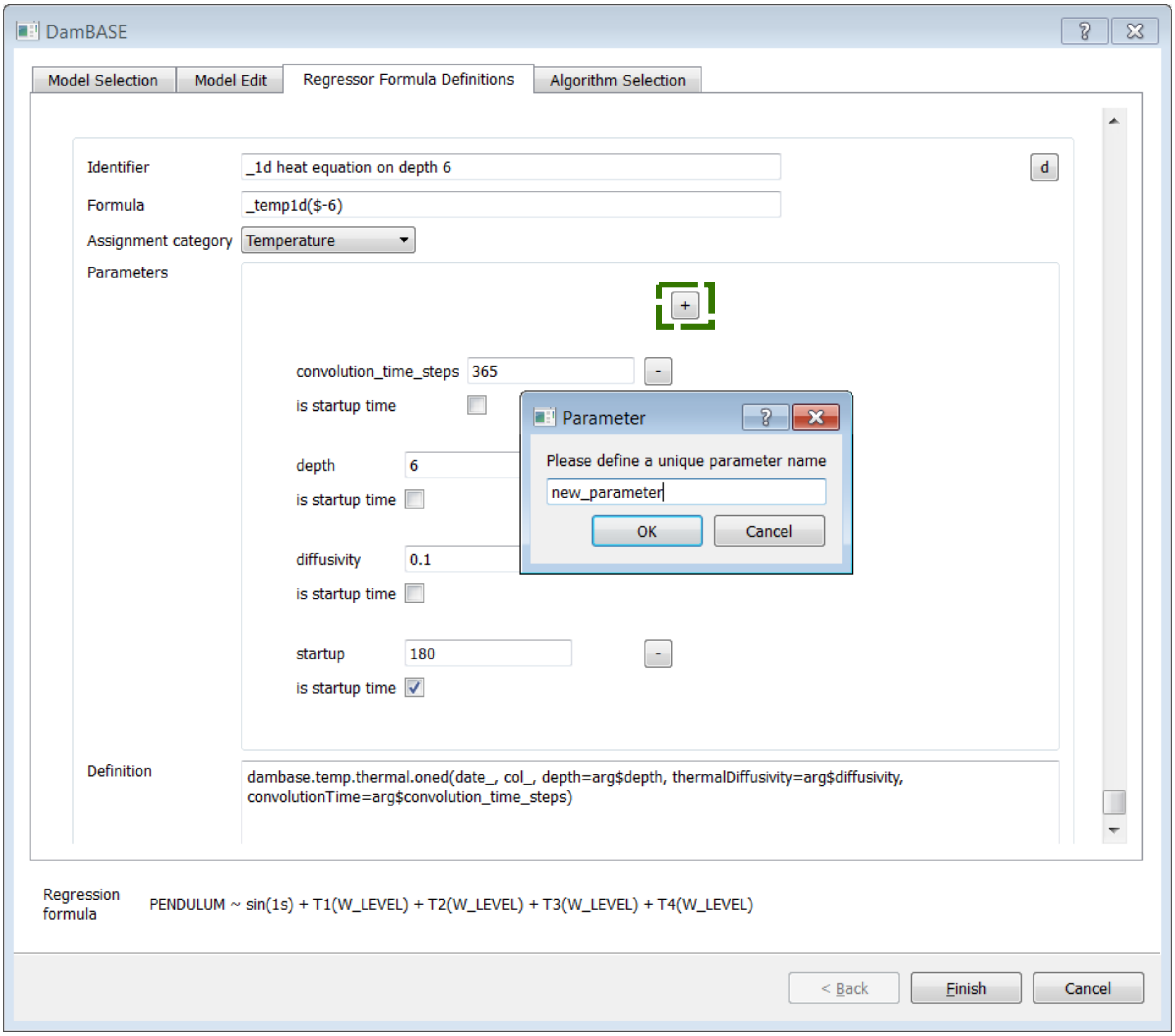


Fig. 4.8 *Create new parameter*

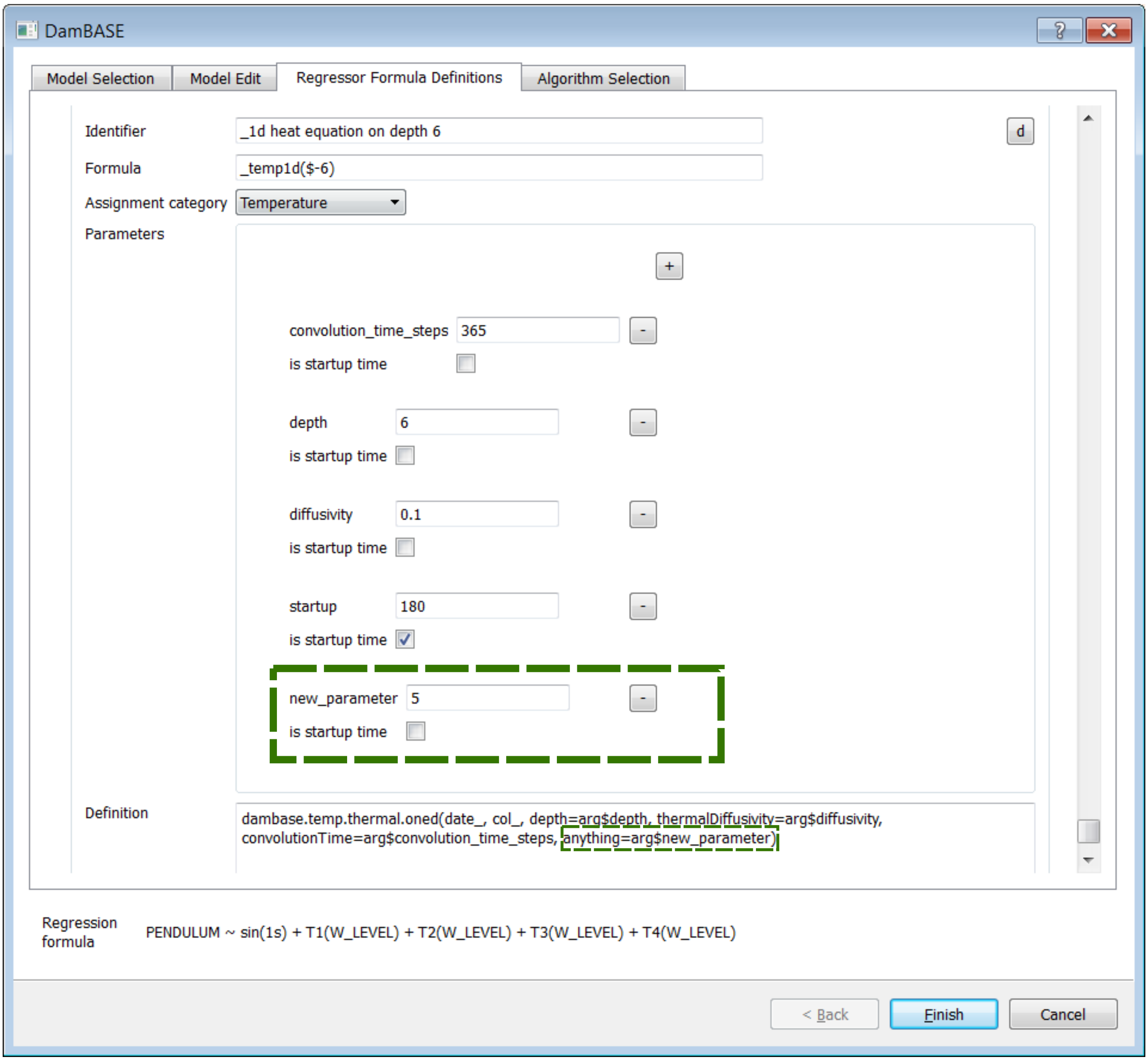
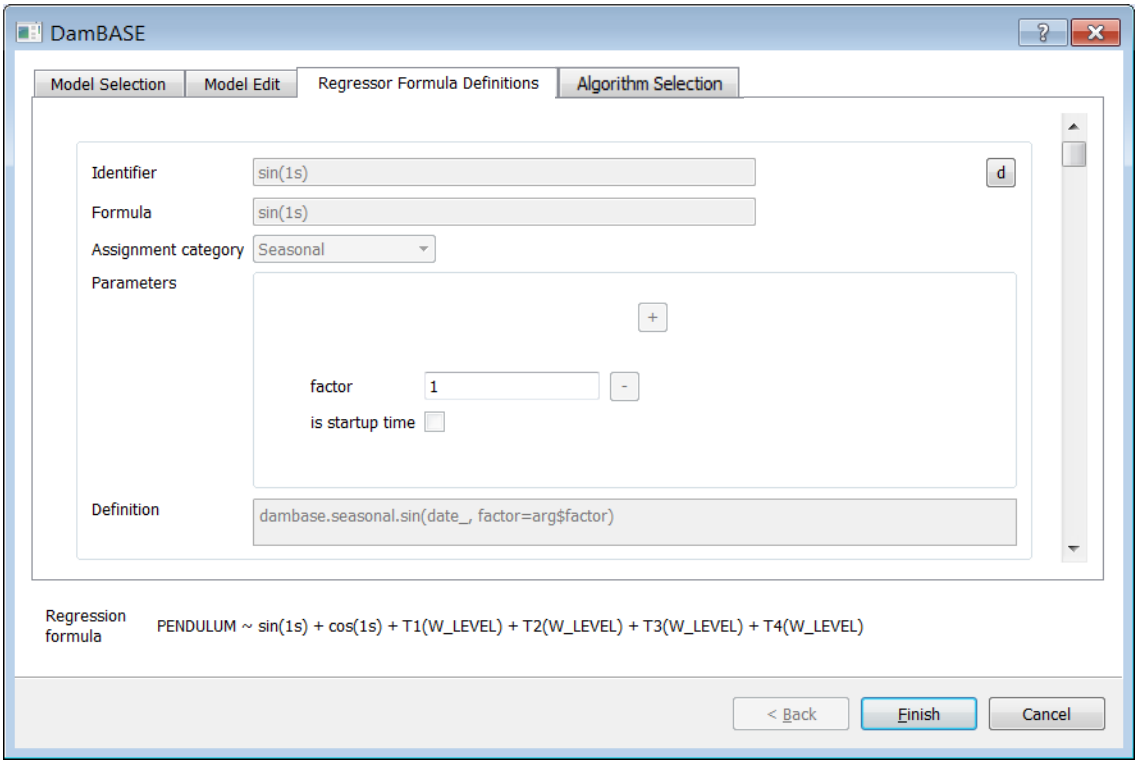


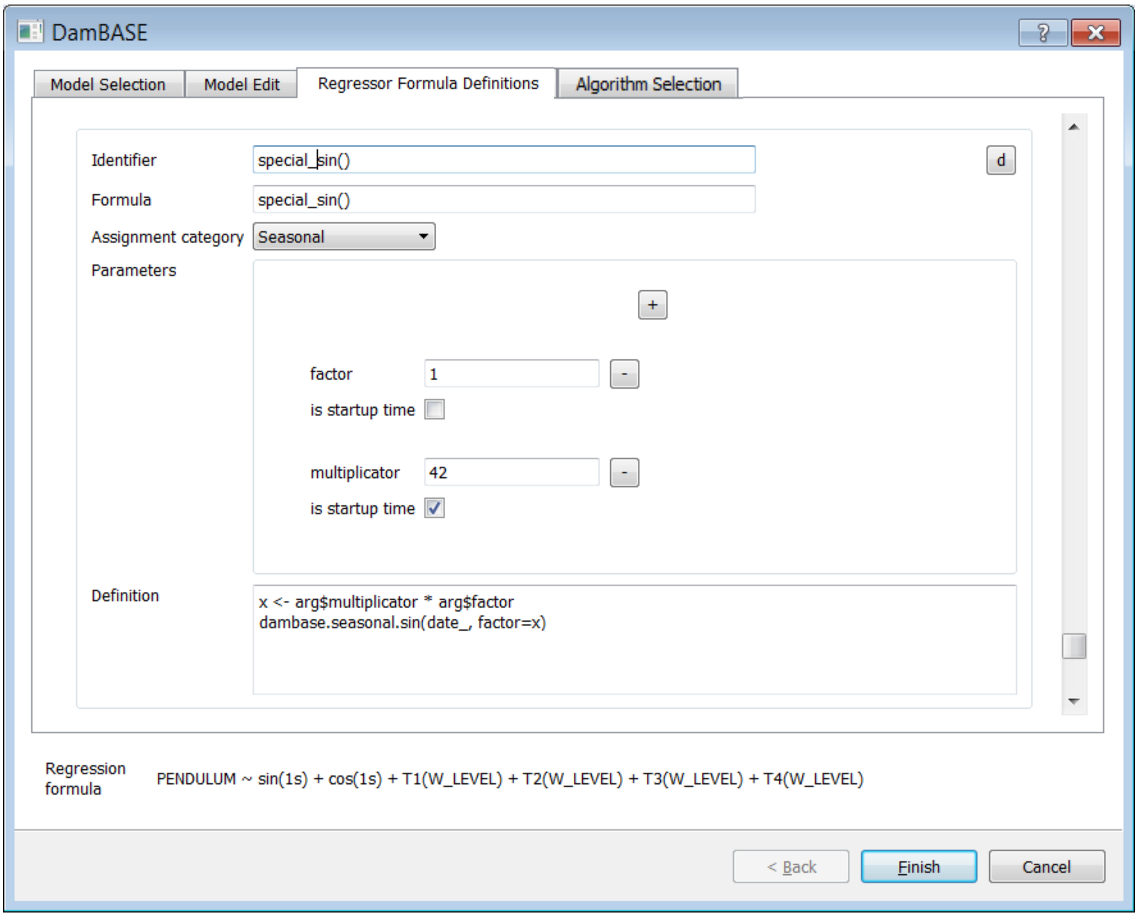
Fig. 4.9 *Use of new parameter in field* Definition

4.7 Creating new regressors

A new regressor can be created by derivation of an existing regressor by click on the “d” button on the *Regressor Formula Definitions* tab of the *Formula editing* wizard page (figure 4.10a). As changing of the assignment category is prevented by the software, it is necessary to choose a regressor of the corresponding assignment category. Note that changing the assignment category after creation of the regressor would cause incomplete regressor definition.



(a) Derivation (d) of existing regressor



(b) New regressor with additional parameters and extended Definition property

Fig. 4.10 *Introduce new regressor by derivation of an existing regressor*

4.8 Algorithm selection

The following algorithms are currently available: Ordinary least squares (OLS), Prais-Winsten estimation (PWE), Ridge regression (RR) and Principal components regression (PCR) which are selectable on the corresponding tab of the model setup wizard page:

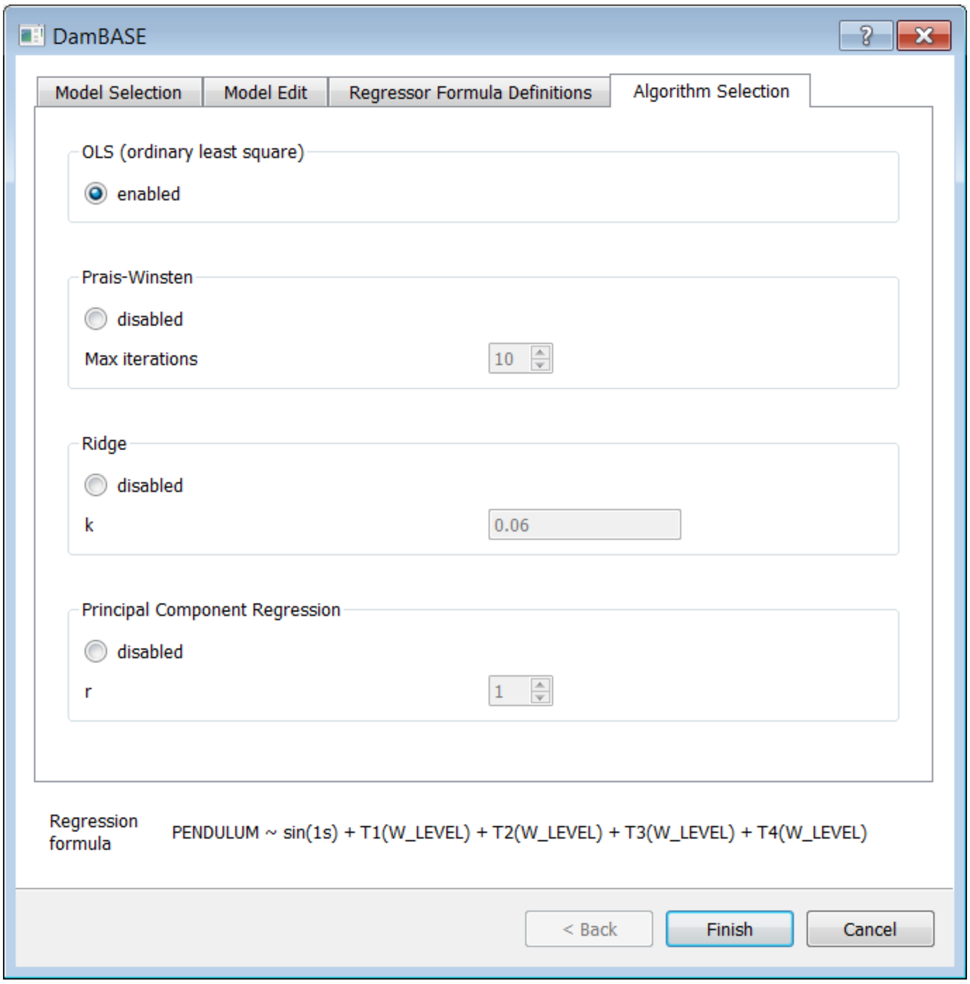


Fig. 4.11 *Algorithm Selection*

4.9 Analysis tools (AT) supporting calibration and validation

In scientific user mode, analysis tools are separated into those which are for analysing least square fit preconditions (tab *second* on the right sided analysis area) and those for calibration and validation (remaining tabs). In basic user mode, only analysis tools related to calibration and validation/prediction are available.

4.9.1 AT *Summary*

modes: basic/scientific

AT Summary provides an overview of the regression fit for both, calibration and validation periods (separated by a gray vertical line) displaying out of limit data points. It is important to note that the 95% interval is an indicator of the goodness-of-fit of the model: the smaller the interval, the better the fit. Therefore, the number of *Out of interval* data points can increase. As a consequence, it is necessary to take *MS Res* from AT *Regression Summary* (see 4.9.3 on the next page) into account.

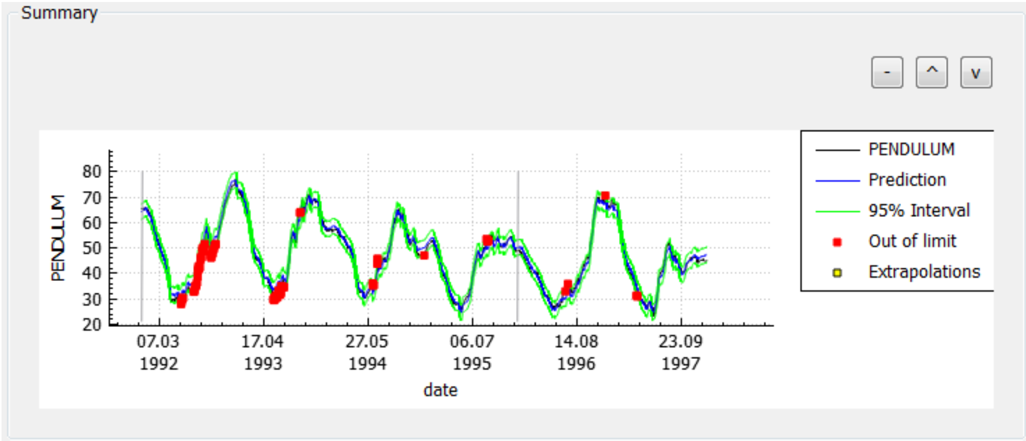


Fig. 4.12 *AT* Summary

4.9.2 AT *Regression overall quality*

modes: basic/scientific

The aim of this analysis tool is to provide some information about regression quality in textual form.

Be careful, because a good fit may be indicated even if there is autocorrelation. Thus, further steps are necessary to remove the autocorrelation.

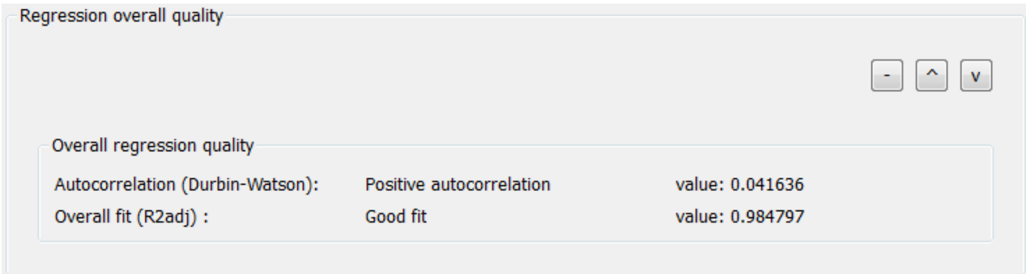


Fig. 4.13 *AT* Regression overall quality

Note that *Overall fit* cannot be used as a single expression about the quality of the model but rather as a hint for how well the model fits on the regression period (see table 4.11 for quality limits).

Tab. 4.11 *Quality limits*

R^2adj	quality expression
1	Perfect fit
$x \geq 0.95$ and x<1	Good fit
$x > 0.5$ and $x < 0.95$	Bad fit
$x > 0$ and $x < 0.5$	Very bad fit

4.9.3 AT *Regression summary*

modes: basic (reduced)/scientific

AT *Regression summary* provides some statistics related information about the created regression model (see section 4.1 on page 30 for more about *MS Res* and *MS Res (prediction)*).

In order to support model optimization, relative changes with regard to the previous model setup are provided within brackets.

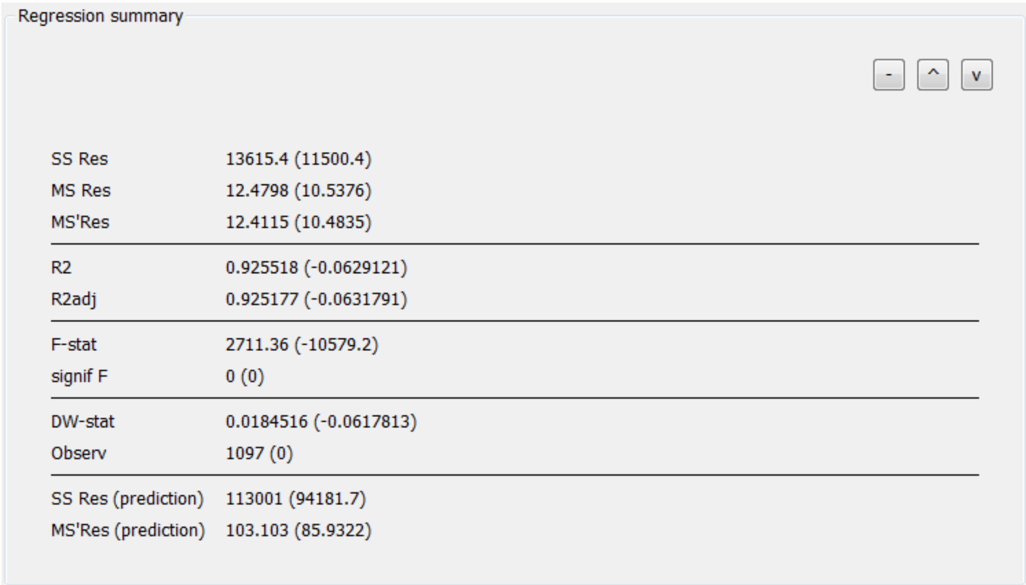
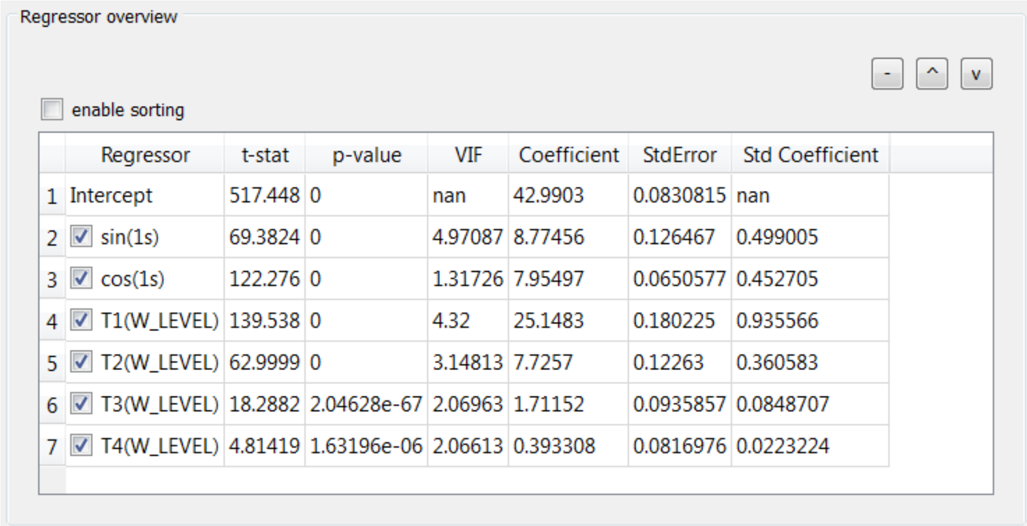


Fig. 4.14 *AT* Regression summary

4.9.4 AT *Regressor overview*

modes: basic (reduced)/scientific

The goal of this analysis tool is to provide not only statistical values of each regressor but also to enable or disable regressors. For the latter, it is important that the model definition is editable (see *Enabling / disabling regressors* on page 34). After enabling sorting on the provided checkbox, the table can be sorted by clicking on the corresponding table header.



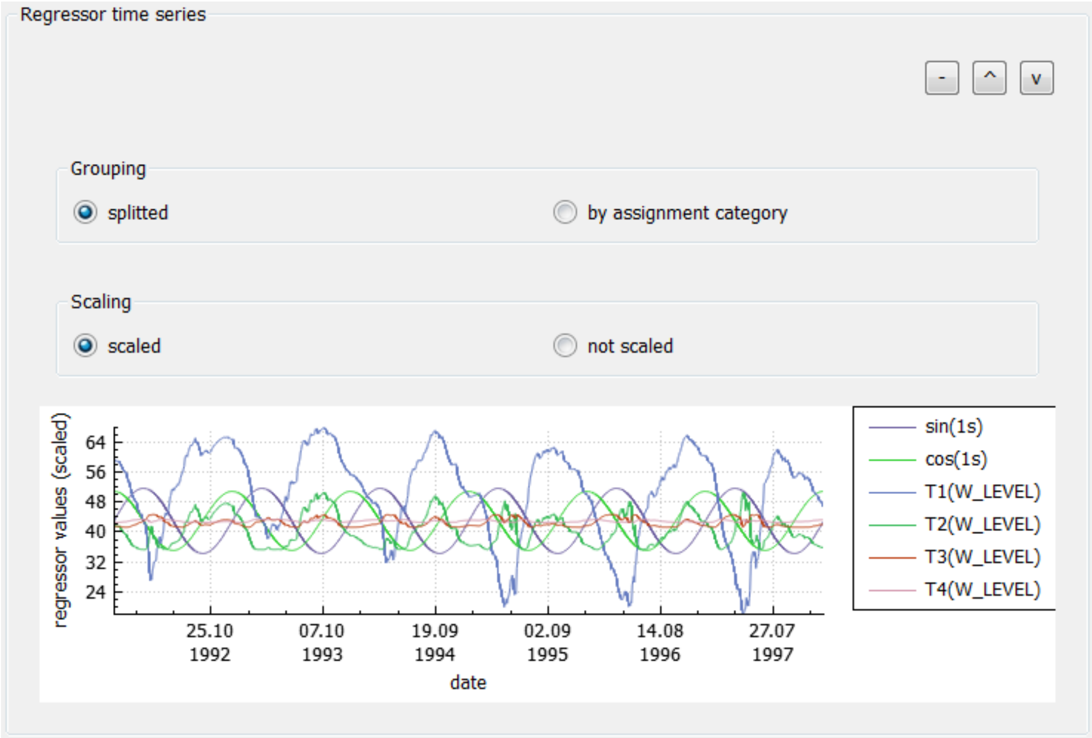
	Regressor	t-stat	p-value	VIF	Coefficient	StdError	Std Coefficient
1	Intercept	517.448	0	nan	42.9903	0.0830815	nan
2	sin(1s)	69.3824	0	4.97087	8.77456	0.126467	0.499005
3	cos(1s)	122.276	0	1.31726	7.95497	0.0650577	0.452705
4	T1(W_LEVEL)	139.538	0	4.32	25.1483	0.180225	0.935566
5	T2(W_LEVEL)	62.9999	0	3.14813	7.7257	0.12263	0.360583
6	T3(W_LEVEL)	18.2882	2.04628e-67	2.06963	1.71152	0.0935857	0.0848707
7	T4(W_LEVEL)	4.81419	1.63196e-06	2.06613	0.393308	0.0816976	0.0223224

Fig. 4.15 *AT* Regressor overview *of a modifiable model definition*

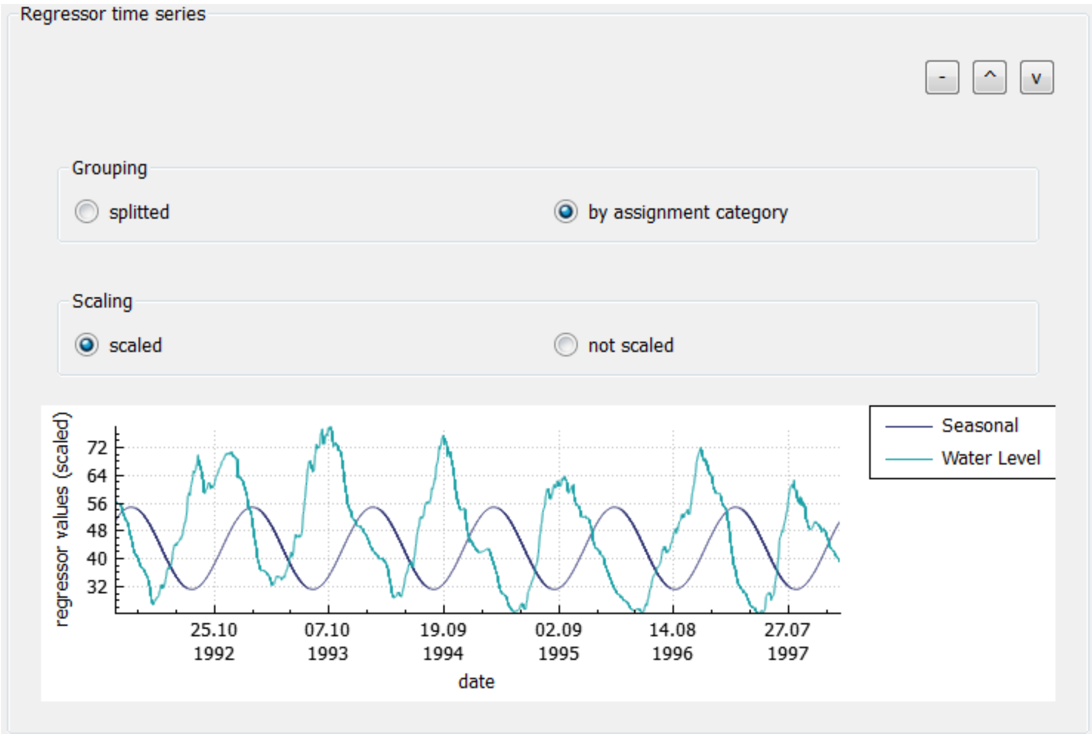
4.9.5 AT *Regressor time series*

modes: basic/scientific

AT Regressor time series displays multiple graphs of either each regressor (figure 4.16a) or parts of the polynomial (regression formula) grouped by assignment category (figure 4.16b). Further, regressor results regardless of whether they are summarized by assignment category or split up into single values can be displayed scaled or not scaled. In the first case, all regressor values are multiplied by their corresponding coefficients.



(a) Split into regressors



(b) By assignment category

Fig. 4.16 *AT* Regressor time series

4.9.6 AT *Physical behaviour*

modes: basic/scientific (no automatic appearance)

The aim of this analysis tool is to provide a way for getting an idea of how the model describes the behaviour of the dam (more precisely the behaviour indicator of the model and therefore the displacement of the dam) from a physical point of view. With this graph the partial displacement and therefore the physical impact of each influence factor grouped by assignment category can be displayed. For example, with this analysis tool it is possible to get the values of the applied water level polynomial opposed to the measured water level displayed on a graph (figure 4.17).

It is important to note that the calculation is only performed on measured values (no extrapolations).

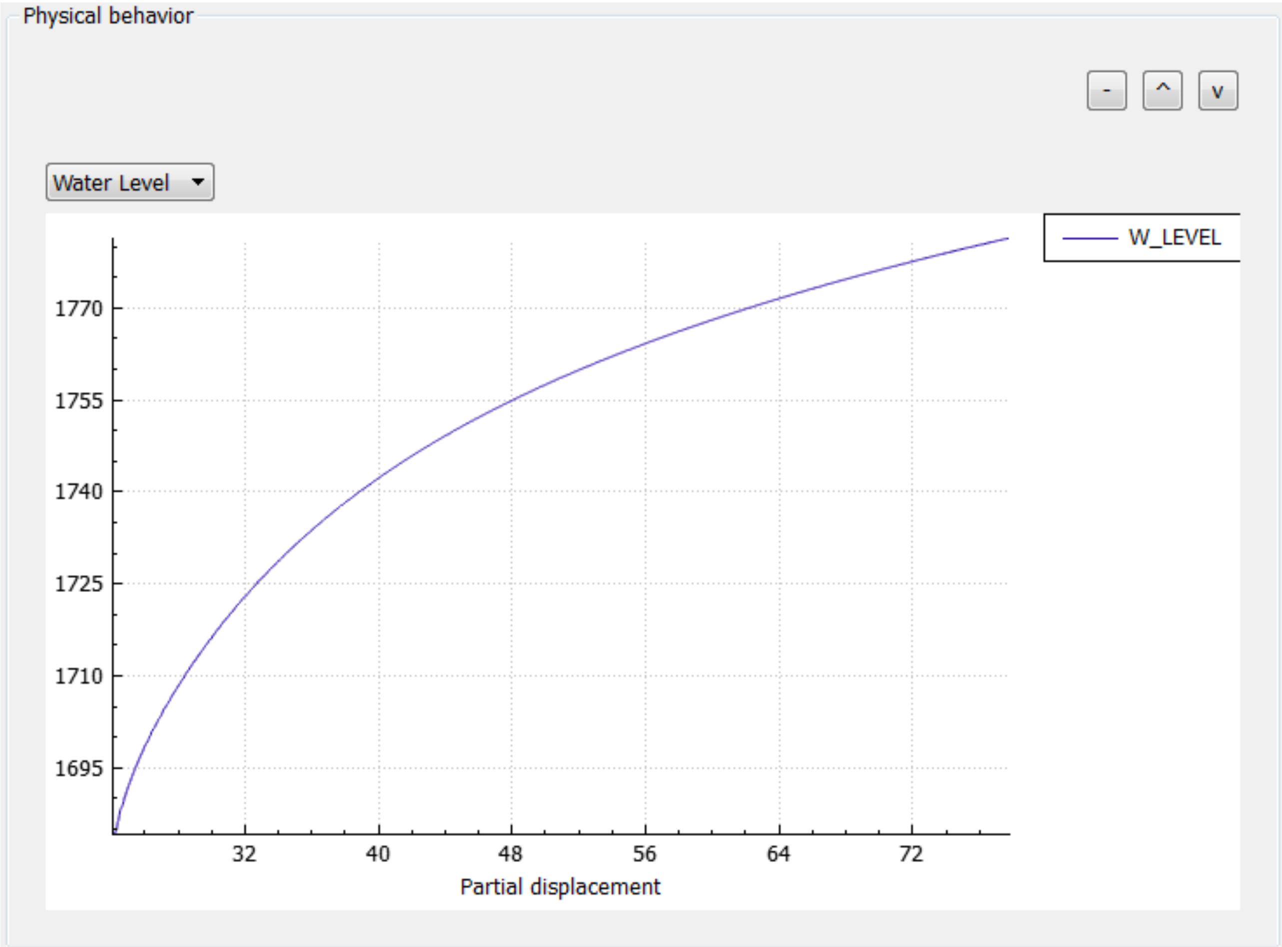


Fig. 4.17 *AT* Physical behaviour

4.9.7 AT *Tukey-Anscombe plot (residual v.s. fitted)*

modes: basic (no automatic appearance)/scientific

This analysis tool displays a plot of fitted values v.s. studentized residuals.

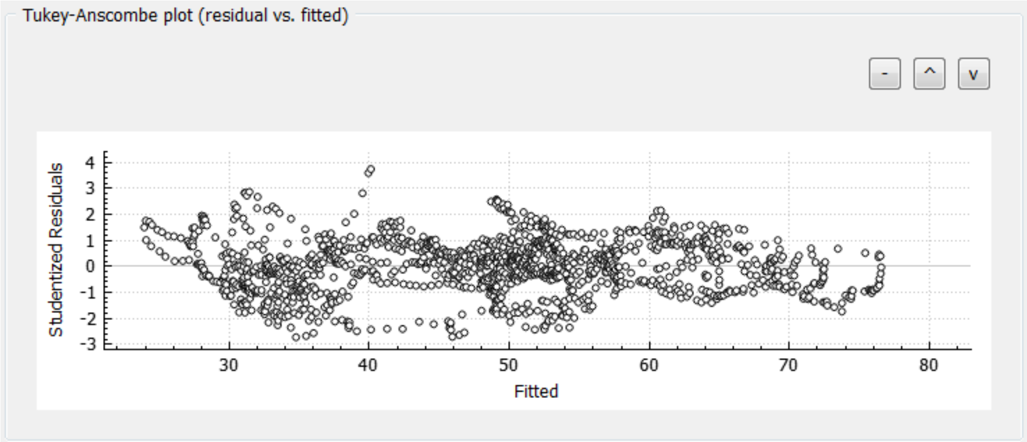


Fig. 4.18 *AT* Tukey-Anscombe plot (residual v.s. fitted)

4.9.8 AT *Residual ACF*

modes: scientific only

This analysis tool displays an ACF (autocorrelation function) plot for auto correlation detection on residuals.

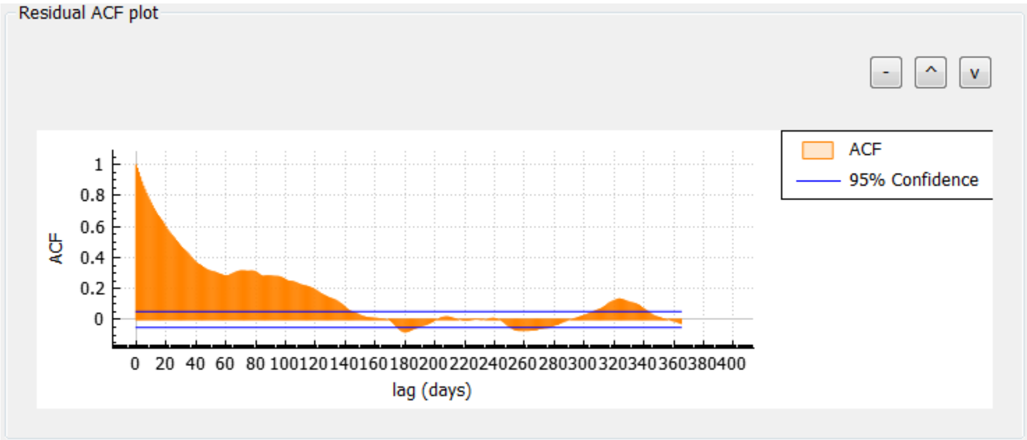


Fig. 4.19 *AT* Residual ACF

4.9.9 AT *Residual normal probability*

modes: scientific only

This analysis tool displays a residual normal probability plot.

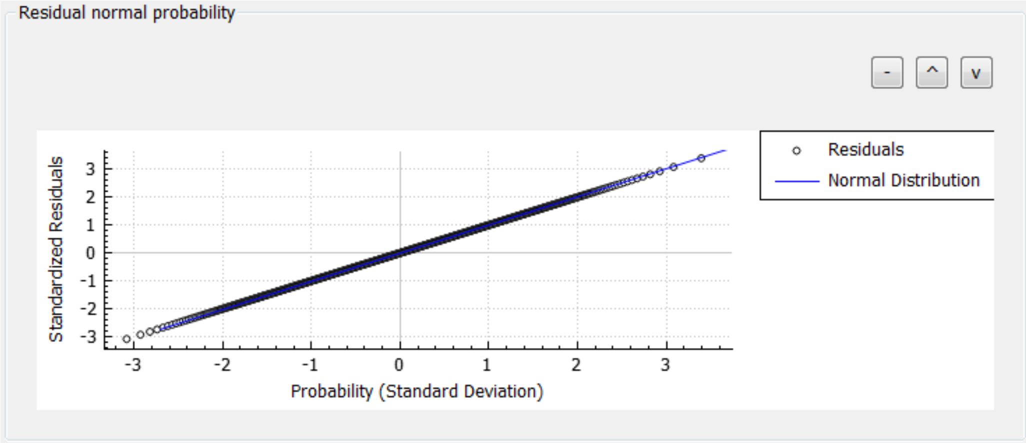


Fig. 4.20 *AT* Residual normal probability

4.9.10 AT *Residual time series plot*

modes: scientific only

The goal of this analysis tool is to display the date related residuals.

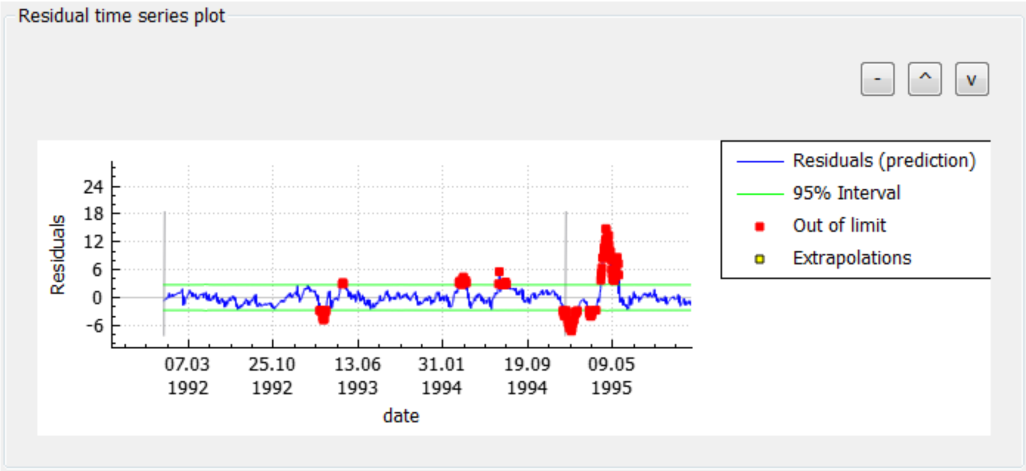


Fig. 4.21 *AT* Residual time series plot

Chapter 5

Regression model setup for prediction

5.1 Introduction

The step *prediction* is the last of the three main work flow steps. The purpose of this step is to use the previously calibrated and validated regression model and to do a prediction based on new measurement data (added to the data of the previous two periods) in order to analyse the dam behaviour (figure 1.2 on page 8).

5.2 Analysis setup and data model selection

As a new data set is required, an analysis object needs to be created with the corresponding data loaded.

When assigning columns on the data model, it is recommended to assign only columns of influence factors that are required by the regression model.

5.3 Regression model setup

An important step here is to choose the correct time range for the regression period covering the periods of the calibration/validation model including the new measurement data. The latter will be used for prediction. To apply the same regression model used for calibration/validation, the identical regression formula has to be set up to make sure that the prediction runs with the calibrated and validated regression model.

5.4 Analysis tools (AT) supporting prediction

Besides all analysis tools available for calibration and validation, the two most important ATs are *AT Regression summary* (page 49) as well as *AT Residual time series plot* (page 54)as they provide information about *MS Res (prediction)*.

However, it is also recommended to check model consistency with other analysis tools especially if deviation appears (i.e. unexpected behaviour) which might be due to model incompleteness or inconsistency, or physically based impacts on the dam.

Part IV

Report generation and export of regression results

Chapter 6

Creating reports

DamBASE supports creating reports in formats pdf, docx, odf, rtf and txt (markdown).

A report can be created out of all loaded analysis tools by right clicking on the selected regression model within the navigation area. All elements of the report can be sorted on the report creation dialog (figure 6.1).

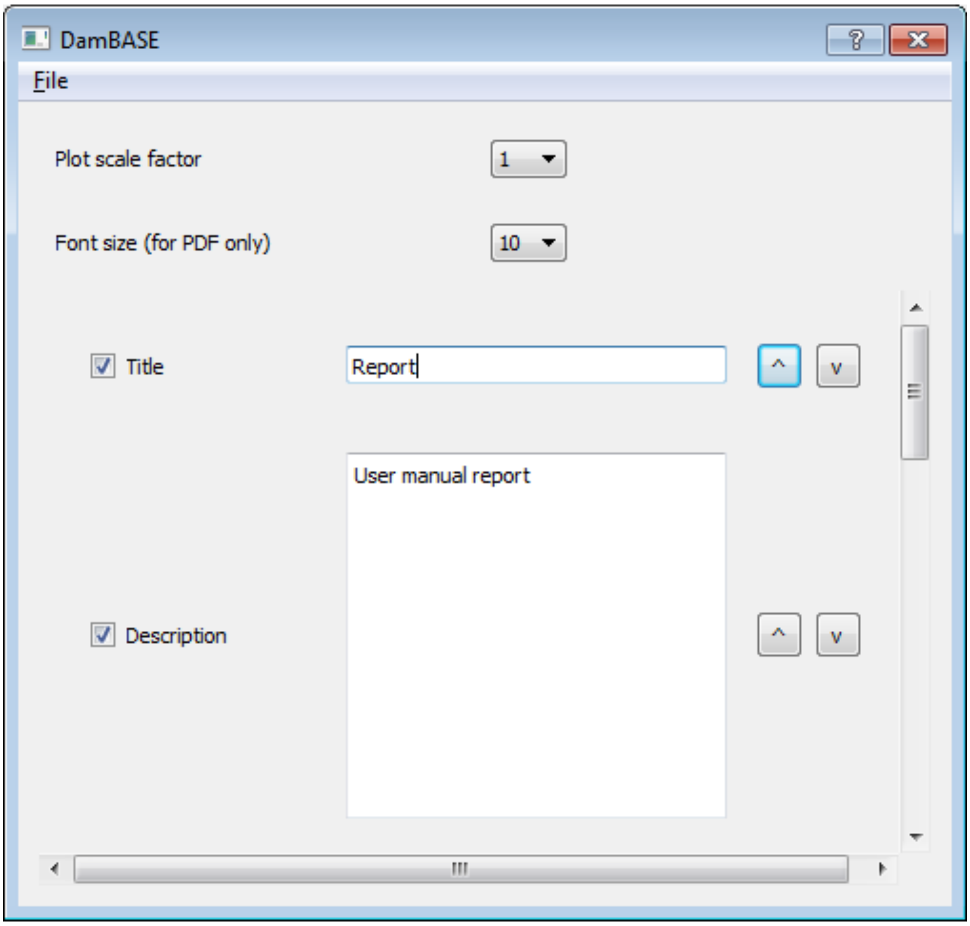


Fig. 6.1 *Report creation dialog*

Chapter 7

Export regression results

For post processing, regression results of all loaded analysis tools can be exported to a tab-separated text file. See figure *Export regression results* on the current page.

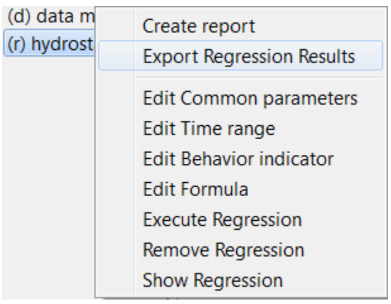


Fig. 7.1 *Export regression results*

Part V

Save / Load project

Chapter 8

Introduction

The ability of DamBASE to save and load projects allows to work continuously on projects by adding new data and regression models. The data format for the project data file is XML that allows for saveing and loading different project snapshots and keeping project management flexible.

As the project file can be of any name, providing enough information on the project object level as well as on all corresponding objects (description field) is key to keep the overview and knowing at any time what and why something was done while modelling.

As the state of analysis tools is not saved, anything related to changes on analysis tools will not result in a pop up dialog to save the project on closing the application. Nevertheless, after loading a project and double clicking on either a data or regression model on the navigation area, all analysis tools are loaded with regard to the current user mode.

Chapter 9

Saving project

A project can be saved by using the menu entry *Project/Save project*.

Chapter 10

Loading project

Projects can only be loaded if no other project is currently opened. This means that if you intend to open a project while working on another, you first have to close the application and load the project after restart by using the menu entry *Project/Load project*.

Bibliography

[ICOLD benchmark, 2001] ICOLD benchmark (2001). ICOLD benchmark workshop on dam safety in Graz.

[Simon *et al.*, 2013] Simon, A., Royer, M., Mauris, F. & Fabre, J. P. (2013). Analysis and interpretation of dam measurements using artificial neural networks. Proc. *9th ICOLD European Club Symposium*, Venezia.

[Swiss Committee on Dams, 2010] Swiss Committee on Dams (2010). *Analysemethoden für die Vorhersage und Kontrolle des Verhaltens von Talsperren*.

[Tatin *et al.*, 2013] Tatin, M., Briffaut, M., Dufour, F., Simon, A., Fabre, J. P. & Rousset, B. (2013). Thermal deformation of concrete dams: justification, clarification and improvement of statistical analysis. Proc. *17th biennial conference of the British Dam Society, Dams: Engineering in a social and environmental context*, A. Pepper (Ed.), Leeds.

[Weber, B., 2002] Weber, B. (2002). Vorhersage des Verhaltens von Talsperren mit Hilfe des Soll Ist Vergleichs Statistischer Teil. Retrieved on 26/05/2015 from `http://www.dambase.ethz.ch/Vorhersage_Verh_Talsperren_Statistik-B_Weber.pdf`.