GroundWater Spatio-Temporal Data Analysis Tool (GWSDAT Version 2.1) User Manual

by

Wayne R. Jones [*] (Wayne.W.Jones@shell.com)
Michael Spence [†] (Mike.Spence@shell.com)

Shell Global Solutions (UK)
Brabazon House, Concord Business Park,
Threapwood Road, Manchester, M22 0RR

---

[*]Statistical Consultant, Statistics and Chemometrics (PTD/TASE).
[†]Environmental Consultant, HSE Technology Soil and Groundwater (PTD/HSGW).

Contents

List of Figures

Acknowledgements

1   Introduction

The GroundWater Spatio-Temporal Data Analysis Tool (GWSDAT), has been developed by Shell Global Solutions to help visualise trends in groundwater monitoring data. It is designed to work with simple time-series data for solute concentration and ground water elevation, but can also plot non-aqueous phase liquid (NAPL) thickness if required. Spatial data is input in the form of well coordinates, and wells can be grouped to separate data from different aquifer units. The software also allows the import of a site basemap in GIS shapefile format. Trend and contour plots generated using GWSDAT can be exported directly to Microsoft PowerPoint and Word to expedite reporting.

GWSDAT uses Microsoft Excel as the primary user interface and data entry platform. The underlying statistical calculations and graphical output are generated using the open source statistical program R [19]. More details on the software architecture and statistical routines can be found in Appendix A.1 and Appendix A.2.

Potential applications where GWSDAT can add value (cost savings and reduction in environmental liabilities) through improved risk-based decision making and response include:

- Early identification of increasing trends or off-site migration.

- Evaluation of groundwater monitoring trends over time and space (i.e., holistic plume evaluation).

- Nonparametric statistical and uncertainty analyses to assess highly variable groundwater data.

- Reduction in the number of sites in long-term monitoring or active remediation through simple, visual demonstrations of groundwater data and trends.

- More efficient evaluation and reporting of groundwater monitoring trends via simple, standardised plots and tables created at the 'click of a mouse'.

Disclaimer: There is no warranty for the Program (GWSDAT), to the extent permitted by applicable law. SHELL, Affiliates of SHELL, the copyright holders and/or any other party provide the Program 'as is' without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the Program is with the LICENSEE. Should the Program prove defective, the LICENSEE assumes the cost of all necessary servicing, repair or correction.

2   Installation Instructions

For up-to-date advice on how to install GWSDAT, additional training materials and software updates please refer to the GWSDAT Livelink site. To access this site please contact your Shell COE Technical Assurance Regional Focal Point.

A minimum screen resolution of 1024 x 768 or better is required for correct display of the GWSDAT graphical user interface. Lower screen resolutions may result in only part of the user interface being visible.

3   Data Input

Groundwater monitoring data is entered into GWSDAT by means of a standardised Excel input sheet (Figure 1). GWSDAT will use the name of the input data sheet as the name of the analysis, so please change accordingly. The template has been designed with simplicity in mind. There are only two input tables that must be completed, namely, the *Historical Monitoring Data* table and the *Well Coordinates* table. The third *GIS Shapefiles* table may be populated with links to the location of GIS shapefiles for use as basemaps or site plans, if required. Please ensure that there are no empty rows in these completed data input tables. Two example data sets are provided with the software for training purposes, the use of which is explained in section 4.

Figure 1: *GWSDAT data input example.*

3.1 Historical Monitoring Data Input Table

Each row of this table corresponds to a unique combination of well, sampling date, and solute type. Groundwater and NAPL (Non-Aqueous Phase Liquids) gauging data may also be entered in this table. Figure 1 displays an example GWSDAT input data set for illustrative purposes. The columns (fields) in the *Historical Monitoring Data* input table are as follows:

- **WellName:** The name or identifier of the well (or soil boring) being sampled. Well names must be consistent and unique. For example: 'MW-1' and 'MW1' would be treated as different wells.

- **Constituent:** Here enter the name of the solute type, e.g. Benzene, Toluene. Again in the same manner as WellName please ensure that the name of a solute is consistent and unique for all entries. The iden-

tifiers 'GW' and 'NAPL' are reserved for Groundwater elevation measurements and NAPL thickness data respectively, see further details below.

- **SampleDate:** The date at which the well was sampled (not the date the results were returned from laboratory analysis). Please use a calendar date format, the preferred format is 'dd/mm/yyyy' and do not include time data.

- **Result:** According to the constituent type the result may be concentration, groundwater elevation or NAPL thickness data:

  Concentration data: The concentration of the constituent is entered here. Non-detect values, should be entered as either '<X' or 'ND<X' where 'X' is the detection limit as specified by the laboratory. For example, if the detection limit is 100ug/l then either '<100' or 'ND<100' is acceptable. The non detect threshold value must be specified, i.e. 'ND' on its own is not permissible. In the absence of known detection limits the user must substitute a sensible value, e.g. the lowest detected value for the solute in the data set.

  Groundwater level data is entered here as an elevation above a common datum, e.g. metres or feet above sea level or some other common reference height. Please ensure that all groundwater measurement entries have the same units (e.g. metres or feet) and that the Constituent field is set to 'GW'. In the presence of NAPL, please ensure that the groundwater elevation has been corrected for NAPL density. See Figure 1 for an example of how to enter groundwater height data.

  NAPL thickness data is also entered here. Please ensure that all NAPL thickness entries have the same units, e.g. feet or metres and that the Constituent field is set to 'NAPL'. If no NAPL is present, do not add a NAPL entry with zero thickness; simply omit from the table.

  Where NAPL is recorded in soil borings that do not reach the water table the NAPL thickness should be entered as zero. Well location markers for soil borings or wells where NAPL has been recorded are highlighted in red.

- **Units:** Here select from the drop-down listbox (see Figure 1) the corresponding units. Solute concentration data can either be 'mg/l' or

'ug/l'. For groundwater elevation and NAPL thickness data please set to one of 'mm', 'cm', 'metres', 'inches', 'feet' or 'level'. Units must be specified for each entry. All entered groundwater elevation measurements must have the same units. Likewise for NAPL thickness.

- **Flags:** Four flags are available in GWSDAT v2.1 that can be used to modify the way in which certain types of data are handled by the software. The 'E-Acc' (Electron Acceptor), 'NotInNAPL' and 'Redox' flags are used to identify input data types which are to be omitted in the event that the user activates the NAPL substitution function (see section 5.5). Note, that it is only necessary to flag one data row in this way for all rows containing that constituent to be excluded from NAPL substitution (see Figure 1). The fourth flag ('Omit') can be used to exclude individual data entry rows from the GWSDAT analysis.

3.2   Well Coordinates Table

The *Well Coordinates Table* is used to store the coordinates of groundwater monitoring wells or soil borings. For most of the purposes of GWSDAT modelling, it is only the relative distances between wells which are important. This means any arbitrary cartesian coordinate system can be used as long as well coordinate values have an aspect ratio very close to 1, i.e. a unit in the x-coordinate is the same distance as a unit in the y-coordinate. Hence, well coordinates can be measured directly from a map, or given in easting and northing, etc.

Note: GWSDAT cannot plot monitoring data in the absence of well coordinates.

- **WellName:** Name or identifier of the well or soil boring. Well names must be identical to those specified in the *Historical Monitoring Data* table.
  Hint: It is better to name wells using the convention of 'MW-01' rather than 'MW1' so that plots in GWSDAT are correctly ordered.

- **XCoord:** The x-coordinate value for the corresponding well.

- **YCoord:** The y-coordinate value for the corresponding well.

- **Aquifer:** The aquifer field allows the user to associate wells or soil borings with particular subsurface features (e.g. aquifers, sub-strata), in the event that data from these needs to be modelled separately. The user can enter the name (maximum of 8 characters) of the aquifer or sub-stratum, or select a letter A-G from the drop-down listbox. The aquifer field can also be used to partition the dataset from a large site, in the event that mutiple unrelated plumes are present or if wells are clustered with large gaps in between. On initiation of a GWSDAT analysis the user is asked to select an aquifer (subsurface feature) to analyse. **Note:** Plots generated using data associated with particular subsurface features have the feature name appended to the title, e.g. Shallow aquifer. If the user leaves the aquifer flag as blank, no such appending will occur.

- **CoordUnits:** Select from the drop-down listbox the length scale of the well coordinates. Either leave this field blank or select 'metres' or 'feet'. The units specified in this field are used in the calculation of plume mass balance parameters (e.g. plume area and solute mass), for further details see Section 7.7 on plume diagnostics.

3.3   GIS ShapeFiles Table

A site plan can be superposed over plots of concentration distribution, NAPL thickness and groundwater elevation (see Figure 5). Site plans are imported into GWSDAT in the form of shapefiles (see http://en.wikipedia.org/wiki/Shapefile for more information). A 'shapefile' is actually a collection of several files, typically created using ARC-GIS.

- **Filenames (*.shp):** The 'GIS Shapefiles Table' is used to store the filepath to site plans, which need to be in shapefile format. The user can either enter the file location manually or use the 'Browse for Shapefile' function in the GWSDAT Excel menu (see Figure 2) for interactive file selection. Only the location of the main shapefile (file ending with a '.shp' extension) needs to be specified in this table - the associated data files (e.g. .dbf, .sbn, .sbx, .shx) will be picked up automatically, **provided they are in the same folder** (see example in Figure 1). It is possible to overlay multiple shapefiles up to a maximum of seven.

4   GWSDAT Main Menu

In MS Excel 2007 or later, the GWSDAT menu is located under the 'Add-Ins'
tab.
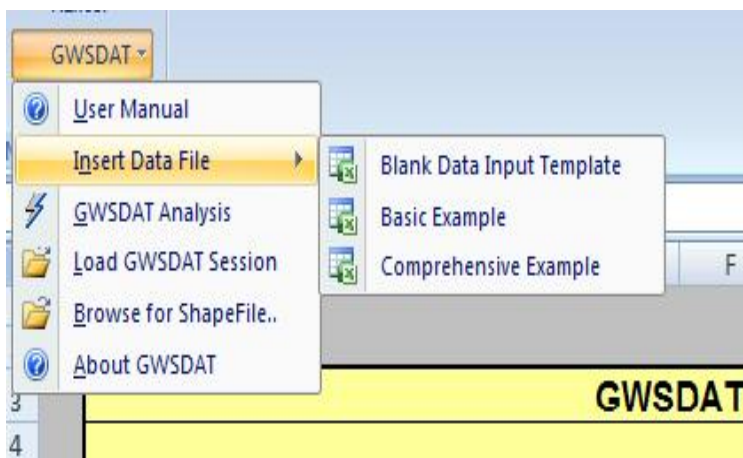
Figure 2: *GWSDAT Main Menu.*

The main menu options, as shown in Figure 2 , are as follows:

- **User Manual:** Displays this (pdf) GWSDAT user manual.

- **Insert Data File:**

  - **Blank Data Input Template:** Inserts a blank GWSDAT Data
    input template worksheet into the active Excel workbook.

  - **Basic Example:** Inserts an example GWSDAT worksheet data
    set into the active Excel workbook.

  - **Comprehensive Example:** Inserts a more detailed example
    GWSDAT worksheet data set which includes a site plan, NAPL
    thickness data, 'Electron Acceptor' flagged solutes and multiple
    aquifers into the active Excel workbook.

- **GWSDAT Analysis:** Begin GWSDAT analysis on the Excel active
  worksheet data.

- **Load GWSDAT Session:** Loads a previously saved GWSDAT session, see Section 7.8 for more information.

- **Browse for Shapefile..:** Interactively browse for a shapefile and add location to GIS Shapefiles table, see Section 3.3 for more information.

- **About GWSDAT:** Displays version information and Terms & Conditions for GWSDAT.

The two example data files are provided for training and demonstration purposes. Select the basic or comprehensive dataset and then 'GWSDAT Analysis' from the main menu. Accept the default data processing parameters (Section 5) and maximise the GWSDAT panel to practise using the data visualization tools (Section 7)

---

5   Data Processing Options

Figure 3: *GWSDAT options form.*

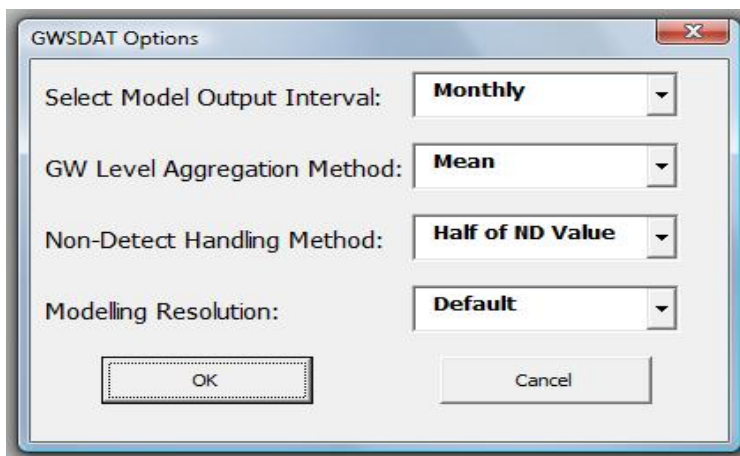On initiation of a GWSDAT analysis, data processing options are displayed, as shown in Figure 3. The data processing options influence how the data is displayed and how non-detects are handled.

5.1 Model Output Interval

The spatiotemporal model can generate predictions at a user specified interval. The three different options are as follows:

- **All Dates:** Concentration and groundwater elevation contour plots are generated for every date represented in the input dataset. This is a good option to choose if each monitoring event comprises samples/ measurements collected within one 24-hour period.

- **Monthly:** Concentration and groundwater elevation contours are generated at monthly intervals, working backwards in time from the latest date in the input dataset. Choosing this option aggregates groundwater elevation data within each monthly interval so that a larger dataset is available for the plotting of elevation contours (by local linear regression).

- **Quarterly:** Concentration and groundwater elevation contours are generated at quarterly (3 month) intervals, working backwards in time from the latest date in the input dataset. Choosing this option aggregates groundwater elevation data within each 3-month interval so that a larger dataset is available for the plotting of elevation contours (by local linear regression).

Note that both the monthly and quarterly model output options only aggregate the dataset used to plot groundwater elevation contours. The solute concentration dataset is **not** aggregated in time because the spatiotemporal model from which concentration contours are generated does not require this, i.e. the underlying spatiotemporal model used to generate the solute concentration smoother plots does not vary with the data aggregation interval.

Note that if no monitoring data is present within a particular monthly or quarterly interval, then GWSDAT will not generate a groundwater elevation contour or spatio-temporal solute concentration smoother plot. This is to avoid producing potentially misleading spatial plots far away in time from any actual data.

5.2   GW Level Aggregation Method:

In the event that there are multiple groundwater elevation measurements from the same well within a given output interval, the user can select how to use this data. The user can select to calculate either the 'Mean', 'Median', 'Min', or 'Max' groundwater elevation. Again, this choice does not affect the spatiotemporal model used to generate the solute concentration smoother plots.

5.3   Non-Detect Handling

Non-Detect Handling Method: GWSDAT handles non-detect data by a method of substitution. In accordance with general convention, the default option is to substitute the non-detect data with half its detection limit, e.g. ND<50ug/l is substituted with 25ug/l. For a more conservative choice, select the alternative of non-detect data to be substituted with its full detection limit, e.g. ND<50ug/l is substituted with 50ug/l.

5.4   Spatiotemporal Modelling Resolution

Modelling Resolution: This option controls the resolution of the spatiotemporal solute concentration smoother (see Appendix A.2.1). The user can select between either a default resolution or a higher resolution model fit. In most instances there will be little difference in the modelling results between the two settings. However, in some rare circumstances with complex data sets, it may well be necessary to use the higher resolution setting. Please note it takes approximately 3-4 times longer to fit a higher resolution model.

5.5   NAPL Handling Method

An additional pop-up box will be displayed after the GWSDAT options box if the input contains NAPL data (i.e. 'NAPL' is entered in the constituent field). Selecting 'Yes' to the question 'Do you wish to substitute NAPL values with maximum observed solute concentrations?' forces GWSDAT to recognise NAPL data in the input dataset as indicative of high dissolved

solute concentrations. This option has been added to provide the user with a more realistic picture of the area of impacted groundwater in the event that NAPL in wells prevents direct measurement of CoC concentrations. *Before using this function the user should, however, be confident that dissolved CoCs are derived from the observed NAPL and not from a different source.* Solutes flagged as 'Electron Acceptors' (see Section 3.1) are omitted from the NAPL substitution process.

6 Data Validation Procedures

In the majority of cases, the default data processing options as displayed in Figure 3 will be acceptable. To continue the GWSDAT analysis, press the 'OK' button. This initialises a series of data validation procedures. GWS-DAT will report back by means of a messagebox on any data anomalies. These include:

- Incorrect Date formatting.

- Unspecified detection limits.

- Missing or non-unique well coordinates.

- Data input case sensitivity issues.

- Solute concentrations entered as zero (detection limit must be specified).

> An excellent way to spot errors in GWSDAT input data is to use the Excel 'Auto-Filter' on the input tables and inspect the unique filtered values in the drop down listbox.

Following successful data input, GWSDAT will begin the model fitting process, see Appendix A.2. Model fitting progress is indicated to the user by a commented progress bar indicator.
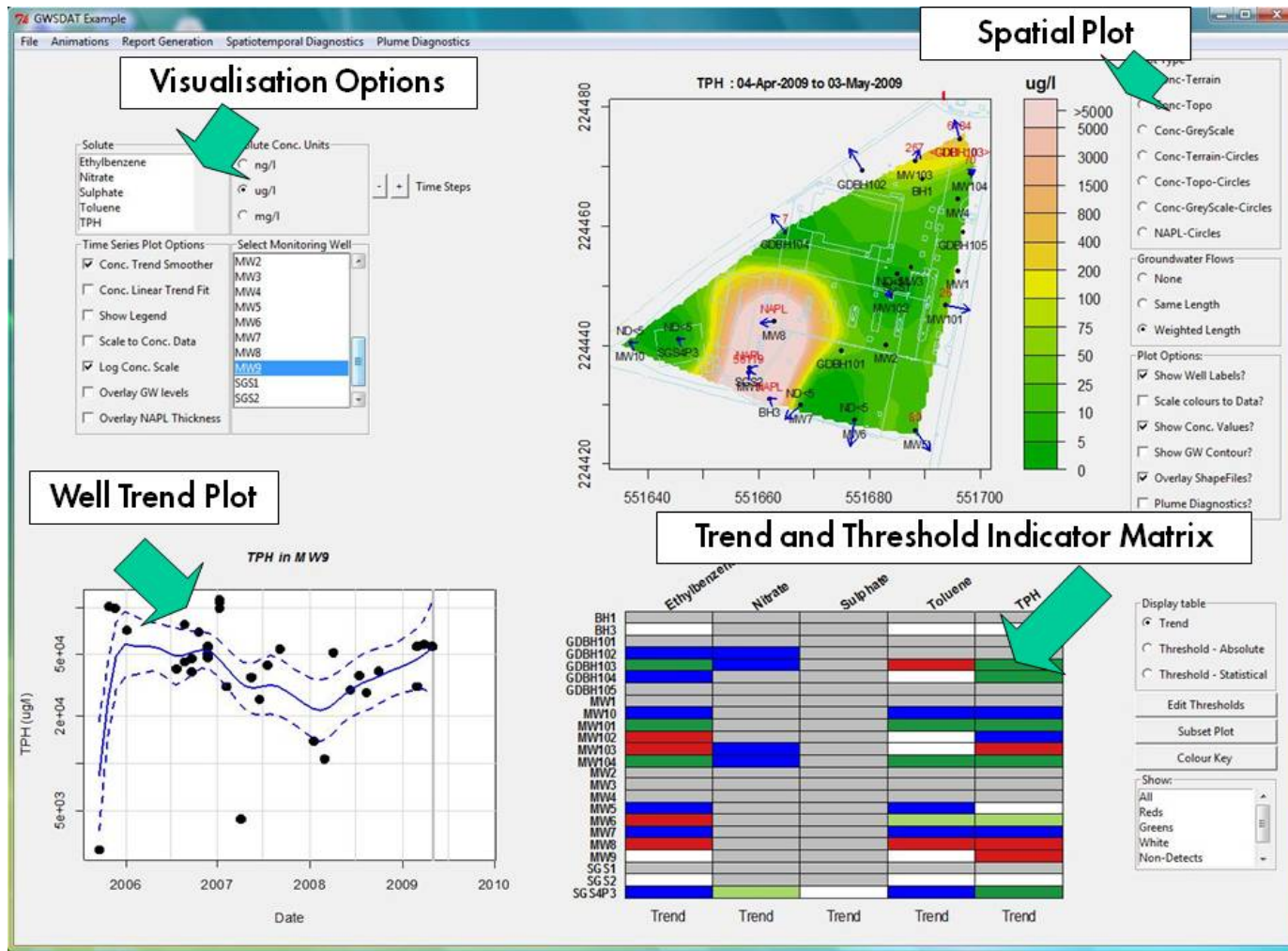
Figure 4: *GWSDAT user interface.*

7 GWSDAT User Interface

Following successful model fitting the GWSDAT user interface will be displayed, see Figure 4. Note: Occasionally, the GWSDAT user interface appears hidden underneath any windows that are currently open. The user can locate GWSDAT from the windows taskbar at the bottom of the screen.

The GWSDAT user interface is a stand-alone point and click graphical user interface (GUI) which the user can interact with in many ways. The following sections will explain the interface in more detail.

7.1 GWSDAT Spatial Plot
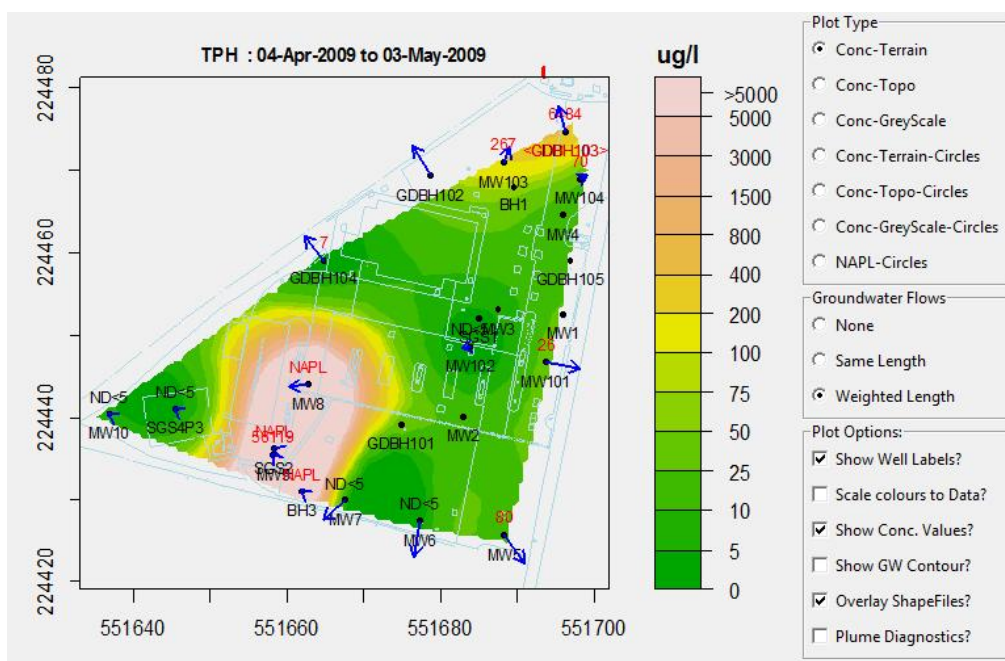
Figure 5: *GWSDAT Spatial Plot.*

The GWSDAT spatial plot (see Figure 5) is for the analysis of spatial trends in solute concentrations, groundwater flow and, if present, NAPL thickness. It displays the locations of the monitoring wells (black solid dots) together
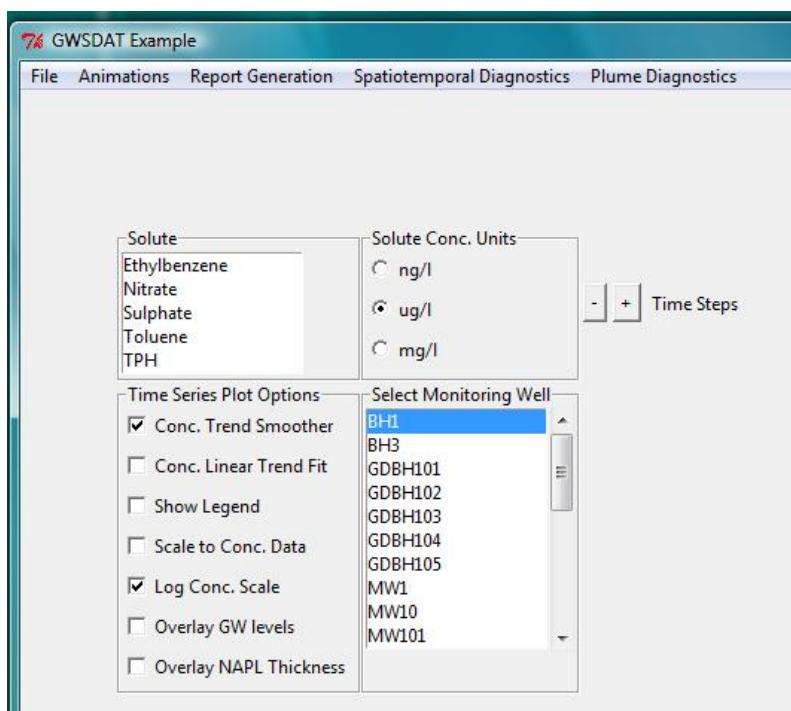
Figure 6: *GWSDAT Visualisation Options.*

with the well names and actual measured solute concentration values (detect data is displayed in a red font; non-detect in a black font). The date interval for the displayed data is indicated above the spatial plot. In this example, a monthly model output interval has been selected and the displayed actual solute concentration values were sampled between the 4th April 2009 and the 3rd May 2009. If a GIS shapefile has been supplied then the major site features (roads, tanks, etc) are overlaid on the spatial plot as light blue lines.

To visualise the spatial plot for a different solute simply select from the *Solute* listbox from the *Visualisation Options* portion of GWSDAT user interface, see Figure 6. The *Solute Conc. Units* radiogroup controls whether to display solute concentrations in 'ug/l' or 'mg/l'.

To increment forward and backwards between time steps, press the '+' and '-' *Time Steps* buttons again located in the *Visualisation Options* portion of GWSDAT user interface. As the user increments through different time

steps the small red horizontal bar at the top of the plot (see Figure 5) changes position to indicate where the time step lies in the monitoring period. The far left position indicates the beginning of the monitoring period and the far right indicates the end of the monitoring period. By holding down the '+' button, the plot will update continuously to generate an animated series of images. If the final time step is reached it will restart at the initial time step. For more methods to generate animations see Section 7.4.

By left-clicking the plot, an identical but expanded plot is generated in a separate window. This is particularly useful for plots which contain a lot of information. This window can be resized and, importantly, the aspect ratio is maintained. The user can save this plot in a variety of different graphical formats by selecting *'File'-> 'Save as'* from the menu located at the top left of the window. Alternatively, the user can simply right click on the plot and choose to copy the image to the clipboard as either a *metafile* or *bitmap* and paste directly into another application. Note that once the image has been saved to a graphical format, the ability to maintain an aspect ratio of one is lost.

The spatiotemporal solute concentration smoother is a function which simultaneously estimates both the spatial and time series trend in site solute concentrations. By smoothing the data in both space and time it provides a clearer interpretation of site solute concentration dynamics than would otherwise be gleaned from the raw data. However, it is important to note that it is a smoother function and as such, the predictions do not necessarily lie on the observed data points. In the event that a sampled concentration value is significantly larger than the predictions of the spatiotemporal smoother, the well label is coloured red and surrounded by braces, e.g. '<MW-1>'. This serves as a very useful method for outlier detection. In addition, the analysis may be skewed if data are input from monitoring wells with disparate construction or screened in different aquifer systems.

Another important point to consider is that the quality of the spatiotemporal smoother is directly influenced by quality of the underlying data. In general, data originating from sites with many evenly spatially distributed wells with a long time history leads to better quality smoother predictions. The converse of a small number of wells or poor well location network design (e.g. wells located in almost a straight line), or short monitoring history, will lead to less reliable smoother predictions, particularly at the edges.

In summary, the 'spatiotemporal solute concentration smoother' plot is provided to help the user visualise the distribution of solutes and as an aid to risk-based decision-making. However, for the reasons stated above, the predictions should be interpreted with care and a more detailed evaluation may be necessary to understand observed trends and outliers.

Further methods for assessing the goodness of fit of the spatiotemporal smoother can be found in Section 7.6. For more details on the spatiotemporal smoothing algorithm, please see Appendix A.2.1.

The following sections describe the various spatial plotting options (see Figure 5).

7.1.1 Plot Type:

- **Conc-Terrain:**

  This option overlays the predictions of the spatiotemporal solute concentration smoother for a particular model output interval using a 'terrain' colour scheme. The above example (Figure 5) displays the spatio-temporal solute concentration smoother for TPH (Total Petrol Hydrocarbons) on the 3rd May 2009. Please note that the output of the spatio-temporal trend smoother is always given for the latest date in the displayed output interval. The dark green colours indicate low solute concentration and the colours are gradated through yellow and brown to almost white, to illustrate increasing estimated solute concentrations. The concentration values can be read off from the key on the right hand side of the plot.

  As the user iterates through time steps, it may be noticed that the area covered by the spatiotemporal solute concentration smoother changes. This is due to the fact that spatiotemporal predictions are only generated between interpolated data and are not extrapolated to regions where no data exists, which could potentially lead to erroneous results. For each time step, the area of the contour is calculated only from the collection of wells for which the monitoring period spans the current model output interval. GWSDAT generates predictions in the convex hull region dictated by these wells. The convex hull (see http://en.wikipedia.org/wiki/Convex_hull) may be visualised as the expected boundary if an elastic band was placed around the locations of these wells.

- **Conc-Topo:** This function is identical to Conc-Terrain but uses a topograghic colour scheme which gradates increasing solute concentrations through blue, green, yellow and beige.

- **Conc-Terrain-Circles:** This selection overlays (terrain) colour coded circles located at the wells which have been monitored within the current model output interval. The size of the circles scales with the log of the observed solute concentration values and the solute concentration range can be read off from the colour key to the right of the plot.

- **Conc-Topo-Circles:** This selection is identical to Conc-Terrain-Circles but uses a topographic colour scheme.

> **Hint:** In the presence of poor well location network design or limited data then it is recommended the user select either the 'Terrain-Circles' or 'Topo-Circles' plot type.

- **NAPL-Circles:** This selection displays the observed NAPL thicknesses within the current model output interval as size scaled and colour coded circles. NAPL thickness ranges are read off from the colour key on the right hand side of the plot. Colours are gradated from dark red through yellow to almost pure white to illustrate increasing NAPL thickness. The location of wells which have recorded NAPL in any part of their monitoring history are coloured with red solid dots instead of the usual black solid dots.

7.1.2   Groundwater Flows:

The blue arrows in Figure 5 display the estimated direction and (relative) hydraulic gradient of groundwater flow at monitoring points across the a site. This is calculated from the combination of well coordinates and recorded groundwater elevations for this particular model output interval (see Appendix A.2.3 for more details). The *Groundwater Flows* radiogroup positioned to the right hand side of the plot allows the user to choose either 'not to display groundwater arrows' or 'direction only arrows' or 'both direction and relative strength arrows'.

7.1.3   Plot Options

- **Show Well Labels?:** This controls whether to display well names/labels immediately below the well locations.

- **Scale colours to Data:** By default the colour key of solute concentrations is subdivided as shown in Figure 5. By using the same subdivisions the spatiotemporal solute concentration smoother plots can be directly compared between different model output intervals. This

control will produce a new colour key whose subdivisions span the concentration predictions for the current model output interval only.

- **Show Conc. Values?:** This controls whether to display actual sampled concentration values immediately above the well locations. If the data is identified as a NAPL measurement the value will be displayed as 'NAPL' in a red font.

- **Show GW Contour?:** To add contour lines of groundwater level data, check the *GW Contour?* checkbox located on the right hand side of the plot. This superposes isobars of smoothed groundwater elevation data on top of the solute concentration plot.

- **Overlay Shapefiles?:** This controls whether to overlay a site plan.
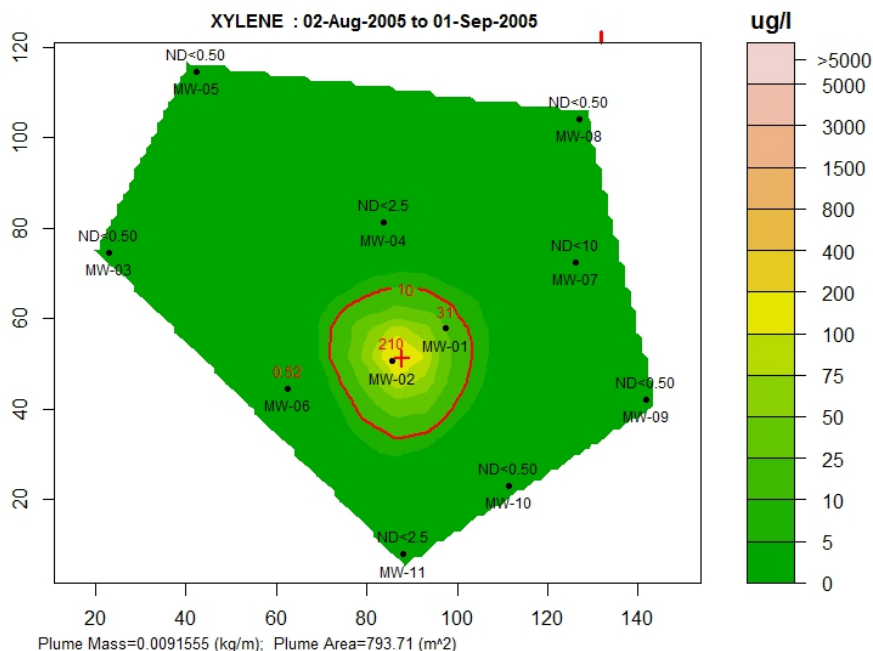
Figure 7: *Example spatial plot with plume diagnostics.*

- **Plume Diagnostics?:** This controls whether to calculate and display plume diagnostic quantities from the predictions of the spatiotemporal solute concentration smoother (see Figure 7). The delineated plume

is displayed with a solid red contour line which also includes a label displaying the plume boundary threshold value. The plume centre of mass is displayed with a red cross and the plume mass and area printed at the bottom left margin of the spatial plot. Note: in order for the correct plume diagnostics units to be used the **CoordUnits** field in the *Well Coordinates* table must be specified, see Section 3.2. More details about plume diagnostics can be found in Section 7.7 and Appendix A.2.2.

7.2 Well Trend Plot

Figure 8: *GWSDAT Well Trend Plot.*

The well trend plot enables the user to investigate time series historical trends

of solute concentrations in individual wells. Figure 8 displays an example GWSDAT Well Trend plot of 'Benzene' in well 'MW-01' using an illustrative example dataset. The actual sampled concentration values are plotted against sampling date and are represented as black solid points. Orange points represent the substituted non-detect values according to the selection chosen in Section 5. Red points represent the NAPL substitued solute concentration values.

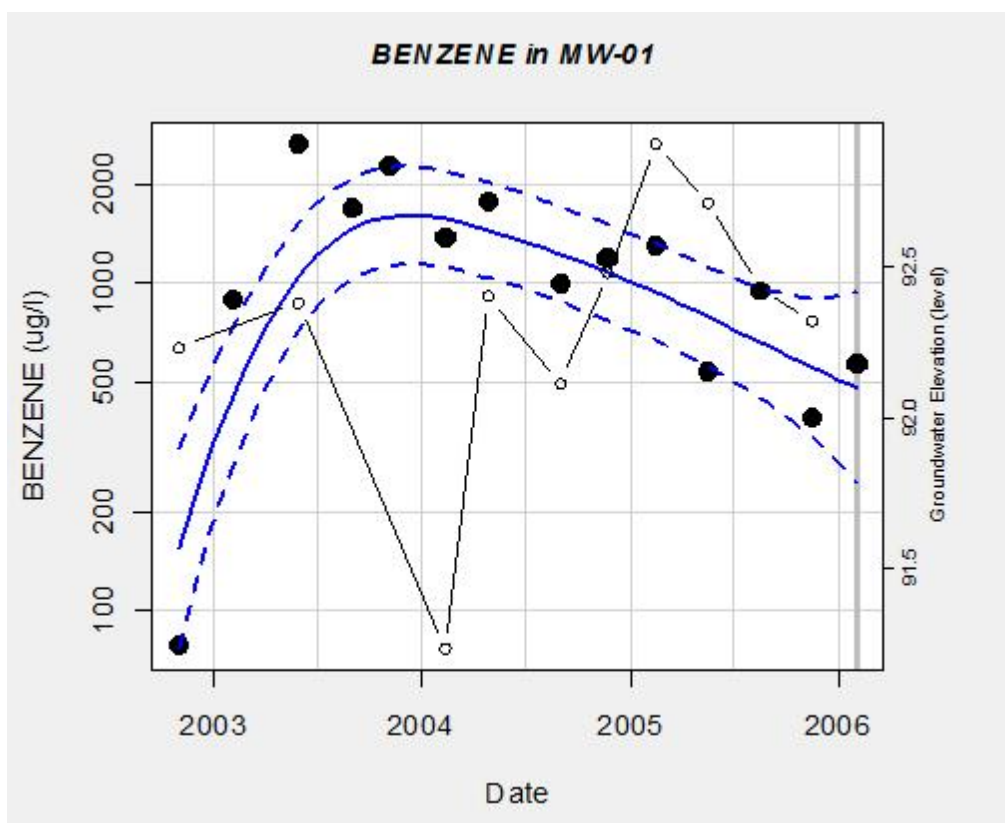The *Visualisation Options* (see Figure 6) portion of the GWSDAT user interface controls well trend plot display options. To switch between different solutes and monitoring wells, simply select from the *Solute* and *Select Monitoring Well* listboxes.

The *Time Series Plot Options* checkbox control includes:

- **Conc. Trend Smoother:** This displays the estimated time series trend in solute concentration using a nonparametric smoother (see Figure 8). The solid blue line displays the estimate of the mean trend level at a particular point in time. The upper and lower dashed blue lines depict a 95% confidence interval around this estimate. This is interpreted as *'one is 95% confident that the actual mean trend level lies within this region'.* The smaller the 95% confidence interval, the more confidence one has in the estimated time series trend. Areas of the trend smoother fit in which the 95% confidence intervals are very large (i.e. very low confidence in the trend smoother fit) are coloured grey instead of blue and are disregarded from the 'Trend' and 'Threshold - Statistical' matrix plot calculations, see Section 7.3.

  The advantage of this nonparametric method is that the trend estimate is not constrained to be monotonic, i.e. the trend can change direction. More details of this nonparametric smoothing algorithm are given in Appendix A.2.4.

- **Conc. Linear Trend:** This displays a traditional linear time series trend estimate (green solid line) together with 95% confidence intervals (green dashed lines) to the log of historical solute concentrations values. This is equivalent to fitting an exponential decay/growth model on a linear scale. The statistical significance of this trend is assessed by means of the well established Mann-Kendall trend test [12].

The Mann-Kendall p-value and the estimated solute concentration half-life is displayed immediately below the main title of the well trend plot. Users should be aware that individual well half-life values should not be used to estimate the plume half-life.

> If the Mann-Kendall p-value is below 0.05, then the estimated trend is deemed statistically significantly different from 0, i.e. there is indeed trend present in the data. A p-value above 0.05 should be interpreted as there is no evidence to suggest that trend is present.

- **Show Legend:** This controls whether to display a legend in the top right hand side of the plot giving a key of the plotting symbols.

- **Scale to Conc. Data:** By default the well trend plot x-axis is scaled such that it spans the sampling dates of all data. The y-axis is scaled to span the current data concentrations and the user-specified trend threshold limit, see Section 7.3. By checking this control the x and y axes are scaled to the span the current combination of well and solute concentration data only.

- **Log Scale:** Controls whether to use a logarithmic or linear scale for the y-axis, i.e. solute concentration values.

- **Overlay GW levels:** Allows the user to overlay the corresponding groundwater level measurements on the well trend plot. The scale is read from the right hand axis. This function is useful for assessing correlations between groundwater levels and solute concentrations.

- **Overlay NAPL Thickness:** Allows the user to overlay the corresponding NAPL thickness level measurements on the well trend plot. The scale is read from the right hand axis. This function is useful for assessing correlations between NAPL thickness and groundwater levels.

As with the GWSDAT spatial plot, left-clicking the plot produces an expanded plot generated in a separate window which can be saved as explained in Section 7.1.

7.3   Trend & Threshold Indicator Matrix

The Trend and Threshold Indicator Matrix plot is a summary of the level and time series trend in solute concentrations at a particular model output interval, see Figure 9. The rows correspond to each well and the columns correspond to the different solutes. The user can select between the options of displaying 'Trend', 'Threshold - Absolute' or 'Threshold - Statistical' by using the radiogroup control called 'Display Table' positioned to the right of this plot (note that the example depicted in Figure 9 is set to display Trend). The 'Trend' and 'Threshold - Statistical' options use the fitted nonparametric time series trend smoother described in Section 7.2.

- **Trend:** This reports the concentration trend for each solute in every well within the selected model output interval. The Trend Threshold Indicator Matrix looks at the instantaneous gradient of the trend smoother (solid blue line) where it crosses the end of the current model output interval (vertical grey line) in the *Well Trend Plot*, see Figure 8. The cells of the Trend Threshold Indicator Matrix are coloured to
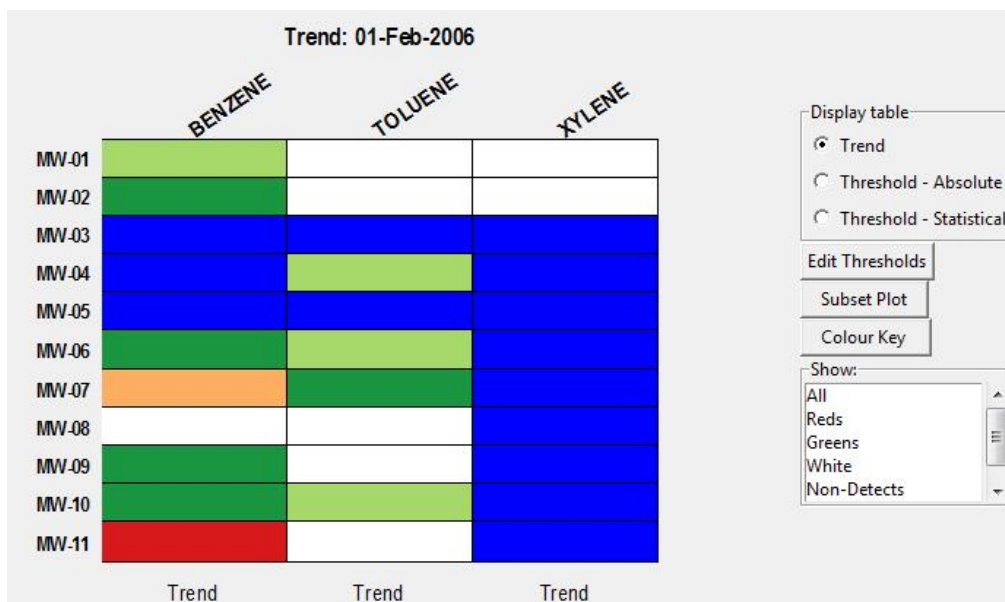

Figure 9: *GWSDAT Trend & Threshold Indicator Matrix*

23

indicate the strength and direction of the current trend. White cells indicate a generally flat trend where the solute concentration is estimated to no more than double or half in the next two years. Light red and light green indicate that solute concentrations will no more than double or half in the next year, respectively. Dark red and Dark green indicate stronger upward and downward trends, respectively. In the event that the trend cannot be calculated, e.g. no data or our confidence in the trend smoother estimate is poor then the corresponding cell is coloured grey. Blue cells represent non-detect data. As an example consider Figures 8 and 9. It can be seen that the trend at the end of the current model output interval (grey vertical line in Figure 8) for 'Benzene' at monitoring well 'MW-01' is decreasing. The corresponding cell in Figure 9 (top left) has been coloured light green to illustrate this.

- **Threshold - Absolute:** This assesses if the **observed** solute concentration values for all well and solute combinations are below a user-specified threshold value (default value of 500 ug/l) within any given model output interval. The threshold value is depicted as a horizontal dashed red line and the end of the current model output interval is depicted by a vertical grey line in the *Well Trend Plot*, see Figure 8. The 'Threshold - Absolute' Indicator Matrix compares the observed concentration values with the threshold value. If any observed concentration values within a model output interval are above the threshold value then the corresponding cell is coloured dark red. If the concentration values within a model output interval are all below the threshold value then the corresponding cell is coloured dark green. In the event that no data exists then the cell is coloured grey. If the current concentration value is classified as non-detect, then the corresponding cell is coloured blue.

  To change the threshold values, press the 'Edit Thresholds' button. This generates a data editor where the user can manually change the threshold limits (in ug/l) for the different solutes. There is no need to save the changes, simply close down the data editor to automatically update threshold limit values. The current solute threshold limit values are displayed directly below the columns of the traffic light plot.

- **Threshold - Statistical:** This assesses if current solute concentration

levels for all well and solute combinations are below a user-specified threshold value **with a statistical degree of confidence**. Again the threshold value is depicted as a horizontal dashed red line and the end of the current model output interval is depicted by a vertical grey line in the *Well Trend Plot*, see Figure 8. The 'Threshold - Statistical' Indicator Matrix looks at the intersection of the end of the current model output interval(vertical grey line) and the trend smoother (solid blue line). If the upper 95% confidence interval (upper dashed blue line) is below the user-specified threshold value, the cell is coloured dark green. If the upper 95% confidence interval is not below the threshold value, the corresponding cell is coloured dark red. In the event that this cannot be calculated, e.g. no data or our confidence in the trend smoother estimate is poor then the cell is coloured grey. If the current concentration value is classified as non-detect, then the corresponding cell is coloured blue.

The 'Subset Plot' button located to the right (see Figure 9) generates a user specified subset of the the Trend and Threshold Indicator Matrix plot in its own separate window. The 'Show:' listbox allows the user to filter the Trend and Threshold Indicator Matrix plot according to the different colours. For example, if the user selects red then the plot will only display the corresponding rows and columns which contain a red entry. These functions are particularly useful when there exists a large number of wells and/or solutes.

The 'Colour Key' button generates a graphic displaying the colour key explained above for the Trend and Threshold Indicator Matrix plot.

7.4 Animations

The animations menu located at the top-left of the GWSDAT user interface (Figure 6) provides three different methods for generating an animated series of solute concentration plots (Section 7.1). These movies give the user an insight into the historical dynamics of site groundwater solute concentrations and distributions. All three animation methods display data for the current solute selected in the 'Solute' listbox (Figure 6) and maintain the current spatial plot options.

- **Contour Animation:** This selection produces a series of spatial plots generated in a separate window from the first to the last model output interval. Once plotting is complete the user can toggle forwards and backwards between the different model output intervals using the **'Page Up'** and **'Page Down'** buttons on the keyboard.

- **Contour Animation -> PPT:** This selection is identical but additionally generates a PowerPoint slidepack of spatial plots for each model output interval. More PowerPoint output options are given in Section 7.5.

- **Contour Animation -> HTML:** This selection generates a html animation page of spatial plots in the user's internet browser. During the short delay whilst html output is being generated, it is recommended that the user not use any of the GWSDAT functionality. To start the animation simply press the play button (>) at the bottom left of the page. The animation will play the spatial plots in time series order to build a movie of the historical solute concentration dynamics. The html animation can be viewed independently of GWSDAT, and hence provides an excellent media for communicating results to individuals who do not have direct access to GWSDAT.

7.5   Report Generation

The Report Generation menu located at the top-left of the GWSDAT user interface (Figure 6) provides options for automatically generating a series of report plots.

- **Current Plot -> PPT:** Left-clicking on any of the plots in the GWSDAT user interface (see Figure 4) results in an expanded plot in its own separate window. This selection copies the current plot into a Microsoft PowerPoint slide.

- **Current Plot -> WORD:** Left-clicking on any of the plots in the GWSDAT user interface (see Figure 4) results in an expanded plot in its own separate window. This selection copies the current plot into a Microsoft Word document.

- **Latest Snapshot:** This selection generates a sequence of plots relating to the latest model output interval. For each solute, the spatial plot for the latest model output interval is generated. This is followed by the latest 'Trend', 'Threshold - Absolute' and 'Threshold - Statistical' Indicator Matrices. The plots are generated one after each other. To toggle between each of the plots use the **'Page Up'** and **'Page Down'** buttons on the keyboard.

- **Latest Snapshot -> PPT:** This selection is identical but additionally generates a PowerPoint slidepack of the plots described immediately above.
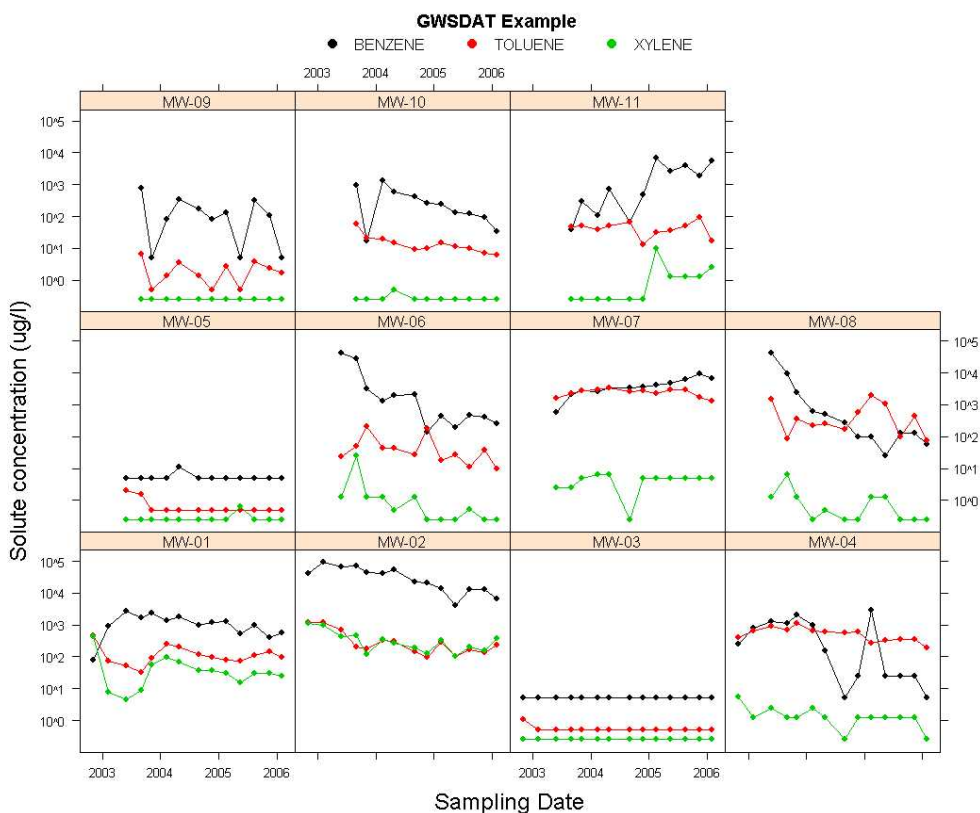

Figure 10: *GWSDAT 'Well Reporting' Report Generation Example.*

- **Well Reporting:** This selection generates a matrix of graphs display-

ing time series solute concentration values on a well by well basis. In contrast to the 'Well Trend Plot' (Section 7.2) it is possible to overlay different solute concentration values within the same graph. Figure 10 is an example 'Well Reporting' output. The colour key at the top identifies each solute and the name of each well is displayed in a banner at the top of each of the individual time series graphs. The 'Well Reporting' output provides a very concise method of visualising a lot of data.

The choice of which solutes and wells to include, together with the choice of whether to use a log-scale for the solute concentration values, is selected by the user from a series of interactive listboxes which are generated when the 'Well Reporting' option is initiated. Multiple solutes and wells can be selected (or deselected) by holding the **'CTRL'** key down whilst selecting, and shift-clicking can be used to select ranges.

If only one solute is selected, then the plotting behaviour is modified such that the detect and non-detect data points are coloured black and orange, respectively. Furthermore, if the 'Display Trend Smoother' is checked in Figure 6 then the corresponding trend smoother with 95% confidence intervals are overlaid as thin black lines onto each graph.

In order to avoid output which is too busy to be comprehensible, the maximum number of wells that can be displayed in the same plot is 16 (i.e. 4X4 matrix). If the number of wells selected exceeds this, then a series of well reporting plots will be generated in the same window. To toggle between these plots use the **'Page Up'** and **'Page Down'** buttons.

- **Well Reporting -> PPT:** This selection is identical but redirects the output directly to Microsoft PowerPoint.

- **GW Well Reporting:** This selection is similar to 'Well Reporting' but additionally allows the overlay of groundwater level and/or NAPL thickness on each plot with the axis being placed on the right hand side. The time series of observed groundwater level is represented by open circles joined by a black solid line. The functionality works by iterating through the selected wells and pasting the Well Trend Plots into a 2X2 matrix. In an identical manner to 'Well Reporting', use
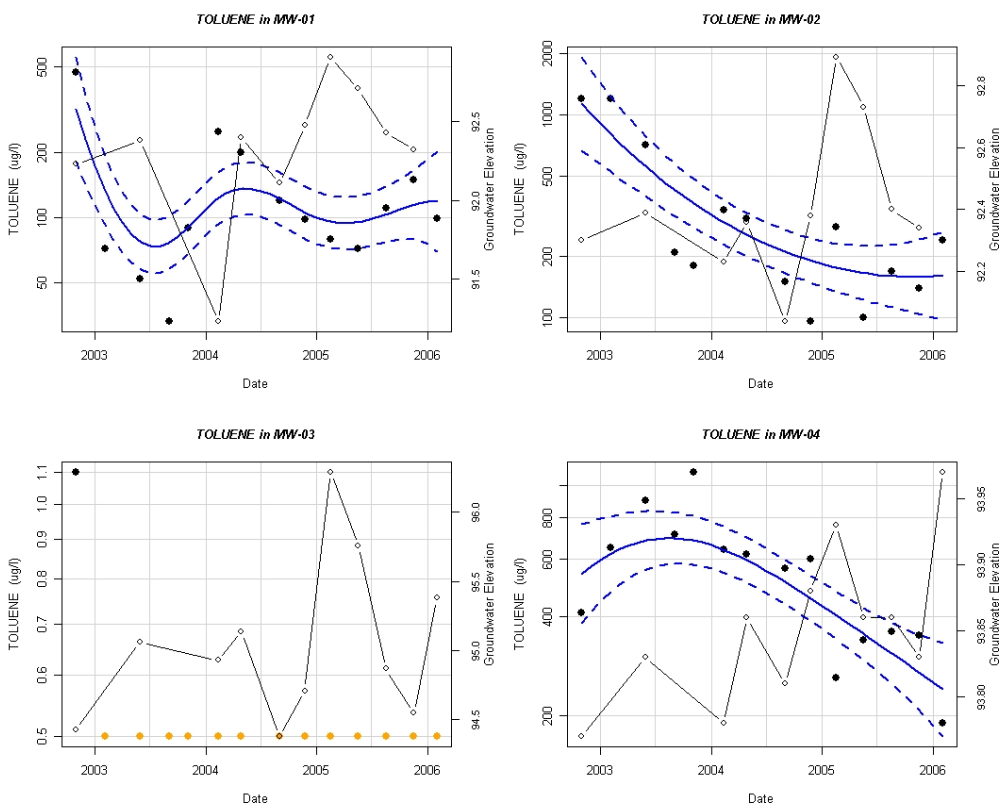
Figure 11: *GWSDAT 'GW Well Reporting' Report Generation Example.*

the **'Page Up'** and **'Page Down'** buttons to toggle between plots if the number of wells selected is greater than 4. The current checkbox options selected in the 'Time Series Plot Options' (see Section 7.2) are respected. Figure 11 is an example 'GW Well Reporting' output.

- **GW Well Reporting -> PPT:** This selection is identical but redirects output directly to Microsoft PowerPoint.

7.6   Spatiotemporal Diagnostics

The spatiotemporal Diagnostics menu located at the top of the GWSDAT user interface (Figure 6) provides options to assess the goodness of fit of the

spatiotemporal solute trend smoother.

- **Output Predictions to Table:** This generates a 'csv' (comma seperated value) Excel spreadsheet with the predictions of spatiotemporal solute smoother together with the actual observed values.

- **Spatiotemporal Diagnostic Plot:** This generates a matrix of graphs displaying the time series of solute concentrations on a well by well basis together with the predictions of the spatiotemporal solute smoother overlaid as a solid light grey line.

- **Spatiotemporal Diagnostic Plot -> PPT:** This selection is identical but redirects output directly to Microsoft PowerPoint.

7.7   Plume Diagnostics

In GWSDAT v2.1 the user can select to automatically delineate a plume based on a user-defined concentration threshold. This functionality, which is accessed via the Plume Diagnostics menu (top right in Figure 6), allows the user to evaluate plume stability based on changes in the plume area, centre of mass and plume mass through time. The plume diagnostics analysis is performed on whichever solute is selected in the solute listbox (see Section 7.1 and Figure 4).

CAUTION: The plume diagnostics functions provide insight into relative changes in plume mass, area and average concentration for a given site and monitoring well network. Plume mass estimates are provided in units of kg/m or kg/ft to reflect this uncertainty. Absolute values of plume mass cannot be calculated because the model does not define the vertical concentration distribution inside the plume. The assumptions made in plume mass estimation are that:

- Vertical variation in plume concentration is averaged across monitoring well screen intervals.

- The monitoring well network samples the full depth and extent of the plume and is also representative of concentration variation across the site.

The assumptions are similar to those presented in the paper by Ricker, [20].

- **Edit Plume Thresholds:** This opens a data editor where the user can manually change the plume boundary threshold concentration limits (in ug/l) for the different solutes. There is no need to save the changes, simply close down the data editor to automatically update the values.

- **Set Ground Porosity:**

  This opens a data editor panel where the user can enter the effective (interconnected) porosity of the subsurface (default=0.25). This value is used in the calculation of plume mass (see Appendix A.2.2).

- **Estimate Plume Boundary:** This function provides a graphic to assist the user in selecting a solute concentration value that will yield a closed plume, which is an essential requirement for the calculation of plume parameters. The plume boundary contour is determined to be closed (i.e. forms a loop) when it does not extend outside the area bounded by monitoring data. The function works by iterating through each time step of the spatiotemporal model and plotting minimum and maximum solute concentrations at the edge of the area bounded by the monitoring data (see Figure 12). The grey shaded area between the minimum and maximum lines defines the region where plume diagnostics can be calculated. The user-defined value of the plume threshold
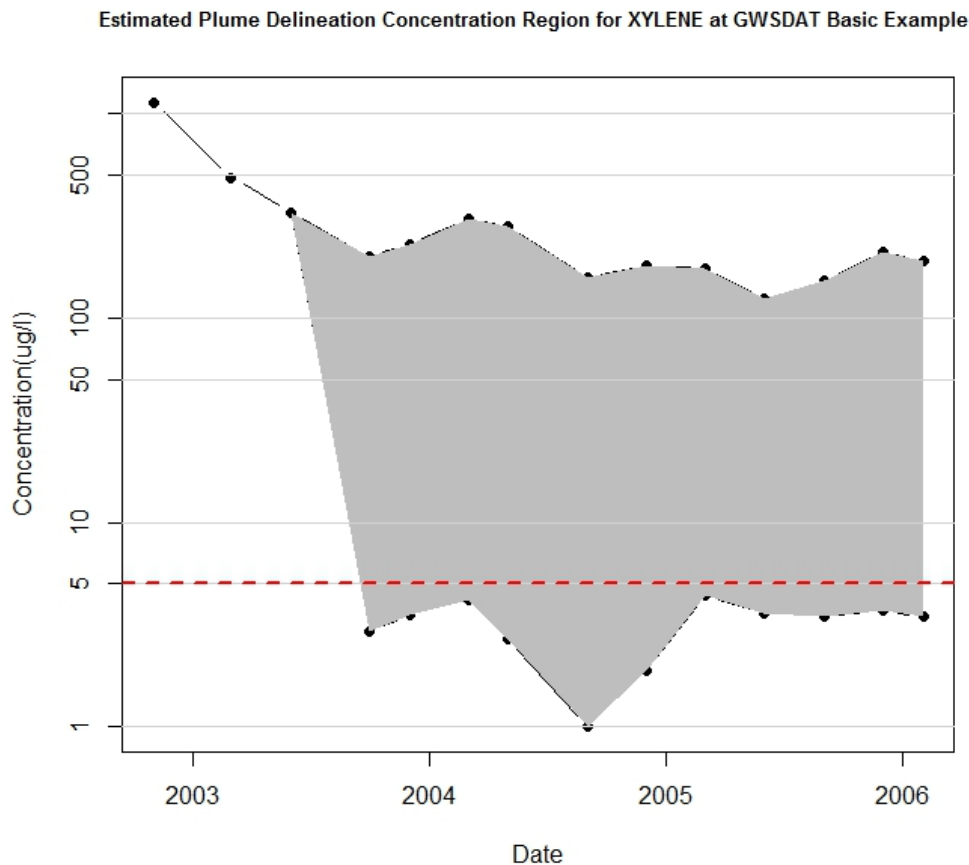
Figure 12: *Example Estimate Plume Boundary plot. The grey shaded area defines a region where plume analytics can be calculated.*

concentration value is displayed as a horizontal red dashed line. The user should select a solute concentration value that is just above the minimum line, for the time period within which calculation of plume parameter is required. It is clear from Figure 12 that a value of 5ug/l works well in this example. The selection of a higher solute concentration value within the grey area, while still producing a closed plume, would exclude part of the plume from the plume mass/ area analysis. The selection of a lower solute concentration value (e.g. 2 ug/l) would,

in this example, greatly reduce the time period over which the plume was closed.

- **Plume Animation:**
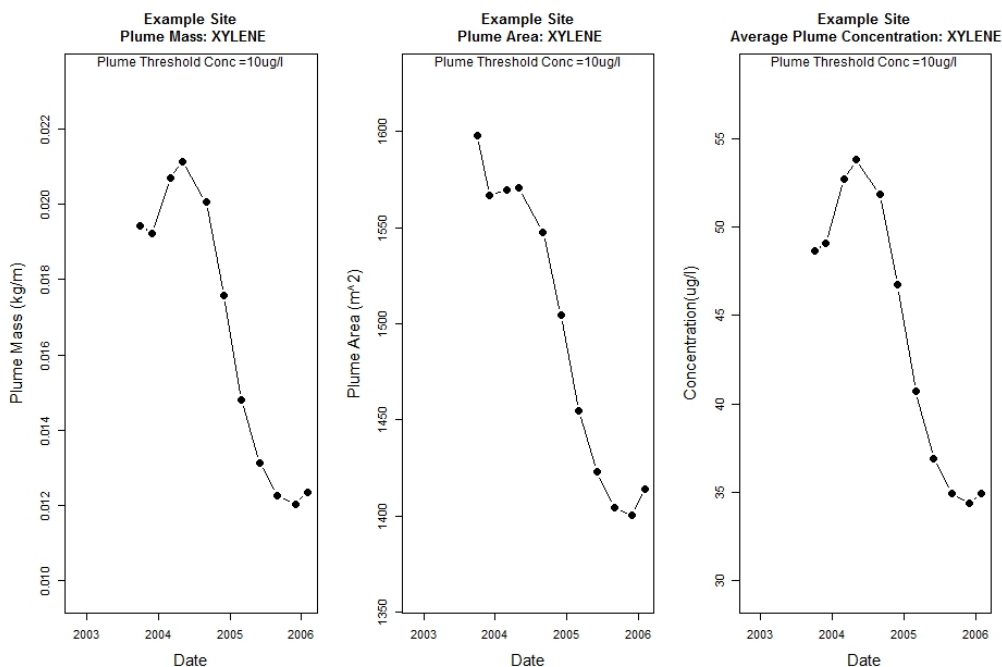
Figure 13: *Example Plume Animation plume diagnostic output.*

This function produces a series of spatial plots incorporating plume diagnostics data from the first to the last spatiotemporal model output interval. In addition, a time series plot of plume mass, plume area and average plume concentration (see Figure 13) is generated. The user can also select to export the complete time series of plume diagnostics data to a Microsoft Excel csv file.

- **Plume Animation -> PPT:** This selection exports a series of spatial plots incorporating the plume diagnostics data directly to Microsoft PowerPoint.

Note: Plume parameters are calculated based on the units of length entered in the CoordUnits field in the Well Coordinates table (Section 3.2): if no units are entered then relative changes in plume parameters are plotted in dimensionless units. Note also that plume mass is calculated per unit of plume depth (e.g. kg/m in Figure 13). To estimate total plume mass the user must multiply this value by the estimated plume thickness (using same units as those entered in the CoordUnits field). Additional information on how plume parameters are calculated can be found in Appendix A.2.2.

7.8   Saving and Loading a GWSDAT Session

The user can save their GWSDAT session at any time by selecting 'Save Session' from the File menu located at the top-left of the GWSDAT user interface, see Figure 6. This is particularly useful for large data sets which take a long time to run a GWSDAT analysis. The GWSDAT session is saved with a '.GWSDATData' file extension. All the current display settings, e.g. Colour Scheme, Plot Options, Current Solute, are saved. This '.GWSDAT-Data' file can be forwarded to another user and viewed once again exactly as it was saved.

To open a previously saved session, select 'Load Session' from the GWSDAT Excel Menu, see Figure 2. Note: Occasionally the GWSDAT user interface appears hidden underneath any windows that are currently open. The user can locate GWSDAT from the windows taskbar at the bottom of the screen.

References

[1] Raftery A., Madigan D., and Hoeting J. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997.

[2] N. Ahuja and B. J. Schacter. *Pattern Models.* John Wiley & Sons, New York, 1983.

[3] Julia A. Aziz, C. J. Newell, Meng Ling, Hanadi S. Rifai, and J. R. Gonzales. Maros: A decision support system for optimizing monitoring plans. *Ground Water*, 41(3), 2003.

[4] A. W. Bowman and E. Crawford. *R package **rpanel**: simple control panels (version 1.0-5).* University of Glasgow, UK. R package, www.stats.gla.ac.uk/~adrian/rpanel.

[5] Adrian W. Bowman and Adelchi Azzalini. *sm: Smoothing methods for nonparametric regression and density estimation.* R package, http://www.stats.gla.ac.uk/~adrian/sm.

[6] A.W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis:the Kernel Approach with S-Plus Illustrations.* Oxford University Press, Oxford, 1997.

[7] Denison D., Holmes C., Mallick B., and Smith A. *Bayesian Methods for Nonlinear Classification & Regression.* John Wiley & Sons, New York, 2002.

[8] Paul H. C. Eilers, Dcmr Milieudienst Rijnmond, and Brian D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–121, 1996.

[9] C.M. Hurvich, J.S. Simonoff, and C.-L. Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society, Series B*, 60:271–293, 1998.

[10] Duncan Temple Lang. *RDCOMClient: R-DCOM client.* R package, http://www.omegahat.org/RDCOMClient.

[11] Nicholas J. Lewin-Koh and Roger Bivand. *maptools: Tools for reading and handling spatial objects*. R package, http://cran.r-project.org/web/packages/maptools/index.html.

[12] H.B. Mann. Nonparametric tests against trend. *Econometrica*, 13:245–259, 1945.

[13] A.I. McLeod. *Kendall: Kendall rank correlation and Mann-Kendall trend test*. R package, http://www.stats.uwo.ca/faculty/aim.

[14] Wood S. N. *Generalized Additive Models - An Introduction with R*. Chapman & Hall/CRC, 2006.

[15] Wood S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:3–36, 2011.

[16] A.A. Oloufa. Triangulation applications in volume calculation. *Journal of Computing in Civil Engineering*, 5(1):103–119, 1991.

[17] Eilers P. and Marx B. *Generalized Linear Models with P-Splines in Advances in GLIM and Statistical Modelling (L.Fahrmeir et al. eds.)*. Springer, New York, 1992.

[18] Edzer J. Pebesma and Roger S. Bivand. Classes and methods for spatial data in R. *R News*, 5(2):9–13, November 2005.

[19] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0, http://www.R-project.org.

[20] Joseph A. Ricker. A practical method to evaluate ground water contaminant plume stability. *Ground Water Monitoring and Remediation*, 28(4):85–94, 2008.

[21] Barry Rowlingson, Peter Diggle, adapted, packaged for R by Roger Bivand, pcp functions by Giovanni Petris, and goodness of fit by Stephen Eglen. *splancs: Spatial and Space-Time Point Pattern Analysis*. R package, http://www.maths.lancs.ac.uk/~rowlings/Splancs/.

[22] Deepayan Sarkar. *lattice: Lattice Graphics*, 2008. R package, http://cran.r-project.org/web/packages/lattice/index.html.

[23] Luke Tierney. *tkrplot: TK Rplot*. R package, http://cran.r-project.org/web/packages/tkrplot/index.html.

[24] Rolf Turner. *deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tesselation.* R package, http://www.math.unb.ca/~rolf/.

[25] Yihui Xie. *animation: Demonstrate Animations in Statistics.* R package, http://animation.yihui.name.

[26] Achim Zeileis and Gabor Grothendieck. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27, 2005.

A   Appendices

A.1   GWSDAT Software Architecture

The user entry point and data input platform to GWSDAT is Microsoft Excel by means of a custom built Excel Add-in application. The statistical engine used to perform geostatistical modelling and display graphical output is the open source statistical programming language called R [19].

Members of the R-community contribute statistical routines and functionality to this collaborative project by means of packages which can be downloaded from http://cran.r-project.org/web/packages/. The greatest majority of these packages are licensed under the GNU General Public License, http://www.gnu.org/copyleft/gpl.html. The authors of GWSDAT would like to gratefully acknowledge and thank the authors of the following packages which GWSDAT makes use of:

- 'animation' [25], 'lattice' [22] , 'rpanel' [4] and 'tkrplot' [23] - used for graphical output in GWSDAT.

- 'deldir' [24], 'sp' [18], 'splancs' [21] and 'maptools' [11] - to provide spatial statistics routines.

- 'Kendall' [13], 'sm' [5] and 'zoo' - [26] to provide time series and trend detection routines.

- 'RDCOMClient' [10], to provide report generation functionality to Microsoft Word and PowerPoint.

A.2   The Mathematics behind GWSDAT

A.2.1   Spatiotemporal Solute Concentration Smoother

The spatiotemporal solute concentration smoother is fitted using a non parametric regression technique known as Penalised Splines (P-Splines). It is beyond the scope of this text to give a full and detailed explanation of this technique here. However, the following outlines some of the most important aspects for the purposes of GWSDAT. For a more detailed explanation the reader is referred to [17], [8].

Let y_i be the solute concentration at $\boldsymbol{x_i} = (x_{i1}, x_{i2}, x_{i3})$ where x_{i1} and x_{i2} stand for the spatial coordinates of the well and x_{i3} represents the corresponding time point for the *i-th* observation with $i = 1, \ldots, n$. We start by modelling the solute concentration as

$$y_i = \sum_{j=1}^{m} b_j(\boldsymbol{x_i})\alpha_j + \epsilon_i \tag{1}$$

where the b_j, $j = 1, \ldots, m$ are m functions, known as *basis functions*, generally second or third order polynomials. The first term in equation (1) is a linear combination of the basis functions b_j, each evaluated at $\boldsymbol{x_i}$, and aims at capturing the deterministic part of the *i-th* observation, generally known as "signal"; the second term, ϵ_i, accounts for the variability in the measurement due to randomness and is usually termed as "noise". The behaviour of ϵ_i is described in terms of a probabilistic model; such a model guarantees that the value of ϵ_i fluctuates around zero conveying the idea that we do not expect to make any systematic error in the measurement. This model also comprises the notion that the expected spread of ϵ_i is given by σ^2, the *variance* of the random component ϵ_i. By using the matrix notation

$$\boldsymbol{B}(\boldsymbol{x}) = \begin{pmatrix} b_1(x_1) & \cdots & b_j(x_1) & \cdots & b_m(x_1) \\ b_1(x_2) & \cdots & b_j(x_2) & \cdots & b_m(x_2) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_1(x_i) & \cdots & b_j(x_i) & \cdots & b_m(x_i) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_1(x_n) & \cdots & b_j(x_n) & \cdots & b_m(x_n) \end{pmatrix}$$

equation (1) can be written in a more compact fashion as $\boldsymbol{y} = \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha} + \boldsymbol{\epsilon}$. Because, as mentioned earlier, we expect the ϵ_i's to oscillate around zero, a sensible choice for the regression parameters $\boldsymbol{\alpha}$ is the one that minimises the norm of the vector $\boldsymbol{\epsilon}$ defined as $\mathrm{S}(\boldsymbol{\alpha}) = \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha}\|^2$. A large value of basis functions is generally chosen to allow the model to capture most of the signal. The downside of this approach is that it tends also to overfit, that is to fit the noise in the observations, with the consequent loss of smoothness. To overcome this hurdle, the objective function modified with the addition of a term that penalises the lack of smoothness of the fit.

The objective function now takes the form $S(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha}\|^2 + \lambda\|D\alpha\|^2$ where λ is a non-negative smoothing parameter and D is the $(m-2) \times m$ matrix

$$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}$$

The additional term in the objective function

$$\|D\alpha\|^2 = (\alpha_1 - 2\alpha_2 + \alpha_3)^2 + \ldots + (\alpha_{m-2} - 2\alpha_{m-1} + \alpha_m)^2$$

controls the smoothness of the fit by applying penalties over adjacent coefficients. By minimising the new objective function for a given value of λ, we obtain the least squares estimator of the parameters $\hat{\boldsymbol{\alpha}} = (\boldsymbol{B}'\boldsymbol{B} + \lambda\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y}$. Consequently, the fitted values are given by:

$$\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{\alpha}} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + \lambda\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \qquad (2)$$

When $\lambda = 0$, the expression for the estimator of the parameters $\hat{\boldsymbol{\alpha}}$ boils down to the classical solution in linear models theory. As $\lambda \to \infty$, the fitted function tends to a linear function. Figure 14 shows the effect of penalisation: it forces the coefficients to yield a smooth pattern. The fitting process of a function using B-Splines is pictured with and without penalisation, together with the basis functions (the columns of the $\boldsymbol{B}$ matrix). The left plot results from not penalising ($\lambda = 0$) the term in the objective function that accounts for the smoothness; it can be noticed that it yields a rather wiggly regression function. In the right plot, a suitable choice for λ constrains the optimisation method to find values for the coefficients $\hat{\alpha}$ which result in a smoother regression curve.

Prior to fitting the regression coefficients $\boldsymbol{\alpha}$ the observed solute concentration values are natural log transformed. This avoids the possibility of predicting
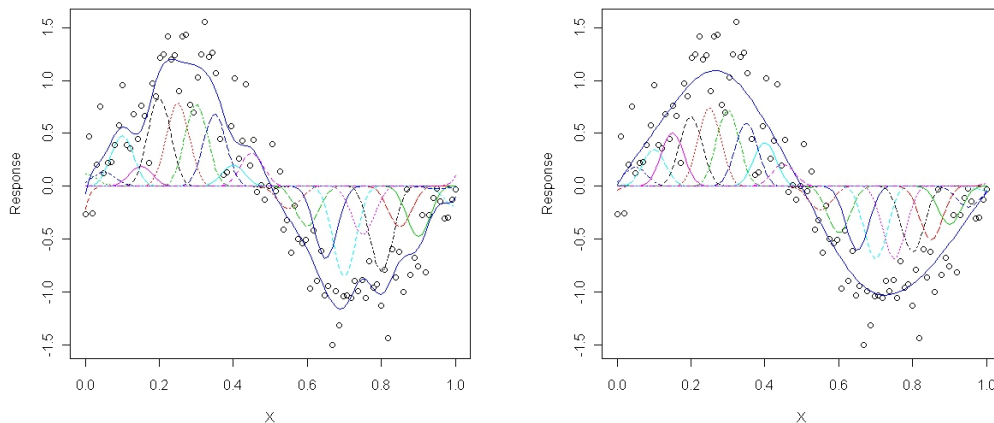
Figure 14: *Curve based on 20 nodes in the basis, without penalisation (left), with penalisation (right)*

negative concentration values and also helps the model cope with data which often spans several orders of magnitude. Furthermore, the uncertainty in the measured concentrations can reasonably be expected to be proportional to the magnitude of the value, e.g. the uncertainty around a measured value of 10ug/l would be expected to be very much less than the uncertainty surrounding a measured value of 10000ug/l. The natural log transformation stabilises the variance.

The choice of the penalisation parameter λ is a crucial matter as a too small value would result in "overfitting", i.e. capturing also the noise, whereas an extremely large value would lead to underfitting (flat predictive function as a result of loss of signal). Several criteria have been traditionally proposed (see [9], [14]) but we tackled the issue by means of a probabilistic framework known as *Bayesian* (after Rev. Thomas Bayes whose rule represents the pivotal theorem in this approach, see [7], [1], [15]).

Under this paradigm, λ is not considered to be a fixed unknown quantity to be estimated but rather a random variable whose value may vary within a given range. This behaviour is described in probabilistic terms which assign a measure of confidence or *probability* to each of the values λ may take on.

The Bayesian framework allows to compute the probability that the random

variable λ may take a particular value, conditional to the fact that $\boldsymbol{y}$ has already been observed. This probability, indicated as $f(\lambda|\boldsymbol{y})$, is known as the *posterior distribution* of λ.

Bayes' rule states that $f(\lambda|\boldsymbol{y}) \propto f(\boldsymbol{y}|\lambda)f(\lambda)$ where $\propto$ stands for "proportional to". $f(\boldsymbol{y}|\lambda)$ is known as the *likelihood function* and expresses the conditional probability of observing data $\boldsymbol{y}$, given that the true value of the parameter is λ; $f(\lambda)$ is known as the *prior distribution* of the random variable λ and comprises our prior beliefs on its uncertainty.

The optimal value of λ is the one that maximises the posterior distribution and is computed using numerical methods.

A.2.2 Plume Diagnostics

GWSDAT calculates plume diagnostic quantities from the predictions of the spatiotemporal solute concentration smoother (see Appendix A.2.1). In common to [3, 20], numerical methods are employed to integrate out the plume diagnostic quantities. For a given model time step a fine spatial mesh grid of predictions is generated. The plume boundary region D, for a given plume threshold concentration value, is calculated using the R function *contourLines* which is included in the base distribution of the R programming language [19]. The plume area, A, is defined as

$$A = \iint\limits_{D} dxdy \tag{3}$$

where where x and y are the spatial coordinates and is calculated numerically using the *areapl* function from the R package 'splancs' [21]. The average plume concentration is defined as

$$\frac{1}{A} \iint\limits_{D} \hat{s}_t\left(x, y\right) dxdy \tag{4}$$

where $\hat{s}_t\left(x, y\right)$ represents the predictions of the spatiotemporal solute concentration smoother evaluated at time t. This integral (and all subsequent integrals in this section) is calculated numerically using a method described in [16]. A Delaunay triangulation is performed using the R package 'deldir'

[24] on the spatial mesh grid of predictions within the plume boundary, D. The integral is numerically approximated by summing up the individual volumes under each prism formed.

Plume mass is calculated from the scaled product of plume area and average concentration. The scaling factor encompasses the user specified value of ground porosity (see Section 7.7) and appropriate scaling values for mapping together the volumetric concentration units (e.g. ug/l) with the length scale (see **CoordUnits** in Section 3.2) of the Well coordinates (e.g. metres or feet). The plume mass is calculated on a mass per unit aquifer depth basis (e.g. kg/m). To calculate the total plume mass the user must multiply this value by the aquifer depth.

The plume center of mass (x, y) is defined as the mean location of the concentration distribution within the plume boundary region D. The x-coordinate of the plume Centre of Mass is evaluated by numerical calculation of

$$
X_c = \frac{\iint\limits_D x \hat{s}_t\left(x, y\right) dxdy}{\iint\limits_D \hat{s}_t\left(x, y\right) dxdy} \tag{5}
$$

where x and y are the spatial coordinates and $\hat{s}_t\left(x, y\right)$ represents the predictions of the spatiotemporal solute concentration smoother evaluated at time t. In a completely analogous manner the y-coordinate of the plume Centre of Mass is evaluated as follows:

$$
Y_c = \frac{\iint\limits_D y \hat{s}_t\left(x, y\right) dxdy}{\iint\limits_D \hat{s}_t\left(x, y\right) dxdy} \tag{6}
$$

In the event that multiple plumes are detected then the above quantities are calculated for each individual plume and aggregated together. The individual plume areas and masses are summed to calculate the total over all plumes. The aggregate average plume concentration and aggregate plume centre of mass is calculated by taking a weighted average of the individual quantities.

A.2.3 Groundwater Flow Calculation

For a given model output interval the Groundwater (GW) flow strength and direction are estimated using available GW level and well coordinates data. The model is based on the simple premise that local GW flow will follow the local direction of steepest descent (hydraulic gradient).

For a given well, a linear plane is fitted to the local GW level data:

$$L_i = a + bx_i + cy_i + \epsilon_i \tag{7}$$

where L_i represents the GW level at location (x_i, y_i). Local data is defined as the neighbouring wells as given by a Delauney triangulation (http://en.wikipedia.org/wiki/Delaunay_triangulation, [2]) of the monitoring well locations. The gradient of this linear surface in both x and y directions is given by the coefficients b and c. Estimated direction of flow is given by:

$$\theta = \tan^{-1}\left(\frac{c}{b}\right) \tag{8}$$

and the relative hydraulic gradient (a measure of relative flow velocity) is given by

$$R = \sqrt{b^2 + c^2} \tag{9}$$

For any given model output interval this algorithm is applied to each and every well where a GW level has been recorded.

A.2.4 Well Trend Plot Smoother

The well trend plot smoother is fitted using a nonparametric method called local linear regression. This involves solving locally the least squares problem:

$$\min_{\alpha,\beta} \sum_i^n \left\{y_i - \alpha - \beta(x_i - x)\right\}^2 w(x_i - x; h) \tag{10}$$

where $w(x_i - x; h)$ is called the kernel function. A normally-distributed probability density function with standard deviation h is used as the kernel. h is also called smoothing parameter that controls the width of the kernel

function, and hence the degree of smoothing applied to the data (the higher the value of h, the smoother the estimates). Within GWSDAT, local linear regression is deployed using the R package 'sm' [5, 6] and the bandwidth is selected using the method published in [9].

A.3 Converting a CAD drawing to a Shapefile

System requirements: ArcGIS comprising ArcMap, ArcEditor, ArcCatalog

1. Open ArcCatalog from the Start Menu ('Start' -> 'All Programs' -> 'ArcGIS' -> 'ArcCatalog')

2. In ArcCatalog navigate to ArcMap (globe with magnifying glass icon)

3. When ArcMap opens a screen will pop-up. Select 'A New Empty Map' then click 'OK'.

4. Go to 'File' -> 'Add Data' (positive sign with yellow triangle underneath) -> Select site CAD drawing saved as a '.dxf' file -> 'Add'.

5. Click on the '+' symbol to expand the sub-layers of the dxf file (e.g. 'Polyline', 'Polygon', 'Multipatch', 'Point').

6. Right click on required layer (e.g. Polyline or an edited & exported shapefile) to open the drop down menu.

7. On drop down menu select 'Data' -> 'Export Data'.

8. On 'Export Data' pop-up menu choose 'Select All Features' + 'This Layers Source Data' and select the folder you wish to save the shapefile into, then click 'OK'.

9. Click 'Yes' to add the exported data as a new layer.

10. Repeat steps 6-9 to convert all the layers required to produce the base-map in GWSDAT into shapefiles.

11. Add the shapefiles into GWSDAT one by one (see Section 3.3) to produce the complete base-map image.

The next section details how to edit layers in ArcMap after their conversion to shapefiles, prior to upload into GWSDAT (useful for removing gridlines etc)

A Uncheck the CAD layer used to produce shapefile to remove image from view window

B Ensure exported Shapefile is selected and visible in view window

C Click 'Start Editing' on the 'Editor' toolbar above the map.

D Use the arrow pointer to select lines and press delete on the keyboard to remove from drawing. (Select 'UnDo' from 'Edit' Toolbar in case of errors)

E 'Editor' -> 'Stop Editing'. Click 'Yes' to save edits.

F Repeat data export as detailed in steps 6-9 above and re-save as new shapefile.